Preface of DD24 Book of Proceedings

This volume contains a selection of 53 papers submitted to the 24th International Conference on Domain Decomposition Methods, hosted by the University of Bergen in cooperation with the Western Norway University of Applied Sciences, and held in Spitsbergen at Svalbard, Norway, February 6–10, 2017.

Background of the Conference Series

With its first meeting in Paris in 1987, the International Conference on Domain Decomposition Methods has been held in 15 countries in Asia, Europe, and North America, and now for the first time north of 78^0 in the kingdom of the Polar Bears. The conference is held at roughly 18-months intervals. A complete list of 25 meetings appears below.

Domain decomposition is often seen as a form of divide-and-conquer for mathematical problems posed over a physical domain, reducing a large problem into a collection of smaller problems, each of which is much easier to solve computationally than the undecomposed problem, and most or all of which can be solved independently and concurrently, and then solving them iteratively in a consistent way. Much of the theoretical interest in domain decomposition algorithms lies in ensuring that the number of iterations required to converge is very small. Domain decomposition algorithms can be tailored to the properties of the physical system as reflected in the mathematical operators, to the number of processors available, and even to specific architectural parameters, such as cache size and the ratio of memory bandwidth to floating point processing rate, proving it to be an ideal paradigm for large-scale simulation on advanced architecture computers.

The principle technical content of the conference has always been mathematical, but the principle motivation has been to make efficient use of distributed memory computers for complex applications arising in science and engineering. While research in domain decomposition methods is presented at numerous venues, the International Conference on Domain Decomposition Methods is the only regularly

occurring international forum dedicated to interdisciplinary technical interactions between theoreticians and practitioners working in the development, analysis, software implementation, and application of domain decomposition methods.

As we approach the dawn of exascale computing, where we will command 1018 floating point operations per second, clearly efficient and mathematically well-founded methods for the solution of large-scale systems become more and more important-as does their sound realization in the framework of modern HPC architectures. In fact, the massive parallelism, which makes exascale computing possible, requires the development of new solutions methods, which are capable of efficiently exploiting this large number of cores as well as the connected hierarchies for memory access. Ongoing developments such as parallelization in time asynchronous iterative methods, or nonlinear domain decomposition methods show that this massive parallelism does not only demand for new solution and discretization methods, but also allowsto fosterthe development of new approaches.

Here is a list of the 25 first conferences on Domain Decomposition:

1. Paris, France, January 7-9, 1987
2. Los Angeles, USA, January 14-16, 1988
3. Houston, USA, March 20-22, 1989
4. Moscow, USSR, May 21-25, 1990
5. Norfolk, USA, May 6-8, 1991
6. Como, Italy, June 15-19, 1992
7. University Park, Pennsylvania, USA, October 27-30□, 1993
8. Beijing, China, May 16-19, 1995
9. Ullensvang, Norway, June 3-8, 1996
10. Boulder, USA, August 10-14, 1997
11. Greenwich, UK, July 20-24, 1998
12. Chiba, Japan, October 25-20, 1999
13. Lyon, France, October 9-12, 2000
14. Cocoyoc, Mexico, January 6-11, 2002
15. Berlin, Germany, July 21-25, 2003
16. New York, USA, January 12-15, 2005
17. St. Wolfgang-Strobl, Austria, July 3-7, 2006
18. Jerusalem, Israel, January 12-17, 2008
19. Zhangjiajie, China, August 17-22, 2009
20. San Diego, California, USA, February 7-11, 2011
21. Rennes, France, June 25-29, 2012
22. Lugano, Switzerland, September 16-20, 2013
23. Jeju Island, Korea, July 6-10, 2015
24. Spitsbergen, Svalbard, Norway, February 6-10, 2017
25. St. John's, Newfoundland, Canada, July 23-27, 2018

International Scientific Committee on Domain Decomposition Methods

- Petter Bjørstad, University of Bergen, Norway
- Susanne Brenner, Louisiana State University, USA
- Xiao-Chuan Cai, CU Boulder, USA
- Martin Gander, University of Geneva, Switzerland
- Laurence Halpern, University Paris 13, France
- David Keyes, KAUST, Saudi Arabia
- Hyea Hyun Kim, Kyung Hee University, Korea
- Axel Klawonn, Universität zu Köln, Germany
- Ralf Kornhuber, Freie Universität Berlin, Germany
- Ulrich Langer, University of Linz, Austria
- Alfio Quarteroni, EPFL, Switzerland
- Olof Widlund, Courant Institute, USA
- Jinchao Xu, Penn State, USA
- Jun Zou, Chinese University of Hong Kong, Hong Kong

About the 24th. Conference

The twenty-fourth International Conference on Domain Decomposition Methods had close to 200 participants from about 30 different countries. The conference contained 12 invited presentation selected by the International Scientific Committee, fostering both experienced and younger scientists, 19 minisymposia around specific topics, 3 contributed sessions, and a poster session. The present proceedings contain a selection of 53 papers grouped into three separate groups: 8 plenary papers, 41 minisymposia papers, and 4 contributed papers.

Sponsoring Organizations

- Department of Informatics, University of Bergen
- Simula Research Laboratory
- Faculty of Engineering and Science, WNUAS
- SparebankenVest Bergen
- The Research Council of Norway

Local Organizing/Program Committee Members

- Liv Rebecca Aae (Insititute for Informatics, University of Bergen)
- Petter E. Bjørstad (Insititute for Informatics, University of Bergen)
- Sushmita Gupta (Insititute for Informatics, University of Bergen)
- Talal Rahman (Faculty of Engineering and Science, WNUAS)

Plenary Presentations

- *An additive Schwarz analysis of multiplicative Schwarz methods*, Sue Brenner (Louisiana State University, USA)
- *On nonlinear adaptivity with heterogeneity*, Jed Brown (University of Colorado Boulder, USA)
- *Overlapping methods for high-contrast multiscale problems*, Juan Carlos Galvis-Arrieta (Universidad Nacional de Colombia)
- *Domain Decomposition for high frequency Helmholtz problems*, Ivan Graham (University of Bath, UK)
- *PDE based mesh generation: domain decomposition approaches*, Ron Haynes (Memorial University, Canada)
- *Robust Preconditioners for Coupled Problems*, Xiaozhe Hu (Tufts University, USA)
- *Modeling and discretization of thin inclusions for flow in deformable porous media*, Jan Nordbotten (University of Bergen, Norway)
- *Domain decomposition based methods for multiphysics problems*, Alfio Quarteroni (Ecole polytechnique fédérale de Lausanne, Switzerland)
- *Recent advances on adaptive multilevel BDDC methods for div- and curl-conforming spaces*, Stefano Zampini (KAUST, Saudi Arabia)
- *Communication avoiding iterative solvers and preconditioners*, Laura Grigori (Inria Paris and Laboratoire J.L. Lions UPMC, France)
- *Impact of high abstraction/high performance finite element software in biomedical computing*, Marie Rognes (Simula Research Laboratory, Norway)
- *Scalable multilevel preconditioners for cardiac electro-mechanics*, Simone Scacchi (University of Milano, Italy)

Acknowledgements

The organizers would like to thank all the participants for their enthusiasm and carefully prepared contributions that made this meeting a very successful event, both scientifically and socially. A warm thank also to our sponsors that made the budget come together. Also, our deep appreciation for the people of Longyearbyen. They were helpful in all respect and allocated the city movie theater to our plenary talks,

thus all movies were cancelled for an entire week. We also would like to acknowledge the efforts of our excursion partner, the Svalbard Adventure Group. Unfortunately, our conference experienced the harsh reality of global warming, with temperatures about 25 degress warmer than normal.

Bergen, May 2018.

Petter E. Bjørstad
Universitetet i Bergen, Norway

Susanne Brenner
Louisiana State University, USA

Laurence Halpern
University Paris 13, France

Hyea Hyun Kim
Kyung Hee University, Korea

Ralf Kornhuber
Freie Universität Berlin, Germany

Talal Rahman
Western Norway University of Appl. Sci.

Contents

Part I
Plenary Talks (PT)

Contents

Robust Block Preconditioners for Biot's Model

James H. Adler, Francisco J. Gaspar, Xiaozhe Hu, Carmen Rodrigo, and Ludmil T. Zikatanov

Abstract In this paper, we design robust and efficient block preconditioners for the two-field formulation of Biot's consolidation model, where stabilized finite-element discretizations are used. The proposed block preconditioners are based on the well-posedness of the discrete linear systems. Block diagonal (norm-equivalent) and block triangular preconditioners are developed, and we prove that these methods are robust with respect to both physical and discretization parameters. Numerical results are presented to support the theoretical results.

1 Introduction

In this work, we study the quasi-static Biot's model for soil consolidation. For linearly elastic, homogeneous, and isotropic porous medium, saturated by an incompressible Newtonian fluid, the consolidation is modeled by the following system of

J. H. Adler
Department of Mathematics, Tufts University, Medford, Massachusetts 02155, USA, e-mail: James.Adler@tufts.edu

F. J. Gaspar
Departamento de Matemática Aplicada, Universidad de Zaragoza, Zaragoza, Spain, e-mail: fjgaspar@unizar.es

X. Hu
Department of Mathematics, Tufts University, Medford, Massachusetts 02155, USA, e-mail: Xiaozhe.Hu@tufts.edu

C. Rodrigo
Departamento de Matemática Aplicada, Universidad de Zaragoza, Zaragoza, Spain, e-mail: carmenr@unizar.es

L. T. Zikatanov
Department of Mathematics, Penn State, University Park, Pennsylvania, 16802, USA, e-mail: ludmil@psu.edu

partial differential equations (see [8]):

$$\text{equilibrium equation:} \quad -\operatorname{div}\boldsymbol{\sigma}' + \alpha\nabla p = \boldsymbol{g}, \quad \text{in}\,\Omega, \tag{1}$$

$$\text{constitutive equation:} \quad \boldsymbol{\sigma}' = 2\mu\boldsymbol{\varepsilon}(\boldsymbol{u}) + \lambda \operatorname{div}(\boldsymbol{u})I, \quad \text{in}\,\Omega, \tag{2}$$

$$\text{compatibility condition:} \quad \boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^t), \quad \text{in}\,\Omega, \tag{3}$$

$$\text{Darcy's law:} \quad \boldsymbol{w} = -K\nabla p, \quad \text{in}\,\Omega, \tag{4}$$

$$\text{continuity equation:} \quad -\alpha \operatorname{div}\partial_t\boldsymbol{u} - \operatorname{div}\boldsymbol{w} = f, \quad \text{in}\,\Omega, \tag{5}$$

where λ and μ are the Lamé coefficients, α is the Biot-Willis constant (assumed to be one without loss of generality), K is the hydraulic conductivity (ratio of the permeability of the porous medium to the viscosity of the fluid), I is the identity tensor, $\boldsymbol{u}$ is the displacement vector, p is the pore pressure, $\boldsymbol{\sigma}'$ and $\boldsymbol{\varepsilon}$ are the effective stress and strain tensors for the porous medium, and $\boldsymbol{w}$ is the percolation velocity of the fluid relative to the soil. The right-hand-side term, $\boldsymbol{g}$, is the density of applied body forces and the source term f represents a forced fluid extraction or injection process. Here, we consider a bounded open subset, $\Omega \subset \mathbb{R}^d$, $d = 2,3$ with regular boundary Γ. This system is often subject to the following set of boundary conditions:

$$\begin{aligned} &p = 0, \quad \text{for} \quad x \in \overline{\Gamma}_t, \quad \boldsymbol{\sigma}'\boldsymbol{n} = \boldsymbol{0}, \qquad \text{for} \quad x \in \Gamma_t, \\ &\boldsymbol{u} = \boldsymbol{0}, \quad \text{for} \quad x \in \overline{\Gamma}_c, \quad \boldsymbol{w}\cdot\boldsymbol{n} = 0, \quad \text{for} \quad x \in \Gamma_c, \end{aligned}$$

where $\boldsymbol{n}$ is the outward unit normal to the boundary, $\overline{\Gamma} = \overline{\Gamma}_t \cup \overline{\Gamma}_c$, with Γ_t and Γ_c being open (with respect to Γ) subsets of Γ with nonzero measure. These, or similar conditions, along with appropriate initial conditions for the displacement and pressure, complete the system.

Suitable discretizations yield a large-scale linear system of equations to solve at each time step, which are typically ill-conditioned and difficult to solve in practice. Thus, iterative solution techniques are usually considered. For the coupled poromechanics equations considered here, there are two typical approaches: fully-coupled or monolithic methods and iterative coupling methods. Monolithic techniques solve the resulting linear system simultaneously for all the involved unknowns. In this context, efficient preconditioners are developed to accelerate the convergence of Krylov subspace methods and special smoothers are designed in a multigrid framework. Examples of this approach for poromechanics are found in [7, 14, 16, 25, 17, 23, 5] and the references therein. Iterative coupling [22, 20], in contrast, is a sequential approach in which either the fluid flow problem or the geomechanics part is solved first, followed by the solution of the other system. This process is repeated until a converged solution within a prescribed tolerance is achieved. The main advantage of iterative coupling methods is that existing software for simulating fluid flow and geomechanics can be reused. These type of schemes have been widely studied [28, 9, 4, 6]. In particular, in [11] and [31] a re-interpretation of the four commonly used sequential splitting methods as preconditioned-Richardson iterations with block-triangular preconditioning is presented. Such analysis indicates that a fully-implicit method outperforms the con-

vergence rate of the sequential-implicit methods. Following this idea a family of preconditioners to accelerate the convergence of Krylov subspace methods was recently proposed for the three-field formulation of the poromechanics problem [10].

In this work, we take the monolithic approach and develop efficient block preconditioners for Krylov subspace methods for solving the linear systems of equations arising from the discretization of the two-field formulation of Biot's model. These preconditioners take advantage of the block structure of the discrete problem, decoupling different fields at the preconditioning stage. Our theoretical results show their efficiency and robustness with respect to the physical and discretization parameters. Moreover, the techniques proposed here can also be used for designing fast solvers for the three-field formulation of Biot's model.

The paper is organized as follows. Section 2 introduces the stabilized finite-element discretizations for the two-field formulation and the basics of block preconditioners. The proposed block preconditioners are introduced in Section 3. Finally, in Section 4, we present numerical experiments illustrating the effectiveness and robustness of the proposed preconditioners and make concluding remarks in Section 5.

2 Two-Field Formulation

First, we consider the two-field formulation of Biot's model (1)-(5), where the unknowns are the displacement $\boldsymbol{u}$ and the pressure p. By considering appropriate Sobolev spaces and integration by parts, we obtain the following variational form: find $\boldsymbol{u}(t) \in \boldsymbol{H}_0^1(\Omega)$ and $p(t) \in H_0^1(\Omega)$, such that

$$a(\boldsymbol{u},\boldsymbol{v}) - \alpha(\operatorname{div}\boldsymbol{v}, p) = (\boldsymbol{g},\boldsymbol{v}), \quad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega), \tag{6}$$

$$-\alpha(\operatorname{div}\partial_t \boldsymbol{u}, q) - a_p(p,q) = (f,q), \quad \forall q \in H_0^1(\Omega), \tag{7}$$

where

$$a(\boldsymbol{u},\boldsymbol{v}) = 2\mu \int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) + \lambda \int_\Omega \operatorname{div}\boldsymbol{u} \operatorname{div}\boldsymbol{v} \quad \text{and} \quad a_p(p,q) \quad = \int_\Omega K\nabla p \cdot \nabla q.$$

Here, we assume the above holds for fixed values of t in some time interval, $(0, t_{max}]$. The system is then completed with suitable initial data $\boldsymbol{u}(0)$ and $p(0)$.

2.1 Finite-Element Method

We consider two stable discretizations for the two-field formulation of Biot's model proposed in [29]: $\mathbb{P}_1$-$\mathbb{P}_1$ elements and the Mini element with stabilization. The fully discretized scheme at time t_n, $n = 1, 2, \ldots$ is as follows:
Find $\boldsymbol{u}_h^n \in \boldsymbol{V_h} \subset \boldsymbol{H}_0^1(\Omega)$ and $p_h^n \in Q_h \subset H_0^1(\Omega)$, such that,

$$a(\boldsymbol{u}_h^n, \boldsymbol{v}_h) - \alpha(\operatorname{div} \boldsymbol{v}_h, p_h^n) = (\boldsymbol{g}(t_n), \boldsymbol{v}_h), \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \tag{8}$$

$$-\alpha(\operatorname{div} \bar{\partial}_t \boldsymbol{u}_h^n, q_h) - a_p(p_h^n, q_h) - \eta h^2 (\nabla \bar{\partial}_t p_h^n, \nabla q_h) = (f(t_n), q_h), \quad \forall q_h \in Q_h, \tag{9}$$

where $\bar{\partial}_t \boldsymbol{u}_h^n := (\boldsymbol{u}_h^n - \boldsymbol{u}_h^{n-1})/\tau$, $\bar{\partial}_t p_h^n := (p_h^n - p_h^{n-1})/\tau$, and η represents the stabilization parameter. Here, $\boldsymbol{V}_h$ and Q_h come from the $\mathbb{P}_1$-$\mathbb{P}_1$ or Mini element. At each time step, the linear system has the following two-by-two block form:

$$\mathscr{A}\boldsymbol{x} = \boldsymbol{b}, \quad \mathscr{A} = \begin{pmatrix} A_{\boldsymbol{u}} & \alpha B^T \\ \alpha B & -\tau A_p - \eta h^2 L_p \end{pmatrix}, \ \boldsymbol{x} = \begin{pmatrix} \boldsymbol{u} \\ p \end{pmatrix}, \text{ and } \boldsymbol{b} = \begin{pmatrix} \boldsymbol{f}_{\boldsymbol{u}} \\ f_p \end{pmatrix}, \tag{10}$$

where $a(\boldsymbol{u}, \boldsymbol{v}) \to A_{\boldsymbol{u}}$, $-(\operatorname{div} \boldsymbol{u}, q) \to B$, $a_p(\nabla p, \nabla q) \to A_p$, and $(\nabla p, \nabla q) \to L_p$ represent the discrete versions of the variational forms.

2.2 Block Preconditioners

Next, we introduce the general theory for designing block preconditioners of Krylov subspace iterative methods [24, 27]. Let $\boldsymbol{X}$ be a real, separable Hilbert space equipped with norm $\|\cdot\|_{\boldsymbol{X}}$ and inner product $(\cdot,\cdot)_{\boldsymbol{X}}$. Also let $\mathscr{A} : \boldsymbol{X} \mapsto \boldsymbol{X}'$ be a bounded and symmetric operator induced by a symmetric and bounded bilinear form $\mathscr{L}(\cdot,\cdot)$, i.e. $\langle \mathscr{A}\boldsymbol{x}, \boldsymbol{y}\rangle = \mathscr{L}(\boldsymbol{x}, \boldsymbol{y})$. We assume the bilinear form is bounded and satisfies an inf-sup condition:

$$|\mathscr{L}(\boldsymbol{x}, \boldsymbol{y})| \le \beta \|\boldsymbol{x}\|_{\boldsymbol{X}} \|\boldsymbol{y}\|_{\boldsymbol{X}}, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X} \quad \text{and} \quad \inf_{\boldsymbol{x} \in \boldsymbol{X}} \sup_{\boldsymbol{y} \in \boldsymbol{X}} \frac{\mathscr{L}(\boldsymbol{x}, \boldsymbol{y})}{\|\boldsymbol{x}\|_{\boldsymbol{X}} \|\boldsymbol{y}\|_{\boldsymbol{X}}} \ge \gamma > 0. \tag{11}$$

2.2.1 Norm-equivalent Preconditioner

Consider a symmetric positive definite (SPD) operator $\mathscr{M} : \boldsymbol{X}' \mapsto \boldsymbol{X}$ as a preconditioner for solving $\mathscr{A}\boldsymbol{x} = \boldsymbol{b}$. We define an inner product $(\boldsymbol{x}, \boldsymbol{y})_{\mathscr{M}^{-1}} := \langle \mathscr{M}^{-1}\boldsymbol{x}, \boldsymbol{y}\rangle$ on $\boldsymbol{X}$ and the corresponding induced norm is $\|\boldsymbol{x}\|^2_{\mathscr{M}^{-1}} := (\boldsymbol{x}, \boldsymbol{x})_{\mathscr{M}^{-1}}$. It is easy to show that $\mathscr{M}\mathscr{A} : \boldsymbol{X} \mapsto \boldsymbol{X}$ is symmetric with respect to $(\cdot,\cdot)_{\mathscr{M}^{-1}}$. Therefore, we can use $\mathscr{M}$ as a preconditioner for the MINRES algorithm and use the following theorem for the convergence rate of preconditioned MINRES.

Theorem 1. *[18] If $\boldsymbol{x}^m$ is the m-th iteration of MINRES and $\boldsymbol{x}$ is the exact solution, then,*

$$\|\boldsymbol{r}^m\|_{\mathscr{M}} \le 2\rho^m \|\boldsymbol{r}^0\|_{\mathscr{M}}, \tag{12}$$

where $\boldsymbol{r}^k = \mathscr{A}(x - x^k)$ is the residual after the k-th iteration, $\rho = \frac{\kappa(\mathscr{M}\mathscr{A})-1}{\kappa(\mathscr{M}\mathscr{A})+1}$, and $\kappa(\mathscr{M}\mathscr{A})$ denotes the condition number of $\mathscr{M}\mathscr{A}$.

In [27], Mardal and Winther show that, if the well-posedness conditions, (11), hold, and $\mathscr{M}$ satisfies

$$c_1\|\boldsymbol{x}\|_{\boldsymbol{X}}^2 \leq \|\boldsymbol{x}\|_{\mathscr{M}^{-1}}^2 \leq c_2\|\boldsymbol{x}\|_{\boldsymbol{X}}^2, \tag{13}$$

then, $\mathscr{A}$ and $\mathscr{M}$ are *norm-equivalent* and $\kappa(\mathscr{M}\mathscr{A}) \leq \frac{c_2\beta}{c_1\gamma}$. This implies that $\rho \leq \frac{c_2\beta - c_1\gamma}{c_2\beta + c_1\gamma}$. Thus, if the original problem is well-posed and the constants c_1 and c_2 are independent of the physical and discretization parameters, then the convergence rate of preconditioned MINRES is uniform, hence $\mathscr{M}$ is a robust preconditioner.

2.2.2 FOV-equivalent Preconditioner

In this section we consider the class of field-of-values-equivalent (FOV-equivalent) preconditioners $\mathscr{M}_L : \boldsymbol{X}' \mapsto \boldsymbol{X}$, for GMRES. We define the notion of FOV-equivalence after the following classical theorem on the convergence rate of the preconditioned GMRES method.

Theorem 2. *[13, 12] If $\boldsymbol{x}^m$ is the m-th iteration of the GMRES method preconditioned with $\mathscr{M}_L$ and $\boldsymbol{x}$ is the exact solution, then*

$$\|\mathscr{M}_L\mathscr{A}(\boldsymbol{x} - \boldsymbol{x}^m)\|_{\mathscr{M}^{-1}}^2 \leq \left(1 - \frac{\Sigma^2}{\Upsilon^2}\right)^m \|\mathscr{M}_L\mathscr{A}(\boldsymbol{x} - \boldsymbol{x}^0)\|_{\mathscr{M}^{-1}}^2, \tag{14}$$

where, for any $\boldsymbol{x} \in \boldsymbol{X}$,

$$\Sigma \leq \frac{(\mathscr{M}_L\mathscr{A}\boldsymbol{x}, \boldsymbol{x})_{\mathscr{M}^{-1}}}{(\boldsymbol{x}, \boldsymbol{x})_{\mathscr{M}^{-1}}}, \quad \frac{\|\mathscr{M}_L\mathscr{A}\boldsymbol{x}\|_{\mathscr{M}^{-1}}}{\|\boldsymbol{x}\|_{\mathscr{M}^{-1}}} \leq \Upsilon. \tag{15}$$

If the constants Σ and Υ are independent of the physical and discretization parameters, then $\mathscr{M}_L$ is a uniform left preconditioner for GMRES and is referred to as an *FOV-equivalent* preconditioner. In [24], a block lower triangular preconditioner has been shown to satisfy (15) based on the well-posedness conditions, (11), for Stokes/Navier-Stokes equations. More recently, the same approach has been generalized to Maxwell's equations [2] and Magnetohydrodynamics [26].

Similar arguments also apply to right preconditioners for GMRES, $\mathscr{M}_U : \boldsymbol{X}' \mapsto \boldsymbol{X}$, where the operators, $\mathscr{M}_U$ and $\mathscr{A}$, are FOV equivalent if, for any $\boldsymbol{x}' \in \boldsymbol{X}'$,

$$\Sigma \leq \frac{(\mathscr{A}\mathscr{M}_U\boldsymbol{x}', \boldsymbol{x}')_{\mathscr{M}}}{(\boldsymbol{x}', \boldsymbol{x}')_{\mathscr{M}}}, \quad \frac{\|\mathscr{A}\mathscr{M}_U\boldsymbol{x}'\|_{\mathscr{M}}}{\|\boldsymbol{x}'\|_{\mathscr{M}}} \leq \Upsilon. \tag{16}$$

Again, if Σ and Υ are independent of the physical and discretization parameters, $\mathscr{M}_U$ is a uniform right preconditioner for GMRES. Such an approach leads to block upper triangular preconditioners.

3 Robust Preconditioners for Biot's Model

In this section, following the framework proposed in [24, 27] and techniques recently developed in [26], we design block diagonal and triangular preconditioners based on the well-posedness of the discretized linear system at each time step. First, we study the well-posedness of the linear system (10). The analysis here is similar to the analysis in [29]. However, we make sure that the constants arising from the analysis are independent of any physical and discretization parameters.

The choice of finite-element spaces give $\boldsymbol{X} = \boldsymbol{V}_h \times Q_h$, and the finite-element pair satisfies the following inf-sup condition (see [30]),

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}_h} \frac{(\operatorname{div}\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_1} \geq \gamma_B^0 \|q\| - \xi^0 h \|\nabla q\|, \quad \forall q \in Q_h. \tag{17}$$

Here, $\gamma_B^0 > 0$ and $\xi^0 \geq 0$ are constants that do not depend on the mesh size. Moreover, if we use the Mini-element, $\xi^0 = 0$.

For $\boldsymbol{x} = (\boldsymbol{u}, p)^T$, we define the following norm,

$$\|\boldsymbol{x}\|_{\boldsymbol{X}}^2 := \|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 + \tau \|p\|_{A_p}^2 + \eta h^2 \|p\|_{L_p}^2 + \frac{\alpha^2}{\zeta^2} \|p\|^2, \tag{18}$$

where $\|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 := a(\boldsymbol{u}, \boldsymbol{u})$, $\|p\|_{A_p}^2 := a_p(\nabla p, \nabla p)$, $\|p\|_{L_p}^2 := (\nabla p, \nabla p)$, $\zeta = \sqrt{\lambda + \frac{2\mu}{d}}$, and $d = 2$ or 3 is the dimension of the problem. With ζ defined as above, it holds that $\|\boldsymbol{v}\|_{A_{\boldsymbol{u}}} \leq \sqrt{d}\zeta \|\boldsymbol{v}\|_1$, and we can reformulate the inf-sup condition, (17), as follows,

$$\sup_{\boldsymbol{v}\in\boldsymbol{V}_h} \frac{(B\boldsymbol{v}, q)}{\|\boldsymbol{v}\|_{A_{\boldsymbol{u}}}} \geq \sup_{\boldsymbol{v}\in\boldsymbol{V}_h} \frac{(B\boldsymbol{v}, q)}{\sqrt{d}\zeta \|\boldsymbol{v}\|_1} \geq \frac{\gamma_B^0}{\sqrt{d}\zeta} \|q\| - \frac{\xi^0}{\sqrt{d}\zeta} h \|\nabla q\| =: \frac{\gamma_B}{\zeta} \|q\| - \frac{\xi}{\zeta} h \|\nabla q\|, \tag{19}$$

where $\gamma_B := \gamma_B^0/\sqrt{d}$ and $\xi = \xi^0/\sqrt{d}$.

Noting that for $d = 2, 3$, $2\mu(\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\varepsilon}(\boldsymbol{v})) \leq a(\boldsymbol{v}, \boldsymbol{v}) \leq (2\mu + d\lambda)(\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\varepsilon}(\boldsymbol{v}))$. Thus, $(\operatorname{div}\boldsymbol{v}, \operatorname{div}\boldsymbol{v}) \leq d(\boldsymbol{\varepsilon}(\boldsymbol{v}), \boldsymbol{\varepsilon}(\boldsymbol{v}))$ and,

$$\zeta^2 \|B\boldsymbol{v}\|^2 = (\lambda + \frac{2\mu}{d}) \|\operatorname{div}\boldsymbol{v}\|^2 \leq \|\boldsymbol{v}\|_{A_{\boldsymbol{u}}}^2 \Longrightarrow \|B\boldsymbol{v}\| \leq \frac{1}{\zeta} \|\boldsymbol{v}\|_{A_{\boldsymbol{u}}}. \tag{20}$$

This allows us to show that linear system (10) is well-posed.

Theorem 3. *For $\boldsymbol{x} = (\boldsymbol{u}, p)$ and $\boldsymbol{y} = (\boldsymbol{v}, q)$, let*

$$\mathscr{L}(\boldsymbol{x}, \boldsymbol{y}) = (A_{\boldsymbol{u}}\boldsymbol{u}, \boldsymbol{v}) + \alpha(B\boldsymbol{v}, p) + \alpha(B\boldsymbol{u}, q) - \tau(K\nabla p, \nabla q) - \eta h^2 (\nabla p, \nabla q). \tag{21}$$

Then, (11) *holds and $\mathscr{A}$ defined in* (10) *is an isomorphism from $\boldsymbol{X}$ to $\boldsymbol{X}'$ provided that the stabilization parameter, η, satisfies $\eta = \delta \frac{\alpha^2}{\zeta^2}$ with $\delta > 0$. Moreover, the constants γ and β are independent of the physical and discretization parameters.*

Proof. Based on the inf-sup condition (17) and (19), for any p, there exists $\boldsymbol{w} \in \boldsymbol{V}_h$ such that $(B\boldsymbol{w}, p) \geq \left(\frac{\gamma_B}{\zeta}\|p\| - \frac{\xi}{\zeta}h\|\nabla p\|\right)\|\boldsymbol{w}\|_{A_{\boldsymbol{u}}}$ and $\|\boldsymbol{w}\|_{A_{\boldsymbol{u}}} = \|p\|$. For given $(\boldsymbol{u}, p) \in \boldsymbol{V}_h \times Q_h$, we choose $\boldsymbol{v} = \boldsymbol{u} + \theta\boldsymbol{w}$, $\theta = \vartheta\frac{\gamma_B\alpha}{\zeta}$ and $q = -p$ and then have,

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{x}, \boldsymbol{y}) &= (A_{\boldsymbol{u}}\boldsymbol{u}, \boldsymbol{u} + \theta\boldsymbol{w}) + \alpha(B(\boldsymbol{u} + \theta\boldsymbol{w}), p) - \alpha(B\boldsymbol{u}, p) \\
&\quad + \tau(K\nabla p, \nabla p) + \eta h^2(\nabla p, \nabla p) \\
&\geq \|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 - \vartheta\|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}\frac{\gamma_B\alpha}{\zeta}\|p\| + \vartheta\frac{\gamma_B^2\alpha^2}{\zeta^2}\|p\|^2 - \vartheta\frac{\gamma_B\alpha^2}{\zeta^2}\xi h\|\nabla p\|\|p\| \\
&\quad + \tau\|p\|_{A_p}^2 + \frac{\delta}{\xi^2}\frac{\alpha^2}{\zeta^2}\xi^2 h^2\|\nabla p\|^2 \\
&\geq \begin{pmatrix} \|\boldsymbol{u}\|_{A_{\boldsymbol{u}}} \\ \frac{\gamma_B\alpha}{\zeta}\|p\| \\ \frac{\alpha}{\zeta}\xi h\|\nabla p\| \\ \sqrt{\tau}\|p\|_{A_p} \end{pmatrix}^T \begin{pmatrix} 1 & -\vartheta/2 & 0 & 0 \\ -\vartheta/2 & \vartheta & -\vartheta/2 & 0 \\ 0 & -\vartheta/2 & \delta/\xi^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \|\boldsymbol{u}\|_{A_{\boldsymbol{u}}} \\ \frac{\gamma_B\alpha}{\zeta}\|p\| \\ \frac{\alpha}{\zeta}\xi h\|\nabla p\| \\ \sqrt{\tau}\|p\|_{A_p} \end{pmatrix}.
\end{aligned}
$$

If $0 < \vartheta < \min\{2, \frac{2\delta}{\xi^2}\}$, the matrix in the middle is SPD and there exists γ_0 such that

$$
\mathscr{L}(\boldsymbol{x}, \boldsymbol{y}) \geq \gamma_0\left(\|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 + \frac{\gamma_B^2\alpha^2}{\zeta^2}\|p\|^2 + \frac{\alpha^2}{\zeta^2}\xi^2 h^2\|\nabla p\|^2 + \tau\|p\|_{A_p}^2\right) \geq \tilde{\gamma}\|\boldsymbol{x}\|_{\boldsymbol{X}}^2,
$$

where $\tilde{\gamma} = \gamma_0\min\{\gamma_B^2, \xi^2/\delta\}$. Also, it is straightforward to verify $\|(\boldsymbol{v}, q)\|_{\boldsymbol{X}}^2 \leq \bar{\gamma}^2\|(\boldsymbol{u}, p)\|_{\boldsymbol{X}}^2$, and the boundedness of $\mathscr{L}$ by continuity of each term and the Cauchy-Schwarz inequality. Therefore, $\mathscr{L}$ satisfies (11) with $\gamma = \tilde{\gamma}/\bar{\gamma}$.

Remark 1. Note that the choice of $\zeta = \sqrt{\lambda + 2\mu/d}$ is essential to the proof, but is consistent with previous implementations [3, 29]. Additionally, choosing *any* $\delta > 0$ is sufficient to show the well-posedness of the stabilized discretization. However, for eliminating non-physical oscillations of the pressure approximation seen in practice [3], this is not sufficient, and δ should be sufficiently large. For example, in 1D, $\delta = 1/4$ is chosen.

3.1 Block Diagonal Preconditioner

Now that we have shown (11) and that the system is well-posed, we find SPD operators such that (13) is satisfied. One natural choice is the Reisz operator corresponding to the inner product $(\cdot,\cdot)_{\boldsymbol{X}}$, $(\mathscr{B}\boldsymbol{f}, \boldsymbol{x})_{\boldsymbol{X}} = \langle \boldsymbol{f}, \boldsymbol{x}\rangle$, $\forall \boldsymbol{f} \in \boldsymbol{X}'$, $\boldsymbol{x} \in \boldsymbol{X}$. For the two-field stabilized discretization and the norm $\|\cdot\|_{\boldsymbol{X}}$ defined in (18), we get

$$
\mathscr{B}_D = \begin{pmatrix} A_{\boldsymbol{u}} & 0 \\ 0 & \tau A_p + \eta h^2 L_p + \frac{\alpha^2}{\zeta^2}M \end{pmatrix}^{-1}, \tag{22}
$$

where M is the mass matrix of the pressure block. Since $\mathscr{B}_D$ satisfies the norm-equivalent condition with $c_1 = c_2 = 1$, by Theorem 3, it holds that $\kappa(\mathscr{B}_D\mathscr{A}) = \mathscr{O}(1)$.

In practice, applying the preconditioner $\mathscr{B}_D$ involves the action of inverting the diagonal blocks exactly, which is very expensive and infeasible. Therefore, we replace the diagonal blocks by their spectrally equivalent SPD approximations,

$$\mathscr{M}_D = \begin{pmatrix} H_{\boldsymbol{u}} & 0 \\ 0 & H_p \end{pmatrix},$$

where

$$c_{1,\boldsymbol{u}}(H_{\boldsymbol{u}}\boldsymbol{u},\boldsymbol{u}) \le (A_{\boldsymbol{u}}^{-1}\boldsymbol{u},\boldsymbol{u}) \le c_{2,\boldsymbol{u}}(H_{\boldsymbol{u}}\boldsymbol{u},\boldsymbol{u}) \tag{23}$$

$$c_{1,p}(H_p p, p) \le ((\tau A_p + \eta h^2 L_p + \frac{\alpha^2}{\zeta^2} M)^{-1} p, p) \le c_{2,p}(H_p p, p). \tag{24}$$

Again, $\mathscr{M}_D$ and $\mathscr{A}$ are norm-equivalent and $\kappa(\mathscr{M}_D\mathscr{A}) = \mathscr{O}(1)$ by Theorem 3.

3.2 Block Triangular Preconditioners

Next, we consider block triangular preconditioners for the stabilized scheme, $\mathscr{A}$. For simplicity of the analysis, we modify $\mathscr{A}$ slightly by negating the second equation.

We consider two kinds of block triangular preconditioners,

$$\mathscr{B}_L = \begin{pmatrix} A_{\boldsymbol{u}} & 0 \\ -\alpha B & \tau A_p + \eta h^2 L_p + \frac{\alpha^2}{\zeta^2} M \end{pmatrix}^{-1} \text{ and } \mathscr{M}_L = \begin{pmatrix} H_{\boldsymbol{u}}^{-1} & 0 \\ -\alpha B & H_p^{-1} \end{pmatrix}^{-1}, \tag{25}$$

and block upper triangular preconditioners,

$$\mathscr{B}_U = \begin{pmatrix} A_{\boldsymbol{u}} & \alpha B^T \\ 0 & \tau A_p + \eta h^2 L_p + \frac{\alpha^2}{\zeta^2} M \end{pmatrix}^{-1} \text{ and } \mathscr{M}_U = \begin{pmatrix} H_{\boldsymbol{u}}^{-1} & \alpha B^T \\ 0 & H_p^{-1} \end{pmatrix}^{-1}. \tag{26}$$

According to Theorem 2, we need to show that these block preconditioners satisfy the FOV-equivalence, (15) and (16). We first consider the block lower triangular preconditioner, $\mathscr{B}_L$.

Theorem 4. *There exist constants Σ and Υ, independent of discretization or physical parameters, such that, for any $\boldsymbol{x} = (\boldsymbol{u}, p)^T \neq \boldsymbol{0}$,*

$$\Sigma \le \frac{(\mathscr{B}_L\mathscr{A}\boldsymbol{x},\boldsymbol{x})_{(\mathscr{B}_D)^{-1}}}{(\boldsymbol{x},\boldsymbol{x})_{(\mathscr{B}_D)^{-1}}}, \ \frac{\|\mathscr{B}_L\mathscr{A}\boldsymbol{x}\|_{(\mathscr{B}_D)^{-1}}}{\|\boldsymbol{x}\|_{(\mathscr{B}_D)^{-1}}} \le \Upsilon,$$

provided that $\eta = \delta\frac{\alpha^2}{\zeta^2}$ with $\delta > 0$.

Proof. By direct computation,

$$
\begin{aligned}
(\mathscr{B}_L\mathscr{A}\boldsymbol{x},\boldsymbol{x})_{(\mathscr{B}_D)^{-1}} &= (\boldsymbol{u},\boldsymbol{u})_{A_{\boldsymbol{u}}} + \alpha(B^T p,\boldsymbol{u}) + \tau(p,p)_{A_p} \\
&\quad + \eta h^2 (L_p p, p) + \alpha^2 (B A_{\boldsymbol{u}}^{-1} B^T p, p) \\
&\geq \Sigma_0 \left(\|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 + \tau \|p\|_{A_p}^2 + \eta h^2 \|p\|_{L_p}^2 + \alpha^2 \|B^T p\|_{A_{\boldsymbol{u}}^{-1}}^2 \right).
\end{aligned}
$$

Note that, due to the inf-sup condition (17),

$$
\|B^T p\|_{A_{\boldsymbol{u}}^{-1}} = \sup_{\boldsymbol{v}} \frac{(B\boldsymbol{v},p)}{\|\boldsymbol{v}\|_{A_{\boldsymbol{u}}}} \geq \frac{\gamma_B}{\zeta}\|p\| - \frac{\xi}{\zeta} h \|\nabla p\|.
$$

Therefore, since $\eta = \delta \frac{\alpha^2}{\zeta^2}$ with $\delta > 0$ and by choosing $\frac{1}{1+\delta/\xi^2} < \theta < 1$,

$$
\begin{aligned}
(\mathscr{B}_L\mathscr{A}\boldsymbol{x},\boldsymbol{x})_{(\mathscr{B}_D)^{-1}} &\geq \Sigma_0 \Big[\|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 + \tau \|p\|_{A_p}^2 + \eta h^2 \|p\|_{L_p}^2 \\
&\qquad + \alpha^2 \left(\frac{\gamma_B}{\zeta}\|p\| - \frac{\xi}{\zeta} h \|\nabla p\| \right)^2 \Big] \\
&\geq \Sigma_0 \Big[\|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 + \tau \|p\|_{A_p}^2 \\
&\qquad + (1-\theta) \frac{\gamma_B^2 \alpha^2}{\zeta^2} \|p\|^2 + \left(1 + \frac{\delta}{\xi^2} - \frac{1}{\theta} \right) \frac{\alpha^2}{\zeta^2} \xi^2 h^2 \|\nabla p\|^2 \Big] \\
&\geq \Sigma_0 \Sigma_1 \left(\|\boldsymbol{u}\|_{A_{\boldsymbol{u}}}^2 + \tau \|p\|_{A_p}^2 + \frac{\alpha^2}{\zeta^2} h^2 \|p\|_{L_p}^2 + \frac{\alpha^2}{\zeta^2} \|p\|^2 \right) \\
&=: \Sigma(\boldsymbol{x},\boldsymbol{x})_{(\mathscr{B}_D)^{-1}},
\end{aligned}
$$

where $\Sigma_1 := \min\{1, (1-\theta)\gamma_B^2, \left(1 + \frac{\delta}{\xi^2} - \frac{1}{\theta}\right) \frac{\xi^2}{\delta}\}$. This gives the lower bound. The upper bound Υ can be obtained directly from the continuity of each term, the Cauchy-Schwarz inequality, and the fact that $\|B^T p\|_{A_{\boldsymbol{u}}^{-1}} \leq \frac{1}{\zeta}\|p\|$ obtained by (20).

Similarly, we can show that the other three block preconditioners are also FOV-equivalent with $\mathscr{A}$ and, therefore, can be used as preconditioners for GMRES. Due to the length constraint of this paper and the fact that the proofs are similar, we only state the results here.

Theorem 5. *If the conditions* (23) *and* (24) *hold and* $\|I - H_{\boldsymbol{u}} A_{\boldsymbol{u}}\|_{A_{\boldsymbol{u}}} \leq \rho$ *with* $0 \leq \rho < 1$*, and there exist constants* Σ *and* Υ*, independent of discretization and physical parameters, such that, for any* $\boldsymbol{x} = (\boldsymbol{u},p)^T \neq \boldsymbol{0}$*, it holds that*

$$
\Sigma \leq \frac{(\mathscr{M}_L\mathscr{A}\boldsymbol{x},\boldsymbol{x})_{(\mathscr{M}_D)^{-1}}}{(\boldsymbol{x},\boldsymbol{x})_{(\mathscr{M}_D)^{-1}}}, \ \frac{\|\mathscr{M}_L\mathscr{A}\boldsymbol{x}\|_{(\mathscr{M}_D)^{-1}}}{\|\boldsymbol{x}\|_{(\mathscr{M}_D)^{-1}}} \leq \Upsilon,
$$

provided that $\eta = \delta \frac{\alpha^2}{\zeta^2}$ *with* $\delta > 0$*.*

Theorem 6. *There exist constants* Σ *and* Υ*, independent of discretization or physical parameters, such that, for any* $\boldsymbol{0} \neq \boldsymbol{x}' \in \boldsymbol{X}'$*, it holds that*

$$\Sigma \leq \frac{(\mathscr{A}\mathscr{B}_U \boldsymbol{x}', \boldsymbol{x}')_{\mathscr{B}_D}}{(\boldsymbol{x}', \boldsymbol{x}')_{\mathscr{B}_D}}, \ \frac{\|\mathscr{A}\mathscr{B}_U \boldsymbol{x}'\|_{\mathscr{B}_D}}{\|\boldsymbol{x}'\|_{\mathscr{B}_D}} \leq \Upsilon,$$

provided that $\eta = \delta \frac{\alpha^2}{\zeta^2}$ *with* $\delta > 0$.

Theorem 7. *If the conditions* (23) *and* (24) *hold and* $\|I - H_{\boldsymbol{u}} A_{\boldsymbol{u}}\|_{A_{\boldsymbol{u}}} \leq \rho$ *with* $0 \leq \rho < 1$*, and there exist constants* Σ *and* Υ*, independent of discretization or physical parameters, such that, for any* $\mathbf{0} \neq \boldsymbol{x}' \in \boldsymbol{X}'$*, it holds that*

$$\Sigma \leq \frac{(\mathscr{A}\mathscr{M}_U \boldsymbol{x}', \boldsymbol{x}')_{\mathscr{M}_D}}{(\boldsymbol{x}', \boldsymbol{x}')_{\mathscr{M}_D}}, \ \frac{\|\mathscr{A}\mathscr{M}_U \boldsymbol{x}'\|_{\mathscr{M}_D}}{\|\boldsymbol{x}'\|_{\mathscr{M}_D}} \leq \Upsilon,$$

provided that $\eta = \delta \frac{\alpha^2}{\zeta^2}$ *with* $\delta > 0$.

Remark 2. The block upper preconditioner $\mathscr{B}_U$ here is related to the well-known *fixed-stress split* scheme [22]. In fact, without the stabilization term, i.e., $\eta = 0$, it is exactly a re-cast of the fixed-stress split scheme [31]. Moreover, $\zeta^2 = \lambda + 2\mu/d =: K_{\text{dr}}$, where K_{dr} is the drained bulk modulus of the solid. This is exactly the choice suggested in [21]. Here, we give a rigorous theoretical analysis when the fixed-stress split scheme is used as a preconditioner. Our analysis is more general in the sense that $\mathscr{M}_U$ is an inexact version of the fixed-stress split scheme, and we have generalized it to the finite-element discretization with stabilizations.

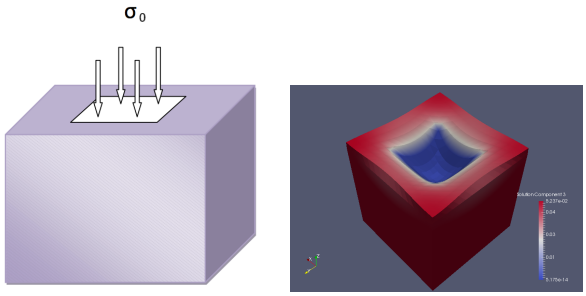
4 Numerical Experiments

Finally, we provide some preliminary numerical results to demonstrate the robustness of the proposed preconditioners. As a discretization, we use the stabilized $\mathbb{P}_1$-$\mathbb{P}_1$ scheme described in [29] and implemented in the `HAZMATH` library [1].

We consider a 3D footing problem as in [15], on the domain, $\Omega = (-32, 32) \times (-32, 32) \times (0, 64)$. This is shown in the left side of Figure 1, and represents a block of porous soil. A uniform load of intensity $0.1N/m^2$ is applied in a square of size $32 \times 32m^2$ at the middle of the top of the domain. The base of the domain is assumed to be fixed while the rest of the domain is free to drain. For the material properties, the Lame coefficients are computed in terms of the Young modulus, E, and the Poisson ratio, ν: $\lambda = \frac{E\nu}{(1-2\nu)(1+\nu)}$ and $\mu = \frac{E}{1+2\nu}$. Since we want to study the robustness of the preconditioners with respect to the physical parameters, we fix $E = 3 \times 10^4 N/m^2$ and let ν change in the experiments. The right side of Figure 1 shows the results of the simulation, demonstrating the deformation due to a uniform load.

We first study the performance of the preconditioners with respect to the mesh size h and time step size τ. Therefore, we fix $K = 10^{-6} m^2$ and $\nu = 0.2$. We use flexible GMRES as the outer iteration with a relative residual stopping criteria of 10^{-6}. For $\mathscr{M}_D$, $\mathscr{M}_L$, and $\mathscr{M}_U$, the diagonal blocks are solved inexactly by preconditioned

Fig. 1 Computational domain and boundary conditions



GMRES with a tolerance of 10^{-2}. The results are shown in Table 1. We see that the block preconditioners are effective and robust with respect to the discretization parameters h and τ.

Table 1 Iteration counts for the block preconditioners ($*$ means the direct method for solving diagonal blocks is out of memory)

	$\mathscr{B}_D$				$\mathscr{B}_L$				$\mathscr{B}_U$				$\mathscr{M}_D$				$\mathscr{M}_L$				$\mathscr{M}_U$			
τ \ h	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
0.1	7	7	8	*	5	5	6	*	4	4	4	*	8	8	9	9	6	6	8	8	6	6	8	8
0.01	7	7	8	*	5	5	6	*	4	4	5	*	8	8	9	9	6	6	8	8	6	6	8	8
0.001	7	7	8	*	5	5	6	*	5	5	6	*	8	8	9	9	6	6	8	8	6	6	8	8
0.0001	7	7	8	*	5	5	6	*	5	5	6	*	8	8	9	9	7	6	8	8	6	7	8	8

Next, we investigate the robustness of the block preconditioners with respect to the physical parameters K and ν. We fix the mesh size $h = 1/16$ and time step size $\tau = 0.01$. The results are shown in Table 2. From the iteration counts, we can see that the proposed preconditioners are quite robust with respect to the physical parameters.

Table 2 Iteration counts when varying K or ν

	$\nu = 0.2$ and varying K						$K = 10^{-6}$ and varying ν					
	1	10^{-2}	10^{-4}	10^{-6}	10^{-8}	10^{-10}	0.1	0.2	0.4	0.45	0.49	0.499
$\mathscr{B}_D$	4	7	8	8	8	8	7	8	11	11	12	12
$\mathscr{B}_L$	2	5	6	6	6	6	5	6	8	8	8	9
$\mathscr{B}_U$	3	4	5	5	5	5	4	5	6	6	5	4
$\mathscr{M}_D$	5	8	9	9	9	9	8	9	12	13	14	13
$\mathscr{M}_L$	5	7	8	8	8	8	7	8	11	11	12	12
$\mathscr{M}_U$	5	7	8	8	9	8	7	8	7	8	17	11

5 Conclusions

We have shown that the stability of the discrete problem, using stabilized finite elements, provides the means for designing robust preconditioners for the two-field formulation of Biot's consolidation model. Our analysis shows uniformly bounded condition numbers and uniform convergence rates of the Krylov subspace methods for the preconditioned linear systems. More precisely, we prove that the convergence is independent of mesh size, time step, and the physical parameters of the model.

Current work includes extending this to non-conforming (and conforming) three-field formulations as in [19]. For discretizations that are stable independent of the physical parameters, uniform block diagonal preconditioners can be designed using the framework developed here. Block lower and upper triangular preconditioners for GMRES can also be constructed in a similar fashion. In addition to their excellent convergence properties, the triangular preconditioners naturally provide an (inexact) fixed-stress split scheme for the three-field formulation.

References

1. James H. Adler, Xiaozhe Hu, and Ludmil T. Zikatanov. `HAZMATH`: A simple finite element, graph, and solver library.
2. James H Adler, Xiaozhe Hu, and Ludmil T Zikatanov. Robust solvers for Maxwell's equations with dissipative boundary conditions. *SIAM J. Sci. Comput.*, to appear.
3. G. Aguilar, F. Gaspar, F. Lisbona, and C. Rodrigo. Numerical stabilization of Biot's consolidation model by a perturbation on the flow equation. *Internat. J. Numer. Methods Engrg.*, 75(11):1282–1300, 2008.
4. T. Almani, K. Kumar, A. Dogru, G. Singh, and M.F. Wheeler. Convergence analysis of multirate fixed-stress split iterative schemes for coupling flow with geomechanics. *Comput. Methods Appl. Mech. Engrg.*, 311(1):180 – 207, 2016.
5. Trygve Baerland, Jeonghun J Lee, Kent-Andre Mardal, and Ragnar Winther. Weakly imposed symmetry and robust preconditioners for biot's consolidation model. *arXiv preprint arXiv:1703.07792*, 2017.
6. M. Bause, F.A. Radu, and U. Kocher. Space-time finite element approximation of the Biot poroelasticity system with iterative coupling. *Computer Methods in Applied Mechanics and Engineering*, 320:745 – 768, 2017.
7. L. Bergamaschi, M. Ferronato, and G. Gambolati. Block-partitioned solvers for coupled poromechanics: A unified framework. *Comput. Methods Appl. Mech. Engrg.*, 196:2647 – 2656, 2007.
8. Maurice A. Biot. General theory of threedimensional consolidation. *Journal of Applied Physics*, 12(2):155–164, 1941.
9. Jakub Wiktor Both, Manuel Borregales, Jan Martin Nordbotten, Kundan Kumar, and Florin Adrian Radu. Robust fixed stress splitting for Biot's equations in heterogeneous media. *Applied Mathematics Letters*, 68:101 – 108, 2017.
10. N. Castelleto, J. A. White, and M. Ferronato. Scalable algorithms for three-field mixed finite element coupled poromechanics. *Journal of Computational Physics*, 327:894 – 918, 2016.
11. N. Castelleto, J. A. White, and H. A. Tchelepi. Accuracy and convergence properties of the fixed-stress iterative solution of two-way coupled poromechanics. *Int. J. Numer. Anal. Meth. Geomech.*, 39:1593 – 1618, 2015.

12. Stanley C Eisenstat, Howard C Elman, and Martin H Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis*, 20(2):345–357, 1983.
13. Howard C Elman. *Iterative methods for large, sparse, nonsymmetric systems of linear equations*. PhD thesis, Yale University New Haven, Conn, 1982.
14. M. Ferronato, L. Bergamaschi, , and G. Gambolati. Performance and robustness of block constraint preconditioners in finite element coupled consolidation problems. *Int. J. Numer. Meth. Engng*, 81:381 – 402, 2010.
15. F. J. Gaspar, J. L. Gracia, F. J. Lisbona, and C. W. Oosterlee. Distributive smoothers in multigrid for problems with dominating grad-div operators. *Numer. Linear Algebra Appl.*, 15(8):661–683, 2008.
16. F. J. Gaspar, F. J. Lisbona, C.W. Oosterlee, and R. Wienands. A systematic comparison of coupled and distributivesmoothing in multigrid for the poroelasticity system. *Numer Linear Algebra Appl*, 11:93–113, 2004.
17. Francisco J. Gaspar and Carmen Rodrigo. On the fixed-stress split scheme as smoother in multigrid methods for coupling flow and geomechanics. *Submitted*, 2017.
18. A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, 1997.
19. Xiaozhe Hu, Carmen Rodrigo, Francisco J Gaspar, and Ludmil T Zikatanov. A nonconforming finite element method for the Biot's consolidation model in poroelasticity. *Journal of Computational and Applied Mathematics*, 310:143–154, 2017.
20. J. Kim. *Sequential methods for coupled geomechanics and multiphase flow*. Stanford University, 2010.
21. J Kim, HA Tchelepi, and R Juanes. Stability and convergence of sequential methods for coupled flow and geomechanics: Fixed-stress and fixed-strain splits. *Computer Methods in Applied Mechanics and Engineering*, 200(13):1591–1606, 2011.
22. Jihoon Kim, Hamdi A Tchelepi, Ruben Juanes, et al. Stability, accuracy and efficiency of sequential methods for coupled flow and geomechanics. In *SPE reservoir simulation symposium*. Society of Petroleum Engineers, 2009.
23. Jeonghun J Lee, Kent-Andre Mardal, and Ragnar Winther. Parameter-robust discretization and preconditioning of biot's consolidation model. *SIAM Journal on Scientific Computing*, 39(1):A1–A24, 2017.
24. D. Loghin and A. J. Wathen. Analysis of preconditioners for saddle-point problems. *SIAM J. Sci. Comput.*, 25(6):2029–2049 (electronic), 2004.
25. P. Luo, C. Rodrigo, F. J. Gaspar, and C.W. Oosterlee. On an Uzawa smoother in multigrid for poroelasticity equations. *Numer Linear Algebra Appl*, page e2074.doi:10.1002/nla.2074, 2017.
26. Yicong Ma, Kaibo Hu, Xiaozhe Hu, and Jinchao Xu. Robust preconditioners for incompressible MHD models. *Journal of Computational Physics*, 316:721–746, 2016.
27. K.A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 2010.
28. Andro Mikelić and Mary F Wheeler. Convergence of iterative coupling for coupled flow and geomechanics. *Computational Geosciences*, 17(3):455–461, 2013.
29. C Rodrigo, FJ Gaspar, Xiaozhe Hu, and LT Zikatanov. Stability and monotonicity for some discretizations of the Biot's consolidation model. *Computer Methods in Applied Mechanics and Engineering*, 298:183–204, 2016.
30. Rolf Stenberg. A technique for analysing finite element methods for viscous incompressible flow. *Internat. J. Numer. Methods Fluids*, 11(6):935–948, 1990. The Seventh International Conference on Finite Elements in Flow Problems (Huntsville, AL, 1989).
31. Joshua A White, Nicola Castelletto, and Hamdi A Tchelepi. Block-partitioned solvers for coupled poromechanics: A unified framework. *Computer Methods in Applied Mechanics and Engineering*, 303:55–74, 2016.

An additive Schwarz analysis for multiplicative Schwarz methods: General case

Susanne C. Brenner

1 Introduction

Multiplicative and additive Schwarz methods are two main classes of iterative methods since the times of Gauss and Jacobi. Traditionally the analyses of these two classes of methods follow different paths. On one hand, the theory for additive Schwarz methods [8, 12, 2, 16, 9, 14, 13, 15, 6, 11], like the theory for the classical Jacobi method, is relatively simple. On the other hand, the theory for multiplicative Schwarz methods [10, 3, 16, 19, 18, 9, 4, 1, 17, 13, 15, 11], like the theory for the classical Gauss-Seidel method, can be quite sophisticated.

An analysis of multiplicative Schwarz methods that is based on the additive theory was carried out in [5]. It is restricted to the case where the subspace corrections are based on symmetric positive definite (SPD) solvers. The goal of this work is to extend the results in [5] to multiplicative Schwarz methods with general subspace corrections. As a by-product we recover the main result in [17], namely a formula for the norm of product operators.

The rest of the paper is organized as follows. First we review the Gauss-Seidel method in Section 2. The analysis of this prototypical multiplicative Schwarz method provides motivations and guidance for the theory in this paper and [5]. We introduce a general framework of multiplicative Schwarz methods in Section 3 and recall the fundamental lemma for additive Schwarz theory in Section 4. The key observation that allows the extension of the formulas in Section 2 to general multiplicative Schwarz methods is presented in Section 5. The main results of the paper are then derived in Section 6. Finally, the connection of our theory to [17] is discussed in Section 7.

Susanne C. Brenner
Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA, e-mail: brenner@math.lsu.edu

2 The Gauss-Seidel Method

The additive Schwarz analysis for multiplicative Schwarz methods is motivated and guided by looking at the analysis of the Gauss-Seidel method through the lens of additive Schwarz theory.

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a SPD matrix and $\mathbf{b} \in \mathbb{R}^n$. The (forward) Gauss-Seidel method for the system $\mathbf{Ax} = \mathbf{b}$ is defined by the iteration step

$$\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{old}} + (\mathbf{L}+\mathbf{D})^{-\mathbf{1}}(\mathbf{b} - \mathbf{A}\mathbf{x}_{\text{old}}), \tag{1}$$

where $\mathbf{L}$ and $\mathbf{D}$ are the strictly lower triangular part and the diagonal part of $\mathbf{A}$ respectively. The error propagation for (1) is described by

$$\mathbf{x} - \mathbf{x}_{\text{new}} = \big(\mathbf{I} - (\mathbf{L}+\mathbf{D})^{-\mathbf{1}}\mathbf{A}\big)(\mathbf{x} - \mathbf{x}_{\text{old}}) = (\mathbf{I} - \mathbf{BA})(\mathbf{x} - \mathbf{x}_{\text{old}}), \tag{2}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{B} = (\mathbf{L}+\mathbf{D})^{-\mathbf{1}}$.

The norm of the iteration matrix $\mathbf{I} - \mathbf{BA}$ in the matrix norm $\|\cdot\|_{\mathbf{A}}$ induced by the inner product $(\mathbf{v},\mathbf{w})_{\mathbf{A}} = \mathbf{w}^{\mathbf{t}}\mathbf{A}\mathbf{v}$ is given by the following standard formula:

$$\|\mathbf{I} - \mathbf{BA}\|_{\mathbf{A}}^{\mathbf{2}} = \|(\mathbf{I} - \mathbf{BA})^*(\mathbf{I} - \mathbf{BA})\|_{\mathbf{A}} = \|(\mathbf{I} - \mathbf{B}^{\mathbf{t}}\mathbf{A})(\mathbf{I} - \mathbf{BA})\|_{\mathbf{A}}, \tag{3}$$

where $(\mathbf{I} - \mathbf{BA})^*$ denotes the adjoint of $\mathbf{I} - \mathbf{BA}$ with respect to $(\cdot,\cdot)_{\mathbf{A}}$. It follows from (3), the spectral theorem and the Rayleigh quotient formula that

$$\begin{aligned} \|\mathbf{I} - \mathbf{BA}\|_{\mathbf{A}}^{\mathbf{2}} &= \lambda_{\max}\big(\mathbf{I} - (\mathbf{B}^{\mathbf{t}} + \mathbf{B} - \mathbf{B}^{\mathbf{t}}\mathbf{AB})\mathbf{A}\big) \\ &= 1 - \lambda_{\min}\big((\mathbf{B}^{\mathbf{t}} + \mathbf{B} - \mathbf{B}^{\mathbf{t}}\mathbf{AB})\mathbf{A}\big) \\ &= 1 - \min_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^{\mathbf{t}}\mathbf{A}\mathbf{v}}{\mathbf{v}^t(\mathbf{B}^{\mathbf{t}} + \mathbf{B} - \mathbf{B}^{\mathbf{t}}\mathbf{AB})^{-\mathbf{1}}\mathbf{v}}\,. \end{aligned} \tag{4}$$

A simple calculation yields

$$(\mathbf{B}^{\mathbf{t}} + \mathbf{B} - \mathbf{B}^{\mathbf{t}}\mathbf{AB})^{-\mathbf{1}} = (\mathbf{I} + \mathbf{D}^{-\mathbf{1}}\mathbf{U})^{\mathbf{t}}\mathbf{D}(\mathbf{I} + \mathbf{D}^{-\mathbf{1}}\mathbf{U}), \tag{5}$$

where $\mathbf{U} = \mathbf{L}^{\mathbf{t}}$ is the strictly upper triangular part of $\mathbf{A}$. It is easy to see that (5) can be rewritten as

$$(\mathbf{B}^{\mathbf{t}} + \mathbf{B} - \mathbf{B}^{\mathbf{t}}\mathbf{AB})^{-\mathbf{1}} = \mathbf{A} + (\mathbf{D}^{-\mathbf{1}}\mathbf{U})^{\mathbf{t}}\mathbf{D}(\mathbf{D}^{-\mathbf{1}}\mathbf{U}). \tag{6}$$

Combining (4) and (5), we have a formula

$$\|\mathbf{I} - \mathbf{BA}\|_{\mathbf{A}}^2 = 1 - \min_{\mathbf{v}\in\mathbb{R}^n} \frac{\mathbf{v}^t\mathbf{A}\mathbf{v}}{\mathbf{v}^t(\mathbf{I} + \mathbf{D}^{-1}\mathbf{U})^t\mathbf{D}(\mathbf{I} + \mathbf{D}^{-1}\mathbf{U})\mathbf{v}} \tag{7}$$

for the norm of the iteration matrix $\mathbf{I} - \mathbf{BA}$. Similarly the formula

$$\|\mathbf{I}-\mathbf{BA}\|_{\mathbf{A}}^2 = 1-\min_{\mathbf{v}\in\mathbb{R}^n}\frac{\mathbf{v^t Av}}{\mathbf{v^t Av}+\mathbf{v^t}(\mathbf{D^{-1}U})^{\mathbf{t}}\mathbf{D}(\mathbf{D^{-1}U})\mathbf{v}}$$
$$= 1-\frac{1}{1+\max\limits_{\mathbf{v}\in\mathbb{R}^n,\|\mathbf{v}\|_{\mathbf{A}}=1}\mathbf{v^t}(\mathbf{D^{-1}U})^{\mathbf{t}}\mathbf{D}(\mathbf{D^{-1}U})\mathbf{v}} \tag{8}$$

follows from (4) and (6).

Since $\mathbf{L}+\mathbf{D}$ is the lower triangular part of $\mathbf{A}$, we can apply forward substitutions to obtain

$$(\mathbf{L}+\mathbf{D})^{-1}\mathbf{A} = \sum_{i=1}^{n}\mathbf{X}_i,$$

where $\mathbf{X}_i \in \mathbb{R}^{n\times n}$ is determined recursively by

$$\mathbf{X}_i = \mathbf{T}_i\Big(\mathbf{I}-\sum_{j=1}^{i-1}\mathbf{X}_j\Big),$$

$\mathbf{T}_i = \mathbf{e}_i(\mathbf{e}_i^t\mathbf{A}\mathbf{e}_i)^{-1}\mathbf{e}_i^t\mathbf{A}$, and $\mathbf{e}_1,\ldots,\mathbf{e}_n$ are the canonical basis vectors in $\mathbb{R}^n$.

It follows that

$$\mathbf{I}-\mathbf{BA} = \Big(\mathbf{I}-\sum_{i=1}^{n-1}\mathbf{X}_i\Big)-\mathbf{X}_n = \Big(\mathbf{I}-\sum_{i=1}^{n-1}\mathbf{X}_i\Big)-\mathbf{T}_n\Big(\mathbf{I}-\sum_{j=1}^{n-1}\mathbf{X}_j\Big)$$
$$= (\mathbf{I}-\mathbf{T}_n)\Big(\mathbf{I}-\sum_{i=1}^{n-1}\mathbf{X}_i\Big) = (\mathbf{I}-\mathbf{T}_n)\cdots(\mathbf{I}-\mathbf{T}_1).$$

Hence (7) and (8) are also formulas for the norm of the product $(\mathbf{I}-\mathbf{T}_n)\cdots(\mathbf{I}-\mathbf{T}_1)$.

Below we will derive similar formulas for general multiplicative Schwarz methods. The key observation is that even though the explicit formula (5) does not exist in the general case, we can find an expression for $\mathbf{v^t}(\mathbf{B^t}+\mathbf{B}-\mathbf{B^tAB})^{-\mathbf{1}}\mathbf{v}$ through the additive Schwarz theory.

3 Multiplicative Schwarz Methods

Let V be a finite dimensional vector space, $a(\cdot,\cdot)$ be a SPD bilinear form on V, and $\alpha \in V'$, the dual space of V. We consider the following problem:

Find $u \in V$ such that

$$a(u,v) = \langle \alpha, v\rangle \qquad \forall\, v \in V, \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes the canonical bilinear form on $V'\times V$.

We can rewrite (9) as

$$Au = \alpha \tag{10}$$

where $A : V \longrightarrow V'$ is defined by

$$\langle Aw, v\rangle = a(w, v) \qquad \forall\, v, w \in V. \tag{11}$$

The operator A is SPD in the sense that

$$\langle Aw, v\rangle = \langle Av, w\rangle \quad \forall\, v, w \in V \quad \text{and} \quad \langle Av, v\rangle > 0 \quad \forall\, v \in V \setminus \{0\}.$$

We will denote by L^* the adjoint of a linear operator $L : V \longrightarrow V$ with respect to $a(\cdot,\cdot)$, i.e,

$$a(Lv, w) = a(v, L^* w) \qquad \forall\, v, w \in V,$$

and for a linear operator $M : W \longrightarrow V$, the operator $M^t : V' \longrightarrow W'$ is defined by

$$\langle M^t \beta, w\rangle = \langle \beta, Mw\rangle \qquad \forall\, \beta \in V,' \, w \in W.$$

Let $V_1, V_2, \ldots, V_J$ be subspaces of V such that

$$V = \sum_{j=1}^{J} V_j, \tag{12}$$

and let $a_j(\cdot,\cdot)$ be a nonsingular bilinear form on V_j, i.e.,

$$A_j : V_j \longrightarrow V_j' \quad \text{is invertible} \tag{13}$$

where

$$\langle A_j v_j, w_j\rangle = a_j(v_j, w_j) \qquad \forall\, v_j, w_j \in V_j.$$

The operator $F_j : V' \longrightarrow V_j$ for $1 \le j \le J$ are defined recursively by

$$a_j(F_j \beta, v_j) = \langle \beta, v_j\rangle - \sum_{k=1}^{j-1} a(F_k \beta, v_j) \qquad \forall\, v_j \in V_j,\, \beta \in V', \tag{14}$$

and we define $B : V' \longrightarrow V$ by

$$B\beta = \sum_{j=1}^{J} (F_j \beta). \tag{15}$$

The multiplicative Schwarz algorithm is then given by the iteration

$$u_{\text{new}} = u_{\text{old}} + B(\alpha - A u_{\text{old}}), \tag{16}$$

As in the case of the Gauss-Seidel method, we have two expressions for the error propagation operator. The first obvious one is given by

$$u - u_{\text{new}} = (I - BA)(u - u_{\text{old}}), \tag{17}$$

where I is the identity operator on V, and the second one, which is responsible for the name of the algorithm, can be derived as follows.

Let $T_j : V \longrightarrow V_j$ be defined by

$$a_j(T_j v, v_j) = a(v, v_j) \qquad \forall\, v \in V,\, v_j \in V_j. \tag{18}$$

Remark 1. Note that (18) implies that $\mathrm{Ker}\, T_j$ is the orthogonal complement of V_j with respect to $a(\cdot,\cdot)$. Therefore $\mathrm{Ker}\, T_j^* = \mathrm{Ker}\, T_j$ and the restrictions of T_j and T_j^* to V_j are isomorphisms. In particular we have $T_j V = V_j = T_j^* V$. It follows that the pseudo-inverse T_j^{-1} (resp., $(T_j^*)^{-1}$) of T_j (resp., T_j^*) with respect to $a(\cdot,\cdot)$ maps V onto V_j.

It follows from (14) and (18) that

$$F_j\beta = T_j(A^{-1}\beta - \sum_{k=1}^{j-1} F_k\beta) = T_j z_j, \tag{19}$$

where

$$\begin{aligned} z_j = A^{-1}\beta - \sum_{k=1}^{j-1} F_k\beta &= \Big(A^{-1}\beta - \sum_{k=1}^{j-2} F_k\beta\Big) - F_{j-1}\beta \\ &= (I - T_{j-1}) z_{j-1} = (I - T_{j-1})\cdots(I - T_1)A^{-1}\beta. \end{aligned} \tag{20}$$

Combining (15), (16), (19) and (20), with $\beta = \alpha - Au_{\mathrm{old}}$, we find

$$\begin{aligned} u - u_{\mathrm{new}} &= u - u_{\mathrm{old}} - \sum_{j=1}^{J} T_j z_j \\ &= \Big(I - \sum_{j=1}^{J} T_j(I - T_{j-1})\cdots(I - T_1)\Big)(u - u_{\mathrm{old}}) \\ &= (I - T_J)\cdots(I - T_1)(u - u_{\mathrm{old}}). \end{aligned} \tag{21}$$

We are interested in formulas for $\|I - BA\|_a = \|(I - T_J)\cdots(I - T_1)\|_a$, where $\|\cdot\|_a$ is the operator norm induced by $a(\cdot,\cdot)$.

4 Additive Schwarz Theory

We need the following fundamental result from the additive Schwarz theory.

Lemma 1. *Let $S_j : V_j \longrightarrow V$ and $B_j : V_j \longrightarrow V_j'$ be linear operators for $1 \le j \le J$, and let B_j be SPD. Then the operator $B = \sum_{j=1}^{J} S_j B_j^{-1} S_j' : V' \longrightarrow V$ is SPD if and only if $V = \sum_{j=1}^{J} S_j V_j$, in which case we have*

$$\langle B^{-1} v, v\rangle = \min_{\substack{v = \sum_{j=1}^{J} S_j v_j \\ v_j \in V_j}} \sum_{j=1}^{J} \langle B_j v_j, v_j\rangle \qquad \forall\, v \in V. \tag{22}$$

Proof. B is clearly symmetric semi-definite, and we have for any $\beta \in V'$,

$$\langle \beta, B\beta \rangle = 0 \Leftrightarrow \sum_{j=1}^{J} \langle S_j^t \beta, B_j^{-1} S_j^t \beta \rangle = 0,$$

which holds if and only if $S_j^t \beta = 0$ for $1 \le j \le J$, since B_j^{-1} is also SPD. We conclude that $\langle \beta, B\beta \rangle = 0$ if and only if

$$\sum_{j=1}^{J} \langle \beta, S_j v_j \rangle = 0 \qquad \forall\, v_j \in V_j,\, 1 \le j \le J.$$

Therefore $\langle \beta, B\beta \rangle = 0$ implies $\beta = 0$ if and only if $V = \sum_{j=1}^{J} S_j V_j$.

The identity (22) comes from the observations that

$$\langle B^{-1} v, \sum_{j=1}^{J} S_j v_j \rangle = \sum_{j=1}^{J} \langle B_j (B_j^{-1} S_j^t B^{-1} v), v_j \rangle = \sum_{j=1}^{J} \langle B_j w_j, v_j \rangle, \tag{23}$$

where $w_j = B_j^{-1} S_j^t B^{-1} v \in V_j$, and

$$\sum_{j=1}^{J} S_j w_j = \sum_{j=1}^{J} S_j (B_j^{-1} S_j^t B^{-1} v) = \Big(\sum_{j=1}^{J} S_j B_j^{-1} S_j^t \Big) B^{-1} v = B B^{-1} v = v. \tag{24}$$

Indeed it follows from (23) that

$$\langle B^{-1} v, v \rangle = \sum_{j=1}^{J} \langle B_j w_j, v_j \rangle \qquad \text{if} \quad \sum_{j=1}^{J} S_j v_j = v, \tag{25}$$

and in particular, because of (24),

$$\langle B^{-1} v, v \rangle = \sum_{j=1}^{J} \langle B_j w_j, w_j \rangle. \tag{26}$$

Subtracting (26) from (25) we find

$$0 = \sum_{j=1}^{J} \langle B_j w_j, v_j - w_j \rangle \qquad \text{if} \quad \sum_{j=1}^{J} S_j v_j = v. \tag{27}$$

The orthogonality condition (27) implies

$$\sum_{j=1}^{J} \langle B_j v_j, v_j \rangle = \sum_{j=1}^{J} \langle B_j w_j, w_j \rangle + \sum_{j=1}^{J} \langle B_j (v_j - w_j), v_j - w_j \rangle \qquad \text{if} \quad \sum_{j=1}^{J} S_j v_j = v,$$

and hence

$$\sum_{j=1}^{J}\langle B_j w_j, w_j\rangle = \min_{\substack{v=\sum_{j=1} S_j v_j \\ v_j\in V_j}} \sum_{j=1}^{J}\langle B_j v_j, v_j\rangle,$$

which together with (26) implies (22).

5 A Fundamental Operator

We begin with the standard formula

$$\|I-BA\|_a^2 = \|(I-BA)^*(I-BA)\|_a, \tag{28}$$

where $(I-BA)^* = I-B^tA$ is the adjoint of $I-BA$ with respect to the bilinear form $a(\cdot,\cdot)$. We can write

$$(I-BA)^*(I-BA) = (I-B^tA)(I-BA) = I-(B^t+B-B^tAB)A. \tag{29}$$

As in the case of the Gauss-Seidel method, the operator B^t+B-B^tAB will play a fundamental role. The key to the additive analysis is to interpret this operator as an additive Schwarz preconditioner. We begin with the following result.

Lemma 2. *We have*

$$\langle\beta,(B^t+B-B^tAB)\beta\rangle = \sum_{j=1}^{J}\Big[2a_j(y_j,y_j)-a(y_j,y_j)\Big] \qquad \forall\,\beta\in V', \tag{30}$$

where $y_j = F_j\beta$.

Proof. From (14) and (15), we have

$$\begin{aligned}\langle\beta,(B^t+B-B^tAB)\beta\rangle &= 2\langle\beta,\sum_{j=1}^{J}y_j\rangle - a(\sum_{\ell=1}^{J}y_\ell,\sum_{j=1}^{J}y_j)\\ &= 2\sum_{j=1}^{J}\Big(a_j(y_j,y_j)+\sum_{\ell=1}^{j-1}a(y_\ell,y_j)\Big) - a(\sum_{\ell=1}^{J}y_\ell,\sum_{j=1}^{J}y_j),\end{aligned}$$

which implies (30) by the symmetry of $a(\cdot,\cdot)$.

We assume that

$$\exists\ \omega_j\in(0,2) \text{ such that} \quad a(v_j,v_j)\le\omega_j a_j(v_j,v_j) \qquad \forall\, v_j\in V_j. \tag{31}$$

Let the operator $B_j: V_j\longrightarrow V_j'$ be defined by

$$\langle B_j v_j, w_j\rangle = a_j(v_j,w_j)+a_j(w_j,v_j)-a(v_j,w_j) \qquad \forall\, v_j,w_j\in V_j. \tag{32}$$

Clearly B_j is symmetric and it is positive definite because of (31).

Remark 2. Since we are in a finite dimensional setting, condition (31) is equivalent to B_j being SPD. It is also equivalent to

$$\|(I-T_j)v\|_a \le \|v\|_a \quad \forall\, v \in V \quad \text{and} \quad \|(I-T_j)v_j\|_a < \|v_j\|_a \quad \forall\, v_j \in V_j \setminus \{0\}.$$

Note that we can write, by (18),

$$\begin{aligned}\langle B_j v_j, w_j\rangle &= a(T_j^{-1}v_j, w_j) + a(w_j, (T_j^{-1})^* v_j) - a(v_j, w_j)\\ &= a((T_j^*)^{-1}(T_j^* + T_j - T_j^* T_j)T_j^{-1}v_j, w_j) = a(\bar{T}_j T_j^{-1} v_j, T_j^{-1} w_j) \end{aligned} \tag{33}$$

for all $v_j, w_j \in V_j$, where

$$\bar{T}_j = T_j^* + T_j - T_j^* T_j. \tag{34}$$

Remark 3. According to Remark 1, we have $\bar{T}_j V \subset V_j$. The relation (33) implies that $a(\bar{T}_j v_j, v_j) = \langle B_j v_j, v_j\rangle > 0$ for $v_j \in V_j \setminus \{0\}$. Therefore the restriction of $\bar{T}_j$ to V_j is an isomorphism and it follows from Remark 1 that $\operatorname{Ker}\bar{T}_j = \operatorname{Ker} T_j = \operatorname{Ker} T_j^*$ is the orthogonal complement of V_j with respect to $a(\cdot,\cdot)$. Consequently the pseudo-inverse $\bar{T}_j^{-1}$ of $\bar{T}_j$ with respect to $a(\cdot,\cdot)$ maps V onto V_j.

From Lemma 2 and (32) we have

$$\langle \beta, (B^t + B - B^t A B)\beta\rangle = \sum_{j=1}^J \langle B_j F_j \beta, F_j \beta\rangle = \sum_{j=1}^J \langle \beta, F_j^t B F_j \beta\rangle \qquad \forall\, \beta \in V'. \tag{35}$$

It then follows from polarization that

$$B^t + B - B^t A B = \sum_{j=1}^J F_j^t B_j F_j = \sum_{j=1}^J (F_j^t B_j) B_j^{-1} (B_j F_j) = \sum_{j=1}^J S_j B_j^{-1} S_j^t, \tag{36}$$

where the operator $S_j : V_j \longrightarrow V$ is given by $S_j = F_j^t B_j = (B_j F_j)^t$.

Remark 4. The identity (36) shows that the operator $B + B^t - B^t A B$ is indeed an additive Schwarz preconditioner. Note that (12) and (14) imply $F_1\beta = \cdots = F_J\beta = 0$ if and only if $\beta = 0$, and hence $B^t + B - B^t A B$ is SPD by (35). Therefore the formula (22) in Lemma 1 is valid.

An explicit formula for S_j is provided by the following lemma.

Lemma 3. *We have*

$$S_j = (I - T_1^*)\cdots(I - T_{j-1}^*)\bar{T}_j T_j^{-1}. \tag{37}$$

Proof. Let $v_j \in V_j$ be arbitrary. It follows from (19), (20) and (33) that

$$\begin{aligned}\langle S_j^t \beta, v_j\rangle = \langle (B_j F_j)\beta, v_j\rangle &= a(\bar{T}_j T_j^{-1} F_j \beta, T_j^{-1} v_j)\\ &= a(z_j, \bar{T}_j T_j^{-1} v_j) = \langle \beta, (I - T_1^*)\cdots(I - T_{j-1}^*)\bar{T}_j T_j^{-1} v_j\rangle,\end{aligned}$$

which implies (37).

6 Formulas for $\|I - BA\|_a$

It follows from (28), (29), the spectral theorem and the Rayleigh quotient formula that

$$\|I - BA\|_a^2 = 1 - \min_{v \in V} \frac{\langle Av, v \rangle}{\langle (B^t + B - B^t AB)^{-1} v, v \rangle},$$

which together with (36) and Lemma 1 (cf. Remark 4) implies

$$\|I - BA\|_a^2 = 1 - \min_{v \in V} \frac{\langle Av, v \rangle}{\min\limits_{\substack{v = \sum_{j=1}^J S_j w_j \\ w_j \in V_j}} \sum\limits_{j=1}^J \langle B_j w_j, w_j \rangle}. \tag{38}$$

Remark 5. Note that we can rewrite (7) as

$$\|\mathbf{I} - \mathbf{BA}\|_{\mathbf{A}}^2 = 1 - \min_{\mathbf{v} \in \mathbb{R}^n} \frac{\mathbf{v^t Av}}{\min\limits_{\mathbf{v} = (\mathbf{I} + \mathbf{D^{-1}U})^{-1}\mathbf{w}} \mathbf{w^t Dw}}, \tag{39}$$

and (38) is precisely the analog of (39).

Next we will replace the implicit decomposition for v that appears in (38) by an explicit decomposition that will lead to an analog of (7). In the case of the Gauss-Seidel method, it is equivalent to inverting the relation $\mathbf{v} = (\mathbf{I} + \mathbf{D}^{-1}\mathbf{U})^{-1}\mathbf{w}$ in (39) to express $\mathbf{w}$ as $(\mathbf{I} + \mathbf{D}^{-1}\mathbf{U})\mathbf{v}$. This motivates the following construction of the explicit decomposition through an "upper triangular" system.

Given $v_j \in V_j$ for $1 \le j \le J$, we want to find $w_j \in V_j$ for $1 \le j \le J$ such that

$$\sum_{j=1}^J S_j w_j = \sum_{j=1}^J v_j. \tag{40}$$

It is easy to check using (37) that the solution of (40) is given by

$$\bar{T}_j T_j^{-1} w_j = v_j + T_j^* \sum_{k=j+1}^J v_k \qquad \text{for} \quad 1 \le j \le J. \tag{41}$$

Combining (33), (38), (40) and (41), we have the following analog of (7):

$$\|I - BA\|_a^2 = 1 - \min_{v \in V} \frac{a(v, v)}{\min\limits_{\substack{v = \sum_{j=1}^J v_j \\ v_j \in V_j}} \sum\limits_{j=1}^J a\Big(v_j + T_j^* \sum\limits_{k=j+1}^J v_k, \bar{T}_j^{-1}\Big(v_j + T_j^* \sum\limits_{k=j+1}^J v_k\Big)\Big)}. \tag{42}$$

Remark 6. In the case where $a_j(\cdot,\cdot)$ is SPD for $1 \le j \le J$, the formula (42) becomes

$$\|I-BA\|_a^2$$
$$= 1-\min_{v\in V}\frac{a(v,v)}{\min\limits_{\substack{v=\sum_{j=1}^J v_j\\ v_j\in V_j}}\sum\limits_{j=1}^J a_j\Big(v_j+T_j\sum\limits_{k=j+1}^J v_k,(2I-T_j)^{-1}\Big(v_j+T_j\sum\limits_{k=j+1}^J v_k\Big)\Big)}. \tag{43}$$

The application of the formula (43) to domain decomposition and multigrid can be found in [5].

To derive the analog of (8), we again seek guidance from the analysis in Section 2. The transition from (7) to (8) involves the difference of $\mathbf{I}+\mathbf{D^{-1}U}$ and $\mathbf{D^{-1}U}$, which is a diagonal matrix. Therefore we look for operators $Q_j : V_j \longrightarrow V_j$ for $1\le j\le J$ such that

$$\sum_{j=1}^J a\Big(v_j+T_j^*\sum_{k=j+1}^J v_k,\bar{T}_j^{-1}\Big(v_j+T_j^*\sum_{k=j+1}^J v_k\Big)\Big)-a(v,v)$$
$$=\sum_{j=1}^J a\Big(v_j+Q_jv_j+T_j^*\sum_{k=j+1}^J v_k,\bar{T}_j^{-1}\Big(v_j+Q_jv_j+T_j^*\sum_{k=j+1}^J v_k\Big)\Big). \tag{44}$$

It is straightforward to check that (44) is equivalent to

$$a(Q_jv_j,\bar{T}_j^{-1}Q_jv_j)+2a(v_j+T_j^*\sum_{k=j+1}^J v_k,\bar{T}_j^{-1}Q_jv_j)=-a\Big(v_j+2\sum_{k=j+1}^J v_k,v_j\Big),$$

which would follow from the relations

$$a(Q_jv_j+2v_j,\bar{T}_j^{-1}Q_jv_j)=-a(v_j,v_j), \tag{45}$$
$$a\Big(T_j^*\sum_{k=j+1}^J v_k,\bar{T}_j^{-1}Q_jv_j\Big)=-a\Big(\sum_{k=j+1}^J v_k,v_j\Big). \tag{46}$$

The relation (46) indicates that we should choose $\bar{T}_j^{-1}Q_j=-T_j^{-1}$ and therefore Q_j should be given by

$$Q_j=-\bar{T}_jT_j^{-1}=-(T_j^*+T_j-T_j^*T_j)T_j^{-1}=-(T_j^*T_j^{-1}+I-T_j^*), \tag{47}$$

and then (45) is also satisfied because

$$\begin{aligned}a(Q_jv_j+2v_j,\bar{T}_j^{-1}Q_jv_j)&=-a\big((I-T_j^*T_j^{-1}+T_j^*)v_j,T_j^{-1}v_j\big)\\&=-a(v_j,T_j^{-1}v_j)+a(T_j^*T_j^{-1}v_j,T_j^{-1}v_j)-a(T_j^*v_j,T_j^{-1}v_j)\\&=-a(v_j,v_j).\end{aligned}$$

In view of (47), we have

$$v_j + Q_j v_j + T_j^* \sum_{k=j+1}^{J} v_k = -T_j^* T_j^{-1} v_j + T_j^* \sum_{k=j}^{J} v_k = T_j^* \Big(\sum_{k=j}^{J} v_k - T_j^{-1} v_j \Big). \quad (48)$$

Putting (42), (44) and (48) together we arrive at the following analog of (8):

$$\begin{aligned} \|I - BA\|_a^2 &= 1 - \min_{v \in V} \frac{a(v,v)}{a(v,v) + \min\limits_{\substack{v = \sum_{j=1}^{J} v_j \\ v_j \in V_j}} a\Big(T_j^* \Big(\sum_{k=j}^{J} v_k - T_j^{-1} v_j\Big), \bar{T}_j^{-1} T_j^* \Big(\sum_{k=j}^{J} v_k - T_j^{-1} v_j\Big)\Big)} \\ &= 1 - \frac{1}{1 + \max\limits_{\substack{v \in V \\ \|v\|_a = 1}} \min\limits_{\substack{v = \sum_{j=1}^{J} v_j \\ v_j \in V_j}} a\Big(T_j^* \Big(\sum_{k=j}^{J} v_k - T_j^{-1} v_j\Big), \bar{T}_j^{-1} T_j^* \Big(\sum_{k=j}^{J} v_k - T_j^{-1} v_j\Big)\Big)}. \end{aligned} \quad (49)$$

7 Connection to the Xu-Zikatanov Theory

The theory in [17] was developed for the product operator $(I - T_J) \cdots (I - T_1)$ on an inner product space $\big(V, a(\cdot,\cdot)\big)$, where $T_j : V \longrightarrow V_j$ and $T_j : V_j \longrightarrow V_j$ is an isomorphism.

A key assumption in [17] is

$$\|T_j v\|_a^2 \le \omega\, a(T_j v, v) \qquad \forall\, v \in V \quad (50)$$

for some $\omega \in (0,2)$.

Lemma 4. *Under assumption* (50)*, we have*

$$T_j v = 0 \Leftrightarrow a(v, v_j) = 0 \quad \forall\, v_j \in V_j.$$

Proof. If $a(v_j, v) = 0$ for all $v_j \in V_j$, then $T_j v = 0$ by (50). Therefore, by a dimension argument, the kernel of T_j is the orthogonal complement of V_j with respect to $a(\cdot,\cdot)$.

In view of Lemma 4, we can define $a_j(\cdot,\cdot)$ by

$$a_j(T_j v, v_j) = a(v, v_j) \qquad \forall\, v \in V,\ v_j \in V_j.$$

Then $a_j(\cdot,\cdot)$ is nonsingular since

$$a_j(w_j, v_j) = 0 \quad \forall\, w_j \in V_j \quad \Rightarrow \quad a(v, v_j) = 0 \quad \forall\, v \in V \quad \Rightarrow v_j = 0.$$

On one hand we have $a(T_j v, T_j v) = \|T_j v\|_a^2$, and on the other hand we have $a(T_j v, v) = a(v, T_j v) = a_j(T_j v, T_j v)$. Hence (50) is equivalent to (31) since $V_j = T_j V$.

We conclude that the framework in [17] is identical to the framework in Section 3 and Section 5, and $\|(I - T_J) \cdots (I - T_1)\|_a$ is given by the formulas (42) and (49). In particular, the formula (49) is identical to the identity (1.1) in [17]. We note that another derivation of this identity can be found in [7].

Acknowledgements

This work is supported in part by the National Science Foundation under Grant No. DMS-16-20273. The author would also like to acknowledge the support provided by the Hausdorff Research Institute of Mathematics at Univeristät Bonn during her visit in Spring 2017.

References

1. M. Benzi, A. Frommer, R. Nabben, and D.B. Szyld. Algebraic theory of multiplicative Schwarz methods. *Numer. Math.*, 89:605–639, 2001.
2. P. Bjørstad and J. Mandel. On the spectra of sums of orthogonal projections with applications to parallel computing. *BIT*, 31:76–88, 1991.
3. J.H. Bramble, J.E. Pasciak, J. Wang, and J. Xu. Convergence estimates for product iterative methods with applications to domain decomposition. *Math. Comp.*, 57:1–21, 1991.
4. J.H. Bramble and X. Zhang. The Analysis of Multigrid Methods. In P.G. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis, VII*, pages 173–415. North-Holland, Amsterdam, 2000.
5. S.C. Brenner. An additive analysis of multiplicative Schwarz methods. *Numer. Math.*, 123:1–19, 2013.
6. S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods* (*Third Edition*). Springer-Verlag, New York, 2008.
7. L. Chen. Deriving the X-Z identity from auxiliary space method. In *Domain decomposition methods in science and engineering XIX*, volume 78 of *Lect. Notes Comput. Sci. Eng.*, pages 309–316. Springer, Heidelberg, 2011.
8. M. Dryja and O.B. Widlund. An additive variant of the Schwarz alternating method in the case of many subregions. Technical Report 339, Department of Computer Science, Courant Institute, 1987.
9. M. Griebel and P. Oswald. On the abstract theory of additive and multiplicative Schwarz algorithms. *Numer. Math.*, 70:163–180, 1995.
10. P.-L. Lions. On the Schwarz alternating method. I. In *First International Symposium on Domain Decomposition Methods for Partial Differential Equations* (*Paris, 1987*), pages 1–42. SIAM, Philadelphia, PA, 1988.
11. T.P.A. Mathew. *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations*. Springer-Verlag, Berlin, 2008.
12. S. Nepomnyaschikh. On the application of the bordering method to the mixed boundary value problem for elliptic equations and on mesh norms in $W_2^{1/2}(S)$. *Sov. J. Numer. Anal. Math. Modelling*, 4:493–506, 1989.
13. Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2003.
14. B. Smith, P. Bjørstad, and W. Gropp. *Domain Decomposition*. Cambridge University Press, Cambridge, 1996.
15. A. Toselli and O.B. Widlund. *Domain Decomposition Methods - Algorithms and Theory*. Springer, New York, 2005.
16. J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581–613, 1992.
17. J. Xu and L. Zikatanov. The method of alternating projections and the method of subspace corrections in Hilbert space. *J. Amer. Math. Soc.*, 15:573–597 (electronic), 2002.
18. H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numerica*, 2:285–326, 1993.
19. X. Zhang. Multilevel Schwarz methods. *Numer. Math.*, 63:521–539, 1992.

Scalable cardiac electro-mechanical solvers and reentry dynamics

P. Colli Franzone, L. F. Pavarino, S. Scacchi, and S. Zampini

Abstract We present a scalable solver for the three-dimensional cardiac electro-mechanical coupling (EMC) model, which represents, currently, the most complete mathematical description of the interplay between the electrical and mechanical phenomena occurring during a heartbeat. The most computational demanding parts of the EMC model are: the electrical current flow model of the cardiac tissue, called Bidomain model, consisting of two non-linear partial differential equations of reaction-diffusion type; the quasi-static finite elasticity model for the deformation of the cardiac tissue. Our finite element parallel solver is based on: Block Jacobi and Multilevel Additive Schwarz preconditioners for the solution of the linear systems deriving from the discretization of the Bidomain equations; Newton-Krylov-Algebraic-Multigrid or Newton-Krylov-BDDC algorithms for the solution of the non-linear algebraic system deriving from the discretization of the finite elasticity equations. Three-dimensional numerical test on two linux clusters show the effectiveness and scalability of the EMC solver in simulating both physiological and pathological cardiac dynamics.

P. Colli Franzone
University of Pavia, Dept. of Mathematics, Via Ferrata 5, 27100 Pavia, Italy, e-mail: colli@imati.cnr.it

L. F. Pavarino
University of Pavia, Dept. of Mathematics, Via Ferrata 5, 27100 Pavia, Italy, e-mail: luca.pavarino@unipv.it

S. Scacchi
University of Milano, Dept. of Mathematics, Via Saldini 50, 20133 Milano, Italy, e-mail: simone.scacchi@unimi.it

S. Zampini
Extreme Computing Research Center, Computer Electrical and Mathematical Sciences & Engineering Department, King Abdullah University of Science and Technology, Saudi Arabia e-mail: stefano.zampini@kaust.edu.sa

1 Introduction

In the last twenty years, computer modeling has become an effective tool to push forward the understanding of the fundamental mechanisms underlying the origin of life-threatening arrhythmias and contractile disorders in the human heart and to provide theoretical support to cardiologists in developing more successful pharmacological and surgical treatments for these pathologies.

The spread of the electrical impulse in the cardiac muscle and the subsequent contraction-relaxation process are quantitatively described by the cardiac electromechanical coupling (EMC) model, which consists of the following four components: the quasi-static finite elasticity model of the deforming cardiac tissue, derived from a strain energy function which characterizes the anisotropic mechanical properties of the myocardium; the active tension model, consisting of a system of non-linear ordinary differential equations (ODEs), describing the intracellular calcium dynamics and cross bridges binding; the electrical current flow model of the cardiac tissue, called Bidomain model, which is a degenerate parabolic system of two non-linear partial differential equations of reaction-diffusion type, describing the evolution in space and time of the intra- and extracellular electric potentials; the membrane model of the cardiac myocyte, i.e. a stiff system of ODEs, describing the flow of the ionic currents through the cellular membrane.

This complex non-linear model poses great theoretical and numerical challenges. At the numerical level, the approximation and simulation of the cardiac EMC model is a very demanding and expensive task, because of the very different space and time scales associated with the electrical and mechanical models, as well as their non-linear and multiphysics interactions.

In this paper, we present the finite element solver that we have developed to simulate the cardiac electro-mechanical activity on parallel computational platforms. The solver is based on a Multilevel Additive Schwarz preconditioner for the linear system arising from the discretization of the Bidomain model and on a Newton-Krylov-BDDC method for the non-linear system arising from the discretization of finite elasticity. Three-dimensional numerical tests show the effectiveness and scalability of the solver on Linux clusters, in both normal physiological and pathological situations.

2 Cardiac electro-mechanical models

a) Mechanical model of cardiac tissue. The deformation of the cardiac tissue is described by the equations of three-dimensional non-linear elasticity

$$\mathrm{Div}(\mathbf{FS}) = \mathbf{0}, \qquad \mathbf{X} \in \widehat{\Omega}, \tag{1}$$

where $\mathbf{X} = (X_1, X_2, X_3)^T$ are the material coordinates of the undeformed cardiac domain $\widehat{\Omega}$ ($\mathbf{x} = (x_1, x_2, x_3)^T$ are the spatial coordinates of the deformed cardiac do-

main $\Omega(t)$ at time t), and $\mathbf{F}(\mathbf{X},t) = \frac{\partial \mathbf{x}}{\partial \mathbf{X}}$ is the deformation gradient. The second Piola-Kirchoff stress tensor $\mathbf{S} = \mathbf{S}^{pas} + \mathbf{S}^{vol} + \mathbf{S}^{act}$ is assumed to be the sum of passive, volumetric and active components. The passive and volumetric components are defined as

$$S_{ij}^{pas,vol} = \frac{1}{2}\left(\frac{\partial W^{pas,vol}}{\partial E_{ij}} + \frac{\partial W^{pas,vol}}{\partial E_{ji}}\right) \quad i,j = 1,2,3, \tag{2}$$

where $\mathbf{E} = \frac{1}{2}(\mathbf{C} - \mathbf{I})$ and $\mathbf{C} = \mathbf{F}^T\mathbf{F}$ are the Green-Lagrange and Cauchy strain tensors, W^{pas} is an exponential strain energy function (derived from [7]) modeling the myocardium as an orthotropic (or transversely isotropic) hyperelastic material, and $W^{vol} = K(J-1)^2$ is a volume change penalization term accounting for the nearly incompressibility of the myocardium, with K a positive bulk modulus and $J = det(\mathbf{F})$.

b) Mechanical model of active tension. The active component of the stress tensor is given by $\mathbf{S}^{act} = T_a \frac{\widehat{\mathbf{a}}_l \otimes \widehat{\mathbf{a}}_l}{\widehat{\mathbf{a}}_l^T \mathbf{C} \widehat{\mathbf{a}}_l}$, where $\widehat{\mathbf{a}}_l$ is the fiber direction and $T_a = T_a\left(Ca_i, \lambda, \frac{d\lambda}{dt}\right)$ is the fiber active tension, obtained by solving a biochemical differential system depending on intracellular calcium concentrations, the myofiber stretch $\lambda = \sqrt{\widehat{\mathbf{a}}_l^T \mathbf{C} \widehat{\mathbf{a}}_l}$ and stretch-rate $\frac{d\lambda}{dt}$ (see [11]).

c) Bioelectrical model of cardiac tissue: the Bidomain model. The evolution of the cardiac extracellular and transmembrane potentials u_e, v, gating variable $\mathbf{w}$, and ionic concentrations $\mathbf{c}$, is given by the Bidomain model. Its parabolic-elliptic formulation on the deformed configuration $\Omega(t)$ reads:

$$\begin{cases} c_m \dfrac{\partial v}{\partial t} - \mathrm{div}(D_i \nabla(v + u_e)) + i_{ion}(v, \mathbf{w}, \mathbf{c}, \lambda) = i_{app} \\ -\mathrm{div}(D_i \nabla v) - \mathrm{div}((D_i + D_e)\nabla u_e) = 0. \end{cases} \tag{3}$$

In the Lagrangian framework, after the pull-back on the reference configuration $\widehat{\Omega} \times (0,T)$, the Bidomain system becomes

$$\begin{cases} c_m J \left(\dfrac{\partial \widehat{v}}{\partial t} - \mathbf{F}^{-T}\,\mathrm{Grad}\,\widehat{v} \cdot \mathbf{V}\right) - \mathrm{Div}(J\,\mathbf{F}^{-1}\widehat{D}_i\mathbf{F}^{-T}\,\mathrm{Grad}(\widehat{v} + \widehat{u}_e)) + J\, i_{ion}(\widehat{v}, \widehat{\mathbf{w}}, \widehat{\mathbf{c}}, \lambda) = J\,\widehat{i}_{app} \\ -\mathrm{Div}(J\,\mathbf{F}^{-1}\widehat{D}_i\mathbf{F}^{-T}\,\mathrm{Grad}\,\widehat{v}) - \mathrm{Div}(J\,\mathbf{F}^{-1}(\widehat{D}_i + \widehat{D}_e)\mathbf{F}^{-T}\,\mathrm{Grad}\,\widehat{u}_e) = 0, \end{cases} \tag{4}$$

where c_m and i_{ion} are the membrane capacitance and ionic current per unit volume, respectively, and $\mathbf{V} = \dfrac{\partial \mathbf{u}}{\partial t}$ is the rate of deformation; see [4] for the detailed derivation. These two partial differential equations (PDEs) are coupled through the reaction term i_{ion} with the ODE system of the membrane model, given in $\Omega(t) \times (0,T)$ by

$$\frac{\partial \mathbf{w}}{\partial t} - \mathbf{R}_w(v, \mathbf{w}) = 0, \quad \frac{\partial \mathbf{c}}{\partial t} - \mathbf{R}_c(v, \mathbf{w}, \mathbf{c}) = 0. \tag{5}$$

This system is completed by prescribing initial conditions, insulating boundary conditions, and the applied current $\widehat{i}_{app}$. Since the extracellular potential $\widehat{u}_e$ is defined up to a time dependent constant in space, we fix it by imposing that $\widehat{u}_e$ has zero aver-

age on the cardiac domain; see [4] for further details. The orthotropic conductivity tensors in the deformed configuration are given by

$$D_{i,e} = \sigma_t^{i,e}\, I + (\sigma_l^{i,e} - \sigma_t^{i,e})\mathbf{a}_l \otimes \mathbf{a}_l + (\sigma_n^{i,e} - \sigma_t^{i,e})\mathbf{a}_n \otimes \mathbf{a}_n,$$

where $\sigma_l^{i,e}$, $\sigma_t^{i,e}$, $\sigma_n^{i,e}$ are the conductivity coefficients in the intra- and extracellular media measured along and across the fiber direction $\mathbf{a}_l, \mathbf{a}_t, \mathbf{a}_n$.

d) Ionic membrane model and stretch-activated channel current. The ionic current in the Bidomain model (3) is $i_{ion} = \chi I_{ion}$, where χ is the membrane surface to volume ratio and $I_{ion}(v, \mathbf{w}, \mathbf{c}, \lambda) = I_{ion}^m(v, \mathbf{w}, \mathbf{c}) + I_{sac}(v, \mathbf{c}, \lambda)$ is the sum of the ionic term $I_{ion}^m(v, \mathbf{w}, \mathbf{c})$ given by the ten Tusscher model (TP06) consisting of 17 ordinary differential equations, [20, 21], available from the cellML depository (models.cellml.org/cellml), and a stretch-activated current I_{sac}. In this work, we adopt the model of I_{sac} proposed in [13] as the sum of non-selective and selective currents $I_{sac} = I_{ns} + I_{Ko}$. We will consider two calibrations where the I_{sac} equilibrium potential (denoted in the following by V_{sac}, i.e. the value such that $I_{sac}(V_{sac}) = 0$) is either $V_{sac} = -60\ mV$ or $V_{sac} = -19\ mV$. We recall that, for $v > V_{sac}$, the stretch-activated current I_{sac} is positive, thus it has a hyperpolarizing effect, while, for $v < V_{sac}$, I_{sac} is negative, resulting in a depolarizing effect. For further details, we refer to [5].

3 Numerical methods

Space discretization. We discretize the cardiac domain with a hexahedral structured grid T_{h_m} for the mechanical model (1) and T_{h_e} for the electrical Bidomain model (4), where T_{h_e} is a refinement of T_{h_m}. We then discretize all scalar and vector fields of both mechanical and electrical models by isoparametric Q_1 finite elements in space.

Time discretization. The time discretization is performed by a semi-implicit splitting method, where the electrical and mechanical time steps can be different. At the $n-$th time step,

a) given v^n, w^n, c^n, solve the ODE system of the membrane model with a first-order IMEX method to compute the new w^{n+1}, c^{n+1}.

b) given the calcium concentration Ca_i^{n+1}, which is included in the concentration variables c^{n+1}, solve the mechanical problems (1) and the active tension differential system to compute the new deformed coordinates $\mathbf{x}^{n+1}$, providing the new deformation gradient tensor $\mathbf{F}_{n+1}$.

c) given w^{n+1}, c^{n+1}, $\mathbf{F}_{n+1}$ and $J_{n+1} = \det(\mathbf{F}_{n+1})$, solve the Bidomain system (4) with a first order IMEX method and compute the new electric potentials v^{n+1}, u_e^{n+1} with an operator splitting method, where the parabolic and elliptic PDEs are decoupled; see [4] for further details.

4 Parallel solver

4.1 Computational kernels

Due to the discretization strategies described above, the main computational kernels of our solver at each time step are the following:

1- solve the non-linear system deriving from the discretization of the mechanical problem (1) using an inexact Newton method. At each Newton step, a non-symmetric Jacobian system $Kx = f$ is solved inexactly by the GMRES iterative method preconditioned by a BDDC preconditioner, described in the next section.
2- solve the two linear systems deriving from the discretization of the parabolic and elliptic equations of the Bidomain model, by using the Conjugate Gradient method preconditioned by the Block Jacobi and Multilevel Additive Schwarz preconditioners, respectively, developed in [14].

4.2 Mechanical solver

Schur Complement System. To keep the notation simple, in the remainder of this section and the next, we denote the reference domain by Ω instead of $\widehat{\Omega}$. Let us consider a decomposition of Ω into N nonoverlapping subdomains Ω_i of diameter H_i (see e.g. [22, Ch. 4]) $\Omega = \bigcup_{i=1}^{N} \Omega_i$, and set $H = \max H_i$. As in classical iterative substructuring, we reduce the problem to the interface $\Gamma := \left(\bigcup_{i=1}^{N} \partial\Omega_i\right) \backslash \partial\Omega$ by eliminating the interior degrees of freedom associated to basis functions with support in the interior of each subdomain, hence obtaining the Schur complement system

$$S_\Gamma x_\Gamma = g_\Gamma, \tag{6}$$

where $S_\Gamma = K_{\Gamma\Gamma} - K_{\Gamma I} K_{II}^{-1} K_{I\Gamma}$ and $g_\Gamma = f_\Gamma - K_{\Gamma I} K_{II}^{-1} f_I$ are obtained from the original discrete problem $Kx = f$ by reordering the finite element basis functions in interior (subscript I) and interface (subscript Γ) basis functions.
BDDC preconditioner. The Schur complement system (6) is solved iteratively by the GMRES method using a BDDC preconditioner M_{BDDC}^{-1}

$$M_{BDDC}^{-1} S_\Gamma x_\Gamma = M_{BDDC}^{-1} g_\Gamma. \tag{7}$$

Once the interface solution x_Γ is computed, the internal values x_I can be recovered by solving local problems on each subdomain Ω_i.

BDDC preconditioners represent an evolution of balancing Neumann-Neumann methods where all local and coarse problems are treated additively due to a choice of so-called primal continuity constraints across the interface of the subdomains. These primal constraints can be point constraints and/or averages or moments over edges or faces of the subdomains. BDDC preconditioners were introduced in [6] and first

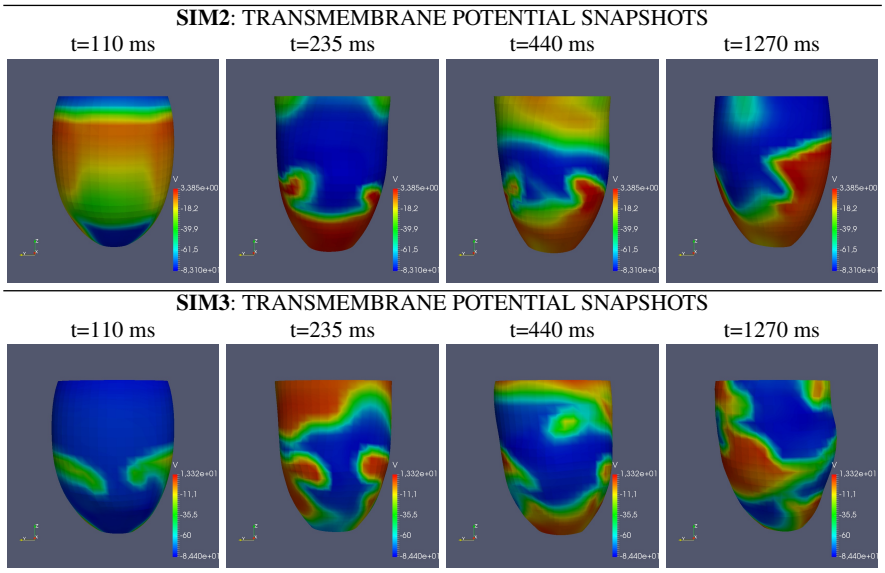


Fig. 1 Test 1: Snapshots of transmembrane potentials computed from **SIM2** (ventricular tachycardia) and **SIM3** (ventricular fibrillation). The units in the colorbars are given in mV.

analyzed in [12]. We remark that BDDC is closely related to FETI-DP algorithms, see, e.g. [10, 9], defined with the same set of primal constraints as BDDC, since it is known that in such a case the BDDC and FETI-DP operators have the same eigenvalues with the exception of zeros and ones. For the construction of BDDC preconditioners applied to the non-linear elasticity system constituting the cardiac electromechanical coupling problem, we refer to [16].

5 Numerical Results

In this section, we present the results of parallel numerical experiments performed on the Linux cluster Marconi (http://www.hpc.cineca.it/hardware/marconi) of the Cineca Consortium (www.cineca.it). Our code is built on top of the FORTRAN90 wrappers of the open source PETSc library [1]. In the mechanical solver, at each Newton iteration, the non-symmetric Jacobian system is solved iteratively by GMRES preconditioned by the BoomerAMG or the BDDC preconditioner, with zero initial guess and stopping criterion a 10^{-8} reduction of the relative residual l_2-norm. The BDDC method is available as a preconditioner in PETSc and it has been contributed to the library by S. Zampini, see [25].

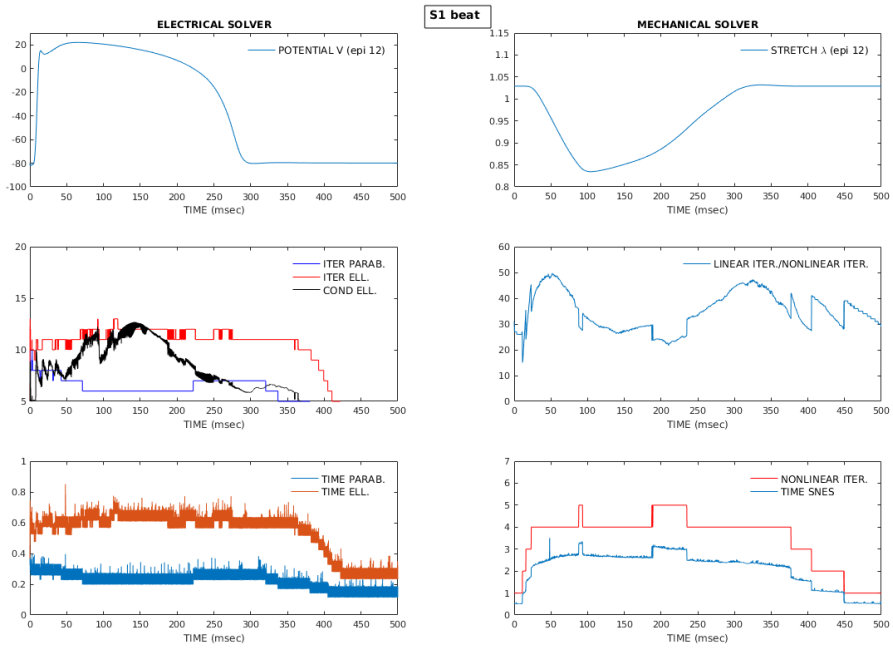


Fig. 2 S1 beat of physiological test **SIM1** over 500 msec.: time plots at an epicardial point of the indicated electrical (left) and mechanical (right) quantities

5.1 Test 1: comparison of solver performance on normal and pathological dynamics

We consider an idealized left ventricle, represented by a truncated ellipsoid discretized by an electrical grid of $384 \times 192 \times 48$ Q^1 finite elements, yielding a total amount of about $3.6 \cdot 10^6$ nodes, thus the degrees of freedom (dofs) of the parabolic and elliptic Bidomain linear systems are $3.6 \cdot 10^6$. The mechanical mesh is eight times coarser than the electrical one, i.e. $48 \times 24 \times 6$ Q^1 finite elements, with a total amount of 8400 nodes, thus the dofs of the finite elasticy non-linear system are 25200. The electrical time step is 0.05 *ms*, while the mechanical time step is 0.5 *ms*. The simulations are run on 24 processors. The tissue is assumed to be axisymmetric. The mechanical non-linear system is solved by the Newton-Krylov-AMG method.

We first compare the performance of the electro-mechanical solver in three different situations:

- a normal physiological heartbeat (**SIM1**) without reentry;
- a ventricular tachycardia dynamics (**SIM2**), with $V_{sac} = -19\ mV$;
- a ventricular fibrillation dynamics (**SIM3**), with $V_{sac} = -60\ mV$.

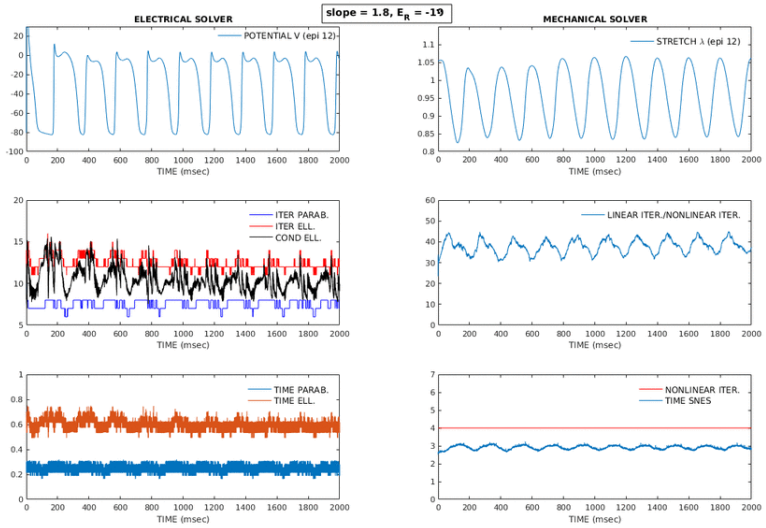


Fig. 3 Periodic test **SIM2** with slope = 1.8, $V_{sac} = -19\ mV$ over 2000 msec.: time plots at an epicardial point of the indicated electrical (left) and mechanical (right) quantities

In **SIM1**, the external stimulus is applied at the endocardial apical region, the interior bottom part of the truncated ellipsoid, and the total simulation run is 500 *ms*. The activation wavefront propagates starting from the endocardial apical regions, where the stimulus is delivered, towards the whole ventricle (not shown, but similar to the propagation displayed in Fig. 6).

In **SIM2** and **SIM3**, we apply first an S1 stimulus as in **SIM1**. 280 *ms* after the S1 stimulus is delivered, we apply a premature S2 cross-gradient stimulation current from the base to the apex and across the wall thickness, covering about a third of ventricular volume, to induce a ventricular reentry consisting of a pair of counter-rotating scroll waves. We run the simulation for 2000 *ms* after the S2 delivery. The SAC parameter V_{sac} is set to $-19\ mV$ and $-60\ mV$ is **SIM2** and **SIM3**, respectively.

In **SIM2**, the two scroll waves generated by the S2 stimulus continue to rotate without breaking, leading to a stable periodic ventricular tachycardia pattern, see Fig. 1.

In **SIM3** instead, after the first rotation, the two scroll waves break up into several smaller scroll waves, generating irregular transmembrane potential distributions characterized by high electrical turbulence, often associated with ventricular fibrillation, as shown in the snapshots of Fig. 1. Thus, the low SAC reversal potential ($V_{sac} = -60\ mV$) seems to induce deterioration of the stability of scroll waves, promoting the onset of ventricular fibrillation.

Figures 2, 3, 4 report the time evolution of the mathematical parameters of the electro-mechanical solver (CG iterations, condition numbers, Newton iterations, GMRES iterations) and the CPU times needed to solve the parabolic, elliptic and

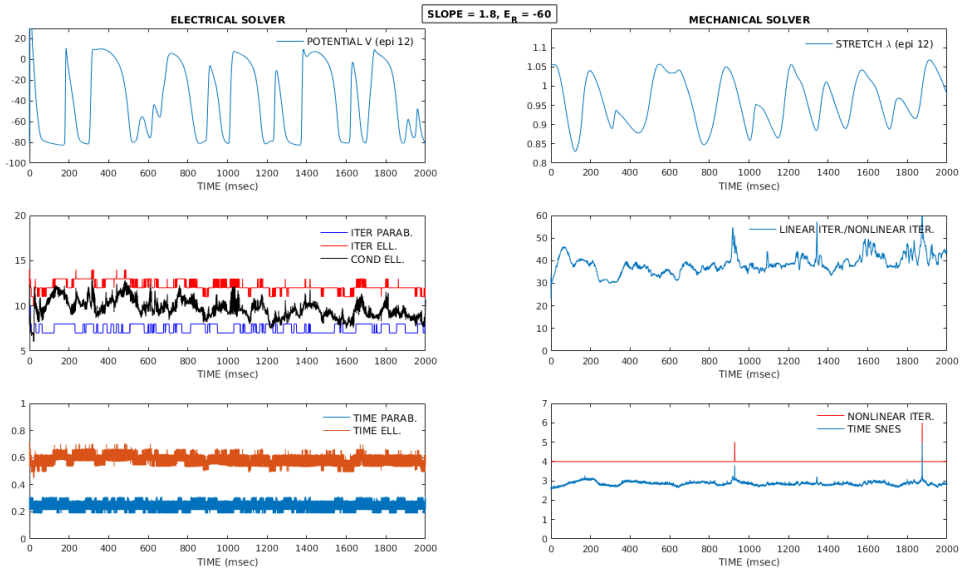


Fig. 4 Turbulent test **SIM3** with slope = 1.8, $V_{sac} = -60\ mV$ over 2000 msec.: time plots at an epicardial point of the indicated electrical (left) and mechanical (right) quantities

$proc$	it_{par}	$time_{par}$	it_{ell}	$time_{ell}$	$\Pi = V$			$\Pi = VE$		
					nit	lit	$time_{snes}$	nit	lit	$time_{snes}$
32	3	2.24e-1	20	9.56e-1	4	39	12.90	4	38	13.23
64	3	1.24e-1	20	5.37e-1	4	48	5.03	4	47	5.36
128	3	7.71e-2	20	3.17e-1	4	48	3.67	4	47	3.50
256	3	3.78e-2	20	2.40e-1	4	45	2.55	4	44	2.88

Table 1 Strong scaling test on a whole heartbeat simulation. it_{par}: CG iteration to solve the parabolic linear system (average per time step). $time_{par}$: CPU time to solve the parabolic linear system (average per time step). it_{ell}: CG iteration to solve the elliptic linear system (average per time step). $time_{ell}$: CPU time to solve the elliptic linear system (average per time step). nit: Newton iteration to solve the mechanical system (average per time step). lit: GMRES iteration to solve the Jacobian system (average per Newton iteration). $time_{snes}$: CPU time to solve the mechanical system (average per time step). All CPU times are given in seconds.

non-linear systems (TIME PARAB., TIME ELL., TIME SNES, respectively) obtained from the **SIM1**, **SIM2**, **SIM3**, respectively. The results show that all the components of the solver are quite robust with respect to the different simulation dynamics considered, physiological and pathological. The condition number of the elliptic solver increases slightly when the contraction is more pronounced, but it always remains bounded betweem 10 and 15.

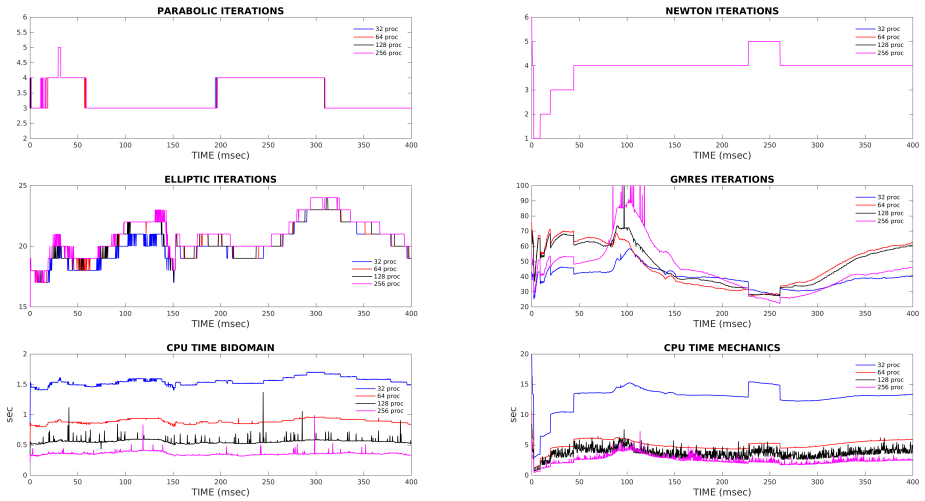


Fig. 5 Strong scaling test on a whole heartbeat simulation. Time evolution of electrical and mechanical solvers parameters.

5.2 Test 2: strong scaling on a normal heartbeat

We then perform a strong scaling test on a whole heartbeat lasting 400 *ms*. The three-dimensional cardiac domain considered is a truncated ellipsoid modeling the left ventricle, discretized by an electrical mesh of $384 \cdot 192 \cdot 48$ Q_1 finite elements, yielding the same Bidomain dofs as in the previous test, about $3.6 \cdot 10^6$. The mechanical mesh size is now four times coarser than the electrical one in each direction, thus the mechanical elements are $96 \cdot 48 \cdot 12$, resulting in 183456 displacement dofs. The number of subdomains (processors) increases from 32 to 256 whereas the number of degrees of freedom per subdomain is reduced as the number of subdomains increases. The tissue is assumed to be orthotropic. The mechanical non-linear system is solved by the Newton-Krylov-BDDC method. We choose as BDDC primal constraints vertices ($\Pi = V$) and vertices + edges ($\Pi = VE$). To start the electrical excitation, the external stimulus is applied at the endocardial apical region, in four points modeling an idealized Purkinje network.

Fig. 6 reports selected snapshots of transmembrane and extracellular potentials on the deforming domain during the entire heartbeat. The results reported in Table 1 (averages) and Fig. 5 (time evolution) show a good scalability of both the electrical and mechanical components of the parallel solver, with linear and non-linear iterations remaining about constant, while the CPU times decrease when the number of processors increases.

References

1. S. Balay et al. PETSc users manual. Tech. Rep. ANL-95/11 - Revision 3.3, Argonne National Laboratory, 2012.
2. D. Chapelle et al. An energy-preserving muscle tissue model: formulation and compatible discretizations. *J. Multiscale Comput. Engrg.*, 10:189–211, 2012.
3. P. Colli Franzone, L. F. Pavarino, and S. Scacchi. *Mathematical Cardiac Electrophysiology.* Springer, MSA Vol. 13, New York, 2014.
4. P. Colli Franzone, L. F. Pavarino, and S. Scacchi. Bioelectrical effects of mechanical feedbacks in a strongly coupled cardiac electro-mechanical model. *Math. Mod. Meth. Appl. Sci.*, 26:27–57, 2016.
5. P. Colli Franzone, L. F. Pavarino, and S. Scacchi. Effects of mechanical feedback on the stability of cardiac scroll waves: A bidomain electro-mechanical simulation study. *Chaos*, 27: 093905, 2017.
6. C. R. Dohrmann. A preconditioner for substructuring based on constrained energy minimization. *SIAM J. Sci. Comput.*, 25:246–258, 2003.
7. T. S. E. Eriksson et al. Influence of myocardial fiber/sheet orientations on left ventricular mechanical contraction. *Math. Mech. Solids*, 18:592–606, 2013.
8. V. Gurev et al. Models of cardiac electromechanics based on individual hearts imaging data: Image-based electromechanical models of the heart. *Biomech. Model Mechanobiol.*, 10:295–306, 2011.
9. A. Klawonn and O. Rheinbach. Highly scalable parallel domain decomposition methods with an application to biomechanics. *ZAMM-Z. Angew. Math. Mech.*, 90:5–32, 2010.
10. A. Klawonn and O. B. Widlund. Dual-primal FETI methods for linear elasticity. *Comm. Pure Appl. Math.*, 59:1523–1572, 2006.
11. S. Land et al. An analysis of deformation-dependent electromechanical coupling in the mouse heart. *J. Physiol.*, 590:4553–4569, 2012.
12. J. Mandel and C. R. Dohrmann. Convergence of a balancing domain decomposition by constraints and energy minimization. *Numer. Lin. Alg. Appl.*, 10:639–659, 2003.
13. S. A. Niederer and N. P. Smith, A mathematical model of the slow force response to stretch in rat ventricular myocites. *Biophys. J.*, 92: 4030–4044,2007.
14. L. F. Pavarino and S. Scacchi. Multilevel additive Schwarz preconditioners for the Bidomain reaction-diffusion system. *SIAM J. Sci. Comput.*, 31:420–443, 2008.
15. L. F. Pavarino, S. Zampini, and O.B. Widlund. BDDC preconditioners for spectral element discretizations of almost incompressible elasticity in three dimensions. *SIAM J. Sci. Comput.*, 32 (6):3604–3626, 2010.
16. L. F. Pavarino, S. Scacchi, and S. Zampini. Newton-krylov-BDDC solvers for non-linear cardiac mechanics. *Comput. Meth. Appl. Mech. Engrg.*, 295:562–580, 2015.
17. G. Plank et al. Algebraic Multigrid Preconditioner for the cardiac bidomain model. *IEEE Trans. Biomed. Engrg.*, 54:585–596, 2007.
18. S. Rossi et al. Orthotropic active strain models for the numerical simulation of cardiac biomechanics. *Int. J. Num. Meth. Biomed. Engrg.*, 28:761–788, 2012.
19. J. Sundnes et al. Improved discretisation and linearisation of active tension in strongly coupled cardiac electro-mechanics simulations. *Comput. Meth. Biomech. Biomed. Engrg.*, 17:604–615, 2014.
20. K. H. W. J. ten Tusscher et al. A model for human ventricular tissue. *Am. J. Phys. Heart. Circ. Physiol.*, 286:H1573–H1589, 2004.
21. K. H. W. J. ten Tusscher and A. V. Panfilov. Alternans and spiral breakup in a human ventricular tissue model. *Am. J. Phys. Heart Circ. Physiol.*, 291: H1088–H1100, 2006.
22. A. Toselli and O. B. Widlund. *Domain Decomposition Methods: Algorithms and Theory*. Springer-Verlag, Berlin, 2004.
23. F. J. Vetter and A. D. McCulloch. Three-dimensional stress and strain in passive rabbit left ventricle: a model study. *Ann. Biomed. Engrg.*, 28:781–792, 2000.

24. S. Zampini. Dual-primal methods for the cardiac bidomain model. *Math. Mod. Meth. Appl. Sci.*, 24:667–696, 2014.
25. S. Zampini. PCBDDC: a class of robust dual-primal preconditioners in PETSc. *SIAM J. Sci. Comput.*, 38(5):S282–S306, 2016.

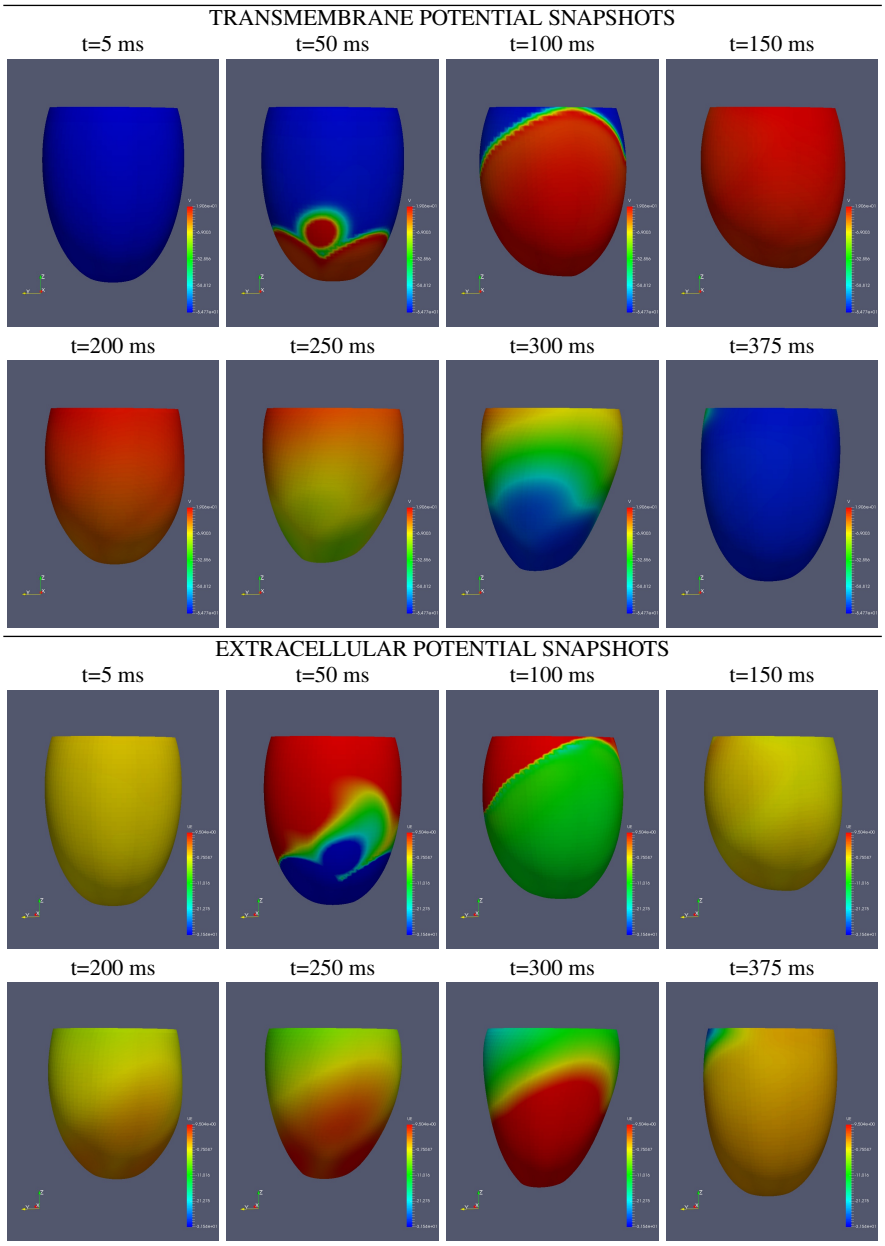


Fig. 6 Snapshots of transmembrane and extracellular potentials during a whole heartbeat. The units in the colorbars are given in mV.

On overlapping domain decomposition methods for high-contrast multiscale problems

Juan Galvis[1], Eric Chung[3], Yalchin Efendiev[2], and Wing Tat Leung[2]

1 Summary

We review some important ideas in the design and analysis of robust overlapping domain decomposition algorithms for high-contrast multiscale problems. In recent years, there have been many contributions to the application of different domain decomposition methodologies to solve high-contrast multiscale problems. We mention two- and multi-levels methods, additive and additive average methods, iterative substructuring and non-overlapping methods and many others. See [11]. Due to page limitation, we focus only on two-levels overlapping methods developed by some of the authors that use a coarse-grid for the construction of the second level. We also propose a domain decomposition method with better performance in terms of the number of iterations. The main novelty of our approaches is the construction of coarse spaces, which are computed using spectral information of local bilinear forms. We present several approaches to incorporate the spectral information into the coarse problem in order to obtain minimal (locally constructed) coarse space dimension. We show that using these coarse spaces, we can obtain a domain decomposition preconditioner with the condition number independent of contrast and small scales. To minimize further the number of iterations until convergence, we use this minimal dimensional coarse spaces in a construction combining them with large overlap local problems that take advantage of the possibility of localizing global fields orthogonal to the coarse space. We obtain a condition number close to 1 for the new method. We discuss possible drawbacks and further extensions.

[1]Departamento de Matemáticas, Universidad Nacional de Colombia, Bogotá, Colombia. [2]Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, USA. [3]Department of Mathematics, The Chinese University of Hong Kong, Hong Kong SAR.

2 High-contrast problems. Introduction

The methods and algorithms, discussed in the paper, can be applied to various PDEs, even though we will focus on Darcy flow equations. Given $D \subset \mathbb{R}^2$, $f : D \to \mathbb{R}$, and $g : \partial D \to \mathbb{R}$, find $u : D \to \mathbb{R}$ such that

$$\frac{\partial}{\partial x_i}\left(\kappa_{ij}\frac{\partial u}{\partial x_j}\right) = f$$

with a suitable boundary condition, for instance $u = 0$ on ∂D. The coefficient $\kappa_{ij}(x) = \kappa(x)\delta_{ij}$ represents the permeability of the porous media D. We focus on two-levels overlapping domain decomposition and use local spectral information in constructing "minimal" dimensional coarse spaces (MDCS) within this setting. After some review on constructing MDCS and their use in overlapping domain decomposition preconditioners, we present an approach, which uses MDCS to minimize the condition number to a condition number closer to 1. This approach requires a large overlap (when comparted to coarse-grid size) and, thus, is more efficient for small size coarse grids. We present the numerical results and state our main theoretical result. We assume that there exists $\kappa_{\min}$ and $\kappa_{\max}$ with $0 < \kappa_{\min} \leq \kappa(x) \leq \kappa_{\max}$ for all $x \in D$. **The coefficient κ has a multiscale structure** (significant local variations of κ occur across D at different scales). We also assume that the coefficient κ is a high-contrast coefficient (the constrast is $\eta = \kappa_{\max}/\kappa_{\min}$). We assume that η is large compared to the coarse-grid size.

It is well known that performance of numerical methods for high-contrast multiscale problems depends on η and local variations of κ across D. For classical finite element methods, the condition to obtain good approximation results is that the finite element mesh has to be fine enough to resolve the variations of the coefficient κ. Under these conditions, finite element approximation leads to the solution of very large (sparse) ill-conditioned problems (with the condition number scaling with h^{-2} and η). Therefore, the performance of solvers depends on η and local variations of κ across D. This was observed in several works, e.g., [8, 10, 1][1].

Let $\mathcal{T}^h$ be a triangulation of the domain D, where h is the size of typical element. We consider only the case of discretization by the classical finite element method $V = P_1(\mathcal{T}^h)$ of piecewise (bi)linear functions. Other discretizations can also be considered. The application of the finite element discretization leads to the solution of a very large ill-conditioned system $Ax = b$, where A is roughly of size h^{-2} and the condition number of A scales with η and h^{-2}. In general, the main goal is to obtain an efficient good approximation of solution u. The two main solution strategies are:

1. Choose h sufficiently small and implement an iterative method. It is important to implement a preconditioner M^{-1} to solve $M^{-1}Au = M^{-1}b$. Then, it is important to have the condition number of $M^{-1}A$ to be small and bounded independently of physical parameters, e.g., η and the multiscale structure of κ.

[1] Due to the page limitation, only a few references are cited throughout.

2. Solve a smaller dimensional linear system ($\mathcal{T}^H$ with $H > h$) so that computations of solutions can be done efficiently[2]. This usually involves the construction of a downscaling operator R_0 (from the coarse-scale to fine-scale $v_0 \mapsto v$) and an upscaling operator (from fine-scale to coarse-scale, $v \mapsto v_0$) (or similar operators). Using these operators, the linear system $Au = b$ becomes a coarse linear system $A_0 u_0 = b_0$ so that $R_0 u_0$ or functionals of it can be computed. The main goal of this approach it to obtain a sub-grid capturing such that $||u - R_0 u_0||$ is small.

The rest of the paper will focus on the design of overlapping domain decomposition methods by constructing appropriate coarse spaces. First, we will review existing results, which construct minimal dimensional coarse spaces, such that the condition number of resulting preconditioner is independent of η. These coarse spaces use local spectral problems to extract the information, which cannot be localized. This information is related to high-conductivity channels, which connect coarse-grid boundaries and it is important for the performance of domain decomposition preconditioners and multiscale simulations. Next, using these and oversampling ideas, we present a "hybrid" domain decomposition approach with a condition number close to 1 by appropriately selecting the oversampling size (i.e., overlapping size). We state our main result, discuss some limitations and show a numerical example. We compare the results to some existing contrast-independent preconditioners.

3 Classical overlapping methods. Brief review

We start with a non-overlapping decomposition $\{D_i\}_{i=1}^{N_S}$ of the domain D and obtain an overlapping decomposition $\{D_i'\}_{i=1}^{N_S}$ by adding a layer of width δ around each non-overlapping subdomain. Let A_j be the Dirichlet matrix corresponding to the overlapping subdomain D_j'. The one level method solves $M_1^{-1}A = M_1^{-1}b$ with $M_1^{-1} = \sum_{j=1}^{N_S} R_j (A_j)^{-1} R_j^T$ and the operators R_j^T, $j = 1, \ldots, N_S$, being the restriction to overlapping subdomain D_j' operator and with the R_j being the extension by zero (outside D_j') operator. We have the bound $\mathrm{Cond}(M_1^{-1}A) \leq C\,(1 + 1/\delta H)$. For high-contrast multiscale problems, it is known that $C \asymp \eta$.

Next, we introduce a coarse space, that is, a subspace $V_0 \subset V$ of small dimension (when compared to the fine-grid finite element space V). We consider A_0 as the matrix form of the discretization of the equation related to subspace V_0. For simplicity of the presentation, let A_0 be the Galerkin projection of A on the subspace V_0. That is $A_0 = R_0 A R_0^T$, where R_0 is a downscaling operator that converts coarse-space coordinates into fine-grid space coordinates. The two-levels preconditioner uses the coarse space and it is defined by $M_2^{-1} = R_0 A_0^{-1} R_0^T + \sum_{j=1}^{N_S} R_j (A_j)^{-1} R_j^T = R_0 A_0^{-1} R_0^T + M_1^{-1}$. It is known that $\mathrm{Cond}(M^{-1}A) \preceq \eta\,(1 + H/\delta)$. The classical two-levels method is robust with respect to the number of subdomains but it is not robust with respect to η.

[2] The coarse mesh does not necessarily resolve all the variations of κ.

The condition number estimates use a Poincaré inequality and small overlap trick; [13]. Without the small overlap trick $\text{Cond}(M^{-1}A) \preceq \eta(1 + H^2/\delta^2)$.

There were several works addressing the performance of classical domain decomposition algorithms for high-contrast problems. Many of these works considered simplified multiscale structures[3], see e.g., [13] for some works by O. Widlund and his collaborators. We also mention the works by Sarkis and his collaborators, where they introduce the assumption of quasi-monotonicity [4]. Sarkis also introduced the idea of using "extra" or additional basis functions as well as techniques that construct the coarse spaces using the overlapping decomposition (and not related to a coarse mesh); [12]. Scheichl and Graham [10] and Hou and Aarnes [1], started a systematic study of the performance of classical overlaping domain decomposition methods for high-contrast problems. In their works, they used coarse spaces constructed using a coarse grid and special basis functions from the family of multiscale finite element methods. These authors designed two-levels domain decomposition methods that were robust (with respect to η) for special multiscale structures. None of the results available in the literature (before the method in papers [8, 9] was introduced) were robust for a coefficient not-aligned with the construction of the coarse space (i.e., not aligned either with the non-overlapping decomposion or the coarse mesh if any), i.e., the condition number of the resulting preconditioner is independent of η for general multiscale coefficients.

4 Stable decomposition and eigenvalue problem. Review

A main tool in obtaining condition number bounds is the construction of a stable decomposition of a global field. That is, if for all $v \in V = P_0^1(D, \mathcal{T}^h)$ there exists a decomposition $v = v_0 + \sum_{j=1}^{N_S} v_j$ with $v_0 \in V_0$ and $v_j \in V_j = P_0^1(D'_j, \mathcal{T}^h)$, $j = 1, \ldots, N$, and

$$\int_D \kappa |\nabla v_0|^2 + \sum_{j=1}^{N_S} \int_{D'_j} \kappa |\nabla v_j|^2 \leq C_0^2 \int_D \kappa |\nabla v|^2$$

for $C_0 > 0$. Then, $\text{cond}(M_2^{-1}A) \leq c(\mathcal{T}^h, \mathcal{T}^H) C_0^2$. Existence of a suitable coarse interpolation $I_0 : V \to V_0 = \text{span}\{\Phi\}$ implies the stable decomposition above. Usually such stable decomposition is constructed as follows.

For the coarse part of the stable decomposition, we introduce a partition of unity $\{\chi_i\}$ subordinated to the coarse mesh (supp $\chi_i \subset \omega_i$ where ω_i is the coarse-block neighborhood of the coarse-node x_i). We begin by restricting the global field v to ω_i. For each coarse node neighborhood ω_i, we identify local field that will contribute to the coarse space $I_0^{\omega_i} v$ so that the coarse space will be defined as $V_0 = \text{Span}\{\chi_i I_0^{\omega_i} v\}$. In classical methods $I_0^{\omega_i} v$ is the average of v in ω_i. Later we present some more general examples for $I_0^{\omega_i}$. We

[3] These works usually assume some alignment between the coefficient heterogeneities and the initial non-overlapping decomposition.

assemble a coarse field as $v_0 = I_0 v = \sum_{i=1}^{N_S} \chi_i (I_0^{\omega_i} v)$. Note that in each block $v - v_0 = \sum_{x_i \in K} \chi_i (v - I_0^{\omega_i} v)$.

For the local parts of the stable decomposition, we introduce a partition of unity $\{\xi_j\}$ subordinated to the non-overlapping decomposition (supp $\xi_j \subset D'_j$). The local part of the stable decomposition is defined by $v_j = \xi_j (v - v_0)$. For instance, to bound the energy of v_j, we have in each coarse-block K,

$$\int_K \kappa |\nabla v_j|^2 \preceq \int_K \kappa |\nabla \xi_j \left(\sum_{x_i \in K} \chi_i (v - I_0^{\omega_i} v) \right)|^2$$
$$\preceq \sum_{i \in K} \int_K \kappa (\xi_j \chi_i)^2 |\nabla (v - I_0^{\omega_i} v)|^2 + \sum_{x_i \in K} \int_K \kappa |\nabla (\xi_j \chi_i)|^2 |v - I_0^{\omega_i} v|^2.$$

Adding up over K, we obtain,

$$\int_{D'_j} \kappa |\nabla v_j|^2 \preceq \sum_{x_i \in \overline{D}'_j} \int_{D'_j} \kappa (\xi_j \chi_i)^2 |\nabla (v - I_0^{\omega_i} v)|^2$$
$$+ \sum_{x_i \in \overline{\omega}_j} \int_{D'_j} \kappa |\nabla (\xi_j \chi_i)|^2 |v - I_0^{\omega_i} v|^2$$

and we would like to bound the last term by $C \int_{D'_j} \kappa |\nabla v|^2$.

For simplicity of our presentation, we consider the case when the coarse elements coincide with the non-overlapping decomposition subdomains. That is, $D'_j = \omega_j$. In this case, we can replace ξ by χ and replace $\nabla(\chi^2)$ by $\nabla \chi$ so that we need to bound $\sum_{x_i \in \overline{\omega}_j} \int_{\omega_j} \kappa |\nabla \chi_i|^2 |v - I_0^{\omega_i} v|^2$. We refer to this design as **coarse-grid based**.

Remark 1 (General case and overlapping decomposition based design). Similar analysis holds in the case when there is no coarse-grid and the coarse space is spanned by a partition of unity $\{\xi_j\}$. We can replace χ by ξ and $\nabla(\xi^2)$ by $\nabla \xi$. In general these two partitions are not related (see Sec. 4.1).

We now review the three main arguments to complete the required bound: 1) A Poincaré inequality. 2) L^∞ estimates. 3) Eigenvalue problem.

1. A Poincaré inequality: Classical analysis uses a Poincaré inequality to obtain the required bound above. That is, the inequality $\frac{1}{H^2} \int_\omega (v - \bar{v})^2 \leq C \int_\omega |\nabla v|^2$ to obtain $\sum_{x_i \in \omega_j} \int_{\omega_j} \kappa |\nabla \chi_i|^2 |v - I_0^{\omega_i} v|^2 \preceq \frac{1}{H^2} \int_{\omega_i} \kappa |v - I_0^{\omega_i} v|^2 \preceq C \int_{\omega_i} \kappa |\nabla v|^2$. In this case, $I_0^{\omega_i} v$ is the average of v on the subdomain. For the case of high-contrast coefficients, C depends on η, in general. For quasi-monotone coefficient it can be obtained that C is independent of the contrast [4]. We also mention [8] for the case **locally connected high-contrast region**. In this case $I_0^{\omega_i} v$ is a weighted average. From the argument given in [8], it was clear that when the high-contrast regions break across the domain,

defining only one average was not enough to obtain contrast independent constant in the Poincaré inequality.

2. L^∞ **estimates:** Another idea is to use an L^∞ estimate of the form

$$\sum_{x_i \in K} \int_{\omega_i} \kappa |\nabla \chi_i|^2 |v - I_0^{\omega_i} v|^2 \preceq \sum_{x_i \in K} ||\kappa |\nabla \chi_i|^2||_\infty \int_{\omega_i} |v - I_0^{\omega_i} v|^2.$$

The idea in [10, 1] was then to construct a partition of unity such that $||\kappa|\nabla\chi_i|^2||_\infty$ is bounded independently of the contrast and then to use classical Poincaré inequality estimates. Instead of minimizing the L^∞, one can intuitively try to minimize $\int_K \kappa |\nabla \chi_i|^2$. This works well when the multiscale structure of the coefficient is confined within the coarse blocks. For instance, for a coefficient and coarse-grid as depicted in Figure 1 (left picture), we have that a two-level domain decomposition method can be proven to be robust with respect the value of the coefficient inside the inclusions. In fact, the coarse space spanned by classical multiscale basis functions with linear boundary conditions ($-\text{div}(\kappa \nabla \chi_i) = 0$ in K and linear on each edge of ∂K) is sufficient and the above proof works. Now consider the coefficient in Figure 1 (center picture). For such cases, the boundary condition of the basis functions is important. In these cases, basis functions can be constructed such that the above argument can be carried on. Here, we can use multiscale basis functions with oscillatory boundary condition in its construction[4].

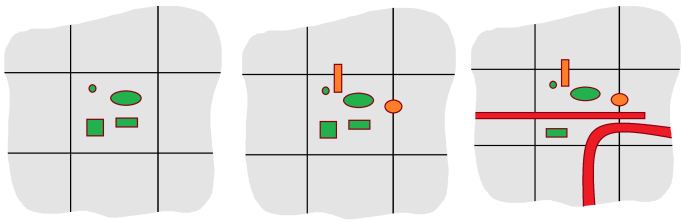


Fig. 1 Examples o multiscale coefficients with interior high-contrast inclusions (left), boundary inclusions (center) and long channels(right).

For the coefficient in Figure 1, right figure, the argument above using L^∞ cannot be carried out unless we can work with larger support basis functions (as large as to include the high-contrast channels of the coefficient). If the support of the coarse basis function does not include the high-contrast region, then $||\kappa|\nabla\chi_i|^2||_\infty$ increases with the contrast leading to non-robust two-level domain decomposition methods.

[4] We can include constructions of boundary conditions using $1D$ solution of the problem along the edges. Other choices include basis functions constructed using oversampling regions, energy minimizing partition of unity (global), constructions using limited global information (global), etc.

3. Eigenvalue problem. We can write $\sum_{x_j \in \omega_i} \int_{\omega_i} \kappa |\nabla \chi_j|^2 |v - I_0^{\omega_i} v|^2 \preceq \frac{1}{H^2} \int_{\omega_i} \kappa |(v - I_0^{\omega_i} v)|^2 \preceq C \int_{\omega_i} \kappa |\nabla v|^2$, where we need to justify the last inequality with constant independent of the contrast. The idea is then to consider the Rayleigh quotient, $\mathcal{Q}(v) := \frac{\int_{\omega_i} \kappa |\nabla v|^2}{\int_{\omega_i} \kappa |v|^2}$ with $v \in P^1(\omega_i)$. This quotient is related to an eigenvalue problem and we can define $I_0^{\omega_i} v$ to be the projection on low modes of this quotient on ω_i. The associated eigenproblem is given by $-\text{div}(k(x)\nabla \psi_\ell) = \lambda_\ell k(x) \psi_i$ in ω_i with homogeneous Neumann boundary condition for floating subdomains and a mixed homogeneous Neumann-Dirichlet condition for subdomains that touch the boundary. It turns out that the low part of the spectrum can be written as $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_L < \lambda_{L+1} \leq ...$ where $\lambda_1, ..., \lambda_L$ are small, asymptotically vanishing eigenvalues and λ_L can be bounded below independently of the contrast. After identifying the local field $I_0^{\omega_i} v$, we then define the coarse space as $V_0 = Span\{I^h \chi_i \psi_j^{\omega_i}\} = Span\{\Phi_i\}$.

Eigenvalue problem with a multiscale partition of unity. Instead of the argument presented earlier, we can include the gradient of the partition of unity in the bounds (somehow similar to the ideas of L^∞ bounds). We then need the following chain of inequalities, $\int_{\omega_i} \underbrace{\left(\sum_{x_j \in \omega_i} \kappa |\nabla \chi_j|^2 \right)}_{:= H^{-2} \widetilde{\kappa}} |v - I_0^{\omega_i} v|^2 = \frac{1}{H^2} \int_{\omega_i} \widetilde{\kappa} |v - I_0^{\omega_i} v)|^2 \preceq \int_{\omega_i} \kappa |\nabla v|^2$. Here we have to consider Rayleigh quotient $\mathcal{Q}_{ms}(v) := \frac{\int_{\omega_i} \kappa |\nabla v|^2}{\int_{\omega_i} \widetilde{\kappa} |v|^2}$, $v \in P^1(\omega_i)$ and define $I_0^{\omega_i} v$ as projection on low modes. Additional modes "complement" the initial space spanned by the partition of unity used so that the resulting coarse space leads to robust methods with minimal dimension coarse spaces; [9].

If we consider the two-level method with the (multiscale) spectral coarse space presented before, then

$$\text{cond}(M^{-1} A) \leq C(1 + (H/\delta)^2), \tag{1}$$

where C is independent of the contrast if enough eigenfunctions in each node neighborhood are selected for the construction of the coarse spaces. The constant C and the resulting coarse-space dimension depend on the partition of unity (initial coarse-grid representation) used.

4.1 Abstract problem eigenvalue problems

We consider an abstract variational problem, where the global bilinear form is obtained by assembling local bilinear forms. That is $a(u,v) = \sum_K a_K(R_K u, R_K v)$, where $a_K(u,v)$ is a bilinear form acting on functions with supports being the coarse block K. Define the subdomain bilinear form $a_{\omega_i}(u,v) = \sum_{K\subset\omega_i} a_K(u,v)$. We consider the abstract problem

$$a(u,v) = F(v) \quad \text{for all } v \in V.$$

We introduce $\{\chi_j\}$, a partition of unity subordinated to coarse-mesh blocks and $\{\xi_i\}$ a partition of unity subordinated to overlapping decomposition (not necessarily related in this subsection). We also define the "Mass" bilinear form (or energy of cut-off) m_{ω_i} and the Rayleigh quotient $\mathcal{Q}_{abs}$ by

$$m_{\omega_i}(v,v) := \sum_{j\in\omega_i} a(\xi_i\chi_j v, \xi_i\chi_j v) \quad \text{and} \quad \mathcal{Q}_{abs}(v) := \frac{a_{\omega_i}(v,v)}{m_{\omega_i}(v,v)}.$$

For the Darcy problem, we have $m_{\omega_i}(v,v) = \sum_{j\in\omega_i}\int_{\omega_i}\kappa|\nabla(\xi_i\chi_j v)|^2 \preceq \int_{\omega_i}\widetilde{\kappa}|v|^2$. The same analysis can be done by replacing the partition of unity functions by partition of degree of freedom (PDoF). Let $\{\boldsymbol{\chi}_j\}$ be PDoF subordianted to coarse mesh neighborhood and $\{\boldsymbol{\xi}_i\}$ be PDoF subordianted to overlapping decomposition. We define the cut-off bilinear form and quotient,

$$m_{\omega_i}(v,v) := \sum_{j\in\omega_i} a(\boldsymbol{\xi}_i\boldsymbol{\chi}_j v, \boldsymbol{\xi}_i\boldsymbol{\chi}_j v) \quad \text{and} \quad \mathcal{Q}_{abs2}(v) := \frac{a_{\omega_i}(v,v)}{m_{\omega_i}(v,v)}.$$

The previous construction alows applying the same design recursively and therefore to use the same ideas in a multilevel method. See [6, 7].

4.2 Generalized Multiscale Finite Element Method (GMsFEM) eigenvalue problem

We can consider the Rayleigh quotients presented before only in a suitable subspace that allows a good approximation of low modes. We call these subspace the snapshot spaces. Denote by W_i the snapshot space corresponding to subdomain ω_i, then we consider the Rayleigh quotient, $\mathcal{Q}_{gm}(v) := \dfrac{a_{\omega_i}(v,v)}{m_{\omega_i}(v,v)}$ with $v \in W_i$. The snapshot space can be obtained by dimension reduction techniques or similar computations. See [5, 2]. For example, we can consider the following simple example. In each subdomain ω_i, $i = 1,\ldots,N_S$:
(1) Generate forcing terms $f_1, f_2, \ldots, f_M$ randomly ($\int_{\omega_i} f_\ell = 0$);
(2) Compute the local solutions $-\mathrm{div}(\kappa\nabla u_\ell) = f_\ell$ with homogeneous Neumann boundary condition;

(3) Generate $W_i = \text{span}\{u_\ell\} \cup \{1\}$;
(4) Consider $\mathcal{Q}_{gm}$ with W_i in 3 and compute important modes.
In Table 1, we see the results of using the local eigenvalue problem versus using the GMsFEM eigenvalue problem.

η	MS	Full	8 rand.	15 rand
10^6	209	35	37	37
10^9	346	38	44	38

Table 1 PCG iterations for different values η. Here $H = 1/10$ with $h = 1/200$. We use the GMsFEM eigevalue problem with $W_i = V_i$ (full local fine-grid space), column 2; W_i spanned by 8 random samples, column 4, and W_i spanned by 15 samples, column 5.

5 Constrained coarse spaces, large overlaps, and DD

In this section, we introduce a hybrid overlapping domain decomposition preconditioner. We use the coarse spaces constructed in [3], which rely on minimal dimensional coarse spaces as discussed above. First, we construct local auxiliary basis functions. For each coarse-block $K \in \mathcal{T}^H$, we solve the eigenvalue problem with Rayleigh quotient $\mathcal{Q}_{ms}(v) := \frac{\int_K \kappa |\nabla v|^2}{\int_K \widehat{\kappa} |v|^2}$, where $\widehat{\kappa} = \kappa \sum_j |\nabla \chi_j|^2$. We assume $\lambda_1^K \leq \lambda_2^K \leq \dots$ and define the local auxiliary spaces,

$$V_{aux}(K) = \text{span}\{\phi_j^K | 1 \leq j \leq L_K\} \text{ and } V_{aux} = \oplus_K V_{aux}(K).$$

Next, define a projection operator π_K as the orthogonal projection on V_{aux} with respect to the inner product $\int_K \widehat{\kappa} uv$ and $\pi_D = \oplus_K \pi_K$. Let K^+ be obtained by adding l layers of coarse elements to the coarse-block K. The coarse-grid multiscale basis $\psi_{j,ms}^K \in V(K^+) = P_0^1(K^+)$ solve

$$\int_{K^+} \kappa \nabla \psi_{j,ms}^K \nabla v + \int_{K^+} \widehat{\kappa} \pi_D(\psi_{j,ms}^K) \pi_D(v) = \int_{K^+} \widehat{\kappa} \phi_j^K \pi_D(v), \ \forall v \in V(K^+).$$

The coarse-grid multiscale space is defined as $V_{ms} = \text{span}\{\psi_{j,ms}^{(i)}\}$. We remark that this space is used as the global coarse solver in our preconditioner. More precisely, we define the (coarse solution) operator $A_{0,ms}^{-1} : H^{-1}(\widehat{\kappa}, D) \mapsto V_{ms}$ by,

$$\int_D \kappa \nabla A_{0,ms}^{-1}(u) \nabla v = u(v) \quad \text{for all } v \in V_{ms}$$

where $H^{-1}(\widehat{\kappa}, D)$ is the space of bounded linear functionals on the weighted sobolev space, $H^1(\kappa, D)$. In our preconditioner, we also need local solution operators which are the operators $A_{i,ms}^{-1} : H^{-1}(\widehat{\kappa}, D) \mapsto V(\omega_i^+)$ defined by,

$$\int_{\omega_i^+} \kappa \nabla A_{i,ms}^{-1}(u_i) \nabla v + \int_{\omega_i^+} \widehat{\kappa} \pi(A_i^{-1}(u_i)) \pi_D(v) = u_i(\chi_i v) \quad \text{for all } v \in V(\omega_i^+),$$

where ω_i^+ is obtained by enlarging ω_i by k coarse-grid layers. Next, we can define the preconditioner[5] M by

$$M^{-1} = (I - A_{0,ms}^{-1}A)\Big(\sum_i A_{i,ms}^{-1}\Big)(I - AA_{0,ms}^{-1}) + A_{0,ms}^{-1}.$$

Note that this is a hybrid preconditioner as defined in [13]. We remark that the constructions of the global coarse space and local solution operators are motivated by [3], where a new multiscale space is developed and analyzed, and it is shown to have a good convergence property independent of the scales of the coefficient of the PDE. In addition, the size of the local problem is dictated by an exponential decay property.

Using some estimates in [3], we can establish the following condition number estimate for $\mathrm{cond}(M^{-1}A)$,

$$\mathrm{cond}(M^{-1}A) \leq \frac{1 + C(1+\Lambda^{-1})^{\frac{1}{2}} E^{\frac{1}{2}} \max\{\tilde{\kappa}^{\frac{1}{2}}\}}{1 - C(1+\Lambda^{-1})^{\frac{1}{2}} E^{\frac{1}{2}} \max\{\tilde{\kappa}^{\frac{1}{2}}\}} \tag{2}$$

where $E = 3(1+\Lambda^{-1})\Big(1+(2(1+\Lambda^{-\frac{1}{2}}))^{-1}\Big)^{1-k}$, C is a constant that depends on the fine and coarse grid only and $\Lambda = \min_K \lambda_{L_K+1}^K$. See [3] for the required estimates of the coarse space. The analysis of the local solvers of the hybrid method above will be presented elsewhere[6]. We see that the condition number is close to 1 if sufficient number of basis functions is selected (i.e., Λ is not close to zero)[7]. The overlap size usually involves several coarse-grid block sizes and thus, the method is effective when the coarse-grid sizes are small. We comment that taking the generous overlap $\delta = kH/2$ in (1), we get the bound $C(1 + 4/k^2)$ with C independent of the contrast. The estimate (2), on the other hand, gives a bound close to 1 if the oversampling is sufficiently large (e.g., the number of coarse-grid layers is related to $\log(\eta)$), which is due to the localization of global fields orthogonal to the coarse space.

Next, we present a numerical result and consider a problem with permeability κ shown in Fig. 2. The fine-grid mesh size h and the coarse-grid mesh size are considered as $h = 1/200$ and $H = 1/20$. In Table 2, we present the number of iterations for using varying numbers of oversampling layers k, values of the contrast η and $\kappa_{M^{-1}A} - 1$, which is the condition number of the preconditioned matrix minus one. We observe that when $k = 3$, the condition number $\kappa_{M^{-1}A}$ is almost one, which confirms (2). In practice, one can choose smaller local problems with a corresponding increase in the number of iterations. This balance can be determined by practical needs.

We would like to emphasize that the proposed method has advantages if the coarse mesh size is not very coarse. In this case, the oversampled coarse regions are still sufficiently small and the coarse-grid solves can be relatively

[5] Here we avoid restriction and extension operators for simplicity of notation.

[6] We mention that the analysis does not use a stable decomposition so, in principle, a new family of robust methods can be obtained.

[7] Having a robust condition number close to 1 is important, especially in applications where the elliptic equation needs to be solved many times.

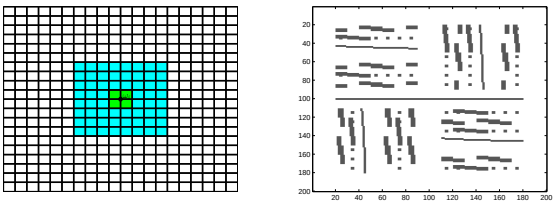


Fig. 2 Left: The coarse mesh used in the numerical experiments. We highlight a coarse neighborhood and the results of adding 3 coarse-block layers to it. Right: The permeability κ used in the experiments. The gray regions indicate high-permeability regions of order η while the white regions indicates a low (order 1) permeability.

# basis per ω	k	# iter	$\kappa_{M^{-1}A} - 1$
3	3	3	5.33e-04
3	4	2	2.57e-05
3	5	2	1.25e-06
3	6	1	5.50e-08

# basis per ω	η	# iter	$\kappa_{M^{-1}A} - 1$
3	1e+3	3	5.68e-04
3	1e+4	3	5.33e-04
3	1e+5	3	6.74e-04

Table 2 Condition number $\kappa_{M^{-1}A}$ and number of iterations until convergence for the PCG with $H = 1/20$, $h = 1/200$ and tol $= 1e-10$. Left: different number of oversampling layers k with $\eta = 1e+4$. Right: different values of the contrast η with $k = 3$.

expensive. Consequently, one wants to minimize the number of coarse-grid solves in addition to local solves. In general, the proposed approach can be used in a multi-level setup, in particular, at the finest levels, while at the coarsest level, we can use original spectral basis functions proposed in [8]. This is object of future research.

6 Conclusions

In this paper, we give an overview of domain decomposition preconditioners for high-contrast multiscale problems. In particular, we review the design of overlapping methods with an emphasis on the stable decomposition for the analysis of the method. We emphasize the use of minimal dimensional coarse spaces in order to construct optimal preconditioners with the condition number independent of physical scales (contrast and spatial scales). We discuss various approaches in this direction. Furthermore, using these spaces and oversampling ideas, we design a new preconditioner with a significant reduction in the number of iterations until convergence if oversampling regions are large enough (several coarse-grid blocks). We note that when using only minimal dimensional coarse spaces in additive Schwarz preconditioner with standard overlap size, we obtain around 19 iterations. in the new method, our main goal is to reduce even further the number of iterations due to large coarse problem sizes. We obtained around 3 iterations until convergence for the new approach. A main point of the new methodology is that after removing the channels we are able to localize the remaining multiscale information

via oversampling. Another interesting aspect of the new approach is that the bound can be obtained by estimating directly operator norms and do not require a stable decomposition.

References

[1] J.E. Aarnes and T. Hou. Multiscale domain decomposition methods for elliptic problems with high aspect ratios. *Acta Math. Appl. Sin. Engl. Ser.*, 18:63–76, 2002.

[2] V. M. Calo, Y. Efendiev, J. Galvis, and G. Li. Randomized oversampling for generalized multiscale finite element methods. *Multiscale Modeling & Simulation*, 14(1):482–501, 2016.

[3] E. T Chung, Y. Efendiev, and W. T. Leung. Constraint energy minimizing generalized multiscale finite element method. *arXiv preprint arXiv:1704.03193*, 2017.

[4] M. Dryja, M. Sarkis, and O. Widlund. Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions. *Numer. Math.*, 72(3):313–348, 1996.

[5] Y. Efendiev, J. Galvis, and T. Hou. Generalized multiscale finite element methods. *Journal of Computational Physics*, 251:116–135, 2013.

[6] Y. Efendiev, J. Galvis, and P.S. Vassilevski. Multiscale Spectral AMGe Solvers for High-Contrast Flow Problems. Preprint available at `http://isc.tamu.edu/resources/preprints/2012/2012-02.pdf`.

[7] Y. Efendiev, J. Galvis, and P.S. Vassilevski. Spectral element agglomerate algebraic multigrid methods for elliptic problems with high-contrast coefficients. In *Domain decomposition methods in science and engineering XIX*, volume 78 of *Lect. Notes Comput. Sci. Eng.*, pages 407–414. Springer, Heidelberg, 2011.

[8] J. Galvis and Y. Efendiev. Domain decomposition preconditioners for multiscale flows in high contrast media. *SIAM J. Multiscale Modeling and Simulation*, 8:1461–1483, 2010.

[9] J. Galvis and Y. Efendiev. Domain decomposition preconditioners for multiscale flows in high contrast media. Reduced dimensional coarse spaces. *SIAM J. Multiscale Modeling and Simulation*, 8:1621–1644, 2010.

[10] I.G. Graham, P. O. Lechner, and R. Scheichl. Domain decomposition for multiscale PDEs. *Numerische Mathematik*, 106(4):589–626, 2007.

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

[12] M. Sarkis. Partition of unity coarse spaces: enhanced versions, discontinuous coefficients and applications to elasticity. In *Domain decomposition methods in science and engineering*, pages 149–158. Natl. Auton. Univ. Mex., México, 2003.

[13] A. Toselli and O. Widlund. *Domain decomposition methods – Algorithms and Theory*, volume 34 of *Computational Mathematics*. Springer-Verlag, 2005.

INTERNODES for heterogeneous couplings

Paola Gervasio and Alfio Quarteroni

Abstract The INTERNODES (INTERpolation for NOnconforming DEcompositionS) method is an interpolation based approach to solve partial differential equations on non-conforming discretizations. In this paper we apply the INTERNODES method to different problems such as the Fluid Structure Interaction problem and the Stokes-Darcy coupled problem that models the filtration of fluids in porous media. Our results highlight the flexibility of the method as well as its optimal rate of convergence.

1 Introduction

The INTERNODES (INTERpolation for NOnconforming DEcompositionS) method is an interpolation based approach to solve partial differential equations on non-conforming discretizations [3, 9]. It is an alternative to projection-based methods like mortar [1], or other interpolation-based method like GFEM/XFEM [10]. Differently than in mortar methods, no cross-mass matrix involving basis functions living on different grids of the interface are required by INTERNODES to build the intergrid operators. Instead, two separate interface mass matrices (separately on either interface) are used. The substantial difference between GFEM/XFEM methods and INTERNODES consists in the fact that the former ones use a partition of unity to enrich the finite element space, while the latter does not add any shape function to those of the local finite element subspaces.

Paola Gervasio
DICATAM, Università degli Studi di Brescia, via Branze 38, 25123 Brescia (Italy), e-mail: paola.gervasio@unibs.it

Alfio Quarteroni
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano (Italy) and École Polytechnique Fédérale de Lausanne (EPFL) (honorary professor), e-mail: alfio.quarteroni@epfl.ch

In this paper we apply the INTERNODES method to different problems such as the Fluid Structure Interaction problem and the Stokes-Darcy coupled problem that models the filtration of fluids in porous media. Our results highlight the flexibility of the method as well as its optimal rate of convergence. Before addressing the two specific problems above, we introduce an abstract formulation for heterogeneous problems. This will also be useful to state the definition of the interface matching operators that will stand at the base of the INTERNODES method.

Let $\Omega \subset \mathbb{R}^d$, with $d = 2,3$, be an open domain with Lipschitz boundary $\partial\Omega$, Ω_1 and Ω_2 be two non-overlapping subdomains with Lipschitz boundary such that $\overline{\Omega} = \overline{\Omega_1 \cup \Omega_2}$, and $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ be their common interface.

Given a function f defined in Ω, we look for u_1 in Ω_1 and u_2 in Ω_2 such that

$$
\begin{aligned}
&L_k(u_k) = f && \text{in } \Omega_k, \quad k = 1,2, && (1)\\
&\Phi_2(u_2) = \Phi_1(u_1) && \text{on } \Gamma \quad \text{(Dirichlet-like condition)}, && (2)\\
&\Psi_1(u_1) + \Psi_2(u_2) = 0 && \text{on } \Gamma \quad \text{(Neumann-like condition)}, && (3)\\
&\text{boundary conditions} && \text{on } \partial\Omega, && (4)
\end{aligned}
$$

where L_1 and L_2 are two differential operators (that may also coincide) while, for $k = 1,2$, Φ_k and Ψ_k are suitable boundary operators restricted to the interface Γ, that depend upon the nature of the differential operators L_1 and L_2. More specifically, Neumann conditions refer here to natural conditions that are enforced *weakly*, whereas Dirichlet conditions identify those *essential* conditions that are enforced directly in the solution subspaces, via the suitable choice of trial functions (see, e.g., [13]). Typically for second order differential operators there is one Dirichlet-like condition and one Neumann-like condition, however more general situations are admissible.

Problem (1)–(4) provides an abstract setting for several kinds of differential problems; here we present two instances of (1)–(4) which the INTERNODES method is applied to.

2 Fluid Structure Interaction problem

When modeling the coupling between fluids and solids, the viscous incompressible Navier-Stokes equations are typically written in ALE (Arbitrarian Lagrangian Eulerian) coordinates in the fluid domain, whereas an elasticity model (either linear or nonlinear, depending on the type of structure) is solved in a reference frame; a third field, the so-called geometry problem, allows to determine the displacement of the fluid domain which defines, in turn, the ALE map, see, e.g., [14, 8, 5].

Let $\widehat{\Omega}_s$ and $\widehat{\Omega}_f$ be two non-overlapping reference configurations for the structure and fluid domains, respectively, and $\widehat{\Gamma} = \partial\widehat{\Omega}_s \cap \partial\widehat{\Omega}_f$ be the fluid-structure reference interface. We assume that the boundaries $\partial\widehat{\Omega}_k$, for $k = s, f$ are Lipschitz continuous and that $(\partial\widehat{\Omega}_k \setminus \widehat{\Gamma})$ is the union of two nonoverlapping subsets $\partial\widehat{\Omega}_k^N$ and $\partial\widehat{\Omega}_k^D$ on which Neumann and Dirichlet boundary conditions will be

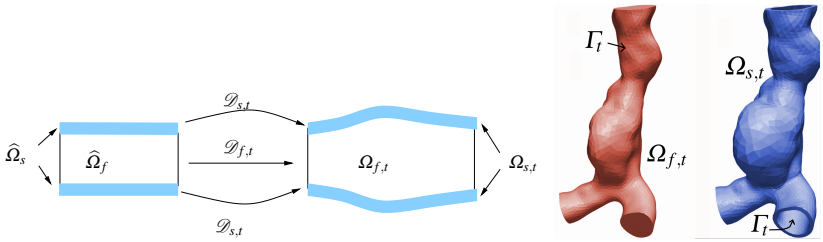


Fig. 1 At left: the ALE frame of reference. At right: the computational domains for the FSI problem: the fluid domain $\Omega_{f,t}$ and the structure domain $\Omega_{s,t}$. $\Gamma_t = \partial\Omega_{f,t} \cap \partial\Omega_{s,t}$

imposed, respectively. Then, for any $t \in (0,T)$ let $\Omega_{s,t}$ and $\Omega_{f,t}$ be the computational structure and fluid domains, respectively, such that $\Omega_{s,0} = \widehat{\Omega}_s$, $\Omega_{f,0} = \widehat{\Omega}_f$ and $\overline{\Omega}_t = \overline{\Omega_{s,t}} \cup \overline{\Omega_{f,t}}$. The current configurations $\Omega_{s,t}$ and $\Omega_{f,t}$ are defined as $\Omega_{k,t} = \{\mathbf{x} = \mathscr{D}_{k,t}(\widehat{\mathbf{x}}) = \widehat{\mathbf{x}} + \widehat{\mathbf{d}}_k(\widehat{\mathbf{x}},t),\ \forall \widehat{\mathbf{x}} \in \widehat{\Omega}_k\}$, with $k = s, f$, where $\widehat{\mathbf{d}}_s$ and $\widehat{\mathbf{d}}_f$ are the displacements induced by the dynamics (see Fig. 1).

We introduce the following entities:

- the outward unit normal vectors $\mathbf{n}_k$ to $\partial\Omega_{k,t}$ (current configuration) and $\widehat{\mathbf{n}}_k$ to $\partial\widehat{\Omega}_k$ (reference configuration),
- the *Arbitrary-Lagrangian-Eulerian (ALE) velocity* $\mathbf{w} = \frac{\partial \widehat{\mathbf{d}}_f}{\partial t}|_{\widehat{\mathbf{x}}}$,
- the deformation gradient tensor for both structure ($k = s$) and fluid ($k = f$) $\mathbf{F}_k = \frac{\partial \mathbf{x}}{\partial \widehat{\mathbf{x}}} = \mathbf{I} + \frac{\partial \widehat{\mathbf{d}}_k}{\partial \widehat{\mathbf{x}}}$ for any $\widehat{\mathbf{x}} \in \widehat{\Omega}_k$,
- the fluid velocity $\mathbf{u}_f$ and the fluid pressure p_f, the dynamic viscosity of the fluid μ, the fluid density ρ_f,
- the Cauchy stress tensor for the fluid $\boldsymbol{\sigma}_f = \boldsymbol{\sigma}_f(\mathbf{u}_f, p_f) = -p_f\mathbf{I} + \mu(\nabla\mathbf{u}_f + (\nabla\mathbf{u}_f)^T)$, and $\widehat{\boldsymbol{\sigma}}_f$ such that $\widehat{\boldsymbol{\sigma}}_f\widehat{\mathbf{n}}_f = det(\mathbf{F}_f)\mathbf{F}_f^{-T}\boldsymbol{\sigma}_f\mathbf{n}_f \circ \mathscr{D}_{f,t}$,
- the Cauchy stress tensor $\boldsymbol{\sigma}_s = \boldsymbol{\sigma}_s(\widehat{\mathbf{d}}_s)$ and the first Piola-Kirchhoff tensor $\widehat{\boldsymbol{\sigma}}_s = \widehat{\boldsymbol{\sigma}}_s(\widehat{\mathbf{d}}_s) = det(\mathbf{F}_s)\boldsymbol{\sigma}_s(\widehat{\mathbf{d}}_s)\mathbf{F}_s^{-T}$ for the structure, the structure density ρ_s.

Then, for any $t \in (0,T)$ the structure and fluid displacements ($\widehat{\mathbf{d}}_s$ and $\widehat{\mathbf{d}}_f$) and the fluid velocity and pressures ($\mathbf{u}_f$ and p_f) are the solution of the FSI system:

structure problem (in reference configuration)

$$\rho_s \frac{\partial^2 \widehat{\mathbf{d}}_s}{\partial t^2} - \nabla \cdot \widehat{\boldsymbol{\sigma}}_s = \mathbf{0} \qquad \text{in } \widehat{\Omega}_s, \qquad (5)$$

fluid problem (in current configuration)

$$\rho_f \frac{\partial \mathbf{u}_f}{\partial t}\Big|_{\widehat{\mathbf{x}}} + \rho_f((\mathbf{u}_f - \mathbf{w}) \cdot \nabla)\mathbf{u}_f - \nabla \cdot \boldsymbol{\sigma}_f = \mathbf{0}, \qquad \text{in } \Omega_{f,t}, \qquad (6)$$

$$\nabla \cdot \mathbf{u}_f = 0 \qquad \text{in } \Omega_{f,t}, \qquad (7)$$

geometry problem (in reference configuration)

$$-\Delta \widehat{\mathbf{d}}_f = \mathbf{0} \qquad \text{in } \widehat{\Omega}_f, \qquad (8)$$

interface conditions (at interface in reference configuration)

$$\widehat{\boldsymbol{\sigma}}_s\widehat{\mathbf{n}}_s+\widehat{\boldsymbol{\sigma}}_f\widehat{\mathbf{n}}_f=\mathbf{0} \qquad \text{(dynamic)} \quad \text{on } \widehat{\Gamma}, \qquad (9)$$

$$\mathbf{u}_f\circ\mathscr{D}_{f,t}=\frac{\partial\widehat{\mathbf{d}}_s}{\partial t} \qquad \text{(kinematic)} \quad \text{on } \widehat{\Gamma}, \qquad (10)$$

$$\widehat{\mathbf{d}}_f=\widehat{\mathbf{d}}_s \qquad \text{(adherence)} \quad \text{on } \widehat{\Gamma}, \qquad (11)$$

completed with: the Dirichlet boundary conditions $\mathbf{u}_f=\mathbf{g}_f^D$ on $\Gamma_{f,t}^D$ and $\widehat{\mathbf{d}}_f=\mathbf{g}_g^D$ on $\widehat{\Gamma}_f^0\subset\partial\widehat{\Omega}_f$, $\widehat{\mathbf{d}}_s=\mathbf{g}_s^D$ on $\widehat{\Gamma}_s^D$, the Neumann conditions $\boldsymbol{\sigma}_f\mathbf{n}_{f,t}=\mathbf{g}_f^N$ on $\Gamma_{f,t}^N$, $\widehat{\boldsymbol{\sigma}}_s\widehat{\mathbf{n}}_s=\mathbf{g}_s^N$ on $\widehat{\Gamma}_s^N$, and the initial conditions $\mathbf{u}_f=\mathbf{u}_0$ in $\Omega_{f,0}$, $\widehat{\mathbf{d}}_s=\widehat{\mathbf{d}}_0$, $\frac{\partial\widehat{\mathbf{d}}_s}{\partial t}=\widehat{\mathbf{d}}_1$ in $\Omega_{s,0}$.

System (5)–(11) can be recast in the form (1)–(4) by associating the structure problem with $L_1(u_1)$ (now representing nonlinear operators, the choices of u_1 and u_2 are obvious), the fluid problem and the geometric problem with $L_2(u_2)$, both the adherence and the kinematic interface conditions are interpreted as Φ-like conditions (they involve the traces of the unknowns functions on $\widehat{\Gamma}$), whereas the dynamic interface condition is interpreted as a Ψ-like condition (as it involves normal stresses on $\widehat{\Gamma}$).

3 Fluids filtration in porous media (Stokes-Darcy coupling)

Flow processes in a free-fluid region adjacent to a porous medium occur in many relevant applications. Under the (realistic) assumption that the Reynolds number in the porous domain is small, the Navier-Stokes equations could be therein up-scaled to a macroscopic level and replaced by the Darcy law.

Consider the case of a tangential flow of a fluid over a porous bed. This situation is known in literature also as *near parallel flows* [12], i.e. flows for which the pressure gradient is not normal to the interface and the Darcy velocity inside the porous domain is much smaller than the velocity in the fluid domain. The most widely used approach to couple the free fluid regime with the porous-medium one consists of:
- the introduction of an artificial sharp interface Γ between the Stokes (or fluid) domain Ω_s and the Darcy (or porous) domain Ω_d;
- the imposition of the mass conservation, the balance of normal forces and the Beavers-Joseph-Saffman (BJS) experimental law on Γ ([7]), see Fig. 2.

To write down the associated mathematical model, we introduce the following entities:

- the outward unit normal vectors $\mathbf{n}_k$ to $\partial\Omega_k$,
- the dynamic viscosity μ, the density ρ, the velocity $\mathbf{u}_s$ and the pressure p_s of the fluid in Ω_s,
- the Cauchy stress tensor for the fluid $\boldsymbol{\sigma}_s=\boldsymbol{\sigma}_s(\mathbf{u}_s,p_s)=-p_s\mathbf{I}+\mu(\nabla\mathbf{u}_s+(\nabla\mathbf{u}_s)^T)$,
- the Darcy velocity $\mathbf{u}_d$ and the intrinsic average pressure p_d in the porous domain, the intrinsic permeability $\boldsymbol{\kappa}=\boldsymbol{\kappa}(\mathbf{x})$ (for any $\mathbf{x}\in\Omega_d$) of the porous media,
- two given body forces $\mathbf{f}_s$ and $\mathbf{f}_d$,
- the normal unit vector $\mathbf{n}_\Gamma$ to Γ directed from Ω_s to Ω_d (then $\mathbf{n}_\Gamma=\mathbf{n}_s=-\mathbf{n}_d$ on Γ) and an orthonormal system of tangent vectors $\boldsymbol{\tau}_j$, with $j=1,\ldots,d-1$ on Γ.

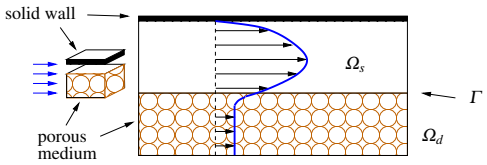


Fig. 2 A typical setting of the Stokes-Darcy coupled problem for a fluid over a porous bed

The coupled problem that we consider reads:

Stokes problem (fluid domain)

$$-\nabla\cdot\boldsymbol{\sigma}_s=\mathbf{f}_s,\quad \nabla\cdot\mathbf{u}_s=0 \qquad \text{in } \Omega_s, \quad (12)$$

Darcy problem (porous domain)

$$\mathbf{u}_d=-\frac{\boldsymbol{\kappa}}{\mu}(\nabla p_d-\mathbf{f}_d),\quad \nabla\cdot\mathbf{u}_d=0 \qquad \text{in } \Omega_d, \quad (13)$$

interface conditions (sharp interface)

$$\mathbf{u}_s\cdot\mathbf{n}_s+\mathbf{u}_d\cdot\mathbf{n}_d=0 \qquad \text{(mass conservation)} \qquad \text{on } \Gamma, \quad (14)$$

$$(\boldsymbol{\sigma}_s\mathbf{n}_s)\cdot\mathbf{n}_s+p_d=0 \qquad \text{(balance of normal forces)} \qquad \text{on } \Gamma, \quad (15)$$

$$(\boldsymbol{\sigma}_s\mathbf{n}_s)\cdot\boldsymbol{\tau}_j+\frac{\alpha\mu}{\sqrt{\boldsymbol{\tau}_j{}^T\boldsymbol{\kappa}\boldsymbol{\tau}_j}}\mathbf{u}_s\cdot\boldsymbol{\tau}_j=0,\ j=1,\ldots,d-1, \text{(BJS condition)} \qquad \text{on } \Gamma, \quad (16)$$

where α is a suitable parameter depending on the porous media. Indeed, the BJS condition is not a coupling condition, as it only involves quantities from one side.

The system (12)–(16) is completed with suitable boundary conditions that read (as usual, D stands for Dirichlet and N for Neumann): $\mathbf{u}_s=\mathbf{g}_s^D$ on $\partial\Omega_s^D$, $\sigma_s\mathbf{n}_s=\mathbf{0}$ on $\partial\Omega_s^N$, $p_d=0$ on $\partial\Omega_d^D$, $\mathbf{u}_d\cdot\mathbf{n}_d=g_d^N$ on $\partial\Omega_d^N$, where we assume that $\partial\Omega_k^N$ and $\partial\Omega_k^D$ are non-intersecting subsets of $\partial\Omega_k\setminus\Gamma$ such that $\overline{\partial\Omega_k^N\cup\partial\Omega_k^D}=\overline{\partial\Omega_k\setminus\Gamma}$.

The coupled system (12)–(16) can be recast in the form (1)–(4) by associating the Stokes problem with $L_2(u_2)$ and the Darcy problem with $L_1(u_1)$. When considering the weak (variational) formulation of the coupled problem (12)–(16), the interface coupling conditions (14) and (15) can be treated in different ways depending on the specific variational form used. In the form used in Sect. 6, the balance of normal forces (15) plays the role of a Φ-like condition (2), while the mass conservation condition (14) will be treated as a Ψ-like condition (3). In specific circumstances, however, for instance when the interface Γ is parallel to one of the cartesian coordinates, condition (14) can be easily enforced as a Dirichlet condition (thus under the form (2)) on the space of trial functions and condition (15) as a Neumann (natural) condition, e.g., like (3).

4 Intergrid operators for non-conforming discretization

We consider two a-priori *independent families of triangulations* $\mathscr{T}_{1,h_1}$ in Ω_1 and $\mathscr{T}_{2,h_2}$ in Ω_2, respectively. The meshes in Ω_1 and in Ω_2 can be non-conforming on Γ and characterized by different mesh-sizes h_1 and h_2. Moreover, different poly-

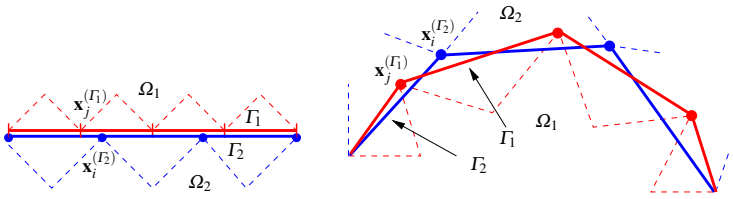


Fig. 3 Γ_1 and Γ_2 induced by the triangulations $\mathscr{T}_{1,h_1}$ and $\mathscr{T}_{2,h_2}$, when $d=2$

nomial degrees $\mathtt{p}_1$ and $\mathtt{p}_2$ can be used to define the finite element spaces. Inside each subdomain Ω_k we assume that the triangulations $\mathscr{T}_{k,h_k}$ are affine, regular and quasi-uniform ([15, Ch.3]).

Then, for $k=1,2$, let $X_{k,h_k}=\{v\in C^0(\overline{\Omega_k}):\ v_{|T}\in\mathbb{P}_{\mathtt{p}_k},\forall T\in\mathscr{T}_{k,h_k}\}$ be the usual Lagrangian finite element spaces associated with $\mathscr{T}_{k,h_h}$, while $Y_{k,h_k}=\{\lambda=v|_\Gamma,\ v\in X_{k,h_k}\}$ are the spaces of traces on Γ of functions in X_{k,h_k}, whose dimension is n_k.

We denote by Γ_1 and Γ_2 the internal boundaries of Ω_1 and Ω_2, respectively, induced by the triangulations $\mathscr{T}_{1,h_1}$ and $\mathscr{T}_{2,h_2}$. If Γ is a straight segment, then $\Gamma_1=\Gamma_2=\Gamma$, otherwise Γ_1 and Γ_2 may not coincide (see Fig. 3).

For $k=1,2$, let $\{\mathbf{x}_1^{(\Gamma_k)},\ldots,\mathbf{x}_{n_k}^{(\Gamma_k)}\}\in\overline{\Gamma}_k$ be the nodes induced by the mesh $\mathscr{T}_{k,h_k}$.

We introduce two independent operators that exchange information between the two independent grids on the interface Γ: $\Pi_{12}:Y_{2,h_2}\to Y_{1,h_1}$ and $\Pi_{21}:Y_{1,h_1}\to Y_{2,h_2}$.

If $\Gamma_1=\Gamma_2$, Π_{12} and Π_{21} are the classical Lagrange interpolation operators defined by the relations:

$$(\Pi_{12}\mu_{2,h_2})(\mathbf{x}_i^{(\Gamma_1)})=\mu_{2,h_2}(\mathbf{x}_i^{(\Gamma_1)}),\quad i=1,\ldots,n_1,\qquad\forall\mu_{2,h_2}\in Y_{2,h_2},\tag{17}$$

$$(\Pi_{21}\mu_{1,h_1})(\mathbf{x}_i^{(\Gamma_2)})=\mu_{1,h_1}(\mathbf{x}_i^{(\Gamma_2)}),\quad i=1,\ldots,n_2,\qquad\forall\mu_{1,h_1}\in Y_{1,h_1}.\tag{18}$$

If, instead, Γ_1 and Γ_2 are geometrically non-conforming, we define Π_{12} and Π_{21} as the Rescaled Localized Radial Basis Function (RL-RBF) interpolation operators introduced in formula (3.1) of [4]. In both cases, the (rectangular) matrices associated with Π_{12} and Π_{21} are, respectively, $R_{12}\in\mathbb{R}^{n_1\times n_2}$ and $R_{21}\in\mathbb{R}^{n_2\times n_1}$ and they are defined by:

$$\begin{aligned}(R_{12})_{ij}&=(\Pi_{12}\mu_j^{(2)})(\mathbf{x}_i^{(\Gamma_1)})\quad i=1,\ldots,n_1,\ j=1,\ldots,n_2,\\(R_{21})_{ij}&=(\Pi_{21}\mu_j^{(1)})(\mathbf{x}_i^{(\Gamma_2)})\quad i=1,\ldots,n_2,\ j=1,\ldots,n_1,\end{aligned}\tag{19}$$

where $\{\mu_i^{(k)}\}$ are the Lagrange basis functions of Y_{k,h_k}, for $k=1,2$ and $i=1,\ldots,n_k$.

In the special conforming case for which $\Gamma_1=\Gamma_2$, $h_1=h_2$ and $\mathtt{p}_1=\mathtt{p}_2$, the interpolation operators Π_{12} and Π_{21} are the identity operator and $R_{12}=R_{21}=I$ (the identity matrix of size $n_1=n_2$). Finally, let M_{Γ_k} such that

$$(M_{\Gamma_k})_{ij}=(\mu_j^{(k)},\mu_i^{(k)})_{L^2(\Gamma_k)},\qquad k=1,2,\tag{20}$$

be the *interface mass matrices*. To assemble both the interface mass matrices and the interpolation matrices, for both the Lagrange and the RL-RBF approaches, the only information that are needed are the coordinates of the interface nodes.

Let $\ell,k=1,2$. If $\boldsymbol{\mu}^{(k)} \in [Y_{k,h_k}]^d$ with $d=2,3$; by writing $\Pi_{\ell k}\boldsymbol{\mu}^{(k)}$ we mean that the interpolation operator $\Pi_{\ell k}$ is applied to each component of the vector-value function $\boldsymbol{\mu}^{(k)}$. Finally, $\mathbf{M}_{\Gamma_k} = diag(M_{\Gamma_k},\ldots,M_{\Gamma_k})$ and $\mathbf{R}_{\ell k} = diag(R_{\ell k},\ldots,R_{\ell k})$ are block diagonal matrices with d blocks.

5 INTERNODES applied to the FSI system

We define the functional spaces:

$$\begin{array}{l}
\mathbf{V}_{f,t} = [H^1(\Omega_{f,t})]^d, \quad Q_{f,t} = L^2(\Omega_{f,t}), \quad \mathbf{V}^D_{f,t} = \{\mathbf{v}\in\mathbf{V}_{f,t}:\ \mathbf{v}=\mathbf{0} \text{ on } \partial\Omega^D_{f,t}\},\\
\mathbf{V}^0_{f,t} = \{\mathbf{v}\in\mathbf{V}_{f,t}:\ \mathbf{v}=\mathbf{0} \text{ on } \partial\Omega^D_{f,t}\cup\Gamma_t\}, \quad \mathbf{V}_s = [H^1(\widehat{\Omega}_s)]^d,\\
\mathbf{V}^D_s = \{\mathbf{v}\in\mathbf{V}_s:\ \mathbf{v}=\mathbf{0} \text{ on } \partial\widehat{\Omega}^D_s\}, \quad \mathbf{V}^0_s = \{\mathbf{v}\in\mathbf{V}_s:\ \mathbf{v}=\mathbf{0} \text{ on } \partial\widehat{\Omega}^D_s\cup\widehat{\Gamma}\},\\
\mathbf{V}_g = [H^1(\widehat{\Omega}_f)]^d, \quad \mathbf{V}^D_g = \{\mathbf{v}\in\mathbf{V}_g:\ \mathbf{v}=\mathbf{0} \text{ on } \partial\widehat{\Omega}^0_f\}, \quad \widehat{\boldsymbol{\Lambda}} = [H^{1/2}_{00}(\widehat{\Gamma})]^d,
\end{array} \tag{21}$$

and the lifting operators $\mathscr{R}_s : \widehat{\boldsymbol{\Lambda}} \to \widehat{\mathbf{V}}^D_s$ s.t. $(\mathscr{R}_s\widehat{\boldsymbol{\lambda}})|_{\widehat{\Gamma}} = \widehat{\boldsymbol{\lambda}}$, $\mathscr{R}_{f,t} : \widehat{\boldsymbol{\Lambda}} \to \mathbf{V}^D_{f,t}$ s.t. $(\mathscr{R}_{f,t}\widehat{\boldsymbol{\lambda}})|_{\Gamma_t} = \widehat{\boldsymbol{\lambda}}\circ\mathscr{D}^{-1}_{f,t}$.

Let us discretize the time derivatives by standard finite difference schemes (e.g. a backward differentiation formula to approximate the first order derivative and the Newmark method to approximate the second one). The weak semi-discrete (continuous in space) counterpart of the FSI system (5)–(11) reads: for any time level t^n, with $n\geq 1$, find $\mathbf{u}^n_f \in \mathbf{V}_{f,t^n}$, $p^n_f \in Q_{f,t^n}$, $\widehat{\mathbf{d}}^n_f \in \mathbf{V}_g$ and $\widehat{\mathbf{d}}^n_s \in \mathbf{V}_s$ satisfying the Dirichlet boundary conditions $\mathbf{u}^n_f = \mathbf{g}^D_f(t^n)$ on Γ^D_{f,t^n} and $\widehat{\mathbf{d}}^n_f = \mathbf{g}^D_g(t^n)$ on $\widehat{\Gamma}^0_f \subset \partial\widehat{\Omega}_f$, $\widehat{\mathbf{d}}^n_s = \mathbf{g}^D_s(t^n)$ on $\widehat{\Gamma}^D_s$ and the initial conditions $\mathbf{u}^0_f = \mathbf{u}_0$ in $\Omega_{f,0}$, $\widehat{\mathbf{d}}^0_s = \widehat{\mathbf{d}}_0$, and $\frac{\partial\widehat{\mathbf{d}}_s}{\partial t}|_{t=0} = \widehat{\mathbf{d}}_1$ in $\Omega_{s,0}$, such that:

$$\mathscr{A}_s(\widehat{\mathbf{d}}^n_s,\widehat{\mathbf{v}}_s) = \mathscr{F}^n_s(\widehat{\mathbf{v}}_s) \qquad \forall\widehat{\mathbf{v}}_s\in\mathbf{V}^0_s, \tag{22}$$

$$\mathscr{A}_f(\mathbf{u}^n_f,\widehat{\mathbf{d}}^n_f;\mathbf{v}_f) + \mathscr{B}_f(\mathbf{v}_f,p^n_f) = \mathscr{F}^n_f(\mathbf{v}_f) \qquad \forall\mathbf{v}_f\in\mathbf{V}^0_{f,t^n}, \tag{23}$$

$$\mathscr{B}_f(\mathbf{u}^n_f,q) = 0 \qquad \forall q\in Q_{f,t^n}, \tag{24}$$

$$\mathscr{G}(\widehat{\mathbf{d}}^n_f,\widehat{\mathbf{v}}_g) = 0 \qquad \forall\widehat{\mathbf{v}}_g\in\mathbf{V}^D_g, \tag{25}$$

$$\begin{array}{ll}
\mathscr{A}_s(\widehat{\mathbf{d}}^n_s,\mathscr{R}_s\widehat{\boldsymbol{\mu}}) + \mathscr{A}_f(\mathbf{u}^n_f,\widehat{\mathbf{d}}^n_f;\mathscr{R}_f\widehat{\boldsymbol{\mu}}) + \mathscr{B}_f(\mathscr{R}_f\widehat{\boldsymbol{\mu}},p^n_f) & \\
\qquad = \mathscr{F}^n_s(\mathscr{R}_s\widehat{\boldsymbol{\mu}}) + \mathscr{F}^n_f(\mathscr{R}_f\widehat{\boldsymbol{\mu}}) & \forall\widehat{\boldsymbol{\mu}}\in\widehat{\boldsymbol{\Lambda}},
\end{array} \tag{26}$$

$$\mathbf{u}^n_f\circ\mathscr{D}_{f,t^n} = a_3\widehat{\mathbf{d}}^n_s + \widehat{\mathbf{b}}^{n-1}_3, \qquad \widehat{\mathbf{d}}^n_f = \widehat{\mathbf{d}}^n_s \qquad \text{on } \widehat{\Gamma}, \tag{27}$$

where

$$\mathscr{A}_s(\widehat{\mathbf{d}}_s,\widehat{\mathbf{v}}_s)=\int_{\widehat{\Omega}_s}(\rho_s a_1\widehat{\mathbf{d}}_s\cdot\widehat{\mathbf{v}}_s+\widehat{\boldsymbol{\sigma}}_s:\nabla_{\hat{\mathbf{x}}}\widehat{\mathbf{v}}_s)\,d\widehat{\Omega},$$
$$\mathscr{F}_s^n(\widehat{\mathbf{v}}_s)=\int_{\partial\widehat{\Omega}_s^N}\mathbf{g}_{s,N}^n\cdot\widehat{\mathbf{v}}_s d\widehat{\gamma}+\int_{\widehat{\Omega}_s}\mathbf{b}_1^{n-1}\,d\widehat{\Omega},\quad \mathscr{G}(\widehat{\mathbf{d}}_f,\widehat{\mathbf{v}}_g)=\int_{\widehat{\Omega}_f}\nabla_{\hat{\mathbf{x}}}\widehat{\mathbf{d}}_f:\nabla_{\hat{\mathbf{x}}}\widehat{\mathbf{v}}_g\,d\Omega,$$
$$\mathscr{A}_f(\mathbf{u}_f,\widehat{\mathbf{d}}_f;\mathbf{v}_f)=\int_{\Omega_{f,t}}\rho_f(a_2\mathbf{u}_f+((\mathbf{u}_f-\mathbf{w})\cdot\nabla)\mathbf{u}_f)\cdot\mathbf{v}_f\,d\Omega$$
$$+\int_{\Omega_{f,t}}\mu(\nabla\mathbf{u}_f+(\nabla\mathbf{u}_f)^T):\nabla\mathbf{v}_f)\,d\Omega,$$
$$\mathscr{B}_f(\mathbf{u}_f,q_f)=\int_{\Omega_{f,t}}(\nabla\cdot\mathbf{u}_f)q\,d\Omega,\quad \mathscr{F}_f^n(\mathbf{v}_f)=\int_{\partial\Omega_{f,t}^N}\mathbf{g}_{f,N}^n\cdot\mathbf{v}_f d\gamma+\int_{\Omega_{f,t}}\mathbf{b}_2^{n-1}\cdot\mathbf{v}_f\,d\Omega,$$

with a_1,a_2,a_3 suitable real values and $\mathbf{b}_1^{n-1},\mathbf{b}_2^{n-1}$, and $\mathbf{b}_3^{n-1}$ (depending on the solution at the previous time levels) suitable vector functions arising from the finite difference discretization of the time derivatives.

Equation (26) is the weak counterpart of the dynamic interface condition (9).

We consider now independent finite element space discretizations (as described in Sect. 4) in $\widehat{\Omega}_f$ and $\widehat{\Omega}_s$ (a suitable inf-sup stable couple of finite elements will be considered in the fluid domain) that may induce two different discrete interfaces $\widehat{\Gamma}_f=\mathscr{T}_{f,h_f}\cap\widehat{\Gamma}$ and $\widehat{\Gamma}_s=\mathscr{T}_{s,h_s}\cap\widehat{\Gamma}$ in the case that $\widehat{\Gamma}$ is curved as in Fig. 3, right. Then we use the subindices h_k, for $k=s,f$, to characterize the subspaces of the functional spaces (21) as well as the discrete counterpart of each variable appearing in the system (22)–(27). From now on, in $\widehat{\mathbf{d}}_{s,h_s}^n$, $\mathbf{u}_{f,h_f}^n$, $\widehat{\mathbf{d}}_{f,h_f}^n$, and p_{f,h_f}^n, the super-index n will be omitted for sake of notations.

In order to apply the INTERNODES method to the discrete counterpart of (22)–(27), we define the scalar quantities:

$$\begin{aligned}
r_{s,i}=&\ \mathscr{A}_s(\widehat{\mathbf{d}}_{s,h_s},\mathscr{R}_s\widehat{\boldsymbol{\mu}}_i^{(s)})-\mathscr{F}_s^n(\mathscr{R}_s\widehat{\boldsymbol{\mu}}_i^{(s)}), && i=1,\dots,d\cdot n_s,\\
r_{f,i}=&\ \mathscr{A}_f(\mathbf{u}_{f,h_f},\widehat{\mathbf{d}}_{f,h_f};\mathscr{R}_f\widehat{\boldsymbol{\mu}}_i^{(f)})+\mathscr{B}_f(\mathscr{R}_f\widehat{\boldsymbol{\mu}}_i^{(f)},p_{f,h_f})\\
&-\mathscr{F}_f^n(\mathscr{R}_f\widehat{\boldsymbol{\mu}}_i^{(f)}), && i=1,\dots,d\cdot n_f
\end{aligned}\tag{28}$$

(where $\{\widehat{\boldsymbol{\mu}}_i^{(k)}\}_{i=1}^{d\cdot n_k}$ are the Lagrange basis functions of $[Y_{k,h_k}]^d$) and

$$z_{k,j}=\sum_{i=1}^{d\cdot n_k}(\mathbf{M}_{\widehat{\Gamma}_k}^{-1})_{ji}r_{k,i},\qquad k=s,f,\ j=1,\dots,d\cdot n_k,\tag{29}$$

and the functions $\mathbf{r}_{k,h_k}=\sum_{j=1}^{d\cdot n_k}z_{k,j}\widehat{\boldsymbol{\mu}}_j^{(k)}$,which are the so called *discrete residuals* and are the discrete counterpart of $\widehat{\boldsymbol{\sigma}}_k\widehat{\mathbf{n}}_k$.

The INTERNODES method applied to system (9)–(11) at any t^n reads:

$$\mathscr{A}_s(\widehat{\mathbf{d}}_{s,h_s},\widehat{\mathbf{v}}_{s,h_s})=\mathscr{F}_s^n(\widehat{\mathbf{v}}_{s,h_s})\qquad\forall\widehat{\mathbf{v}}_{s,h_s}\in\mathbf{V}_{s,h_s}^0,\tag{30}$$
$$\mathscr{A}_f(\mathbf{u}_{f,h_f},\widehat{\mathbf{d}}_{f,h_f};\mathbf{v}_{f,h_f})+\mathscr{B}_f(\mathbf{v}_{f,h_f},p_{f,h_f})=\mathscr{F}_f^n(\mathbf{v}_{f,h_f})\qquad\forall\mathbf{v}_{f,h_f}\in\mathbf{V}_{f,h_f,t^n}^0,\tag{31}$$
$$\mathscr{B}_f(\mathbf{u}_{f,h_f},q_{f,h_f})=0\qquad\forall q_{f,h_f}\in Q_{f,h_f,t^n},\tag{32}$$
$$\mathscr{G}(\widehat{\mathbf{d}}_{f,h_f},\widehat{\mathbf{v}}_{g,h_g})=0\qquad\forall\widehat{\mathbf{v}}_{g,h_g}\in\mathbf{V}_{g,h_g}^D,\tag{33}$$

$$\mathbf{r}_{s,h_s}+\Pi_{sf}\mathbf{r}_{f,h_f}=\mathbf{0}\qquad\text{(dynamic)}\qquad\text{on }\widehat{\Gamma}_s,\tag{34}$$
$$\mathbf{u}_{f,h_f}\circ\mathscr{D}_{f,t^n}=\Pi_{fs}(a_3\widehat{\mathbf{d}}_{s,h_s}+\widehat{\mathbf{b}}_3^{n-1})\qquad\text{(kynematic)}\qquad\text{on }\widehat{\Gamma}_f,\tag{35}$$

$$\widehat{\mathbf{d}}_{f,h_f} = \Pi_{fs}\widehat{\mathbf{d}}_{s,h_s} \qquad \text{(adherence)} \quad \text{on } \widehat{\Gamma}_f. \tag{36}$$

The conditions (34)–(36) are the INTERNODES counterpart of the interface condition (9)–(11), obtained by applying the intergrid operators Π_{12} and Π_{21} defined in Sect. 4. More precisely, if we make the associations $s \leftrightarrow 1$ and $f \leftrightarrow 2$, the operator $\Pi_{fs}(=\Pi_{21})$ is used to interpolate on $\widehat{\Gamma}_f$ each component of the discrete traces $\widehat{\mathbf{d}}_{s,h_s}$ and (the discretization of) $\frac{\partial \widehat{\mathbf{d}}_{s,h_s}}{\partial t}|_{t^n}$ that are known on $\widehat{\Gamma}_s$, while $\Pi_{sf}(=\Pi_{12})$ is used to interpolate on $\widehat{\Gamma}_s$ each component of the discrete counterpart of the normal stress $\widehat{\boldsymbol{\sigma}}_f\widehat{\mathbf{n}}_f$ that is known on $\widehat{\Gamma}_f$.

By construction, $\mathbf{r}_{k,h_k} \in \mathbf{Y}_k = [Y_{k,h_k}]^d$, for $k = s, f$, and then $\mathbf{r}_{f,h_f}$ has the sufficient regularity to be interpolated.

Remark 1. The scalar values (28), typically computed as algebraic residuals at the interface of the finite element system, *are not* the coefficients of the function $\mathbf{r}_{k,h_k}$ w.r.t. the Lagrange expansion $\{\widehat{\boldsymbol{\mu}}_j^{(k)}\}$, rather the coefficients of $\mathbf{r}_{k,h_k}$ w.r.t. the canonical basis $\{\widehat{\boldsymbol{\psi}}_i^{(k)}\}_{i=1}^{d\cdot n_k}$ of $\mathbf{Y}'_{k,h_k}$. The latter is the dual to $\{\widehat{\boldsymbol{\mu}}_j^{(k)}\}$, that is it satisfies the relations $(\widehat{\boldsymbol{\psi}}_i^{(k)}, \widehat{\boldsymbol{\mu}}_j^{(k)})_{L^2(\widehat{\Gamma}_k)} = \delta_{ij}$, for $i, j = 1, d\cdot\ldots, n_k$, with δ_{ij} the Kronecker delta. It can proved (see [2]) that $\widehat{\boldsymbol{\psi}}_i^{(k)} = \sum_{j=1}^{d\cdot n_k}(\mathbf{M}_{\Gamma_k}^{-1})_{ji}\widehat{\boldsymbol{\mu}}_j^{(k)}$, i.e., the interface mass matrix $\mathbf{M}_{\Gamma_k}$ and its inverse play the role of transfer matrices from the Lagrange basis to the dual one and viceversa, respectively.

Denoting by r_f, r_s, u_f, d_s, d_f, b_3^{n-1}, and d_f the arrays whose entries are the Lagrangian degrees of freedom of $\mathbf{r}_{f,h_f}$, $\mathbf{r}_{s,h_s}$, $\mathbf{u}_{f,h_f}$, $\widehat{\mathbf{d}}_{s,h_s}$, $\widehat{\mathbf{d}}_{f,h_f}$, and $\mathbf{b}_3^{n-1}$, respectively, the algebraic form of the INTERNODES conditions (34)–(36) reads:

$$\mathbf{M}_{\Gamma_s}^{-1}\mathsf{r}_s + \mathbf{R}_{sf}\mathbf{M}_{\Gamma_f}^{-1}\mathsf{r}_f = 0, \tag{37}$$
$$\mathsf{u}_f = \mathbf{R}_{fs}(a_3\mathsf{d}_s + \mathsf{b}_3^{n-1}), \tag{38}$$
$$\mathsf{d}_f = \mathbf{R}_{fs}\mathsf{d}_s. \tag{39}$$

Notice that (37) can be equivalently written as $\mathsf{r}_s + \mathbf{M}_{\Gamma_s}\mathbf{R}_{sf}\mathbf{M}_{\Gamma_f}^{-1}\mathsf{r}_f = 0$.

The INTERNODES method has been successfully applied to the FSI system in [8, 5].

6 INTERNODES applied to the Stokes-Darcy system

We define the functional spaces:

$$\mathbf{V}_s = [H^1(\Omega_s)]^d, \quad \mathbf{V}_s^D = \{\mathbf{v} \in \mathbf{V}_s :\ \mathbf{v} = \mathbf{0} \text{ on } \partial\Omega_s^D\}, \tag{40}$$
$$\mathbf{V}_d = \{\mathbf{v} \in [L^2(\Omega_d)]^d :\ \nabla\cdot\mathbf{v} \in L^2(\Omega_d)\}, \quad \mathbf{V}_d^N = \{\mathbf{v} \in \mathbf{V}_d :\ \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega_d^N\},$$
$$Q_s = L^2(\Omega_s), \quad Q_d = L^2(\Omega_d), \quad \Lambda = H_{00}^{1/2}(\Gamma).$$

Then we consider the following weak form of the Stokes-Darcy coupled problem (12)–(16) ([11]): find $\mathbf{u}_s \in \mathbf{V}_s$, $p_s \in Q_s$, $\mathbf{u}_d \in \mathbf{V}_d$, $p_d \in Q_d$, and $\lambda \in \Lambda$ with $\mathbf{u}_s = \mathbf{g}_s^D$ on $\partial\Omega_s^D$, $\mathbf{u}_d \cdot \mathbf{n}_d = g_d^N$ on $\partial\Omega_d^N$ such that:

$$
\begin{aligned}
&2\mu \int_{\Omega_s} D(\mathbf{u}_s) : D(\mathbf{v}_s)\, d\Omega - \int_{\Omega_s} p_s \nabla \cdot \mathbf{v}_s\, d\Omega + \int_{\Gamma} \lambda \mathbf{v}_s \cdot \mathbf{n}_s\, d\Gamma \qquad (41)\\
&\quad + \sum_{j=1}^{d-1} \int_{\Gamma} \alpha_j (\mathbf{u}_s \cdot \boldsymbol{\tau}_j)(\mathbf{v}_s \cdot \boldsymbol{\tau}_j)\, d\Gamma = \int_{\Omega_s} \mathbf{f}_s \cdot \mathbf{v}_s\, d\Omega \qquad \forall \mathbf{v}_s \in \mathbf{V}_s^D,\\
&\int_{\Omega_s} q_s \nabla \cdot \mathbf{u}_s\, d\Omega = 0 \qquad \forall q_s \in Q_s,\\
&\mu \int_{\Omega_d} (\boldsymbol{\kappa}^{-1} \mathbf{u}_d) \cdot \mathbf{v}_d\, d\Omega - \int_{\Omega_d} p_d \nabla \cdot \mathbf{v}_d\, d\Omega + \int_{\Gamma} \lambda \mathbf{v}_d \cdot \mathbf{n}_d\, d\Gamma \qquad (42)\\
&\quad = \int_{\Omega_d} \mathbf{f}_d \cdot \mathbf{v}_d\, d\Omega \qquad \forall \mathbf{v}_d \in \mathbf{V}_d^N,\\
&\int_{\Omega_d} q_d \nabla \cdot \mathbf{u}_d\, d\Omega = 0 \qquad \forall q_d \in Q_d,\\
&\int_{\Gamma} \mathbf{u}_s \cdot \mathbf{n}_s \eta + \int_{\Gamma} \mathbf{u}_d \cdot \mathbf{n}_d \eta = 0 \qquad \forall \eta \in \Lambda, \qquad (43)
\end{aligned}
$$

where $D(\mathbf{v}) = (\nabla \mathbf{v} + (\nabla \mathbf{v})^T)/2$, while $\alpha_j = \alpha\mu / \sqrt{\boldsymbol{\tau}_j^T \boldsymbol{\kappa} \boldsymbol{\tau}_j}$.

The Lagrange multiplier $\lambda \in \Lambda$ is in fact $\lambda = p_d = -(\boldsymbol{\sigma}_s \mathbf{n}_s) \cdot \mathbf{n}_s$ on Γ.

We discretize both Stokes problem (12) and Darcy problem (13) by inf-sup stable (or stabilized) couples of finite elements (see, e.g., [6]). Independent finite element space discretizations (as described in Sect. 4) are considered in Ω_s and Ω_d that may induce two different discrete interface $\Gamma_s = \mathscr{T}_{s,h_s} \cap \Gamma$ and $\Gamma_d = \mathscr{T}_{d,h_d} \cap \Gamma$ in the case that Γ is curved as in Fig. 3, right. Then we use the subindices h_k, for $k = s, d$, to characterize the subspaces of the functional spaces (40) as well as the discrete counterpart of each variable appearing in the system (41)–(43). For $k = s, d$, $\Lambda_{k,h_k} = \Lambda \cap Y_{k,h_k}$.

In order to apply the INTERNODES method to the discrete counterpart of (41)–(43), we define the scalar quantities:

$$
r_{k,i} = \int_{\Gamma} (\mathbf{u}_{k,h_k} \cdot \mathbf{n}_k) \mu_i^{(k)}, \qquad i = 1, \ldots, n_k,\ k = s, d, \qquad (44)
$$

(where $\{\mu_i^{(k)}\}_{i=1}^{n_k}$ are the Lagrange basis functions of Y_{k,h_k}) and

$$
z_{k,j} = \sum_{i=1}^{n_k} (M_{\Gamma_k}^{-1})_{ji} r_{k,i}, \qquad j = 1, \ldots, n_k,\ k = s, d, \qquad (45)
$$

and the discrete functions (belonging to Y_{k,h_k})

$$
w_{k,h_k} = \sum_{j=1}^{n_k} z_{k,j} \mu_j^{(k)}. \qquad (46)
$$

The INTERNODES form of problem (41)–(43) reads: find $\mathbf{u}_{s,h_s} \in \mathbf{V}_{s,h_s}$, $p_{s,h_s} \in Q_{s,h_s}$, $\mathbf{u}_{d,h_d} \in \mathbf{V}_{d,h_d}$, $p_{d,h_d} \in Q_{d,h_d}$, $\lambda_{s,h_s} \in \Lambda_{s,h_s}$ and $\lambda_{d,h_d} \in \Lambda_{d,h_d}$ (satisfying the given boundary conditions) such that:

$$2\mu \int_{\Omega_s} D(\mathbf{u}_{s,h_s}) : D(\mathbf{v}_{s,h_s})\, d\Omega - \int_{\Omega_s} p_{s,h_s} \nabla \cdot \mathbf{v}_{s,h_s}\, d\Omega + \int_{\Gamma} \lambda_{s,h_s} \mathbf{v}_{s,h_s} \cdot \mathbf{n}_s\, d\Gamma \tag{47}$$
$$+ \sum_{j=1}^{d-1} \int_{\Gamma} \alpha_j (\mathbf{u}_{s,h_s} \cdot \boldsymbol{\tau}_j)(\mathbf{v}_{s,h_s} \cdot \boldsymbol{\tau}_j)\, d\Gamma = \int_{\Omega_s} \mathbf{f}_s \cdot \mathbf{v}_{s,h_s}\, d\Omega \qquad \forall \mathbf{v}_{s,h_s} \in \mathbf{V}^D_{s,h_s},$$
$$\int_{\Omega_s} q_{s,h_s} \nabla \cdot \mathbf{u}_{s,h_s}\, d\Omega = 0 \qquad \forall q_{s,h_s} \in Q_{s,h_s},$$
$$\mu \int_{\Omega_d} (\boldsymbol{\kappa}^{-1} \mathbf{u}_{d,h_d}) \cdot \mathbf{v}_{d,h_d}\, d\Omega - \int_{\Omega_d} p_{d,h_d} \nabla \cdot \mathbf{v}_{d,h_d}\, d\Omega \tag{48}$$
$$+ \int_{\Gamma} \lambda_{d,h_d} \mathbf{v}_{d,h_d} \cdot \mathbf{n}_d\, d\Gamma = \int_{\Omega_d} \mathbf{f}_d \cdot \mathbf{v}_{d,h_d}\, d\Omega \qquad \forall \mathbf{v}_{d,h_d} \in \mathbf{V}^N_{d,h_d},$$
$$\int_{\Omega_d} q_{d,h_d} \nabla \cdot \mathbf{u}_{d,h_d}\, d\Omega = 0 \qquad \forall q_{d,h_d} \in Q_{d,h_d},$$
$$\Pi_{ds} w_{s,h_s} + w_{d,h_d} = 0 \qquad \text{on } \Gamma_d, \tag{49}$$
$$\lambda_{s,h_s} = \Pi_{sd} \lambda_{d,h_d} \qquad \text{on } \Gamma_s. \tag{50}$$

The conditions (49)–(50) are the INTERNODES counterpart of the interface condition (14)–(15), obtained by applying the intergrid operators Π_{12} and Π_{21} defined in Sect. 4. More precisely, if we make the associations $d \leftrightarrow 1$ and $s \leftrightarrow 2$, the operator $\Pi_{sd} (= \Pi_{21})$ is used to interpolate on Γ_s the discrete trace of p_{d,h_d} that is known on Γ_d, while $\Pi_{ds} (= \Pi_{12})$ is used to interpolate on Γ_d the weak counterpart of $\mathbf{u}_{s,h_s} \cdot \mathbf{n}_s$ that is known on Γ_s.

Denoting by $\mathtt{w}_s$, $\mathtt{w}_d$, $\mathtt{t}_s$, and $\mathtt{t}_d$, the arrays whose entries are the Lagrangian degrees of freedom of w_{s,h_s}, w_{d,h_d}, λ_{s,h_s}, and λ_{d,h_d} respectively, the algebraic form of the INTERNODES conditions (49)–(50) reads:

$$R_{ds} M_{\Gamma_s}^{-1} \mathtt{w}_s + M_{\Gamma_d}^{-1} \mathtt{w}_d = 0, \qquad \mathtt{t}_s = R_{sd} \mathtt{t}_d. \tag{51}$$

We test the accuracy of INTERNODES by solving problem (12)–(16) with: $\Omega_s = (0,1) \times (1,2)$, $\Omega_d = (0,1) \times (0,1)$, $\mu = 1$, $\kappa = 10^{-2}$, $\boldsymbol{\kappa} = \kappa I$, boundary data and $\mathbf{f}_s = \mathbf{f}_d$ are such that the exact solution is $\mathbf{u}_s = \kappa[-\sin(\frac{\pi}{2}x)\cos(\frac{\pi}{2}y) - y + 1, \cos(\frac{\pi}{2}x)\sin(\frac{\pi}{2}y) - 1 + x]$, $p_s = 1 - x$, $\mathbf{u}_d = \kappa[\sin(\frac{\pi}{2}x)\cos(\frac{\pi}{2}y) + y, \cos(\frac{\pi}{2}x)\sin(\frac{\pi}{2}y) - 1 + x]$, $p_d = \frac{2}{\pi}\cos(\frac{\pi}{2}x)\sin(\frac{\pi}{2}y) - y(x-1)$. The approximation in each subdomain is performed with stabilized hp-fem on quadrilaterals ([6]). The errors $e_s = \|\mathbf{u}_s - \mathbf{u}_{s,h_s}\|_{H^1(\Omega_s)} + \|p_s - p_{s,h_s}\|_{L^2(\Omega_s)}$ and $e_d = \|\mathbf{u}_d - \mathbf{u}_{d,h_d}\|_{L^2(\Omega_d)} + \|p_d - p_{d,h_d}\|_{H^1(\Omega_d)}$ are shown in Figure 4, versus either the mesh sizes h_s, h_d and the polynomial degrees $\mathtt{p}_s$ and $\mathtt{p}_d$, they decay exponentially w.r.t. the polynomial degrees (Fig. 4, at left) and with order $q = \mathtt{p}_s = \mathtt{p}_d$ w.r.t. the mesh sizes (Fig. 4, at center and at right).

In Fig. 5 we show the INTERNODES solution computed for the *cross-flow membrane filtration* test case with non-flat interface Γ. The setting of the problem is given in Sect. 5.3 of [6]. We have considered either a cubic spline interface (Fig. 5 at the left) and a piece-wise interface (Fig. 5 at the right). Quadrilaterals hp-fem are used for the discretization in either Ω_s and Ω_d. The solution at left is obtained with

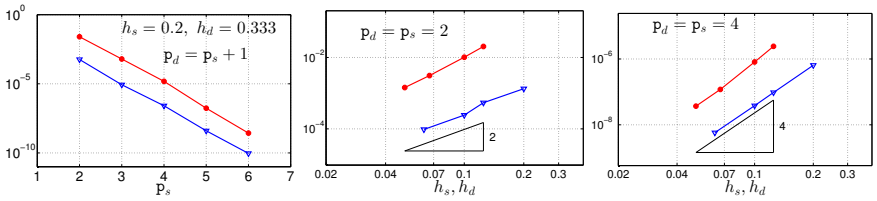


Fig. 4 Errors e_s (red) and e_d (blue) for the Stoked-Darcy problem (12)–(16) solved on non-conforming meshes by the INTERNODES method

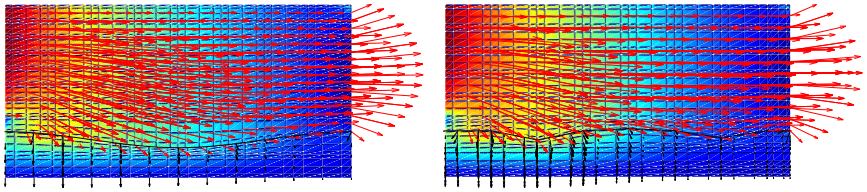


Fig. 5 INTERNODES solution of the Stokes-Darcy coupling. The velocity field $\mathbf{u}_s$ is red in Ω_s and black in Ω_d, the underground colored scalar field the hydrodynamic pressure. Γ is curved at left and piece-wise linear at right

$h_s = 3/8$, $h_d = 1/2$, and $\mathrm{p}_s = \mathrm{p}_d = 4$, that at right with $h_s = h_d = 3/8$, $\mathrm{p}_s = 4$ and $\mathrm{p}_d = 3$. RL-RBF interpolation is used to build the intergrid operators (17) when Γ is curved, and Lagrange interpolation when Γ is piece-wise linear.

Numerical results show that INTERNODES keeps the optimal accuracy of the local discretizations and that it is a versatile method to deal with non-conforming interfaces.

References

1. C. Bernardi, Y. Maday, and A.T. Patera. A new nonconforming approach to domain decomposition: the mortar element method. In *Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. XI (Paris, 1989–1991)*, volume 299 of *Pitman Res. Notes Math. Ser.*, pages 13–51. Longman Sci. Tech., Harlow, 1994.
2. H. J. Brauchli and J. T. Oden. Conjugate approximation functions in finite-element analysis. *Quart. Appl. Math.*, 29:65–90, 1971.
3. S. Deparis, D. Forti, P. Gervasio, and A. Quarteroni. INTERNODES: an accurate interpolation-based method for coupling the Galerkin solutions of PDEs on subdomains featuring non-conforming interfaces. *Computers & Fluids*, 141:22–41, 2016.
4. S. Deparis, D. Forti, and A. Quarteroni. A rescaled localized radial basis function interpolation on non-Cartesian and nonconforming grids. *SIAM J. Sci. Comput.*, 36(6):A2745–A2762, 2014.
5. S. Deparis, D. Forti, and A. Quarteroni. *A fluid-structure interaction algorithm using radial basis function interpolation between non-conforming interfaces*, pages 439–450. Modeling and Simulation in Science, Engineering and Technology. Springer, 2016.

6. M. Discacciati, P. Gervasio, A. Giacomini, and A. Quarteroni. Interface Control Domain Decomposition (ICDD) Method for Stokes-Darcy coupling. *SIAM J. Numer. Anal.*, 54(2):1039–1068, 2016.
7. M. Discacciati and A. Quarteroni. Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. *Rev. Mat. Complut.*, 22(2):315–426, 2009.
8. D. Forti. *Parallel Algorithms for the Solution of Large-Scale Fluid-Structure Interaction Problems in Hemodynamics*. PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne (Switzerland), 4 2016.
9. P. Gervasio and A. Quarteroni. Analysis of the INTERNODES method for non-conforming discretizations of elliptic equations. Technical report, MATHICSE, EPFL, Lausanne (Switzerland), 2016. Submitted.
10. V. Gupta, C.A. Duarte, I. Babuška, and U. Banerjee. Stable GFEM (SGFEM): improved conditioning and accuracy of GFEM/XFEM for three-dimensional fracture mechanics. *Comput. Methods Appl. Mech. Engrg.*, 289:355–386, 2015.
11. W.J. Layton, F. Schieweck, and I. Yotov. Coupling fluid flow with porous media flow. *SIAM J. Numer. Anal.*, 40(6):2195–2218 (2003), 2002.
12. T. Levy and E. Sánchez-Palencia. On boundary conditions for fluid flow in porous media. *Internat. J. Engrg. Sci.*, 13(11):923–940, 1975.
13. A. Quarteroni. *Numerical Models for Differential Problems, 2nd ed.* Springer, 2014.
14. A. Quarteroni, A. Manzoni, and C. Vergara. The cardiovascular system: Mathematical modeling, numerical algorithms, clinical applications. *Acta Numerica*, 26:–, 2017. in press.
15. A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer Verlag, Heidelberg, 1994.

Domain Decomposition Approaches for PDE Based Mesh Generation

Ronald D. Haynes

Abstract Adaptive, partial differential equation (PDE) based, mesh generators are introduced. The mesh PDE is typically coupled to the physical PDE of interest and one has to be careful not to introduce undue computational burden. Here we provide an overview of domain decomposition approaches to reduce this computational overhead and provide a parallel solver for the coupled PDEs. A preview of a new analysis for optimized Schwarz methods for the mesh generation problem using the theory of M–functions is given. We conclude by introducing a two-grid method with FAS correction for the grid generation problem.

1 Introduction

Automatically adaptive and possibly dynamic meshes are often introduced to solve partial differential equations (PDEs) whose solutions evolve on disparate space and time scales. In this paper we will review a class of PDE based mesh generators in 1D and 2D – a PDE is formulated and its solution provides the mesh used to approximate the solution of the physical PDE of interest. The physical PDE and mesh PDE are coupled and are solved in a simultaneous or decoupled manner. The hope is that the cost of computing the mesh, by solving the mesh PDE, should not substantially increase the total computational burden and ideally the mesh solution strategy should fit within the overall solution framework.

Meshes which automatically react to the solution of the physical PDE fall into (at least) two broad categories: hp-refinement and r–refinement — PDE based mesh generation which evolves a fixed number of mesh points with a fixed topology. The choice of mesh generator is often predicated on the class of problem and experience of the practitioner. The PDE based mesh generators, motivated by r-refinement, discussed here, can be designed to capture dynamical physics, Lagrangian behaviour,

Ronald D. Haynes
Memorial University, St. John's, Newfoundland, Canada, e-mail: rhaynes@mun.ca

symmetries, conservation laws or self–similarity features of the physical solution, and achieve global mesh regularity.

In this overview paper, we review parallel solution strategies for the mesh PDE and the coupled system using domain decomposition (DD) and survey various known theoretical results. The analysis of the optimized Schwarz method (OSM) uses several classical tools including Peaceman-Rachford iterations and monotone convergence using the theory of M-functions. We present previews of two extensions of our previous work. We provide an analysis of OSM on two subdomains using the theory of M–functions. We also introduce a coarse correction for the mesh PDE to improve convergence of DD as the number of subdomains increases.

In this paper we provide a brief review of PDE based mesh generation (Section 2), an overview of, and theoretical convergence results for, Schwarz methods to solve the mesh PDE (Section 3), a new strategy for the analysis of OSM and a new coarse correction algorithm to solve the nonlinear mesh PDE (Section 4).

2 PDE based mesh generation

We consider PDEs whose numerical solution can benefit from automatically chosen non-uniform meshes. r-refinement adapts an initial grid by relocating a fixed number of mesh nodes. The mesh is determined by solving a mesh PDE simultaneously, or in an iterative fashion, with the physical PDE. Suppose the PDE defined on the physical domain $x \in \Omega_p = [0,1]$ is difficult to solve in the physical co-ordinate x. We compute a mesh transformation, $x = x(\xi,t)$, so that solving the problem on a uniform mesh $\xi_i = \frac{i}{N}, \quad i = 0,1,\ldots,N$, with moderate N, is sufficient. In one dimension, such a mesh transformation can be constructed by the equidistribution principle of de Boor [4]. Given some measure of the error in the physical solution, M (called the mesh density function), we require

$$\int_{x_{i-1}(t)}^{x_i(t)} M(t,\tilde{x},u)\,d\tilde{x} = \frac{1}{N}\int_0^1 M(t,\tilde{x},u)\,d\tilde{x},$$

which says that the error in the solution is equally distributed across all intervals.

If we assume some approximation to the physical solution u is given, then in the steady case a continuous form of the mesh transformation can be found by solving the nonlinear boundary value problem (BVP)

$$\frac{\partial}{\partial\xi}\left\{M(x(\xi))\frac{\partial}{\partial\xi}x(\xi)\right\} = 0, \quad \text{subject to} \quad x(0) = 0 \text{ and } x(1) = 1. \tag{1}$$

The boundary conditions ensure mesh points at the boundaries of the physical domain. This is equivalent to minimizing the functional $I[x] = \frac{1}{2}\int_0^1 \left(M(x)\frac{dx}{d\xi}\right)^2 d\xi$. Discretizing and solving gives the physical mesh locations directly, however the

Euler-Lagrange (EL) equations are nonlinear, and a system of nonlinear algebraic equations must be solved upon discretization.

As an example, consider constructing an equidistributing grid for the function $u(x) = (e^{\lambda x} - 1)/(e^{\lambda} - 1)$ for large λ. A uniform grid in the physical co-ordinate x would require a large number of mesh points to resolve the boundary layer near $x = 1$. Instead we solve the nonlinear BVP above on a uniform grid $\xi_i = \frac{i}{N}$ with $M(x,u) \sim \sqrt{1+|u_{xx}|^2}$ and we obtain the grid locations corresponding to the abscissa of the green circles in Figure 1. The solution on a uniform grid (white squares) is shown for comparison.

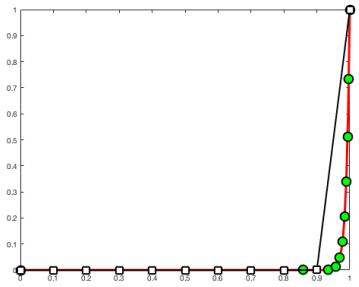


Fig. 1 An example of an equidistributing grid for a boundary layer function.

Alternatively, we can solve for the the inverse transformation, $\xi(x)$, as the solution of

$$\frac{d}{dx}\left(\frac{1}{M(x)}\frac{d\xi}{dx}\right) = 0, \qquad \xi(0) = 0, \quad \xi(1) = 1,$$

or as the minimizer of the functional $I[\xi] = \frac{1}{2}\int_0^1 \frac{1}{M(x)}\left(\frac{d\xi}{dx}\right)^2 dx$.

The EL equations are now linear, and discretizing on a uniform grid in x gives a linear system for the now non-uniform points in the computational co-ordinate ξ. We have to invert the transformation to find the required physical mesh locations. It is easier to ensure well-posedness in higher dimensions ($d \geq 2$) for this formulation.

In two dimensions, solution independent, but boundary fitted meshes, can be found by generalizing the formulations above, but setting the mesh density to be the identity function. The mesh transformation $\boldsymbol{x} = [x(\xi,\eta), y(\xi,\eta)] : \Omega_c \to \Omega_p$ can be found by minimizing

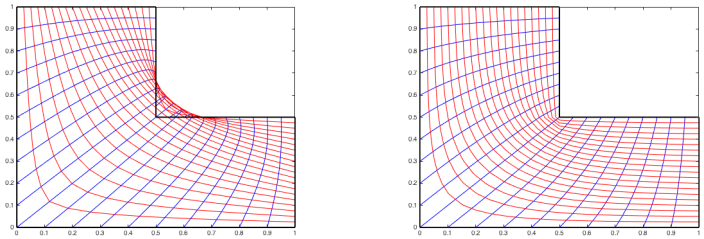
$$I[x,y] = \frac{1}{2}\int_{\Omega_c}\left[\left(\frac{\partial x}{\partial \xi}\right)^2 + \left(\frac{\partial x}{\partial \eta}\right)^2 + \left(\frac{\partial y}{\partial \xi}\right)^2 + \left(\frac{\partial y}{\partial \eta}\right)^2\right] d\xi d\eta.$$

The EL eqns are

$$\frac{\partial^2 x}{\partial \xi^2} + \frac{\partial^2 x}{\partial \eta^2} = 0, \quad \frac{\partial^2 y}{\partial \xi^2} + \frac{\partial^2 y}{\partial \eta^2} = 0.$$

Solving the EL equations subject to boundary conditions, which ensure mesh points on the boundary of Ω_p, gives a boundary fitted co-ordinate system. Care is required, however, as folded meshes may result if Ω_p is concave (see the left of Figure 2 where Ω_p is L-shaped and $\Omega_c = [0,1]^2$).

If instead we solve for the inverse mesh transformation $\boldsymbol{\xi} = [\xi(x,y), \eta(x,y)] : \Omega_p \to \Omega_c$ by minimizing



$x, y : \Omega_c \to \Omega_p$ $\qquad$ $\xi, \eta : \Omega_p \to \Omega_c$

Fig. 2 PDE generated physical grid lines on L–shaped domains.

$$I[\xi,\eta] = \frac{1}{2}\int_{\Omega_p}\left[\left(\frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial \xi}{\partial y}\right)^2 + \left(\frac{\partial \eta}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial y}\right)^2\right] dxdy,$$

or solving the EL equations

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} = 0, \quad \frac{\partial^2 \eta}{\partial x^2} + \frac{\partial^2 \eta}{\partial y^2} = 0,$$

subject to appropriate boundary conditions, we obtain the mesh on the right of Figure 2. This is the *equipotential* mesh generation method of Crowley [7]. The physical grid lines are obtained as level curves $\xi = C$, $\eta = K$. This approach is more robust – well-posed if the domain Ω_c (which we get to choose) is convex, see [8]. But as mentioned previously it is more complicated to get the physical mesh.

Solution dependent meshes in higher dimensions can be constructed by specifying a scalar mesh density function $M = M(u, \boldsymbol{x}) > 0$, characterizing where additional mesh resolution is needed, and minimizing $I[x] = \frac{1}{2}\int_{\Omega_p} \frac{1}{M}\sum_i (\nabla \xi_i)^T \nabla \xi_i dx$. The EL equations give the variable diffusion mesh generator of Winslow [26], which requires the solution of the elliptic PDEs $-\nabla \cdot \left(\frac{1}{M}\nabla \xi_i\right) = 0, \quad i = 1, 2, \ldots, d$. This gives an *isotropic* mesh generator. Godunov and Prokopov [10], Thompson et al. [25] and Anderson [2], for example, add terms to the mesh PDEs to better control the mesh distribution and quality. As an example, in Figure 3, we illustrate the mesh obtained by adapting a mesh for a solution with a rapid transition at $x = 3/4$ and using an arc-length based M.

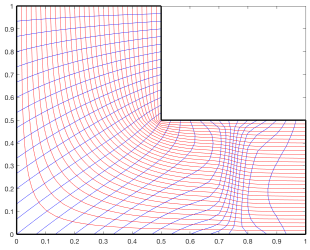


Fig. 3 A mesh generated using a Winslow generator on an L–shaped domain.

If the physical solution has strong *anisotropic* behaviour, corresponding mesh adaptation is desired. This can

be achieved by using a matrix–valued diffusion coefficient [6] and minimizing $I[\boldsymbol{x}] = \frac{1}{2}\int_{\Omega_p} \sum_i (\nabla\xi_i)^T \boldsymbol{M}^{-1}(\nabla\xi_i) d\boldsymbol{x}$ where $\boldsymbol{M}$ is a spd matrix.

These approaches can be extended to the time dependent situation, where $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\xi},t)$ or $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{x},t)$; we obtain moving mesh PDEs as the modified gradient flow equations for the adaptation functionals.

In addition to the variational approach to derive the mesh PDEs mentioned above, there are other PDE based approaches including harmonic maps, Monge–Ampère, and geometric conservation laws, see [15] for a recent extensive overview.

3 Domain Decomposition approaches and analysis for nonlinear mesh generation

We wish to design and analyze parallel approaches to solve the continuous (and discrete forms) of the PDE mesh generators discussed above. Our research goal is to systematically analyze DD based implementations to solve mesh PDEs and coupled mesh–physical PDE systems.

3.1 Mesh/Physical PDE solution strategies

There are several approaches to introduce parallelism, by domain decomposition, while solving PDEs which require or benefit from a PDE based mesh generator. As an example we consider generating a time dependent mesh for a moving interior layer problem. In [14] we apply DD in the physical co-ordinates by partitioning Ω_p, and use an adaptive, moving mesh solver in each physical domain. This is illustrated in the left of Figure 4 for two overlapping subdomains; the solver tracks a front which develops and moves to the right. In each physical subdomain, the mesh points react and follow the incoming front. In general, this approach needs *hr*–refinement to predict the number of mesh points in each subdomain and could result in a severe load balancing issue. Alternatively, one could fix the total number of mesh points and apply DD in the fixed, typically uniform, computational co-ordinates, by partitioning Ω_c. This gives rise to time dependent or moving subdomains, as viewed in the physical co-ordinate system, as shown in the right of Figure 4 for a similar moving front. The subdomains are shaded dark and light gray, with the overlap in between.

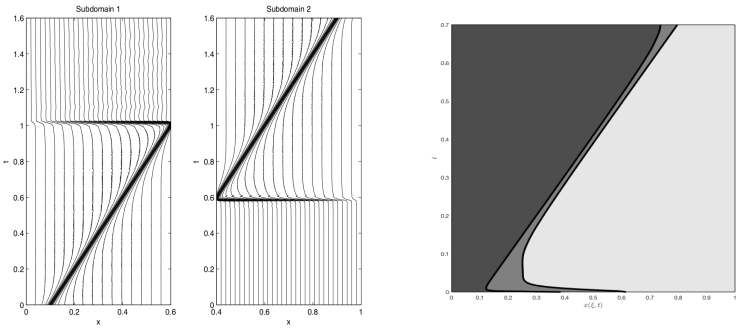


Fig. 4 DD in Ω_p (left) and DD in Ω_c (right).

In Figure 5 we illustrate a two dimensional mesh computed using a classical Schwarz iteration applied in Ω_c, on two overlapping subdomains (the overlap is shown in green). DD is applied to the two dimensional nonlinear mesh generator of [16]. Here the mesh is adapted to the physical solution given by

$$u = \tanh(R(\frac{1}{16} - (x-\frac{1}{2})^2 - (y-\frac{1}{2})^2))$$

and

$$M = \frac{a^2 \nabla u \cdot \nabla u^T}{1 + b \nabla u^T \nabla u} + I,$$

where $a = 0.2$ and $b = 0$.

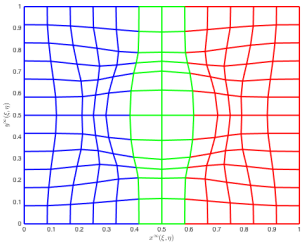


Fig. 5 DD solution of two-dimensional mesh generator

3.2 PDE Based Mesh Generation using Schwarz methods

Here we will focus on the analysis of DD methods for the mesh PDE applied in the computational co-ordinates, assuming an approximation to the solution of the physical PDE is given. To generate the physical mesh locations directly, we are interested in the solution of the nonlinear BVP (1).

A general parallel Schwarz approach would partition $\xi \in \Omega_c$ into two subdomains $\Omega_1 = (0, \beta)$ and $\Omega_2 = (\alpha, 1)$ with $\alpha \leq \beta$. Let x_1^n and x_2^n solve

$$\frac{d}{d\xi}\left(M(x_1^n)\frac{dx_1^n}{d\xi}\right) = 0 \quad \text{on } \Omega_1 \qquad \frac{d}{d\xi}\left(M(x_2^n)\frac{dx_2^n}{d\xi}\right) = 0 \quad \text{on } \Omega_2$$
$$x_1^n(0) = 0 \qquad B_2(x_2^n(\alpha)) = B_2(x_1^{n-1}(\alpha))$$
$$B_1(x_1^n(\beta)) = B_1(x_2^{n-1}(\beta)) \qquad x_2^n(1) = 1,$$

where $B_{1,2}$ are transmission operators between the subdomains.

If $0 < \check{m} \le M(x) \le \hat{m} < \infty$, we show in [9] the overlapping ($\beta > \alpha$) parallel classical Schwarz iteration ($B_{1,2} = I$) converges for any initial guess $x_1^0(\alpha)$, $x_2^0(\beta)$, with a contraction factor $\rho := \frac{\alpha}{\beta}\frac{1-\beta}{1-\alpha} < 1$ which improves with the size of the overlap. As expected $\alpha < \beta$ is needed for convergence. A multidomain result is also given in [9] with a contraction rate that deteriorates as the number of subdomains increases. This result motivates the need for a coarse correction (see Section 4). Optimal Schwarz methods using non-local transmission conditions (TCs) giving finite convergence have been proposed and analyzed in [9, 12]. This comes at a cost as nonlocal TCs are expensive!

We can recover a local algorithm, an OSM, on 2 subdomains by approximating the non-local TCs. We decompose $\xi \in [0,1]$ into two non–overlapping subdomains $\Omega_1 = [0,\alpha]$ and $\Omega_2 = [\alpha,1]$ and approximate the optimal TCs with nonlinear Robin TCs. Using the notation above, we choose $B_1(\cdot) = M(\cdot)\partial_\xi(\cdot) + p(\cdot)$ and $B_2(\cdot) = M(\cdot)\partial_\xi(\cdot) - p(\cdot)$, where p is a constant chosen to improve the convergence rate. The OSM is equivalent to a nonlinear Peaceman–Rachford interface iteration for the interface values

$$\begin{aligned}(pI - R_2)x_2^{n+1}(\alpha) &= (pI - R_1)x_1^n(\alpha),\\ (pI + R_1)x_1^{n+1}(\alpha) &= (pI + R_2)x_2^n(\alpha),\end{aligned} \tag{2}$$

where the operators R_1 and R_2, given by $R_1(x) = \frac{1}{\beta}\int_0^x M\,d\tilde{x}$ and $R_2(x) = \frac{1}{1-\beta}\int_x^1 M\,d\tilde{x}$, are strictly monotonic (increasing and decreasing respectively). This type of iteration has been analyzed by Kellogg and Caspar [17] and Ortega & Rheinboldt [18]. In [9] we show convergence for all $p > 0$ and the contraction rate can be minimized by an appropriate choice of p.

An analysis of the classical Schwarz algorithm at the discrete level has been provided in [13] in the steady and time dependent cases using a θ method to discretize in time. Using the notion of M–functions, which we will revisit in the next section, we have shown convergence of nonlinear Jacobi and Gauss–Seidel (and block versions) starting from super and sub solutions or from a uniform initial guess.

A dramatically different parallel technique for PDE mesh generation has been considered by Haynes and Bihlo in [3]. Motivated by the possible lower accuracy requirements for mesh generation we have investigated stochastic domain decomposition (SDD) methods, proposed by Acebrón et al. [1], Spigler [24], and Peirano and Talay [19]. These methods use the Feynmac-Kac formula (and Monte-Carlo) to approximate the linear mesh generator in 2D/3D along artificial interfaces. These interface solutions then provide boundary conditions for the deterministic solves in the subdomains. No iteration is required, and the method is fully parallel. The

method may be expensive in the relatively rare situation that the mesh is needed with high-accuracy due to the slow convergence of the Monte Carlo evaluations.

4 Some Extensions

In this section, we provide previews of two extensions of the work described above.

4.1 Optimized Schwarz on Many Subdomains

Her we show an alternate approach to obtain a sufficient condition for convergence of the OSM for the grid generation problem. This approach, which guarantees a monotonic convergence result, is generalizable to an arbitrary number of subdomains. Here we will give a flavour of the analysis on two subdomains. The general result was studied by Sarker [22] and will be published elsewhere.

To demonstrate the difficulty of generalizing the OSM analysis to an arbitrary number of subdomains, consider partitioning Ω_c into three non–overlapping subdomains, $[0,\alpha_1],[\alpha_1,\alpha_2]$ and $[\alpha_2,1]$. The analysis of the parallel OSM to generate equidistributing grids requires us to study the interface iteration

$$\begin{aligned} py_1^n + R_1(x_1^n, y_1^n) &= px_2^{n-1} + R_2(x_2^{n-1}, y_2^{n-1}),\\ px_2^n - R_2(x_2^n, y_2^n) &= py_1^{n-1} - R_1(x_1^{n-1}, y_1^{n-1}),\\ py_2^n + R_2(x_2^n, y_2^n) &= px_3^{n-1} + R_3(x_3^{n-1}, y_3^{n-1}),\\ px_3^n - R_3(x_3^n, y_3^n) &= py_2^{n-1} - R_2(x_2^{n-1}, y_2^{n-1}), \end{aligned} \tag{3}$$

where $x_1^n = 0$ and $y_3^n = 1$, $R_i(x_i,y_i) = \frac{1}{\alpha_i - \alpha_{i-1}} \int_{x_i}^{y_i} M(\sigma)\, d\sigma$, and we define $\alpha_0 \equiv 0$ and $\alpha_3 \equiv 1$.

The Peaceman–Rachford analysis relies on the monotonicity of the operators which define the subdomain solutions. The difficulty in the analysis of (3) lies in the coupled system of equations which arise from the middle subdomain. This coupled system involves the operator $pI + H$. The operator $H = (-R_2, R_2)^T$ is not monotonic and hence the two subdomain analysis can not be repeated, at least not in a straightforward way.

We pursue an alternate tack to obtain a sufficient condition for convergence. It is well known that for linear systems, $Ax = b$, Gauss–Seidel and Jacobi will converge for any initial vector if A is symmetric positive definite, or if A is an M–matrix (for example if $a_{ij} \le 0, i \ne j$, $a_{ii} > 0$ and A is strictly diagonally dominant). Analogous results for nonlinear systems, $Fx = b$, where

$$Fx \equiv (f_1(x_1,\ldots,x_n), f_2(x_1,\ldots,x_n),\ldots,f_n(x_1,\ldots,x_n))^T \quad \text{and} \quad b = (b_1,b_2,\ldots,b_n)^T,$$

were obtained by Schechter [23] who showed if F has a continuous, symmetric, and uniformly positive definite (Frechet) derivative then nonlinear Gauss–Seidel converges. The analogous M–matrix condition for convergence was extended to the nonlinear case by Rheinboldt [21], with the introduction of M–functions. To be an M–function requires F to have certain monotonicity, sign and diagonal dominance properties. Rheinboldt gives the following sufficient condition to guarantee a nonlinear map F is an M–function.

Theorem 1. *Let $\mathbb{D}$ be a convex and open subset of $\mathbb{R}^n$. Assume $F : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}^n$ is off-diagonally non-increasing, and that for any $x \in \mathbb{D}$, the functions $q_i : S_i \subset \mathbb{R} \to \mathbb{R}^n$ defined as*

$$q_i(\tau) = \sum_{j=1}^{n} f_j(x + \tau e^i),\ i = 1, \ldots, n, \quad \textit{with} \quad S_i = \{\tau : x + \tau e^i \in \mathbb{D}\},$$

are strictly increasing. Then F is an M-function.

If F is an M–function and if $Fx = b$ has a solution then it is unique. Moreoever, Ortega and Rheinboldt [18] show that if F is a continuous, surjective M–function then for any initial vector the nonlinear Jacobi and Gauss–Seidel processes will converge to the unique solution. Results for the convergence of block versions of these iterations exist [20]. This result generalizes the classical result of Varga for M–matrices. We note that the parallel OSM (3) is a nonlinear block Jacobi iteration.

As an application of this theory we reconsider the two subdomain iteration (2). The technique generalizes to an arbitrary number of subdomains. The iteration (2) is well–posed. Existence and uniqueness for a given right hand side is trivial since the functions are uniformly monotone and tend to $\pm\infty$ as $x_{1,2} \to \pm\infty$. The two subdomain interface solution would solve the system $F = (f_1, f_2)^T = 0$ where $f_1(x,y) = R_1(x) - R_2(y) + px - py = 0$ and $f_2(x,y) = -R_2(y) + R_1(x) + py - px = 0$. In [22] Sarker obtains the following result.

Theorem 2. *The function $F = (f_1, f_2)^T$ above is a surjective M–function if $p > max\{1/\alpha, 1/(1-\alpha)\}\hat{m}$. Hence, the iteration (2) will converge to the unique solution of $F = 0$ for any initial vector. The convergence will be monotone if we start from a super or sub solution.*

Proof. Clearly the function F is continuous. By direct calculation and the bounds on M we have $\frac{\partial f_1}{\partial x} = \frac{1}{\alpha}M(x) + p > 0$ and $\frac{\partial f_2}{\partial y} = \frac{1}{1-\alpha}M(y) + p > 0$, for all $p > 0$. Hence f_1 and f_2 are strictly increasing. Therefore, F is strictly diagonally increasing. Furthermore, $\frac{\partial f_1}{\partial y} = \frac{1}{1-\alpha}M(y) - p$ and $\frac{\partial f_2}{\partial x} = \frac{1}{\alpha}M(x) - p$. Hence, if $p > \{\frac{\hat{m}}{\alpha}, \frac{\hat{m}}{1-\alpha}\}$ then F is off-diagonally decreasing. A super (sub) solution, a vector $(\hat{x}, \hat{y})$ satisfying $F(\hat{x}, \hat{y}) \geq 0\,(\leq 0)$, can easily be constructed [22]. Monotone convergence from $(\hat{x}, \hat{y})$ follows from Theorem 13.5.2 of [18].

To show that F is an M–function, we now consider the functions $q_i(t) = \sum_{j=1}^{2} f_j(X + te^i)$ where $e^i \in R^2$ is the i-th standard basis vector, for $i = 1, 2$. The functions $q_1(t)$ and $q_2(t)$ are given by $q_1(t) = f_1(x+t, y) + f_2(x+t, y) = 2R_1(x+$

$t) - 2R_2(y)$ and $q_2(t) = f_1(x, y+t) + f_2(x,, y+t) = 2R_1(x) - 2R_2(y+t)$. Hence $\frac{dq_1}{dt} = \frac{2}{\alpha}M(t) > 0$ and $\frac{dq_2}{dt} = \frac{2}{1-\alpha}M(t) > 0$ and we conclude that q_i is strictly increasing, for $i = 1, 2$. Hence F is an M-function from Theorem 1. Surjectivity requires a super and sub solution for $Fx = b$ for a general b, see [22]. The convergence from any initial vector then follows from Theorem 13.5.9 of [18].

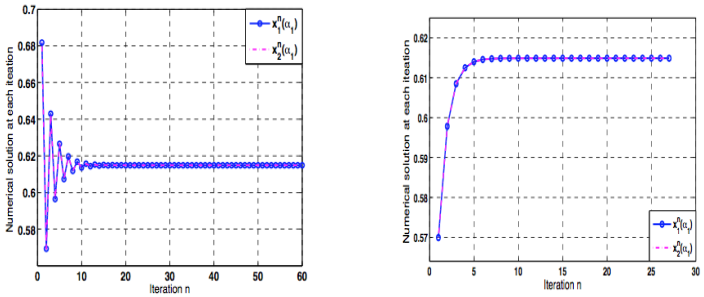


Fig. 6 Convergence history of the interface iteration for small and large p values.

In Figure 6 we see monotonic convergence (consistent with the M–function theory) if p is large enough and non-monotonic convergence for small p (consistent with the Peaceman–Rachford theory).

4.2 A Coarse Correction

The convergence rate of Schwarz methods suffers as the number of subdomains increases, see the left plot in Figure 7. A coarse correction is able to improve the situation dramatically by providing a global transfer of solution information. Here we propose a coarse correction for the (nonlinear) PDE based mesh generation problem by using a two-grid method with a full approximation scheme (FAS) correction applied in the computational co-ordinates. This work was completed by Grant in [11] and will be published in full elsewhere.

FAS [5] provides a solution strategy for nonlinear PDEs. FAS restricts an approximation (and corresponding residual) of the PDE, obtained on a fine grid, to a coarse grid. The error in the approximation is found by solving a coarse problem. This error is then interpolated back to the fine grid and used to update the solution approximation.

FAS may be combined with a DD approach in a very natural way. We perform one classical Schwarz iteration to obtain approximate subdomain solutions on a fine grid. FAS is then applied to update the subdomain solutions before proceeding with the next Schwarz iteration. As shown in the right plot of Figure 7, the effect

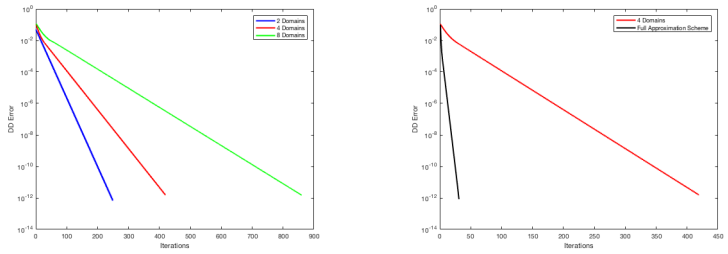


Fig. 7 Schwarz convergence results on multidomains and multidomains with a coarse correction.

is dramatic. This promising result for the nonlinear PDE mesh generator suggests the possibility of a two-grid FAS DD approach for the coupled mesh and physical PDEs.

5 Conclusions

PDE based mesh generators can be useful for problems which would benefit from automatically adaptive spatial grids. It is possible to analyze DD approaches for nonlinear mesh generators which directly give the physical mesh locations. We can then incorporate DD, within the coupled physical PDE/mesh PDE solution frameworks in a theoretically sound way.

Acknowledgements I would like to thank my former students Alex Howse, Devin Grant and Abu Sarker for their assistance and some of the plots included in this paper, and also Felix Kwok for several discussions related to this work.

References

1. Acebrón, J.A., Busico, M.P., Lanucara, P., Spigler, R.: Domain decomposition solution of elliptic boundary-value problems via Monte Carlo and quasi-Monte Carlo methods. SIAM Journal on Scientific Computing **27**(2), 440–457 (2005)
2. Anderson, D.A.: Equidistribution schemes, Poisson generators, and adaptive grids. Applied Mathematics and Computation **24**(3), 211–227 (1987)
3. Bihlo, A., Haynes, R.D.: Parallel stochastic methods for PDE based grid generation. Computers & Mathematics with Applications **68**(8), 804–820 (2014)
4. de Boor, C.: Good approximation by splines with variable knots. Spline Functions and Approximation Theory **21**(Chapter 3), 57–72. Internat. Ser. Numer. Math., Vol. 21 (1973)
5. Brandt, A.: Multi-level adaptive solutions to boundary-value problems. Mathematics of Computation **31**(138), 333–390 (1977)
6. Cao, W., Huang, W., Russell, R.D.: A study of monitor functions for two-dimensional adaptive mesh generation. SIAM Journal on Scientific Computing **20**(6), 1978–1994 (1999)

7. Crowley, W.P.: An equipotential zoner on a quadrilateral mesh. Memo, Lawrence Livermore National Lab **5** (1962)
8. Dvinsky, A.S.: Adaptive grid generation from harmonic maps on Riemannian manifolds. Journal of Computational Physics **95**(2), 450–476 (1991)
9. Gander, M.J., Haynes, R.D.: Domain decomposition approaches for mesh generation via the equidistribution principle. SIAM Journal on Numerical Analysis **50**(4), 2111–2135 (2012)
10. Godunov, S.K., Prokopov, G.P.: The use of moving meshes in gas-dynamical computations. USSR Computational Mathematics and Mathematical Physics **12**(2), 182–195 (1972)
11. Grant, D.: Acceleration techniques for mesh generation via domain decomposition methods. B.Sc., Memorial University of Newfoundland, St. John's, Newfoundland, Canada (2015)
12. Haynes, R.D., Howse, A.J.M.: Alternating Schwarz methods for partial differential equation-based mesh generation. International Journal of Computer Mathematics **92**(2), 349–376 (2014)
13. Haynes, R.D., Kwok, F.: Discrete analysis of domain decomposition approaches for mesh generation via the equidistribution principle. Mathematics of Computation **86**(303), 233–273 (2017)
14. Haynes, R.D., Russell, R.D.: A Schwarz waveform moving mesh method. SIAM Journal on Scientific Computing **29**(2), 656–673 (2007)
15. Huang, W., Russell, R.D.: Adaptive moving mesh methods, *Applied Mathematical Sciences*, vol. 174. Springer, New York, NY (2011)
16. Huang, W., Sloan, D.M.: A simple adaptive grid method in two dimensions. SIAM Journal on Scientific Computing **15**(4), 776–797 (1994)
17. Kellogg, R.B.: A nonlinear alternating direction method. Mathematics of Computation **23**, 23–27 (1969)
18. Ortega, J.M., Rheinboldt, W.C.: Iterative solution of nonlinear equations in several variables. SIAM (2000)
19. Peirano, E., Talay, D.: Domain decomposition by stochastic methods. In: Domain decomposition methods in science and engineering, pp. 131–147. Natl. Auton. Univ. Mex., México (2003)
20. Rheinboldt, W.: On classes of n-dimensional nonlinear mappings generalizing several types of matrices. In: B. Hubbard (ed.) Numerical Solution of Partial Differential Equations - II, Synspade 1970, pp. 501–545. Academic Press, New York-London (1971)
21. Rheinboldt, W.C.: On M-functions and their application to nonlinear Gauss-Seidel iterations and to network flows. Journal of Mathematical Analysis and Applications **32**(2), 274–307 (1970)
22. Sarker, A.: Optimized Schwarz domain decomposition approaches for the generation of equidistributing grids. M.Sc., Memorial University of Newfoundland, St. John's, Newfoundland, Canada (2015)
23. Schechter, S.: Iteration methods for nonlinear problems. Trans. Amer. Math. Soc **104**(1), 179–189 (1962)
24. Spigler, R.: A probabilistic approach to the solution of PDE problems via domain decomposition methods (1991). ICIAM 1991, The Second International Conference on Industrial and Applied Mathematics
25. Thompson, J.F., Thames, F.C., Mastin, C.W.: Automatic numerical generation of body-fitted curvilinear coordinate system for field containing any number of arbitrary two-dimensional bodies. Journal of Computational Physics **15**(3), 299–319 (1974)
26. Winslow, A.M.: Adaptive-mesh zoning by the equipotential method. Tech. rep., Lawrence Livermore National Lab., CA (USA) (1981)

Proceedings of 24th International Conference on Domain Decomposition Methods, Svalbard,
Lecture Notes in Computational Science and Engineering

Modeling, Structure and Discretization of Hierarchical Mixed-dimensional Partial Differential Equations

J. M. Nordbotten and W. M. Boon
Department of Mathematics
University of Bergen
Norway

Abstract

Mixed-dimensional partial differential equations arise in several physical applications, wherein parts of the domain have extreme aspect ratios. In this case, it is often appealing to model these features as lower-dimensional manifolds embedded into the full domain. Examples are fractured and composite materials, but also wells (in geological applications), plant roots, or arteries and veins.

In this manuscript, we survey the structure of mixed-dimensional PDEs in the context where the sub-manifolds are a single dimension lower than the full domain, including the important aspect of intersecting sub-manifolds, leading to a hierarchy of successively lower-dimensional sub-manifolds. We are particularly interested in partial differential equations arising from conservation laws. Our aim is to provide an introduction to such problems, including the mathematical modeling, differential geometry, and discretization.

1. Introduction

Partial differential equations (PDE) on manifolds are a standard approach to model on high-aspect geometries. This is familiar in the setting of idealized laboratory experiments, where 1D and 2D representations are used despite the fact that the physical world is 3D. Similarly, it is common to consider lower-dimensional models in applications ranging from geophysical applications. Some overview expositions for various engineering problems can be found in [1, 2, 3].

Throughout this paper we will consider the ambient domain to be 3D, and our concern is when models on 2D submanifolds are either coupled to the surrounding domain, and/or intersect on 1D and 0D submanifolds. Such models are common in porous media, where the submanifolds may represent either fractures (see e.g. [4]) or thin porous strata (see [1]), but also appear in materials [3]. In all these examples, elliptic differential equations representing physical conservation laws are applicable on all subdomains, and the domains of different dimensionality are coupled via discrete jump conditions. These systems form what we will consider as mixed-dimensional elliptic PDEs, and we will limit the exposition herein to this case.

In order to establish an understanding for the physical setting, we will in section 2 present a short derivation of the governing equations for fractured porous media, emphasizing the conservation

structure and modeling assumptions. This derivation will lead to familiar models from literature (see e.g. [4, 5, 6, 7] and references therein).

We develop a unified treatment of mixed-dimensional differential operators on submanifolds of various dimensionality, using the setting of exterior calculus, and thus recast the physical problem in the sense of differential forms. We interpret the various subdomains as an imposed structure on the original domain, and provide a decomposition of differential forms onto the mixed-dimensional structure. By introducing a suitable inner product, we show that this mixed-dimensional space is a Hilbert space. On this decomposition we define a semi-discrete exterior derivative, which leads to a de Rham complex with the same co-homology structure as the original domain. It is interesting to note that the differential operators we define were independently considered by Licht who introduced the concept of discrete distributional differential forms [8]. A co-differential operator can be defined via the inner product, and it is possible to calculate an explicit expression for the co-differential operator. This allows us to establish a Helmholtz decomposition on the mixed-dimensional geometry. We also define the mixed-dimensional extensions of the familiar Sobolev spaces.

Having surveyed the basic ingredients of a mixed-dimensional calculus, we are in a position to discuss elliptic minimization problems. Indeed, the mixed dimensional minimization problems are well-posed with unique solutions based on standard arguments, and we also state the corresponding Euler equations (variational equations). With further regularity assumptions, we also give the strong form of the minimization problems, corresponding to conservation laws and constitutive laws for mixed-dimensional problems.

This paper aims to provide a general overview and roadmap for the concepts associated with hierarchical mixed-dimensional partial differential equations, more complete and detailed analysis will necessarily due to space be considered in subsequent publications.

2. Fractured porous media as a mixed-dimensional PDE

This section gives the physical rationale for mixed-dimensional PDE. As the section is meant to be motivational, we will omit technical details whenever convenient. We will return to these details in the following sections.

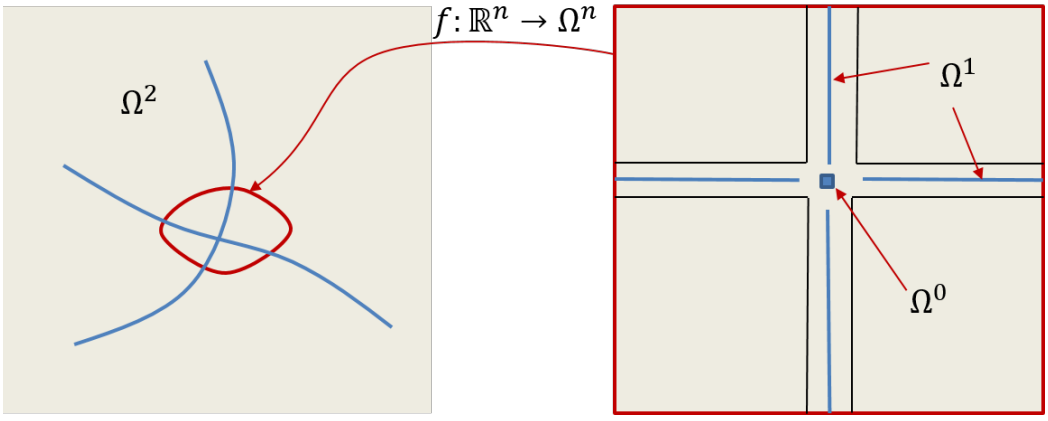


Figure 1: Example geometry of two intersecting fractures in 2D, and the logical representation of the intersection after mapping to a local coordinate system.

We consider the setting of a domain $D \in \mathbb{R}^n$. In sections 3 and onwards we will consider arbitrary n, however in this section we will for simplicity of exposition consider only $n = 3$. We consider a fractured media, where we are given explicit knowledge of the fractures, thus we consider the domains Ω_i^d as given, where $i \in I$ is an index and $d = d(i)$ represents the dimensionality of the domain. We denote by $i \in I^d$ the subset of indexes in I for which $d(i) = d$. In particular, intact material lies in domains of $d = 3$, while $d = 2$ represents fracture segments, and $d = 1$ represents intersections, see Figure 1. For each domain Ω_i^d we assign an orientation based on $n - d$ outer normal vectors $\boldsymbol{\nu}_{ij}$.

In order to specify the geometry completely, we consider the index sets $\hat{S}_i$ and $\check{S}_i$ as the $d + 1$ dimensional and $d - 1$ dimensional neighbors of a domain i. Thus for $d = 2$, the set $\hat{S}_i$ contains the domain(s) Ω_l^3 which are on the positive (and negative) side of Ω_i^2. On the other hand, the set $\check{S}_i$ contains the lines that form (parts of) the boundary of Ω_i. Additionally, the set of all lower-dimensional neighbors is defined as $\check{\mathfrak{S}}_i = \left[\check{S}_i, \check{S}_{\check{S}_i}, \dots\right]$ We will define $\Omega^d = \sum_{i \in I^d} \Omega_i^d$ as all subdomains of dimension d, while similarlty $\Omega = \sum_{d=0}^{n} \Omega^d$ is the full mixed-dimensional stratification. Note that since the superscript indicating dimension is redundant when the particular domain is given, we will (depending what offers more clarity) use $\Omega_i = \Omega_i^d$ interchangeably.

For steady-state flows in porous media, the fluid satisfies a conservation law, which for intact rock and an n-dimensional fluid flux vector $\boldsymbol{u}$ takes the form

$$\nabla \cdot \boldsymbol{u} = \phi \qquad \text{on} \qquad D \tag{2.1}$$

We wish to express this conservation law with respect to our geometric structure. To this end, let us first define the mixed-dimensional flux $\mathbf{u}$, which is simply a d-dimensional vector field on each Ω_i^d. We write $\mathbf{u} = \left[\boldsymbol{u}_i^d\right]$ when we want to talk about specific components of $\mathbf{u}$. We similarly define other mixed-dimensional variables, such as the source-term f.

Now clearly, for $d = n$, we recover equation (2.1). Now consider $d = n - 1$, and a fracture Ω_1 of variable Lipschitz-continuous aperture (illustrated for $d(1) = 1$ in figure 2).

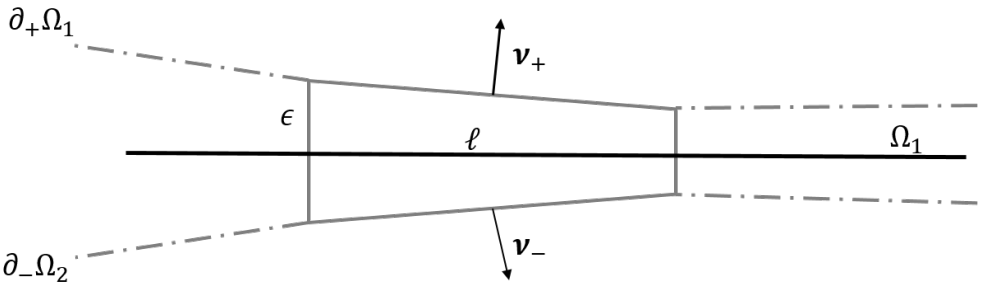


Figure 2: Example of local geometry for derivation of mixed-dimensional conservation law.

Here the dashed lines indicate a fracture boundary, the solid black line is the lower-dimensional representation, and the solid gray line indicates the region of integration, ω, of length ℓ and width $\epsilon(x)$. Evaluating the conservation law over ω leads to

$$\int_\omega \nabla \cdot \boldsymbol{u} \; da = \int_{\partial\omega} \boldsymbol{u} \cdot \boldsymbol{\nu} \; ds = \int_\omega \phi$$

where $\boldsymbol{\nu}$ are the external normal vectors. Since our integration area is in the limiting case of $\ell \to 0$ a quadrilateral, we split the last integral into parts where $\boldsymbol{\nu}$ is constant,

$$\int_{\partial\omega} \boldsymbol{u}\cdot\boldsymbol{\nu}\ ds = \boldsymbol{\nu}_+ \cdot \int_{\partial_+\omega} \boldsymbol{u}_{l_1}\ ds + \boldsymbol{\nu}_- \cdot \int_{\partial_-\omega} \boldsymbol{u}_{l_2}\ ds + \int_{\partial_R\omega} \boldsymbol{\tau}\cdot\boldsymbol{u}\ ds - \int_{\partial_L\omega} \boldsymbol{\tau}\cdot\boldsymbol{u}\ ds$$

where $[l_1, l_2] \in \hat{S}_i$ is the domain on the "+" and "-" side of Ω_1, respectively, and denote the Left and Right side of the integration boundary by subindexes. The notation $\boldsymbol{\tau}$ is the tangential vector to Ω_1. Clearly, letting the length ℓ be infinitesimal, the last two terms satisfy

$$\lim_{\ell\to 0} \frac{\int_{\partial\omega_R} \boldsymbol{\tau}\cdot\boldsymbol{u}\ ds - \int_{\partial\omega_L} \boldsymbol{\tau}\cdot\boldsymbol{u}\ ds}{\ell} = \nabla_{\Omega_1}\cdot \int_{\partial_-\omega}^{\partial_+\omega} \boldsymbol{\tau}\cdot\boldsymbol{u}\ ds = \nabla_{\Omega_1}\cdot(\epsilon\boldsymbol{u}_1)$$

where $\nabla_{\Omega_1}\cdot$ is the in-plane divergence and

$$\boldsymbol{u}_1 \equiv \frac{1}{\epsilon}\int_{\partial_-\omega}^{\partial_+\omega} \boldsymbol{\tau}\cdot\boldsymbol{u}\ ds \tag{2.3}$$

Considering similarly the limits of $\ell \to 0$ for the two first terms, we obtain for the positive side

$$\lim_{\ell\to 0} \boldsymbol{\nu}_+ \cdot \ell^{-1} \int_{\partial_+\omega} \boldsymbol{u}_{l_1}^{d(l_1)}\ ds = \left(1 + \left|\frac{d}{dx}\nabla_{\Omega_1}(\partial_+\omega)\right|^2\right)^{1/2} \boldsymbol{\nu}_+ \cdot \boldsymbol{u}_{l_1}^{d(l_1)}$$

Combining the above, we thus have

$$\lim_{\ell\to 0} \ell^{-1}\int_\omega \nabla\cdot\boldsymbol{u}\ da = \lambda_{l_1} + \lambda_{l_2} + \nabla_{\Omega_1}\cdot(\epsilon\boldsymbol{u}_1) = [\![\lambda]\!]_i + \nabla_{\Omega_1}\cdot(\epsilon\boldsymbol{u}_1) \tag{2.4}$$

where λ is defined as

$$\lambda_{l_1} = \left(1 + \left|\frac{d}{dx}\nabla_{\Omega_1}(\partial_+\omega)\right|^2\right)^{1/2} \boldsymbol{\nu}_+ \cdot \boldsymbol{u}_{l_1}^{d(l_1)} \tag{2.5}$$

and (using the analogous definition for λ_{l_2})

$$[\![\lambda]\!]_i = -\sum_{l\in\hat{S}_i} \lambda_l \tag{2.6}$$

Note that we have made no approximations in obtaining equation (2.4) – the left-hand side is an exact expression of conservation. The model approximations appear later when deriving suitable constitutive laws. Nevertheless, since the fractures have a high aspect ratio by definition, the pre-factor in equation (2.5) is in practice often approximated by identity, for which (2.5) simplifies to

$$\lambda_l \approx \boldsymbol{\nu}_\pm \cdot \boldsymbol{u}_l \tag{2.7}$$

The derivation above (including the definition in equation (2.4)), generalizes in the same way to intersection lines and intersection points, thus we find that for all $d < n$ it holds that

$$[\![\hat{\epsilon}\lambda]\!]_i + \nabla_{\Omega_i}\cdot(\epsilon_i\boldsymbol{u}_i) = \phi_i \tag{2.8}$$

Here the hat again denotes the next higher-dimensional domains, so that $\hat{\epsilon} = \epsilon_l$. Since $\hat{S}_i = \emptyset$ for $i \in I^n$, equation (2.8) reduces to (2.1) for $d = n$, and thus it represents the mixed-dimensional conservation

law for all Ω_i^d. In this more general setting, ϵ denotes the cross-sectional width (2D), area (1D) and volume (0D) for successively lower-dimensional intersections.

For porous materials, the conservation law (2.1) is typically closed by introducing Darcy's law as a modeling assumption, stated in terms of a potential p on the domain D as

$$\boldsymbol{u} = -K\nabla p \tag{2.9}$$

The coefficient K is in general a tensor. Unlike for the conservation law, it is not possible to derive an exact expression for the mixed-dimensional constitutive law, but by making some (reasonable) assumptions on the structure of the solution, it is usually accepted that Darcy's law is inherited for each subdomain (see extended discussion in [1], but also [9]), i.e.

$$\boldsymbol{u}_i = -K_i\nabla_{\Omega_i} p_i \tag{2.10}$$

To close the model, it is also necessary to specify an additional constraint, where the two most common choices are that either the potential is continuous (see discussion in [10])

$$\hat{p}_+ = \hat{p}_- \tag{2.11}$$

or, more generally, that the pressure is discontinuous but related to the normal flux above

$$\lambda_l = -2\widehat{K}_{i,\pm}\frac{p_i - \hat{p}_\pm}{(\epsilon_i)^{\frac{1}{n-d}}} \tag{2.12}$$

The model equations (2.8-2.12) are typical of those used in practical applications [11]. However, to the authors' knowledge, our work is the first time they are explicitly treated as a mixed-dimensional PDE (see also [12, 13]).

3. Exterior calculus for mixed-dimensional geometries

We retain the same geometry as in the previous section, but continue the exposition in the language of exterior calculus (for introductions, see [14, 15, 16]). Throughout the section, we will assume that all functions are sufficiently smooth for the derivatives and traces to be meaningful. We also point out that similar structures to those discussed in this section have been considered previously by Licht in a different context [8].

First, we note that the components of the mixed-dimensional flux discussed in section 2 all correspond to $d-1$ forms, $\boldsymbol{u}_i^d \in \Lambda^{d-1}(\Omega_i^d)$, while the components of pressure all correspond to d-forms, $p_i^d \in \Lambda^d(\Omega_i^d)$. This motivates us to define the following mixed-dimensional k-form

$$\mathfrak{L}^k(\Omega) = \prod_{i\in I}\Lambda^{k-(n-d(i))}(\Omega_i^d) \tag{3.1}$$

From here on, it is always assumed that $\mathfrak{L}^k$ is defined over Ω, and the argument is suppressed.

Moreover, we note that equation (2.7) is (up to a sign) the trace with respect to the inclusion map of the submanifold, thus for a mixed-dimensional variable $\mathfrak{a} \in \mathfrak{L}^k$ the jump operator is naturally written as

$$(\mathbb{d}\mathfrak{a})_i^d = (-1)^{d+k}\sum_{j\in\hat{S}_i}\varepsilon(\Omega_i^d, \partial_i\Omega_j^{d+1})\mathrm{Tr}_{\Omega_i^d}\, a_j^{d+1} \tag{3.2}$$

Here we have exchanged the bracket notation of equation (2.5b), which is common in applications, with a simpler notation, $\mathbb{d}$, which more clearly emphasizes that this is a (discrete) differential operator, in the normal direction(s) with respect to the submanifold. We use the notation $\varepsilon\left(\Omega_i^d, \partial_i \Omega_j^{d+1}\right)$ to indicate the relative orientation (positive or negative) of the arguments.

We obtain a mixed-dimensional exterior derivative, which we denote $\mathfrak{d}$, by combining the jump operator with the exterior derivative on the manifold, such that for $\mathfrak{a} \in \mathfrak{Q}^k$

$$(\mathfrak{d}\mathfrak{a})_i^d = d a_i^d + (\mathbb{d}\mathfrak{a})_i^d \tag{3.3}$$

This expression is meaningful, since both $d a_i^d, (\mathbb{d}\mathfrak{a})_i^d \in \Lambda^{k-(n-d)+1}\left(\Omega_i^d\right)$, and thus clearly $\mathfrak{d}\mathfrak{a} \in \mathfrak{Q}^{k+1}$. A straight-forward calculation shows that $d(\mathbb{d}\mathfrak{a})_i^d = -(\mathbb{d} d\mathfrak{a})_i^d$, thus for all $\mathfrak{a}$

$$\mathfrak{d}\mathfrak{d}\mathfrak{a} = 0 \tag{3.4}$$

and it can furthermore be shown that if $\mathfrak{a} = 0$, and if D is contractible, then there exists $\mathfrak{b} \in \mathfrak{Q}^{k-1}$ such that $\mathfrak{a} = \mathfrak{d}\mathfrak{b}$. Thus the mixed-dimensional exterior derivative forms a de Rham complex,

$$0 \to \mathbb{R} \xrightarrow{\subset} \mathfrak{Q}^0 \xrightarrow{\mathfrak{d}} \mathfrak{Q}^1 \xrightarrow{\mathfrak{d}} \dots \xrightarrow{\mathfrak{d}} \mathfrak{Q}^n \to 0 \tag{3.5}$$

which is exact (for the proof of this, and later assertions, please confer [13]).

Due to the jump terms in the differential operators, the natural inner product for the mixed-dimensional geometry must take into account the traces on boundaries, and thus takes the form for $\mathfrak{a}, \mathfrak{b} \in \mathfrak{Q}^k$

$$(\mathfrak{a}, \mathfrak{b}) = \sum_{i \in I} \left(\left(a_i^d, b_i^d\right) + \sum_{l \in \breve{\Xi}_i^d} \left(\mathrm{Tr}_{\Omega_l^{d(l)}} a_i^d, \mathrm{Tr}_{\Omega_l^{d(l)}} b_i^d \right) \right) \tag{3.6}$$

Note that $\Lambda^k\left(\Omega_i^d\right) = \emptyset$ whenever $k \notin [0, d]$, thus many of the terms in (3.6) are void. It is easy to verify that equation (3.6) indeed defines an inner product, and thus forms the norm on $\mathfrak{Q}^k$

$$\|\mathfrak{a}\| = (\mathfrak{a}, \mathfrak{a})^{1/2} \tag{3.7}$$

The codifferential $\mathfrak{d}^*: \mathfrak{Q}^k \to \mathfrak{Q}^{k-1}$ is defined as the dual of the exterior derivative with respect to the inner product, such that for $\mathfrak{a} \in \mathfrak{Q}^k$

$$(\mathfrak{d}^*\mathfrak{a}, \mathfrak{b}) = (\mathfrak{a}, \mathfrak{d}\mathfrak{b}) + (\mathrm{Tr}\,\mathfrak{b}, \mathrm{Tr}^*\mathfrak{a})_{\partial \mathrm{D}} \qquad \text{for all } \mathfrak{b} \in \mathfrak{Q}^{k-1} \tag{3.8}$$

It follows from the properties of inner product spaces that the codifferential also forms an exact de Rham sequence. Thus, when D is contractible, we have the following Helmholtz decomposition: For all $\mathfrak{a} \in \mathfrak{Q}^k$, there exist $\mathfrak{a}_\mathfrak{d} \in \mathfrak{Q}^{k-1}$ and $\mathfrak{a}_{\mathfrak{d}^*} \in \mathfrak{Q}^{k+1}$ such that

$$\mathfrak{a} = \mathfrak{d}\mathfrak{a}_\mathfrak{d} + \mathfrak{d}^*\mathfrak{a}_{\mathfrak{d}^*} \tag{3.9}$$

In view of the uncertainty in the modeling community of the correct constitutive laws for mixed-dimensional problems (as per the discussion of equation (2.11) and (2.12)), it is of great practical utility to be able to explicitly calculate the co-differential, since this will have the structure of the constitutive law. Utilizing equations (3.6) and (3.8), we obtain

$$(\mathfrak{d}^*\mathfrak{b})_i^d = d^* b_i^d \qquad \text{on } \Omega_i^d \tag{3.10}$$

and

$$\mathrm{Tr}_{\partial_l\Omega_i^d}\,(\mathfrak{d}^*\mathfrak{b})_i^d = d^*\mathrm{Tr}_{\partial_l\Omega_i^d}\,b_i^d + \left(\mathrm{Tr}^*_{\partial_l\Omega_i^d}\,b_i^d - \sum_{j\in\check{S}_i^d}(-1)^{d+k}\varepsilon\left(\Omega_j^{d-1},\partial_j\Omega_i^d\right)b_j^{d-1}\right) \qquad \text{on } \partial\Omega_i^d \tag{3.11}$$

We close this section by noting that the differential operators provide the basis for extending Hilbert spaces to the mixed-dimensional setting. In particular, we are interested in the first order differential spaces, and therefore the norms of $H\mathfrak{L}^k$ and $H^*\mathfrak{L}^k$ by

$$\|\mathfrak{a}\|_H = \|\mathfrak{a}\| + \|\mathfrak{d}\mathfrak{a}\| \qquad \text{and} \qquad \|\mathfrak{a}\|_{H^*} = \|\mathfrak{a}\| + \|\mathfrak{d}^*\mathfrak{a}\| \tag{3.12}$$

from which we obtain the spaces

$$H\mathfrak{L}^k := \{\mathfrak{a}\in\mathfrak{L}^k \,\big|\, \|\mathfrak{a}\|_H < \infty\} \qquad \text{and} \qquad H^*\mathfrak{L}^k := \{\mathfrak{a}\in\mathfrak{L}^k \,\big|\, \|\mathfrak{a}\|_{H^*} < \infty\} \tag{3.13}$$

We use the convention that a circle above the function space denotes homogeneous boundary conditions, i.e. $\mathring{H}\mathfrak{L}^k$: $\{\mathfrak{a}\in H\mathfrak{L}^k \,|\mathrm{Tr}_{\partial D}\,\mathfrak{a}=0\}$ and $\mathring{H}^*\mathfrak{L}^k$: $\{\mathfrak{a}\in H^*\mathfrak{L}^k \,|\mathrm{Tr}^*_{\partial D}\,\mathfrak{a}=0\}$. The spaces $H\mathfrak{L}^k$ and $H^*\mathfrak{L}^k$ can be characterized in terms of product spaces of functions defined on domains Ω_i^d and its boundary components $\partial_j\Omega_i^d$, see e.g. [13, 12].

Then, the Poincaré inequality holds for contractible domains in the mixed-dimensional setting for either $\mathfrak{a}\in\mathring{H}\mathfrak{L}^k\cap H^*\mathfrak{L}^k$ or $\mathfrak{a}\in H\mathfrak{L}^k\cap\mathring{H}^*\mathfrak{L}^k$:

$$\|\mathfrak{a}\| \le C_\Omega(\|\mathfrak{d}\mathfrak{a}\| + \|\mathfrak{d}^*\mathfrak{a}\|) \tag{3.14}$$

4. Mixed-dimensional elliptic PDEs

Based on the extension of the exterior derivative and its dual to the mixed-dimensional setting, we are now prepared to define the generalization of elliptic PDEs. We start by considering the minimization problem equivalent to the Hodge Laplacian for $\mathfrak{a}\in\mathfrak{L}^k$

$$\mathfrak{a} = \arg\inf_{\mathfrak{a}\in\mathring{H}\mathfrak{L}^k\cap H^*\mathfrak{L}^k} J_{\mathfrak{K}}(\mathfrak{a}') \tag{4.1}$$

where we define the functional by

$$J_{\mathfrak{K}}(\mathfrak{a}') = \frac{1}{2}(\mathfrak{K}\mathfrak{d}^*\mathfrak{a}',\mathfrak{d}^*\mathfrak{a}') + \frac{1}{2}(\mathfrak{K}^*\mathfrak{d}\mathfrak{a}',\mathfrak{d}\mathfrak{a}') - (\mathfrak{f},\mathfrak{a}') \tag{4.2}$$

The material coefficients $\mathfrak{K}$ are spatially variable mappings from $\Lambda^{k-(n-d(i))}\left(\Omega_i^d\right)$ onto itself, defined independently for all terms in the inner product (3.6). In particular, with reference to section 2, $\mathfrak{K}$ contains all instances of the proportionality constants K appearing in (2.9), (2.10) and (2.12).

For equation (4.1) to be well-posed and have a unique solution, we need $(\mathfrak{K}\mathfrak{d}^*\mathfrak{a}',\mathfrak{d}^*\mathfrak{a}') + (\mathfrak{K}^*\mathfrak{d}\mathfrak{a}',\mathfrak{d}\mathfrak{a}')$ to be continuous and coercive, i.e. we need to impose constraints on $\mathfrak{K}$ and $\mathfrak{K}^*$. Indeed, by reverting to the definition of the inner product, we define the ellipticity constant $\alpha_{\mathfrak{K}}$ as the minimum eigenvalue of $\mathfrak{K}$, and similarly for $\alpha_{\mathfrak{K}^*}$. We require both these constants to be bounded above zero, such that

$$(\mathfrak{K}\mathfrak{d}^*\mathfrak{a}',\mathfrak{d}^*\mathfrak{a}') + (\mathfrak{K}^*\mathfrak{d}\mathfrak{a}',\mathfrak{d}\mathfrak{a}') \ge \min(\alpha_{\mathfrak{K}},\alpha_{\mathfrak{K}^*})\,(1+C_\Omega)^2\|\mathfrak{a}'\|^2$$

The minimum of equation (4.1) must satisfy the Euler-Lagrange equations, thus $\mathfrak{a} \in \mathring{H}\Lambda^k \cap H^*\Lambda^k$ satisfies

$$(\mathfrak{K}\mathfrak{d}^*\mathfrak{a}, \mathfrak{d}^*\mathfrak{a}') + (\mathfrak{K}^*\mathfrak{d}\mathfrak{a}, \mathfrak{d}\mathfrak{a}') = (\mathfrak{f}, \mathfrak{a}') \qquad \text{for all } \mathfrak{a}' \in \mathring{H}\Lambda^k \cap H^*\Lambda^k \tag{4.3}$$

From the perspective of applications, and mirroring the distinctions between conservation laws and constitutive laws discussed in Section 2, we will be interested in the mixed formulation of equation (4.3) obtained by introducing the variable $\mathfrak{b} = \mathfrak{K}\mathfrak{d}^*\mathfrak{a}$, where $\mathfrak{b}$ is the generalization of the various fluxes $\boldsymbol{u}$. Then we may either consider a constrained minimization problem derived from equation (4.1), or for the sake of brevity, proceed directly to the Euler-Lagrange formulation: Find $(\mathfrak{a}, \mathfrak{b}) \in H\Lambda^k \times H\Lambda^{k-1}$ which satisfy

$$(\mathfrak{K}^{-1}\mathfrak{b}, \mathfrak{b}') - (\mathfrak{a}, \mathfrak{d}\mathfrak{b}') = 0 \qquad \text{for all } \mathfrak{b}' \in H\Lambda^{k-1} \tag{4.4}$$

$$(\mathfrak{d}\mathfrak{b}, \mathfrak{a}') + (\mathfrak{K}^*\mathfrak{d}\mathfrak{a}, \mathfrak{d}\mathfrak{a}') = (\mathfrak{f}, \mathfrak{a}') \qquad \text{for all } \mathfrak{a}' \in H\Lambda^k \tag{4.5}$$

The saddle-point formulation is well-posed subject to Babuška-Aziz inf-sup condition. Due to the presence of a Helmholtz decomposition, this follows by standard arguments. From equations (4.4) and (4.5) we deduce the strong form of the Hodge Laplacian on mixed form, corresponding to the equations

$$\mathfrak{b} = \mathfrak{K}\mathfrak{d}^*\mathfrak{a} \qquad \text{and} \qquad \mathfrak{d}\mathfrak{b} + \mathfrak{d}^*(\mathfrak{K}^*\mathfrak{d}\mathfrak{a}) = \mathfrak{f} \tag{4.6}$$

Of the various formulations, equations (4.4) and (4.5) are particularly appealing from the perspective of practical computations, as they do not require the coderivative.

An important remark is that the relative simplicity of the well-posedness analysis for the mixed-dimensional equations relies on the definition of the function spaces and norms. In particular, due to the definition of $H\Lambda^k$ via the mixed-dimensional differential $\mathfrak{d}$, the norm on the function space is inherently also mixed-dimensional, and cannot simply be decomposed into, say norms on the function spaces $H\Lambda^{k-(n-d)}(\Omega_i^d)$. For this reason, analysis in terms of "local norms" becomes significantly more involved [17, 18, 11].

5. Finite-dimensional spaces

In order to exploit the mixed-dimensional formulations from the preceding section, and in particular equations (4.4-4.5) we wish to consider finite-dimensional subspaces of $H\Lambda^k$. These spaces should be constructed to inherit the de Rham structure of equation (3.5), and with bounded projection operators. A natural approach is to consider the polynomial finite element spaces as a starting point [15].

From the finite element exterior calculus (FEEC - [15]), we know that on the highest-dimensional domains Ω_i^d, we may choose any of the finite element de Rham sequences, and in particular, we may consider the standard spaces from applications for a simplicial tessellation $\mathcal{T}_i^n = \mathcal{T}(\Omega_i^n)$

$$\mathcal{P}_r\Lambda^k(\mathcal{T}_i^n) \qquad \text{and} \qquad \mathcal{P}_r^-\Lambda^k(\mathcal{T}_i^n) \tag{5.1}$$

These correspond to the full and reduced polynomial spaces of order r, respectively, in the sense of [15]. In order to build a finite element de Rham sequence, we recall that (while still commuting with bounded projection operators) the full polynomial spaces reduce order

$$\mathcal{P}_r\Lambda^k(\mathcal{T}^n) \xrightarrow{d} \mathcal{P}_{r-1}\Lambda^{k+1}(\mathcal{T}^n) \tag{5.2}$$

while the reduced spaces preserve order

$$\mathcal{P}_r^-\Lambda^k(\mathcal{T}^n) \xrightarrow{d} \mathcal{P}_r^-\Lambda^{k+1}(\mathcal{T}^n) \tag{5.3}$$

Thus, any of these combinations of spaces are acceptable for Ω_i^n, and consider therefore the choice as given, and denoted by $\Lambda_h^{k,n}$ and $\Lambda_h^{k+1,n}$.

For $d < n$, we must consider not only the continuous differential operator d, but also the discrete jump operator $\mathbb{d}$. It is therefore clear that for i.e. $d = n - 1$, we must consider the traces of the finite element spaces of higher dimensions. In particular, we require for all pairs of dimensions $0 \leq e < d \leq n$,

$$\mathrm{Tr}_{\Omega_i^e}\, \Lambda_h^k\big(\mathcal{T}^d\big) \subseteq \Lambda_h^{k+(n-e)}(\mathcal{T}^e) \tag{5.4}$$

In contrast to the continuous differential order, the discrete differential operator preserves order for both the full and reduced spaces, since [15]:

$$\mathrm{Tr}_{\Omega^e}\, \mathcal{P}_r\Lambda^k\big(\mathcal{T}^d\big) = \mathcal{P}_r\Lambda^{k+(n-e)}(\mathcal{T}^e) \qquad \text{and} \qquad \mathrm{Tr}_{\Omega^e}\, \mathcal{P}_r^-\Lambda^k\big(\mathcal{T}^d\big) = \mathcal{P}_r^-\Lambda^{k+(n-e)}(\mathcal{T}^e) \tag{5.5}$$

We now define the polynomial subspaces $\mathcal{P}_{\mathfrak{r}}^{\mathfrak{m}}\mathfrak{L}^k \in H\mathfrak{L}^k$ as

$$\big(\mathcal{P}_{\mathfrak{r}}^{\mathfrak{m}}\mathfrak{L}^k\big)_i^d = \mathcal{P}_{r_i^d}^{p_i^d}\Lambda^{k-(n-d)}\big(\mathcal{T}_i^d\big) \tag{5.6}$$

where the multi-indexes $\mathfrak{r}$ and $\mathfrak{m}$ have values $r_i^d \in \mathbb{P}$ and $m_i^d \in [\,, -]$, respectively. When the multi-indexes are chosen to satisfy both (5.2-5.3) as well as (5.4), we obtain the discrete de Rham complex

$$0 \to \mathbb{R} \hookrightarrow \mathcal{P}_{\mathfrak{r}}^{\mathfrak{m}}\mathfrak{L}^0 \xrightarrow{\mathfrak{d}} \mathcal{P}_{\mathfrak{r}}^{\mathfrak{m}}\mathfrak{L}^1 \xrightarrow{\mathfrak{d}} \dots \xrightarrow{\mathfrak{d}} \mathcal{P}_{\mathfrak{r}}^{\mathfrak{m}}\mathfrak{L}^n \to 0 \tag{5.7}$$

Due to the existence of stable projections for all finite element spaces in $\mathcal{P}_{\mathfrak{r}}^{\mathfrak{m}}\mathfrak{L}^k$, the discrete de Rham sequence can be shown to be exact, thus equations (4.4) and (4.5) have stable approximations.

The discrete spaces for $H^*\mathfrak{L}^k$ must satisfy similar properties. Equations (5.2-5.3) hold in the dual sense, i.e. we write $\mathcal{P}_r^*\Lambda^k\big(\mathcal{T}_i^d\big) = \mathcal{P}_r^*\Lambda^k\big(\mathcal{T}_i^d\big) =\star\, (\mathcal{P}_r\Lambda^{d-k}\big(\mathcal{T}_i^d\big)$, and $d^*\mathcal{P}_r^*\Lambda^k\big(\mathcal{T}_i^d\big) \subset \mathcal{P}_r^{-*}\Lambda^{k-1}\big(\mathcal{T}_i^d\big) \subset \mathcal{P}_{r-1}^*\Lambda^{k-1}\big(\mathcal{T}_i^d\big)$. Furthermore, the coderivative $\mathfrak{d}^*$ imposes the inverted condition $\Lambda_h^{k+(n-e)}(\mathcal{T}^e) \subseteq \mathrm{Tr}_{\Omega_i^{n-1}}^*\, \Lambda_h^k\big(\mathcal{T}^d\big)$ on boundaries.

6. Implications in terms of classical calculus

We take a moment to untangle the notation from Sections 3-5 in order to extract insight into modeling and discretization for the original physical problem.

Our initial task is to express simplest form of the mixed-dimensional Hodge Laplacian in terms of conventional notation. We limit the discussion to the case where $k = n$, the function spaces $H^*\mathfrak{L}^n$ and $H\mathfrak{L}^{n-1}$ correspond to H_1 scalars and $H(div)$ vectors on each dimension $d \geq 1$. For $d = 0$, only the scalars are defined. Furthermore, the term $\mathfrak{d}\mathfrak{a} \in \mathfrak{L}^{n+1} = \emptyset$, and thus we arrive from (4.6) to the simpler problem

$$\mathfrak{b} = \mathfrak{K}\mathfrak{d}^*\mathfrak{a} \qquad \text{and} \qquad \mathfrak{d}\mathfrak{b} = \mathfrak{f} \tag{6.1}$$

In this case, the exterior derivative is the negative divergence plus jumps for each domain, while the codifferential is the gradient parallel to each domain, and the difference from boundaries perpendicular. As such, we arrive exactly at the model equations of Section 2, with the second choice of modeling assumption (2.12).

Turning our attention to the finite element spaces, the lowest order spaces for discretizing (4.4-4.5) are the reduced spaces obtained by choosing $r_i^d = 1$ and $m_i^d = -$, from which we obtain piecewise constants for $\mathfrak{a}$ on all domains, while we obtain for $\mathfrak{b}$ the Nedelec 1st kind (div) – Raviart-Thomas – continuous Lagrange elements for domains with dimensions $d = 3,2,1$, respectively – all of the lowest order [12] (this method will be referred to as "Mixed Reduced" in the next section). Interestingly, if we choose Nedelec 2nd kind (div) elements of lowest order for $d = 3$, equations (5.2) and (5.5) implies that we should increase the order in the lower-dimensional domains, obtaining dG elements of order $n - d$ for pressure, with BDM (2nd order) – continuous Lagrange (3rd order) for fluxes in domains with $d = 2,1$. This is a new method resulting from the analysis herein. We refer to this method as "Mixed Full".

The mixed finite element discretization has the advantage of a strong conservation principle, and may be hybridized to obtain a cheaper numerical scheme (see [12] for a direct approach in this context, but also [6, 5] for direct constructions in the finite volume setting). Alternatively, we consider discretizing the Euler-variation of the unconstrained minimization problem, equations (4.3). The natural finite element spaces are $\mathcal{P}_{\mathfrak{r}}^{\mathfrak{m},*}\Omega^n$, with $r_i^d = 1$ and $\mathfrak{m}$ does come into play, corresponding to 1st-order continuous Lagrange elements in all dimensions. From an engineering perspective, a similar formulation has been described in [19], we refer to this method as "Primal" in the next section.

7. Computational example

In order to illustrate the concepts discussed in the preceding sections, we will continue to consider $k = n$, and thus fractured porous media as a computational example, using the three numerical methods obtained using the lowest-order elements of the families described in the previous section.

The example consists of the unit square with two fractures crossing through the domain, intersecting at a right angle, as illustrated in figures 3. We impose unit permeability in the surroundings, set the normal and tangential permeability of the fractures to 100 and assume the apertures of both fractures as $\epsilon = 10^{-3}$. The boundary conditions are chosen as zero pressure at the bottom and no-flux conditions on the sides. Moreover, a boundary pressure of one is imposed on the fracture crossing the top boundary. All computations were performed with the use of FEniCS [21].

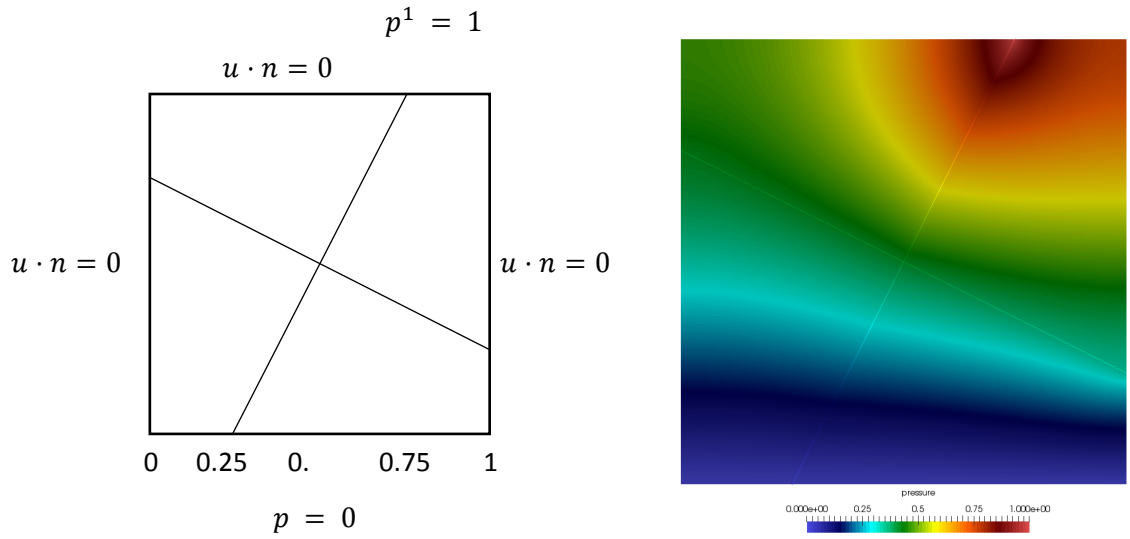


Figure 3: (Left) Domain of computation and associated boundary conditions. The pressure boundary condition is only imposed on the fracture pressure. (Right) Example of calculated solution (pressure).

The results show that all three methods are stable and convergent (Table 1). The relative errors and L^2-convergence rates after four consecutive refinements (identified by the characteristic grid size h) are given in the following table. Here, we compare the results to a fine-scale solution, obtained after a fifth refinement. In this example, all grids are matching.

		Primal		Mixed Reduced				Mixed Full			
Domain	Grid size	Pressure		Pressure		Flux		Pressure		Flux	
	h	Error	Rate	Error	Rate	Error	Rate	Error	Rate	Error	Rate
Ω^0	2^{-4}	2.66e-03	1.53	2.21e-03	1.52	N/A	N/A	2.89e-04	1.61	N/A	N/A
	2^{-5}	8.45e-04	1.65	7.18e-04	1.62			8.99e-05	1.69		
	2^{-6}	2.15e-04	1.97	1.87e-04	1.94			2.26e-05	1.99		
Ω^1	2^{-4}	2.54e-03	1.46	1.89e-02	1.01	6.32e-03	1.22	3.01e-04	1.71	1.84e-03	1.28
	2^{-5}	9.57e-04	1.41	9.22e-03	1.04	2.49e-03	1.34	8.99e-05	1.74	7.44e-04	1.30
	2^{-6}	3.23e-04	1.57	4.12e-03	1.16	7.82e-04	1.67	2.37e-05	1.92	2.61e-04	1.51
Ω^2	2^{-4}	4.25e-03	1.53	1.89e-02	1.02	8.21e-02	0.74	1.86e-02	1.01	3.16e-02	0.75
	2^{-5}	1.36e-03	1.64	9.17e-03	1.05	4.75e-02	0.79	9.11e-03	1.03	1.87e-02	0.75
	2^{-6}	3.60e-04	1.92	4.08e-03	1.17	2.47e-02	0.94	4.07e-03	1.16	1.04e-02	0.86

Table 1: Convergence rates for the three FE and MFEM discussed for the fracture problem in Section 6. With reference to Figure 3, the domain Ω^0 is the intersection point, Ω^1 represents the four fracture segments, while Ω^2 is the remaining ambient geometry.

Each method captures the intersection pressure well, with second order convergence over all. In the surroundings, the pressure convergence with second order for the primal formulation and first order for both mixed formulations, as expected. The Mixed Full method has higher-order elements in the fracture, and this is reflected in higher convergence rates for both pressure and flux.

Acknowledgments

The authors wish to thank Gunnar Fløystad, Eirik Keilegavlen, Jon Eivind Vatne and Ivan Yotov for valuable comments and discussions on this topic. The authors also with to thank the two anonymous reviewers who provided helpful comments on the initial version of this manuscript. This research is funded in part by the Norwegian Research Council grants: 233736 and 250223.

References

[1] J. M. Nordbotten and M. A. Celia, Geological Storage of CO2: Modeling Approaches for Large-Scale Simulation, Hoboken, N. J.: Wiley, 2012.

[2] J. Bear, Hydraulics of Groundwater, McGraw-Hill, 1979.

[3] P. G. Ciarlet, Mathematical Elasticity Volume II: Theory of Plates, Amsterdam: Elsevier, 1997.

[4] C. Alboin, J. Jaffré, J. E. Roberts and C. Serres, "Domain decomposition for flow in porous media with fractures," in *14th conference on Domain Decomposition Methods in Sciences and Engineering*, Cocoyoc, Mexico, 1999.

[5] T. H. Sandve, I. Berre and J. M. Nordbotten, "An efficient multi-point flux approximation method for discrete fracture-matrix simulations," *Journal of Computational Physics,* vol. 231, pp. 3784-3800, 2012.

[6] M. Karimi-Fard, L. J. Durlofsky and K. Aziz, "An effcient discrete-fracture model applicable for general-purpose reservoir simulations," *SPE Journal,* pp. 227-236, 2004.

[7] V. Martin, J. Jaffré and J. E. Roberts, "Modeling fractures and barriers as interfaces for flow in porous media," *SIAM Journal of Scientiffic Computing,* vol. 26, pp. 1557-1691, 2005.

[8] M. W. Licht, "Complexes of discrete distributional differential forms and their homology theory," *Foundations of Computational Mathematics,* 2016.

[9] Y. C. Yortsos, "A theoretical analysis of vertical flow equilibrium," *Transport in Porous Media,* vol. 18, pp. 107-129, 1995.

[10] N. Schwenk, B. Flemisch, R. Helmig and B. I. Wohlmuth, "Dimensionally reduced flow models in fractured porous media," *Computational Geosciences,* vol. 16, pp. 277-296, 2012.

[11] L. Formaggia, A. Fumagalli, A. Scotti and P. Ruffo, "A Reduced Model for Darcy's Problem in Networks of Fractures," *ESAIM: Mathematical Modelling and Numerical Analysis,* vol. 48, pp. 1089-1116, 2014.

[12] W. M. Boon, J. M. Nordbotten and I. Yotov, "Robust discretization of flow in fractured porous media," in preparation.

[13] W. M. Boon, J. M. Nordbotten and J. E. Vatne, "Exterior calculus for mixed-dimensional partial differential equations," arXiv:1710.00556.

[14] M. Spivak, Calculus on Manifolds, Reading, Massachusetts: Addison-Wesley Publishing Company, 1965.

[15] D. N. Arnold, R. S. Falk and R. Winther, "Finite element exterior calculus, homological techniques, and applications," *Acta Numerica,* vol. 15, pp. 1-155, 2006.

[16] R. Hiptmair, "Finite elements in computational electromagnetism," *Acta Numerica,* vol. 11, pp. 237-339, 2002.

[17] C. D'Angelo and A. Scotti, "A Mixed Finite Element Method for Darcy Flow in Fractured Porous Media with Non-Matching Grids," *ESAIM: Mathematical Modelling and Numerical Analysis,* pp. 465-489, 2012.

[18] N. Frih, V. Martin, J. E. Roberts and A. Saada, "Modeling Fractures as Interfaces with Nonmatching Grids," *Computational Geosciences,* vol. 16, pp. 1043-10060, 2012.

[19] R. Helmig, C. Braun and M. Emmert, "MUFTE: A Numerical Model for Simulation of Multiphase Flow Processes in Porous and Fractured Porous Media," Universität Stuttgart, 1994.

[20] X. Tunc, F. I., T. Gallouët, M. C. Cacas and P. Havé, "A model for conductive faults with non-matching grids," *Computational Geosciences,* vol. 16, pp. 277-296, 2012.

[21] A. Logg, K.-A. Mardal, G. N. Wells and e. al, Automated Solution of Differential Equations by the Finite Element Method, Springer, 2012.

[22] X. Claeys and R. Hiptmair, "Integral equations on multi-screens," *Integral Equations and Operator Theory,* 2013.

Balancing Domain Decomposition by Constraints algorithms for curl-conforming spaces of arbitrary order

Stefano Zampini, Panayot Vassilevski, Veselin Dobrev and Tzanio Kolev

Abstract We construct Balancing Domain Decomposition by Constraints methods for the linear systems arising from arbitrary order, finite element discretizations of the $H(\mathrm{curl})$ model problem in three-dimensions. Numerical results confirm that the proposed algorithm is quasi-optimal in the coarse-to-fine mesh ratio, and poly-logarithmic in the polynomial order of the curl-conforming discretization space. Additional numerical experiments, including higher-order geometries, upscaled finite elements, and adaptive coarse spaces, prove the robustness of our algorithm. A scalable three-level extension is presented, and it is validated with large scale experiments using up to 16384 subdomains and almost a billion of degrees of freedom.

1 Introduction

We construct Balancing Domain Decomposition by Constraints (BDDC) methods [8] for the linear systems arising from three-dimensional, arbitrary order finite element discretizations of the $H(\mathrm{curl})$ bilinear form

$$\int_{\Omega} \alpha\, \nabla \times \boldsymbol{u} \cdot \nabla \times \boldsymbol{v} + \beta\, \boldsymbol{u} \cdot \boldsymbol{v}\, dx, \quad \alpha \geq 0,\ \beta > 0. \tag{1}$$

The proposed algorithm is quasi-optimal in the coarse-to-fine mesh ratio, and poly-logarithmic in the polynomial order of the finite element discretization space, which is confirmed by the numerical results in Section 3. Our results will be equally valid

Panayot Vassilevski, Veselin Dobrev and Tzanio Kolev
Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, P.O. Box 808, L-561, Livermore, CA 94551, U.S.A e-mail: {panayot,veselin,tzanio}@llnl.gov

Stefano Zampini
King Abdullah University of Science and Technology, Computer, Electrical and Mathematical Sciences and Engineering Division, Extreme Computing Research Center, e-mail: stefano.zampini@kaust.edu.sa

for the Finite Element Tearing and Interconnecting Dual-Primal (FETI-DP) method [12], due to the well known duality between BDDC and FETI-DP [25].

The bilinear form (1) originates from implicit time-stepping schemes of the quasi-static approximation of the Maxwell's equations in the time domain [30]. The coefficient α is the reciprocal of the magnetic permeability, assumed constant, whereas β is proportional to the conductivity of the medium; positive definite anisotropic tensor conductivities can be handled as well. We only present results for essential boundary conditions, but the generalization of the algorithms to natural boundary conditions is straightforward. Magnetostatic problems with $\beta = 0$ are not covered in the present work, and they can be the subject of future research. The same bilinear form appears in block preconditioning techniques for the frequency domain case [13], mixed form of Brinkman (Darcy-Stokes) [39], and in incompressible magneto-hydrodynamics [24].

The operator $\nabla\times$ is the *curl* operator, defined, e.g., in [4]; the vector fields belong to $H(\mathrm{curl})$, the Sobolev space of square-integrable vector fields having a square-integrable curl. The space $H(\mathrm{curl})$ is often discretized using Nédélec elements [26]; those of lowest order use polynomials with continuous tangential components along the edges of the elements. While most existing finite element codes for electromagnetics use lowest order elements, those of higher order have shown to require fewer degrees of freedom (dofs) for a fixed accuracy [31, 13].

The design of solvers for edge-element approximations of (1) poses significant difficulties, since the kernel of the curl operator is non-trivial. An even greater challenge for domain decomposition solvers consists in finding logarithmically stable decompositions in three dimensions, due to the strong coupling that exists between the dofs located on the subdomain edges and those lying on the subdomain faces. Among non-overlapping methods, it is worth citing the wirebasket algorithms [9, 18, 19], Neumann-Neumann [32], and one-level FETI [36, 28]. Overlapping Schwarz methods have also been studied [33, 6].

The edge-element approximation of (1) has also received a lot of attention by the multigrid community; Algebraic Multigrid (AMG) methods have been proposed in [29], [5], and [17]. For geometric multigrid, see [14]. Robust and efficient multigrid solvers can be obtained combining AMG and auxiliary space techniques, that require some extra information on the mesh connectivity and on the dofs [15, 16, 22].

In this work, we follow the approach proposed by Toselli for three-dimensional FETI-DP with the lowest order Nédélec elements [34], where a stable decomposition is obtained by using a change of basis for the dofs located on the subdomain edges. The same approach has been pursued recently by Dohrmann and Widlund [11], who were able to improve Toselli's results, and obtain sharp and quasi-optimal condition number bounds (in the lowest order case) by using the deluxe variant of BDDC [10]. This is critical for obtaining iteration counts and condition number estimates independent of the jumps of α and β aligned with the subdomain interface. Finally, it has to be noted that BDDC deluxe algorithms for high-order Nédélec elements in two dimensions, and for the lowest order Nédélec elements in three dimensions have been already presented by the first author in [40, 42].

In Section 2, we complement the results in [34, 11] by proposing an algorithm for the change of basis that does not make any assumption on the mesh, the associated discretization space, and the domain decomposition. Inspired by the success of the auxiliary space technique [15], we construct the change of basis by using the so-called *discrete gradient*, a linear operator that maps gradients of scalar functions to their representation in the curl-conforming discretization space. Numerical experiments, provided in Section 3, confirm that the robustness of our approach is not confined to the more standard Nédélec elements, but it also extends to the case of elements with curved boundaries, and to upscaled $H(\mathrm{curl})$ spaces constructed by preserving the de Rham sequence on agglomerations of fine scale elements. Due to page restrictions, we refer the interested reader to [23] for a thorough description of these kind of elements.

2 Design of the algorithm

2.1 Domain decomposition and discrete spaces

We follow the framework of iterative substructuring [37, Chapters 4-6], and we decompose the domain Ω into N non-overlapping open Lipschitz subdomains Ω_i,

$$\overline{\Omega} = \bigcup_{i=1}^{N} \overline{\Omega}_i, \quad \Gamma := \bigcup_{i \neq j} \partial\Omega_j \cap \partial\Omega_i,$$

with Γ the interface between the subdomains. We further assume that Ω and each Ω_i are simply connected (does not contain any holes). We denote by $\mathbf{V}_h(\Omega)$ and $S_h(\Omega)$ the curl- and H^1- conforming finite element spaces of polynomial order p, respectively, together with their subdomain counterparts $\mathbf{V}_h^{(i)} := \mathbf{V}_h(\Omega_i)$ and $S_h^{(i)} := S_h(\Omega_i)$. We denote by $\mathbf{W}$ the global finite element space in which we seek the solution of problems coming from the bilinear form (1), and by $\mathbf{W}^{(i)}$ the corresponding subdomain spaces. We note that $\mathbf{V}_h$ coincides with $\mathbf{W}$ when using Nédélec elements; however, our algorithm covers also the case $\mathbf{V}_h \subset \mathbf{W}$, as it is the case of upscaled finite elements that preserve the de Rham sequence [23], or of three-level extensions of the BDDC algorithm for (1) (see Section 2.4).

The success of the algorithm depends on the analysis of the interface, that leads to the detection of equivalence classes such as the subdomain faces, i.e. sets of connected dofs shared by the same two subdomains, and the subdomain edges, i.e. sets of connected dofs shared by 3 or more subdomains. We assume that a subdivision of Γ in face and edge disjoint subsets has been found; moreover, we assume that each subdomain edge has exactly two endpoints, and none of the edge endpoints lie in the interior of another subdomain edge. As noted in [11, Section 5], this guarantees that the change of basis (defined in the next section) leads to a new well-posed problem.

2.2 BDDC method

The recipe for the construction of a BDDC preconditioner consists in the design of a partially continuous interface space $\widetilde{\mathbf{W}}_\Gamma$, the direct sum of a continuous *primal* space $\mathbf{W}_\Pi$ and a discontinuous *dual* space $\mathbf{W}_\Delta$, and in the choice of an averaging operator E_D for the partially continuous dofs, which drives the analysis and the design of robust primal spaces [25].

Following Toselli [34], we characterize the primal space $\mathbf{W}_\Pi$ by using two primal constraints per subdomain edge E as given by

$$s_{0,E}(\boldsymbol{w}) := \tfrac{1}{|E|}\int_E \ \boldsymbol{w}\cdot\boldsymbol{t}_E \ ds, \quad \boldsymbol{w}\in\mathbf{V}_h, \tag{2}$$

$$s_{1,E}(\boldsymbol{w}) := \tfrac{1}{|E|}\int_E s\boldsymbol{w}\cdot\boldsymbol{t}_E \ ds, \quad \boldsymbol{w}\in\mathbf{V}_h, \tag{3}$$

where $\boldsymbol{t}_{E_{|e}} := \boldsymbol{t}_e$, with $\boldsymbol{t}_e$ the vector oriented in the direction of a fine mesh edge e belonging to E. For implementation details of the primal space, see Remark 3.

2.3 Change of basis

As in [34, 11], we consider a change of basis for the dofs of $\mathbf{V}_h$ that are located on each subdomain edge E, and we split a finite element function $\boldsymbol{w}$ into a *constant* component $\boldsymbol{\Phi}_E$ and *gradient* components $\nabla\phi_{jE}$ associated with the nodal dofs of S_h lying in the interior of the edge, i.e.

$$\boldsymbol{w}|_E = s_{0,E}(\boldsymbol{w})\boldsymbol{\Phi}_E + \sum_{j=1}^{n_E-1} w_{jE}(\boldsymbol{w})\nabla\phi_{jE} + \boldsymbol{w}_{Ec},$$

with n_E the number of $\mathbf{V}_h$ dofs on E, and $\boldsymbol{w}_{Ec}$ the finite element function (if any) identified by the dofs of $\mathbf{W}$ that lie on E and are not in $\mathbf{V}_h$.

The change of basis in BDDC methods is performed by projection as T^TAT, where the columns of T represents the new basis in terms of the old dofs [21], and A results from the discretization of the bilinear form (1). The structure of T for three-dimensional curl-conforming spaces is as follows [34, 11, 40, 42]

$$T = \begin{bmatrix} I_C & 0 & 0 & \dots & 0 \\ 0 & I_F & T_{FE_1} & \dots & T_{FE_n} \\ 0 & 0 & T_{E_1E_1} & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & T_{E_nE_n} \end{bmatrix},$$

where I_C and I_F are identity matrices of appropriate sizes. Here, F denotes the set of dofs of $\mathbf{V}_h$ that belong to the subdomain faces, and C denotes all the remaining dofs of $\mathbf{W}$ that do not belong to F or to any of the subdomain edge dofs of $\mathbf{V}_h$.

Differently from a conventional change of basis in BDDC, the one used for curl-conforming spaces is not local to the subdomain edges, as it also involves, through the sparse off-diagonal blocks T_{FE_i}, the dofs of F that are located on the fine mesh edges sharing a mesh vertex with any of the E_i.

In our code, we use T^TAT as iteration matrix; however, in order to preserve a one-to-one correspondence between old and new subdomain dofs, we construct the preconditioner using the subdomain matrices $\tilde{A}^{(i)} := R^{(i)}T^TR^{(i)T}A^{(i)}R^{(i)}TR^{(i)T}$, where $A^{(i)}$ is the discretization matrix on $\mathbf{W}^{(i)}$, and $R^{(i)}$ is the usual restriction operator from $\mathbf{W}$ to $\mathbf{W}^{(i)}$. Note that $T^TAT = \sum_{i=1}^N (T^TR^{(i)T})A^{(i)}(R^{(i)}T) \neq \sum_{i=1}^N R^{(i)T}\tilde{A}^{(i)}R^{(i)}$.

The functions $\boldsymbol{\Phi}_E$ and $\nabla\phi_{jE}$ are explicitly constructed in [34, 11]; however, the procedures used therein possess strong limitations, as they need to access the underlying mesh and to understand how the edge dofs are related with the orientation of the fine mesh edges; moreover, they are limited to the lowest order Nédélec space only. In this work, we propose a construction of the change of basis by using the information contained in the discrete gradient operator G, the matrix representation of the mapping $\phi \in S_h \to \nabla\phi \in \mathbf{V}_h$, that is also used by the auxiliary space method, see [15] and [22, Section 4]. We note that, when using Nédélec elements, there are p dofs associated to each fine mesh edge, and that the number of nonzeros per row of G is $p+1$, with p the polynomial order of the finite element space used for S_h. For upscaled elements, the number p_e of dofs of $\mathbf{V}_h$ associated to each fine mesh edge e may vary from one fine mesh edge to another, but the number of nonzeros of the corresponding rows of G is always p_e+1.

For each subdomain edge E, we construct the corresponding column block of T as follows. We first extract the matrix $G_{E\mathring{E}}$, where $\mathring{E}$ is the set of dofs of S_h that is associated with those basis functions being nonzero on the nodes in the interior of E; note that $G_{E\mathring{E}}$ has full-column rank, and that $n_E = n_{\mathring{E}} + 1$. We then compute the representation of the subdomain edge constant function $\boldsymbol{\Phi}_E$ in $\mathbf{V}_h$ as the eigenvector corresponding to the nonzero eigenvalue of the orthogonal complement of $G_{E\mathring{E}}$, i.e. $I - G_{E\mathring{E}}(G_{E\mathring{E}}^TG_{E\mathring{E}})^{-1}G_{E\mathring{E}}^T$. The dofs defining $\nabla\phi_{jE}$ are simply given by the columns of G that correspond to $\mathring{E}$. The change of basis block relative to E is

$$\begin{bmatrix} T_{FE} \\ T_{EE} \end{bmatrix} := \begin{bmatrix} 0 & G_{E^c\mathring{E}} \\ \boldsymbol{\Phi}_E & G_{E\mathring{E}} \end{bmatrix},$$

with $E \cup E^c$ the set of row indices corresponding to the nonzero values in the $\mathring{E}$ columns.

Remark 1. The construction of our change of basis just needs sub-matrix extraction operations and the computation of the orthogonal complement of $G_{E\mathring{E}}$, which can be obtained by doing a singular value decomposition of the same matrix, of size $n_E \times n_{\mathring{E}}$: note that n_E is usually very small, on the order of ten, and we can thus efficiently use algorithms for dense matrix storages. After having changed the basis, the sparsity pattern of the local matrices is not spoiled, and optimal nested dissection orderings for the direct solves of the subdomain problems can be found.

Remark 2. For the lowest order Nédélec elements, G has two nonzeros per row; the values are $+1$ or -1 depending on the orientation of the element edge. When hexahedral meshes and box subdomains are considered, our change of basis is the same as that proposed by Toselli [34].

Remark 3. The constraint given in eq. (2) is obtained by selecting the dofs corresponding to each $\boldsymbol{\Phi}_E$ as primal; arithmetic averages for the remaining dofs on the subdomain edges are used to impose the constraint (3), see also [11, Section 2.2].

Remark 4. Our algorithm does not require the user to input the mesh connectivity. From G, we can infer the dofs connectivity which will lead to a well-posed change of basis, since the sparsity pattern of the matrices $G^T G$ and GG^T carry the information of a vertex-to-vertex, and an edge-to-edge mesh connectivity graph, respectively.

2.4 Three-level extension of the algorithm

Three-level extensions of the algorithm [38] are crucial for large scale simulations, as the solution of the coarse problem in BDDC (as with all two-level methods) can become a bottleneck when many subdomains are considered, see [41, Section 3.6] and the references therein for additional details. The minimal coarse space presented in Section 2.2 can be naturally split in two disjoint subsets; the one arising from the constraints given in eq. (2) resembles a lowest-order Nédélec space defined on the coarse element (i.e., the subdomain). The rest of the coarse dofs are instead generated by gradients of scalar functions, and a scalable coarse space can be thus obtained by considering arithmetic averages defined on the coarse subdomain edges.

We thus propose an approximate coarse discrete gradient to obtain a stable decomposition of the coarse dofs generated by eq. (2), obtained by projecting the fine discrete gradient G on the $\boldsymbol{\Phi}_E$ functions. The resulting coarse discrete gradient will have two nonzero entries per row, with entries given by $G^T_{E\partial E}\boldsymbol{\Phi}_E$, with ∂E the indices of the basis functions of S_h associated with the two endpoints of E. We then construct the primal space of the coarse problem as outlined in the previous sections. Numerical results confirm that such an approach provides an optimal coarse space for the second level of the BDDC operator, and leads to scalable three-level algorithms in terms of number of iterations. We note that multilevel extensions, with an arbitrary number of levels, can be obtained by recursion arguments.

3 Numerical results

Here we present numerical experiments that confirm the robustness of our algorithm; we test the quasi-optimality, the dependence on the polynomial order of the curl-conforming spaces, and the proposed three-levels extension. In addition, we test the case of elements with curved boundaries. We also provide results for adaptive

enrichment (see [41] and the references therein) of the minimal coarse space given by eqs. (2) and (3) in the presence of heterogeneous coefficients. As quality metrics, we consider the experimental condition number (denoted by κ) and the number of conjugate gradient iterations needed to reduce by eight orders of magnitude the initial residual norm, starting from zero initial guess and randomly distributed right-hand side. Unless otherwise stated, the primal space consists of two dofs per subdomain edge as described in Section 2.2, $\alpha = \beta = 1$, and $\Omega = [0,1]^3$.

All the numerical results have been obtained using the discretization packages MFEM [1] (for Nédélec elements and high-order geometries) and ParElag [2] (for upscaled finite elements) developed at Lawrence Livermore National Laboratory, and by using the BDDC implementation developed by the first author in the PETSc library [3, 41]. Irregular decompositions of tetrahedral (TET) or hexahedral (HEX) meshes obtained from the graph partitioner ParMETIS [20] are always considered; deluxe scaling is always used to accommodate for spurious eigenvalues of the preconditioned operator arising from possibly jagged subdomain interfaces [7].

In Figure 1 we report the results of a quasi-optimality test, performed by considering successive uniform refinements of a mesh decomposed in 40 subdomains, and by using Nédélec elements of order $p = 1$ (lowest-order) and $p = 2$. The domain decomposition is kept fixed, in order to fix the value of the maximum subdomain diameter H. The results show a $(1+\log H/h)^2$ dependence in all the cases considered.

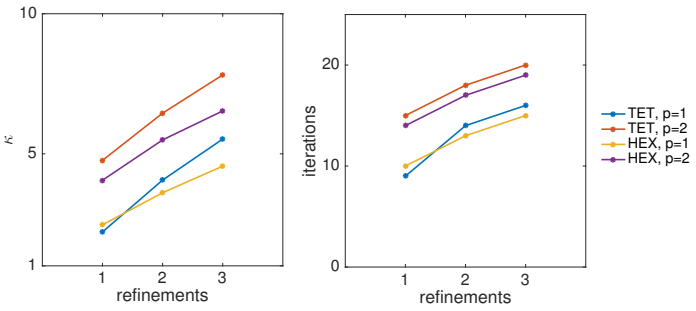


Fig. 1: Quasi-optimality test. κ (left) and number of iterations (right) for successive uniform refinements for Nédélec elements on hexahedra (HEX) and tetrahedra (TET); polynomial orders $p = 1$ and $p = 2$.

We then fix the mesh and the domain decomposition (i.e. H/h), and we increase the polynomial order of the discretization spaces. Figure 2 contains results for the Nédélec elements, going from $p = 1$ to $p = 6$; we note that we obtained the same results when considering statically condensed spaces (relevant when $p > 1$ for the HEX and $p > 2$ for the TET case, data not shown). In the same spirit, Figure 3 contains the results for upscaled curl-conforming elements, obtained by considering two successive levels of structured aggregation (UP1 and UP2 respectively), and with polynomial orders ranging from $p = 1$ to $p = 4$; results for Nédélec elements (NED) on the same mesh are given for comparison. In both cases, Nédélec or up-

scaled elements, our algorithm shows to be robust with the higher degree of the polynomial space, and it leads to a poly-logarithmic convergence rate. The results of this test, together with those related with the quasi-optimality, suggest a condition number bound of the type $(1+\log(p^2H/h))^2$ for the preconditioned operator.

Fig. 2: Polynomial order test. κ and number of iterations as a function of the polynomial order for Nédélec elements on hexahedra (HEX) and tetrahedra (TET).

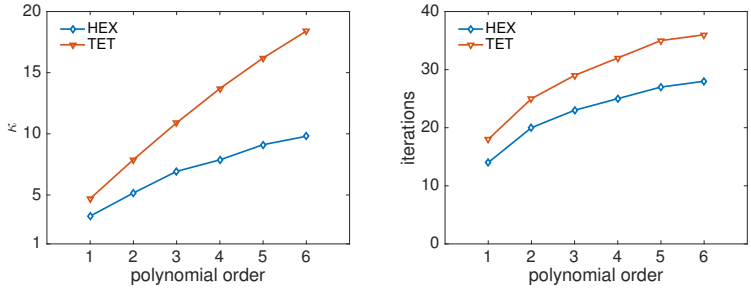
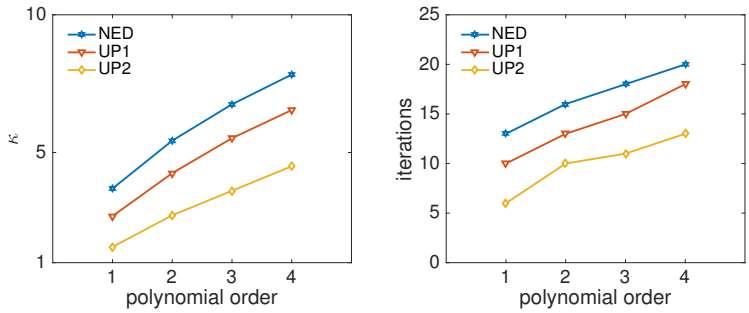


Fig. 3: Polynomial order test. κ and number of iterations as a function of the polynomial order for Nédélec elements (NED), and upscaled curl-conforming elements. UP1 one level of element aggregation with structured coarsening, UP2 two levels.



Further numerical evidence for the robustness of our approach is given by the results shown in Table 1, where condition numbers and number of iterations are reported by testing against third-order geometries, in combination with Nédélec elements of order $p = 1,2$. The meshes used to run the tests have been obtained from 2 levels of uniform refinements of those shown in Figure 4, and they are available with the MFEM source code as `escher-p3.mesh` and `fichera-q3.mesh`. The number of subdomains considered is 40.

Fig. 4: Third-order meshes used for the results in Table 1.

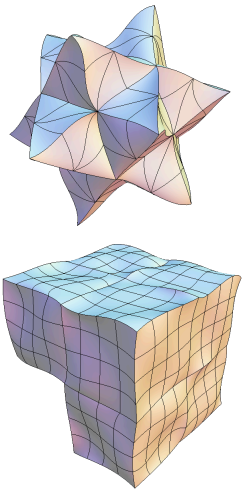


Table 1: High-order geometry test. Size of linear systems (dofs), condition number, (κ) and number of iterations (it) for Nédélec elements of degree 1 and 2 with the meshes shown in Figure 4.

TET, $p=1$, TET, $p=2$		
dofs	27K	144K
κ	11.8	23.7
it	26	37

HEX, $p=1$, TET, $p=2$		
dofs	12K	92K
κ	5.7	8.4
it	21	26

We next consider the case of heterogeneous coefficients; we fix $\alpha = 1$, and vary the distribution of β as pictured in Figure 5. For this test, we adaptively enrich the minimal coarse space by means of the adaptive selection of constraints algorithm described in [41, 40]; results have been obtained using either tetrahedral or hexahedral meshes, 40 subdomains, and with Nédélec elements of order $p = 1$ and $p = 2$. The number of dofs in the tetrahedral case is approximately 200 thousand (K) for $p = 1$, and 1.2 million (M) for $p = 2$; in the hexahedral case, the number of dofs are 330K and 3.5M, respectively. Results are reported in Table 2, together with the adaptive threshold used (λ), and the ratio between the number of generated coarse dofs and the number of interface dofs (C/Γ).

Fig. 5: Heterogeneous β distribution used for testing adaptive coarse spaces.

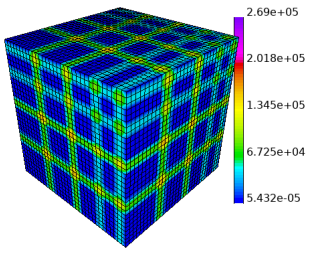


Table 2: Adaptive coarse spaces. Condition number (κ), number of iterations (it) and coarse-to-fine ratio (C/Γ) for different eigenvalue thresholds λ.

TET, $p=1$

λ	-	10	5	2.5
κ	150.2	7.5	4.6	2.2
it	54	15	12	8
C/Γ	0.01	0.05	0.06	0.09

TET, $p=2$

λ	-	10	5	2.5
κ	413.3	5.9	4.3	2.3
it	113	15	12	9
C/Γ	0.01	0.02	0.02	0.04

HEX, $p=1$

λ	-	10	5	2.5
κ	203.4	5.8	3.2	2.0
it	62	13	10	7
C/Γ	0.02	0.05	0.06	0.09

HEX, $p=2$

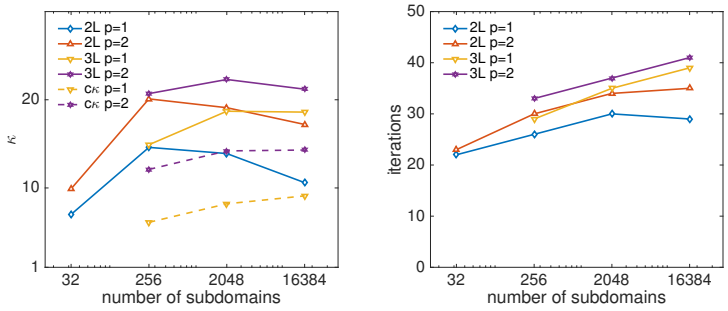
λ	-	10	5	2.5
κ	330.8	5.1	3.4	2.0
it	97	14	11	8
C/Γ	0.01	0.01	0.02	0.04

Without adaptive coarse spaces, the algorithm performs poorly (as expected) since the jumps in β are not aligned with the (irregular) subdomain boundaries;

on the other hand, the number of iterations and the condition numbers are consistently (and constantly) reduced when considering adaptive coarse spaces associated with smaller and smaller tolerances λ. The ratio of coarse-to-fine dofs remain bounded for all the tolerance values considered; interestingly, the coarsening procedure is more effective for $p = 2$ than for $p = 1$, as also observed experimentally with Raviart-Thomas vector fields [27, 43].

We close this section by reporting the results of a weak scalability test. Since we consider unstructured domain decompositions, we obtain subdomain problems of approximately the same size by using uniform refinements of an hexahedral mesh; at each level of refinement, we multiply by eight the number of subdomains used. As a consequence, we cannot guarantee that the shape of the subdomains remains the same. The total number of dofs in the test ranges from 186K to 94M with Nédélec elements of degree $p = 1$, and from 1.5M to 742M for $p = 2$. In Figure 6, we compare the results using a standard two-level BDDC algorithm (2L) and a three-level approach (3L), where the coarse subdomains have been obtained by aggregating 32 fine subdomains using ParMETIS; the condition number of the coarse BDDC preconditioned operator is also provided (κ_c, left panel, dashed lines). The number of iterations are scalable up to 16384 subdomains in both cases; condition numbers and number of iterations are slightly larger for the 3L case, but the algorithm preserves the convergence properties of the 2L case.

Fig. 6: Weak scalability test. κ and number of iterations as a function of the number of subdomains for two-level (2L) and three-level BDDC (3L). Coarse condition number (κ_c) is also shown.



4 Conclusions

We have constructed BDDC methods for arbitrary order, finite element discretizations of the $H(\mathrm{curl})$ model problem. Numerical results have shown that the proposed algorithm leads to a poly-logarithmic condition number bound, with a mild dependence on the polynomial order of the approximation space, of the type $(1+\log(p^2H/h))^2$. The robustness of our approach has been confirmed for various cases, including high-order geometries, upscaled curl-conforming finite elements,

and heterogeneous distributions of the coefficients. A scalable, three-level extension of the method has also been proposed; large scale parallel experiments using up to 16384 subdomains and almost a billion of dofs have been provided to validate the algorithm.

Acknowledgements This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory (LLNL) under Contract DE-AC52-07NA27344 and was supported by the U.S. DOE ASCR program. The research was performed during a visit of the first author to the LLNL, Center for Applied Scientific Computing. The authors are grateful to Umberto Villa for fruitful discussions. For computer time, this research used also the resources of the Supercomputing Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia.

References

1. MFEM: Modular finite element methods, http://mfem.org.
2. ParElag: Parallel element agglomeration algebraic multigrid upscaling and solvers, https://github.com/LLNL/parelag
3. Balay, S., et. al.: PETSc Users Manual, ANL-95/11 - Revision 3.7, Argonne National Lab (2016)
4. Boffi, D., Brezzi, F., Fortin, M.: Mixed finite element methods and applications. Springer, Heidelberg (2013)
5. Bochev, P. B., Garasi, C. J., Hu, J. J., Robinson, A. C., Tuminaro, R. S.: An improved algebraic multigrid method for solving Maxwell's equations. SIAM J. Sci. Comput., **25**, 623–642 (2003)
6. Calvo, J.: A two-level overlapping Schwarz algorithm for H(curl) in two dimensions with irregular subdomains. Electron. Trans. Numer. Anal. **44**, 497–521 (2015).
7. Calvo, J.: A BDDC algorithm with deluxe scaling for H(curl) in two dimensions with irregular subdomains. Math. Comp. **85**, 1085–1111 (2016)
8. Dohrmann, C. R.: A preconditioner for substructuring based on constrained energy minimization. SIAM J. Sci. Comput. **25**, 246–258 (2003)
9. Dohrmann, C. R.. Widlund, O. B.: An iterative substructuring algorithm for two-dimensional problems in $H(\mathrm{curl})$. SIAM J. Numer. Anal. **50**, 1004–1028 (2012)
10. Dohrmann, C. R., Widlund, O. B.: Some recent tools and a BDDC algorithm for 3D problems in $H(\mathrm{curl})$. Domain Decomposition Methods in Science and Engineering XX, Lect. Notes Comput. Sci. Eng. **91**, 15–25 (2013)
11. Dohrmann, C. R., Widlund, O. B.: A BDDC Algorithm with Deluxe Scaling for Three-Dimensional H(curl) Problems. Comm. Pure Appl. Math. **69**, 745–770 (2016)
12. Farhat, C., Lesoinne, M., LeTallec, P. , Pierson, K., Rixen, D.: FETI-DP: a dual-primal unified FETI method. I. A faster alternative to the two-level FETI method. Internat. J. Numer. Methods Engrg. **50**, 1523–1544 (2001)
13. Grayver, A. V., Kolev, T. V.: Large-scale 3D geoelectromagnetic modeling using parallel adaptive high-order finite element method. Geophysics **80**, E277–E291 (2015)
14. Hiptmair, R.: Multigrid method for Maxwell's equations. SIAM J. Numer. Anal. **36**, 204–225 (1999)
15. Hiptmair, R., Widmer, G., Zou, J.: Auxiliary space preconditioning in $H_0(\mathrm{curl};\ \Omega)$. Numer. Math. **103**, 435–459 (2006)
16. Hiptmair, R., Xu, J.: Nodal auxiliary space preconditioning in H(curl) and H(div) spaces. SIAM J. Numer. Anal. **45**, 2483–2509 (2007)

17. Hu, J. J., Tuminaro, R. S., Bochev, P. B., Garasi, C. J., Robinson, A. C.: Toward an h-independent algebraic multigrid method for Maxwell's equations. SIAM J. Sci. Comput. **27**, 1669–1688 (2006)
18. Hu, Q., Zou, J.: A nonoverlapping domain decomposition method for Maxwell's equations in three dimensions. SIAM J. Numer. Anal. **41**, 1682–1708 (2003)
19. Hu, Q., Shu, S., Zou, J.: A substructuring preconditioner for three-dimensional Maxwell's equations, Domain decomposition methods in science and engineering XX, Lect. Notes Comput. Sci. Eng. **91**, 73–84 (2013)
20. Karypis, G.: METIS and ParMETIS. Encyclopedia of Parallel Computing, 1117–1124 (2011)
21. Klawonn, A., Widlund, O. B.: Dual-primal FETI methods for linear elasticity. Comm. Pure Appl. Math. **59**, 1523–1572 (2006)
22. Kolev, T. V., Vassilevski, P. S.: Parallel auxiliary space AMG for $H(\mathrm{curl})$ problems. J. Comput. Math. **27**, 604–623 (2009)
23. Lashuk, I. V., Vassilevski, P. S.: The construction of the coarse de Rham complexes with improved approximation properties. Comput. Methods Appl. Math. **14**, 257–303 (2014)
24. Ma, Y., Hu, K., Hu, X., Xu, J.: Robust preconditioners for incompressible MHD models. J. Comp. Phys. **316**, 721–746 (2016)
25. Mandel, J., Dohrmann, C. R., Tezaur, R.: An algebraic theory for primal and dual substructuring methods by constraints. Appl. Numer. Math. **54**, 167–193 (2005)
26. Nédélec, J.-C.: Mixed finite elements in R^3. Numer. Math. **35**, 315–341 (1980)
27. Oh, D.-S., Widlund, O. B., Zampini, S., Dohrmann, C.: BDDC Algorithms with deluxe scaling and adaptive selection of primal constraints for Raviart-Thomes Vector Fields. Math. Comp., published electronically (2017)
28. Rapetti, F. and Toselli, A.: A FETI preconditioner for two-dimensional edge element approximations of Maxwell's equations on nonmatching grids. SIAM J. Sci. Comput. **23**, 92–108 (2001)
29. Reitzinger, S., Schöberl, J.: An algebraic multigrid method for finite element discretizations with edge elements. Numer. Linear Algebra Appl **9**, 223–238 (2002)
30. Rieben, R. N., White, D. A.: Verification of high-order mixed finite-element solution of transient magnetic diffusion problems. IEEE Trans. Magnetics **42**, 25–39 (2006)
31. Schwarzbach, C., Börner, R.-U., Spitzer, K.: Three-dimensional adaptive higher order finite element simulation for geo-electromagnetics: a marine CSEM example. Geophys. J. Internat., **187**, 63–74 (2011)
32. Toselli, A.: Neumann-Neumann methods for vector field problems. Electron. Trans. Numer. Anal. **11**, 1–24 (2000)
33. Toselli, A.: Overlapping Schwarz methods for Maxwell's equations in three dimensions. Numer. Math. **86**, 733–752 (2000)
34. Toselli, A.: Dual-primal FETI algorithms for edge finite-element approximations in 3D. IMA J. Numer. Anal. **26**, 96–130 (2006)
35. Toselli, A., Vasseur, X.: Dual-primal FETI algorithms for edge element approximations: two-dimensional h and p finite elements on shape-regular meshes. SIAM J. Numer. Anal. **42**, 2590–2611 (2005)
36. Toselli, A., Vasseur, X.: Robust and efficient FETI domain decomposition algorithms for edge element approximations. COMPEL **24**, 396–407 (2005)
37. Toselli, A., Widlund, O. B.: Domain decomposition methods—algorithms and theory. Springer-Verlag (2005)
38. Tu, X: Three-level BDDC in three dimensions. SIAM J. Sci. Comput. **29**, 1759–1780 (2007)
39. Vassilevski, P. S., Villa, U.: A Mixed Formulation for the Brinkman Problem, SIAM J. Numer. Anal. **5**2, 258–281 (2014).
40. Zampini, S., Keyes, D. E.: On the Robustness and Prospects of Adaptive BDDC Methods for Finite Element Discretizations of Elliptic PDEs with High-Contrast Coefficients. Proceedings of the Platform for Advanced Scientific Computing Conference, 6:1–6:13 (2016)
41. Zampini, S.: PCBDDC: a class of robust dual-primal methods in PETSc. SIAM J. Sci. Comp. **38**, S282–S306 (2016)

42. Zampini, S.: Adaptive BDDC Deluxe Methods for H(curl). Domain Decomposition Methods in Science and Engineering XXIII, Lect. Notes Comput. Sci. Eng. **116**, 285–292 (2017)
43. Zampini, S., Tu, X.: Multilevel BDDC deluxe algorithms with adaptive coarse spaces for flow in porous media. SIAM J. Sci. Comp. **39**, A1389–A1415 (2017)

Restricted additive Schwarz method for some inequalities perturbed by a Lipschitz operator

Lori Badea[1]

1 Introduction

The first restricted additive Schwarz methods have been introduced for algebraic linear systems in Cai et al. [1998], Cai and Sarkis [1999] and Frommer and Szyld [2001]. In Frommer et al. [2002] and Nabben and Szyld [2002] the restricted variant of the multiplicative Schwarz method is also analyzed. Numerical experiments have proven that these restricted methods, besides the fact that they sometimes converge faster and also preserve the good properties of the usual additive methods, they reduce the communication time when they are implemented on distributed memory computers. In Efstathiou and Gander [2003], it is explained this fact by showing that even if the restricted method is defined at the matrix level, it can be interpreted as an iteration at the continuous level of the given problem. Restricted additive Schwarz methods for complementarity problems have been introduced in Yang and Li [2012], Zhang et al. [2015], Xu et al. [2014] and Xu et al. [2011].

In the above papers, the methods are approached by a matricial point of view. In this paper, we introduce and analyze a restricted additive method for inequalities perturbed by a Lipschitz operator in the functional framework of the PDEs. Such an approach is not new in the case of the additive and multiplicative Schwarz methods, including the multilevel and multigrid methods for inequalities (see Badea [2008b], Badea [2015] and Badea [2008a], for instance).

In the next section, like in Badea [2008a], we give an existence and uniqueness result concerning the solution of the inequalities we consider¿ Also, we introduce the method as a subspace correction algorithm, prove the convergence and estimate the error in a general framework of a finite dimensional Hilbert space. In Section 3, by introducing the finite element spaces,

Institute of Mathematics of the Romanian Academy and Francophone Center for Mathematics in Bucharest, P.O. Box 1-764, RO-014700 Bucharest, Romania lori.badea@imar.ro

we conclude that both the convergence condition and convergence rate are independent of the mesh parameters, the number of subdomains and of the parameters of the domain decomposition, but the convergence condition is a little more restrictive than the existence and uniqueness condition of the solution.

In a forthcoming paper, by considering the perturbing operator of a particular form, we introduce and analyze some restricted additive Schwarz-Richardson methods for inequalities which do not arise from the minimization of a functional. Also, we shall compare the convergence of these restricted additive methods with the convergence of the corresponding additive methods.

2 Convergence result in a Hilbert space

Let V be finite dimensional real Hilbert space with the basis φ_j, $j = 1, \ldots, d$, and let c_d and C_d be two constants such that, for any $v = \sum_{j=1}^{d} v_j\varphi_j \in V$, we have

$$c_d \sum_{j=1}^{d} ||v_j\varphi_j||^2 \leq ||v||^2 \leq C_d \sum_{j=1}^{d} ||v_j\varphi_j||^2 \tag{1}$$

Also, let $V_1, \ldots, V_m$ be some closed subspaces of V and $K \subset V$ be a non empty closed convex set. We consider a Gâteaux differentiable functional $F : V \to \mathbf{R}$ and assume that there exist two real numbers $\alpha,\ \beta > 0$ for which

$$\alpha||v-u||^2 \leq \langle F'(v) - F'(u), v-u\rangle \text{ and } ||F'(v) - F'(u)||_{V'} \leq \beta||v-u|| \tag{2}$$

for any $u, v \in V$. Above, we have denoted by F' the Gâteaux derivative of F. Following the way in Glowinski et al. [1981], we can prove that for any $u, v \in V$, we have

$$\langle F'(u), v-u\rangle + \frac{\alpha}{2}||v-u||^2 \leq F(v) - F(u) \leq \langle F'(u), v-u\rangle + \frac{\beta}{2}||v-u||^2 \tag{3}$$

Also, we consider an operator $T : V \to V'$ with the property that there exists $\gamma > 0$ such that

$$||T(v) - T(u)||_{V'} \leq \gamma||v-u|| \text{ for any } u,\ v \in V. \tag{4}$$

By using the above functional $F : V \to \mathbf{R}$, we also introduce the functional $\mathcal{F} : V \to \mathbf{R}$ defined as $\mathcal{F}(v) = \sum_{j=1}^{d} F(v_j\varphi_j)$. Evidently, the derivative $\mathcal{F}'$ of $\mathcal{F}$ at $u = \sum_{j=1}^{d} u_j\varphi_j$ in the direction $v = \sum_{j=1}^{d} v_j\varphi_j$ is written as $\langle \mathcal{F}'(u), v\rangle = \sum_{j=1}^{d} \langle F'(u_j\varphi_j), v_j\varphi_j\rangle$ and, in view of (3), we have

$$\begin{array}{l} \langle \mathcal{F}'(u), v-u\rangle + \frac{\alpha}{2}\sum_{j=1}^{d} ||(v_j-u_j)\varphi_j||^2 \leq \mathcal{F}(v) - \mathcal{F}(u) \\ \leq \langle \mathcal{F}'(u), v-u\rangle + \frac{\beta}{2}\sum_{j=1}^{d} ||(v_j-u_j)\varphi_j||^2 \end{array} \tag{5}$$

for any $u = \sum_{j=1}^{d} u_j\varphi_j,\ v = \sum_{j=1}^{d} v_j\varphi_j \in V$. Evidently, from the convexity of F we get that $\mathcal{F}$ is also a convex functional. Finally, we assume that if K is not bounded then the functional $\mathcal{F}$ is coercive in the sense that $\mathcal{F}(v)/||v|| \to \infty$ as $||v|| \to \infty,\ v \in V$.

Now, we define an operation $* : V \times V \to V$ by

$$u * v = \sum_{j=1}^{d} u_j v_j \varphi_j \text{ for any } u = \sum_{j=1}^{d} u_j\varphi_j \text{ and } v = \sum_{j=1}^{d} v_j\varphi_j \in V \quad (6)$$

We fix some functions $\theta_i = \sum_{j=1}^{d} \theta_{ij}\varphi_j \in V_i$, $i = 1, \ldots, m$, and assume that they have the property

$$0 \le \theta_{ij} \le 1 \text{ and } \sum_{i=1}^{m} \theta_{ij} = 1 \text{ for any } j = 1, \ldots, m \quad (7)$$

i.e., in some sense, they supply a unity decomposition associated with the subspaces $V_1, \ldots, V_m$. Also, we assume that the convex set K has the property

Property 1. If $v, w \in K$ and $\theta = \sum_{j=1}^{d} \theta_j\varphi_j \in V$ with $0 \le \theta_j \le 1, j = 1, \ldots, d$, then $\theta * v + (\bar{1} - \theta) * w \in K$.

Above and in what follows in this section, $\sum_{j=1}^{d} \varphi_j$ is denoted by $\bar{1}$. Using (6), we have $\bar{1} * v = v$ for any $v \in V$. Finally, we consider the problem

$$u \in K : \langle \mathcal{F}'(u), v - u\rangle - \langle T(u), v - u\rangle \ge 0, \text{ for any } v \in K. \quad (8)$$

which is a variational inequality perturbed by the operator T. Concerning the existence and the uniqueness of the solution of this problem we have the following result (see Badea [2008a], for the proof of a similar result).

Proposition 1. *If $\frac{\gamma}{\alpha}C_d < 1$, then problem (8) has a unique solution.*

Since the functional $\mathcal{F}$ is convex and differentiable, problem (8) is equivalent with the minimization problem

$$u \in K : \mathcal{F}(u) - \langle T(u), u\rangle \le \mathcal{F}(v) - \langle T(u), v\rangle, \text{ for any } v \in K. \quad (9)$$

We write the restricted additive algorithm for the solution of problem (8) as

Algorithm 1 *We start the algorithm with an arbitrary $u^0 \in K$. At iteration $n+1$, having $u^n \in K$, $n \ge 0$, we solve the inequalities: find $w_i^{n+1} \in V_i$, $u^n + w_i^{n+1} \in K$ such that*

$$\begin{aligned} &\langle \mathcal{F}'(u^n + w_i^{n+1}), v_i - w_i^{n+1}\rangle - \langle T(u^n), v_i - w_i^{n+1}\rangle \ge 0, \\ &\text{for any } v_i \in V_i,\ u^n + v_i \in K, \end{aligned} \quad (10)$$

for $i = 1, \ldots, m$, and then we update $u^{n+1} = u^n + \sum_{i=1}^{m} \theta_i * w_i^{n+1}$.

Now we prove

Theorem 1. *Let u be the solution of problem (8), and u^n, $n \ge 1$, be its approximations obtained from Algorithm 1. If $\frac{\gamma}{\alpha}C_d \le \vartheta_{\max}$, where $\vartheta_{\max}$ is*

defined in (27), then Algorithm 1 is convergent for any initial guess $u^0 \in K$ *and the error estimates*

$$\begin{aligned}&\mathcal{F}(u^n) - \langle T(u), u^n\rangle - \mathcal{F}(u) + \langle T(u), u\rangle\\ &\leq \left(\frac{\tilde{C}}{\tilde{C}+1}\right)^n \left[\mathcal{F}(u^0) - \langle T(u), u^0\rangle - \mathcal{F}(u) + \langle T(u), u\rangle\right]\end{aligned} \tag{11}$$

and

$$\begin{aligned}&\textstyle\sum_{j=1}^d ||(u_j^n - u_j)\varphi_j||^2 \leq \frac{2}{\alpha}\left(\frac{\tilde{C}}{\tilde{C}+1}\right)^n \left[\mathcal{F}(u^0) - \langle T(u), u^0\rangle\right.\\ &\left.-\mathcal{F}(u) + \langle T(u), u\rangle\right]\end{aligned} \tag{12}$$

hold for any $n \geq 1$*, where constant* $\tilde{C}$ *is given in (28).*

Proof. Using (5), (7) and (10), we get

$\mathcal{F}(u^{n+1}) - \mathcal{F}(u) + \langle T(u), u - u^{n+1}\rangle + \frac{\alpha}{2}\sum_{j=1}^d ||(u_j^{n+1} - u_j)\varphi_j||^2$
$\leq \langle \mathcal{F}'(u^{n+1}), u^{n+1} - u\rangle + \langle T(u), u - u^{n+1}\rangle$
$\leq \sum_{i=1}^m \langle \mathcal{F}'(u^n + w_i^{n+1}) - \mathcal{F}'(u^{n+1}), \theta_i * (u - u^n) + (\bar{1} - \theta_i) * w_i^{n+1} - w_i^{n+1}\rangle$
$- \sum_{i=1}^m \langle T(u^n), \theta_i * (u - u^n) + (\bar{1} - \theta_i) * w_i^{n+1} - w_i^{n+1}\rangle + \langle T(u), u - u^{n+1}\rangle$

Above, we have used the fact that $\theta_i*(u-u^n)+(\bar{1}-\theta_i)*w_i^{n+1} \in V_i$ and, in view of Property 1, $u^n+\theta_i*(u-u^n)+(\bar{1}-\theta_i)*w_i^{n+1} = (\bar{1}-\theta_i)*(u^n+w_i^{n+1})+\theta_i*u \in K$ and therefore, we can replace v_i by $\theta_i * (u - u^n) + (\bar{1} - \theta_i) * w_i^{n+1}$ in (10). Consequently, we have

$$\begin{aligned}&\mathcal{F}(u^{n+1}) - \mathcal{F}(u) - \langle T(u), u^{n+1} - u\rangle + \tfrac{\alpha}{2}\textstyle\sum_{j=1}^d ||(u_j^{n+1} - u_j)\varphi_j||^2\\ &\leq \textstyle\sum_{i=1}^m \langle \mathcal{F}'(u^n + w_i^{n+1}) - \mathcal{F}'(u^{n+1}), \theta_i * (u - u^n - w_i^{n+1})\rangle\\ &+ \textstyle\sum_{i=1}^m \langle T(u) - T(u^n), \theta_i * (u - u^n - w_i^{n+1})\rangle\end{aligned} \tag{13}$$

In view of (2) and (7), we have

$\sum_{i=1}^m \langle \mathcal{F}'(u^n + w_i^{n+1}) - \mathcal{F}'(u^{n+1}), \theta_i * (u - u^n - w_i^{n+1})\rangle$
$\leq \beta \sum_{i=1}^m \sum_{j=1}^d \theta_{ij} ||((1-\theta_{ij})w_{ij}^{n+1} - \sum_{k=1,\ k\neq i}^m \theta_{kj} w_{kj}^{n+1})\varphi_j||$
$\cdot ||(u_j - u_j^{n+1} - (1-\theta_{ij})w_{ij}^{n+1} + \sum_{k=1,\ k\neq i}^m \theta_{kj} w_{kj}^{n+1})\varphi_j||$
$\leq \beta \sum_{i=1}^m \sum_{j=1}^d \theta_{ij} \left((1-\theta_{ij})||w_{ij}^{n+1}\varphi_j|| + \sum_{k=1,\ k\neq i}^m \theta_{kj}||w_{kj}^{n+1}\varphi_j||\right)$
$\cdot \left(||(u_j - u_j^{n+1})\varphi_j|| + (1-\theta_{ij})||w_{ij}^{n+1}\varphi_j|| + \sum_{k=1,\ k\neq i}^m \theta_{kj}||w_{kj}^{n+1}\varphi_j||\right)$
$\leq \beta \sum_{i=1}^m \sum_{j=1}^d \theta_{ij} \left[(1 + \frac{1}{2\varepsilon_1})\left((1-\theta_{ij})||w_{ij}^{n+1}\varphi_j|| \right.\right.$
$\left.\left. + \sum_{k=1,\ k\neq i}^m \theta_{kj}||w_{kj}^{n+1}\varphi_j||\right)^2 + \frac{\varepsilon_1}{2}||(u_j - u_j^{n+1})\varphi_j||^2\right] \leq 2\beta(1 + \frac{1}{2\varepsilon_1})$
$\cdot \sum_{i=1}^m \sum_{j=1}^d \theta_{ij}(1-\theta_{ij}) \left((1-\theta_{ij})||w_{ij}^{n+1}\varphi_j||^2 + \sum_{k=1,\ k\neq i}^m \theta_{kj}||w_{kj}^{n+1}\varphi_j||^2\right)$
$+\beta\frac{\varepsilon_1}{2}\sum_{j=1}^d ||(u_j - u_j^{n+1})\varphi_j||^2 = 2\beta(1 + \frac{1}{2\varepsilon_1})\sum_{i=1}^m \sum_{j=1}^d \theta_{ij}(1-\theta_{ij})$
$\cdot(1 - 2\theta_{ij})||w_{ij}^{n+1}\varphi_j||^2 + 2\beta(1 + \frac{1}{2\varepsilon_1})\sum_{i=1}^m \sum_{j=1}^d \theta_{ij}(1-\theta_{ij})$
$\cdot \sum_{k=1}^m \theta_{kj}||w_{kj}^{n+1}\varphi_j||^2 + \beta\frac{\varepsilon_1}{2}\sum_{j=1}^d ||(u_j - u_j^{n+1})\varphi_j||^2$

or

$$\begin{aligned}&\textstyle\sum_{i=1}^m \langle \mathcal{F}'(u^n + w_i^{n+1}) - 4\beta(1\\ &+\tfrac{1}{2\varepsilon_1})\textstyle\sum_{i=1}^m \sum_{j=1}^d \theta_{ij}||w_{ij}^{n+1}\varphi_j||^2 + \beta\tfrac{\varepsilon_1}{2}\textstyle\sum_{j=1}^d ||(u_j - u_j^{n+1})\varphi_j||^2\end{aligned} \tag{14}$$

for any $\varepsilon_1 > 0$. Also, from (4) and (1), we get
$\sum_{i=1}^{m}\langle T(u)-T(u^n), \theta_i * (u-u^n-w_i^{n+1})\rangle = \langle T(u)-T(u^n), u-u^n$
$-\sum_{i=1}^{m}\theta_i * w_i^{n+1}\rangle = \langle T(u)-T(u^n), u-u^{n+1})\rangle \le \gamma||u-u^n||\ ||u-u^{n+1}||$
$\le \gamma\left(||u-u^{n+1}|| + ||\sum_{j=1}^{d}\sum_{i=1}^{m}\theta_{ij}w_{ij}^{n+1}\varphi_j||\right)||u-u^{n+1}||$
$\le \gamma C_d\left((1+\frac{\varepsilon_2}{2})\sum_{j=1}^{d}||(u_j-u_j^{n+1})\varphi_j||^2 + \frac{1}{2\varepsilon_2}\sum_{j=1}^{d}||\sum_{i=1}^{m}\theta_{ij}w_{ij}^{n+1}\varphi_j||^2\right)$
i.e., using (7), we have

$$\begin{array}{l}\sum_{i=1}^{m}\langle T(u)-T(u^n), \theta_i * (u-u^n-w_i^{n+1})\rangle \le \gamma C_d\left((1+\frac{\varepsilon_2}{2})\right.\\ \left.\cdot\sum_{j=1}^{d}||(u_j-u_j^{n+1})\varphi_j||^2 + \frac{1}{2\varepsilon_2}\sum_{j=1}^{d}\sum_{i=1}^{m}\theta_{ij}||w_{ij}^{n+1}\varphi_j||^2\right)\end{array} \tag{15}$$

for any $\varepsilon_2 > 0$. From (13), (14) and (15), we get

$$\begin{array}{l}\mathcal{F}(u^{n+1}) - \mathcal{F}(u) - \langle T(u), u^{n+1}-u\rangle + \left(\frac{\alpha}{2} - \beta\frac{\varepsilon_1}{2} - \gamma C_d(1+\frac{\varepsilon_2}{2})\right)\\ \cdot\sum_{j=1}^{d}||(u_j^{n+1}-u_j)\varphi_j||^2 \le \left[4\beta(1+\frac{1}{2\varepsilon_1}) + \gamma C_d\frac{1}{2\varepsilon_2}\right]\\ \cdot\sum_{i=1}^{m}\sum_{j=1}^{d}\theta_{ij}||w_{ij}^{n+1}\varphi_j||^2\end{array} \tag{16}$$

for any $\varepsilon_1,\ \varepsilon_2 > 0$.

Now, by taking $v_i = (\bar{1}-\theta_i) * w_i^{n+1}$ in (10), for $i = 1, \ldots, m$, we get

$$\sum_{j=1}^{d}\theta_{ij}\left[\langle F'((u_j^n + w_{ij}^{n+1})\varphi_j), -w_{ij}^{n+1}\varphi_j\rangle - \langle T(u^n), -w_{ij}^{n+1}\varphi_j\rangle\right] \ge 0 \tag{17}$$

In view of (7), the convexity of F, (2) and the above equation, we have
$\mathcal{F}(u^{n+1}) - \mathcal{F}(u^n) \le \sum_{j=1}^{d}\sum_{i=1}^{m}\theta_{ij}[F((u_j^n + w_{ij}^{n+1})\varphi_j) - F(u_j^n\varphi_j)]$
$\le \sum_{j=1}^{d}\sum_{i=1}^{m}\theta_{ij}\left[-\frac{\alpha}{2}||w_{ij}^{n+1}||^2 - \langle F'((u_j^n + w_{ij}^{n+1})\varphi_j), -w_{ij}^{n+1}\varphi_j\rangle\right]$
$= \sum_{j=1}^{d}\sum_{i=1}^{m}\theta_{ij}\left[-\frac{\alpha}{2}||w_{ij}^{n+1}\varphi_j||^2 - \langle T(u^n), -w_{ij}^{n+1}\varphi_j\rangle\right.$
$\left.-\langle F'((u_j^n + w_{ij}^{n+1})\varphi_j), -w_{ij}^{n+1}\varphi_j\rangle + \langle T(u^n), -w_{ij}^{n+1}\varphi_j\rangle\right]$
$\le \sum_{j=1}^{d}\sum_{i=1}^{m}\theta_{ij}\left[-\frac{\alpha}{2}||w_{ij}^{n+1}\varphi_j||^2 - \langle T(u^n), -w_{ij}^{n+1}\varphi_j\rangle\right]$
Consequently, we have

$$\begin{array}{l}\frac{\alpha}{2}\sum_{i=1}^{m}\sum_{j=1}^{d}\theta_{ij}||w_{ij}^{n+1}\varphi_j||^2 \le \mathcal{F}(u^n) - \mathcal{F}(u^{n+1})\\ +\langle T(u), u^{n+1}-u^n\rangle + \langle T(u^n)-T(u), u^{n+1}-u^n\rangle\end{array} \tag{18}$$

With a proof similar to that of (15), we get

$$\begin{array}{l}\langle T(u^n)-T(u), u^{n+1}-u^n\rangle \le \gamma C_d\left[\frac{\varepsilon_3}{2}\sum_{j=1}^{d}||(u_j^{n+1}-u_j)\varphi_j||^2\right.\\ \left.+(1+\frac{1}{2\varepsilon_3})\sum_{i=1}^{m}\sum_{j=1}^{d}\theta_{ij}||w_{ij}^{n+1}\varphi_j||^2\right]\end{array} \tag{19}$$

for any $\varepsilon_3 > 0$.

Consequently, from (18) and (19), we get

$$\begin{array}{l}\left[\frac{\alpha}{2} - \gamma C_d(1+\frac{1}{2\varepsilon_3})\right]\sum_{i=1}^{m}\sum_{j=1}^{d}\theta_{ij}||w_{ij}^{n+1}\varphi_j||^2 \le \mathcal{F}(u^n) - \mathcal{F}(u^{n+1})\\ +\langle T(u), u^{n+1}-u^n\rangle + \gamma C_d\frac{\varepsilon_3}{2}\sum_{j=1}^{d}||(u_j^{n+1}-u_j)\varphi_j||^2\end{array} \tag{20}$$

for any $\varepsilon_3 > 0$. Let us write

$$C_1 = \frac{\alpha}{2} - \gamma C_d(1 + \frac{1}{2\varepsilon_3}) \tag{21}$$

For values of γ, α and ε_3 such that $C_1 > 0$, from (16) and (20), we have

$$\begin{array}{l} \mathcal{F}(u^{n+1}) - \mathcal{F}(u) - \langle T(u), u^{n+1} - u\rangle + C_2 \sum_{j=1}^{d} ||(u_j^{n+1} - u_j)\varphi_j||^2 \\ \leq \tilde{C}\left[\mathcal{F}(u^n) - \mathcal{F}(u^{n+1}) + \langle T(u), u^{n+1} - u^n\rangle\right] \end{array} \tag{22}$$

where

$$\tilde{C} = \frac{1}{C_1}\left(4\beta(1 + \frac{1}{2\varepsilon_1}) + \gamma C_d \frac{1}{2\varepsilon_2}\right) \tag{23}$$

and

$$C_2 = \frac{\alpha}{2} - \beta\frac{\varepsilon_1}{2} - \gamma C_d(1 + \frac{\varepsilon_2}{2}) - \gamma C_d \frac{\varepsilon_3}{2}\tilde{C} \tag{24}$$

In view of (22), assuming that $C_2 \geq 0$, we easily get (11). Estimation (12) follows from (11) and (3) and (8). Indeed, we have

$$\begin{array}{l} \mathcal{F}(u^n) - \mathcal{F}(u) - \langle T(u), u^n - u\rangle = \sum_{j=1}^{d} F(u_j^n\varphi_j) - \sum_{j=1}^{d} F(u_j\varphi_j) \\ -\langle T(u), u^n - u\rangle \geq \sum_{j=1}^{d}\langle F'(u_j\varphi_j), (u_j^n - u_j)\varphi_j\rangle \\ +\frac{\alpha}{2}\sum_{j=1}^{d}||(u_j^n - u_j)\varphi_j||^2 - \langle T(u), u^n - u\rangle = \langle \mathcal{F}'(u), u^n - u\rangle \\ -\langle T(u), u^n - u\rangle + \frac{\alpha}{2}\sum_{j=1}^{d}||(u_j^n - u_j)\varphi_j||^2 \geq \frac{\alpha}{2}\sum_{j=1}^{d}||(u_j^n - u_j)\varphi_j||^2 \end{array} \tag{25}$$

Using (23), (24) and (21), condition $C_2 \geq 0$ can be written as $C_2 = A - 4B\beta - \frac{\beta}{2}(\varepsilon_1 + 4\frac{B}{\varepsilon_1}) - \frac{\gamma C_d}{2}(\varepsilon_2 + \frac{B}{\varepsilon_2}) \geq 0$ with $A = \frac{\alpha}{2} - \gamma C_d$ and $B = \frac{\gamma C_d \frac{\varepsilon_3}{2}}{A - \gamma C_d \frac{1}{2\varepsilon_3}}$
The maximum value of C_2 is obtained for

$$\varepsilon_1 = 2\frac{\gamma C_d}{A} \quad \varepsilon_2 = \varepsilon_3 = \frac{\gamma C_d}{A} \tag{26}$$

Consequently, for these values, we should have
$C_{2\max} = \frac{\alpha^3}{A^2}\left[\frac{1}{2}(\frac{1}{2} - \frac{\gamma C_d}{\alpha})(\frac{1}{2} - 2\frac{\gamma C_d}{\alpha}) - \frac{2\beta}{\alpha}\frac{\gamma C_d}{\alpha}(\frac{1}{2} + \frac{\gamma C_d}{\alpha})\right] \geq 0$, or

$$C_d\frac{\gamma}{\alpha} \leq \frac{1}{\sqrt{16\frac{\beta^2}{\alpha^2} + 40\frac{\beta}{\alpha} + 1} + 4\frac{\beta}{\alpha} + 3}} = \vartheta_{\max} \tag{27}$$

By a simple calculus, we see that if (27) holds, then condition $C_1 > 0$ is satisfied for the value of ε_3 in (26). Finally, by replacing ε_1, ε_2 and ε_3 in (23) with their values in (26), we get

$$\tilde{C} = 1 + \frac{2\beta}{\alpha}\frac{6\frac{\gamma C_d}{\alpha} + 1}{\frac{\gamma C_d}{\alpha}\left(1 - 2\frac{\gamma C_d}{\alpha}\right)} \geq 1 + \frac{2\beta}{\alpha}\frac{6\vartheta_{\max} + 1}{\vartheta_{\max}(1 - 2\vartheta_{\max})} \tag{28}$$

It should be noted that the convergence condition and the convergence rate are independent of the number m of subspaces.

3 Restricted additive Schwarz method in a finite element space

Let Ω be an open bounded domain in $\mathbf{R}^N$, $N = 1, 2$ or 3, and we consider a simplicial regular mesh partition $\mathcal{T}_h$. We assume that domain Ω is decomposed in m subdomains, $\Omega = \bigcup_{i=1}^m \Omega_i$, and that $\mathcal{T}_h$ supplies a mesh partition for each subdomain Ω_i, $i = 1, \ldots, m$. We associate to the mesh partition $\mathcal{T}_h$ the piecewise linear finite element space $V_h \subset H_0^1(\Omega)$ and to the domain decomposition the subspaces $V_h^i \subset H_0^1(\Omega_i)$. We assume that the convex set $K_h \subset V_h$ has the following

Property 2. If $v, w \in K_h$, and if $\theta \in V_h$, $0 \leq \theta \leq 1$, then $L_h(\theta v + (1-\theta)w) \in K_h$.

Above and also in the following, we denote by L_h the P_1-Lagrangian interpolation operator which uses the function values at the nodes of the mesh $\mathcal{T}_h$. It is easy to see that the convex sets of two-obstacle type have Property 2.

Now, we estimate C_d in (1). Given a triangle $\tau \in \mathcal{T}_h$, let $J_\tau = \{1 \leq j \leq d : \tau \subset \text{ supp } \varphi_j\}$. Then, for a $v = \sum_{j=1}^d v_j \varphi_j \in V_h$, and using the norm of $H^1(\Omega)$ we have

$||v||^2 = \sum_\tau ||v||_\tau^2 = \sum_\tau \left(\sum_{j \in J_\tau} v_j \varphi_j, \sum_{j \in J_\tau} v_j \varphi_j \right)_\tau \leq$
$\sum_\tau |J_\tau| \sum_{j \in J_\tau} ||v_j \varphi_j||_\tau^2 \leq \sum_\tau |J_\tau| \sum_{j=1}^d ||v_j \varphi_j||_\tau^2 \leq C_d \sum_{j=1}^d \sum_\tau ||v_j \varphi_j||_\tau^2$
$= C_d \sum_{j=1}^d ||v_j \varphi_j||^2$

where we have denoted $C_d = \max_{\tau \in \mathcal{T}_h} |J_\tau|$. Since $\mathcal{T}_h$ are simplicial meshes, then $\max_\tau |J_\tau|$ is independent of the mesh parameters when $h \to 0$. Therefore, we can consider that C_d *is independent of the domain or mesh parameters.*

Finally, it is evident that $*$ in (6) can be written as $u * v = L_h(uv)$ for any $u,\ v \in V_h$. Moreover, if $\{\theta_1, \ldots, \theta_m\} \subset V_h$ is a unity partition associated with the domain decomposition, then (7) holds for any $v \in V_h$. Besides that, in view of Property 2 of the convex set K_h, this convex set also has Property 1. In the matricial description of the method, some restriction operators, $R_1^0, \ldots, R_m^0$, are used instead of our unity partition $\{\theta_1, \ldots, \theta_m\}$. If we associate to a $v = \sum_{j=1}^d v_j \varphi_j \in V_h$ the vector $(v_1, \ldots, v_d)$ then $\theta_i * v$ is associated with $R_i^0(v_1, \ldots, v_d)$. In general, these restriction operators supply a minimum overlap i.e., with our notations, the components θ_{ij} of the functions $\theta_i = \sum_{j=0}^m \theta_{ij} \varphi_j$ satisfy either $\theta_{ij} = 1$ or $\theta_{ij} = 0$. A PDEs definition of the method using a unity partition associated to the domain decomposition and which is very close to that introduced by us is given in Dolean et al. [2015].

From (27), (28) and the above comments we can conclude that the convergence condition and convergence rate of Algorithm 1 are independent of the mesh parameters and of both the number of subdomains and the parameters of the domain decomposition, but the convergence condition is more restrictive than the existence and uniqueness condition of the solution given in Proposition 1.

References

L. Badea. Schwarz methods for inequalities with contraction operators. *Journal of Computational and Applied Mathematics*, 215(1):196–219, 2008a.

L. Badea. Additive Schwarz method for the constrained minimization of functionals in reflexive banach spaces. In *U. Langer et al. (eds.), Domain decomposition methods in science and engineering XVII, LNSE 60*, pages 427–434. Springer, 2008b.

L. Badea. Convergence rate of some hybrid multigrid methods for variational inequalities. *Journal of Numerical Mathematics*, 23(3):195–210, 2015.

X.-C. Cai and M. Sarkis. A restricted additive Schwarz preconditioner for general sparse linear systems. *SIAM J. Sci. Comput.*, 21:792–797, 1999.

X.-C. Cai, C. Farhat, and M. Sarkis. A minimum overlap restricted additive Schwarz preconditioner and applications to 3d flow simulations. In *Contemporary Mathematics, 218*, pages 479–485. AMS, 1998.

V. Dolean, P. Jolivet, and F. Nataf. *An introduction to domain decomposition methods: algorithms, theory, and parallel implementation.* SIAM, 2015.

E. Efstathiou and M.J. Gander. Why restricted additive Schwarz converges faster than additive Schwarz. *BIT Numer. Math.*, 43:945–959, 2003.

A. Frommer and D. B. Szyld. An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms. *SIAM J. Numer. Anal.*, 39(2):463–479, 2001.

A. Frommer, R. Nabben, and D. B. Szyld. An algebraic convergence theory for restricted additive and multiplicative Schwarz methods. In *N. Debit et. al (eds.), Domain Decomposition Methods in Science and Engineering*, pages 371–377. CIMNE, UPS, Barcelona, 2002.

R. Glowinski, J. L. Lions, and R. Trémolières. *Numerical Analysis of Variational Inequalities.* J. L. Lions, G. Papanicolau, R. T. Rockafellar (Eds.), Studies in Mathematics and its Applications, vol. 8, J. L. Lions, G. Papanicolau, R. T. Rockafellar (Eds.), Studies in Mathematics and its Applications, vol. 8,, 1981.

R. Nabben and D. B. Szyld. Convergence theory of restricted multiplicative Schwarz methods. *SIAM J. Numer. Anal.*, 40(6):2318–2336, 2002.

H. Xu, K. Huang, S. Xie, and Z. Sun. Restricted additive Schwarz method for a kind of nonlinear complementarity problem. *Journal of Computational Mathematics*, 32(5):547–559, 2014.

H.-R. Xu, K.-K. Huang, and S.-L. Xie. Restricted additive Schwarz method for nonlinear complementarity problem with an m-function. *Communications in Computer and Information Science*, 158:46–50, 2011.

H. Yang and Q. Li. Overlapping restricted additive Schwarz method applied to the linear complementarity problem with an h-matrix. *Comput. Optim. Appl.*, 51:223–239, 2012.

L.-T. Zhang, T.-X. Gu, and X.-P. Liu. Overlapping restricted additive Schwarz method with damping factor for h-matrix linear complementarity problem. *Applied Mathematics and Computation*, 271:1–10, 2015.

Does SHEM for Additive Schwarz work better than predicted by its condition number estimate ?

Petter E. Bjørstad[2], Martin J. Gander[1], Atle Loneland[2], and Talal Rahman[3]

1 Introduction and Model Problem

The SHEM (Spectral Harmonically Enriched Multiscale) coarse space is a new coarse space for arbitrary overlapping or non-overlapping domain decomposition methods. In contrast to recent new coarse spaces like GenEO [13] or the one in [12] that improve certain Rayleigh quotients in the convergence analysis of the underlying domain decomposition method, SHEM is based on understanding the stationary iterates of the domain decomposition method itself (see [6] for details), and can thus be constructed and used also for domain decomposition methods which do not (yet) have such a convergence analysis, like for example Restricted Additive Schwarz (RAS) [7], or optimized Schwarz [4]. SHEM is based on the approximation of an optimal coarse space which was discovered in [3], and further studied in [5, 4, 7], see [6] for a general introduction, and also [9] for the specific case of Additive Schwarz (AS). SHEM can use spectral information, as its name indicates, but can also be constructed avoiding eigenvalue problems, for examples, see [8]. If a convergence analysis for the domain decomposition method is available, SHEM can improve the corresponding convergence estimate, see [8] for a condition number estimate when SHEM is used with AS. We are interested here to test numerically if in this case

1. the hypothesis of small overlap (one or two mesh sizes) in the proof in [9] is necessary for the condition number estimate to hold in practice;
2. the quadratic growth in the factor H/h in the condition number estimate from [9] is really present when the method is used numerically.

Section of Mathematics, University of Geneva, 1211 Geneva 4, Switzerland Martin.Gander@unige.ch · Department of Informatics, University of Bergen, 5020 Bergen, Norway Atle.Loneland@ii.uib.no · Department of Computing, Mathematics and Physics, Western Norway University of Applied Sciences, 5063 Bergen, Norway Talal.Rahman@hvl.no

We consider as our model problem the following variational formulation of a second order elliptic boundary value problem with Dirichlet boundary conditions: find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = \int_\Omega \alpha(x)\nabla u \cdot \nabla v \, dx = \int_\Omega fv \, dx \quad \forall v \in H_0^1(\Omega), \tag{1}$$

where Ω is a bounded convex domain in $\mathbb{R}^2$, $f \in L^2(\Omega)$ and $\alpha \in L^\infty(\Omega)$ such that $\alpha \geq \alpha_0$ for some positive constant α_0. Discretizing this problem using P1 finite elements from the finite element space V_h with associated mesh $\mathcal{T}_h(\Omega)$ leads to the linear system

$$A\mathbf{u} = \mathbf{f}. \tag{2}$$

Let Ω be partitioned into non-overlapping open, connected Lipschitz polytopes $\{\Omega_i : i = 1, \ldots, N\}$ such that $\overline{\Omega} = \bigcup_{i=1}^N \overline{\Omega}_i$, where each $\overline{\Omega}_i$ is assumed to consist of elements from $\mathcal{T}_h(\Omega)$. We assume that this partitioning is shape-regular. By extending each subdomain Ω_i with a distance δ in each direction, we create a further decomposition of Ω into overlapping subdomains $\{\Omega_i'\}_{i=1}^N$. As usual, we assume that each point $x \in \Omega$ is contained in at most N_0 subdomains (finite covering). The layer of elements in Ω_i touching the boundary $\partial\Omega_i$ is denoted by Ω_i^h and we assume that the triangles corresponding to this layer are shape regular with minimum diameter $h_i := \min_{K \in \mathcal{T}_h(\Omega_i^h)} h_K$, where h_K is the diameter of the triangle K. The interfaces between two subdomains, Ω_i and Ω_j, are defined as $\overline{\Gamma}_{ij} := \overline{\Omega}_i \cap \overline{\Omega}_j$. The sets of vertices of elements in $\mathcal{T}_h(\Omega)$ (nodal points) belonging to Ω, Ω_i, $\partial\Omega$, $\partial\Omega_i$ and Γ_{ij} are denoted by Ω_h, Ω_{ih}, $\partial\Omega_h$, $\partial\Omega_{ih}$ and Γ_{ijh}. With each interface we define the space of finite element functions restricted to Γ_{ij} and zero on $\partial\Gamma_{ij}$ as $V_h^0(\Gamma_{ij})$.

We define the restriction of the bilinear form $a(\cdot,\cdot)$ to an interface Γ_{ij} shared by two subdomains as

$$a_{\Gamma_{ij}}(u,v) := \left(\alpha_{|\Gamma_{ij}}(x) D_\tau u, D_\tau v\right)_{L^2(\Gamma_{ij})},$$

where $\alpha_{|\Gamma_{ij}}(x) := \lim_{y \in \Omega_i \to x} \alpha(y)$ and D_τ denotes the tangent derivative with respect to Γ_{ij}. In order to obtain continuous basis functions across subdomain interfaces, we define a second bilinear form on each interface Γ_{ij},

$$\bar{a}_{\Gamma_{ij}}(u,v) := \left(\overline{\alpha}_{ij}(x) D_\tau u, D_\tau v\right)_{L^2(\Gamma_{ij})},$$

where $\overline{\alpha}_{ij}$ is taken as the maximum of $\alpha_{|\Gamma_{ij}}$ and $\alpha_{|\Gamma_{ji}}$.

Given a partition of unity $\{\chi_i\}_{i=1}^N$ subordinate to the overlapping decomposition defined above and corresponding restriction matrices R_i, as well as a suitable coarse space V_0 with restriction operator R_0, the two-level additive Schwarz method may be defined for $i = 0, \ldots N$ as

$$M_{AS,2}^{-1} = \sum_{i=0}^{N} R_i^T A_i^{-1} R_i \qquad \text{where} \quad A_i := R_i A R_i^T. \tag{3}$$

Classically, coarse spaces for Additive Schwarz methods consist of finite elements on a coarser triangulation $\mathcal{T}_H$ of Ω. This type of choice for the coarse space, however, is not robust with respect to large variations in the coefficient α.

2 The SHEM Coarse Space

SHEM is based on enriching a particular underlying coarse space, which in the case of high contrast problems is the multiscale finite element coarse space, see [1, 10]. We use the variant that generates the multiscale elements by solving lower dimensional problems along the edges, and then extending the result harmonically into the interior of the element. In the case of Laplace's equation on a rectangular domain decomposition, this underlying coarse space would just be Q1 finite elements on the subdomains, see [5]. Note that SHEM is also interesting in this case, since it systematically improves the overall convergence of the underlying domain decomposition method in an optimized way, see [9]. We choose here for SHEM a harmonic enrichment based on solutions of local eigenvalue problems along the interfaces between subdomains[1]:

Definition 1 (Generalized Interface Eigenvalue Problem). For each interface Γ_{ij}, we define the generalized eigenvalue problem: find ψ and λ, such that

$$\bar{a}_{\Gamma_{ij}}(\psi, v) := \lambda b_{\Gamma_{ij}}(\psi, v) \qquad \forall v \in V_h^0(\Gamma_{ij}), \tag{4}$$

where $b_{\Gamma_{ij}}(\psi, v) := h_i^{-1} \sum_{k \in \Gamma_{ijh}} \beta_k \psi_k v_k$ and $\beta_k = \sum_{\substack{K \in \mathcal{T}_h(\Omega) \\ k \in \mathrm{dof}(K)}} \alpha_K$.

We will test the following two types of SHEM coarse spaces:

- SHEM$_m$, where m is an integer: here we choose the m eigenfunctions associated with the smallest m eigenvalues of (4), and extend each of them harmonically into the two subdomains Ω_i and Ω_j adjacent to the interface Γ_{ij} with zero Dirichlet boundary conditions on the remaining part of the subdomain boundaries. These functions are then added to the underlying multiscale coarse space to form SHEM$_m$.
- SHEM$_\tau$, where τ is a given tolerance: here we choose adaptively on each interface Γ_{ij} to include all eigenfunctions associates with eigenvalues smaller

[1] Any other Sturm Liuville problem could be used as well to get a different variant of SHEM, for example more expensive Schur complements corresponding to the Dirichlet to Neumann maps [11], or one could construct even cheaper interface basis functions without eigenvalue problem, see [8].

than τ, extend them harmonically like above and add them to the underlying multiscale coarse space to form SHEM_τ.

Theorem 1 (Condition Number Estimate [8]). *If the overlap is one or two mesh sizes, then the condition number of the two level Schwarz operator (3) with the SHEM_m coarse space can be bounded by*

$$\kappa(M_{AS,2}^{-1}A) \preceq C_0^2(N_0+1), \tag{5}$$

where $C_0^2 \simeq \left(1+\frac{1}{\lambda_{m+1}}\right)$ *and* $\lambda_{m+1} := \min_i \min_{\Gamma_{ij}\subset\partial\Omega_i} \lambda^{ij}_{m_{ij}+1}$.

The restriction on the overlap size is necessary in the proof based on the abstract Schwarz framework. The convergence estimate in Theorem 1 also indicates a quadratic dependence of the condition number on the mesh ratio H/h, even for the case without enrichment, because the inverse of the smallest eigenvalues of (4) have a quadratic dependence on the ratio H/h. In the case of Laplace's equation and without enrichment, such that our coarse space is just the normal Q1 coarse space, standard domain decomposition theory says that the condition number of additive Schwarz should depend linearly on the mesh ratio H/h. We investigate now numerically if these restrictions are really also properties of SHEM_m, or just artefacts in the analysis.

3 Numerical Investigation of the SHEM coarse space

We solve problem (1) with $f = 1$ on a unit square domain $\Omega = (0,1)^2$, and the coefficient $\alpha(x)$ represents various (possibly discontinuous) distributions. We use AS with SHEM_m as a preconditioner for the conjugate gradient method, and stop the iteration when the l_2 norm of the residual is reduced by a factor of 10^{-6}. If not stated otherwise, the coefficient $\alpha(x)$ is equal to 1 for all the numerical examples, except in the areas marked with red where the value of $\alpha(x)$ is equal to $\hat{\alpha}$. All the experiments were carried out using Matlab 9.0 on a serial workstation. For the interface eigenvalue problems, we have in our implementation exploited the fact that we are able to extract exactly the 1D stiffness and mass matrix corresponding to the bilinear forms in Definition 1 algebraically from the global problem.

3.1 Is small overlap necessary for SHEM?

We start by studying the dependence on the overlap for the contrast function $\alpha(x)$ shown in Figure 1. For the case of overlap $\delta = 2h$ and $\delta = 8h$, we show the iteration counts and condition number estimates in Table 1 for

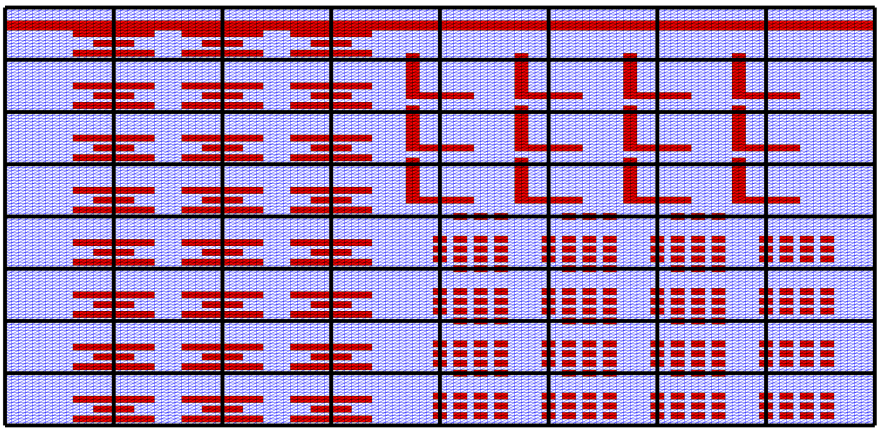


Fig. 1 Distribution of α for a geometry with $h = \frac{1}{128}$, $H = 16h$. The regions marked with red are where α has a large value $\hat{\alpha}$.

the classical multiscale coarse space (MS), SHEM$_m$ and the adaptive variant SHEM$_{\tau=6e-3}$. We see that even though the theory only addressed small overlap, SHEM$_m$ works very well also with larger overlap, and overlap improves the performance like usual. We even see that independence of the contrast arrives for the large overlap already with two enrichment functions instead of three. This is because the middle of the three channels crossing the interfaces in Figure 1 is shorter, and for the large overlap case included in the overlap, and thus not a convergence problem any more for the underlying AS; there are therefore only two channels left the coarse space has to treat, see [6] presented at this conference. In the current adaptive variant SHEM$_{\tau=6e-3}$ it is not clear how to take into account the overlap, and thus the same number of

	MS	SHEM$_1$	SHEM$_2$	SHEM$_3$	SHEM$_4$	SHEM$_{\tau=6e-3}$	
dim.	49	161	273	385	497		
$\hat{\alpha}$	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)	dim.
10^0	21 (1.29e1)	16 (7.45e0)	15 (5.99e0)	13 (5.19e0)	13 (5.15e0)	21 (1.29e1)	49
10^2	122 (3.74e2)	70 (1.17e2)	47 (6.70e1)	19 (6.77e0)	16 (5.66e0)	25 (1.10e1)	233
10^4	367 (3.64e4)	248 (1.10e4)	187 (6.22e3)	19 (6.78e0)	17 (5.73e0)	25 (1.09e1)	233
10^6	610 (3.64e6)	423 (1.10e6)	290 (6.22e5)	19 (6.78e0)	17 (5.73e0)	25 (1.09e1)	233
10^0	16 (5.57e0)	15 (4,88e0)	15 (4.82e0)	15 (4.94e0)	15 (4.95e0)	16 (5.47e0)	49
10^2	47 (4.08e1)	28 (1.53e1)	19 (5.58e0)	18 (5.02e0)	18 (4.99e0)	21 (6.26e0)	233
10^4	145 (3.48e3)	55 (1.08e3)	20 (6.03e0)	18 (5.06e0)	18 (4.99e0)	21 (6.55e0)	233
10^6	241 (3.48e5)	78 (1.08e5)	20 (6.03e0)	18 (5.06e0)	18 (4.99e0)	21 (6.56e0)	233

Table 1 Top half: overlap $\delta = 2h$. Bottom half: overlap $\delta = 8h$. Iteration count and condition number estimate for the channel distribution in Figure 1 for the classical multiscale coarse space, SHEM$_m$, $m = 1, 2, 3, 4$ and SHEM$_{\tau=6e-3}$ for $h = \frac{1}{128}$, $H = 16h$. Here 'dim' denotes the dimension of the coarse space.

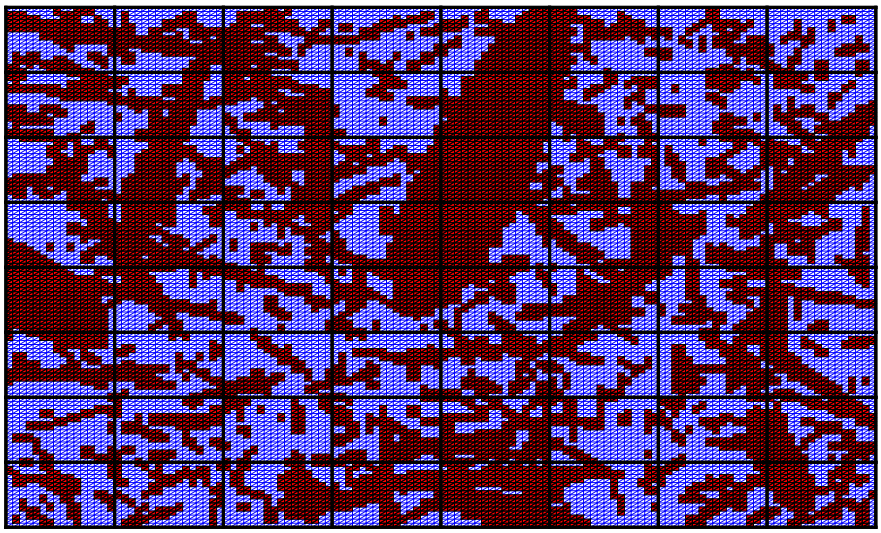


Fig. 2 Distribution of α for a geometry with $h = \frac{1}{128}$, $H = 16h$. The regions marked with red are where α has a large value $\hat{\alpha}$.

enrichment functions was chosen. Larger overlap can however also be taken into account by a different construction of SHEM for AS, see [9].

We next perform the same test also on the irregular high contrast structure shown in Figure 2. The corresponding results in Table 2 show that also in this case SHEM works very well with larger overlap, and that difficulties can be either remedied by increasing the overlap, or enriching the coarse space: SHEM with one enrichment function is enough to get robust convergence with large overlap, but with small overlap, SHEM needs 2-3 enrichment functions.

	MS	SHEM_1	SHEM_2	SHEM_3	SHEM_4	$\text{SHEM}_{\tau=6e-3}$	
dim.	49	161	273	385	497		
$\hat{\alpha}$	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)	dim.
10^0	21 (1.29e1)	16 (7.45e0)	15 (5.99e0)	13 (5.19e0)	13 (5.15e0)	21 (1.29e1)	49
10^2	72 (1.09e2)	53 (6.49e1)	27 (1.52e1)	22 (9.47e0)	20 (6.45e0)	36 (2.14e1)	165
10^4	288 (9.43e3)	98 (5.46e3)	29 (1.60e1)	23 (9.60e0)	21 (6.54e0)	38 (2.44e1)	169
10^6	524 (9.41e6)	156 (5.49e5)	32 (1.60e1)	24 (9.59e0)	22 (6.28e0)	39 (2.44e1)	169
10^0	16 (5.57e0)	15 (4,88e0)	15 (4.82e0)	15 (4.94e0)	15 (4.95e0)	16 (5.47e0)	49
10^2	29 (1.31e1)	22 (7.75e0)	19 (5.54e0)	18 (5.10e0)	18 (5.05e0)	22 (7.89e0)	165
10^4	72 (7.56e2)	28 (1.36e1)	20 (5.68e0)	19 (5.12e0)	19 (5.07e0)	25 (9.97e0)	169
10^6	121 (7.50e4)	32 (1.43e2)	21 (5.41e0)	20 (5.05e0)	20 (5.02e0)	26 (1.01e1)	169

Table 2 Top half: overlap $\delta = 2h$. Bottom half: overlap $\delta = 8h$. Iteration count and condition number estimate for the distribution in Figure 2 for the classical multiscale coarse space, SHEM_m, $m = 1, 2, 3, 4$ and $\text{SHEM}_{\tau=6e-3}$ for $h = \frac{1}{128}$, $H = 16h$. Here 'dim' denotes the dimension of the coarse space.

	MS	SHEM_1	SHEM_2	SHEM_3	SHEM_4
$\frac{H}{h}$	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)	#it. (κ)
8	18 (7.67e0)	14 (5.36e0)	14 (5.02e1)	14 (5.07e0)	13 (5.12e0)
16	21 (1.29e1)	16 (7.45e0)	15 (5.99e0)	13 (5.19e0)	13 (5.15e0)
32	29 (2.37e1)	20 (1.22e1)	18 (8.97e5)	15 (7.52e0)	14 (6.55e0)
64	41 (4.52e1)	26 (2.23e1)	22 (1.56e1)	19 (1.32e1)	18 (1.03e1)
128	58 (8.85e1)	36 (4.25e1)	30 (2.88e1)	25 (2.23e1)	23 (1.82e1)
256	80 (1.75e2)	50 (8.83e1)	41 (5.57e1)	34 (4.24e1)	31 (3.42e1)
16	367 (3.64e4)	248 (1.10e4)	187 (6.78e3)	19 (6.78e0)	17 (5.73e0)
32	525 (7.47e4)	326 (2.32e4)	252 (1.32e4)	22 (9.33e0)	19 (7.74e0)
64	740 (1.51e5)	458 (4.76e4)	329 (2.72e4)	28 (1.70e1)	22 (1.25e1)
128	1062 (3.05e5)	665 (9.62e4)	457 (5.52e4)	38 (3.15e1)	29 (2.25e1)
256	1522 (6.12e5)*	980 (1.94e5)*	679 (1.11e5)*	52 (6.06e1)	41 (4.28e1)
16	288 (9.43e3)	98 (5.46e3)	29 (1.60e1)	23 (9.60e0)	21 (6.54e0)
32	443 (1.97e4)	129 (1.14e4)	38 (2.75e1)	28 (1.53e0)	23 (8.00e0)
64	612 (4.03e4)	170 (2.31e4)	51 (5.07e1)	36 (2.73e1)	29 (1.27e1)
128	856 (8.17e4)	232 (4.65e4)	70 (9.82e1)	48 (5.20e1)	38 (2.26e1)
256	1207 (1.64e5)	315 (9.33e4)	98 (1.94e2)	66 (1.02e2)	52 (4.30e1)

* Stagnation.

Table 3 Top: $\alpha = 1$. Middle: Distribution of α from Figure 1 with $\hat{\alpha} = 10^4$. Bottom: Distribution of α from Figure 2 with $\hat{\alpha} = 10^4$. Iteration count and condition number estimate for the classical multiscale coarse space and SHEM_m, $m = 1, 2, 3, 4$, solving Problem 1 for decreasing h, $H = \frac{1}{8}$ and overlap $\delta = 2h$.

3.2 What is the condition number growth in H/h?

We now test numerically the dependence on the mesh ratio H/h for the case where $\alpha = 1$ and for the high contrast cases given in Figure 1 and 2 with $\hat{\alpha} = 10^4$. The iteration counts and condition number estimates are given in Table 3 for decreasing h while the subdomain diameter is kept fixed at $H = 1/8$. We clearly see that the convergence rate is linearly dependent on the mesh ratio H/h, for both the constant coefficient case and the high contrast cases. This confirms that the restrictions in the analysis in [9] are not a property of SHEM itself, but rather restrictions of the analysis. We also see that for very high contrast, SHEM can even fix stagnation when using the appropriate amount of enrichment.

4 Conclusions

The numerical experiments we presented indicate that the first convergence estimate for SHEM in Theorem 1 might not need the technical assumption of small overlap, and also that the convergence bound with the square dependence on the mesh ratio H/h is too pessimistic. Another important observation is that the dimension of the coarse space is not larger than the

dimension of the largest subdomain in our experiments, and thus the coarse space solve remains less expensive than the subdomain solves. Based on this numerical investigation, we are currently carefully studying the technical estimates in the proof of Theorem 1 to see under which conditions on the high contrast parameter α the overlap restriction and the quadratic dependence on the mesh ratio in the condition number estimate can be removed. We are also working on the extension to three dimensional problems, see [2], and on a parallel implementation.

References

1. Jørg Aarnes and Thomas Y. Hou. Multiscale domain decomposition methods for elliptic problems with high aspect ratios. *Acta Math. Appl. Sin. Engl. Ser.*, 18(1):63–76, 2002.
2. Erik Eikeland, Leszek Marcinkowski, and Talal Rahman. Overlapping Schwarz Methods with Adaptive Coarse Spaces for Multiscale Problems in 3D. *arXiv:1611.00968*, November 2016.
3. Martin J Gander and Laurence Halpern. Méthodes de décomposition de domaine. *Encyclopédie électronique pour les ingénieurs*, 2012.
4. Martin J. Gander, Laurence Halpern, and Kévin Santugini Repiquet. Discontinuous coarse spaces for DD-methods with discontinuous iterates. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 607–615. Springer, 2014.
5. Martin J. Gander, Laurence Halpern, and Kévin Santugini Repiquet. A new coarse grid correction for RAS/AS. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 275–283. Springer, 2014.
6. Martin J. Gander, Laurence Halpern, and Kévin Santugini Repiquet. On optimal coarse spaces for domain decomposition and their approximation. In *these proceedings (submitted)*. 2017.
7. Martin J. Gander and Atle Loneland. SHEM: An optimal coarse space for RAS and its multiscale approximation. In *Domain Decomposition Methods in Science and Engineering XXIII*. Springer, 2016.
8. Martin J Gander, Atle Loneland, and Talal Rahman. Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods. *arXiv preprint arXiv:1512.05285*, 2015.
9. Martin J. Gander and Bo Song. Complete, optimal and optimized coarse spaces for AS. In *these proceedings (submitted)*. 2017.
10. I. G. Graham, P. O. Lechner, and R. Scheichl. Domain decomposition for multiscale PDEs. *Numer. Math.*, 106(4):589–626, 2007.
11. A. Heinlein, A. Klawonn, J. Knepper, and O. Rheinbach. Multiscale coarse spaces for overlapping Schwarz methods based on the ACMS space in 2d. *ETNA*, page submitted, 2017.
12. Axel Klawonn, Patrick Radtke, and Oliver Rheinbach. Feti-dp methods with an adaptive coarse space. *SIAM Journal on Numerical Analysis*, 53(1):297–320, 2015.
13. Nicole Spillane, Victorita Dolean, Patrice Hauret, Frédéric Nataf, Clemens Pechstein, and Robert Scheichl. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numerische Mathematik*, 126(4):741–770, 2014.

Two-level preconditioners for the Helmholtz equation

Marcella Bonazzoli[1], Victorita Dolean[1,2], Ivan G. Graham[3], Euan A. Spence[3], and Pierre-Henri Tournier[4]

1 Introduction

Solving the Helmholtz equation $-\Delta u - k^2 u = f$ is a challenging task because of its indefinite nature and its highly oscillatory solution when the wavenumber k is high. Although there have been different attempts to solve it efficiently, we believe that there is no established and robust preconditioner, whose behavior is independent of k, for general decompositions into subdomains. In Conen et al. [2014] a two-level preconditioner was introduced, where the coarse correction involves local eigenproblems of Dirichlet-to-Neumann (DtN) maps. This method proved to be very robust with respect to heterogeneous coefficients compared to the reference preconditioner based on plane waves, and its construction is completely automatic without the need for parameter tuning. Another method was developed in Graham et al. [2017b,a], where two-level domain decomposition approximations of the Helmholtz equation with absorption $-\Delta u - (k^2 + \mathrm{i}\varepsilon)u = f$ were used as preconditioners for the pure Helmholtz equation without absorption; there the coarse correction is based on a coarse mesh with diameter constrained by k. Our purpose is to compare numerically the performance of the latter with the two-level method based on DtN maps, both in two and three dimensions.

2 Definition of the problem

Consider the interior Helmholtz problem of the following form: let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a polyhedral, bounded domain; find $u \colon \Omega \to \mathbb{C}$ such that

[1] Université Côte d'Azur, CNRS, LJAD, France, e-mail: marcella.bonazzoli@unice.fr
[2] University of Strathclyde, Glasgow, UK, e-mail: Victorita.Dolean@strath.ac.uk
[3] University of Bath, UK, e-mail: I.G.Graham@bath.ac.uk,E.A.Spence@bath.ac.uk
[4] UPMC Univ Paris 06, LJLL, Paris, France, e-mail: tournier@ljll.upmc.fr

$$-\Delta u - (k^2 + \mathtt{i}\varepsilon)u = f \qquad \text{in } \Omega, \tag{1a}$$

$$\frac{\partial u}{\partial n} - \mathtt{i}\eta u = 0 \qquad \text{on } \Gamma = \partial\Omega. \tag{1b}$$

The wavenumber k is given by $k(\mathbf{x}) = \omega/c(\mathbf{x})$, where ω is the angular frequency and c is the speed of propagation that might depend on $\mathbf{x} \in \Omega$; we take $\eta = \text{sign}(\varepsilon)k$ if $\varepsilon \neq 0$, $\eta = k$ if $\varepsilon = 0$,[R2] as Robin boundary condition parameter. We are interested in solving the problem in the case $\varepsilon = 0$, using ε as a parameter when building the preconditioner. The variational formulation of Problem (1) is: find $u \in V = H^1(\Omega)$ such that $a_\varepsilon(u, v) = F(v)$, $\forall v \in V$, where $a_\varepsilon(.,.) \colon V \times V \to \mathbb{C}$ and $F \colon V \to \mathbb{C}$ are defined by

$$a_\varepsilon(u, v) = \int_\Omega \left(\nabla u \cdot \overline{\nabla v} - (k^2 + \mathtt{i}\varepsilon)u\overline{v}\right) - \int_\Gamma \mathtt{i}\eta u\overline{v}, \quad F(v) = \int_\Omega f\overline{v}.$$

Note that if $\varepsilon \neq 0$ and $\eta = \text{sign}(\varepsilon)k$, a_ε is coercive (see §2 in Graham et al. [2017b]). We consider a discretization of the variational problem using piecewise linear finite elements on a uniform simplicial mesh $\mathcal{T}_h$ of Ω. Denoting by $V_h \subset V$ the corresponding finite element space and by $\{\phi_k\}_{k=1}^n$ its basis functions, $n := \dim(V_h)$, the discretized problem reads: find $u_h \in V_h$ such that $a_\varepsilon(u_h, v_h) = F(v_h)$, $\forall v_h \in V_h$, that is, in matrix form,

$$A_\varepsilon \mathbf{u} = \mathbf{f}, \tag{2}$$

where the coefficients of the matrix $A_\varepsilon \in \mathbb{C}^{n\times n}$ and the right-hand side $\mathbf{f} \in \mathbb{C}^n$ are given by $(A_\varepsilon)_{k,l} = a(\phi_l, \phi_k)$ and $(\mathbf{f})_k = F(\phi_k)$. The matrix A_ε is complex, symmetric (but not Hermitian), and indefinite if $\varepsilon = 0$.

3 Two-level domain decomposition preconditioners

In the following we will define the domain decomposition preconditioners for the linear system $A_0\mathbf{u} = \mathbf{f}$ resulting from the discretization of the Helmholtz problem without absorption ($\varepsilon = 0$). These are two-level Optimized Restricted Additive Schwarz (ORAS) algorithms, where "optimized" refers to the use of Robin boundary conditions at the interface between subdomains. In the terminology of Graham et al. [2017b], the prefix O is replaced with Imp, which stands for impedance (i.e. Robin) boundary conditions.

First of all, consider a decomposition of the domain Ω into a set of overlapping subdomains $\{\Omega_j\}_{j=1}^{N_{\text{sub}}}$, with each subdomain consisting of a union of elements of the mesh $\mathcal{T}_h$. Let $V_h(\Omega_j) = \left\{v|_{\Omega_j} : v \in V_h\right\}$, $1 \leq j \leq N_{\text{sub}}$, denote the space of functions in V_h restricted to the subdomain Ω_j. Let n_j be the dimension of $V_h(\Omega_j)$, $1 \leq j \leq N_{\text{sub}}$. ~~Let $n_j := \#\,\text{dof}\,(\Omega_j)$, $1 \leq j \leq N_{\text{sub}}$, where $\text{dof}(D) := \left\{k : \text{supp}\,(\phi_k) \subset \overline{D}\right\}$ represents the degrees of freedom (dofs) of $V_h(\Omega_j)$.~~[R1] For $1 \leq j \leq N_{\text{sub}}$, we define a restriction operator $\mathcal{R}_j \colon V_h \to$

$V_h(\Omega_j)$ by injection, i.e. for $u \in V_h$ we set $(\mathcal{R}_j u)(\mathbf{x}_i) = u(\mathbf{x}_i)$ for all $\mathbf{x}_i \in \Omega_j$. We denote by R_j the corresponding Boolean matrix in $\mathbb{R}^{n_j \times n}$ that maps coefficient vectors of functions in V_h to coefficient vectors of functions in $V_h(\Omega_j)$. Let $D_j \in \mathbb{R}^{n_j \times n_j}$ be a diagonal matrix corresponding to a partition of unity in the sense that $\sum_{i=1}^{N_{\text{sub}}} \tilde{R}_i^T R_i = I$, where $\tilde{R}_j := D_j R_j$. Then the *one-level ORAS* preconditioner (which is also the one-level ImpRAS ~~ImpHRAS~~[R2] of Graham et al. [2017b]) reads

$$M_{1,\varepsilon}^{-1} := \sum_{j=1}^{N_{\text{sub}}} \tilde{R}_j^T A_{j,\varepsilon}^{-1} R_j. \tag{3}$$

We define the matrices $A_{j,\varepsilon}$ in (3) to be the matrices stemming from the discretization of the following local Robin problems with absorption

$$\begin{aligned} -\Delta u_j - (k^2 + \mathtt{i}\varepsilon) u_j &= f && \text{in } \Omega_j, \\ \frac{\partial u_j}{\partial n_j} - \mathtt{i}\eta u_j &= 0 && \text{on } \partial\Omega_j. \end{aligned}$$

In order to achieve weak dependence on the wavenumber k and number of subdomains, we add a coarse component to (3). The *two-level* preconditioner can be written in a generic way as follows

$$M_{2,\varepsilon}^{-1} = Q M_{1,\varepsilon}^{-1} P + Z E^{-1} Z^*, \tag{4}$$

where $*$ denotes the conjugate transpose, $M_{1,\varepsilon}^{-1}$ is the one-level preconditioner (3), Z is a rectangular matrix with full column rank, $E = Z^* A_\varepsilon Z$ is the so-called coarse grid matrix, $\Xi = Z E^{-1} Z^*$ is the so-called coarse grid correction matrix. If $P = Q = I$ this is an *additive* two-level preconditioner (which would be called two-level ImpRAS in Graham et al. [2017b]). If $P = I - A_\varepsilon \Xi$ and $Q = I - \Xi A_\varepsilon$, this is a *hybrid* two-level preconditioner (ImpHRAS in Graham et al. [2017b]), also called the *Balancing* Neumann Neumann (BNN) preconditioner. Preconditioner (4) is characterized by the choice of Z, whose columns span the *coarse space* (CS). We will consider the following two cases:

The grid coarse space The most natural coarse space would be one based on a coarser mesh, we subsequently call it "grid coarse space". Let us consider $\mathcal{T}_{H_{\text{coarse}}}$ a simplicial mesh of Ω with mesh diameter H_{coarse} and $V_{H_{\text{coarse}}} \subset V$ the corresponding finite element space. Let $\mathcal{I}_0 \colon V_{H_{\text{coarse}}} \to V_h$ be the nodal interpolation operator and define Z as the corresponding matrix. Then ~~Let $\mathcal{R}_0 \colon V_h \to V_{H_{\text{coarse}}}$ be the nodal interpolation operator and R_0 the corresponding matrix. Define $Z = R_0^T$, then~~[R2] in this case $E = Z^* A_\varepsilon Z$ is really the stiffness matrix of the problem (with absorption) discretized on the coarse mesh. Related preconditioners without absorption are used in Kimn and Sarkis [2007].

The DtN coarse space This coarse space (see Conen et al. [2014]) is based on local Dirichlet-to-Neumann (DtN) eigenproblems on the subdomain interfaces. For a subdomain Ω_i, first of all consider $a^{(i)}\colon H^1(\Omega_i)\times H^1(\Omega_i)\to\mathbb{R}$

$$a^{(i)}(v,w)=\int_{\Omega_i}\left(\nabla v\cdot\overline{\nabla w}-(k^2+\mathtt{i}\varepsilon)v\overline{w}\right)-\int_{\partial\Omega_i\cap\partial\Omega}\mathtt{i}\eta u\overline{v}.$$

Let $(A^{(i)})_{kl}=a^{(i)}(\phi_k,\phi_l)$, and let I and Γ_i be the sets of indices corresponding, resp., to the interior and boundary dofs on Ω_i, with n_I and n_{Γ_i} their cardinalities. With the usual block notation, the subscripts I and Γ_i for the matrices A and $A^{(i)}$ denote the entries of these matrices associated with the respective dofs. Let $M_{\Gamma_i}=\left(\int_{\Gamma_i}\phi_k\phi_l\right)_{k,l\in\Gamma_i}$ be the mass matrix on the interface $\Gamma_i=\partial\Omega_i\setminus\partial\Omega$ of subdomain Ω_i. We need to solve the following eigenproblem: find $(\mathbf{u},\lambda)\in\mathbb{C}^{n_{\Gamma_i}}\times\mathbb{C}$, s.t.

$$(A^{(i)}_{\Gamma_i\Gamma_i}-A_{\Gamma_iI}A_{II}^{-1}A_{I\Gamma_i})\mathbf{u}=\lambda M_{\Gamma_i}\mathbf{u}. \tag{5}$$

Now, the matrix Z of the DtN coarse space is a rectangular, block-diagonal matrix with blocks W_i, associated with the subdomain Ω_i, $1\le i\le N_{\text{sub}}$, given by Algorithm 3.1. If m_i is the number of eigenvectors selected by the automatic criterion in Line 2 of Algorithm 3.1, the block W_i has dimensions $n_i\times m_i$, and the matrix Z has dimensions $n\times\sum_{j=1}^{N_{\text{sub}}}m_i$. Due to the overlap in the decomposition, the blocks may share some rows inside the matrix Z.

Algorithm 3.1 Construction of the block W_i of the DtN CS matrix Z

1: Solve the discrete DtN eigenproblem (5) on subdomain Ω_i for the eigenpairs $(\lambda_j,\mathbf{g}_i^j)$.
2: Choose the m_i eigenvectors $\mathbf{g}_i^j\in\mathbb{C}^{n_{\Gamma_i}}$ such that $\Re(\lambda_j)<k$, $1\le j\le m_i$.
3: **for** $j=1$ to m_i **do**
4: Compute the discrete Helmholtz extension $\mathbf{u}_i^j\in\mathbb{C}^{n_i}$ to Ω_i of $\mathbf{g}_i^j$ as $\mathbf{u}_i^j=[-A_{II}^{-1}A_{I\Gamma_i}\mathbf{g}_i^j,\quad\mathbf{g}_i^j]^T$.
5: **end for**
6: Define the matrix $W_i\in\mathbb{C}^{n_i\times m_i}$ as $W_i=\left(D_i\mathbf{u}_i^1,\ldots,D_i\mathbf{u}_i^{m_i}\right)$.

4 Numerical experiments

We solve (2) with $\varepsilon=0$ on the unit square/cube, with a uniform simplicial mesh of diameter $h\sim k^{-3/2}$, which is believed to remove the pollution effect. The right-hand side is given by $f=-\exp(-100((x-0.5)^2+(y-0.5)^2))$ for $d=2$, $f=-\exp(-400((x-0.5)^2+(y-0.5)^2+(z-0.5)^2))$ for $d=3$.

We use GMRES with right preconditioning (with a tolerance $\tau=10^{-6}$), starting with a *random initial guess*, which ensures, unlike a zero initial

guess, that all frequencies are present in the error;[R1] the stopping criterion is based on the relative residual. We consider a regular decomposition into subdomains (squares/cubes), the overlap for each subdomain is of size $\mathcal{O}(2h)$ in all directions and the two-level preconditioner (4) is used in the hybrid way. All the computations are done in the open source language FreeFem++ (`http://www.freefem.org/ff++/`). The $3d$ code is parallelized and run on the TGCC Curie supercomputer. We assign each subdomain to one processor. So in our experiments the number of processors increases if the number of subdomains increases. To apply the preconditioner, the local problems in each subdomain (with matrices $A_{j,\varepsilon}$ in (3)) and the coarse space problem (with matrix E in (4)) are solved with a direct solver.

As in Graham et al. [2017b,a], in the experiments we take the subdomain diameter H_{sub} and the coarse mesh diameter H_{coarse} constrained by k: $H_{\text{sub}} \sim k^{-\alpha}$ and $H_{\text{coarse}} \sim k^{-\alpha'}$, for some choices of $0 < \alpha, \alpha' <= 1$ detailed in the following; if not differently specified, we take $\alpha = \alpha'$, which is the setting of all numerical experiments in Graham et al. [2017b]. Note that H_{coarse} does not appear as a parameter in the DtN coarse space. We denote by n_{CS} the size of the coarse space. For the grid coarse space $n_{\text{CS}} = (1/H_{\text{coarse}} + 1)^d$, the number of dofs for the nodal linear finite elements in the unit square/cube. For the DtN coarse space $n_{\text{CS}} = \sum_{j=1}^{N_{\text{sub}}} m_i$, the total number of computed eigenvectors for all the subdomains. While we solve the pure Helmholtz problem without absorption, both the one-level preconditioner (3) and the two-level preconditioner (4) are built from problems which can have non zero absorption given by $\varepsilon_{\text{prec}} = k^\beta$. In the experiments we put $\beta = 1$ or $\beta = 2$.

In the following tables we compare the one-level preconditioner, the two-level preconditioners with the grid coarse space and with the DtN coarse space in terms of number of iterations of GMRES and size of the coarse space (n_{CS}), for different values of the wavenumber k and of the parameters α, β. We also report the number of subdomains N_{sub}, which is controlled by k and α as mentioned above. Since $h \sim k^{-3/2}$, the dimension n of the linear system matrix is of order $k^{3d/2}$; for 3d experiments we report n explicitly. Tables 1, 2 concern the $2d$ problem, Table 3 the $3d$ problem.

In Table 1, we let the DtN coarse space size be determined by the automatic choice criterion in Line 2 of Algorithm 3.1 (studied in Conen et al. [2014]) and the grid coarse space size by $H_{\text{coarse}} \sim k^{-\alpha}$. We see that the DtN coarse space is much larger than the grid coarse space and gives fewer iterations. The preconditioners with absorption $\varepsilon_{\text{prec}} = k^2$ perform much worse than those with absorption $\varepsilon_{\text{prec}} = k$ independently of n_{CS}. For $\varepsilon_{\text{prec}} = k$, when $\alpha = 1$ the number of iterations grows as $k^{0.9}$, respectively $k^{1.1}$, for the grid coarse space, respectively DtN coarse space (excluding the first two values for k small where the asymptotic behaviour is not reached yet) ~~the number of iterations grows mildly with the wavenumber k for both coarse spaces (but at the cost of an increasing coarse space size),~~[R2] while the one-level preconditioner performs poorly (for $k = 80$ it needs more than 500 iterations to converge)[R2]. When $\alpha < 1$, i.e. for coarser coarse meshes, the

		$\beta = 1$					$\beta = 2$				
		$\alpha = 0.6$					$\alpha = 0.6$				
k	N_{sub}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}
10	9	22	19	16	11	39	28	27	16	23	40
20	36	48	46	49	26	204	67	56	49	40	220
40	81	78	98	100	37	531	121	114	100	72	578
60	121	109	114	144	43	1037	169	165	144	109	920
80	169	139	138	196	93	1588	223	216	196	126	1824
		$\alpha = 0.8$					$\alpha = 0.8$				
k	N_{sub}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}
10	36	35	19	49	10	122	39	27	49	28	86
20	100	71	35	121	13	394	83	51	121	41	362
40	361	158	88	400	22	1440	182	95	400	71	1370
60	676	230	187	729	39	2700	268	150	729	103	2698
80	1089	304	331	1156	68	4352	355	214	1156	138	4350
		$\alpha = 1$					$\alpha = 1$				
k	N_{sub}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}
10	100	65	26	121	11	324	57	30	121	23	324
20	400	122	26	441	14	1120	130	49	441	42	1120
40	1600	286	33	1681	20	4640	296	80	1681	72	4640
60	3600	445	45	3721	29	10560	455	112	3721	101	10560
80	6400	>500	62	6561	44	18880	>500	149	6561	134	18880

Table 1: ($d = 2$) Number of iterations (and coarse space size n_{CS}) for the one-level preconditioner and the two-level preconditioners with the grid coarse space/DtN coarse space, with $H_{\text{sub}} = H_{\text{coarse}} \sim k^{-\alpha}$, $\varepsilon_{\text{prec}} = k^{\beta}$.

growth with k is higher, and for $\alpha = 0.6$ the two-level preconditioner is not much better than the one-level preconditioner because the coarse grid problem is too coarse; for $\alpha = 0.8$ with the DtN coarse space the growth with k degrades less than with the grid coarse space.

We have seen in Table 1 that the DtN coarse space gives fewer iterations than the grid coarse space, but their sizes differed significantly. Therefore, in Table 2 we compare the two methods forcing n_{CS} to be similar. On the left, we force the DtN coarse space to have a smaller size, similar to the one of the grid coarse space, by taking just $m_i = 2$ eigenvectors for each subdomain. On the right, we do the opposite, we force the grid coarse space to have the size of the DtN coarse space obtained in Table 1, by prescribing a smaller coarse mesh diameter H_{coarse}, while keeping the same number of subdomains as in Table 1 with $H_{\text{sub}} \sim k^{-\alpha}$. We can observe that for smaller coarse space sizes (left) the grid coarse space gives fewer iterations than the DtN coarse space, while for larger coarse space sizes (right) the result is reversed.

We have seen that the coarse mesh obtained with $H_{\text{coarse}} \sim k^{-\alpha'}$, $\alpha' = \alpha$ can be too coarse if $\alpha = 0.6$. At the same time, for $\alpha = 1$ the number of subdomains gets quite large since $H_{\text{sub}} \sim k^{-\alpha}$, especially in $3d$; this is not desirable because in our parallel implementation we assign each subdomain to one processor, so communication among them would prevail and each

		n_{CS} forced by grid CS				n_{CS} forced by DtN CS			
		$\alpha = 0.6$				$\alpha = 0.6$			
k	N_{sub}	grid CS	n_{CS}	DtN CS	n_{CS}	grid CS	n_{CS}	DtN CS	n_{CS}
10	9	19	16	18	18	17	36	11	39
20	36	46	49	44	72	24	196	26	204
40	81	98	100	85	162	50	529	37	531
60	121	114	144	109	242	104	841	43	1037
80	169	138	196	140	338	173	1521	93	1588
		$\alpha = 0.8$				$\alpha = 0.8$			
k	N_{sub}	grid CS	n_{CS}	DtN CS	n_{CS}	grid CS	n_{CS}	DtN CS	n_{CS}
10	36	19	49	26	72	15	121	10	122
20	100	35	121	61	200	20	361	13	394
40	361	88	400	139	722	35	1369	22	1440
60	676	187	729	191	1352	52	2601	39	2700
80	1089	331	1156	250	2178	78	4225	68	4352
		$\alpha = 1$				$\alpha = 1$			
k	N_{sub}	grid CS	n_{CS}	DtN CS	n_{CS}	grid CS	n_{CS}	DtN CS	n_{CS}
10	100	26	121	52	200	17	324	11	324
20	400	26	441	43	800	23	1089	14	1120
40	1600	33	1681	157	3200	22	4624	20	4640
60	3600	45	3721	338	7200	26	10404	29	10560
80	6400	62	6561	>500	12800	30	18769	44	18880

Table 2: ($d = 2$) Number of iterations (and coarse space size n_{CS}) for the two-level preconditioners with the grid coarse space/DtN coarse space *forcing similar* n_{CS}, with $H_{sub} \sim k^{-\alpha}$, $\varepsilon_{prec} = k$.

processor would not be fully exploited since the subdomains would become very small. Therefore, to improve convergence with the grid coarse space while maintaining a reasonable number of subdomains, we consider separate H_{coarse} and H_{sub}, taking $\alpha' \neq \alpha$. For load balancing (meant as local problems having the same size as the grid coarse space problem), in $3d$ we choose $\alpha' = 3/2 - \alpha$. The DtN coarse space is still built by keeping the eigenvectors verifying the automatic choice criterion; note that in $3d$ the number of selected eigenvectors is larger than in $2d$, but we only keep a maximum of 20 eigenvectors in each subdomain. ~~The DtN coarse space size is still determined by the automatic choice criterion (among 20 computed local eigenvectors) in each subdomain.~~[R2] In Table 3 we report the results of this experiment. As expected, for the grid coarse space the best iteration counts are obtained for $\alpha = 0.5$ because then $\alpha' = 1$ gives the coarse mesh with the smallest diameter among the experimented ones: the number of iterations grows slowly, with $\mathcal{O}(k^{0.61}) \cong \mathcal{O}(n^{0.13})$. With higher α the iteration counts get worse quickly, and $\alpha = 0.8$ is not usable. For the DtN coarse space, the larger coarse space size is obtained by taking α bigger (recall that α' is not a parameter in the DtN case): for $\alpha = 0.8$ the number of iterations grows slowly, with $\mathcal{O}(k^{0.2}) \cong \mathcal{O}(n^{0.04})$, but this value may be optimistic, there is a decrease in iteration number between $k = 20$ and 30. We believe that for the other

			$\alpha = 0.5$, $\alpha' = 1$				
k	n	N_{sub}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}
10	39304	27	25	12	1331	14	316
20	704969	64	39	17	9261	31	1240
30	5000211	125	55	21	29791	54	2482
40	16194277	216	74	29	68921	80	4318
			$\alpha = 0.6$, $\alpha' = 0.9$				
k	n	N_{sub}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}
10	39304	27	25	15	512	14	316
20	912673	216	61	24	3375	41	2946
30	4826809	343	73	34	10648	65	6226
40	16194277	729	98	48	21952	108	13653
			$\alpha = 0.7$, $\alpha' = 0.8$				
k	n	N_{sub}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}
10	46656	125	34	19	343	11	896
20	912673	512	73	35	1331	18	4567
30	5929741	1000	103	57	4096	65	12756
40	17779581	2197	139	89	8000	116	30603
			$\alpha = 0.8$, $\alpha' = 0.7$				
k	n	N_{sub}	1-level	grid CS	n_{CS}	DtN CS	n_{CS}
10	50653	216	39	23	216	19	1354
20	1030301	1000	46	86	729	23	7323
30	5929741	3375	137	116	1331	21	26645
40	28372625	6859	189	200	2744	27	54418

Table 3: ($d = 3$) Number of iterations (and coarse space size n_{CS}) for the one-level preconditioner and the two-level preconditioners with the grid coarse space/DtN coarse space, with $H_{\text{sub}} \sim k^{-\alpha}$, $H_{\text{coarse}} \sim k^{-\alpha'}$, $\varepsilon_{\text{prec}} = k$.

values of α, where the iteration counts are not much better or worse than with the one-level preconditioner, we did not compute enough eigenvectors in each subdomain to build the DtN coarse space.

5 Conclusion

We tested numerically two different coarse space definitions for two-level domain decomposition preconditioners for the pure Helmholtz equation (discretized with piecewise linear finite elements), both in $2d$ and $3d$, reaching more than 15 million degrees of freedom in the resulting linear systems. The preconditioners built with absorption $\varepsilon_{\text{prec}} = k^2$ appear to perform much worse than those with absorption $\varepsilon_{\text{prec}} = k$. We have seen that in most cases for smaller coarse space sizes the grid coarse space gives fewer iterations than the DtN coarse space, while for larger coarse space sizes the grid coarse space gives generally more iterations than the DtN coarse space. The best iteration counts for the grid coarse space are obtained by separating the coarse

mesh diameter $H_{\text{coarse}} \sim k^{-\alpha'}$ from the subdomain diameter $H_{\text{sub}} \sim k^{-\alpha}$, taking $\alpha' > \alpha$. Both for the grid coarse space ~~the coarse grid space~~[R2] and the DtN coarse space, for appropriate choices of the method parameters we have obtained iteration counts which grow quite slowly with the wavenumber k. Further experiments to compare the two coarse spaces ~~the two definitions of coarse space~~[R2] should be carried out in the heterogenous case.

Acknowledgement This work has been supported in part by the French National Research Agency (ANR), project MEDIMAX, ANR-13-MONU-0012.

References

L. Conen, V. Dolean, R. Krause, and F. Nataf. A coarse space for heterogeneous Helmholtz problems based on the Dirichlet-to-Neumann operator. *J. Comput. Appl. Math.*, 271:83–99, 2014.

I. G. Graham, E. A. Spence, and E. Vainikko. *Recent Results on Domain Decomposition Preconditioning for the High-Frequency Helmholtz Equation Using Absorption*, pages 3–26. Geosystems Mathematics. Springer, 2017a.

I. G. Graham, E. A. Spence, and E. Vainikko. Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption. *Math. Comp.*, 86(307):2089–2127, 2017b.

J.-H. Kimn and M. Sarkis. Restricted overlapping balancing domain decomposition methods and restricted coarse problems for the Helmholtz problem. *Comput. Methods Appl. Mech. Engrg.*, 196(8):1507–1514, 2007.

A two-level domain-decomposition preconditioner for the time-harmonic Maxwell's equations

Marcella Bonazzoli[1], Victorita Dolean[1,2], Ivan G. Graham[3], Euan A. Spence[3], and Pierre-Henri Tournier[4]

1 Introduction

The construction of fast iterative solvers for the indefinite time-harmonic Maxwell's system at mid- to[R2] high-frequency is a problem of great current interest. Some of the difficulties that arise are similar to those encountered in the case of the mid- to[R2] high-frequency Helmholtz equation. Here we investigate how domain-decomposition (DD) solvers recently proposed for the ~~high-frequency~~[R2] Helmholtz equation work in the Maxwell case.

The idea of preconditioning discretisations of the Helmholtz equation with discretisations of the corresponding problem with absorption was introduced in Erlangga et al. [2004]. In Graham et al. [2017a], a two-level domain-decomposition method was proposed that uses absorption, along with a wavenumber dependent coarse space correction. Note that, in this method, the choice of absorption is motivated by the analysis in both Graham et al. [2017a] and the earlier work Gander et al. [2015].

Our aim is to extend these ideas to the time-harmonic Maxwell's equations, both from the theoretical and numerical points of view. These results will appear in full in the forthcoming paper Bonazzoli et al. [2017].

Our theory will apply to the boundary value problem (BVP)

$$\begin{cases} \nabla \times (\nabla \times \mathbf{E}) - (k^2 + \mathrm{i}\kappa)\mathbf{E} = \mathbf{J} & \text{in } \Omega \\ \mathbf{E} \times \mathbf{n} = \mathbf{0} & \text{on } \Gamma := \partial\Omega \end{cases} \tag{1}$$

where Ω is a bounded Lipschitz polyhedron in $\mathbb{R}^3$ with boundary Γ and outward-pointing unit normal vector $\mathbf{n}$, k is the wave number, and $\mathbf{J}$ is the source term. The PDE in (1) is obtained from Maxwell's equations by as-

[1] Université Côte d'Azur, CNRS, LJAD, France, e-mail: marcella.bonazzoli@unice.fr
[2] University of Strathclyde, Glasgow, UK, e-mail: Victorita.Dolean@strath.ac.uk
[3] University of Bath, UK, e-mail: I.G.Graham@bath.ac.uk,E.A.Spence@bath.ac.uk
[4] UPMC Univ Paris 06, LJLL, Paris, France, e-mail: tournier@ljll.upmc.fr

suming that the electric field $\boldsymbol{\mathcal{E}}$ is of the form $\boldsymbol{\mathcal{E}}(\mathbf{x},t) = \Re(\mathbf{E}(\mathbf{x})\mathrm{e}^{-\mathrm{i}\omega t})$, where $\omega > 0$ is the angular frequency. The boundary condition in (1) is called Perfect Electric Conductor (PEC) boundary condition. The parameter κ dictates the absorption/damping in the problem; in the case of a conductive medium, $\kappa = k\sigma Z$, where σ is the electrical conductivity of the medium and Z the impedance. If $\sigma = 0$, the solution is not unique for all $k > 0$ but a sufficient condition for existence of a solution is $\nabla \cdot \mathbf{J} = 0$.

We will also give numerical experiments for the BVP (1) where the PEC boundary condition is replaced by an impedance boundary condition, i.e. the BVP

$$\begin{cases} \nabla \times (\nabla \times \mathbf{E}) - (k^2 + \mathrm{i}\kappa)\mathbf{E} = \mathbf{J} & \text{in } \Omega \\ (\nabla \times \mathbf{E}) \times \mathbf{n} - \mathrm{i}\, k\, \mathbf{n} \times (\mathbf{E} \times \mathbf{n}) = \mathbf{0} & \text{on } \Gamma := \partial\Omega \end{cases} \tag{2}$$

In contrast to the PEC problem, the solution of the impedance problem is unique for every $k > 0$. There is large interest in solving (1) and (2) *both* when $\kappa = 0$ *and* when $\kappa \neq 0$. We will consider both these cases, in each case constructing preconditioners by using larger values of κ. Indeed, a higher level of absorption makes the problems involved in the preconditioner definition more "elliptic" (in a sense more precisely explained in Bonazzoli et al. [2017]), thus easier to solve. Note that the absorption cannot increase too much, otherwise the problem in the preconditioner is "too far away" from the initial problem.[R1]

2 Variational formulation and discretisation

Let $H_0(\mathrm{curl};\Omega) := \{\mathbf{v} \in L^2(\Omega), \nabla \times \mathbf{v} \in L^2(\Omega), \mathbf{v} \times \mathbf{n} = \mathbf{0}\}$. We introduce the k-weighted inner product on $H_0(\mathrm{curl};\Omega)$:

$$(\mathbf{v},\mathbf{w})_{\mathrm{curl},k} \;=\; (\nabla \times \mathbf{v}, \nabla \times \mathbf{w})_{L^2(\Omega)} + k^2 (\mathbf{v},\mathbf{w})_{L^2(\Omega)}.$$

The standard variational formulation of (1) is: Given $\mathbf{J} \in L^2(\Omega)$, $\kappa \in \mathbb{R}$ and $k > 0$, find $\mathbf{E} \in H_0(\mathrm{curl};\Omega)$ such that

$$a_\kappa(\mathbf{E},\mathbf{v}) = F(\mathbf{v}) \quad \text{for all } \mathbf{v} \in H_0(\mathrm{curl};\Omega), \tag{3}$$

where

$$a_\kappa(\mathbf{E},\mathbf{v}) := \int_\Omega \nabla \times \mathbf{E} \cdot \overline{\nabla \times \mathbf{v}} - (k^2 + \mathrm{i}\kappa) \int_\Omega \mathbf{E} \cdot \overline{\mathbf{v}} \tag{4}$$

and $F(\mathbf{v}) := \int_\Omega \mathbf{J} \cdot \overline{\mathbf{v}}$. When $\kappa > 0$, it is well-known that the sesquilinear form is coercive (see, e.g., Bonazzoli et al. [2017] and the references therein) and so existence and uniqueness follow from the Lax–Milgram theorem.

Nédélec edge elements are particularly suited for the approximation of electromagnetic fields. They provide a conformal discretisation of $H(\mathrm{curl},\Omega)$, since their tangential component across faces shared by adjacent tetrahedra

of a simplicial mesh $\mathcal{T}^h$ is continuous. We therefore define our approximation space $\mathcal{V}^h \subset H_0(\mathrm{curl};\Omega)$ as the lowest-order edge finite element space on the mesh $\mathcal{T}^h$ with functions whose tangential trace is zero on Γ. More precisely, over each tetrahedron τ, we write the discretised field as $\mathbf{E}_h = \sum_{e\in\tau} c_e \mathbf{w}_e$, a linear combination with coefficients c_e of the basis functions $\mathbf{w}_e$ associated with the edges e of τ, and the coefficients c_e will be the unknowns of the resulting linear system. The Galerkin method applied to the variational problem (3) is

$$\text{find } \mathbf{E}_h \in \mathcal{V}^h \;\text{ such that }\; a_\kappa(\mathbf{E}_h, \mathbf{v}_h) = F(\mathbf{v}_h) \;\text{ for all }\; \mathbf{v}_h \in \mathcal{V}^h. \tag{5}$$

The Galerkin matrix A_κ is defined by $(A_\kappa)_{ij} := a_\kappa(\mathbf{w}_{e_i}, \mathbf{w}_{e_j})$ and the Galerkin method is then equivalent to solving the linear system $A_\kappa \mathbf{U} = \mathbf{F}$, where $F_i := F(\mathbf{w}_{e_i})$ and $U_j := c_{e_j}$.

3 Domain decomposition

To define appropriate subspaces of $\mathcal{V}^h$, we start with a collection of open subsets $\{\widetilde{\Omega}_\ell : \ell = 1,\dots,N\}$ of $\mathbb{R}^d$ of maximum diameter H_{sub} that form an overlapping cover of $\overline{\Omega}$, and we set $\Omega_\ell = \widetilde{\Omega}_\ell \cap \overline{\Omega}$. Each $\overline{\Omega}_\ell$ is assumed to be non-empty and is assumed to consist of a union of elements of the mesh $\mathcal{T}_h$. Then, for each $\ell = 1,\dots,N$, we set

$$\mathcal{V}_\ell := \mathcal{V}^h \cap H_0(\mathrm{curl}, \Omega_\ell),$$

where $H_0(\mathrm{curl},\Omega_\ell)$ is considered as a subset of $H_0(\mathrm{curl};\Omega)$ by extending functions in $H_0(\mathrm{curl},\Omega_\ell)$ by zero, thus the tangential traces of elements of $\mathcal{V}_\ell$ vanish on the internal boundary $\partial\Omega_\ell\backslash\Gamma$ (as well as on $\partial\Omega_\ell \cap \Gamma$). Thus a solve of the Maxwell problem (3) in the space $\mathcal{V}_\ell$ involves a PEC boundary condition on $\partial\Omega_\ell$ (including any external parts of $\partial\Omega_\ell$). When $\kappa \neq 0$, such solves are always well-defined by uniqueness of the solution of the BVP (1).

Let $\mathcal{I}^h$ be the set of interior edges of elements of the triangulation; this set can be identified with the degrees of freedom of $\mathcal{V}^h$. Similarly, let $\mathcal{I}^h(\Omega_\ell)$ be the set of edges of elements contained in (the interior of) Ω_ℓ (corresponding to degrees of freedom on those edges). We then have that $\mathcal{I}^h = \cup_{\ell=1}^N \mathcal{I}^h(\Omega_\ell)$. For $e \in \mathcal{I}^h(\Omega_\ell)$ and $e' \in \mathcal{I}^h$, we define the restriction matrices $(R_\ell)_{e,e'} := \delta_{e,e'}$. We will assume that we have matrices $(D_\ell)_{\ell=1}^N$ satisfying

$$\sum_{\ell=1}^N R_\ell^T D_\ell R_\ell = I; \tag{6}$$

such matrices $(D_\ell)_{\ell=1}^N$ are called a *partition of unity*.

For two-level methods we need to define a coarse space. Let $\{\mathcal{T}^H\}$ be a sequence of shape-regular, tetrahedral meshes on $\overline{\Omega}$, with mesh diameter H. We assume that each element of $\mathcal{T}^H$ consists of the union of a set of fine grid elements. Let $\mathcal{I}^H$ be an index set for the coarse mesh edges. The coarse basis functions $\{\mathbf{w}_e^H\}$ are taken to be Nédélec edge elements on $\mathcal{T}^H$ with zero tangential traces on Γ. From these functions we define the coarse space $\mathcal{V}_0 := \text{span}\{\mathbf{w}_{e_p}^H : p \in \mathcal{I}^H\}$, and we define the "restriction matrix"

$$(R_0)_{pj} := \psi_{e_j}(\mathbf{w}_{e_p}^H) = \int_{e_j} \mathbf{w}_{e_p}^H \cdot \mathbf{t}, \quad j \in \mathcal{I}^h, \quad p \in \mathcal{I}^H, \tag{7}$$

where ψ_e are the degrees of freedom on the fine mesh.

With the restriction matrices $(R_\ell)_{\ell=0}^N$ defined above, we define

$$A_{\kappa,\ell} \;:=\; R_\ell A_\kappa R_\ell^T, \quad \ell = 0, \ldots, N$$

For $\ell = 1, \ldots, N$, the matrix $A_{\kappa,\ell}$ is then just the minor of A_κ corresponding to rows and columns taken from $\mathcal{I}^h(\Omega_\ell)$. That is $A_{\kappa,\ell}$ corresponds to the Maxwell problem on Ω_ℓ with homogeneous PEC boundary condition on $\partial\Omega_\ell \backslash \Gamma$. The matrix $A_{\kappa,0}$ is the Galerkin matrix for the problem (1) discretised in $\mathcal{V}_0$. In a similar way as for the global problem it can be proven that matrices $A_{\kappa,\ell}$, $\ell = 0, \ldots, N$, are invertible for all mesh sizes h and all choices of $\kappa \neq 0$.

In this paper we consider two-level preconditioners, i.e. those involving both local and coarse solves, except if '1-level' is specified in the numerical experiments. The classical *two-level Additive Schwarz* (AS) and *Restricted Additive Schwarz* (RAS) preconditioners for A_κ are defined by

$$M_{\kappa,\text{AS}}^{-1} := \sum_{\ell=0}^{N} R_\ell^T A_{\kappa,\ell}^{-1} R_\ell \quad M_{\kappa,\text{RAS}}^{-1} := \sum_{\ell=0}^{N} R_\ell^T D_\ell A_{\kappa,\ell}^{-1} R_\ell. \tag{8}$$

In the numerical experiments we will also consider two other preconditioners: (i) $M_{\kappa,\text{ImpRAS}}^{-1}$, which is similar to $M_{\kappa,\text{RAS}}^{-1}$, but the solves with $A_{\kappa,\ell}$ are replaced by solves with matrices corresponding to the Maxwell problem on Ω_ℓ with homogeneous impedance boundary condition on $\partial\Omega_\ell \backslash \Gamma$, and (ii) the hybrid version of RAS

$$M_{\kappa,\text{HRAS}}^{-1} := (I - \Xi A_\kappa)\left(\sum_{\ell=1}^{N} R_\ell^T D_\ell A_{\kappa,\ell}^{-1} R_\ell\right)(I - A_\kappa \Xi) + \Xi, \; \Xi = R_0^T A_{\kappa,0}^{-1} R_0. \tag{9}$$

In a similar manner we can define $M_{\kappa,\text{HAS}}^{-1}$, $M_{\kappa,\text{ImpHRAS}}^{-1}$, the hybrid versions of AS and ImpRAS.

4 Theoretical results

The following result is the Maxwell-analogue of the Helmholtz-result in [Graham et al., 2017b, Theorem 5.6] and appears in Bonazzoli et al. [2017]. We state a version of this result for $\kappa \sim k^2$, but note that Bonazzoli et al. [2017] contains a more general result that, in particular, allows for smaller values of the absorption κ.

Theorem 1 (GMRES convergence for left preconditioning with $\kappa \sim k^2$). *Assume that Ω is a convex polyhedron. Let C_k be the matrix representing the $(\cdot,\cdot)_{curl,k}$ inner product on the finite element space $\mathcal{V}^h$ in the sense that if $v_h, w_h \in \mathcal{V}^h$ with coefficient vectors $\mathbf{V}, \mathbf{W}$ then*

$$(v_h, w_h)_{curl,k} \;=\; \langle \mathbf{V}, \mathbf{W} \rangle_{C_k}. \tag{10}$$

Consider the weighted GMRES method where the residual is minimised in the norm induced by C_k. Let $\mathbf{r}^m$ denote the mth residual of GMRES applied to the system A_κ, left preconditioned with $M^{-1}_{\kappa,AS}$. Then

$$\frac{\|\mathbf{r}^m\|_{C_k}}{\|\mathbf{r}^0\|_{C_k}} \;\lesssim\; \left(1 - \left(1 + \left(\frac{H}{\delta}\right)^2\right)^{-2}\right)^{m/2}, \tag{11}$$

provided the following condition holds:

$$\max\{kH_{sub},\; kH\} \;\leq\; \mathcal{C}_1 \left(1 + \left(\frac{H}{\delta}\right)^2\right)^{-1}. \tag{12}$$

where H_{sub} and H are the typical diameters of a subdomain and of the coarse grid, δ denotes the size of the overlap, and $\mathcal{C}_1$ is a constant independent of all parameters.

As a particular example we see that, provided $\kappa \sim k^2$, $H \sim H_{\text{sub}} \sim k^{-1}$ and $\delta \sim H$ ("generous overlap"), then GMRES will converge with a number of iterations independent of all parameters. This property is illustrated in the numerical experiments in the next section. A result analogous to Theorem 1 for right-preconditioning appears in Bonazzoli et al. [2017].

5 Numerical results

In this section we will perform several numerical experiments in a cube domain with PEC boundary conditions (Experiments 1-2) or impedance boundary conditions (Experiments 3-4). The right-hand side is given by $\mathbf{J} = [f, f, f]$, where $f = -\exp(-400((x-0.5)^2 + (y-0.5)^2 + (z-0.5)^2))$.

We solve the linear system with GMRES with right preconditioning, starting with a *random initial guess*, which ensures, unlike a zero initial guess, that all frequencies are present in the error; the stopping criterion, with a tolerance of 10^{-6}, is based on the relative residual. The maximum number of iterations allowed is 200. We consider a regular decomposition into subdomains (cubes), the overlap for each subdomain is of size $\mathcal{O}(2h)$ (except in Experiment 1, where we take generous overlap) in all directions. All the computations are done in FreeFem++, an open source domain specific language (DSL) specialised for solving BVPs with variational methods (`http://www.freefem.org/ff++/`). The code is parallelised and run on the TGCC Curie supercomputer and the CINES Occigen supercomputer. We assign each subdomain to one processor. Thus in our experiments the number of processors increases if the number of subdomains increases. To apply the preconditioner, the local problems in each subdomain and the coarse space problem are solved with a direct solver (MUMPS on one processor). In all the experiments the fine mesh diameter is $h \sim k^{-3/2}$, which is believed to remove the pollution effect.

In our experiments we will often choose $H_{\text{sub}} \sim H$ and our preconditioners are thus determined by choices of H and κ, which we denote by H_{prec} and κ_{prec}. The absorption parameter of the problem to be solved is denoted κ_{prob}. The coarse grid problem is of size $\sim H_{\text{prec}}^{-2}$ and there are $\sim H_{\text{prec}}^{-2}$ local problems of size $(H_{\text{prec}}/h)^2$ (case $H_{\text{sub}} \sim H$). In the tables of results, n denotes the size of the system being solved, n_{CS} the size of the coarse space, the figures in the tables denote the GMRES iterations corresponding to a given method (e.g. #AS is the number of iterations for the AS preconditioner), whereas Time denotes the total time (in seconds) including both setup and GMRES solve times[R2]. For some of the experiments we compute (by linear least squares) the approximate value of γ so that the entries of this column grow with k^γ. We also compute ξ so that the entries of the column grow with n^ξ (here $\xi = \gamma \cdot 2/9$, because $n \sim (h^{3/2})^3 = k^{9/2}$).

Experiment 1. The purpose of this experiment is to test the theoretical result which says that even with AS (i.e. when solving PEC local problems), provided $H \sim H_{\text{sub}} \sim k^{-1}$, $\delta \sim H$ (generous overlap), $\kappa_{\text{prob}} = \kappa_{\text{prec}} = k^2$, the number of GMRES iterations should be bounded as k increases. In Table 1 we compare three two-level preconditioners: additive Schwarz, restricted additive Schwarz, and the hybrid version of restricted additive Schwarz. Note that in theory we would expect AS to be eventually robust, although its inferiority compared to the other methods is to be expected Graham et al. [2017a].

Experiment 2. In this experiment (Table 2) we set $\kappa_{\text{prob}} = \kappa_{\text{prec}} = k^2$ and $H \sim H_{\text{sub}} \sim k^{-0.8}$ and the overlap is $\mathcal{O}(2h)$ in all directions. As we are not in the case $H_{\text{prec}} \sim k^{-1}$ and we do not have generous overlap, we do not expect a bounded number of iterations here. Nevertheless, the method still performs well. Not surprisingly, the best method is ImpHRAS, as better

k	n	N_{sub}	n_{CS}	#AS	#RAS	#HRAS
10	4.6×10^5	1000	7.9×10^3	53	26	12
15	1.5×10^6	3375	2.6×10^4	59	28	12
20	1.2×10^7	8000	6.0×10^4	76	29	17

Table 1 $\delta \sim H$ (generous overlap), $H \sim H_{\text{sub}} \sim k^{-1}$, $\kappa_{\text{prob}} = \kappa_{\text{prec}} = k^2$.

transmission conditions at the interfaces between subdomains are used in the preconditioner. It is important to note that the time is growing very much slower than the dimension of the problem being solved.

k	n	N_{sub}	n_{CS}	#RAS (#HRAS)	#ImpRAS (#ImpHRAS)	Time ImpHRAS
10	3.4×10^5	216	1.9×10^3	34 (23)	27 (20)	11.0
20	7.1×10^6	1000	7.9×10^3	43 (31)	35 (28)	42.6
30	4.1×10^7	3375	2.6×10^4	47 (34)	39 (32)	100.9
40	1.3×10^8	6859	5.1×10^4	49 (36)	42 (35)	264.5
γ	4.5					2.23

Table 2 $\delta \sim 2h$, $H \sim H_{\text{sub}} \sim k^{-0.8}$, $\kappa_{\text{prob}} = \kappa_{\text{prec}} = k^2$.

Experiment 3 In this case we take $\kappa_{\text{prob}} = k$. Moreover, we take impedance boundary conditions on $\partial\Omega$. We take $H \sim H_{\text{sub}} \sim k^{-\alpha}$, $\kappa_{\text{prec}} = k^{\beta}$, and we use ImpHRAS as a preconditioner.

	$\alpha = 0.6$				$\alpha = 0.8$			
k	n	N_{sub}	n_{CS}	#2-level	n	N_{sub}	n_{CS}	#2-level
10	2.6×10^5	27	2.8×10^2	31	3.4×10^5	216	1.8×10^3	29
20	6.3×10^6	216	1.9×10^3	87	7.1×10^6	1000	7.9×10^3	60
30	3.3×10^7	343	2.9×10^3	148	4.1×10^7	3375	2.5×10^4	90
40	1.1×10^8	729	5.9×10^3	200	1.3×10^8	6859	5.1×10^4	154

				$\beta = 1$	$\beta = 2$
k	n	N_{sub}	n_{CS}	#2-level(Time)	#2-level(Time)
10	3.4×10^5	216	1.8×10^3	29 (12.9)	37 (13.1)
20	7.1×10^6	1000	7.9×10^3	60 (63.7)	70 (69.8)
30	4.1×10^7	3375	2.5×10^4	90 (200.4)	101 (221.2)
40	1.3×10^8	6859	5.1×10^4	154 (771.7)	137 (707.6)
γ	4.5		2.4	1.2 (2.9)	0.94 (2.8)
ξ	1.0		0.5	0.3 (0.6)	0.2 (0.6)

Table 3 $\kappa_{\text{prob}} = k$, $\delta \sim 2h$, $H \sim H_{\text{sub}} \sim k^{-\alpha}$, $\kappa_{\text{prec}} = k^{\beta}$; Top:~~Left:~~[R2] $\beta = 2, \alpha = 0.6, 0.8$; Bottom:~~Right:~~[R2] $\alpha = 0.8$, $\beta = 1, 2$.

In Table 3 on the bottom we see that the dimension of the coarse space is

$$n_{\text{CS}} = (k^{-0.8})^{-3} = k^{2.4} = \mathcal{O}(n^{0.5}).$$

This is reflected in the γ and ξ figures in the n_{CS} column. For this method the reduction factor n_{CS}/n is substantial (about 3.9×10^{-4} when $k = 40$). The computation time grows only slightly faster than the dimension of the coarse space, showing (a) weak scaling and (b) MUMPS is still performing close to optimally for Maxwell systems of size 5×10^4. Iteration numbers are growing with about $n^{0.3}$ at worst. Note that the iteration numbers may be improved by separating the coarse grid size from the subdomain size, making the coarse grid finer and the subdomains bigger.

Experiment 4. Here we solve the pure Maxwell problem without absorption, i.e. $\kappa_{\mathrm{prob}} = 0$, with impedance boundary conditions on $\partial\Omega$. In the preconditioner we take $\kappa_{\mathrm{prec}} = k$. Results are given in Table 4, where $H_{\mathrm{sub}} \sim k^{-\alpha}$, $H \sim k^{-\alpha'}$. These methods are close to being load balanced in the sense that the coarse grid and subdomain problem size are very similar when $\alpha + \alpha' = 3/2$.

Out of the methods tested, the 2-level method (ImpHRAS) with $(\alpha, \alpha') = (0.6, 0.9)$ gives the best iteration count, but is more expensive. The method $(\alpha, \alpha') = (0.7, 0.8)$ is faster but its iteration count grows more quickly, so its advantage will diminish as k increases further. ~~We have no explanation for the curious reduction in iterations in the 2-level method as k increases for $(\alpha, \alpha') = (0.6, 0.9)$.~~[MB] For $(\alpha, \alpha') = (0.6, 0.9)$ the coarse grid size grows with $\mathcal{O}(n^{0.64})$ while the time grows with $\mathcal{O}(n^{0.65})$~~$\mathcal{O}(n^{0.80})$~~[MB]. For $(\alpha, \alpha') = (0.7, 0.8)$ the rates are $\mathcal{O}(n^{0.54})$ and $\mathcal{O}(n^{0.69})$~~$\mathcal{O}(n^{0.75})$~~[MB]. The subdomain problems are solved on individual processors so the number of processors used grows as k increases. In the current implementation a sequential direct solver on one processor is used to factorize the coarse problem matrix, which is clearly a limiting factor for the scalability of the algorithm. The timings could be significantly improved by using a distributed direct solver, or by adding a further level of domain decomposition for the coarse problem solve.

Acknowledgement This work has been supported in part by the French National Research Agency (ANR), project MEDIMAX, ANR-13-MONU-0012.

References

M. Bonazzoli, V. Dolean, I. G. Graham, E.A. Spence, and P-H. Tournier. Domain Decomposition preconditioning for the high-frequency time-harmonic Maxwell equations with absorption. *Submitted, arXiv:1711.03789*, 2017.

Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Applied Numerical Mathematics*, 50 (3):409–425, 2004.

M. J. Gander, I. G. Graham, and E. A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the

			$\alpha = 0.6$, $\alpha' = 0.9$				
k	n	N_{sub}	#2-level	n_{CS}	Time	#1-level	Time
10	2.6×10^5	27	20	2.9×10^3	16.2	37	13.7
15	1.5×10^6	125	26	1.0×10^4	25.5	70	26.1
20	5.2×10^6	216	29	2.1×10^4	52.0	94	60.6
25	1.4×10^7	216	33	4.4×10^4	145.5	105	191.2
30	3.3×10^7	343	38	6.9×10^4	380.4	132	673.5
			$\alpha = 0.7$, $\alpha' = 0.8$				
k	n	N_{sub}	#2-level	n_{CS}	Time	#1-level	Time
10	3.1×10^5	125	28	1.9×10^3	8.2	58	7.7
15	1.5×10^6	216	39	4.2×10^3	19.0	82	20.1
20	6.3×10^6	512	58	7.9×10^3	42.4	123	49.7
25	1.4×10^7	729	60	1.7×10^4	80.6	148	94.1
30	3.5×10^7	1000	80	2.6×10^4	251.9	179	328.0
			$\alpha = 0.8$, $\alpha' = 0.8$				
k	n	N_{sub}	#2-level	n_{CS}	Time	#1-level	Time
10	3.4×10^5	216	31	1.9×10^3	12.6	67	11.7
20	7.1×10^6	1000	70	7.9×10^3	76.9	147	58.3
30	4.1×10^7	3375	109	2.6×10^4	238.0	>200	-
40	1.3×10^8	6859	193	5.1×10^4	948.9	>200	-

Table 4 $\kappa_{\text{prob}} = 0$, $\kappa_{\text{prec}} = k$, $\delta \sim 2h$, $H_{\text{sub}} \sim k^{-\alpha}$, $H \sim k^{-\alpha'}$.

largest shift for which wavenumber-independent convergence is guaranteed? *Numer. Math.*, 131(3):567–614, 2015.

I. G. Graham, E. A. Spence, and E. Vainikko. *Recent Results on Domain Decomposition Preconditioning for the High-Frequency Helmholtz Equation Using Absorption*, pages 3–26. Geosystems Mathematics. Springer, 2017a.

I. G. Graham, E. A. Spence, and E. Vainikko. Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption. *Math. Comp.*, 86(307):2089–2127, 2017b.

A Coarse Space to Remove the Logarithmic Dependancy in Neumann-Neumann Methods

Faycal Chaouqui[1], Martin J.Gander[1], and Kévin Santugini-Repiquet[2]

1 Introduction

Domain Decomposition Methods are the most widely used methods for solving large linear systems that arize from the discretization of partial differential equations. The one level versions of these method are in general not scalable[1], since communication is just between neighboring subdomains, as it was pointed out already in [15], and one must add an additional coarse correction in order to share global information between subdomains. Examples of early such coarse corrections are proposed in [5, 6] for the additive Schwarz method, and in [12, 13, 14, 12, 7] for Neumann-Neumann and FETI methods, for a comprehensive treatement, see [16].

We are interested here in Neumann-Neumann methods, for which the one level condition number κ_1 and the two-level condition number κ_2 with a piecewise constant coarse space satisfy the estimates

$$\kappa_1 \leq \frac{C}{H^2}\left(1+\log^2(\frac{H}{h})\right), \quad \kappa_2 \leq C\left(1+\log^2(\frac{H}{h})\right), \tag{1}$$

where H is the typical size of a subdomain, h is the mesh size, and the constant C is indepandent of h and H, see [4, 12, 13]. These condition number estimates guarantee robust convergence when Neumann-Neumann is used as a preconditioner for a Krylov method, up to the logarithmic term.

We are interested here in understanding precisely where this logarithmic term is coming from, and how it can be removed using an appropriately chosen coarse space. To this end, we study the Neumann-Neumann method directly as an iterative method, not as a preconditioner, and consider the Laplace equation and two

[1] Université de Genève, Section de mathématiques, e-mail: {Faycal.Chaouqui}{Martin.Gander}@unige.ch .[2] Université Bordeaux, IMB, CNRS UMR5251, MC2, INRIA Bordeaux - Sud-Ouest, e-mail: Kevin.Santugini@math.u-bordeaux1.fr

[1] Notable exceptions are the time dependent wave equation with finite speed of propagation [8], and the Laplace equation in certain molecular simulations with specific geometry [2, 3].

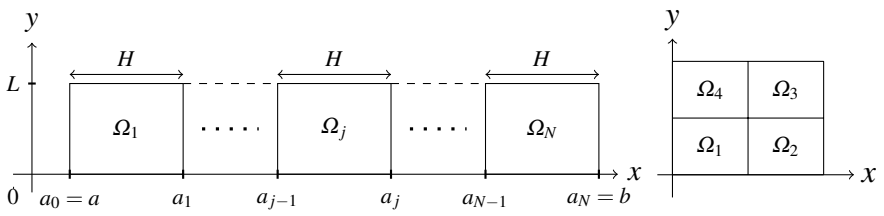


Fig. 1: Left: Strip decomposition. Right: Decomposition with a cross point

specific decompositions: a strip decomposition into a one dimensional sequence of subdomains, and a decomposition including cross points, see Figure 1.

For the strip decomposition, we will show that in the case of Dirichlet boundary conditions, the one level iterative Neumann-Neumann algorithm is convergent and can be weakly scalable, even without coarse grid, for a specific setting, and there are no polylogarithmic terms in the convergence estimate. In the case of Neumann boundary conditions, a coarse space of constant functions is needed to make the Neumann-Neumann method weakly scalable, and again there are no polylogarithmic terms in the convergence estimate. For a decomposition with cross points, we show that the iterative Neumann-Neumann algorithm does not converge, due to logarithmically growing modes at the cross point, and following ideas in [9, 11, 10], we enrich the coarse space with the corresponding modes to obtain a convergent iterative Neumann-Neumann algorithm without polylogarithmic growth.

2 Neumann-Neumann algorithm for a strip decomposition

We start by studying the convergence and weak scalability of the Neumann-Neumann algorithm for the Laplace equation,

$$\begin{aligned} -\Delta u &= f, \quad \text{in } \Omega, \\ u(a,\cdot) &= 0, \quad u(b,\cdot) = 0, \\ u(\cdot,0) &= 0, \quad u(\cdot,L) = 0, \end{aligned} \tag{2}$$

on the rectangular domain $\Omega := (a,b)\times(0,L)$ decomposed into strips, as shown in Figure 1 on the left, where $a_j = a + jH$ for $j = 0,\ldots,N$, and $\Omega_j := (a_{j-1},a_j)\times(0,L)$ for $j = 1,\ldots,N$. Given an initial guess g_j^0 at the interfaces, where we define $g_0^n = g_N^n = 0$ for convenience, the Neumann-Neumann algorithm computes for iteration index $n = 0,1,\ldots$ first solutions of the Dirichlet problems

$$\begin{aligned} -\Delta u_j^n &= f_j \quad \text{in } \Omega_j, \\ u_j^n(a_{j-1},\cdot) = g_{j-1}^n,\ u_j^n(a_j,\cdot) &= g_j^n, \end{aligned} \tag{3}$$

with outer boundary conditions $u_j^n(\cdot,0) = u_j^n(\cdot,L) = 0$, followed by solving Neumann problems on interior domains Ω_j, $j = 2,3,\ldots,N-1$, given by

$$\begin{aligned} -\Delta \psi_j^n &= 0 \quad \text{in } \Omega_j, \\ \partial_x \psi_j^n(a_{j-1},\cdot) &= (\partial_x u_j^n(a_{j-1},\cdot) - \partial_x u_{j-1}^n(a_{j-1},\cdot))/2, \\ \partial_x \psi_j^n(a_j,\cdot) &= (\partial_x u_j^n(a_j,\cdot) - \partial_x u_{j+1}^n(a_j,\cdot))/2, \end{aligned} \tag{4}$$

and on the left and right most subdomains the Neumann problems are

$$\begin{aligned} -\Delta \psi_1^n &= 0 \quad \text{in } \Omega_1, \\ \psi_1^n(a,\cdot) &= 0, \quad \partial_x \psi_1^n(a_1,\cdot) = (\partial_x u_1^n(a_1,\cdot) - \partial_x u_2^n(a_1,\cdot)/2, \\ -\Delta \psi_N^n &= 0 \quad \text{in } \Omega_N, \\ \psi_N^n(b,\cdot) &= 0, \quad \partial_x \psi_N^n(a_{N-1},\cdot) = (\partial_x u_N^n(a_{N-1},\cdot) - \partial_x u_{N-1}^n(a_{N-1},\cdot))/2, \end{aligned}$$

all with outer boundary conditions $\psi_j^n(\cdot,0) = 0$ and $\psi_j^n(\cdot,L) = 0$, $j = 1,\ldots,N$. The new interface traces are then obtained by the updating formula

$$g_j^{n+1} := g_j^n - (\psi_j^n(a_j,\cdot) + \psi_{j+1}^n(a_j,\cdot))/2, \quad j = 1,\ldots,N-1. \tag{5}$$

To study the convergence of this iterative Neumann-Neumann method, it suffices by linearity to apply the algorithm to Equation (2) with $f = 0$, and to study the convergence of the approximate solution u^n to the zero solution. Since the subdomains are rectangles, the iterates can be expanded in a sine series,

$$u_j^n(x,y) = \sum_{m=1}^{\infty} \widehat{u}_j^n(x,m)\sin(k_m y), \quad \psi_j^n(x,y) = \sum_{m=1}^{\infty} \widehat{\psi}_j^n(x,m)\sin(k_m y), \tag{6}$$

where $k_m := \frac{m\pi}{L}$, which allows us to study the convergence based on the Fourier coefficients.

Lemma 1. *Let* $\widehat{\boldsymbol{u}}^n(m) = \left[\widehat{u}_1^n(a_1,m), \widehat{u}_2^n(a_2,m), \ldots, \widehat{u}_{N-1}^n(a_{N-1},m)\right]^T \in \mathbb{R}^{N-1}$, *then for*[2] $N \geq 3$ *we have* $\widehat{\boldsymbol{u}}^n(m) = T(m,H)\widehat{\boldsymbol{u}}^{n-1}(m)$, *where* $T(m,H) \in \mathbb{R}^{(N-1)\times(N-1)}$ *is given by*

$$T(m,H) = -\frac{1}{4\sinh^2(k_m H)} \begin{bmatrix} 1 & \frac{1}{\cosh(k_m H)} & -1 & 0 & \cdots & \cdots & 0 \\ 0 & 2 & 0 & -1 & \ddots & & \vdots \\ -1 & 0 & 2 & 0 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 2 & 0 & -1 \\ \vdots & & \ddots & -1 & 0 & 2 & 0 \\ 0 & \cdots & \cdots & 0 & -1 & \frac{1}{\cosh(k_m H)} & 1 \end{bmatrix}.$$

[2] For $N = 2$ the structure of $T(m,H)$ is not the same since there are no inner subdomains.

Proof. For each $m \geq 1$ and $j = 2, \ldots, N-1$, $u_j^n(x,m)$ and $\psi_j^n(x,m)$ satisfy

$$\begin{array}{ll} k_m^2 \widehat{u}_j^n - \partial_{xx} \widehat{u}_j^n = 0, & k_m^2 \widehat{\psi}_j^n - \partial_{yy} \widehat{\psi}_j^n = 0, \\ \widehat{u}_j^n(a_{j-1},m) = \widehat{g}_{j-1}^n(m), & \widehat{\psi}_j^n(a_{j-1},m) = (\partial_x u_j^n(a_{j-1},m) - \partial_x u_{j-1}^n(a_{j-1},m))/2, \\ \widehat{u}_j^n(a_j,m) = \widehat{g}_j^n(m), & \widehat{\psi}_j^n(a_j,m) = (\partial_x u_j^n(a_j,m) - \partial_x u_{j+1}^n(a_j,m))/2. \end{array}$$

The solution of the Dirichlet problems on interior subdomains are thus

$$\widehat{u}_j^n(x,m) = \widehat{g}_j^n(m) \frac{\sinh(k_m(x-a_{j-1}))}{\sinh(k_m H)} + \widehat{g}_{j-1}^n(m) \frac{\sinh(k_m(a_j - x))}{\sinh(k_m H)}, \ j = 2, \ldots, N-1,$$

and on the subdomains on the left and right we get

$$\widehat{u}_1^n(x,m) = \widehat{g}_1^n(m) \frac{\sinh(k_m(x-a_0))}{\sinh(k_m H)}, \quad \widehat{u}_N^n(x,m) = \widehat{g}_{N-1}^n(m) \frac{\sinh(k_m(a_N - x))}{\sinh(k_m H)}.$$

Similarly for the Neumann problems on the interior subdomains, we obtain

$$\begin{aligned} \widehat{\psi}_j^n(x,m) = & \left(2\,\widehat{g}_j^n(m) \frac{\cosh(k_m H)}{\sinh(k_m H)} - \frac{\widehat{g}_{j-1}^n(m)}{\sinh(k_m H)} - \frac{\widehat{g}_{j+1}^n(m)}{\sinh(k_m H)} \right) \frac{\cosh(k_m(x-a_{j-1}))}{2\sinh(k_m H)} \\ & + \left(2\,\widehat{g}_{j-1}^n(m) \frac{\cosh(k_m H)}{\sinh(k_m H)} - \frac{\widehat{g}_{j-2}^n(m)}{\sinh(k_m H)} - \frac{\widehat{g}_j^n(m)}{\sinh(k_m H)} \right) \frac{\cosh(k_m(a_j - x))}{2\sinh(k_m H)}, \end{aligned}$$

and for the first and last subdomains we find

$$\begin{aligned} \widehat{\psi}_1^n(x,m) &= \left(2\,\widehat{g}_1^n(m) \frac{\cosh(k_m H)}{\sinh(k_m H)} - \frac{\widehat{g}_2^n(m)}{\sinh(k_m H)} \right) \frac{\sinh(k_m(x-a_0))}{2\cosh(k_m H)}, \\ \widehat{\psi}_N^n(x,m) &= \left(2\,\widehat{g}_{N-1}^n(m) \frac{\cosh(k_m H)}{\sinh(k_m H)} - \frac{\widehat{g}_{N-2}^n(m)}{\sinh(k_m H)} \right) \frac{\sinh(k_m(a_N - x))}{2\cosh(k_m H)}. \end{aligned}$$

Using now (5) and the fact that $\widehat{u}_j^n(a_j,m) = \widehat{g}_j^n(m)$ for each $m \geq 1$, we get the stated recurrence relation.

Lemma 2. *If $H/L > \ln(1+\sqrt{2})/\pi$ then for any $m \geq 1$ we have $\|T(m,H)\|_\infty < 1$.*

Proof. It is straightforward to see that $\|T(m,H)\|_\infty \leq \frac{1}{\sinh^2(k_m H)}$ for each m and since $m \mapsto \frac{1}{\sinh^2(k_m H)}$ is strictly decreasing for $m \geq 1$, we have that $\frac{1}{\sinh^2(k_m H)} < \frac{1}{\sinh^2(k_1 H)}$ which is strictly smaller than 1 if $H/L > \ln(1+\sqrt{2})/\pi$, which concludes the proof.

Theorem 1. *For $N \geq 3$ Neumann-Neumann satisfy the L^2 error bound*

$$\left(\sum_{j=1}^{N-1} \|u_j^n(a_j,\cdot)\|_2^2 \right)^{1/2} \leq \frac{1}{\sinh^{2n}(k_1 H)} \left(\sum_{j=1}^{N-1} \|u_j^0(a_j,\cdot)\|_2^2 \right)^{1/2}.$$

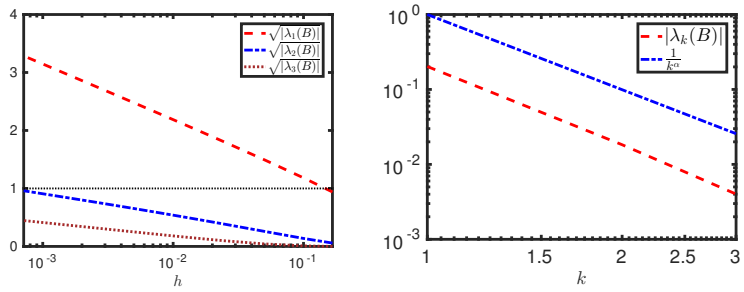


Fig. 2: Left: dependence of $\sqrt{|\lambda_k(B)|}$, $k = 1,\ldots,3$ on the mesh size h in semi-log scale. Right: dependence of the semi-log scale slope of $|\lambda_k(B)|$, $k = 1,\ldots,3$ on k, with $\alpha := \frac{10}{3}$.

Proof. Since for $N \geq 3$ we have that $\|T(m,H)\|_2 \leq \sqrt{\|T(m,H)\|_\infty \|T(m,H)\|_1} \leq \frac{1}{\sinh^2(k_1 H)}$, and using the Parseval identity $\|u_j^n(a_j,\cdot)\|_2^2 = \frac{L}{2}\sum_{m=1}^{\infty} \widehat{u}_i^n(a_j,m)^2$, we get the result stated.

Theorem 1 shows that under a minimal assumption, the one level Neumann-Neumann algorithm for the strip decomposition is weakly scalable, provided H remains fixed, i.e. more and more subdomains of the same size are added, see also [2, 3] for the corresponding Schwarz scaling. If the original Laplace problem (2) has however Neumann conditions at $x = 0$ and $x = L$, then the interior subdomains become floating in the Neumann-Neumann algorithm, and a minimal coarse space consisting of piecewise constant functions is required in order to remove the kernel, and this is sufficient to make the algorithm weakly scalable as in previous case with an L^2 bound as in Theorem 1, see [1].

3 Neumann-Neumann algorithm with cross points

We now study the convergence properties of the iterative Neumann-Neumann algorithm for decompositions with cross points, like the one shown in Figure 1 on the right. Since in this case the algorithm might be undefined at the continuous level due to possible discontinuity at the cross point, we study numerically the convergence of the fixed point iteration

$$\mathbf{u}_{n+1} = B\mathbf{u}_n + \mathbf{f}, \tag{7}$$

where $B \in \mathbb{R}^{d\times d}$ and $\mathbf{f} \in \mathbb{R}^d$ are obtained by discretizing the Neumann-Neumann algorithm using five-points stencil central finite differences. We first show in Figure 2 on the left the three largest (double) eigenvalues in modulus of B when the mesh is refined. We clearly see logarithmic growth, and the iterative Neumann-Neumann method will diverge as soon as the mesh size h is small enough, in our example $h = 0.12$. Hence, in contrast to the classical alternating and parallel Schwarz meth-

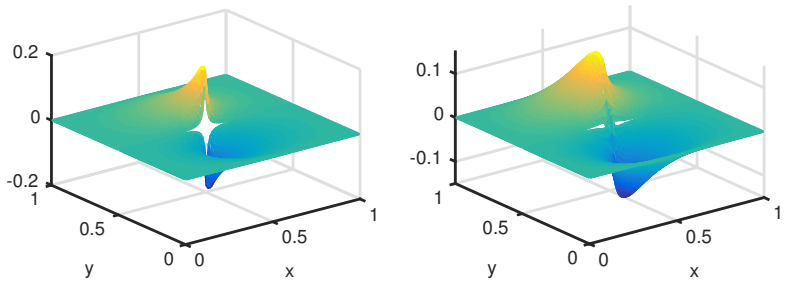


Fig. 3: Left: dominant eigenfunction of B. Right: second eigenfunction of B.

ods, the Neumann-Neumann method can then not be used as an iterative solver. We note however also that the logarithmic growth of the first dominant eigenvalue is faster than the second and the third one. On the right in Figure 2, we show how the growth rate (the slope) of these diverging modes depends on the eigenvalue index k. We see that the growth decays very rapidly, like $1/k^{\alpha}$ with $\alpha = 10/3$, so when h goes to zero, there are only $O(k)$ divergent modes (those with corresponding eigenvalues greater than 1 in the absolute value), where $1/k^{\alpha}\log^2(h) \lesssim 1$, i.e $k \sim (\log^2(h))^{1/\alpha}$.

We next show in Figure 3 the two corresponding dominant eigenmodes of B for a mesh size $h = 0.01$. Since their eigenvalues are double eigenvalues, we chose from the two dimensional subspace of eigenfunctions the one vanishing at the interface aligned with the x axis; the other eigenmode has the same shape, just rotated by 90 degrees. We see that the cross point causes the iterative Neumann-Neumann method to generate eigenmodes with a singular behavior at the cross point, and these modes lead to divergence of the iterative Neumann-Neumann method.

To avoid such logarithmic growth, and obtain an convergent iterative Neumann-Neumann method, one can remove the few divergent modes using an enriched coarse space. Let F be a subspace of $\mathbb{R}^d$ and $F^{\perp}$ its orthogonal complement with standard inner product. Then we can use the reordering

$$B = \begin{array}{c} \\ F \\ F^{\perp} \end{array} \begin{array}{c} \begin{array}{cc} F & F^{\perp} \end{array} \\ \begin{bmatrix} \widetilde{B} & C \\ G & \widehat{B} \end{bmatrix} \end{array}, \quad \mathbf{u} = \begin{array}{c} F \\ F^{\perp} \end{array} \begin{bmatrix} \widetilde{\mathbf{u}} \\ \widehat{\mathbf{u}} \end{bmatrix}, \quad \mathbf{f} = \begin{array}{c} F \\ F^{\perp} \end{array} \begin{bmatrix} \widetilde{\mathbf{f}} \\ \widehat{\mathbf{f}} \end{bmatrix}, \tag{8}$$

and the iterative Neumann-Neumann algorithm (7) becomes

$$\begin{bmatrix} \widetilde{\mathbf{u}}_{n+1} \\ \widehat{\mathbf{u}}_{n+1} \end{bmatrix} = \begin{bmatrix} \widetilde{B} & C \\ G & \widehat{B} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{u}}_{n} \\ \widehat{\mathbf{u}}_{n} \end{bmatrix} + \begin{bmatrix} \widetilde{\mathbf{f}} \\ \widehat{\mathbf{f}} \end{bmatrix}. \tag{9}$$

To correct the problem of the divergent modes, we propose to use the iteration

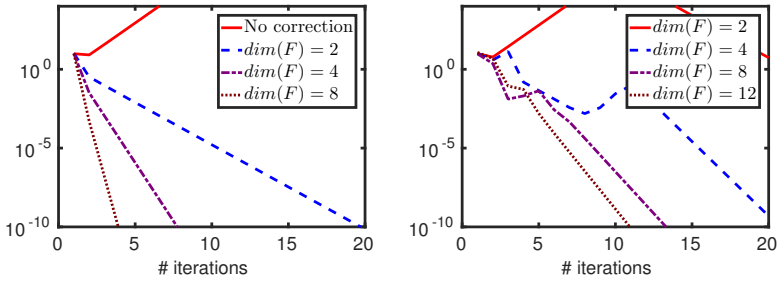


Fig. 4: Left: error of iteration (10) for different dimension of F. Right: same, but using orthogonal iteration to approximate F.

$$\widehat{\mathbf{u}}_{n+1} = \widehat{B}\widehat{\mathbf{u}}_n + \widehat{\mathbf{f}} + G\widetilde{\mathbf{u}}_n, \tag{10a}$$

$$(I - \widetilde{B})\widetilde{\mathbf{u}}_{n+1} = C\widehat{\mathbf{u}}_{n+1} + \widetilde{\mathbf{f}}, \tag{10b}$$

where (10b) is solved exactly.

Theorem 2. *If F consists of all eigenfunctions of B with respective eigenvalues greater than 1 in absolute value, then iteration (10) converges for any $\mathbf{u}_0 \in \mathbb{R}^d$.*

Proof. From (10), we obtain

$$\widehat{\mathbf{u}}_{n+1} = (\widehat{B} + G(I - \widetilde{B})^{-1}C)\widehat{\mathbf{u}}_n + \widehat{\mathbf{f}} + G(I - \widetilde{B})^{-1}\widetilde{\mathbf{f}},$$
$$\widetilde{\mathbf{u}}_{n+1} = (I - \widetilde{B})^{-1}(C\widehat{\mathbf{u}}_{n+1} + \widetilde{\mathbf{f}}),$$

and hence the method is convergent iff $\rho(\widehat{B} + G(I - \widetilde{B})^{-1}C) < 1$. Since F consists of the divergent eigenmodes of B we have that G is zero and the condition for convergence becomes $\rho(\widehat{B}) < 1$, which is satisfied since $\widehat{B}$ does not contain the divergent eigenmodes of B.

We show in Figure 4 on the left the error of iteration (10) with a random initial guess u_0 as a function of the iteration number n for different choices of the dimension of F, using the same mesh size $h = 0.01$ in a semi-log scale. We see that with $\dim(F) = 2$, the iterations start already to converge while without correction the iteration diverges. Increasing the dimension of F improves convergence further. Using just orthogonal iterations to approximate F gives already satisfactory results, as shown on the right in Figure 4.

4 Conclusion

We showed that the logarithmic growth in the condition number estimate of the Neumann-Neumann method comes from modes which are generated at cross points

in the decomposition. Without cross points, the iterative Neumann-Neumann method is convergent and can be made scalable just using a constant per subdomain in the coarse space. With cross points, one can add the logaritmically divergent modes to the coarse space to obtain a convergent iterative Neumann-Neumann method, without logarithmic term in the convergence estimate. We also showed that orthogonal iteration permits already to include such modes numerically, and we are currently trying to determine these coarse functions analytically.

References

1. Chaouqui, F., Gander, M.J., Santugini-Repiquet, K.: A new coarse space for Neumann-Neumann Methods. In preparation (2017)
2. Ciaramella, G., Gander, M.: Analysis of the parallel Schwarz method for growing chains of fixed-size subdomains: part I. To appear in SIAM J. Numer. Anal (2016)
3. Ciaramella, G., Gander, M.: Analysis of the parallel Schwarz method for growing chains of fixed-sized subdomains: Part II. submitted (2017)
4. De Roeck, Y.H., Le Tallec, P.: Analysis and test of a local domain decomposition preconditioner. In: Fourth international symposium on domain decomposition methods for partial differential equations, vol. 4 (1991)
5. Dryja, M., Widlund, O.: An additive variant of the Schwarz alternating method for the case of many subregions. Ultracomputer Research Laboratory, Univ., Courant Inst. of Mathematical Sciences, Division of Computer Science (1987)
6. Dryja, M., Widlund, O.B.: Additive Schwarz methods for elliptic finite element problems in three dimensions. New York University. Courant Institute of Mathematical Sciences. Computer Science Department (1991)
7. Farhat, C., Lesoinne, M., LeTallec, P., Pierson, K., Rixen, D.: Feti-dp: a dual–primal unified feti methodpart i: A faster alternative to the two-level feti method. International journal for numerical methods in engineering **50**(7), 1523–1544 (2001)
8. Gander, M.J., Halpern, L., Nataf, F.: Optimal Schwarz waveform relaxation for the one dimensional wave equation. SIAM Journal of Numerical Analysis **41**(5), 1643–1681 (2003)
9. Gander, M.J., Halpern, L., Santugini-Repiquet, K.: A new coarse grid correction for ras/as. In: Domain Decomposition Methods in Science and Engineering XXI, pp. 275–283. Springer (2014)
10. Gander, M.J., Loneland, A.: SHEM: An optimal coarse space for RAS and its multiscale approximation. Domain Decomposition Methods in Science and Engineering XXIII, Springer (2016)
11. Gander, M.J., Loneland, A., Rahman, T.: Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods. arXiv preprint arXiv:1512.05285 (2015)
12. Mandel, J., Brezina, M.: Balancing domain decomposition for problems with large jumps in coefficients. Mathematics of Computation of the American Mathematical Society **65**(216), 1387–1401 (1996)
13. Mandel, J., Dohrmann, C.R.: Convergence of a balancing domain decomposition by constraints and energy minimization. Numerical linear algebra with applications **10**(7), 639–659 (2003)
14. Mandel, J., Dohrmann, C.R., Tezaur, R.: An algebraic theory for primal and dual substructuring methods by constraints. Applied numerical mathematics **54**(2), 167–193 (2005)
15. Nicolaides, R.A.: Deflation of conjugate gradients with applications to boundary value problems. SIAM Journal on Numerical Analysis **24**(2), 355–365 (1987)
16. Toselli, A., Widlund, O.B.: Domain decomposition methods: algorithms and theory, vol. 34. Springer (2005)

A Crank-Nicholson domain decomposition method for optimal control problem of parabolic partial differential equation

Jixin Chen and Danping Yang

Abstract A parallel domain decomposition algorithm is considered for solving an optimal control problem governed by a parabolic partial differential equation. The proposed algorithm relies on non-iterative and non-overlapping domain decomposition, which uses some implicit sub-domain problems and explicit flux approximations at each time step in every iteration. In addition, outer iterations are introduced to achieve the parallelism. Numerical experiments are supplied to show the efficiency of our proposed method.

1 Introduction

In [1], Dawson and Dupont presented non-overlapping domain decomposition schemes to solve parabolic equation by some explicit flux exchange on inner boundaries and implicit conservative Galerkin procedures in each sub-domain. Here, explicit flux prediction are simple to compute for the unit outward normal vector (see definition in Section 2). A time step limitation, which is less severe than that of a fully explicit method, is induced to maintain stability because of the explicit prediction. Recently, an improved strategy was considered in [2] to avoid the loss of $H^{-\frac{1}{2}}$ factor for space variable in the work of Dawson and Dupont. We would like to mention that another two calculation methods on inner boundaries were studied by Ma and Sun (see [6] and sequent research papers) based on the integral mean value or extrapolation. In previous work [3], we have shown that explict/implict domain

Jixin Chen
Department of Mathematics, East China Normal University, Shanghai, China; Department of Mathematics, Université libre de Bruxelles, Brussels, Belgium. e-mail: cjxhmj22344457@163.com

Danping Yang
Department of Mathematics, East China Normal University, Shanghai, China. e-mail: dpyang@math.ecnu.edu.cn

decomposition method in [2] could be applied in optimal control problems governed by partial differential equations. The main goal of this paper is to develop the corresponding results for second order procedures based on the analysis and schemes designed to solve single PDE in [4].

2 Model problem and optimality condition

We consider the following distributed convex optimal control problems

$$\min_{u\in\mathscr{K}}\Big\{\int_0^T \big(\|y-y_d\|^2_{L^2(\Omega)}+\|u\|^2_{L^2(\Omega)}\big)dt\Big\} \tag{1}$$

subject to

$$\begin{cases} \partial_t y-\Delta y=f+u, & \text{in } \Omega, \quad 0<t\le T;\\ y=0, & \text{on } \partial\Omega, \quad 0<t\le T;\\ y=y_0, & \text{in } \Omega, \quad t=0, \end{cases} \tag{2}$$

where $u\in\mathscr{K}$ is the control and $\mathscr{K}$ is a convex admissible set for control, y is the state variable, y_d is the observation, y_0 is the initial function. Fix $V=H_0^1(\Omega)$ and $U=L^2(\Omega)$. In the following, we will write state space $\mathscr{W}=\{y\in L^2(0,T;V);\ y_t\in L^2(0,T;\ H^{-1}(\Omega))\}$ and the control space $\mathscr{U}=L^2(0,T;U)$. In addition, K is a closed convex set in U and $\mathscr{K}=L^2(0,T;K)$ is a closed convex set in the space $\mathscr{U}$.

2.1 Optimality Condition and discretization

We use standard notation for Sobolev spaces. Define $A(u,v):V\times V\to\mathbb{R}$ to be a bilinear form satisfying

$$A(u,v)=(\nabla u,\nabla v)\quad \forall\, u,v\in V. \tag{3}$$

Then the optimal control problem can be transformed into optimality condition in the following lemma:

Theorem 1. *A pair (y,u) in $\mathscr{W}\times\mathscr{K}$ is the solution of* (1)-(2) *if and only if there is a co-state $p\in\mathscr{W}$ such that the triplet (y,p,u) in $\mathscr{W}\times\mathscr{W}\times\mathscr{K}$ satisfies the following optimality conditions:*

$$\begin{cases} (\partial_t y,w)+A(y,w)=(f+u,w),\ \ \forall\, w\in V;\\ y|_{t=0}=y_0; \end{cases} \tag{4}$$

$$\begin{cases} -(\partial_t p,q)+A(q,p)=(y-y_d,q),\ \ \forall\, q\in V;\\ p|_{t=T}=0; \end{cases} \tag{5}$$

$$\int_0^T (u+p, v-u) \geq 0, \ \ \forall\, v \in \mathscr{K}. \tag{6}$$

Here only the case $K = \{u \geq 0\}$ are considered. Therefore, the third inequality in the optimality conditions is equivalent to

$$(u+p, v-u) \geq 0, \ \ \forall\, v \in K, \ \ 0 \leq t \leq T. \tag{7}$$

In general, for time-dependent optimal control problems, optimality condition, which is a large scale of nonlinear coupled system with respect to time and spacial variables, contains forward and backward PDEs with the variational inequality under consideration. It is very difficult and challenging to solve directly this non-linear system. Domain decomposition method, which could save huge time in calculation by solving the question at the same time, is especially suitable for this kind of complicated problem. To use domain decomposition method, we divide Ω into many non-overlapping sub-domains $\{\Omega_i\}_{i=1}^I$ such that $\bar{\Omega} = \bigcup_{i=1}^I \bar{\Omega}_i$. Set $\Gamma_i = \partial\Omega_i \backslash \partial\Omega$ and $\Gamma = \bigcup_{i=1}^I \Gamma_i$, which is the set of inner boundaries of sub-domains. We recall some definitions which are necessary for deriving the discrete form of (4)-(6). Introduce

$$\phi(x) = \begin{cases} (x-2)/12, & 1 \leq x \leq 2, \\ -5x/4+7/6, & 0 \leq x \leq 1, \\ 5x/4+7/6, & -1 \leq x \leq 0, \\ -(x+2)/12, & -2 \leq x \leq -1, \\ 0. & |x| > 2. \end{cases}$$

For some $H > 0$, define

$$\phi(\tau) = H^{-1}\varphi\big(\frac{\tau}{H}\big), \ \tau \in \mathbb{R}^1.$$

where H is the width of the local averaging interval, which plays an important role for stability of explicit/implicit scheme. Following Dawson-Dupont's idea, we do not use the exact normal derivative along inner boundaries. A proper approximation is (see [1, 2]):

$$B(\psi)(\boldsymbol{x}) = -\int_{-2H}^{2H} \phi'(\tau)\psi(\boldsymbol{x}+\tau\boldsymbol{n}_\Gamma)\mathrm{d}\tau, \ \ x \in \Gamma_i \cap \Gamma_j, \ \ 1 \leq i < j \leq I. \tag{8}$$

From definitions above, we note that function v has a well-defined jump

$$[v](\boldsymbol{x}) = v(\boldsymbol{x}^+) - v(\boldsymbol{x}^-), \quad \forall\, \boldsymbol{x} \text{ on } \Gamma \tag{9}$$

where

$$v(\boldsymbol{x}^\pm) \triangleq \lim_{t \to 0^\pm} v(\boldsymbol{x}+t\boldsymbol{v}_\Gamma) \tag{10}$$

Make a time partition: $0 = t^0 < t^1 < \cdots < t^N = T$ and set $\Delta t^n = t^n - t^{n-1}$ and $\Delta t = \max\limits_{1 \leq n \leq N} \Delta t^n$. For simplicity, we may take $\Delta t^n = \Delta t$ for $n = 1, 2, \ldots, N$. For a given function $g(\boldsymbol{x}, t)$, let $g^n = g(\boldsymbol{x}, t^n)$ and

$$\bar{\partial}_t g^n = \frac{g^n - g^{n-1}}{\Delta t}, \;\; \bar{g}^{n-\frac{1}{2}} = \frac{g^n + g^{n-1}}{2},$$

$$\hat{g}^{n-\frac{1}{2}} = 2\bar{g}^{n-\frac{3}{2}} - \bar{g}^{n-\frac{5}{2}}, \;\; \tilde{g}^{n+\frac{1}{2}} = 2\bar{g}^{n+\frac{3}{2}} - \bar{g}^{n+\frac{5}{2}}.$$

For $i = 1,2,\ldots,I$, denote $M_i^h \subset V$ be the corresponding continuous piecewise linear finite element space associated with conforming trangualtion $\mathscr{T}_i^h$. Let M^h be the subspace of V such that $w_h \in M^h$ if and only if $w_h|_{\Omega_i} \in M_i^h$ for each $1 \le i \le I$. Similarly, we can define piecewise constant finite element space $U^{h_U} \subset U$ for control variable u. Let $K^{h_U} = K \bigcap U^{h_U}$. Then the discrete form that we want to solve is:

$$\begin{cases} Y^0 = y_0; \quad Y^1 = y_0 + \Delta t(f^0 + \Delta y_0 + U^0); \quad Y^2 = y_0 + 2\Delta t(f^0 + \Delta y_0 + U^0); \\ (\bar{\partial}_t Y^n, V) + A(Y^{n-\frac{1}{2}}, V) - (B(\hat{Y}^{n-\frac{1}{2}}), [V])_\Gamma - (B(V), [\hat{Y}^{n-\frac{1}{2}}])_\Gamma \\ \quad = (\bar{f}^{n-\frac{1}{2}} + \bar{U}^{n-\frac{1}{2}}, V), \;\; \forall\, V \in M^h, \;\; n = 3,4,\ldots,N; \end{cases} \tag{11}$$

$$\begin{cases} P^N = 0; \quad P^{N-1} = \Delta t(Y^N - y_d^N); \quad P^{N-2} = 2\Delta t(Y^N - y_d^N); \\ -(\bar{\partial}_t P^{n-2}, V) + A(V, \bar{P}^{n-\frac{5}{2}}) - (B(\tilde{P}^{n-\frac{5}{2}}), [V])_\Gamma - (B(V), [\tilde{P}^{n-\frac{5}{2}}])_\Gamma \\ \quad = (\bar{Y}^{n-\frac{5}{2}} - \bar{y}_d^{n-\frac{5}{2}}, V), \;\; \forall\, V \in M^h, \;\; n = N, N-1, \ldots, 3; \end{cases} \tag{12}$$

$$(\bar{U}^{n-\frac{1}{2}} + \bar{P}^{n-\frac{5}{2}}, \bar{Z}^{n-\frac{1}{2}} - \bar{U}^{n-\frac{1}{2}}) \ge 0, \quad \forall\, Z \in K^{h_U},\, n = 3,4,\ldots,N; \tag{13}$$

$$U^0 = \max\{0, -P^0\}, \quad U^1 = \max\{0, -P^1\}, \quad U^2 = \max\{0, -P^2\}. \tag{14}$$

We see that the original optimal control problem (4)-(6), which is normally large in size, is now decomposed into a set of subproblems with much smaller sizes. In fact, discrete solution of (11)-(14) does not always exist. One could use contraction mapping principle to ensure the existence and uniqueness of system. Taking the limitation of the length into consideration, we will give a rigorous analysis on this and convergence of the following iterative algorithm in a forthcoming paper [5]. In addition, a priori estimates will also be included.

2.2 Parallel iterative algorithm

We note that discrete system (11)-(14) is still a nonlinear system of a forward system for the state variable and a backward system for the co-state variable, which are coupled by the control variable. We introduce outer iterations to decouple the system. Thus, the proposed algorithm could be performed in parallel once domain decomposition is used. Then, fully parallel iterative algorithm is formulated as follows:

PARALLEL DOMAIN DECOMPOSITION ITERATIVE ALGORITHM (PDDIA)

Step 1. *Given initial approximation* $\{U_0^n\}_{n=1}^N \subset U^{h_U}$ *and* $Y^0 \in M^h$. *Take the* $\varepsilon > 0$ *as a tolerance and set* $k := 0$.

Step 2. *Update* $\{Y_{k+1}^n\}_{n=0}^N \subset M^h$ *in parallel on each* Ω_i *for* $1 \le i \le I$:

$$\begin{cases} Y_{k+1}^0 = Y^0;\ Y_{k+1}^1 = Y_0 + \Delta t(f^0 + \Delta Y^0 + U_k^0);\ Y_{k+1}^2 = Y_0 + 2\Delta t(f^0 + \Delta Y^0 + U_k^0); \\ (\bar{\partial}_t Y_{k+1}^n, V) + A(Y_{k+1}^{n-\frac{1}{2}}, V) - (B(\hat{Y}_{k+1}^{n-\frac{1}{2}}), [V])_\Gamma - (B(V), [\hat{Y}_{k+1}^{n-\frac{1}{2}}])_\Gamma \\ \quad = (\bar{f}^{n-\frac{1}{2}} + \bar{U}_k^{n-\frac{1}{2}}, V), \qquad \forall\, V \in M^h, \quad n = 3,4,\ldots,N; \end{cases} \tag{15}$$

Step 3. *Update* $\{P_{k+1}^n\}_{n=0}^N \subset M^h$ *in parallel on each* Ω_i *for* $1 \le i \le I$:

$$\begin{cases} P_{k+1}^N = 0;\ P_{k+1}^{N-1} = \Delta t(Y^N - y_d^N);\ P_{k+1}^{N-2} = 2\Delta t(Y^N - y_d^N); \\ -(\bar{\partial}_t P_{k+1}^{n-2}, V) + A(V, \bar{P}_{k+1}^{n-\frac{5}{2}}) - (B(\bar{P}_{k+1}^{n-\frac{5}{2}}), [V])_\Gamma - (B(V), [\tilde{P}_{k+1}^{n-\frac{5}{2}}])_\Gamma \\ \quad = (\bar{Y}_{k+1}^{n-\frac{5}{2}} - \bar{y}_d^{n-\frac{5}{2}}, V), \quad \forall\, V \in M^h, \quad n = N, N-1,\ldots,3; \end{cases} \tag{16}$$

Step 4. *Update* $\{\bar{U}_{h_U,k+1}^{n-\frac{1}{2}}\}_{n=1}^N \subset U^{h_U}$ *such that*

$$\begin{cases} \bar{U}_{k+\frac{1}{2}}^{n-\frac{1}{2}} = (1-\rho)\bar{U}_k^{n-\frac{1}{2}} - \rho\bar{P}_{k+1}^{n-\frac{5}{2}}, \\ \bar{U}_{k+1}^{n-\frac{1}{2}} = Q^{h_U}\bar{U}_{k+\frac{1}{2}}^{n-\frac{1}{2}}. \end{cases} \qquad n = 3,4,\ldots,N; \tag{17}$$

where ρ *is a constant with* $0 < \rho < 1$ *and* Q^{h_U} *is the projection from* U^{h_U} *to* K^{h_U}.

Define U_{k+1}^0, U_{k+1}^1 *and* U_{k+1}^2 *such that*

$$U_{k+1}^0 = \max\{0, -P_{k+1}^0\}, \quad U_{k+1}^1 = \max\{0, -P_{k+1}^1\}, \quad U_{k+1}^2 = \max\{0, -P_{k+1}^2\}, \tag{18}$$

Step 5. *Compute the iterative error:*

$$\text{eps} = \sum_{n=0}^{N} \left(\|\bar{U}_k^{n-\frac{1}{2}} - \bar{U}_{k+1}^{n-\frac{1}{2}}\|_{L^2(\Omega)} + \|\bar{Y}_k^{n-\frac{1}{2}} - \bar{Y}_{k+1}^{n-\frac{1}{2}}\|_{L^2(\Omega)} + \|\bar{P}_k^{n-\frac{1}{2}} - \bar{P}_{k+1}^{n-\frac{1}{2}}\|_{L^2(\Omega)} \right.$$

If $eps \le \varepsilon$, *then stop the iteration and output*

$$U^n = U_{k+1}^n,\ \ Y^n = Y_{k+1}^n,\ \ P^n = P_{k+1}^n, \quad n = 0,1,2,\ldots,N. \tag{19}$$

Else set $k := k+1$ *and return step 2 to restart new iteration.*

Compared to first order scheme proposed in [3], the computation on Γ requires explicitly the value of three-level solutions, while only little computational cost will be added. We also remark that the algorithm PDDIA is fully parallel.

3 Numerical experiments

In this section, we test the performance and convergence of the proposed PDDIA with respect to the exact solutions:

$$\begin{cases} y = \sin(2\pi x)\sin(2\pi y)t, \\ p = \sin(2\pi x)\sin(2\pi y)(T-t), \\ u = \max(-p, 0), \\ y_d = y + \dfrac{\partial p}{\partial t} + \Delta p, \\ f = -u + \dfrac{\partial y}{\partial t} - \Delta y. \end{cases}$$

Let $T = 0.5$. Domain $\Omega = [0,2] \times [0,1]$ is partitioned into two uniform non-overlapping areas with the inner-domain boundary are $\Gamma = \{1\} \times [0,1]$. The mesh in the x-axis and y-axis varies uniformly from $1/36$, $1/49$, $1/64$ to $1/81$ in each sub-domain, respectively.

Table 1 $L^2(0,T;L^2(\Omega))$-norm error for PDDIA ($r = 1$)

Grids	$y-Y$	order	$u-U$	order	$p-P$	order
36×36	1.625×10^{-3}		7.324×10^{-3}		1.597×10^{-3}	
49×49	8.770×10^{-4}	2.00	5.382×10^{-3}	0.99	8.473×10^{-4}	2.06
64×64	5.294×10^{-4}	1.89	4.153×10^{-3}	0.97	5.020×10^{-4}	1.96
81×81	3.376×10^{-4}	1.91	3.283×10^{-3}	1.00	3.129×10^{-4}	2.01

For domain decomposition, we set $\Delta t = 0.1h$ and $H^2 = rh$ to balance error accuracy, where parameter r is a constant. The algorithm stops after that error of adjacent iterative step defined in step 5 of the algorithm is less than 10^{-6}.

In all of the numerical tests, the state variable y and co-state variable p are approximated by using piecewise linear functions while control solution u are treated with piecewise constant functions. Compared to the scheme proposed in [3], the number presented in Table 1 to Table 3 are the sum of average value of two neighbouring layer, which is a good approximation for exact solution evaluating at the middle of two adjacent time layer. We present numerical simulations in Table 1 for $r = 1$. The L^2-norm error of the numerical solutions are listed in Table 2 for $r = 4$. We present the corresponding results when $r = 9$ in Table 3.

Table 2 $L^2(0,T;L^2(\Omega))$-norm error for PDDIA ($r=4$)

Grids	$y-Y$	order	$u-U$	order	$p-P$	order
36×36	4.835×10^{-3}		8.069×10^{-3}		4.856×10^{-3}	
49×49	2.525×10^{-3}	2.11	5.654×10^{-3}	1.15	2.523×10^{-3}	2.12
64×64	1.389×10^{-3}	2.24	4.258×10^{-3}	1.06	1.379×10^{-3}	2.26
81×81	8.077×10^{-4}	2.30	3.325×10^{-3}	1.05	7.952×10^{-4}	2.34

Inferred from the tables, we can see that the error of the state variable y and co-state variable p are the second order accuracy with respect to the time and space sizes, whereas the error of the control variable u is only first order to the spatial variable because of the modeling space.

Table 3 $L^2(0,T;L^2(\Omega))$-norm error for PDDIA ($r=9$)

Grids	$y-Y$	order	$u-U$	order	$p-P$	order
36×36	1.939×10^{-2}		1.588×10^{-2}		1.964×10^{-2}	
49×49	1.173×10^{-2}	1.63	1.006×10^{-2}	1.48	1.186×10^{-2}	1.63
64×64	7.137×10^{-3}	1.86	6.623×10^{-3}	1.57	7.202×10^{-3}	1.87
81×81	4.429×10^{-3}	2.03	4.678×10^{-3}	1.47	4.453×10^{-3}	2.04

In addition, we could get a brief relationship about the Δt-H constraint. Because one can take more larger H than h for keeping the optimal order accuracy for the spatial variable, the constraint $\Delta t = O(H^2)$ is less severe than that for fully explicit algorithms.

4 Conclusion

In this paper, an efficient domain decomposition algorithm for an optimal control problem governed by a linear parabolic partial differential equation has been proposed. The algorithm can solve coupled optimality condition accurately and efficiently based on the non-overlapping domain decomposition scheme given in [4]. The efficient calculation strategy on the inner boundaries and the outer iterations enable excellent extensibility and usability in parallel. Because of the implict/explict strategy, it is necessary to preserve stability from the explicit prediction, but less severe than that for fully explicit algorithms. Further, second order convergence in time allow us to use larger time step in calculations.

References

1. C.N.Dawson and T.F.Dupont, Explicit/implict conservative Galerkin domain decomposition procedures for parabolic Problems, Math.Comput. 58(197)(1992), 21—34.
2. D.Yang, Parallel domain decomposition procedures of improved D-D type for parabolic problems, Comput. Appl. Math., 233(11)(2010), 2779—2794.
3. B.Zhang, J.Chen and D.Yang, Parallel D-D type domain decomposition algorithm for optimal control problem governed by parabolic partial differential equation, J. Numer. Math., 25(1) (2017), 35—53.
4. J.Chen, D.Yang, Parallel Crank-Nicolson and Dawson-Dupont domain decomposition procedures for parabolic equations, in preparation.
5. J.Chen, D.Yang, Parallel Crank-Nicolson and Dawson-Dupont type domain decomposition procedure for optimal control problem governed by parabolic equation, in preparation.
6. K.Ma, T.Sun, Galerkin domain decomposition procedures for parabolic equations on rectangular domain, Internat. J. Numer. Methods Fluids 62(4)(2010), 449—472.

Partition of Unity Methods for Heterogeneous Domain Decomposition

Gabriele Ciaramella and Martin J. Gander

1 Heterogeneous problems and partition of unity decomposition

We are interested in solving linear PDEs of the form

$$\mathscr{L}(u) = f \text{ in } \Omega,\ u = \widetilde{g} \text{ on } \partial\Omega, \tag{1}$$

where Ω is a bounded domain in $\mathbb{R}^d$ with $d = 1, 2$, $\mathscr{L}$ is a linear (elliptic) differential operator, f and $\widetilde{g}$ are the data, and u is the solution to (1). The weak form of (1) with a Hilbert space $(V, \langle\cdot,\cdot\rangle)$ of functions $v : \Omega \to \mathbb{R}$ is

$$a(u,v) = \ell(v)\ \forall v \in V_0, \text{ with } u = \widetilde{g} \text{ on } \partial\Omega, \tag{2}$$

where $V_0 := \{v \in V\ :\ v = 0 \text{ on } \partial\Omega\}$, $a : V \times V \to \mathbb{R}$ is the bilinear form corresponding to the operator $\mathscr{L}$, and $\ell : V \to \mathbb{R}$ is the linear functional induced by f. We assume that (2) has a unique solution $u \in \{v \in V\ :\ v = \widetilde{g} \text{ on } \partial\Omega\}$, and that u is "heterogeneous", behaving very differently in different parts of Ω. Typical examples are advection-diffusion problems, where there are advection dominated and diffusion dominated regions (subdomains), and the boundaries in between are not clearly defined, see [8, 10] and references therein. Apart from the χ-method [6, 1], there are no methods to determine such subdomain decompositions, and our goal is to present and study a new such method. We thus introduce (see [18, 11])

Definition 1 (Membership function). Let $\Omega \subset \mathbb{R}^d$ be a set. A membership function φ is a map $\varphi : \overline{\Omega} \to [0,1]$, and its support $S \subset \overline{\Omega}$ is $S := \overline{\{x \in \Omega\ :\ \varphi(x) \neq 0\}}$.

Given two membership functions $\varphi_1, \varphi_2 : \overline{\Omega} \to [0,1]$ that form a partition of unity on Ω, $\varphi_1(x) + \varphi_2(x) = 1$ for all $x \in \overline{\Omega}$, their supports provide then a domain decompo-

G. Ciaramella
Universität Konstanz, Germany, e-mail: gabriele.ciaramella@uni-konstanz.de

M. J. Gander
Université de Genève, Genève, Switzerland e-mail: martin.gander@unige.ch

sition $\overline{\Omega} = \overline{\text{supp}\varphi_1 \cup \text{supp}\varphi_2}$. We introduce the approximation $u_{dd} := \varphi_1 u_1 + \varphi_1 u_2 \approx u$, where u_1 and u_2 represent two different possible behaviors of u, and we assume that $u_{dd} = \widetilde{g}$ on $\partial\Omega$. We proceed as follows to define the spaces that u_{dd}, u_1 and u_2 are to be sought in: first, we introduce two approximate problems,

$$\mathscr{L}_1(u_1) = f_1 \text{ in } \Omega, \mathscr{B}_1(u_1, \widetilde{g}) = g, \quad \text{and} \quad \mathscr{L}_2(u_2) = f_2 \text{ in } \Omega, \mathscr{B}_2(u_2, \widetilde{g}) = 0. \quad (3)$$

Here $\mathscr{L}_j$ are approximation operators of $\mathscr{L}$, f_j are approximations of f, and $\mathscr{B}_j$ are operators to define the boundary conditions of (3), see Section 3 for concrete examples. The function g represents a control and belongs to an appropriate Hilbert space W. Notice that g is different from the actual boundary data $\widetilde{g}$: the latter is defined on $\partial\Omega$, while we will define the former only on a subset of $\partial\Omega$. We assume[1] that (3) (left) is uniquely solvable in V for any $g \in W$ and (3) (right) has a unique solution $u_2 \in V$. To reformulate (3) (left), we introduce two operators $A : V \to V^*$ and $B_{\widetilde{g}} : W \to V^*$, such that (3) (left) becomes $Au_1 = B_{\widetilde{g}} g + f_1$. Notice that $B_{\widetilde{g}}$ represents the boundary conditions of (3) (left) and takes into account also $\widetilde{g}$. This problem is formally solved by $u_1 = A^{-1} B_{\widetilde{g}} g + A^{-1} f_1$, where A^{-1} is well defined if (3) (left) is well posed. Now, we define the spaces $V_1 := \{v \in V : v = A^{-1}(B_{\widetilde{g}} q + f_1), q \in W\}$, and $V_2 := \{u_2\}$. Here V_1 represents the space of all possible solutions to the first problem in (3) generated by all the possible (control) functions in W, while V_2 is a singleton containing only the unique solution u_2 to (3) (right). Finally, we use the definition of a "partition of unity method" space (PUM-space [2, 13]) $V_{PUM} := \varphi_1 V_1 + \varphi_2 V_2 \subset V$, where φ_1, φ_2 are membership functions. V_{PUM}, V_1 and V_2 are the spaces that the approximations u_{dd}, u_1 and u_2 have to be sought in. In particular, for the approximation u_{dd} the functions φ_1, φ_2 and g have to be computed. These are defined as solutions to optimal control problems, as described in Section 2. Here we need to remark that our approach could be computationally expensive. However, it is motivated by applications in astrophysics governed by hyperbolic equations like the Boltzmann equation. In many cases, like for supernova explosion, physical phenomena are modeled using two different (limiting) regimes. However, this would require an a-priori knowledge of the transition regime; see, e.g. [8, 3, 11] and references therein. This is exactly the role of the partition of unity functions obtained by our computational framework. In practice, one could use our computationally expensive approach to obtain the partition of unity functions for one representative case and then reuse them (as approximations) in a domain decomposition fashion to compute approximate solutions of other cases of interest.

[1] This specific approximation is motivated by asymptotic expansion techniques providing in general two problems, one that is uniquely determined and a second one that is determined up to some constants for asymptotic matching [15].

2 Optimal control approaches

To compute φ_1, φ_2 and g, we embed the PUM formulation into an optimal control framework. We begin by inserting u_{dd} into (2) and obtain the bounded linear functional $r : V \to \mathbb{R}$ defined by $r(v) := a(\varphi_1 u_1 + \varphi_2 u_2, v) - \ell(v)$, where $v \in V$. In the case that $v = \varphi_1 w$ and $v = \varphi_2 w$ with $w \in V$, we get the functionals

$$r_j(w) := r(\varphi_j w) = a(\varphi_1 u_1 + \varphi_2 u_2, \varphi_j w) - \ell(\varphi_j w) \quad \forall w \in V, \text{ for } j = 1,2.$$

Since $w \in V \mapsto r_j(w)$, $j = 1,2$, are bounded linear functionals, they are elements in V^*, and by the Riesz representation theorem [7], there exist R_1 and R_2 in V such that

$$\langle R_j, v\rangle = a(\varphi_1 u_1 + \varphi_2 u_2, \varphi_j v) - \ell(\varphi_j v) \quad \forall v \in V_0, \quad j = 1,2, \tag{4}$$

where we used V_0, since u_{dd} is exact on $\partial\Omega$ and thus R_1 and R_2 must vanish there. Now, we define $\varphi := \varphi_1$ with $\varphi_2 = 1 - \varphi$, and recall that $\|r_j\|_{V^*} = \|R_j\|_V$. Minimizing the norms of the residuals $\|R_j\|_V$ leads to the optimal control problem

$$\begin{aligned} \min_{R_1,R_2,u_1,g,\varphi} \; & J(R_1,R_2,g,\varphi) := \frac{1}{2}\|R_1\|_V^2 + \frac{1}{2}\|R_2\|_V^2 + \frac{\alpha}{2}\|\varphi\|_V^2 + \frac{\beta}{2}\|g\|_W^2 \\ \text{s.t. } & \langle R_1, v\rangle = a(\varphi u_1 + (1-\varphi)u_2, \varphi v) - \ell(\varphi v) \quad \forall v \in V_0, \\ & \langle R_2, v\rangle = a(\varphi u_1 + (1-\varphi)u_2, (1-\varphi)v) - \ell((1-\varphi)v) \quad \forall v \in V_0, \\ & Au_1 = B_{\tilde{g}} g + f_1,\ g \in W,\ u_2 \in V_2,\ \varphi \in V,\ 0 \le \varphi \le 1 \text{ a.e. in } \Omega, \end{aligned} \tag{5}$$

where $\alpha, \beta > 0$ are two regularization parameters used to tune the cost of φ and g, and f_1 is the same approximation to f introduced in (3).

Solving (5) by an iterative procedure [5, 17] requires at each iteration to solve the two equations (4) for R_1 and R_2, and (3) for u_1. A less expensive optimal control problem is obtained by summing (4) for $j = 1,2$, and we obtain with $R := R_1 + R_2$

$$\langle R, v\rangle = a(\varphi u_1 + (1-\varphi)u_2, v) - \ell(v) \quad \forall v \in V_0, \tag{6}$$

which is a Petrov-Galerkin type equation that we could have obtained directly applying a Petrov-Galerkin method to (2) using V_{PUM} and V as trial and test spaces. Using (6), we get the less expensive optimal control problem

$$\begin{aligned} \min_{R,u_1,g,\varphi} \; & J(R,g,\varphi) := \frac{1}{2}\|R\|_V^2 + \frac{\alpha}{2}\|\varphi\|_V^2 + \frac{\beta}{2}\|g\|_W^2 \\ \text{s.t. } & \langle R, v\rangle = a(\varphi u_1 + (1-\varphi)u_2, v) - \ell(v) \quad \forall v \in V_0, \\ & Au_1 = B_{\tilde{g}} g + f_1,\ g \in W,\ u_2 \in V_2,\ \varphi \in V,\ 0 \le \varphi \le 1 \text{ a.e. in } \Omega. \end{aligned} \tag{7}$$

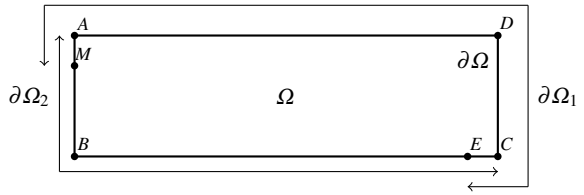


Fig. 1 Example of a boundary decomposition $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$.

3 Optimal control for elliptic boundary-layer problems

As main test cases we consider elliptic problems of the form

$$\mathscr{L}(u) := -\mu\Delta u + \mathbf{a}\cdot\nabla u + c\,u = f \text{ in } \Omega,\ u = \widetilde{g} \text{ on } \partial\Omega, \tag{8}$$

where Ω is a bounded domain in $\mathbb{R}^d$, for $d = 1,2$, $\widetilde{g} \in C(\partial\Omega)$, f is sufficiently smooth, and the components of $\mathbf{a}$ are assumed to be strictly-positive. The assumption on $\mathbf{a}$ is restrictive, but it simplifies the presentation below and can be relaxed. The corresponding weak problem is to find a $u \in \left\{v \in H^1(\Omega)\,|\,v = \widetilde{g} \text{ on } \partial\Omega\right\}$ such that

$$a(u,v) := \int_\Omega \mu\nabla u\cdot\nabla v + \mathbf{a}\cdot\nabla u v + c\,u\,v\,d\mathbf{x} = \int_\Omega f\,v\,d\mathbf{x} =: \ell(v)\ \ \forall v \in H^1_0(\Omega).$$

We also assume that Ω is such that the boundary $\partial\Omega$ can be decomposed into $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, where the intersection $\partial\Omega_1 \cap \partial\Omega_2$ has a non-zero measure, as illustrated in Figure 1. To obtain $u_{dd} = \varphi u_1 + (1-\varphi)u_2 \approx u$, we define $\Gamma := \partial\Omega \setminus \partial\Omega_1$ and introduce the operator $\mathscr{L}_1 := -\mu\Delta + c$. Then, as in (3), for any choice of the control $g \in H^1_0(\Gamma)$ the corresponding approximate problem for u_1 is

$$\begin{aligned} &\int_\Omega \mu\nabla u_1\cdot\nabla v + c\,u_1\,v\,d\mathbf{x} = 0 \quad \forall v \in H^1_0(\Omega),\\ &u_1 \in \left\{w \in H^1(\Omega)\,|\,w = \widetilde{g} \text{ on } \partial\Omega_1,\ w = \widetilde{g} + g \text{ on } \Gamma,\ \tau(w) \in C(\partial\Omega)\right\}, \end{aligned} \tag{9}$$

where τ is the trace operator on $\partial\Omega$. Notice that we have chosen $f_1 = 0$. As before, we introduce the operator $A : H^1(\Omega) \to H^{-1}(\Omega)$ defined as $\langle Au, v\rangle_{H^{-1},H^1} := \int_\Omega \mu\nabla u\cdot\nabla v + c\,u\,v\,d\mathbf{x}$ for all $v \in H^1_0(\Omega)$, and the operator $B_{\widetilde{g}} : H^1_0(\Gamma) \to H^{-1}(\Omega)$ such that $v \mapsto (B_{\widetilde{g}}g)(v)$ is a bounded linear functional in $H^{-1}(\Omega)$. The operator $B_{\widetilde{g}}$ represents the Dirichlet boundary conditions of (9). $Au_1 = B_{\widetilde{g}}g$ is then equivalent to (9). The corresponding set V_1 is given by

$$V_1 = \{v \in H^1(\Omega)\ :\ Av = B_{\widetilde{g}}q \text{ for any } q \in H^1_0(\Gamma)\}.$$

Now, consider the operator $\mathscr{L}_2 := \mathbf{a}\cdot\nabla + c$ and $f_2 = f$. The problem for u_2 is then

$$\mathscr{L}_2(u_2) = \mathbf{a}\cdot\nabla u_2 + c\,u_2 = f \text{ in } \Omega,\ u_2 = \widetilde{g} \text{ on } \partial\Omega_2, \tag{10}$$

which we assume uniquely solvable in $H^1(\Omega)\cap C(\overline{\Omega})$. Notice that (10) is a pure advection problem and the subset $\partial\Omega_2$ is given as the set of points where the characteristic curves enter the domain Ω. This is the main assumption we make on $\partial\Omega_2$ for the problem (10) to be well posed. The set V_2 contains only the solution to (10), i.e. $V_2=\{u_2\}$. The approximation $u_{dd}\approx u$ is then obtained as $u_{dd}=\varphi_1 u_1+\varphi_2 u_2$, where the membership functions $\varphi_1=\varphi, \varphi_2=1-\varphi\in H^1(\Omega)$ form a partition of unity, and φ is such that

$$\varphi(\mathbf{x})\in\begin{cases}\{1\} & \text{if } \mathbf{x}\in\partial\Omega\setminus\partial\Omega_2,\\ [0,1] & \text{if } \mathbf{x}\in\partial\Omega_1\cap\partial\Omega_2,\\ \{0\} & \text{if } \mathbf{x}\in\Gamma,\end{cases} \tag{11}$$

with $\tau(\varphi)\in C(\partial\Omega)$. Notice that this definition of φ makes u_{dd} exact on the boundary $\partial\Omega$, $\tau(u_{dd})=\tau(\varphi_1 u_1+\varphi_2 u_2)=\widetilde{g}$.

In what follows, we study the control problem (7) ((5) would have a similar structure) to optimize φ and g for computing the approximation u_{dd} to the solution to (8). In particular, we first show well-posedness, and then we derive the first-order optimality system. We consider directly a 2-dimensional problem ($d=2$), since the analysis of the 1-dimensional version is simpler and relies on the same arguments. To define our optimal control problem, as in (7), we consider the cost functional $J(R,g,\varphi):=\frac{1}{2}\|R\|^2_{H^1(\Omega)}+\frac{\alpha}{2}\|\varphi\|^2_{H^1(\Omega)}+\frac{\beta}{2}\|g\|^2_{H^1(\Gamma)}$. Now, we introduce the control-to-state maps $g\mapsto u_1(g)$ and $(g,\varphi)\mapsto R(u_1(g),\varphi)$, where $u_1(g)$ and $R(u_1(g),\varphi)$ solve (9) and

$$\langle R,v\rangle_{H^1(\Omega)}=\int_\Omega \mu\nabla u_{dd}\cdot\nabla v+\mathbf{a}\cdot\nabla u_{dd}\,v+c\,u_{dd}\,v-f\,v\,d\mathbf{x}\quad\forall v\in H^1_0(\Omega). \tag{12}$$

Notice that the left-hand side of (12), that is $\langle R,v\rangle_{H^1(\Omega)}=\int_\Omega \nabla R\cdot\nabla v+Rv\,d\mathbf{x}$, is of a similar form to the left-hand side in (9). These maps are well defined according to the lemmas below and allow us to define the reduced cost functional $\widetilde{J}(g,\varphi):=J(R(u_1(g),\varphi),g,\varphi)$ and the optimal control problem

$$\min_{g,\varphi}\ \widetilde{J}(g,\varphi)\ \text{s.t. } 0\le\varphi(\mathbf{x})\le 1 \text{ in } \Omega \text{ and (11) holds.} \tag{13}$$

For well-posedness of this optimization problem, we need four Lemmas:

Lemma 1. *Let $z\in H^1(\partial\Omega)$ with $\Omega\subset\mathbb{R}^2$ convex and $\partial\Omega$ Lipschitz. Then the problem*

$$\int_\Omega \mu\nabla u_1\cdot\nabla v+c\,u_1\,v\,d\mathbf{x}=0\quad\forall v\in H^1_0(\Omega) \tag{14}$$

with $u_1=z$ on $\partial\Omega$ is uniquely solvable by $u_1\in H^1(\Omega)\cap C(\overline{\Omega})$, and there exists a positive constant c such that $\|u_1\|_{H^1(\Omega)}\le c\|z\|_{H^1(\partial\Omega)}$.

Proof. To show that there exists a unique $u_1\in C(\overline{\Omega})$, we define w as the harmonic extension of z in Ω. Recalling the embedding $H^1\hookrightarrow C$ for one-dimensional domains, we have that $z\in C(\partial\Omega)$. Therefore, since Ω is a Lipschitz domain,

$w \in C^2(\Omega) \cap C(\overline{\Omega})$; see, e.g., [12]. Now, consider the problem $-\mu\Delta v + cv = -cw$ in Ω with $v = 0$ on $\partial\Omega$. Since Ω is convex, Theorems 3.2.1.2-3 in [14] ensure that this problem is uniquely solved by $v \in H^2(\Omega) \cap H_0^1(\Omega)$. Since $\Omega \subset \mathbb{R}^2$, the Sobolev embedding $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$ [7] ensures that $v \in C(\overline{\Omega})$. Noticing that the function $w+v$ solves (14), $u_1 \in C(\overline{\Omega})$ and is unique by the linearity of (14). Next, we show that $u_1 \in H^1(\Omega)$ with $\|u_1\|_{H^1(\Omega)} \leq c\|z\|_{H^1(\partial\Omega)}$. Consider the trace operator $\tau : H^1(\Omega) \to H^{1/2}(\partial\Omega)$. Since Ω is a Lipschitz domain, by [16, Theorem 3.37, page 102] this operator has a bounded right-inverse $\tau^{-1} : H^{1/2}(\partial\Omega) \to H^1(\Omega)$. Now, we define $w := \tau^{-1}z$ and note that $w \in H^1(\Omega)$. So, if we decompose u_1 as $u_1 = w + \widetilde{v}$, then $\widetilde{v}$ must solve in a weak sense the problem $-\mu\Delta\widetilde{v} + c\widetilde{v} = -(-\mu\Delta w + cw)$ in Ω with $\widetilde{v} = 0$ on $\partial\Omega$. By the Lax-Milgram theorem we have that the unique solution is $\widetilde{v} \in H_0^1(\Omega)$ and there exists a constant C such that $\|\widetilde{v}\|_{H^1(\Omega)} \leq C\|w\|_{H^1(\Omega)}$. Therefore, $u_1 \in H^1(\Omega)$ and using the decomposition $u_1 = w + \widetilde{v}$ we get $\|u_1\|_{H^1(\Omega)} \leq (1+C)\|w\|_{H^1(\Omega)} = (1+C)\|\tau^{-1}z\|_{H^1(\Omega)} \leq K\|z\|_{H^1(\partial\Omega)}$, for some positive constant K, where we used the boundedness of τ^{-1} [16].

Lemma 2. *Let $\varphi \in H^1(\Omega)$ such that $0 \leq \varphi(\mathbf{x}) \leq 1$ a.e. in Ω. Then for any function $v \in H^1(\Omega) \cap C(\overline{\Omega})$ it holds that $\varphi v \in H^1(\Omega)$.*

Proof. An application of Theorem 1 in [9, page 247] shows that $\nabla(v\varphi) = v\nabla\varphi + \varphi\nabla v$. Then a simple estimate of the norm $\|\nabla(v\varphi)\|_{L^2(\Omega)}$ allows us to obtain the result.

Lemma 3. *Let $\{z_n\}_n$ be a sequence that converges weakly in $H^1(\partial\Omega)$ to a weak limit $\widehat{z} \in H^1(\partial\Omega)$, i.e. $z_n \rightharpoonup \widehat{z}$ in $H^1(\partial\Omega)$. Define the sequence $\{u_{1,n}\}_n$ by $u_{1,n} := u_1(z_n)$, where $u_1(z_n)$ solves (14) with $u_1 = z_n$ on $\partial\Omega$. Then there exists a subsequence u_{1,n_j} that converges weakly in $H^1(\Omega)$ and strongly in $L^2(\Omega)$ to the limit $\widehat{u}_1 = u_1(\widehat{z}) \in H^1(\Omega)$, i.e., $u_{1,n_j} \rightharpoonup \widehat{u}_1$ in $H^1(\Omega)$ and $u_{1,n_j} \to \widehat{u}_1$ in $L^2(\Omega)$.*

Proof. Since the sequence $\{z_n\}_n$ converges weakly in $H^1(\partial\Omega)$, it is bounded in the norm $\|\cdot\|_{H^1(\partial\Omega)}$. By Lemma 1, we have that $\|u_{1,n}\|_{H^1(\Omega)} \leq c\|z_n\|_{H^1(\partial\Omega)} \leq K$, for some positive constant K, and the sequence $u_{1,n}$ is bounded in $H^1(\Omega)$. Since $H^1(\Omega)$ is reflexive, there exists a weakly convergent subsequence $u_{1,n_j} \rightharpoonup \widehat{u}_1$ in $H^1(\Omega)$. Now, from (14), we have that for any $v \in H_0^1(\Omega)$

$$\int_\Omega \mu\nabla u_{1,n_j} \cdot \nabla v + c\, u_{1,n_j} v\, d\mathbf{x} \to \int_\Omega \mu\nabla\widehat{u}_1 \cdot \nabla v + c\,\widehat{u}_1 v\, d\mathbf{x}.$$

Moreover, the weak convergence $z_{n_j} \rightharpoonup \widehat{z}$ and the continuity of the trace operator $\tau : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ [16, Theorem 3.37] implies that $z_{n_j} = \tau(u_{1,n_j}) \rightharpoonup \tau(\widehat{u}_1) = \widehat{z}$, weakly in $H^{1/2}(\partial\Omega)$. Therefore, $\widehat{u}_1 = u_1(\widehat{z})$. We conclude by recalling the Sobolev compact embedding $H^1(\Omega) \Subset L^2(\Omega)$; see, e.g., [7].

Lemma 4. *Let $\{u_{1,n}\}_n$ be the sequence defined in Lemma 3 such that $u_{1,n_j} \rightharpoonup \widehat{u}_1$ (weakly) in $H^1(\Omega)$. Consider a sequence $\{\varphi_n\}_n$ in $H^1(\Omega)$ such that $0 \leq \varphi_n(\mathbf{x}) \leq 1$*

and $\varphi_n \rightharpoonup \widehat{\varphi}$ *(weakly) in* $H^1(\Omega)$ *with* $0 \leq \widehat{\varphi}(\mathbf{x}) \leq 1$. *Then there exist two subsequences* $\{\varphi_{n_j}\}_j$ *and* $\{u_{1,n_j}\}_j$ *such that* $\varphi_{n_j} \to \widehat{\varphi}$ *and* $u_{1,n_j} \to \widehat{u}_1$ *(strongly) in* $L^2(\Omega)$, *and for any* $v \in H_0^1(\Omega)$

$$\int_\Omega \nabla(\varphi_{n_j} u_{1,n_j}) \cdot \nabla v + \varphi_{n_j} u_{1,n_j} v d\mathbf{x} \to \int_\Omega \nabla(\widehat{\varphi}\widehat{u}_1) \cdot \nabla v + \widehat{\varphi}\widehat{u}_1 v d\mathbf{x}.$$

Proof. The existence of the subsequences $\{\varphi_{n_j}\}_j$ and $\{u_{1,n_j}\}_j$ such that $\varphi_{n_j} \to \widehat{\varphi}$ and $u_{1,n_j} \to \widehat{u}_1$ (strongly) in $L^2(\Omega)$ follows from the fact that $\varphi_n \rightharpoonup \widehat{\varphi}$ (weakly in $H^1(\Omega)$), Lemma 3, and the Sobolev (compact) embedding $H^1(\Omega) \Subset L^2(\Omega)$ [7]. Now, recalling Lemma 1 and according to the proof of Lemma 2 it holds that $\nabla(u_{1,n_j}\varphi_{n_j}) = u_{1,n_j}\nabla\varphi_{n_j} + \varphi_{n_j}\nabla u_{1,n_j}$. Therefore, to treat the products of sequences $\widehat{u}_{1.n_j}\nabla\widehat{\varphi}_{n_j}$, $\varphi_{n_j}\nabla u_{1,n_j}$, and $\varphi_{n_j}u_{1,n_j}$, we use [7, Theorem 5.12-4] to obtain for any $v \in H_0^1(\Omega)$ that

$$\begin{aligned}
&\int_\Omega \nabla(\varphi_{n_j} u_{1,n_j})\nabla v + \varphi_{n_j} u_{1,n_j} v d\mathbf{x} = \int_\Omega u_{1,n_j}\nabla\varphi_{n_j}\nabla v + \varphi_{n_j}\nabla u_{1,n_j}\nabla v + \varphi_{n_j} u_{1,n_j} v d\mathbf{x} \\
&\to \int_\Omega \widehat{u}_1\nabla\widehat{\varphi}\nabla v + \widehat{\varphi}\nabla\widehat{u}_1\nabla v + \widehat{u}_1\widehat{\varphi} v d\mathbf{x} = \int_\Omega \nabla(\widehat{\varphi}\widehat{u}_1)\cdot\nabla v + \widehat{\varphi}\widehat{u}_1 v + \widehat{u}_1\widehat{\varphi} v d\mathbf{x}.
\end{aligned}$$

We are now ready to prove that (13) is well posed.

Theorem 1. *Let* $\alpha, \beta > 0$, *then there exists a solution to problem* (13).

Proof. Consider a minimizing sequence $\{(R_n, \varphi_n, u_{1,n}, g_n)\}_n$, where g_n is extended by zero on $\partial\Omega$. Since J is coercive in φ and g we have the bounds $\|\varphi_n\|_{H^1(\Omega)} \leq c$ and $\|g_n\|_{H^1(\partial\Omega)} \leq c'$, for two positive constants c, c'; see, e.g., [17]. The reflexivity of $H^1(\Omega)$ and $H^1(\partial\Omega)$ ensures the existence of weakly convergent subsequences: $\varphi_{n_j} \rightharpoonup \widehat{\varphi}$ in $H^1(\Omega)$ and $g_{n_j} \rightharpoonup \widehat{g}$ in $H^1(\partial\Omega)$. By the Sobolev (compact) embedding $H^1(\Omega) \Subset L^2(\Omega)$ [7], the sequence $\{\varphi_{n_j}\}_j$ converges strongly in $L^2(\Omega)$ to $\widehat{\varphi}$. Since the set $\{v \in L^2(\Omega) : 0 \leq v(\mathbf{x}) \leq 1 \text{ a.e. in } \Omega\}$ is (weakly) closed in $L^2(\Omega)$ [17], we have $0 \leq \widehat{\varphi}(\mathbf{x}) \leq 1$. Consider now the sequence $\{u_{1,n}\}_n$ and the corresponding subsequence $u_{1,n_j} = u_1(g_{n_j})$. By Lemma 3, we have that $u_{1,n_j} \rightharpoonup \widehat{u}_1 = u_1(\widehat{g})$ weakly in $H^1(\Omega)$ and $u_{1,n_j} \to \widehat{u}_1 = u_1(\widehat{g})$ strongly in $L^2(\Omega)$. Consider the sequence $\{R_n\}_n$. Since R_n satisfies

$$\langle R_n, v\rangle_{H^1(\Omega)} = \int_\Omega \mu\nabla u_{dd,n}\cdot\nabla v + \mathbf{a}\cdot\nabla u_{dd,n} v + c\, u_{dd,n} v - f v d\mathbf{x} \;\forall v \in H_0^1(\Omega),$$

where $u_{dd,n} = \varphi_n u_{1,n} + (1-\varphi_n)u_2$, from the Lax-Milgram theorem we have that $\|R_n\|_{H^1(\Omega)} \leq K(\|u_{1,n}\|_{H^1(\Omega)}, \|\varphi_n\|_{H^1(\Omega)})$, where the constant K depends on $\|u_{1,n}\|_{H^1(\Omega)}$ and $\|\varphi_n\|_{H^1(\Omega)}$, which are bounded. Therefore, R_n is bounded as well, and by Lemma 4, one can show that $R_{n_j} \rightharpoonup \widehat{R} = R(\widehat{u}_1, \widehat{\varphi})$ weakly in $H^1(\Omega)$. Now, the weak-lower semi-continuity of J implies the claim [17, 4].

To obtain the first-order optimality system, we rely on the Lagrange multiplier approach and work in the reduced space of solutions of constraint and adjoint equa-

tions; see, e.g., [5, 17]. We first recall the control-to-state maps $g \mapsto u_1(g)$ and $(g,\varphi) \mapsto R(u_1(g),\varphi)$ and the reduced cost functional $\widetilde{J}(g,\varphi)$. Then we notice that its derivatives, for $\delta g \in H_0^1(\Gamma)$ and $\delta\varphi \in H_0^1(\Omega)$, are

$$D_g\widetilde{J}(g,\varphi)(\delta g) = \langle \beta g + R_g, \delta g\rangle_{H^1(\Gamma)}, \quad D_\varphi\widetilde{J}(w,\varphi)(\delta\varphi) = \langle \alpha\varphi + R_\varphi, \delta\varphi\rangle_{H^1(\Omega)}. \tag{15}$$

Here R_g is the solution of the problem

$$\langle R_g, \delta g\rangle_{H_0^1(\Gamma)} = \langle B_{\widetilde{g}}\delta g, \lambda\rangle_{H^{-1},H^1}, \tag{16}$$

where $\langle\cdot,\cdot\rangle_{H^{-1},H^1} : H^{-1}(\Omega)\times H_0^1(\Omega) \to \mathbb{R}$ denotes the duality pairing, and R_φ is the Riesz representative of the linear functional

$$\delta\varphi \mapsto \int_\Omega \mu\nabla\big[(u_1-u_2)\delta\varphi\big]\cdot\nabla R\, d\mathbf{x} + \mathbf{a}\cdot\nabla\big[(u_1-u_2)\delta\varphi\big]R + c\,(u_1-u_2)\delta\varphi\, R\, d\mathbf{x}.$$

In (16), $\lambda \in H_0^1(\Omega)$ is a Lagrange multiplier that solves the adjoint equation

$$\int_\Omega \nabla\lambda\cdot\nabla v + c\,\lambda\, v\, d\mathbf{x} = \int_\Omega \mu\nabla(v\varphi)\cdot\nabla R + \mathbf{a}\cdot\nabla(v\varphi)\,R + c\,v\,\varphi\, R\, d\mathbf{x}, \tag{17}$$

for all $v \in H_0^1(\Omega)$. Therefore, the first-order optimality system is given by (9), (12), (17) and (16) together with the conditions [4, 17]

$$D_g\widetilde{J}(g,\varphi)(\delta g) = 0,$$

for all $\delta g \in H_0^1(\Gamma)$, and for any arbitrary $\theta > 0$

$$\varphi = \mathbb{P}_{V_{ad}}\Big(\varphi - \theta\big(\alpha\varphi + R_\varphi\big)\Big),$$

where $\mathbb{P}_{V_{ad}}$ is the projection onto $V_{ad} := \{v \in H^1(\Omega) : 0 \le v(\mathbf{x}) \le 1 \text{ a.e. in } \Omega\}$.

4 Numerical experiments

We present now numerical experiments for the one-dimensional elliptic problem

$$-\mu\partial_{xx}u - \partial_x u = 1 \text{ in } (0,1), \text{ with } u(0) = 0,\, u(1) = 0, \tag{18}$$

for given $\mu = 0.01$, computing $u_{dd} = \varphi_1 u_1 + \varphi_2 u_2$, with

$$\begin{aligned} -\mu\partial_{xx}u_1 &= 0 \text{ in } (0,1), \\ u_1(0) &= 0,\, u_1(1) = g, \end{aligned} \quad \text{and} \quad \begin{aligned} -\partial_x u_2 &= 1 \text{ in } [0,1), \\ u_2(1) &= 0. \end{aligned}$$

We solve both the PUM and Petrov-Galerkin optimality systems discretized by linear finite-elements with a projected-LBFGS method with stopping tolerance $5\cdot 10^{-5}$

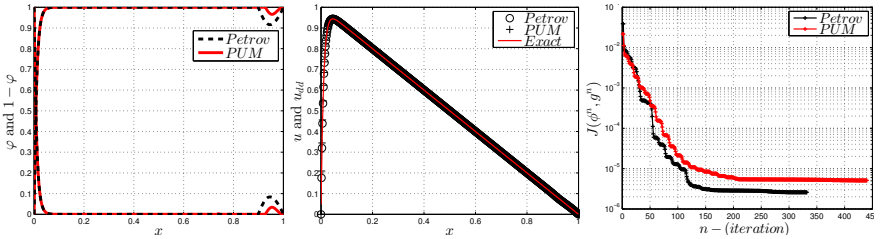


Fig. 2 Comparison of the Petrov and PUM approaches: Left: partition of unity functions φ and $1-\varphi$. Middle: exact solution and approximations. Right: Decay of the cost functional.

on the (relative) residual norm. The regularization parameters are $\alpha = \beta = 10^{-7}$. In Figure 2 (left) we see that the φ and $1-\varphi$ obtained by the two approaches are very similar, and catch well the boundary layer on the left. The small bumps in the right part (close to $x = 1$) are due numerical effects and we checked that they disappear for smaller tolerances. In Figure 2 (middle) the exact solution is compared with the two approximations u_{dd}, and we see good agreement. In Figure 2 (right), we show the decay of the cost functional with respect to the number of iterations, and we see that the Petrov-Galerkin approach converges a bit faster.

References

1. Y. Achdou and O. Pironneau. The χ-method for the Navier-Stokes equations. *IMA journal of numerical analysis*, 13(4):537–558, 1993.
2. I. Babuska and J. M. Melenk. The partition of unity method. *International Journal of Numerical Methods in Engineering*, 40:727–758, 1996.
3. H. Berninger, E. Frnod, M. Gander, M. Liebendrfer, and J. Michaud. Derivation of the isotropic diffusion source approximation (idsa) for supernova neutrino transport by asymptotic expansions. *SIAM Journal on Mathematical Analysis*, 45(6):3229–3265, 2013.
4. A. Borzì, G. Ciaramella, and M. Sprengel. *Formulation and Numerical Solution of Quantum Control Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.
5. A. Borzì and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. SIAM, Philadelphia, 2012.
6. F. Brezzi, C. Canuto, and A. Russo. A self-adaptive formulation for the Euler-Navier Stokes coupling. *Comput. Methods Appl. Mech. Eng.*, 73:317–330, 1989.
7. P. G. Ciarlet. *Linear and Nonlinear Functional Analysis with Applications*. SIAM, Philadelphia, 2013.
8. P. Degond and S. Jin. A smooth transition model between kinetic and diffusion equations. *SIAM J. Numerical Analysis*, 42(6):2671–2687, 2005.
9. L. C. Evans. *Partial differential equations*. Graduate studies in mathematics. American Mathematical Society, Providence (R.I.), 2002.
10. M. J. Gander, L. Halpern, and V. Martin. A new algorithm based on factorization for heterogeneous domain decomposition. *Numerical Algorithms*, 73(1):167–195, 2016.
11. M.J. Gander and J. Michaud. Fuzzy domain decomposition: a new perspective on heterogeneous DD methods. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 265–273. Springer, 2014.
12. D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin, New York, 1983.

13. M. Griebel and M. A. Schweitzer. A particle-partition of unity method for the solution of elliptic, parabolic, and hyperbolic pdes. *SIAM Journal on Scientific Computing*, 22(3):853–890, 2000.
14. P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Monographs and studies in mathematics 24. Pitman Advanced Publishing Program, Boston, London, Melbourne, 1985.
15. M.H. Holmes. *Introduction to Perturbation Methods*. Texts in Applied Mathematics. Springer New York, 2013.
16. W.C.H. McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2000.
17. F. Tröltzsch. *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*. Grad. Stud. Math. 112. American Mathematical Society, Providence, RI, 2010.
18. L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

Integral equation based optimized Schwarz method for electromagnetics

Xavier Claeys[1], Bertrand Thierry[2], and Francis Collino[3]

1 Introduction

The optimized Schwarz method (OSM) is recognized as one of the most efficient domain decomposition strategies without overlap for the solution to wave propagation problems in harmonic regime. For the Helmholtz equation, this approach originated from the seminal work of Després [4, 5], and led to the development of an abundant literature offering more elaborated but more efficient transmission conditions, see [1, 6, 7, 8] and references therein. Most contributions focus on transmission conditions based on local operators.

In [2, 9, 10], the authors introduced non-local transmission conditions that can improve the convergence rate of OSM. In [9, Chap.8] the performance of this strategy was shown to remain robust up to GHz frequency range. Such an approach was proposed only for the Helmholtz equation, and has still not been adapted to electromagnetics.

In the present contribution we investigate such an approach for Maxwell's equations in a simple spherical geometry that allows explicit calculus by means of separation of variables. We study an Optimized Schwarz Method (OSM) where the transmission conditions are based on impedance type traces. The novelty lies in our impedance operator that we choose to be non-local. More precisely, it is chosen as a variant of the so-called Electric Field integral operator (see [11, §5.5]) where the wave number is purely imaginary. We show that the iterative solver associated to our strategy converges at an exponential rate.

1. Université Pierre et Marie Curie, and INRIA Paris, France. claeys@ann.jussieu.fr
2. CNRS and Université Pierre et Marie Curie, Paris, France. thierry@ljll.math.upmc.fr
3. POEMS (ENSTA ParisTech, CNRS, INRIA, Universit Paris-Saclay), Palaiseau, France. francis.collino@orange.fr

2 Maxwell's equations in harmonic regime

As a model problem we consider an electromagnetic transmission problem stemming from Maxwell's equations in harmonic regime where the whole space $\mathbb{R}^3$ is partitioned in two sub-domains $\mathbb{R}^3 = \overline{\Omega}_+ \cup \overline{\Omega}_-$ with Ω_- being the unit open ball centered at $\mathbf{0}$, and $\Omega_+ = \mathbb{R}^3 \setminus \overline{\Omega}_-$. Denote by $\boldsymbol{n}_\sigma$ the vector field normal to Γ directed toward the exterior of $\Omega_\sigma, \sigma = \pm$. With a constant wave number $\kappa > 0$, this is written

$$\begin{aligned}
&\boldsymbol{curl}(\boldsymbol{E}_\pm) - \imath\kappa \boldsymbol{H}_\pm = 0, \qquad \boldsymbol{curl}(\boldsymbol{H}_\pm) + \imath\kappa \boldsymbol{E}_\pm = 0 \quad \text{in } \Omega_\pm,\\
&\lim_{\rho\to\infty} \int_{\partial \mathrm{B}_\rho} |\boldsymbol{H}_+ \times \hat{\boldsymbol{x}} - \boldsymbol{E}_+|^2 \mathrm{d}\sigma_\rho = 0,\\
&\gamma_{\mathrm{T}}^+(\boldsymbol{E}) = +\gamma_{\mathrm{T}}^-(\boldsymbol{E}) + \boldsymbol{g}_{\mathrm{T}}\,, \qquad \text{with } \gamma_{\mathrm{T}}^\pm(\boldsymbol{E}) := \boldsymbol{n}_\pm \times (\boldsymbol{E}_\pm|_\Gamma \times \boldsymbol{n}_\pm),\\
&\gamma_{\mathrm{R}}^+(\boldsymbol{H}) = -\gamma_{\mathrm{R}}^-(\boldsymbol{H}) + \boldsymbol{g}_{\mathrm{R}}\,, \qquad \text{with } \gamma_{\mathrm{R}}^\pm(\boldsymbol{H}) := \boldsymbol{n}_\pm \times \boldsymbol{H}_\pm|_\Gamma,
\end{aligned} \tag{1}$$

with $\mathrm{B}_\rho := \{\boldsymbol{x} \in \mathbb{R}^3, |\boldsymbol{x}| < \rho\}$ and $\hat{\boldsymbol{x}} := \boldsymbol{x}/|\boldsymbol{x}|$. In this problem, $\boldsymbol{g}_{\mathrm{T}}, \boldsymbol{g}_{\mathrm{R}}$ are given source terms assumed to be supported on Γ only. Considering some invertible impedance operator $\mathcal{Z}$ that we shall define in Section 4, the transmission conditions in (1) can be reformulated as

$$\begin{aligned}
\gamma_{\mathrm{T}}^+(\boldsymbol{E}) + \mathcal{Z}\gamma_{\mathrm{R}}^+(\boldsymbol{H}) &= \gamma_{\mathrm{T}}^-(\boldsymbol{E}) - \mathcal{Z}\gamma_{\mathrm{R}}^-(\boldsymbol{H}) + \boldsymbol{g}_{\mathrm{T}} + \mathcal{Z}\boldsymbol{g}_{\mathrm{R}},\\
\gamma_{\mathrm{T}}^-(\boldsymbol{E}) + \mathcal{Z}\gamma_{\mathrm{R}}^-(\boldsymbol{H}) &= \gamma_{\mathrm{T}}^+(\boldsymbol{E}) - \mathcal{Z}\gamma_{\mathrm{R}}^+(\boldsymbol{H}) - \boldsymbol{g}_{\mathrm{T}} + \mathcal{Z}\boldsymbol{g}_{\mathrm{R}}.
\end{aligned} \tag{2}$$

For any tangential vector field $\boldsymbol{v}$ and $\sigma = \pm$ define the magnetic-to-electric operator $\mathcal{T}_\sigma(\boldsymbol{v}) := \gamma_{\mathrm{T}}^\sigma(\mathbf{U})$ where $(\mathbf{U}, \mathbf{V})$ is the unique solution to $\boldsymbol{curl}(\mathbf{U}) - \imath\kappa\mathbf{V} = 0$ in Ω_σ, $\boldsymbol{curl}(\mathbf{V}) + \imath\kappa\mathbf{U} = 0$ in Ω_σ and $\gamma_{\mathrm{R}}^\sigma(\mathbf{V}) = \boldsymbol{v}$ (and Silver-Müller's radiation condition if $\sigma = +$). Taking $\boldsymbol{u}_\sigma = \gamma_{\mathrm{T}}^\sigma(\boldsymbol{E}) + \mathcal{Z}\gamma_{\mathrm{R}}^\sigma(\boldsymbol{H}), \sigma = \pm$ as unknowns of our iterative procedure, Problem (1) is then equivalent to

$$\begin{aligned}
&\boldsymbol{u}_{-\sigma} = \mathcal{A}_\sigma(\boldsymbol{u}_\sigma) + \boldsymbol{f}_\sigma, \quad \sigma = \pm,\\
&\text{with } \mathcal{A}_\sigma := (\mathcal{T}_\sigma - \mathcal{Z})(\mathcal{T}_\sigma + \mathcal{Z})^{-1},
\end{aligned} \tag{3}$$

and $\boldsymbol{f}_\pm := (\mathcal{Z}(\boldsymbol{g}_{\mathrm{R}}) \pm \boldsymbol{g}_{\mathrm{T}})$. An optimized Schwarz strategy to solve Problem (1) now consists in a fixed point iterative method applied to (3), using the approximation $\boldsymbol{u}_\pm = \gamma_{\mathrm{T}}^\pm(\boldsymbol{E}) + \mathcal{Z}\gamma_{\mathrm{R}}^\pm(\boldsymbol{H}) = \lim_{n\to\infty} \boldsymbol{u}_\pm^{(n)}$ where $\boldsymbol{u}_\pm^{(n)}$ follows the recurrence

$$\begin{bmatrix} \boldsymbol{u}_+^{(n+1)} \\ \boldsymbol{u}_-^{(n+1)} \end{bmatrix} = \begin{bmatrix} 1-r & r\mathcal{A}_+ \\ r\mathcal{A}_- & 1-r \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{u}_+^{(n)} \\ \boldsymbol{u}_-^{(n)} \end{bmatrix} + \begin{bmatrix} r\boldsymbol{f}_+ \\ r\boldsymbol{f}_- \end{bmatrix}. \tag{4}$$

In this iterative method, $r > 0$ is a relaxation parameter whose effective value shall be discussed in the sequel.

3 Separation of variables on the sphere

To study the convergence of (4), we rely on the spherical symmetry of our model problem, and decompose the fields by means of vector spherical harmonics. According to e.g. [11, Thm.2.4.8], any tangential vector field $\boldsymbol{u} \in \mathrm{L}^2_{\mathrm{T}}(\Gamma) := \{\boldsymbol{v} : \Gamma \to \mathbb{C},\ \int_\Gamma |\boldsymbol{v}|^2 d\sigma < +\infty,\ \boldsymbol{x} \cdot \boldsymbol{v}(\boldsymbol{x}) = 0 \text{ on } \Gamma\}$ can be decomposed as

$$\begin{aligned}\boldsymbol{u}(\boldsymbol{x}) &= \sum_{n=0}^{+\infty} \sum_{|m|\le n} u^{\mathrm{D}}_{n,m}\, \mathbf{X}^{\mathrm{D}}_{n,m}(\boldsymbol{x}) + u^{\mathrm{C}}_{n,m}\, \mathbf{X}^{\mathrm{C}}_{n,m}(\boldsymbol{x}),\\ \text{with}\quad \mathbf{X}^{\mathrm{D}}_{n,m} &:= \tfrac{1}{\sqrt{n(n+1)}} \nabla_\Gamma \mathrm{Y}^m_n\ , \quad \mathbf{X}^{\mathrm{C}}_{n,m} := \hat{\boldsymbol{x}} \times \mathbf{X}^{\mathrm{D}}_{n,m},\end{aligned}$$

where $\hat{\boldsymbol{x}} := \boldsymbol{x}/|\boldsymbol{x}|$ and ∇_Γ is the surface gradient. Denoting $(\theta, \phi) \in [0, \pi] \times [0, 2\pi]$ the spherical coordinates on Γ, spherical harmonics are defined by

$$\mathrm{Y}^m_n(\theta, \phi) := \sqrt{\frac{2n+1}{4\pi}} \sqrt{\frac{(n-|m|)!}{(n+|m|)!}}\, \mathrm{P}^{|m|}_n(\cos\theta) e^{\imath m\phi},$$

where $\mathrm{P}^{|m|}_n(t)$ are the associated Legendre functions, see e.g. [3, §2.3]. The tangent fields $\mathbf{X}^{\mathrm{D}}_{n,m}, \mathbf{X}^{\mathrm{C}}_{n,m}, 0 \le |m| \le n$ yield an orthonormal Hilbert basis of $\mathrm{L}^2_{\mathrm{T}}(\Gamma)$. The operators $\mathcal{T}_\pm$ are diagonalized by the functions $\mathbf{X}^{\mathrm{D}}_{n,m}, \mathbf{X}^{\mathrm{C}}_{n,m}$. Indeed we have $\mathcal{T}_\pm(\mathbf{X}^\star_{n,m}) = t^\star_{n,\pm} \mathbf{X}^\star_{n,m}$ for $\star = \mathrm{D}, \mathrm{C}$ where, according to Formula (53) in [13],

$$\begin{aligned} t^{\mathrm{D}}_{n,-} &= 1/t^{\mathrm{C}}_{n,-} = +\imath \mathbb{J}'_n(\kappa)/\mathbb{J}_n(\kappa),\\ t^{\mathrm{D}}_{n,+} &= 1/t^{\mathrm{C}}_{n,+} = -\imath \mathbb{H}'_n(\kappa)/\mathbb{H}_n(\kappa).\end{aligned} \tag{5}$$

Here $\mathbb{J}_n(x) := \sqrt{\pi x/2}\ J_{n+1/2}(x)$ with $J_n(x)$ denoting the Bessel function of the first kind of order n, and $\mathbb{H}_n(x) := \sqrt{\pi x/2}\ H^{(1)}_{n+1/2}(x)$ with $H^{(1)}_n(x)$ denoting the Hankel function of the first kind of order n. The following result follows from [11, Thm.5.3.5].

Proposition 1.
We have $\Re e\{\int_\Gamma \overline{\boldsymbol{u}}\, \mathcal{T}_-(\boldsymbol{u}) d\sigma\} = 0$ *and* $\Re e\{\int_\Gamma \overline{\boldsymbol{u}}\, \mathcal{T}_+(\boldsymbol{u}) d\sigma\} > 0$ *for all* $\boldsymbol{u} \in \mathrm{L}^2_{\mathrm{T}}(\mathrm{div}, \Gamma) \setminus \{0\}$ *where* $\mathrm{L}^2_{\mathrm{T}}(\mathrm{div}, \Gamma) := \{\boldsymbol{v} \in \mathrm{L}^2_{\mathrm{T}}(\Gamma),\ \mathrm{div}_\Gamma(\boldsymbol{v}) \in \mathrm{L}^2(\Gamma)\ \}$.

This result is related to energy balance in $\Omega_\pm$. With $\Re e\{\int_\Gamma \overline{\boldsymbol{u}}\, \mathcal{T}_-(\boldsymbol{u}) d\sigma\} = 0$, the energy coming in Ω_- equals the outgoing energy. On the other hand, in Ω_+, there is energy radiated toward infinity as $\Re e\{\int_\Gamma \overline{\boldsymbol{u}}\, \mathcal{T}_+(\boldsymbol{u}) d\sigma\} > 0$. A direct consequence in terms of separation of variables is

$$\Re e\{t^\star_{n,+}\} > \Re e\{t^\star_{n,-}\} = 0 \quad \text{for } \star = \mathrm{D}, \mathrm{C},\ \forall n \ge 0. \tag{6}$$

That $\Re e\{t^\star_{n,-}\} = 0$ can also be seen directly from expression (5) since the $\mathbb{J}_n(z)$ are proportional to Bessel functions hence real valued. Assuming that the impedance is chosen so that $\mathcal{Z}(\mathbf{X}^\star_{n,m}) = z^\star_n \mathbf{X}^\star_{n,m}$ for $\star = \mathrm{D}, \mathrm{C}$ and $n \geq 0$ where $z_{n,\star} \in \mathbb{C}$, we have

$$\mathcal{A}_\pm(\mathbf{X}^\star_{n,m}) = a^\star_{n,\pm}\, \mathbf{X}^\star_{n,m} \quad \text{with} \quad a^\star_{n,\sigma} = \frac{t^\star_{n,\sigma} - z^\star_n}{t^\star_{n,\sigma} + z^\star_n}. \tag{7}$$

The exponential convergence of the optimized Schwarz method is guaranteed provided that the spectral radius ϱ_{OSM} of the iteration operator in (4) is strictly smaller than 1,

$$\varrho_{\mathrm{OSM}} = \sup_{n\geq 0} \varrho_n < 1, \quad \text{with } \varrho_n := \max_{\sigma=\pm,\star=\mathrm{D},\mathrm{C}} |1 - r \pm r\sqrt{a^\star_{n,+}a^\star_{n,-}}|. \tag{8}$$

First observe that, for any $r \in (0,1)$, we have $|1 - r + r\lambda| < 1$ as soon as $\lambda \neq 1$ and $|\lambda| \leq 1$. Since $|(z-1)/(z+1)| \leq 1$ if and only if $\Re e\{z\} \geq 0$, a necessary condition of convergence is that $\varrho_n < 1$ for each n which boils down to $\Re e\{t^\star_{n,\sigma}/z^\star_n\} \geq 0$ for each $n, \sigma, \star$. According to (6), the later condition holds provided that $z^\star_n \in (0, +\infty)$.

4 Non-local impedance operator

Now let us discuss our construction of the impedance operator $\mathcal{Z}$. Compared to existing literature on optimized Schwarz strategies in the context of electromagnetics, the peculiarity of the present contribution lies in our choice of $\mathcal{Z}$ that is non-local. We choose

$$\mathcal{Z}(\boldsymbol{u}) := \alpha \int_\Gamma \mathcal{G}_\alpha(\boldsymbol{x}-\boldsymbol{y})\boldsymbol{u}(\boldsymbol{y})d\sigma(\boldsymbol{y}) - \frac{1}{\alpha}\nabla_\Gamma \int_\Gamma \mathcal{G}_\alpha(\boldsymbol{x}-\boldsymbol{y})\mathrm{div}_\Gamma \boldsymbol{u}(\boldsymbol{y})d\sigma(\boldsymbol{y}) \tag{9}$$

where the kernel $\mathcal{G}_\alpha(\boldsymbol{x}) := \exp(-\alpha|\boldsymbol{x}|)/(2\pi|\boldsymbol{x}|)$ satisfies $-\Delta\mathcal{G}_\alpha + \alpha^2\mathcal{G}_\alpha = 2\delta_0$ in $\mathbb{R}^3$, and $\alpha > 0$ is a parameter whose value shall be discussed later. The operator given by (9) is a classical object of potential theory that can be understood as a dissipative version of the so-called Electric Field Integral operator (EFIE). Defined in this manner, the operator $\mathcal{Z}$ is diagonalized by the $\mathbf{X}^\star_n$. According to Formula (54) in [13] we have

$$z^{\mathrm{D}}_n = 2\mathbb{J}'_n(\imath\alpha)\mathbb{H}'_n(\imath\alpha) \quad \text{and} \quad z^{\mathrm{C}}_n = 2\mathbb{J}_n(\imath\alpha)\mathbb{H}_n(\imath\alpha). \tag{10}$$

According to Rayleigh's formulas, see [12, Chap.10], we have $\mathbb{J}_n(\imath x) = (\imath x)^{n+1}(x^{-1}\partial_x)^n(\sinh(x)/x)$ and $\mathbb{H}_n(\imath x) = -(\imath x)^{n+1}(x^{-1}\partial_x)^n(\exp(-x)/x)$. It is clear from (10) that $z^{\mathrm{D}}_n, z^{\mathrm{C}}_n > 0$ for all $n \geq 0$.

Satisfying $\varrho_n < 1$ for each n is necessary but not sufficient for (8) to be fulfilled. We must also verify that $\limsup_{n\to\infty} \varrho_n < 1$. Let us study the asymptotic behaviour of ϱ_n for $n \to \infty$. First, observe that (5) and (10) provide explicit expressions for $z_n^\star$ and $t_{n,\sigma}^\star$ where $\star = \mathrm{D}, \mathrm{C}$ and $\sigma = \pm$. According to [3, §2.4], we have $\mathbb{J}_n(x) \sim x^{n+1} n! 2^n/(2n+1)!$ and $\mathbb{H}_n(x) \sim -\imath x^{-n}(2n)!/(n!2^n)$ for $n \to +\infty$, and these asymptotics hold for both $x \in \mathbb{R}$ and $x \in \imath\mathbb{R}$. Plugging this inside (5) and (10) yields, for $n \to +\infty$,

$$z_n^{\mathrm{D}} \underset{n\to\infty}{\sim} \frac{n}{\alpha}, \quad z_n^{\mathrm{C}} \underset{n\to\infty}{\sim} \frac{\alpha}{n} \quad \text{and} \quad t_{n,\pm}^{\mathrm{D}} \underset{n\to\infty}{\sim} \frac{\imath n}{\kappa}.$$

We also deduce the asymptotics of $t_{n,\pm}^{\mathrm{C}} = 1/t_{n,\pm}^{\mathrm{D}}$. From this we obtain $t_{n,\pm}^{\mathrm{D}}/z_n^{\mathrm{D}} \sim \imath\alpha/\kappa$ and $t_{n,\pm}^{\mathrm{C}}/z_n^{\mathrm{C}} \sim -\imath\kappa/\alpha$. With (7) we conclude that

$$\lim_{n\to\infty} a_{n,\pm}^{\mathrm{D}} = +\phi(\alpha/\kappa) \quad \text{and} \quad \lim_{n\to\infty} a_{n,\pm}^{\mathrm{C}} = -\phi(\alpha/\kappa) \quad \text{where } \phi(\gamma) := \frac{\imath\gamma - 1}{\imath\gamma + 1}.$$

Now we have $\lim_{n\to\infty} \varrho_n = \max |\, 1 - r \pm r\phi(\alpha/\kappa)\,|$. A natural idea for choosing the parameters r and α consists in minimizing this quantity. The minimum is obtained for $\alpha = \kappa$ and $r = 1/2$ and we have in this case (note that this limit does not depend on κ)

$$\lim_{n\to\infty} \varrho_n = 1/\sqrt{2} \qquad \text{for } \alpha = \kappa,\ r = 1/2. \tag{11}$$

The control of ϱ_n when n goes to infinity is crucial to obtain geometrical convergence. It cannot be obtained when the impedance operator is a combination of local operators (with Padé approximants of the true impedance for instance). The use of non-local and positive impedance operator is the price to pay to achieve geometrical convergence.

5 Numerical illustration

Below we illustrate our analysis with effective numerical calculation[1] of the eigenvalues of the iteration operator of (4), taking systematically $\alpha = \kappa$. In Fig.1 below, we plot these eigenvalues for $\kappa = 10$. We see that the whole spectrum is contained in the unit disc. The values $\pm\imath$ clearly appear as the accumulation points of the spectrum with no relaxation ($r = 1$).

For eigenvalues associated to the relaxation parameter $r = 1/2$, we see that the accumulation points are located at $(1/2, \pm 1/2)$ whose modulus is $1/\sqrt{2}$, which agrees with (11). Next, in Fig.2 we show the same plots at higher frequency $\kappa = 100$. Once again, the whole spectrum is contained in the unit disc.

[1] Matlab scripts are available at: `http://gitlab.lpma.math.upmc.fr/IEOSM/Matlab`

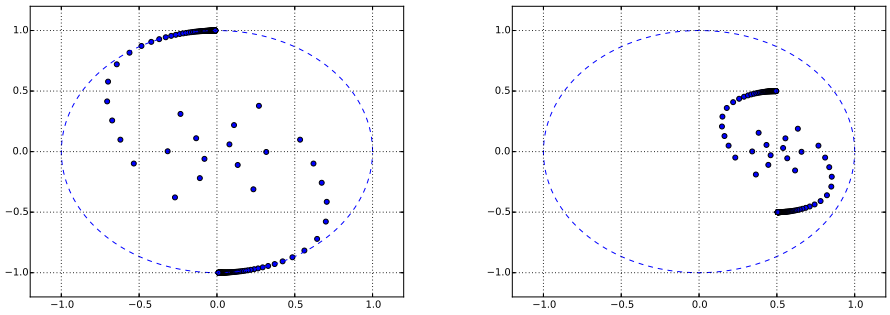


Fig. 1: Iteration eigenvalues with $\kappa = 10$ for $r = 1$ (left) and $r = 1/2$ (right)

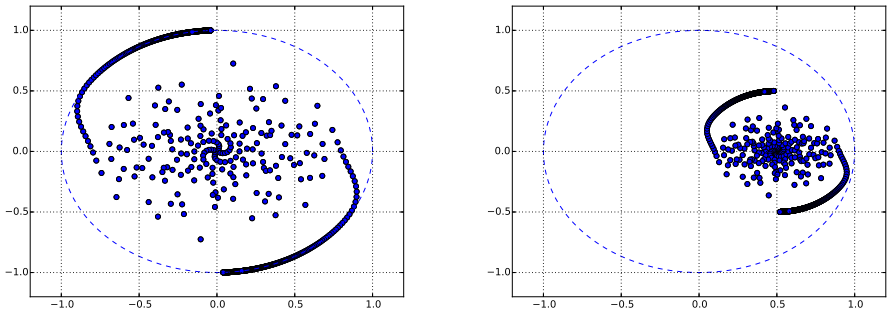


Fig. 2: Iteration eigenvalues with $\kappa = 100$ for $r = 1$ (left) and $r = 1/2$ (right)

Finally in Fig.3 we plot the values ϱ_n versus the modal index n for $\kappa = 10, 30, 100$. For low modal indices, it oscillates with growing amplitude until it reaches a pick located around $n \sim \kappa$. Then ϱ_n smoothly decays to $1/\sqrt{2}$. This scenario does not change as κ grows.

Although $\lim_{n\to\infty} \varrho_n$ remains independent of κ, the spectral radius $\sup_{n\geq 0} \varrho_n$ (reached around $n = \kappa$) does depend on κ, and we see in Fig.3 that this maximum grows closer to 1 as $\kappa \to \infty$. This suggests us that the values $\alpha = \kappa$ and $r = \frac{1}{2}$ may not be the optimal choice.

6 Conclusion

We have shown the convergence of the domain decomposition algorithm based on a dissipative EFIE transmission condition. How to choose the parameter α in a more optimal way should be further investigated. Moreover, it would be worth examining variants of the transmission operator (9). Augmenting it with additional local terms based on Padé approximants, in the manner of [6], seems promising.

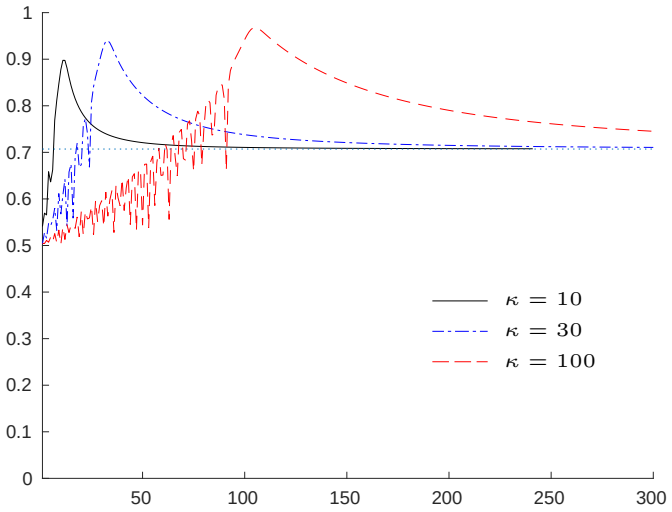


Fig. 3: Values of ϱ_n versus n with $r = 0.5$ for $\kappa = 10, 30, 100$.

Besides, in a finite element context, the use of a non-local operator is expensive in terms of both CPU time and memory storage. Various approaches could be considered for overcoming this problem. A possible solution may consist in truncating the Green kernel so as to (quasi)-localize the operator. The choice of the truncation and how it impacts the iteration operator should then be further investigated.

Other extensions of the present work are possible. For non-spherical interfaces, using the approach developed in [2], a convergent strategy would be obtained by choosing the impedance operator according to (9). This remark also holds in the case of multiple sub-domains, as long as there is no junction point at interfaces. Our strategy can also be adapted to the case of piecewise constant material characteristics. For this case also, the theory in [2] suggests that our method is convergent although, this time, a choice of impedance operator that varies according to the sub-domains may be more optimal. Finally the case of fully heterogeneous media seems to be still a widely open question.

Acknowledgment This work received support from the ANR research Grant ANR-15-CE23-0017-01.

References

[1] Y. Boubendir, X. Antoine, and C. Geuzaine. A quasi-optimal non-overlapping domain decomposition algorithm for the Helmholtz equation. *J. Comput. Phys.*, 231(2):262–280, 2012.

[2] F. Collino, S. Ghanemi, and P. Joly. Domain decomposition method for harmonic wave propagation: a general presentation. *Comput. Methods Appl. Mech. Engrg.*, 184(2-4):171–211, 2000.

[3] D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93 of *Applied Mathematical Sciences*. Springer, New York, third edition, 2013.

[4] B. Després. Décomposition de domaine et problème de Helmholtz. *C. R. Acad. Sci. Paris Sér. I Math.*, 311(6):313–316, 1990.

[5] B. Després. *Méthodes de décomposition de domaine pour les problèmes de propagation d'ondes en régime harmonique. Le théorème de Borg pour l'équation de Hill vectorielle*. 1991. PhD thesis, Univ. de Paris IX (Dauphine), 1991.

[6] M. El Bouajaji, B. Thierry, X. Antoine, and C. Geuzaine. A quasi-optimal domain decomposition algorithm for the time-harmonic Maxwell's equations. *J. Comput. Phys.*, 294:38–57, 2015.

[7] M. J. Gander. Optimized Schwarz methods for Helmholtz problems. In *Domain decomposition methods in science and engineering (Lyon, 2000)*, Theory Eng. Appl. Comput. Methods, pages 247–254. Internat. Center Numer. Methods Eng. (CIMNE), Barcelona, 2002.

[8] M.J. Gander, F. Magoulès, and F. Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24(1):38–60, 2002.

[9] M. Lecouvez. *Iterative domain decomposition method without overlap with geometric convergence for Helmholtz equation*. Phd thesis, École Polytechnique, 2015.

[10] M. Lecouvez, B. Stupfel, P. Joly, and F. Collino. Quasi-local transmission conditions for non-overlapping domain decomposition methods for the helmholtz equation. *Comptes Rendus Physique*, 15(5):403 – 414, 2014.

[11] J.-C. Nédélec. *Acoustic and electromagnetic equations*, volume 144 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2001.

[12] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors. *NIST handbook of mathematical functions*. Cambridge University Press, Cambridge, 2010.

[13] F. Vico, L. Greengard, and Z. Gimbutas. Boundary integral equation analysis on the sphere. *Numer. Math.*, 128(3):463–487, 2014.

Analysis of the shifted Helmholtz expansion preconditioner for the Helmholtz equation

Pierre-Henri Cocquet[1], Martin J. Gander[2]

1 Introduction

Solving Helmholtz problem numerically is challenging [**?**] mainly because of the lack of coercivity of the continuous operator or highly oscillatory solutions. Krylov subspaces methods like GMRES are still used in regards of their robustness but they require a good preconditioner[1] to be fast enough. Among many proposed preconditioners like Incomplete LU, Analytic ILU or domain decomposition based preconditioner, the shifted Helmholtz preconditioner [**?**, **?**, **?**, **?**] has received a lot of attention over the last decade thanks to its simplicity and its relevance to heterogeneous media.

This paper focus on the recent idea of expansion preconditioner [**?**, **?**] which is based on the fact that the inverse of the discrete Helmholtz operator can be written as a superposition of inverse of discrete shifted Helmholtz operator only. This is achieved using the Taylor's expansion, around $\beta = 0$, of the matrix-valued function $f(\beta) = (-\Delta_h - (1+\mathrm{i}\beta)k^2)^{-1}$, where Δ_h corresponds to a finite difference discretization of the usual Laplace operator. The expansion-preconditioner is then defined as the truncation of the Taylor's series hence converging to the exact inverse of the discrete Helmholtz operator if the Taylor series actually converges. They also proposed to compute each inverse of shifted Helmholtz with some iteration of multigrid which is known to converge with a number of iterations independent of the wavenumber (see e.g. [**?**, **?**]). We emphasize that the rate of convergence of the expansion preconditioner toward $A_0^{-1} = f(0)$ is computed in [**?**] and is given to be a $O(\beta^n)$. However, the latter does not involves bounds on the higher derivative of f which can deteriorate the performance of the proposed preconditioner and no additional analysis is performed.

(1) Université de la Réunion, PIMENT, 2 rue Joseph Wetzell, 97490 Sainte-Clotilde.
(2) University of Geneva, 2-4 rue du Lièvre, CP 64, 1211 Genève, Switzerland,
{martin.gander@unige.ch}{pierre-henri.cocquet@univ-reunion.fr}

[1] For a linear system $C\mathbf{x} = \mathbf{y}$, a good preconditioner refer to a matrix B for which the spectrum of $B^{-1}C$ is clustered around 1 (see e.g Elman's estimate [**?**]).

The goal of this paper is to give a theoretical insight of the performances of the expansion preconditioner and to extend its definition to Finite Element discretization. We first build the expansion preconditioner using the generalized resolvent formula and study its performances. We next show, as proved in [**?**], that it is mandatory to have a shift of the order of the wavenumber to get wavenumber independant convergence of GMRES. This paper ends with some numerical simulations.

2 General analysis of the expansion preconditioner

Let Ω be a convex polygon of $\mathbb{R}^d$, with $d = 1,2,3$. The shifted Helmholtz equation with impedance boundary conditions is

$$\begin{cases} -\Delta u(x) - (k^2 + \mathrm{i}\varepsilon)u(x) = f(x),\ x \in \Omega, \\ \partial_{\mathbf{n}} u - \mathrm{i}\eta u = 0,\ \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\mathbf{n}$ is the unitary normal vector directed outward $\partial\Omega$, $\varepsilon > 0$ is the so-called shift, and $\eta > 0$ is the impedance parameter. The Helmholtz equation with approximate radiation condition is recovered from (**??**) by setting $\varepsilon = 0$ and $\eta = k$.

The variational form of (**??**) is given below

$$\begin{cases} \text{Find } u \in H^1(\Omega) \text{ such that for all } v \in H^1(\Omega) : \\ a_\eta(u,v) := \displaystyle\int_\Omega \nabla u \cdot \overline{\nabla v} - (k^2 + \mathrm{i}\varepsilon) u\bar{v} dx - \mathrm{i}\eta \int_{\partial\Omega} u\bar{v} d\sigma = \int_\Omega f\bar{v} dx. \end{cases} \tag{2}$$

Let $\mathscr{V}_l$ be the finite element space obtained with piecewise linear polynomials

$$\mathscr{V}_l = \left\{ v \in \mathscr{C}(\overline{\Omega}) \mid v|_T \in \mathbb{P}_1 \text{ for all } T \in \mathscr{T}_l \right\} = \mathrm{Span}(\phi_1, \cdots, \phi_N),$$

where $\left\{\phi_j\right\}_{j=1}^N$ is the finite element nodal basis. The discrete problem is then

$$\begin{cases} \text{Find } u_l \in \mathscr{V}_l \text{ such that :} \\ a_\varepsilon(u_l, v_l) = \displaystyle\int_\Omega f\overline{v_l} dx,\ \forall v_l \in \mathscr{V}_l. \end{cases} \tag{3}$$

The latter is equivalent to the linear system $A_\varepsilon \mathbf{z}_l = \mathbf{b}_l$ where $u_l = F_h \mathbf{z}_l$ is the Galerkin solution and

$$F_h : x = (x_1, \cdots, x_N) \in \mathbb{C}^N \mapsto \sum_{j=1}^N x_j \phi_j \in \mathscr{V}_h.$$

Denoting by S, M, N respectively the stiffness, mass and boundary mass matrix, one gets

$$A_\varepsilon = S - (k^2 + \mathrm{i}\varepsilon)M - \mathrm{i}\eta N.$$

We denote by A_0 the discrete Helmholtz operator obtained with $\varepsilon = 0$ and $\eta = k$. We emphasize that this matrix is invertible thanks to the impedance boundary

condition. Also, if Dirichlet or Neumann's boundary conditions are used, we assume throughout this paper that A_0 is invertible.

We now give a generalized resolvent formula whose proof can be done by routine computations.

Lemma 1. *Let $A,B \in \mathrm{Hom}(\mathbb{C}^n)$ with B invertible and $p,z \in \mathbb{C}$ be two complex numbers in the resolvent set of AB^{-1}. Let $R(z) = (A - zB)^{-1}$ be the generalized resolvent of A. The following formula then holds*

$$R(p) - R(z) = (z-p)R(z)BR(p).$$

Using Neumann's series, Lemma **??** allow to rewrite the inverse of the discrete Helmholtz operator as a superposition of discrete shifted Helmholtz operator.

Theorem 1. *The inverse of the discrete Helmholtz operator is given as follows*

$$A_0^{-1} = \left(\sum_{j \geq 0} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M \right)^j \right) A_\varepsilon^{-1},$$

where the serie converges with respect to the norm $\|x\|_M = \sqrt{\langle Mx, \bar{x} \rangle} = \|F_h x\|_{L^2(\Omega)}$.

Proof. Lemma **??** applied with $A = A_0$, $B = M$, $p = 0$ and $z = \mathrm{i}\varepsilon$ yields

$$A_0^{-1} = (I_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M)^{-1} A_\varepsilon^{-1}.$$

Note that $A_\varepsilon^{-1} M = (M^{-1} A_\varepsilon)^{-1}$. Let $\mathbf{z} \in \mathbb{C}^N$ such that $A_\varepsilon \mathbf{z} = M\mathbf{b}$ for some $\mathbf{b} \in \mathbb{C}^N$. From the definition of the mass matrix M, the operator F_h and A_ε, one gets

$$a_\eta(F_h\mathbf{z}, F_h\mathbf{z}) = \langle M\mathbf{b}, \bar{\mathbf{z}} \rangle = \left(F_h\mathbf{b}, \overline{F_h\mathbf{z}} \right)_{L^2(\Omega)}.$$

Cauchy-Schwartz inequality and the next lower bound

$$|a_\eta(F_h\mathbf{z}, F_h\mathbf{z})| > |\mathscr{I}\, a_\eta(F_h\mathbf{z}, F_h\mathbf{z})| = \varepsilon \|F_h\mathbf{z}\|^2_{L^2(\Omega)} + \eta \|F_h\mathbf{z}\|^2_{L^2(\partial\Omega)},$$

show that $\|\mathbf{z}\|_M < \|\mathbf{b}\|_M \varepsilon^{-1}$, and thus $\left\|\varepsilon A_\varepsilon^{-1} M\right\|_M < 1$. Finally, $(I_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M)^{-1}$ can be expanded as a Neumann's serie and the proof is finished.

Remark 2 *The mass matrix is symmetric and positive definite so it admits a square root $M^{1/2}$. For any $B \in \mathrm{Hom}(\mathbb{C}^N)$, the matrix norm induced by $\|\cdot\|_M$ is then defined by $\|B\|_M = \left\|M^{1/2} B M^{-1/2}\right\|_2$. This yields*

$$\left\|\varepsilon A_\varepsilon^{-1} M\right\|_M = \varepsilon \left\| M^{1/2} A_\varepsilon^{-1} M^{1/2} \right\|_2 = \varepsilon \left\| A_\varepsilon^{-1} M \right\|_2 < 1,$$

and thus the series from Theorem **??** *converges with respect to the* 2*-norm as well.*

Following [**?**], the expansion preconditioner of order $n \in \mathbb{N}^*$ is defined as a truncation of the Neumann's serie given in Theorem **??**

$$EX(n) = \left(\sum_{j=0}^{n-1} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M\right)^{j+1} \right) M^{-1} = \left(\sum_{j=0}^{n-1} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M\right)^{j} \right) A_\varepsilon^{-1}. \quad (4)$$

The preconditioned problem is thus given as follow

$$EX(n) A_0 \mathbf{z}_l = EX(n) \mathbf{b}_l. \quad (5)$$

From Elman's estimate (see e.g. Theorem 1.8 [**?**]), the rate of convergence of GMRES used for solving $C\mathbf{x} = \mathbf{y}$ only depend on the upper bound of $\|\mathbb{I} - C\|_2$. We now compute this term for the expansion preconditioner.

Theorem 3. *For any shift $\varepsilon > 0$, impedance parameter $\eta > 0$, meshsize h and $n \in \mathbb{N}$, the expansion preconditioner satisfies the following bounds*

$$\mathscr{N}\left(\mathbb{I}_d - EX(1)A_0\right) \leq \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right),$$
$$\forall n \geq 1,\ \mathscr{N}\left(\mathbb{I}_d - EX(n)A_0\right) \leq \frac{1 + \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)}{1 - \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)} (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^n,$$

where $\mathscr{N}(B)$ denotes any matrix norm or $\rho(B)$.

Proof. The first item follows from $\mathbb{I} - EX(1)A_0 = \mathbb{I} - A_\varepsilon^{-1} A_0 = \mathrm{i}\varepsilon A_\varepsilon^{-1} M$. For the second one, we compute

$$\mathbb{I} - EX(n)A_0 = (A_0^{-1} - EX(n))A_0 = \left(\sum_{j \geq n} (-\mathrm{i}\varepsilon)^j (A_\varepsilon^{-1} M)^j \right) A_\varepsilon^{-1} A_0.$$

Note that $A_\varepsilon^{-1} A_0 = \mathbb{I}_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M$ and thus $A_\varepsilon^{-1} A_0$ and $A_\varepsilon^{-1} M$ commute. Now, using that $\varepsilon \rho(A_\varepsilon^{-1} M) \leq \varepsilon \left\| A_\varepsilon^{-1} M \right\|_2 < 1$, we can use Gelfand's formula to get the convergence of the Neumann series with respect to any matrix norm. Majoring and expanding using geometric serie then give

$$\begin{aligned} \mathscr{N}\left(\mathbb{I} - EX(n)A_0\right) &\leq \mathscr{N}\left(\mathbb{I}_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M\right) (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^n \sum_{j \geq 0} (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^j \\ &\leq \frac{1 + \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)}{1 - \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)} (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^n. \end{aligned}$$

Remark 4 *The construction of the expansion preconditioner as well as Theorem* **??** *hold without any changes for high order Finite Element discretization.*

The upper bound from Theorem **??** involves only $\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)$. If the latter is bounded away from 1, the expansion preconditioner can greatly reduce the number of GMRES iterations by considering a large enough n.

3 Wavenumber-independance convergence of GMRES

We show in this section that, as proved in [**?**], taking $\varepsilon \sim k$ is mandatory to ensure wavenumber-independant convergence of GMRES when using an expansion preconditioner. This is done in the next result for two types of meshes: one for which one has pollution-free FEM[2] and one for $h \sim k^{-2}$.

Theorem 5. *Assume that one of the following assumptions holds*

(A1) $\eta \sim k$ *and* $k^3h^2 \leq C_0$ *holds with* C_0 *small enough.*

(A2) $\eta \lesssim k$, $k \geq k_0$ *for a given* $k_0 > 0$ *and* $kh\sqrt{|k^2-\varepsilon|} \leq C_0$ *holds with* C_0 *small enough.*

Then there exists a constant $C_2 > 0$ *depending only on* Ω *such that for any* $\varepsilon > 0$ *with* $\varepsilon C_2 < k$, *one has*

$$\forall\, n \geq 1,\ \mathscr{N}\left(\mathbb{I}_d - EX(n)A_0\right) \leq \left(\frac{C_2\varepsilon}{k}\right)^n \frac{k+C_2\varepsilon}{k-C_2\varepsilon},$$

where $\mathscr{N}(.) = \rho(.)$ *if (A1) hold and* $\mathscr{N}(.) = \|.\|_2$ *if (A2) hold.*

Proof. Assume that (A1) hold. Let $\lambda \in \mathbb{C}$ be an eigenvalue of $M^{-1}A_\varepsilon = (A_\varepsilon^{-1}M)^{-1}$ and $\mathbf{v} \in \mathbb{C}^N$ the associated eigenvector. One has

$$M^{-1}A_\varepsilon \mathbf{v} = \left(M^{-1}(S - \mathrm{i}\eta N) - (k^2 + \mathrm{i}\varepsilon)\mathbb{I}_N\right)\mathbf{v} = \lambda \mathbf{v}.$$

Therefore, the spectrum of $M^{-1}A_\varepsilon$ is given by

$$\sigma(M^{-1}A_\varepsilon) = \left\{\lambda_j + \mathrm{i}\varepsilon \mid \lambda_j \in \sigma(M^{-1}A_0)\right\},$$

from which we infer that

$$\varepsilon\rho(A_\varepsilon^{-1}M) = \max_{\lambda_j \in \sigma(M^{-1}A_0)} \frac{\varepsilon}{|\lambda_j + \mathrm{i}\varepsilon|} \tag{6}$$

Let $\mathbf{b} \in \mathbb{C}^N$ be fixed and $\varphi_h \in \mathbb{C}^N$ be the solution to $A_0\mathbf{v}_h = M\mathbf{b}$. Note that $\varphi_h = F_h\mathbf{v}_h \in \mathscr{V}_h$ corresponds to the FEM discretization of the solution to (**??**) with $f = F_h\mathbf{b}$. Since $f \in L^2(\Omega)$ and Ω is assumed to be convex, the solution to the Helmholtz equation (**??**) belongs to $H^2(\Omega)$. Since (A1) hold, one can apply [**?**, Corollary 4.4 p.12] to get

$$\|\nabla\varphi_h\|_{L^2(\Omega)} + k\|\varphi_h\|_{L^2(\Omega)} \lesssim \|f\|_{L^2(\Omega)}. \tag{7}$$

Then (**??**) shows that

$$\|F_h\mathbf{v}_h\|_{L^2(\Omega)} \lesssim \frac{1}{k}\|F_h\mathbf{b}\|_{L^2(\Omega)}.$$

Using [**?**, Eq. (4.2) p. 24], one has $\|F_h\|_{\mathbb{C}^N \to \mathscr{V}_h} \sim h^{d/2}$ which gives

$$\|\mathbf{v}_h\|_2 = \left\|A_0^{-1}M\mathbf{b}\right\|_2 \lesssim \frac{\|\mathbf{b}\|}{k}.$$

[2] According to [**?**] no pollution effect occurs if $k^3h^2 \leq C_0$ holds with C_0 small enough.

The above estimate holds for any $\mathbf{b} \in \mathbb{C}^N$ and thus

$$\left\|A_0^{-1}M\right\|_2 \lesssim \frac{1}{k}. \tag{8}$$

The upper bound (**??**) proves that, for any $\mu \in \sigma(A_0^{-1}M)$, $|\mu| \lesssim k^{-1}$. Since any $\lambda \in \sigma(M^{-1}A_0)$ can be written as $\lambda = 1/\mu$, one gets $k \lesssim |\lambda|$. We finally infer that there exists $C_2 > 0$ depending only on Ω such that

$$\rho(A_\varepsilon^{-1}M) \leq \frac{C_2}{k}. \tag{9}$$

Assuming now that (A2) hold allow to apply [**?**, Lemma 3.5 p.595] that gives the quasi-optimality of the bilinear form a_ε on $\mathscr{V}_h$ with respect to the weighted norm $\|u\|_{1,k}^2 = \|\nabla u\|_{L^2(\Omega)}^2 + k^2\|u\|_{L^2(\Omega)}^2$. Using this, they proved [**?**, Lemma 4.1 p. 598] that there exists a constant C_2 depending only on Ω such that

$$\left\|A_\varepsilon^{-1}M\right\|_2 \leq \frac{C_2}{k}. \tag{10}$$

Using now (**??**) and (**??**) together with the bound proved in Theorem **??** ends the proof.

4 Numerical simulations

5 Conclusions

References

1. Cai, X. C., & Widlund, O. B. (1992). Domain decomposition algorithms for indefinite elliptic problems. SIAM Journal on Scientific and Statistical Computing, 13(1), 243-258.
2. Cocquet, P. H., & Gander, M. J. (2016). On the minimal shift in the shifted Laplacian preconditioner for multigrid to work. In Domain Decomposition Methods in Science and Engineering XXII (pp. 137-145). Springer International Publishing.
3. Cocquet, P. H., & Gander, M. J. (2017). How Large a Shift is Needed in the Shifted Helmholtz Preconditioner for its Effective Inversion by Multigrid?. SIAM Journal on Scientific Computing, 39(2), A438-A478.
4. Gander, M.J., Graham, I.G., & Spence, E.A. (2015). Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed?. Numerische Mathematik, 131(3), 567-614.
5. Cools, S., & Vanroose, W. (2015). Generalization of the complex shifted Laplacian: on the class of expansion preconditioners for Helmholtz problems. ArXiv e-prints.
6. Cools, S., & Vanroose, W. (2017). On the Optimality of Shifted Laplacian in a Class of Polynomial Preconditioners for the Helmholtz Equation. In Modern Solvers for Helmholtz Problems (pp. 53-81). Springer International Publishing.
7. Y.A Erlangga, C. Vuik, C.W. Oosterlee. On a class of preconditioners for solving the discrete Helmholtz equation, *Applied Numerical Mathematics*, p. 409-425, 2004.

8. Ernst, O. G., & Gander, M. J. (2012). Why it is difficult to solve Helmholtz problems with classical iterative methods. In Numerical analysis of multiscale problems (pp. 325-363). Springer Berlin Heidelberg.
9. Du, Y., & Wu, H. (2015). Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number. SIAM Journal on Numerical Analysis, 53(2), 782-804.

A finite difference method with optimized dispersion correction for the Helmholtz equation

Pierre-Henri Cocquet, Martin J. Gander, Xueshuang Xiang

1 Introduction

We propose a new finite difference method (FDM) with optimized dispersion correction for the Helmholtz equation

$$L_k u := -\Delta u - k^2 u = f, \quad \text{in } \Omega \subset \mathbb{R}^2, \tag{1}$$

where Δ is the Laplacian, k is the wave number, and we assume boundary conditions such that the problem is well posed. The Helmholtz equation has important applications in many fields of science and engineering, e.g., acoustic and electromagnetic waves, and obtaining more accurate numerical discretizations has attracted significant research interest, see [2, 1, 8, 9, 12] and the references therein.

It is well known that all grid based numerical methods, e.g. finite element or finite difference methods, suffer from the so called *pollution effect*, which can not be eliminated [2], and the wave number of the numerical solution is different from that of the exact solution, leading to *numerical dispersion* [7, 6]. To keep the pollution effect and numerical dispersion under control, classical discretizations require a very fine mesh, which leads to very large discrete systems, especially when the frequency increases. To reduce the numerical dispersion of the standard 5-point finite difference scheme, a rotated 9-point FDM was proposed in [8] which minimizes the numerical dispersion, see also [3, 10, 13, 4] for more recent such approaches. Minimizing numerical dispersion is also important for effective coarse grid corrections in domain decomposition and for constructing efficient multigrid solvers: in 1D it is

Pierre-Henri Cocquet
Université de la Réunion, PIMENT, 2 rue Joseph Wetzell, 97490 Sainte-Clotilde e-mail: Pierre-Henri.Cocquet@univ-reunion.fr
M. J. Gander
Université de Genève, 2-4 rue du Lièvre, 1211 Genève e-mail: Martin.Gander@unige.ch
Xueshuang Xiang (corresponding author)
Qian Xuesen Laboratory of Space Technology, Beijing e-mail: xiangxueshuang@qxslab.cn

even possible to obtain perfect multigrid efficiency using dispersion correction [5], see also [11] for an approximation in higher dimensions.

We develop here a new dispersion minimizing FDM for the Helmholtz equation (1) using as a new idea a modified wave number. Compared with the finite difference scheme of [8] which minimizes already the numerical dispersion, our new scheme using the same stencil, but a modified wave number, has substantially less dispersion error and thus much more accurate phase speed. Our examples also indicate that for plane wave solutions, our new FDM is sixth-order accurate.

2 Dispersion correction for standard FDM

We first recall the definition of the dispersion relation and some notation. Given an operator P, e.g. the continuous operator L_k in (1) or any finite difference approximation for L_k, its symbol is

$$\sigma_P(\xi) := e^{-\mathbf{i}\xi\cdot x}(Pe^{\mathbf{i}\xi\cdot x}). \tag{2}$$

The dispersion relation of the operator P is then defined to be the set

$$\{\xi \in \mathbb{R}^2 | \sigma_P(\xi) = 0\}, \tag{3}$$

where $\xi = (\xi_1, \xi_2)$ denotes the wave vector. A direct computation using (2) gives for the continuous operator L_k in (1) the dispersion relation set $\{\xi \in \mathbb{R}^2 | \xi_1^2 + \xi_2^2 = k^2\}$.

For ξ such that $\sigma_P(\xi) = 0$, the number $v = \frac{k}{\|\xi\|}$ is called the (normalized) phase speed associated with a plane wave with angle θ given by $\tan\theta = \xi_2/\xi_1$. For the operator L_k in (1), the phase speed is equal to 1 for any angle. For any discretization scheme, we will consider the phase speed as a function of a dimensionless quantity, the *number of points per wavelength* $G = \frac{2\pi}{kh}$, or its inverse $1/G$.

For any discretization L_k^h of L_k, numerical dispersion can be defined as the difference between the dispersion relation of L_k and L_k^h. The numerical dispersion can also be evaluated by the difference of phase speed of L_k^h and 1 for different angles. A key idea for dispersion correction is to use a different numerical wave number $\hat{k}$ in the discretized operator $L_{\hat{k}}^h$ to minimize the numerical dispersion [5]. Take for example the 1D Helmholtz equation

$$-\frac{\partial^2 u}{\partial x^2} - k^2 u = f, \tag{4}$$

where the dispersion relation is $\{\xi \,|\, |\xi| = k\}$. The standard 3-point FDM of (4) is

$$(L_k^{h,fd3} u)_i = h^{-2}(2u_i - u_{i-1} - u_{i+1}) - k^2 u_i. \tag{5}$$

Using (2), the dispersion relation of $L_k^{h,fd3}$ is $\{\xi \in \mathbb{R} \,|\, 2h^{-2}(1-\cos(\xi h)) = k^2\}$, which is quite different from $\{\xi \,|\, |\xi| = k\}$. In order to make (5) have the same dis-

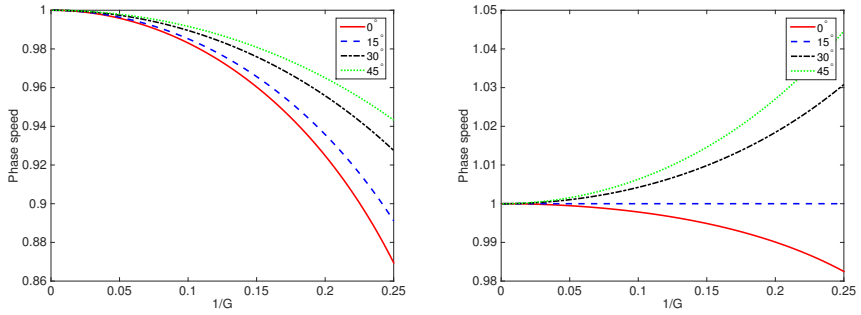


Fig. 1 Phase speed curves for 5-point FDM. Left: no dispersion correction. Right: dispersion correction for $\theta = 20^\circ$.

persion as (4), it was proposed in [5] to use a different wave number in (5), denoted by $\hat{k}$. Choosing $\hat{k} = |\sqrt{2h^{-2}(1-\cos(kh))}|$ implies

$$\{\xi \in \mathbb{R} \,|\, 2h^{-2}(1-\cos(\xi h)) = \hat{k}^2\} = \{\xi \,|\, |\xi| = k\},$$

and hence there is no numerical dispersion!

We investigate now if a similar approach can be used for the 2D Helmholtz equation (1), whose standard 5-point FDM is given by

$$(L_k^{h,fd5}u)_{i,j} = h^{-2}(4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}) - k^2 u_{i,j}, \tag{6}$$

Using (2), its dispersion relation of $L_k^{h,fd5}$ can readily be computed to be

$$\{\xi \in \mathbb{R}^2 | h^{-2}(4 - 2\cos(h\xi_1) - 2\cos(h\xi_2)) = k^2\}. \tag{7}$$

We show in Figure 1 (left) the phase speed v^{fd5} we computed using (7) for the angles $0^\circ, 15^\circ, 30^\circ$ and 45° when $k = 10$. We can clearly see that the numerical dispersion increases as we decrease the number of points per wavelength G. Using the dispersion correction idea from the 1D Helmholtz equation, we can do dispersion correction as well, but only for a specific direction. Given an angle θ, for wave number k and mesh size h, we choose the numerical wave number to be

$$\hat{k}(\theta,k,h) = |\sqrt{h^{-2}(4 - 2\cos(kh\cos(\theta)) - 2\cos(kh\sin(\theta)))}|. \tag{8}$$

The 5-point FDM with dispersion correction is then given by

$$(L_{\hat{k}}^{h,fd5}u)_{i,j} = h^{-2}(4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}) - \hat{k}^2 u_{i,j}, \tag{9}$$

and its dispersion relation is

$$\{\xi \in \mathbb{R}^2 | h^{-2}(4 - 2\cos(h\xi_1) - 2\cos(h\xi_2)) = \hat{k}^2\}. \tag{10}$$

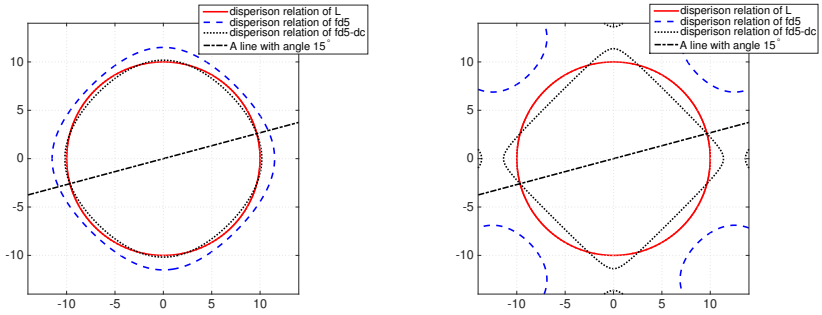


Fig. 2 Dispersion relation of operator L, fd5 and fd5-dc with dispersion correction for angle 15° and $G = 4$ (left) and $G = 2.5$ (right).

By the definition of $\hat{k}$ in (8), one can see that the dispersion correction is used to ensure that the phase speed $v^{fd5-dc} = 1$ for the specific angle θ, i.e. there is no dispersion error in that direction. However, for other angles, we still have numerical dispersion, as shown in Figure 1 (right), where we did dispersion correction for $\theta = 15°$, and then computed the phase speed v^{fd5-dc} for the angles $0°, 15°, 30°$ and $45°$ when $k = 10$. In Figure 2, we show the dispersion relation of L, $L_k^{h,fd5}$ and $L_{\hat{k}}^{h,fd5}$, where $\hat{k}$ is again the dispersion correction with $\theta = 15°$. We see that the discrete corrected dispersion relation is much closer to that of the continuous operator L_k than the uncorrected one. However, numerical dispersion still exists, it is not possible to make the phase speed $v^{fd5-dc} = 1$ for all angles using a modified wave number alone.

3 An optimized 9-point FDM with dispersion correction

To improve dispersion errors, a parametrized 9-point FDM was introduced in [8], where $-\Delta$ is discretized by a tensor product of a 1D mass matrix with stencil $[(1-a)/2, a, (1-a)/2]^T$ and the standard second order difference with stencil $[-h^{-2}, 2h^{-2}, -h^{-2}]^T$, and the the mass term $-k^2$ is discretized by the symmetric 9-point stencil

$$\begin{bmatrix} (1-b-c)/4 & c/4 & (1-b-c)/4 \\ c/4 & b & c/4 \\ (1-b-c)/4 & c/4 & (1-b-c)/4 \end{bmatrix}.$$

This leads with $\alpha = [a, b, c]$ and our numerical wave number $\hat{k}$ to the new 9-point FDM

$$\begin{aligned}(L_{\hat{k}}^{h,\alpha}u)_{i,j} = &(\tfrac{4a}{h^2} - \hat{k}^2 b)u_{i,j} + (\tfrac{1-2a}{h^2} - \tfrac{\hat{k}^2 c}{4})(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) \\ &-(\tfrac{1-a}{h^2} + \hat{k}^2 \tfrac{1-b-c}{4})(u_{i-1,j-1} + u_{i+1,j-1} + u_{i-1,j+1} + u_{i+1,j+1}).\end{aligned} \tag{11}$$

Computing its dispersion relation $\{\xi \in \mathbb{R}^2 | (e^{-\mathbf{i}\xi\cdot x})_{i,j}(L_{\hat{k}}^{h,\alpha} e^{\mathbf{i}\xi\cdot x})_{i,j} = 0\}$ gives

$$\begin{aligned}(4ah^{-2} - \hat{k}^2 b) + 2(\tfrac{1-2a}{h^2} - \tfrac{\hat{k}^2 c}{4})(\cos(h\xi_1) + \cos(h\xi_2)) \\ -2(\tfrac{1-a}{h^2} + \hat{k}^2 \tfrac{1-b-c}{4})(\cos(h(\xi_1+\xi_2)) + \cos(h(\xi_1-\xi_2))) = 0.\end{aligned} \tag{12}$$

For a vector ξ that satisfies the dispersion relation (12), we define

$$\eta_{\hat{k}}^{\alpha} := \|\xi\|, \tag{13}$$

which is a function that depends on θ. Then the phase speed of the operator $L_{\hat{k}}^{h,\alpha}$ is $v_{\hat{k}}^{\alpha} = \frac{k}{\eta_{\hat{k}}^{\alpha}}$. For the phase speed $v_{\hat{k}}^{\alpha}$ to be close to 1, we need that $\eta_{\hat{k}}^{\alpha}$ is close to k. We thus would need to solve the L^2 minimization problem[1]

$$\min_{\alpha,\hat{k}} \int_0^{2\pi} (\eta_{\hat{k}}^{\alpha}(\theta) - k)^2 d\theta. \tag{14}$$

Because we can not explicitly compute (13) from the transcendental relation (12), we propose a different minimization approach based on the reasonable

Assumption 3.1 *Given a mesh size h, there exist sets $\mathscr{K}$ and $\mathscr{P}$ such that*

- *$\forall \hat{k} \in \mathscr{K}$, $\forall \alpha \in \mathscr{P}$, the set of the dispersion relation* (12) *is not empty;*
- *Given $\alpha \in \mathscr{P}$, the mapping of $\mathscr{K}$ to $\{\eta_{\hat{k}}^{\alpha} \,|\, \hat{k} \in \mathscr{K}\}$ is injective.*

Let $\mathscr{F}^{h,\alpha} : p(\theta) \to q(\theta)$ be the operator which computes for given $p(\theta)$, $\theta \in [0, 2\pi]$ the solution $q(\theta)$ of

$$(e^{-\mathbf{i}[p(\theta)\cos(\theta), p(\theta)\sin(\theta)]^T \cdot x})_{i,j}(L_q^{h,\alpha} e^{\mathbf{i}[p(\theta)\cos(\theta), p(\theta)\sin(\theta)]^T \cdot x})_{i,j} = 0. \tag{15}$$

Since $\hat{k}^2$ appears only linearly in the 9-point FDM (11), $\mathscr{F}^{h,\alpha}$ is easy to compute numerically. In addition, by the definition of $\eta_{\hat{k}}^{\alpha}$ in (13) and Assumption 3.1, we have $\mathscr{F}^{h,\alpha}(\eta_{\hat{k}}^{\alpha}) = \hat{k}$. Thus, instead of solving (14), we solve $\min_{\alpha\in\mathscr{P},\hat{k}\in\mathscr{K}} \int_0^{2\pi} (\mathscr{F}^{h,\alpha}(\eta_{\hat{k}}^{\alpha}(\theta)) - \mathscr{F}^{h,\alpha}(k))^2 d\theta$, which, combined with $\mathscr{F}^{h,\alpha}(\eta_{\hat{k}}^{\alpha}) = \hat{k}$, yields

$$\min_{\alpha\in\mathscr{P},\hat{k}\in\mathscr{K}} \int_0^{2\pi} (\hat{k} - \mathscr{F}^{h,\alpha}(k))^2 d\theta, \tag{16}$$

where k can be interpreted as a constant function in θ. Using that $\hat{k}$ does not depend on θ, the objective function in (16) becomes by a direct calculation

$$\begin{aligned}\int_0^{2\pi} (\hat{k} - \mathscr{F}^{h,\alpha}(k))^2 d\theta = 2\pi \left(\hat{k} - \frac{1}{2\pi}\int_0^{2\pi} \mathscr{F}^{h,\alpha}(k) d\theta\right)^2 \\ + \int_0^{2\pi} \mathscr{F}^{h,\alpha}(k)^2 d\theta - \frac{1}{2\pi}\left(\int_0^{2\pi} \mathscr{F}^{h,\alpha}(k) d\theta\right)^2.\end{aligned}$$

[1] We could also use different norms leading to different optimized dispersion corrections.

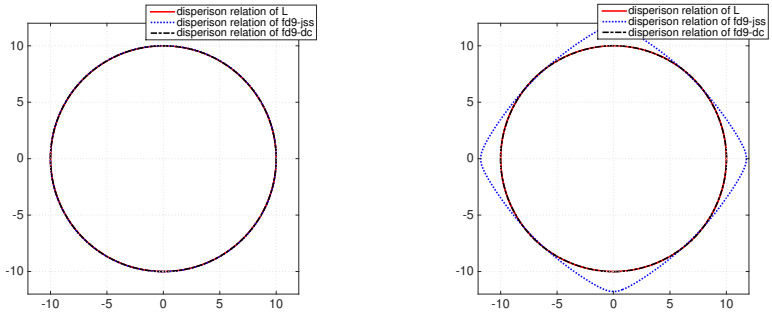


Fig. 3 Dispersion relation of L, fd9-jss and fd9-dc when $G = 4$ (left) and $G = 2.5$ (right).

For any fixed α, we can then take $\hat{k} = \frac{1}{2\pi}\int_0^{2\pi} \mathscr{F}^{h,\alpha}(k)d\theta$ to make the objective function reach its minimum, since the other terms do not depend on $\hat{k}$, and thus the minimization problem (16) is after a short calculation equivalent to minimizing the variance,

$$\min_{\alpha\in\mathscr{P}} \int_0^{2\pi} (\frac{1}{2\pi}\int_0^{2\pi} \mathscr{F}^{h,\alpha}(k)d\theta - \mathscr{F}^{h,\alpha}(k))^2 d\theta. \tag{17}$$

This leads to the following algorithm to compute optimized α^* and $\hat{k}^*$:

Algorithm 3.1 *(Optimized parameters α^* and $\hat{k}^*$ for dispersion correction)*
1° *Input wave number k and mesh size h;*
2° *Construct operator $\mathscr{F}^{h,\alpha}$ in* (15)*;*
3° *Solve minimization problem* (17) *to obtain α^*;*
4° *Compute $\hat{k}^* = \frac{1}{2\pi}\int_0^{2\pi} \mathscr{F}^{h,\alpha^*}(k)d\theta$;*
5° *Output α^* and $\hat{k}^*$.*

4 Numerical examples

We use a Riemann sum, discretizing θ in Algorithm 3.1 from 0 to 2π with step size $\pi/100$, and solve (17) using Nelder Mead with initial guess $\alpha^0 = [1,1,0]$, which corresponds to the standard 5-point FDM. We denote our new 9-point FDM with dispersion correction by fd9-dc, and compare it to the the FDM of Jo, Shin and Suh in [8] denoted by fd9-jss. The parameters for fd9-jss do not depend on h and k and are given by $\alpha = [0.7731, 0.6248, 0.3752]$.

We first compare in Figure 3 the dispersion relation of fd9-jss and fd9-dc when $k = 10$. Algorithm 3.1 gives as optimized parameters for $G = 4$ the values $\alpha^* = [0.8027, 1.0532, 0.0002]$, $\hat{k}^* = 8.7725$, and for $G = 2.5$ the values $\alpha^* = [0.7662, 1.0553, 0.0003]$, $\hat{k}^* = 7.2186$. We see on the left that both schemes seem very good for $G = 4$, compared to the five point schemes in Figure 2, but on the

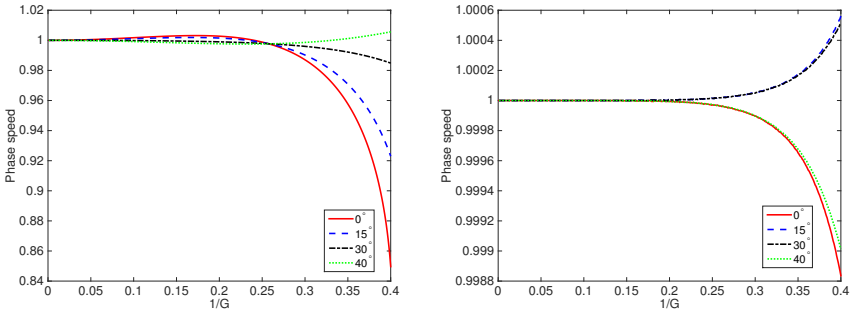


Fig. 4 Phase speed curves for fd9-jss (left) and fd9-dc (right) when $k = 10$.

h	fd5	fd9-jss	fd9-dc
0.05	1.30E5	5.41E3	5.35E3
0.1	5.57E2	1.70E3	1.36E3
0.2	1.28E16	4.37E2	3.77E2

Table 1 Condition number comparison for the linear systems obtained with the different schemes for varying mesh size when $k = 10$.

right for $G = 2.5$ the dispersion relation of fd9-dc is much better, still looking perfect for only 2.5 points per wavelength!

Figure 4 shows the phase speed curves $v_{\hat{k}^*}^{\alpha^*}$ for fd9-jss and fd9-dc for the angles $0^\circ, 15^\circ, 30^\circ$ and 45° when $k = 10$ as a function of $1/G$. We can clearly see that the phase speed of fd9-dc is much closer to 1 than for fd9-jss (note the different scales).

We next investigate the accuracy in h. We consider the Helmholtz equation on $\Omega = (-1,1)\times(-1,1)$ with the exact plane wave solution $u_e^\theta(x) = e^{\mathbf{i}(k\cos(\theta)x_1 + k\sin(\theta)x_2)}$ and Dirichlet boundary conditions. The corresponding numerical solutions of fd9-jss and fd9-dc are $u_h^{\text{fd9-jss},\theta}$ and $u_h^{\text{fd9-dc},\theta}$, and the interpolated exact solution u_e^θ on the mesh with size h by $u_{i,h}^\theta$. We then measure the relative error of fd9-jss and fd9-dc by

$$err_{\text{fd9-jss}}(h,\theta) = \frac{\|u_h^{\text{fd9-jss},\theta} - u_{i,h}^\theta\|}{\|u_{i,h}^\theta\|},\ err_{\text{fd9-dc}}(h,\theta) = \frac{\|u_h^{\text{fd9-dc},\theta} - u_{i,h}^\theta\|}{\|u_{i,h}^\theta\|}.$$

In Figure 5, we show how the θ averaged errors

$$err_{\text{fd9-jss}}(h) = \frac{1}{2\pi}\int_0^{2\pi} err_{\text{fd9-jss}}(h,\theta)d\theta,\ err_{\text{fd9-dc}}(h) = \frac{1}{2\pi}\int_0^{2\pi} err_{\text{fd9-dc}}(h,\theta)d\theta,$$

behave when h becomes small, for $k = 5, 10$. We can clearly see that fd9-dc is 6-th order accurate, while fd9-jss is just second order accurate. We show in Table 1 the condition number of the corresponding linear systems for the different schemes for

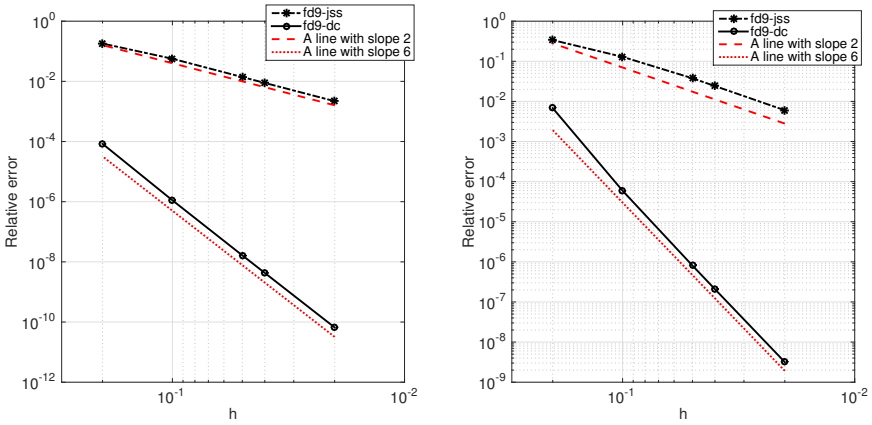


Fig. 5 Averaged relative errors of fd9-jss and fd9-dc for different mesh size h when $k = 5$ (left) and $k = 10$ (right).

different mesh sizes when $k = 10$. We can clearly see that our new method (fd9-dc) also reduces the condition number compared to the original FDM (fd5) or the FDM proposed by Jo, Shin and Suh (fd9-jss).

References

1. Ivo Babuška, Frank Ihlenburg, Ellen T Paik, and Stefan A Sauter. A generalized finite element method for solving the helmholtz equation in two dimensions with minimal pollution. *Computer methods in applied mechanics and engineering*, 128(3-4):325–359, 1995.
2. Ivo M Babuska and Stefan A Sauter. Is the pollution effect of the fem avoidable for the helmholtz equation considering high wave numbers? *SIAM Journal on numerical analysis*, 34(6):2392–2423, 1997.
3. Zhongying Chen, Dongsheng Cheng, and Tingting Wu. A dispersion minimizing finite difference scheme and preconditioned solver for the 3d helmholtz equation. *Journal of Computational Physics*, 231(24):8152–8175, 2012.
4. Dongsheng Cheng, Xu Tan, and Taishan Zeng. A dispersion minimizing finite difference scheme for the helmholtz equation based on point-weighting. *Computers & Mathematics with Applications*, 2017.
5. Oliver G Ernst and Martin J Gander. Multigrid methods for helmholtz problems: A convergent scheme in 1d using standard components.
6. Frank Ihlenburg and Ivo Babuška. Dispersion analysis and error estimation of galerkin finite element methods for the helmholtz equation. *International journal for numerical methods in engineering*, 38(22):3745–3774, 1995.
7. Frank Ihlenburg and Ivo Babuška. Finite element solution of the helmholtz equation with high wave number part i: The h-version of the fem. *Computers & Mathematics with Applications*, 30(9):9–37, 1995.
8. Churl-Hyun Jo, Changsoo Shin, and Jung Hee Suh. An optimal 9-point, finite-difference, frequency-space, 2-d scalar wave extrapolator. *Geophysics*, 61(2):529–537, 1996.

9. Changsoo Shin and Heejeung Sohn. A frequency-space 2-d scalar wave extrapolator using extended 25-point finite-difference operator. *Geophysics*, 63(1):289–296, 1998.
10. Christiaan C Stolk. A dispersion minimizing scheme for the 3-d helmholtz equation based on ray theory. *Journal of computational Physics*, 314:618–646, 2016.
11. Christiaan C Stolk, Mostak Ahmed, and Samir Kumar Bhowmik. A multigrid method for the helmholtz equation with optimized coarse grid corrections. *SIAM Journal on Scientific Computing*, 36(6):A2819–A2841, 2014.
12. Eli Turkel, Dan Gordon, Rachel Gordon, and Semyon Tsynkov. Compact 2d and 3d sixth order schemes for the helmholtz equation with variable wave number. *Journal of Computational Physics*, 232(1):272–287, 2013.
13. Tingting Wu. A dispersion minimizing compact finite difference scheme for the 2d helmholtz equation. *Journal of Computational and Applied Mathematics*, 311:497–512, 2017.

Optimized Schwarz methods for elliptic optimal control problems

Bérangère Delourme[1], Laurence Halpern[1], Binh Thanh Nguyen[1]

Abstract
The present paper deals with the design of optimized Robin-Schwarz methods for the algorithm of optimal control proposed in [1]. In both overlapping and non-overlapping cases, a full analysis of the problem is provided, and is illustrated with numerical tests.

1 Introduction

Let Ω be a bounded open set of $\mathbb{R}^2$, $z \in L^2(\Omega)$, and $\nu > 0$. We consider the following elliptic control problem described in [1] (see also [9, Chapter 2])

$$\min_{u \in L^2(\Omega)} \int_\Omega |y(u) - z|^2 dx + \nu \int_\Omega |u|^2 dx, \tag{1}$$

where, for a given function $f \in L^2(\Omega)$, $y(u)$ is the unique $H_0^1(\Omega)$ solution to

$$-\Delta y = f + u \text{ in } \Omega, \quad y = 0 \text{ on } \partial\Omega. \tag{2}$$

It is well known that the optimal control u (solution to (1)) is related to the adjoint state p by $u = -\frac{p}{\nu}$, and $(y, p) \in H_0^1(\Omega)^2$ is solution of the coupled problem

$$-\Delta y = f - \frac{p}{\nu} \quad -\Delta p = y - z \tag{3}$$

Introducing the new unknown $w = y + \frac{i}{\sqrt{\nu}} p$ (see [1]), Problem (3) is equivalent to the complex Helmholtz problem: find $w \in H_0^1(\Omega)$ such that

University Paris 13, Villetaneuse, France delourme@math.univ-paris13.fr

$$-\Delta w - \frac{i}{\sqrt{\nu}} w = g \text{ in } \Omega \qquad g = f - \frac{i}{\sqrt{\nu}} z. \tag{4}$$

In [2], Benamou and Després proposed a Robin's non-overlapping domain decomposition algorithm. Let us describe this algorithm (written here also for overlapping subdomains like in the original Schwarz algorithm). We consider the case where $\Omega = \mathbb{R}^2$ is split into two subdomains $\Omega_1 =]-\infty, \frac{L}{2}[\times\mathbb{R}$ and $\Omega_2 =]-\frac{L}{2}, +\infty[\times\mathbb{R}$. Here, L is a non-negative parameter that corresponds to the width of the overlapping zone between Ω_1 and Ω_2. We denote by n_j the outward unit normal vector to Ω_j, ∂_{n_j} the normal derivative on the boundary of Ω_j. Letting $\lambda^0 \in H^{1/2}(\partial\Omega_1)$ and $\ell \in \mathbb{C}$, we construct iteratively the sequences $(w_1^n)_{n\in\mathbb{N}}$, $(w_2^n)_{n\in\mathbb{N}}$ as follows: for any $n \in \mathbb{N}\backslash\{0\}$, find $w_1^n \in H^1(\Omega_1)$ and $w_2^n \in H^1(\Omega_2)$ such that

$$\begin{cases} -\Delta w_1^n - \frac{i}{\sqrt{\nu}} w_1^n = g & \text{in } \Omega_1, \\ \partial_{n_1} w_1^n + \ell w_1^n = \lambda^{n-1} & \text{on } \partial\Omega_1, \end{cases} \quad \begin{cases} -\Delta w_2^n - \frac{i}{\sqrt{\nu}} w_2^n = g & \text{in } \Omega_2, \\ \partial_{n_2} w_2^n + \ell w_2^n = \partial_{n_2} w_1^n + \ell w_1^n & \text{on } \partial\Omega_2, \end{cases} \tag{5}$$

$$\lambda^n = \partial_{n_1} w_2^n + \ell w_2^n\Big|_{\partial\Omega_1}.$$

It is easily seen ([1, Theorem 1]) that the problems defining w_1^n and w_2^n are well-posed if ℓ belongs to the angular sector $\mathcal{A}$ defined by

$$\mathcal{A} = \{z \in \mathbb{C} \text{ such that } \mathrm{Im}(z) < 0,\ \mathrm{Im}(z) + \mathrm{Re}(z) > 0\}. \tag{6}$$

Moreover, it is proved in [1, Theorem 2] (see also [2]), in the non-overlapping case, that the algorithm (5) converges, namely the sequence w_1^n (resp. w_2^n) tends to w (solution to (4)) in $H^1(\Omega_1)$ (resp. w in $H^1(\Omega_2)$).

The objective of the present work is to find a parameter $\ell \in \mathcal{A}$ that optimizes the rate of convergence of this algorithm. In the case of strongly elliptic real equation, this problem has been solved in [7] for Robin and Ventcel transmission conditions. In the former case, explicit values of the coefficients were given, whereas in the Ventcel case, only asymptotic formulas in terms of the mesh size are available. Extension to real Helmholtz equations were given in [6, 8]. Following these approaches, we consider the errors $e_1^n = w_1^n - w$ and $e_2^n = w_2^n - w$ and we denote by $\hat{e}_1^n$ and $\hat{e}_2^n$ their Fourier transform with respect to y, with Fourier variable k. It is easily seen that $\hat{e}_1^n$ and $\hat{e}_2^n$ follow a geometrical progression: more specifically, there exists two complex constants a_1 and a_2 such that

$$\hat{e}_j^n = a_j\, \delta(\ell,k)^{2n}\, e^{-\omega(k)|x|}, \quad \delta(\ell,k) = e^{-\omega(k)L}\, \frac{\omega(k)-\ell}{\omega(k)+\ell}, \ \omega(k) = \sqrt{k^2 - \frac{i}{\sqrt{\nu}}},$$

In the previous formulas, . Moreover, here and all over the text, the complex number $\sqrt{z}$ corresponds to the square root of z belonging to $\mathcal{A}$. As a result,

it suffices to minimize the modulus of δ (the square root of the convergence factor) in order to accelerate the convergence of the domain decomposition algorithm (5). As explained in [7, Section 4], we are interested in optimizing δ over a bounded interval $[k_{\min}, k_{\max}]$ (i.e. $k \in [k_{\min}, k_{\max}]$). In practice, the interval depends on the geometry of the domain and the mesh size ($k_{\max} = \frac{\pi}{h}$ where h denotes the characteristic length of the mesh). It leads us to investigate the following homographic best approximation problem (see [7, Section 4.2], [3] for the name in a time-dependent context): find $\delta^* \in \mathbb{R}$ such that

$$\delta^* = \inf_{\ell \in \mathbb{C}} \sup_{k \in [k_{\min}, k_{\max}]} |\delta(\omega(k), \ell)| \tag{7}$$

2 General results of well-posedness

The existence and uniqueness of an optimal parameter ℓ^* are direct consequences of the general results of [3, 4]:

Theorem 1 *For L sufficiently small, there exists a unique $\ell^* \in \mathcal{A}$ such that*

$$\delta^* = \inf_{\ell \in \mathbb{C}} \sup_{k \in [k_{\min}, k_{\max}]} |\delta(\omega(k), \ell)| = \max_{k \in [k_{\min}, k_{\max}]} |\delta(\omega(k), \ell^*)|. \tag{8}$$

Moreover, there exists at least two distinct real numbers $(k_1, k_2) \in [k_{\min}, k_{\max}]^2$ such that

$$\max_{k \in [k_{\min}, k_{\max}]} |\delta(\omega(k), \ell^*)| = |\delta(\omega(k_1), \ell^*)| = |\delta(\omega(k_2), \ell^*)|. \tag{9}$$

Proof (Sketch of the proof of Theorem 1). By contradiction, one can verify that if there exists $\ell^* \in \mathbb{C}$ satisfying (8), then $\ell^* \in \mathcal{A}$ (see e.g. [3, Lemma 4.5] for a similar proof). Then, the existence of ℓ^* ([3, Theorem 2.2 and Theorem 2.8]) results from a compactness argument (k belongs to the compact set $[k_{\min}, k_{\max}]$). Finally, in the non-overlapping case ($L = 0$), the uniqueness is proved in [3, Theorem 2.6]. For $L \neq 0$ and sufficiently small, the uniqueness proof results from an adaptation of [4, Theorem 8]. In both cases, the uniqueness is a consequence of convexity properties and the equi-oscillation property (9)([3, Theorem 2.5 and Theorem 2.11]).

3 Characterization of the optimal parameter in the non-overlapping case

Theorem 2 *The best parameter ℓ^* defined by (8) is given by*

$$\ell^* = \sqrt{\omega_{\min}\omega_{\max}}, \quad \delta^* = \left| \frac{\sqrt{\omega_{\min}} - \sqrt{\omega_{\max}}}{\sqrt{\omega_{\min}} + \sqrt{\omega_{\max}}} \right| \tag{10}$$

where $\omega_{\min} = \omega(k_{\min})$ *and* $\omega_{\max} = \omega(k_{\max})$. *Moreover, if* $k_{\max} = \frac{\pi}{h}$, δ^* *and* ℓ^* *admit the following asymptotic expansion*

$$\delta^* = 1 \; - \; 2\,h^{1/2}\,\frac{\mathrm{Re}(\sqrt{\omega_{\min}})}{\sqrt{\pi}} \; + \; o(h^{1/2}), \;\; \ell^* \; = \; h^{-1/2}\left(\sqrt{\pi}\,\sqrt{\omega_{\min}} \; + \; o(1)\right). \tag{11}$$

We remark that Formula (10) is the same as in the real positive case (see [7, Theorem 4.4]). The reminder of this section is dedicated to the proof of Theorem 2. First, we remark that in the non-overlapping case (and as in the real case), the equi-oscillation property (9) holds for exactly two points that are nothing but $k_{\min}$ and $k_{\max}$ (the proof of this result may be done using either a geometrical argument or a direct investigation of the derivative of $|\delta(k,\ell)|^2$ with respect to k, see [5]):

Lemma 1 *Let* ℓ^* *be defined by* (8). *Then,*

$$\max_{k\in[k_{\min},k_{\max}]} |\delta(\omega,\ell^*)| = |\delta(\omega_{\min},\ell^*)| = |\delta(\omega_{\max},\ell^*)|, \tag{12}$$

and, for any $k \in]k_{\min}, k_{\max}[$, $|\delta(\omega(k),\ell^*)| < |\delta(\omega_{\min},\ell^*)|$.

The previous lemma motivates us to consider the curve of equioscillation Π defined by

$$\Pi = \left\{ \ell = re^{i\theta} \in \mathcal{A} \text{ such that } |\delta(\omega_{\min},\ell)| = |\delta(\omega_{\max},\ell)| \right\}, \tag{13}$$

so that the optimization problem (8) can then be rewritten as follows: find $\ell^* \in \Pi$ such that

$$\delta^* = \min_{\ell\in\Pi} |\delta(\omega_{\min},\ell)| = \min_{\ell\in\Pi} |\delta(\omega_{\max},\ell)|. \tag{14}$$

Note that, unlike in the real case, the set Π is not reduced to the singleton $\{p = \sqrt{\omega_{\min}\omega_{\max}}\}$. Nevertheless, $\sqrt{\omega_{\min}\omega_{\max}}$ still belongs to Π. To continue the proof, it is useful to introduce the perpendicular bisector Δ of the segment $[\omega_{\min},\omega_{\max}]$, i.e. $\Delta = \{z = x+iy \in \mathbb{C} \text{ s.t. } y = ax+b\}$ where $a = -\frac{\mathrm{Re}(\omega_{\max}-\omega_{\min})}{\mathrm{Im}(\omega_{\max}-\omega_{\min})}$ and $b = \frac{|\omega_{\max}|^2-|\omega_{\min}|^2}{2\mathrm{Im}(\omega_{\max}-\omega_{\min})}$. For any $\ell \in \mathbb{C}$, we also consider the signed distance between ℓ and Δ, namely the function $d(\ell) = \frac{a\mathrm{Re}(\ell)-\mathrm{Im}(\ell)+b}{\sqrt{1+a^2}}$. Using the intercept theorem, it is easily seen that the best parameter ℓ^* corresponds to the point of Π for which the distance between Π and Δ is minimal:

Lemma 2 *The function* $\eta : \Pi \to \mathbb{R}$, *defined by* $\eta(\ell) = |\delta(\ell,\omega_{\min})| = |\delta(\ell,\omega_{\max})|$ *is a strictly increasing function of the signed distance* d: *for any* $(\ell_1,\ell_2) \in \Pi^2$ *such that* $d(\ell_1) < d(\ell_2)$, $\eta(\ell_1) < \eta(\ell_2)$.

In other words it suffices to study the variations of the distance function d over Π in order to characterize the best parameter ℓ. By a standard investigation of d we prove the following lemma:

Lemma 3 *The function d reaches its minimum over Π for $\ell^* = \sqrt{\omega_{\min}\omega_{\max}}$.*

The proof of Theorem 2 is completed by a standard asymptotic expansion of δ^* for $k_{\max}$ large.

4 Asymptotics of the optimal parameter in the overlapping case

In the overlapping case ($L > 0$), we are not able to obtain an explicit characterization of the best parameter ℓ^*. Nevertheless, we are able to compute its asymptotic behaviour for h small when the overlapping parameter $L = h$ and $k_{\max} = \frac{\pi}{h}$,

Theorem 3 *Assume that $L = h$ and $k_{\max} = \frac{\pi}{h}$.*

- *For h sufficiently small, there exists $k^* \in]k_{\min}, k_{\max}[$ such that*

$$\max_{k\in[k_{\min},k_{\max}]} |\delta(\omega, \ell^*)| = |\delta(\omega_{\min}, \ell^*)| = |\delta(\omega(k^*), \ell^*)|, \tag{15}$$

and, for any $k \in]k_{\min}, k^*[\cup]k^*, k_{\max}]$, $|\delta(\omega(k), \ell^*)| < |\delta(\omega_{\min}, \ell^*)|$.

- *The optimal parameter ℓ^* and the corresponding convergence factor δ^* admit the following asymptotic expansion:*

$$\ell^* = h^{-1/3}\left((c_x - ic_y) + o(1)\right) \quad \text{and} \quad \delta^* = 1 - c_r h^{1/3} + o(h^{1/3}), \tag{16}$$

where, introducing $r_{\min} = \mathrm{Re}(\omega_{\min})$ and $i_{\min} = \mathrm{Im}(\omega_{\min})$,

$$c_x = \left(\frac{r_{\min} + \sqrt{r_{\min}^2 + i_{\min}^2}}{2\sqrt{2}}\right)^{2/3}, \quad c_y = -\frac{i_{\min}}{2\sqrt{2c_x}}, \text{ and } \quad c_r = 2\sqrt{2c_x}. \tag{17}$$

Proof. The proof of Theorem 3 is divided into two main parts. We first construct a formal asymptotic expansion of ℓ^* that we justify *a posteriori*. To start with, we make an 'ansatz' on the asymptotic behaviour of the optimal parameter ℓ^*. We assume that

$$\ell^* \sim ch^{-\alpha} \text{ with } \alpha \in]0, 1[\text{ and } c = c_x - ic_y \quad (c_x > 0, c_y > 0).$$

Then, computing explicitly the derivative of $|\delta(\ell, k)|^2$, we prove that, in this asymptotic regime, the equi-oscillation property (9) holds for exactly two points $k_1 = k_{\min}$ and $k_2 = k_*$, where k_* admits the following asymptotic:

$$k_* = 2^{1/4}(c_x)^{1/4}h^{(-\alpha-1)/4} + o(h^{(-\alpha-1)/4}), \text{ and}$$

$$|\delta(\omega(k_*),\ell^*)|^2 \sim 1 - 4(2c_x)^{1/2}h^{\frac{1-\alpha}{2}} \quad |\delta(\omega_{\min},\ell^*)|^2 \sim 1 - 4h^{\alpha}\frac{(c_x r_{\min} - c_y i_{\min})}{|c|^2}$$

Identifying the previous two expansions leads to

$$\alpha = \frac{1}{3} \quad \text{and} \quad \sqrt{2c_x}(c_x^2 + c_y^2) - (c_x r_{\min} - c_y i_{\min}) = 0. \tag{18}$$

Thus, in order to minimize the convergence factor (in this asymptotic regime), it suffices to find the couple (c_x, c_y) satisfying (18)(right) and such that c_x is maximal. A direct analysis of equation (18) leads to (17).

It remains to justify the obtained formal asymptotic. For $h \in (0,1)$ and $\varepsilon > 0$ sufficiently small, let

$$\mathscr{L}_h = \left\{\ell \in \mathbb{C},\ \text{s. t. } h^{1/3}(\ell_x,\ell_y) \in\ [c_x - \varepsilon\,, c_x + \varepsilon] \times [-c_y - \varepsilon, -c_y + \varepsilon]\right\},$$

where c_x and c_y are defined by (17). Then, for h sufficiently small (in order to be able to define k^*), let $\mathit{\Gamma}_h = \{\ell \in \mathscr{L}_h, |\delta(\omega_{\min},\ell)| = |\delta(\omega(k^*,\ell)|\}$. Because $\mathit{\Gamma}_h$ is closed and non empty, there exists ℓ_*^h such that

$$|\delta(\omega_{\min},\ell_*^h)| = \inf_{\ell\in\mathit{\Gamma}_h} |\delta(\omega_{\min},\ell)|. \tag{19}$$

It is not difficult to prove that ℓ_*^h admits the asymptotic expansion (16). The end the proof of Theorem 3 consists in showing that $\ell_*^h = \ell^*$. This is done by proving the following lemma:

Lemma 4 *ℓ_*^h is a strict local minimum for $\ell \mapsto \|R(\omega(k),\ell)\|_{L^\infty(k_{\min},k_{\max})}$.*

Indeed, Corollary 2.16 in [3] guarantees that any strict local minimum of the function $\ell \mapsto \|R(\omega(k),\ell)\|_{L^\infty(k_{\min},k_{\max})}$ is the global minimum. Consequently $\ell_*^h = \ell^*$ and the proof is complete. The proof of Lemma (4) is an adaptation of the proof of [3, Theorem 4.2].

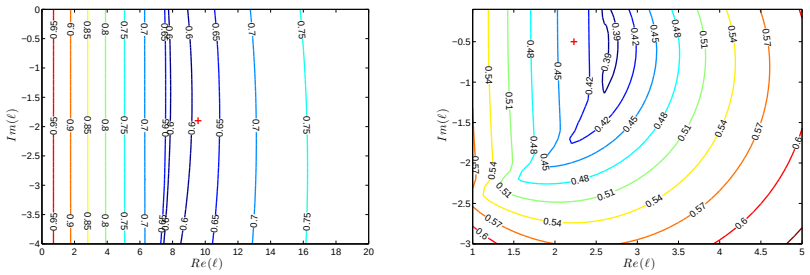
5 Numerical illustration

Let $\Omega =]0,\pi[^2$, $\nu = 1$ and $f = z = 0$ (hence $g = 0$), so that the exact solution is 0. The discretization is done using a standard second order finite difference scheme. We choose a similar discretization in the x and y directions ($h_x = h_y = h$) and we set $k_{\min} = 1$ and $k_{\max} = \frac{\pi}{h}$. In the non-overlapping case, we split the domain Ω into two domains Ω_1 and Ω_2 of equal size: $\Omega_1 =]0,\pi/2[\times]0,\pi[$ and $\Omega_2 =]\pi/2,\pi[\times]0,\pi[$. In the overlapping case, we take $\Omega_1 =]0,\pi/2[\times]0,\pi[$ and $\Omega_2 =]\pi/2 - h,\pi[\times]0,\pi[$ (i.e. $L = h$). The domain

decomposition algorithm is initialized with a uniform (over $]0,1[$) random data λ_1^0. In the next experiments, we evaluate the numerical (or observed) convergence rate $\delta_{\text{num}}(\ell, N)$ defined by

$$\delta_{\text{num}}(\ell, N) = \left(\frac{e_N}{e_{N-1}}\right)^{1/2}, \quad e_n = \sqrt{\|u_{h,1}^n\|^2 + \|u_{h,2}^n\|^2} \tag{20}$$

On Figure 1, we evaluate $\delta_{\text{num}}(\ell, N)$ for different values of ℓ taking $N = 60$ and $h = \pi/80$. The red cross corresponds to theoretical optimal parameter ℓ^*: in the non-overlapping case, $\ell^* = \sqrt{\omega_{\min}\omega_{\max}}$ while in the overlapping case, ℓ^* is numerically computed. Although the theoretical analysis is done for a two dimensional unbounded domain, we remark that the theoretical optimal parameter ℓ^* and the observed optimal parameter are relatively closed. Moreover, for $L = 0$, the convergence factor slowly varies with respect to the imaginary part of ℓ (cf. [5]). Then, Figure 2a presents the evolution of the error e_n with respect to the number n of iterations of the domain decomposition algorithm for two different values of ℓ: $\ell = \ell^*$ and $\ell = \ell^*_{\text{num}}$, where ℓ^*_{num} denotes the numerical optimized coefficient obtained by optimizing $\delta_{\text{num}}(\ell, N)$. Finally, Figure 2b shows the evolution of $1 - \delta_{\text{num}}(\ell, N)$ with respect to the discretization parameter h. The introduction of the overlap perceptibly improves the observed convergence rate (although the asymptotic regime is not entirely reached in this case).

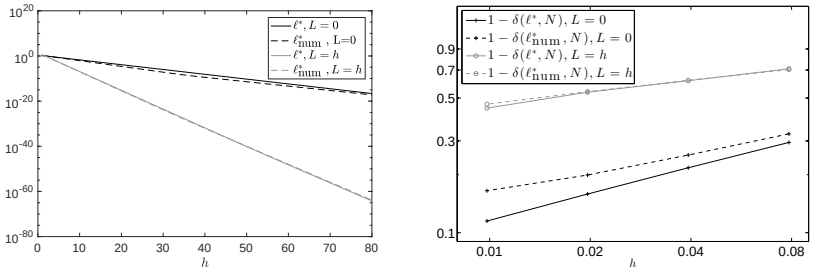


(a) $L = 0$. (b) $L = h$

Fig. 1: Contour plot of $\delta_{\text{num}}(\ell, N)$ for $h = \pi/80$, $N = 60$

References

[1] Jean-David Benamou. A domain decomposition method with coupled transmission conditions for the optimal control of systems governed by



(a) Error w.r.t the num. of iter.

(b) $\delta_{\rm num}(\ell, N)$ w.r.t h

Fig. 2: Error and convergence factor in the overlapping and non-overlapping cases

elliptic partial differential equations. *SIAM J. Numer. Anal.*, 33(6):2401–2416, 1996.

[2] Jean-David Benamou and Bruno Després. A domain decomposition method for the Helmholtz equation and related optimal control problems. *J. Comput. Phys.*, 136(1):68–82, 1997.

[3] D. Bennequin, M. J. Gander, and L. Halpern. A homographic best approximation problem with application to optimized Schwarz waveform relaxation. *Math. Comp.*, 78(265):185–223, 2009.

[4] Daniel Bennequin, Martin J. Gander, Loic Gouarin, and Laurence Halpern. Optimized Schwarz waveform relaxation for advection reaction diffusion equations in two dimensions. *Numer. Math.*, 134(3):513–567, 2016.

[5] B. Delourme and L. Halpern. Optimized Schwarz method for control problems. In preparation.

[6] M. J. Gander, L. Halpern, and F. Magoulès. An optimized Schwarz method with two-sided Robin transmission conditions for the Helmholtz equation. *Internat. J. Numer. Methods Fluids*, 55(2):163–175, 2007.

[7] Martin J. Gander. Optimized Schwarz methods. *SIAM J. Numer. Anal.*, 44(2):699–731 (electronic), 2006.

[8] Martin J. Gander, Frédéric Magoulès, and Frédéric Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24(1):38–60, 2002.

[9] J.-L. Lions. *Optimal control of systems governed by partial differential equations*. Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer-Verlag, New York-Berlin, 1971.

Auxiliary space preconditioners for a DG discretization of $H(\mathbf{curl};\Omega)$- elliptic problem on hexahedral meshes

B. Ayuso de Dios, R. Hiptmair, and C. Pagliantini

Abstract We present a family of preconditioners based on the *auxiliary space method* for a discontinuous Galerkin discretization on cubical meshes of $H(\mathbf{curl};\Omega)$- elliptic problems with possibly discontinuous coefficients. We address the influence of possible discontinuities in the coefficients on the asymptotic performance of the proposed solvers and present numerical results in two dimensions.

1 Introduction

Let $\Omega \subset \mathbb{R}^3$ be a simply connected bounded domain with Lipschitz boundary and let $\boldsymbol{f} \in L^2(\Omega)^3$. We consider the following $H(\mathbf{curl};\Omega)$-elliptic problem

$$\begin{cases} \nabla \times (\nu \nabla \times \boldsymbol{u}) + \beta \boldsymbol{u} = \boldsymbol{f} & \text{in } \Omega, \\ \boldsymbol{u} \times \mathbf{n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{1}$$

where $\nu = \nu(\mathbf{x}) \geq \nu_0 > 0$ and $\beta = \beta(\mathbf{x}) \geq \beta_0 > 0$ are assumed to be in $L^\infty(\Omega)$ but possibly discontinuous, and represent properties of the medium or material: ν is typically the inverse of the magnetic permeability and β is proportional to the ratio of electrical conductivity and the time step. Problem (1) arises in the modelling of magnetic diffusion and also after implicit time discretization of

Blanca Ayuso de Dios
Dipartimento di Matematica, Università di Bologna, Piazza di Porta San Donato 5, Bologna, & IMATI-CNR, Pavia, Italy, e-mail: blanca.ayuso@unibo.it

Ralf Hiptmair
Seminar for Applied Mathematics, ETH Zürich, Rämistrasse 101, Zürich, Switzerland, e-mail: hiptmair@sam.math.ethz.ch

Cecilia Pagliantini
EPFL-SB-MATH-MCSS, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland, e-mail: cecilia.pagliantini@epfl.ch

resistive magneto-hydrodynamics (MHD). In connection with the MHD application the use of hexahedral meshes is typically preferred over to family partitions made of simplices [Pagliantini(2016)].
Finite element discretizations using edge elements of the first family [Nédélec(1980)] are probably the most satisfactory methods to approximate (1) from a theoretical point of view. Only recently, a new *compatible* element (corresponding to an edge element of the second family) has been introduced in [Arnold and Awanou(2014)]. Discontinuous Galerkin (DG) methods offer an attractive alternative to conforming FE edge elements [Houston et al.(2005)] and allow for great flexibility in incorporating the discontinuities of the medium. For both methods, the condition number of the resulting linear systems degrades with mesh refinement and the size of the variations of the coefficients. Hence, designing a preconditioner able to cope with the combined effect of the mesh width and of highly varying coefficients turns out to be essential. For constant coefficients, efficient solvers for FE edge discretizations have been successfully developed using domain decomposition (DD) and the Auxiliary Space (AS) method [Hiptmair and Xu(2007)]. For discontinuous coefficients, a non-overlapping BDDC algorithm has been proposed and analyzed in [Dohrmann and Widlund(2016)], improving previous results in the DD literature, see e.g. [Toselli(2006)]. Recently, in [Ayuso de Dios et al.(2017)], we have developed a family of AS preconditioners for DG discretizations of (1), providing the analysis for simplicial meshes and in the case of cubical meshes when edge elements of the first kind are used as local spaces. In this paper, we report on the construction of the AS preconditioners focusing on the case of cubical meshes, discussing also their performance in the case of jumping coefficients. The proposed preconditioners rely on $H(\mathbf{curl};\Omega)$-conforming auxiliary spaces (as *auxiliary space*) and hence is presumed the availability of a (direct) solver for standard $H(\mathbf{curl};\Omega)$-conforming Galerkin discretizations.

2 SIPG Discretization on Hexahedral Meshes

Let $\mathscr{T}_h$ be a family of shape-regular partitions of Ω into cubes T. For each $T \in \mathscr{T}_h$, let $h_T = \mathrm{diam}(T)$ and set $h = \max_{T\in\mathscr{T}_h} h_T$. We assume that $\mathscr{T}_h$ is conforming and resolves the piece-wise constant coefficients β and ν. (i.e., $\nu_T, \beta_T \in \mathbb{P}^0(T)$ for all $T \in \mathscr{T}_h$).We denote by $\mathscr{F}_h$ the set of all faces of the partition; $\mathscr{F}_h^o$ and $\mathscr{F}_h^\partial$ refer respectively, to the collection of all interior and boundary faces. Similarly, $\mathscr{E}_h = \mathscr{E}_h^o \cup \mathscr{E}_h^\partial$ denote the set of all edges of the skeleton of $\mathscr{T}_h$; with $\mathscr{E}_h^o$ and $\mathscr{E}_h^\partial$ referring to interior and boundary edges, respectively. We define the sets:

$$\mathscr{T}(e) := \{T \in \mathscr{T}_h : e \subset \partial T\}; \qquad \mathscr{E}(T) := \{e \in \mathscr{E}_h : e \subset \partial T\};$$
$$\mathscr{F}(T) := \{f \in \mathscr{F}_h : f \subset \partial T\}; \qquad \mathscr{F}(e) := \{f \in \mathscr{F}_h : e \subset \partial f\}.$$

We introduce the (family of) DG finite element spaces

$$\mathbf{V}_h^{DG} = \{\boldsymbol{v} \in L^2(\Omega)^3 : \ \boldsymbol{v} \in \mathscr{M}(T), T \in \mathscr{T}_h\}, \qquad \mathscr{M}(T) \subseteq \mathbb{Q}_k(T)^3$$

where the local space $\mathscr{M}(T)$ of vector-valued polynomials can be of three types:

1. Nédélec elements of first family on cubical meshes [Nédélec(1980)]

$$\mathscr{M}(T) = \mathscr{N}^I(T) := \mathbb{Q}_{k-1,k,k}(T) \times \mathbb{Q}_{k,k-1,k}(T) \times \mathbb{Q}_{k,k,k-1}(T), \qquad k \geq 1,$$

where $\mathbb{Q}_{\ell,m,n}(T)$ is the space of polynomials of degree at most ℓ, m, n in each vector variable.

2. Compatible elements (of second kind) [Arnold and Awanou(2014)]:

$$\begin{aligned}\mathscr{M}(T) = \mathscr{S}_k(T) := (\mathbb{P}_k(T))^3 + \operatorname{span}\{&[yz(w_2(x,z) - w_3(x,y)), zx(w_3(x,y) - w_1(y,z)), \\ &xy(w_1(y,z) - w_2(x,z))] + \nabla s(x,y,z)\},\end{aligned}$$

where each $w_i \in \mathbb{P}_k$ and $s \in \mathbb{P}_k(T)$ has superlinear degree (ordinary degree ignoring variables which appear linearly) at most $k+1$, with $k \geq 1$.

3. Full polynomials: We set the local space $\mathscr{M}(T) = (\mathbb{Q}_k(T))^3$, and $k \geq 1$.

For each choice of the resulting $\mathbf{V}_h^{DG}$ space, the corresponding $\mathrm{H}_0(\mathbf{curl},\Omega)$-conforming finite element spaces are defined as:

$$\mathbf{V}_h^c := \mathbf{V}_h^{DG} \cap \mathrm{H}_0(\mathbf{curl},\Omega) = \{\boldsymbol{v} \in \mathrm{H}_0(\mathbf{curl},\Omega) : \ \boldsymbol{v} \in \mathscr{M}(T), T \in \mathscr{T}_h\}. \tag{2}$$

For a piecewise smooth vector-valued function $\boldsymbol{v}$, we denote by $\boldsymbol{v}^{\pm}$ the traces of $\boldsymbol{v}$ taken from within $T^{\pm}$. The tangential jump, indicated by $[\![\cdot]\!]_\tau$, is defined by

$$[\![\boldsymbol{v}]\!]_\tau := \mathbf{n}^+ \times \boldsymbol{v}^+ + \mathbf{n}^- \times \boldsymbol{v}^- \quad \text{on} \quad f \in \mathscr{F}_h^o, \qquad [\![\boldsymbol{v}]\!]_\tau := \mathbf{n} \times \boldsymbol{v} \text{ on } f \in \mathscr{F}_h^\partial$$

where $\mathbf{n}^+$ and $\mathbf{n}^-$ denote the unit normal vectors on $f = \partial T^+ \cap \in \partial T^-$ pointing outwards from T^+ and T^-, respectively. We will also use the notation

$$(\theta \boldsymbol{u}, \boldsymbol{v})_{\mathscr{T}_h} = \sum_{T \in \mathscr{T}_h} \int_T \theta_T \boldsymbol{u}\boldsymbol{v} d\mathbf{x}, \qquad \langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathscr{F}_h} = \sum_{f \in \mathscr{F}_h} \int_f \boldsymbol{u}\boldsymbol{v} ds \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathbf{V}_h^{DG}$$

where $\theta \in \mathbb{P}^0(\mathscr{T}_h)$ will be either $\theta = \nu$ or $\theta = \beta$.

The SIPG-DG method. We consider a symmetric Interior Penalty method (SIPG) introduced recently in [Ayuso de Dios et al.(2017)] for approximating (1) robustly (w.r.t the discontinuous coefficients). The method reads:

$$\text{Find} \quad \boldsymbol{u}_h \in \mathbf{V}_h^{DG} \quad \text{such that} \quad a_{\mathrm{DG}}(\boldsymbol{u}_h, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v})_{\mathscr{T}_h} \quad \forall \boldsymbol{v} \in \mathbf{V}_h^{DG}, \tag{3}$$

with $a_{\mathrm{DG}}(\cdot,\cdot)$ defined by

$$\begin{aligned}a_{\mathrm{DG}}(\boldsymbol{u}, \boldsymbol{v}) := &(\nu \nabla \times \boldsymbol{u}, \nabla \times \boldsymbol{v})_{\mathscr{T}_h} + (\beta \boldsymbol{u}, \boldsymbol{v})_{\mathscr{T}_h} - \langle \{\{\nu \nabla \times \boldsymbol{u}\}\}_\gamma, [\![\boldsymbol{v}]\!]_\tau \rangle_{\mathscr{F}_h} \\ &- \langle [\![\boldsymbol{u}]\!]_\tau, \{\{\nu \nabla \times \boldsymbol{v}\}\}_\gamma \rangle_{\mathscr{F}_h} + \sum_{T \in \mathscr{T}_h} \alpha_T(\nu) \sum_{e \in \mathscr{E}(T)} \sum_{f \in \mathscr{F}(e)} (s_f [\![\boldsymbol{u}]\!]_\tau, [\![\boldsymbol{v}]\!]_\tau)_{0,f}.\end{aligned} \tag{4}$$

In (4), the *weighted average* $\{\{\cdot\}\}_\gamma$ is defined as the plain trace for a boundary face, whereas for $\partial T^+ \cap \partial T^- = f \in \mathscr{F}_h^o$, is given by

$$\{\{\boldsymbol{u}\}\}_\gamma := \gamma_f^+ \boldsymbol{u}^+ + \gamma_f^- \boldsymbol{u}^- \quad \text{with} \quad \gamma_f^\pm = \frac{\nu^\mp}{\nu^+ + \nu^-}, \qquad \nu^\pm := \nu_{|_{T^\pm}} .$$

The penalization is defined by $s_f := c h_f^{-1}$ on all $f \in \mathscr{F}_h$ with some $c > 0$ and the mesh function $h_f = \min\{h_{T^+}, h_{T^-}\}$ on $f \in \mathscr{F}_h^o$ and $h_f = h_T$ on $f = \partial T \cap \partial\Omega$.
The coefficient function $(\alpha_T(\nu))_{T\in\mathscr{T}_h} \in \mathbb{P}_0(\mathscr{T}_h)$ is defined by

$$\alpha_T(\nu) := \max_{f\in\mathscr{F}(T)} \{\{\nu\}\}_{*,f} \quad \text{with} \quad \{\{\nu\}\}_{*,f} := \begin{cases} \max\limits_{\substack{T\in\mathscr{T}(e)\\ e\subset\partial f}} \nu_T & f \in \mathscr{F}_h^o , \\ \nu_T & f \in \mathscr{F}_h^\partial . \end{cases}$$

Notice that $\alpha_T(\nu)$ picks the maximum conductivity coefficient over a patch of elements surrounding T. In Figure 1 a 2D sketch of such patch is given.

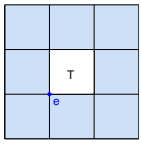


Fig. 1: 2D sketch of the patch involved in definition of $\alpha_T(\nu)$.

We stress that the weighted average $\{\{\cdot\}\}_\gamma$ together with $\{\{\cdot\}\}_{*,f}$ and the definition of $\alpha_T(\nu)$ ensure robustness (with respect to the coefficients) of both the approximation (3) in the energy norm (see [Ayuso de Dios et al.(2017), Proposition 2.1], and [Pagliantini(2016), Proposition 5.1.1]) and the preconditioners, see Theorem 1 and [Ayuso de Dios et al.(2017), Pagliantini(2016)] for details in the analysis. Observe that when the variational formulation (3) is restricted to $\mathbf{V}_h^c$ in (2), the corresponding $\mathrm{H}_0(\mathbf{curl},\Omega)$-conforming discretization of (1) is obtained. In fact,

$$a_{\mathscr{W}}(\boldsymbol{u},\boldsymbol{v}) := (\nu\nabla\times\boldsymbol{u}, \nabla\times\boldsymbol{v})_\Omega + (\beta\boldsymbol{u},\boldsymbol{v})_\Omega = a_{DG}(\boldsymbol{u},\boldsymbol{v}) \quad \forall\, \boldsymbol{u},\boldsymbol{v} \in \mathbf{V}_h^c. \tag{5}$$

We denote by $\mathscr{A} : \mathbf{V}_h^{DG} \longrightarrow (\mathbf{V}_h^{DG})'$ the discrete operator $(\mathscr{A}\boldsymbol{u},\boldsymbol{w}) = a_{DG}(\boldsymbol{u},\boldsymbol{w})$ and by $\mathbb{A}$ the matrix representation of $\mathscr{A}$ with respect to a localized "nodal" basis of $\mathbf{V}_h^{DG}$ (using any of the choices for $\mathscr{M}(T)$). It can be verified that the spectral condition number $\kappa(\mathbb{A})$ is proportional to

$$h^{-2}\frac{\max_T \alpha_T(\nu)}{\min_T \nu_T} + \frac{\max_T \beta_T}{\min_T \beta_T} .$$

3 Auxiliary Space Preconditioning

The Auxiliary Space Method (ASM) was introduced in [Xu(1996), Oswald(1996)] as an expansion of the Fictitious Space Method [Nepomnyaschikh(1991)] providing a neat methodology for developing and analysing preconditioners. To describe the

preconditioners we propose, based on the AS methodology, we first review the basic ingredients behind the Fictitious Space Method:

(1) the *fictitious space*: a real finite dimensional Hilbert space $\overline{\mathscr{V}}$, endowed with an inner product $\bar{a}(\cdot,\cdot)$, induced operator $\overline{\mathscr{A}} : \overline{\mathscr{V}} \to \overline{\mathscr{V}}'$ and norm $\|\cdot\|_{\overline{\mathscr{A}}}$.
(2) A continuous, linear and surjective transfer operator $\Pi : \overline{\mathscr{V}} \to \mathbf{V}_h^{DG}$.

By virtue of [Nepomnyaschikh(1991)], an optimal preconditioner for $\overline{\mathscr{A}}$ would result in an optimal preconditioner for $\mathscr{A}$. The distinguishing feature of ASM is the particular choice of $\overline{\mathscr{V}}$ as a product space, including the original space as one of the components. Here, we set $\overline{\mathscr{V}} = \mathbf{V}_h^{DG} \times \mathscr{W}$, endowed with the inner product

$$\overline{a}(\overline{\boldsymbol{v}},\overline{\boldsymbol{v}}) = s(\boldsymbol{v}_0,\boldsymbol{v}_0) + a_{\mathscr{W}}(\boldsymbol{w},\boldsymbol{w}), \qquad \forall \overline{\boldsymbol{v}} = (\boldsymbol{v}_0,\boldsymbol{w}),\ \boldsymbol{v}_0 \in \mathbf{V}_h^{DG},\ \boldsymbol{w} \in \mathscr{W}, \tag{6}$$

where $\mathscr{W}$ is the (truly) so-called auxiliary space and $a_{\mathscr{W}}(\cdot,\cdot)$ is the auxiliary bilinear form. We will always take as $\mathscr{W}$ an $\mathrm{H}_0(\mathbf{curl},\Omega)$-conforming space $\mathbf{V}_h^c$. In (6), $s(\cdot,\cdot)$ is the bilinear form associated with a relaxation operator $\mathscr{S}$ on $\mathbf{V}_h^{DG}$. Denoting by $\mathscr{A}_{\mathscr{W}}$ the operator associated with $a_{\mathscr{W}}(\cdot,\cdot)$, the auxiliary space preconditioner operator is $\mathscr{B} = \mathscr{S}^{-1} + \Pi_{\mathscr{W}} \circ \mathscr{A}_{\mathscr{W}}^{-1} \circ \Pi_{\mathscr{W}}^*$ where the linear transfer operator $\Pi_{\mathscr{W}} : \mathscr{W} \to \mathbf{V}_h^{DG}$ is the standard inclusion and its adjoint $\Pi_{\mathscr{W}}^* : \mathbf{V}_h^{DG} \to \mathscr{W}$ is defined by $a_{\mathscr{W}}(\Pi_{\mathscr{W}}^* \boldsymbol{v},\boldsymbol{w}) = a(\boldsymbol{v},\Pi_{\mathscr{W}}\boldsymbol{w})$, $\boldsymbol{v} \in \mathbf{V}_h^{DG}$, $\boldsymbol{w} \in \mathscr{W}$. If $\mathbb{S} \in \mathbb{R}^{N\times N}$ with $N := \dim \mathbf{V}_h^{DG}$ and $\mathbb{A}_W \in \mathbb{R}^{N_W \times N_W}$, $N_W := \dim \mathscr{W}$, then the preconditioner in algebraic form reads

$$\mathbb{B} = \mathbb{S}^{-1} + \mathbb{P}\mathbb{A}_W^{-1}\mathbb{P}^{\mathsf{T}}, \tag{7}$$

where $\mathbb{P} \in \mathbb{R}^{N\times N_W}$ is the matrix representation of the transfer operator $\Pi_{\mathscr{W}}$.

We now specify the precise components for the two preconditioners we propose:

1. Natural Preconditioner: We set $\mathscr{W} = \mathbf{V}_h^c = \mathbf{V}_h^{DG} \cap \mathrm{H}_0(\mathbf{curl},\Omega)$ for any choice of the local space $\mathscr{M}(T)$ and $a_{\mathscr{W}}(\cdot,\cdot)$ is as in (5). Hence, $\mathscr{A}_{\mathscr{W}} : \mathbf{V}_h^c \to (\mathbf{V}_h^c)'$ is self-adjoint and positive definite. As relaxation operator $\mathscr{S}$ it is sufficient to use a simple Jacobi or block Jacobi smoother.

2. Coarser or Economical Preconditioner: When the local space is either $\mathscr{M}(T) = \mathscr{S}_k(T)$ or $\mathscr{M}(T) = (\mathbb{Q}_k(T))^3$ in the construction of the $\mathbf{V}_h^{DG}$-space, we consider a second possibility for the AS preconditioner. We take $\mathscr{W}$ as

$$\mathscr{W} := \mathscr{W}_h^c = \{\boldsymbol{w} \in \mathrm{H}_0(\mathbf{curl},\Omega) : \boldsymbol{w}|_T \in \mathscr{N}^I(T), T \in \mathscr{T}_h\} \subset \mathbf{V}_h^c \subset \mathbf{V}_h^{DG}.$$

As to the relaxation operator, we demonstrate numerically that a non-overlapping Schwarz smoother is not able to resolve the components in the kernel of $\mathbf{curl}(\mathscr{W})$ and as a consequence an overlapping smoother is necessary. We will show numerically that in the case $\mathscr{M}(T) = (\mathbb{Q}_k(T))^3$, the resulting AS preconditioner is not effective, independently of the choice of the smoother and the amount of domain overlaps involved in its construction. We suspect that this is connected to the fact that the DG method using $\mathscr{M}(T) = (\mathbb{Q}_k(T))^3$ is not spectrally correct, while $\mathscr{W}_h^c$ is.

Next result provides the convergence of the *Natural Preconditioner*.

Theorem 1. *Let $\mathbb{B}$ be the auxiliary space preconditioner in* (7)*, with $\mathscr{W} = \mathbf{V}_h^c$ and simple Jacobi smoother on $\mathbf{V}_h^{DG}$. Let Δ_h and Δ_h' denote the set of elements in the* ***curl****-dominated regime and* **reaction***-dominated region, respectively:*

$$\Delta_h := \{T \in \mathscr{T}_h : h_T^2 \beta_T < \alpha_T(\nu)\}, \quad \Delta_h' := \{T \in \mathscr{T}_h : h_T^2 \beta_T \geq \alpha_T(\nu)\}.$$

Then, the spectral condition number of the resulting preconditioned system satisfies

$$\kappa(\mathbb{B}\mathbb{A}) \lesssim \max\{1, \Theta(\nu,\beta)\},$$

$$\textit{with } \Theta(\nu,\beta) := \min\left\{ \max_{T\in\mathscr{T}_h} \frac{h_T^2\beta_T}{\nu_T}, \max_{\substack{T,T'\in\mathscr{T}_h \\ \partial T\cap\partial T'\neq\emptyset}} \frac{\beta_T}{\beta_{T'}}, \max_{\substack{T\in\Delta_h, T'\in\Delta_h' \\ \partial T\cap\partial T'\neq\emptyset}} \frac{\alpha_T(\nu)}{\alpha_{T'}(\nu)} \right\}.$$

The proof can be found in [Ayuso de Dios et al.(2017), Pagliantini(2016)] as well as the analysis of the *Coarser AS Preconditioner* on simplicial meshes. The analysis of a *Coarser AS Preconditioner* on hexahedral meshes is still an open problem.

4 Numerical Results

In the following numerical simulations we will restrict to the two dimensional problem (1) on a square. We set the constant entering in the penalty parameter s_f in (4) to $c = 10$. The tolerance for the CG and PCG is set to 10^{-7}. In the tables we always report the number of iterations required for convergence. We refer to the AS preconditioners by $\mathbf{V}_h^{DG} - \mathscr{W}$, or more precisely by the local spaces $\mathscr{M}(T)$ in the construction of each $\mathbf{V}_h^{DG}$ and $\mathscr{W}$. Since the experiments are in 2D we use the rotated Nédélec elements of the first family $\mathscr{N}^1(T) = \mathscr{R}\mathscr{T}_0$; the rotated version of the space $\mathscr{S}_1 := \mathscr{R}\mathscr{T}_0 + \{\mathbf{curl}(x^2y), \mathbf{curl}(xy^2), \mathbf{curl}(x^2), \mathbf{curl}(y^2)\}$, and the 2D full polynomials space $\mathbb{Q}_1(T)^2$. For the *Natural AS Preconditioner* a simple Jacobi smoother is always used. For the *Coarser or Economical AS Preconditioner* we will specify the smoother used at each time.

Test Cases with Continuous Coefficients. We consider first the constant coefficient case $\beta = \nu = 1$. As shown in Table 1, the natural AS preconditioner is indeed optimal in all the cases, as predicted by Theorem 1. In contrast, the coarser AS preconditioner performs optimally for $\mathscr{S}_1 - \mathscr{R}\mathscr{T}_0$ only if an overlapping smoother is included. However, the coarser AS preconditioner $\mathbb{Q}_1 - \mathscr{R}\mathscr{T}_0$ is not efficacious regardless the smoother involved in the construction.

To get some insight on the failure of the coarser AS preconditioner for $\mathbb{Q}_1$, we explore the spectral approximation of the considered DG methods to (1) on $\Omega = [0,\pi]^2$ with $\nu = 1$ and $\beta = 0$. The exact eigenvalues are given by $n^2 + m^2$ for n and m positive integers. In Figure 2 is given the lower part of the spectrum using a DG discretization based on the three possible choices of local spaces $\mathscr{M}(T)$. As it can be observed in in Figure 2, the DG discretization based on the full polynomial space $(\mathbb{Q}_1)^2$, is not spectrally correct. Therefore, a preconditioner built on an auxiliary

$\sharp\mathscr{T}_h$	16×16	32×32	64×64	128×128	256×256
$\mathscr{RT}_0$ Unpreconditioned	128	204	376	753	1504
$(Q_1)^2$ Unpreconditioned	410	815	1454	2796	4554
$\mathscr{S}_1$ Unpreconditioned	543	1083	2031	4056	7316
$\mathscr{RT}_0$-$\mathscr{RT}_0$ Jacobi	9	9	9	9	9
Q_1-Q_1 Jacobi	22	21	20	19	19
Q_1-$\mathscr{RT}_0$: Jacobi | overlapping	259 | 61	471 | 113	844 | 202	1622 | 337	2936 | 618
$\mathscr{S}_1$-$\mathscr{RT}_0$: Jacobi | overlapping	88 | 18	72 | 19	49 | 20	34 | 20	36 | 19

Table 1: Number of iterations for test case with constant coefficients.

space where the $\mathrm{H}_0(\mathbf{curl},\Omega)$-conforming discretization is spectrally correct (e.g. Nédélec elements of the first family) is not effective.

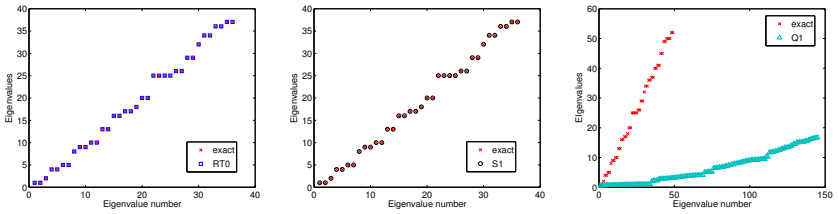


Fig. 2: Lower part of the spectrum for different DG discretizations: rotated Nédélec elements of the first family $\mathscr{RT}_0$ (left), rotated $\mathscr{S}_1$ (center), and the full polynomial space $(Q_1)^2$ (right).

Test Case with Discontinuous Coefficients. We consider now the more challenging case of β and ν both discontinuous following a checkerboard distribution according to the partition $\Omega_1 := [0,0.5]^2 \cup [0.5,1]^2 \subset \Omega = [0,1]^2$. We define

$$\nu(\mathbf{x}) = \begin{cases} 10^2 & \text{if } \mathbf{x} \in \Omega_1\,, \\ 1 & \text{otherwise}\,, \end{cases} \qquad \text{and} \qquad \beta(\mathbf{x}) = \begin{cases} 10^{-3} & \text{if } \mathbf{x} \in \Omega_1\,, \\ 10 & \text{otherwise}\,. \end{cases}$$

In Table 2 we report the iteration counts of the different preconditioners and in Figure 3 are given graphically the estimated condition numbers of the preconditioned systems. As it can be observed in Figure 3 and Table 2, the natural AS preconditioner performs optimally in the presence of discontinuous coefficients, as predicted by Theorem 1. The coarser AS preconditioner $\mathscr{S}_1$-$\mathscr{RT}_0$ is also efficacious in this case, when using an overlapping relaxation. As regards the $(\mathbb{Q}_1)^2$ DG discretization, the coarser AS preconditioner is totally ineffective.

$\sharp\mathscr{T}_h$	16×16	32×32	64×64	128×128	256×256
$\mathscr{RT}_0$-$\mathscr{RT}_0$ Jacobi	11	10	10	10	10
Q_1-Q_1 Jacobi	23	22	21	21	20
$\mathscr{S}_1$-$\mathscr{RT}_0$: overlapping	24	24	24	25	24
Q_1-$\mathscr{RT}_0$: overlapping	69	129	248	425	–

Table 2: Number of iterations for test case with discontinuous coefficients.

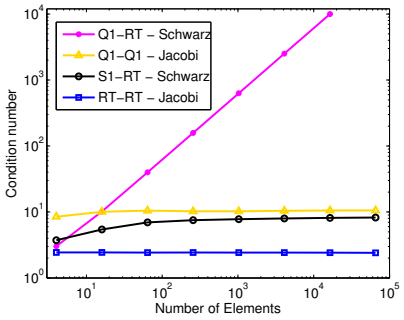


Fig. 3: Test case with discontinuous coefficients. Condition number vs. number of elements: $\mathscr{S}_1$ DG discretization with ASM based on rotated $\mathscr{RT}_0$ elements with overlapping additive Schwarz smoother (black); DG discretization with rotated $\mathscr{RT}_0$ discontinuous elements and rotated $\mathscr{RT}_0$ as auxiliary space with pointwise Jacobi smoother (blue); discontinuous bilinear Lagrangian elements with $\mathrm{H}(\mathbf{curl},\Omega)$-conforming full polynomial auxiliary space and Jacobi smoother (orange).

References

[Arnold and Awanou(2014)] D. N. Arnold and G. Awanou. Finite element differential forms on cubical meshes. *Math. Comp.*, 83(288):1551–1570, 2014.

[Ayuso de Dios et al.(2017)] B. Ayuso de Dios, R. Hiptmair, and C. Pagliantini. Auxiliary space preconditioners for SIP-DG discretizations of H(curl)-elliptic problems with discontinuous coefficients. *IMA J. Numer. Anal.*, 37(2):646–686, 2017.

[Dohrmann and Widlund(2016)] C. R. Dohrmann and O. B. Widlund. A BDDC algorithm with deluxe scaling for three-dimensional H(curl) problems. *Comm. Pure Appl. Math.*, 69(4): 745–770, 2016.

[Houston et al.(2005)] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau. Interior penalty method for the indefinite time-harmonic Maxwell equations. *Numer. Math.*, 100(3):485–518, 2005.

[Nédélec(1980)] J.-C. Nédélec. Mixed finite elements in $\mathbf{R}^3$. *Numer. Math.*, 35(3):315–341, 1980. .

[Nepomnyaschikh(1991)] S. V. Nepomnyaschikh. Mesh theorems on traces, normalizations of function traces and their inversion.*Soviet J. Numer. Anal. Math. Modelling*, 6(3):223–242, 1991.

[Hiptmair and Xu(2007)] R. Hiptmair, and J. Xu Nodal auxiliary space preconditioning in $\mathbf{H}(\mathbf{curl})$ and $\mathbf{H}(\mathrm{div})$ spaces. *SIAM J. Numer. Anal.*, **45**, 2483–2509 , 2007.

[Oswald(1996)] P. Oswald. Preconditioners for nonconforming discretizations. *Math. Comp.*, 65 (215):923–941, 1996.

[Pagliantini(2016)] C. Pagliantini. Computational magnetohydrodynamics with discrete differential forms. PhD dissertation No. 23781, Seminar for Applied Mathematics, ETH Zürich, 2016.

[Toselli(2006)] A. Toselli. Dual-primal FETI algorithms for edge finite-element approximations in 3D, IMA J. Numer. Anal., 26(1):96–130, 2006.

[Xu(1996)] J. Xu. The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids. *Computing*, 56(3):215–235, 1996.

Is minimising the convergence rate a good choice for efficient Optimized Schwarz preconditioning in heterogeneous coupling? The Stokes-Darcy case

Marco Discacciati and Luca Gerardo-Giorda

Abstract Optimized Schwarz Methods (OSM) are domain decomposition techniques based on Robin-type interface condition that have become increasingly popular in the last two decades. Ensuring convergence also on non-overlapping decompositions, OSM are naturally advocated for the heterogeneous coupling of multiphysics problems. Classical approaches optimize the coefficients in the Robin condition by minimizing the effective convergence rate of the resulting iterative algorithm. However, when OSM are used as preconditioners for Krylov solvers of the resulting interface problem, such parameter optimization does not necessarily guarantee the fastest convergence. This drawback is already known for homogeneous decomposition, but in the case of heterogeneous decomposition, the poor performance of the classical optimization approach becomes utterly evident. In this paper, we highlight this drawback for the Stokes/Darcy problem and propose a more effective optimization procedure.

1 Problem settings

The Stokes-Darcy problem, a classical model for the filtration of an incompressible fluid in a porous media [2], is a good example of a multi-physics problem where two different boundary value problems are coupled into a global heterogeneous one.

The problem is defined on a bounded domain $\Omega \subset \mathbb{R}^{\mathrm{D}}$ ($\mathrm{D} = 2,3$) formed by two non overlapping subregions: the fluid domain Ω_f and the porous medium Ω_p separated by an interface Γ. If the fluid is incompressible with constant viscosity and density, and low Reynolds' number, it can be described by the Stokes equations

Marco Discacciati
Department of Mathematical Sciences, Loughborough University, LE11 3TU, Loughborough UK, e-mail: m.discacciati@lboro.ac.uk

Luca Gerardo-Giorda
BCAM - Basque Center for Applied Mathematics, Bilbao, Spain, e-mail: lgerardo@bcamath.org

in Ω_f and by Darcy's law in Ω_p. The physics of the problem naturally drives the decomposition of the domain and, at the same time, imposes interface conditions across Γ to describe filtration phenomena. The coupled problem reads as follows: Find the fluid velocity $\mathbf{u}_f$ and pressure p_f, and the pressure p_p such that

$$\begin{aligned}
-\nabla\cdot\sigma(\mathbf{u}_f,p_f) &= \mathbf{f}_f \quad \text{in } \Omega_f && \text{Stokes equations}\\
\nabla\cdot\mathbf{u}_f &= 0 \quad \text{in } \Omega_f &&\\
-\nabla\cdot(\eta_p\nabla p_p) &= g_p \quad \text{in } \Omega_p && \text{Darcy's equation}\\
-(\eta_p\nabla p_p)\cdot\mathbf{n} &= \mathbf{u}_f\cdot\mathbf{n} \quad \text{on } \Gamma && \text{continuity of the normal velocity}\\
-\mathbf{n}\cdot\sigma(\mathbf{u}_f,p_f)\cdot\mathbf{n} &= p_p \quad \text{on } \Gamma && \text{continuity of the normal stresses}\\
-\tau\cdot\sigma(\mathbf{u}_f,p_f)\cdot\mathbf{n} &= \xi\,\mathbf{u}_f\cdot\tau \quad \text{on } \Gamma && \text{BJS condition on the tangential stresses}
\end{aligned} \tag{1}$$

where $\sigma(\mathbf{u}_f,p_f)=\mu_f(\nabla\mathbf{u}_f+(\nabla\mathbf{u}_f)^T-p_fI$ is the Cauchy stress tensor, while $\mathbf{f}_f$ and g_p are given external forces. The Beaver-Joseph-Saffman (BJS, [1]) condition does not play any role in the coupling of the local problems. Thus, coupling on Γ can be obtained by linear combination of the first two conditions:

$$\begin{aligned}
-\mathbf{n}\cdot\sigma(\mathbf{u}_f,p_f)\cdot\mathbf{n}-\alpha_f\,\mathbf{u}_f\cdot\mathbf{n} &= p_p+\alpha_f\,(\eta_p\nabla p_p)\cdot\mathbf{n}\\
p_p-\alpha_p\,(\eta_p\nabla p_p)\cdot\mathbf{n} &= -\mathbf{n}\cdot\sigma(\mathbf{u}_f,p_f)\cdot\mathbf{n}+\alpha_p\,\mathbf{u}_f\cdot\mathbf{n}
\end{aligned} \tag{2}$$

Using the interface conditions (2) a Robin-Robin method can be formulated. Such method requires solving iteratively the Stokes problem in Ω_f with boundary condition $(2)_1$ on Γ and Darcy's equation in Ω_p with boundary condition $(2)_2$ on Γ. More details can be found in [3].

2 Optimization of the Robin parameters α_p and α_f

Classical approaches in the Optimized Schwarz literature derive, through Fourier analysis, the convergence rate $\rho(\alpha_f,\alpha_p,k)$ of the iterative algorithm as a function of the parameters α_f, α_p and of the frequency k, and they aim at optimizing α_f and α_p by minimization of $\rho(\alpha_f,\alpha_p,k)$ over all the relevant frequencies of the problem. This amounts to solve the min-max problem

$$\min_{\alpha_f,\alpha_p\in\mathbb{R}^+}\ \max_{k\in[k_{min},k_{max}]}\ \rho(\alpha_f,\alpha_p,k), \tag{3}$$

where k_{min} and k_{max} are the minimal frequency relevant to the problem and the maximal frequency supported by the numerical grid (of the order of π/h).

However, when the OSM is used as a preconditioner for a Krylov method to solve the interface problem, such a choice does not necessarily guarantee the fastest convergence. Minimising the effective convergence rate ($\rho_{eff}(\alpha_f,\alpha_p)=\max_k\rho(\alpha_f,\alpha_p,k)$) does not make the convergence rate automatically small for all frequencies, and the Krylov type solver can then suffer from slow convergence. Such an issue can be particularly relevant in the presence of heterogeneous cou-

pling. In the rest of the section, we first introduce the exact interface conditions, then present three different approaches to optimize the interface parameters. The first one is based on a classical equioscillation principle, the second one exploits the peculiar characteristics of the Stokes/Darcy problem, while the third one aims to globally minimize the convergence rate for all frequencies.

2.1 Convergence rate and exact interface conditions

The convergence rate of the Robin-Robin algorithm does not depend on the iteration and, for positive parameters $\alpha_p, \alpha_f > 0$, is given by [3]

$$\rho(\alpha_f, \alpha_p, k) = \left| \frac{2\mu_f k - \alpha_p}{2\mu_f k + \alpha_f} \right| \cdot \left| \frac{1 - \alpha_f \eta_p k}{1 + \alpha_p \eta_p k} \right|. \tag{4}$$

(As shown in [3], by symmetry we can restrict to the case $k > 0$.)

The optimal parameters force the reduction factor $\rho(\alpha_f, \alpha_p, k)$ to be identically equal to zero for all k, so that convergence is attained in a number of iterations equal to the number of subdomains. They can be easily derived from (4) as

$$\alpha_p^{exact}(k) = 2\mu_f k \qquad \alpha_f^{exact}(k) = \frac{1}{\eta_p k}. \tag{5}$$

Their direct use is unfortunately not viable: both depend on the frequency k, and their back transforms in the physical space are either introducing an imaginary coefficient which multiplies a first order tangential derivative ($\alpha_p^{exact}(k)$) or result in a nonlocal operator ($\alpha_f^{exact}(k)$). The use of approximations based on low-order Taylor expansions of the optimal values (5) (around $k = k_{min}$ for α_p and $k = k_{max}$ for α_f) would not help either, as they would suffer from the same drawbacks (see [3]).

2.2 The equioscillation approach

The convergence rate (4) is continuous, has two positive roots, $k_1 = (\alpha_f \eta_p)^{-1}$ and $k_2 = \alpha_p/(2\mu_f)$, and a maximum between k_1 and k_2, given by (setting $\delta = 2\mu_f \eta_p$)

$$k_* = \frac{2\delta(\alpha_p - \alpha_f) + \sqrt{4\delta^2(\alpha_p - \alpha_f)^2 + 4\delta(2\mu_s + \alpha_f \alpha_p \eta_p)^2}}{2\delta(2\mu_s + \alpha_f \alpha_p \eta_p)}. \tag{6}$$

The natural approach to solve the min-max problem (3) would resort to an equioscillation principle, where one seeks for α_f^{eq} and α_p^{eq} such that

$$\rho(\alpha_f^{eq}, \alpha_p^{eq}, k_{min}) = \rho(\alpha_f^{eq}, \alpha_p^{eq}, k_*) = \rho(\alpha_f^{eq}, \alpha_p^{eq}, k_{max}). \tag{7}$$

This approach ensures that all other frequencies exhibit a smaller convergence rate.

Proposition 1. *The solution to problem* (7) *is given by the two pairs of optimal coefficients* $(\alpha_{f,i}^{eq}, \alpha_{p,i}^{eq})$, $i = 1,2$*:*

$$\alpha_{f,i}^{eq} = \frac{1}{2}\left(X_i + \sqrt{X_i^2 + 4Y_i}\right), \quad \alpha_{p,i}^{eq} = \frac{1}{2}\left(-X_i + \sqrt{X_i^2 + 4Y_i}\right), \qquad i = 1,2, \quad (8)$$

with $Y_i \in \mathbb{R}^+$ *and* $X_i \in \mathbb{R}$ *defined as follows:*

$$Y_i = \frac{2\mu_f}{\eta_p}\left(\frac{b}{a} - 1 + (-1)^{i+1}\sqrt{\left(\frac{b}{a} - 1\right)^2 - 1}\right) \qquad i = 1,2, \quad (9)$$

$$X_i = \frac{1 - \delta k_{min} k_{max}}{\eta_p (k_{min} + k_{max})}\left(\frac{\eta_p}{2\mu_f} Y_i + 1\right) \qquad i = 1,2, \quad (10)$$

where $a > 0$ *and* $b > 0$ *are the positive quantities*

$$\begin{aligned} a = \frac{1 + \delta k_{max}^2}{(k_{min} + k_{max})^2}\big(&k_{min}(k_* + k_{max}) + k_*(k_{max} - k_*) \\ &+ \delta k_{min} k_{max}(k_*(k_{min} + k_*) + k_{max}(k_* - k_{min}))\big), \end{aligned} \quad (11)$$

$$b = (1 + \delta k_{max}^2)(1 + \delta k_*^2), \quad (12)$$

and $k_* > 0$ *becomes*

$$k_* = \frac{\delta k_{min} k_{max} - 1 + \sqrt{(\delta k_{min} k_{max} - 1)^2 + \delta (k_{min} + k_{max})^2}}{\delta (k_{min} + k_{max})}. \quad (13)$$

Proof. We consider the first condition of equioscillation in (7): $\rho(\alpha_f, \alpha_p, k_{min}) = \rho(\alpha_f, \alpha_p, k_{max})$. With the help of some algebra, we obtain

$$\alpha_p - \alpha_f = (\delta k_{min} k_{max} - 1)(\eta_p \alpha_f \alpha_p + 2\mu_f)(\delta (k_{min} + k_{max}))^{-1}. \quad (14)$$

Substituting (14) into (6) we obtain the expression (13) for k_* which is now independent of α_f and α_p. It can be easily verified that the obtained value of k_* satisfies $k_{min} < k_* < k_{max}$ so that we can proceed imposing the second condition of equioscillation in (7): $\rho(\alpha_f, \alpha_p, k_{max}) = \rho(\alpha_f, \alpha_p, k_*)$, that is:

$$\begin{aligned} &-\delta(k_*^2 + k_{max}^2)(\alpha_f - \alpha_p)^2 + 2\eta_p k_* k_{max}(\alpha_f \alpha_p)^2 \\ &\quad + \eta_p (k_* + k_{max})(1 - \delta k_* k_{max})(\alpha_f - \alpha_p)\alpha_f \alpha_p \\ &\quad + 2\mu_f (k_* + k_{max})(1 - \delta k_* k_{max})(\alpha_f - \alpha_p) \\ &\quad - 2(1 + \delta^2 k_*^2 k_{max}^2 + \delta (k_{max} - k_*)^2)\alpha_f \alpha_p + 8\mu_f^2 k_* k_{max} = 0. \end{aligned} \quad (15)$$

We introduce now the change of variables: $X = \alpha_f - \alpha_p$ and $Y = \alpha_f \alpha_p$. We substitute the expression of X from (14) into (15) to get

$$Y^2\left(a\frac{\eta_p}{2\mu_f}\right)+2Y(a-b)+a\frac{2\mu_f}{\eta_p}=0 \tag{16}$$

where a and b are as in (11) and (12), respectively. Since $k_{min}<k_*<k_{max}$, $a>0$ and we can rewrite (16) as

$$Y^2\frac{\eta_p}{2\mu_f}-2Y\left(\frac{b}{a}-1\right)+\frac{2\mu_f}{\eta_p}=0 \tag{17}$$

whose roots are (9). By a simple algebraic manipulation, it can be verified that $b-2a>0$ which also implies that $b-a>0$, so that the discriminant of (17) is positive and both its roots are positive as well: $Y_i>0$, $i=1,2$. Finally, (10) follows from (14) and (8) is obtained reversing the change of variables. □

2.3 Exploiting the problem characteristics

From (5), we observe that the product of the optimal values $\alpha_f^{exact}(k)$ and $\alpha_p^{exact}(k)$ is constant and equals $2\mu_f/\eta_p$. We exploit such peculiarity of the problem (not occurring in homogeneous decomposition, see e.g. [5]), and restrict our search for optimized parameters to the curve

$$\alpha_f\,\alpha_p=2\mu_f/\eta_p. \tag{18}$$

Notice that such curve is the subset of the (α_f,α_p) upper-quadrant where the zeros k_1 and k_2 of the convergence rate ρ coincide.

Proposition 2 ([3]). *The solution of the min-max problem*

$$\min_{\alpha_f\alpha_p=\frac{2\mu_f}{\eta_p}}\ \max_{k\in[k_{min},k_{max}]}\rho(\alpha_f,\alpha_p,k)$$

is given by the pair

$$\begin{aligned}\alpha_f^* &= \frac{1-2\mu_f\eta_p k_{min}k_{max}}{\eta_p(k_{min}+k_{max})}+\sqrt{\left(\frac{1-2\mu_f\eta_p k_{min}k_{max}}{\eta_p(k_{min}+k_{max})}\right)^2+\frac{2\mu_f}{\eta_p}}\\ \alpha_p^* &= -\frac{1-2\mu_f\eta_p k_{min}k_{max}}{\eta_p(k_{min}+k_{max})}+\sqrt{\left(\frac{1-2\mu_f\eta_p k_{min}k_{max}}{\eta_p(k_{min}+k_{max})}\right)^2+\frac{2\mu_f}{\eta_p}}\end{aligned} \tag{19}$$

Moreover, $\rho(\alpha_f^*,\alpha_p^*,k)<1$ *for all* $k\in[k_{min},k_{max}]$.

2.4 Minimisation of the mean convergence rate

The reduction factor along (18) is given by

$$\rho(\alpha_f,k) = \frac{2\mu_f}{\eta_p}\left(\frac{\eta_p\alpha_f k - 1}{2\mu_f k + \alpha_f}\right)^2. \tag{20}$$

To further exploit the characteristics of the problem, we consider the set

$$\mathscr{A}_f = \{\alpha_f > 0 : \rho(\alpha_f,k) \leq 1 \quad \forall k \in [k_{min},k_{max}]\}.$$

Notice that the convergence of the Robin-Robin method in the iterative form would be ensured only if the inequality in the definition of $\mathscr{A}_f$ is strict. From [3] we know that the convergence rate can equal 1 in at most one frequency, either in k_{min} or in k_{max}. When using the OSM as a preconditioner for a Krylov method, the latter can handle isolated problems in the spectrum (see, e.g., [4, 6, 7]).

In order to improve the overall convergence for a Krylov method, we minimize, on the set $\mathscr{A}_f$, the expected value of $\rho(\alpha_f,k)$ in the interval $[k_{min},k_{max}]$:

$$E(\alpha_f) := \mathbb{E}[\rho(\alpha_f,k)] = \frac{1}{k_{max}-k_{min}}\int_{k_{min}}^{k_{max}} \rho(\alpha_f,k)\,dk.$$

Owing to (20), $E(\alpha_f)$ can be explicitly computed: it is positive in $\alpha_f = 0$, and has a minimum in the point $\widehat{\alpha}_f$ after which it is always increasing (see [3]). As a consequence, the minimum α_f^{opt} of $E(\alpha_f)$ is attained in $\widehat{\alpha}_f$ if the latter belongs to $\mathscr{A}_f$, or in one extremum of $\mathscr{A}_f$ otherwise, namely:

$$\alpha_f^{opt} = \begin{cases} \min\limits_{\alpha_f\in\mathscr{A}_f} \alpha_f & \text{if } \widehat{\alpha}_f < \min\limits_{\alpha_f\in\mathscr{A}_f} \alpha_f \\ \widehat{\alpha}_f & \text{if } \widehat{\alpha}_f \in \mathscr{A}_f \\ \max\limits_{\alpha_f\in\mathscr{A}_f} \alpha_f & \text{if } \widehat{\alpha}_f > \max\limits_{\alpha_f\in\mathscr{A}_f} \alpha_f. \end{cases} \tag{21}$$

3 Numerical results

We compare here the three approaches (8), (19) and (21) considering a test with analytic solution: $\mathbf{u}_f = (\sqrt{\mu_f\eta_p}, \alpha_{BJ}x)$, $p_f = 2\mu_f(x+y-1) + (3\eta_p)^{-1}$, $p_p = (-\alpha_{BJ}x(y-1) + y^3/3 - y^2 + y)/\eta_p + 2\mu_f x$. We set $\Omega_f = (0,1)\times(1,2)$, $\Omega_p = (0,1)\times(0,1)$ and interface $\Gamma = (0,1)\times\{1\}$. The computational grids are uniform, structured, made of triangles with $h = 2^{-(s+2)}$, $s \geq 0$; $\mathbb{P}_2$-$\mathbb{P}_1$ finite elements are used for Stokes and $\mathbb{P}_2$ elements for Darcy's law; η_p is constant, $\alpha_{BJ} = 1$, $k_{min} = \pi$, $k_{max} = \pi/h$. The interface system associated to the OSM [3] is solved by GMRES with tolerance 1e-9. In Table 1 we report the parameters obtained for various coefficients μ_f and η_p. Figure 1 shows the convergence rates versus k for the three possible choices of α_f and α_p and two pairs of values (μ_f,η_p). The number of iterations for α_f and α_p at fixed h is computed for two pairs of values (μ_f,η_p) and is

Table 1 Parameters obtained in (8), (19) and (21) for different values of μ_f, η_p and $h = 2^{-5}$.

μ_f	η_p	α_f^{eq}	α_p^{eq}	α_f^{*}	α_p^{*}	α_f^{opt}	α_p^{opt}
1	1	0.27	36.93	0.16	12.33	0.036	56.04
1	1e-2	23.00	68.59	9.91	20.17	5.44	36.75
1	1e-4	852.50	157.10	258.19	77.46	217.34	92.01
1e-1	1	0.26	4.19	0.15	1.35	0.03	5.48
1e-1	1e-2	15.71	12.01	4.84	4.13	3.37	5.93
1e-1	1e-4	613.00	17.02	201.61	9.92	195.90	10.21

shown in Figure 2. The parameters devised in (8) feature both the smallest convergence rate and the worst preconditioning performance in terms of iteration counts. Notice also that α_f^{opt} in (21), minimizing the mean convergence rate, always ensures the best performance in terms of iteration counts. Figure 3 displays the number of iterations versus h for different combinations of μ_f and η_p: α_f^{opt} consistently exhibits the best convergence properties, in particular when the ratio μ_f/η_p increases.

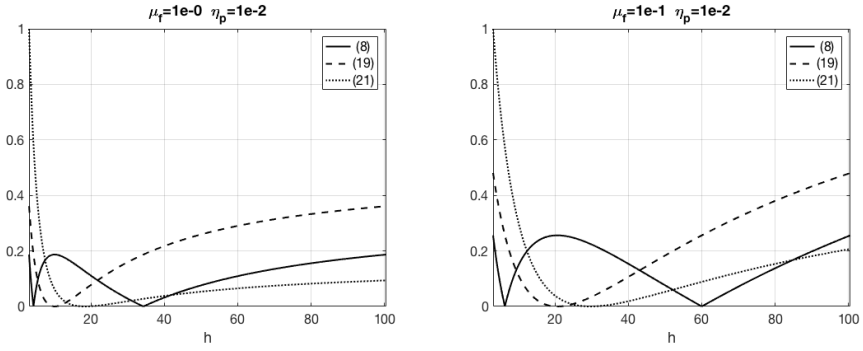


Fig. 1 Convergence rates as a function of k for the parameters (8) (solid line), (19) (dashed line), and (21) (dotted line). Left: $\mu_f = 1$, $\eta_p =$ 1e-2. Right: $\mu_f =$ 1e-1, $\eta_p =$ 1e-2. $h = 2^{-5}$.

4 Conclusions

Using the Stokes/Darcy coupling as a testbed for heterogeneous problems, we show that minimizing the convergence rate of the corresponding iterative algorithm leads to poor convergence when an Optimized Schwarz Method is used as preconditioner for a Krylov method applied to the interface equation. On the other hand, taking advantage of the problem characteristics and minimizing the mean of the convergence rate provides effective preconditioning.

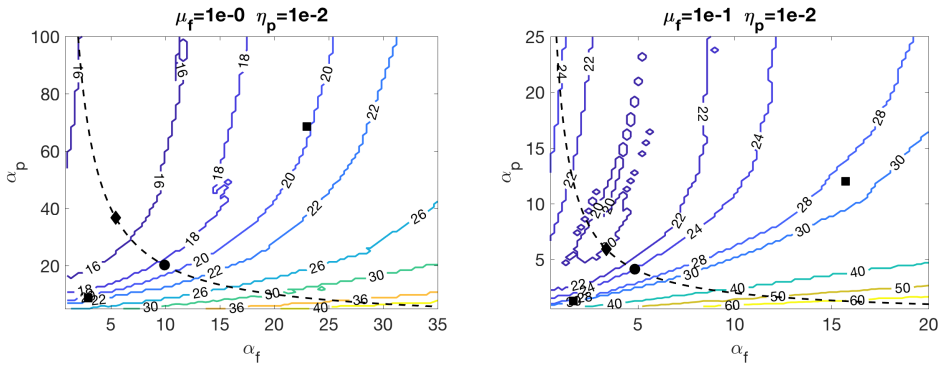


Fig. 2 Number of iterations for $h = 2^{-5}$ and parameters α_f and α_p as in (8) (squares), (19) (circle) and (21) (diamond). Left: $\mu_f = 1$, $\eta_p =$ 1e-2; right: $\mu_f =$ 1e-1, $\eta_p =$ 1e-2.

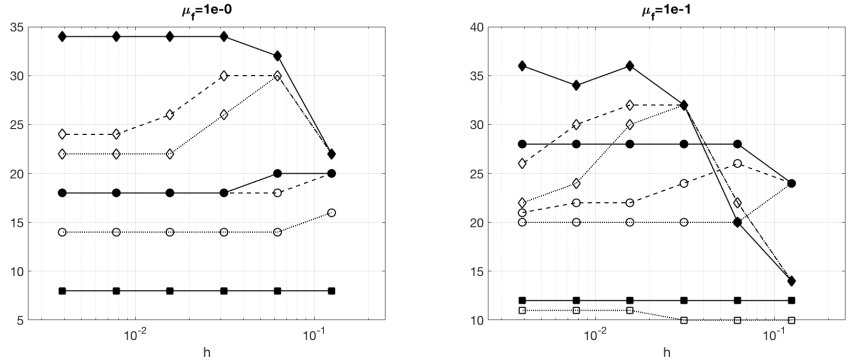


Fig. 3 Number of iterations versus h. Solid lines refer to (8), dashed lines to (19) and dotted lines (21). Squares refer to $\eta_p = 1$, circles $\eta_p =$ 1e-2, diamonds $\eta_p =$ 1e-4. Left: $\mu_f = 1$; right: $\mu_f =$ 1e-1. All values obtained for $\eta_p = 1$ and $\mu_f = 1$ coincide (left plot), while for $\eta_p = 1$ and $\mu_f =$ 1e-1 they coincide only when computed using (8) and (19) (right plot).

References

1. G. Beavers and D. Joseph (1967) Boundary conditions at a naturally permeable wall. *J. Fluid Mech.*, **30**, 197–207.
2. M. Discacciati and A. Quarteroni (2009) Navier-Stokes/Darcy coupling: modeling, analysis, and numerical approximation. *Rev. Mat. Complut.*, **22**, 315–426.
3. M. Discacciati and L. Gerardo-Giorda (2017) Optimized Schwarz Methods for the Stokes-Darcy coupling *IMA J. Num. Anal.*, (submitted)
4. V. Dolean, M.J. Gander, and L. Gerardo-Giorda (2009) Optimized Schwarz methods for Maxwell's equations. *SIAM J. Sci. Comput.*, **31**, 2193–2213.
5. M.J. Gander. Optimized Schwarz methods. *SIAM J. Num. Anal.*, 44(2), pp. 699–731, 2006.
6. M.J. Gander, F. Magoulès, and F. Nataf (2002) Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, **21**, 38–60.
7. L. Gerardo-Giorda and M. Perego (2013) Optimized Schwarz methods for the Bidomain system in electrocardiology. *M2AN*, **75**, 583–608.

Preconditioned space-time boundary element methods for the one-dimensional heat equation

Stefan Dohr and Olaf Steinbach

1 Introduction

Space–time discretization methods, see, e.g., [8], became very popular in recent years, due to their ability to drive adaptivity in space and time simultaneously, and to use parallel iterative solution strategies for time–dependent problems. But the solution of the global linear system requires the use of some efficient preconditioner.

In this note we describe a space–time boundary element discretization of the heat equation and an efficient and robust preconditioning strategy which is based on the use of boundary integral operators of opposite orders, but which requires a suitable stability condition for the boundary element spaces used for the discretization. We demonstrate the method for the simple spatially one-dimensional case. However, the presented results, particularly the stability analysis of the boundary element spaces, can be used to extend the method to the two- and three-dimensional problem [2].

Let $\Omega = (a,b) \subset \mathbb{R}$, $\Gamma := \partial\Omega = \{a,b\}$ and $T > 0$. As a model problem we consider the Dirichlet boundary value problem for the heat equation,

$$\alpha\partial_t u - \Delta_x u = 0 \text{ in } Q := \Omega \times (0,T),\ u = g \text{ on } \Sigma := \Gamma \times (0,T),\ u = u_0 \text{ in } \Omega \quad (1)$$

with the heat capacity constant $\alpha > 0$, the given initial datum u_0, and the boundary datum g. The solution of (1) can be expressed by using the representation formula for the heat equation [1], i.e. for $(x,t) \in Q$ we have

$$\begin{aligned} u(x,t) = &\int_\Omega U^\star(x-y,t)u_0(y)\mathrm{d}y + \frac{1}{\alpha}\int_\Sigma U^\star(x-y,t-s)\frac{\partial}{\partial n_y}u(y,s)\mathrm{d}s_y\mathrm{d}s \\ &- \frac{1}{\alpha}\int_\Sigma \frac{\partial}{\partial n_y}U^\star(x-y,t-s)g(y,s)\mathrm{d}s_y\mathrm{d}s, \end{aligned} \quad (2)$$

Stefan Dohr, Olaf Steinbach
Institut für Angewandte Mathematik, TU Graz, Steyrergasse 30, 8010 Graz
e-mail: stefan.dohr@tugraz.at, o.steinbach@tugraz.at

where $U^\star$ denotes the fundamental solution of the heat equation given by

$$U^\star(x-y,t-s) = \begin{cases} \left(\frac{\alpha}{4\pi(t-s)}\right)^{1/2} \exp\left(\frac{-\alpha|x-y|^2}{4(t-s)}\right), & s<t, \\ 0, & \text{else.} \end{cases}$$

Hence it suffices to determine the yet unknown Cauchy datum $\partial_n u_{|\Sigma}$ to compute the solution of (1). It is well known [5] that for $u_0 \in L^2(\Omega)$ and $g \in H^{1/2,1/4}(\Sigma)$ the problem (1) has a unique solution $u \in H^{1,1/2}(Q,\alpha\partial_t - \Delta_x)$ with the anisotropic Sobolev space

$$H^{1,1/2}(Q,\alpha\partial_t - \Delta_x) := \left\{ u \in H^{1,1/2}(Q) : (\alpha\partial_t - \Delta_x)u \in L^2(Q) \right\}.$$

In the one-dimensional case the spatial component of the space–time boundary Σ collapses to the points $\{a,b\}$ and therefore we can identify the anisotropic Sobolev spaces $H^{r,s}(\Sigma)$ with $H^s(\Sigma)$. The unknown density $w := \partial_n u_{|\Sigma} \in H^{-1/4}(\Sigma)$ can be found by applying the interior Dirichlet trace operator $\gamma_0^{\rm int} : H^{1,1/2}(Q) \to H^{1/4}(\Sigma)$ to the representation formula (2),

$$g(x,t) = (M_0 u_0)(x,t) + (Vw)(x,t) + ((\frac{1}{2}I - K)g)(x,t) \quad \text{for } (x,t) \in \Sigma.$$

The initial potential $M_0 : L^2(\Omega) \to H^{1/4}(\Sigma)$, the single layer boundary integral operator $V : H^{-1/4}(\Sigma) \to H^{1/4}(\Sigma)$, and the double layer boundary integral operator $\frac{1}{2}I - K : H^{1/4}(\Sigma) \to H^{1/4}(\Sigma)$ are obtained by composition of the potentials in (2) with the Dirichlet trace operator $\gamma_0^{\rm int}$, see, e.g., [1, 6]. In fact, we have to solve the variational formulation to find $w \in H^{-1/4}(\Sigma)$ such that

$$\langle Vw,\tau\rangle_\Sigma = \langle(\frac{1}{2}I + K)g,\tau\rangle_\Sigma - \langle M_0 u_0,\tau\rangle_\Sigma \quad \text{for all } \tau \in H^{-1/4}(\Sigma), \tag{3}$$

where $\langle\cdot,\cdot\rangle_\Sigma$ denotes the duality pairing on $H^{1/4}(\Sigma) \times H^{-1/4}(\Sigma)$. The single layer boundary integral operator V is bounded and elliptic, i.e. there exists a constant $c_1^V > 0$ such that

$$\langle Vw,w\rangle_\Sigma \geq c_1^V \|w\|^2_{H^{-1/4}(\Sigma)} \quad \text{for all } w \in H^{-1/4}(\Sigma).$$

Thus, the variational formulation (3) is uniquely solvable. When applying the Neumann trace operator $\gamma_1^{\rm int} : H^{1,1/2}(Q,\alpha\partial_t - \Delta_x) \to H^{-1/4}(\Sigma)$ to the representation formula (2) we obtain the second boundary integral equation

$$w(x,t) = (M_1 u_0)(x,t) + ((\frac{1}{2}I + K')w)(x,t) + (Dg)(x,t) \quad \text{for } (x,t) \in \Sigma$$

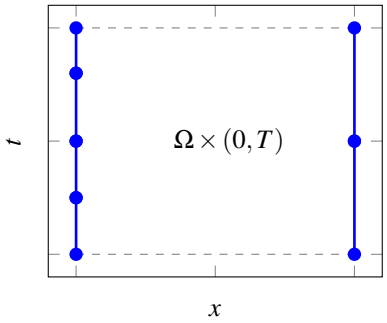


Fig. 1 Sample BE mesh. We consider an arbitrary decomposition of the space–time boundary Σ. Note that there is no time-stepping scheme involved

with the hypersingular boundary integral operator $D : H^{1/4}(\Sigma) \to H^{-1/4}(\Sigma)$, and with the adjoint double layer boundary integral operator $K' : H^{-1/4}(\Sigma) \to H^{-1/4}(\Sigma)$. Moreover, $M_1 : L^2(\Omega) \to H^{-1/4}(\Sigma)$.

2 Boundary element methods

For the Galerkin boundary element discretization of the variational formulation (3) we consider a family $\{\Sigma_N\}_{N\in\mathbb{N}}$ of arbitrary decompositions of the space–time boundary Σ into boundary elements σ_l, i.e. we have

$$\overline{\Sigma}_N = \bigcup_{\ell=1}^{N} \overline{\sigma}_\ell \,.$$

In the one-dimensional case the boundary elements σ_ℓ are line segments in temporal direction with fixed spatial coordinate $x_\ell \in \{a,b\}$ as shown in Fig. 1. Let (x_ℓ, t_{ℓ_1}) and (x_ℓ, t_{ℓ_2}) be the nodes of the boundary element σ_ℓ. The local mesh size is then given as $h_\ell := |t_{\ell_2} - t_{\ell_1}|$ while $h := \max_{\ell=1,\dots,N} h_\ell$ is the global mesh size.
For the approximation of the unknown Cauchy datum $w = \gamma_1^{\text{int}} u \in H^{-1/4}(\Sigma)$ we consider the space $S_h^0(\Sigma) := \operatorname{span}\{\varphi_\ell^0\}_{\ell=1}^N$ of piecewise constant basis functions φ_ℓ^0, which is defined with respect to the decomposition Σ_N. The Galerkin-Bubnov variational formulation of (3) is to find $w_h \in S_h^0(\Sigma)$ such that

$$\langle V w_h, \tau_h\rangle_\Sigma = \langle (\frac{1}{2} I + K) g, \tau_h\rangle_\Sigma - \langle M_0 u_0, \tau_h\rangle_\Sigma \quad \text{for all } \tau_h \in S_h^0(\Sigma)\,. \tag{4}$$

This is equivalent to the system of linear equations $V_h \mathbf{w} = \mathbf{f}$ where

$$V_h[\ell,k] = \langle V\varphi_k^0, \varphi_\ell^0\rangle_\Sigma, \quad \mathbf{f}[\ell] = \langle (\frac{1}{2} I + K) g, \varphi_\ell^0\rangle_\Sigma - \langle M_0 u_0, \varphi_\ell^0\rangle_\Sigma, \quad k,\ell = 1,\dots,N.$$

Due to the ellipticity of the single layer operator V the matrix V_h is positive definite and therefore the variational formulation (4) is uniquely solvable as well. Moreover,

when assuming $w \in H^s(\Sigma)$ for some $s \in [0,1]$, there holds the error estimate

$$\|w - w_h\|_{H^{-1/4}(\Sigma)} \leq ch^{1/4+s}|w|_{H^s(\Sigma)}.$$

Using standard arguments we also conclude the error estimate

$$\|w - w_h\|_{L^2(\Sigma)} \leq ch^s|w|_{H^s(\Sigma)}$$

which implies linear convergence of the $L^2(\Sigma)$–error of the Galerkin approximation w_h if $w \in H^1(\Sigma)$ is satisfied.

3 Preconditioning strategies

Since the boundary element discretization is done with respect to the whole space–time boundary Σ we need to have an efficient iterative solution technique. In fact, the linear system $V_h\mathbf{w} = \mathbf{f}$ with the positive definite but nonsymmetric matrix V_h can be solved by using a preconditioned GMRES method. Here we will apply a preconditioning technique based on boundary integral operators of opposite order [10], also known as operator or Calderon preconditioning [3]. Since the single layer integral operator $V : H^{-1/4}(\Sigma) \to H^{1/4}(\Sigma)$ and the hypersingular integral operator $D : H^{1/4}(\Sigma) \to H^{-1/4}(\Sigma)$ are both elliptic, the operator $DV : H^{-1/4}(\Sigma) \to H^{-1/4}(\Sigma)$ behaves like the identity. Hence we can use the Galerkin discretization of D as a preconditioner for V_h. But for the Galerkin discretization D_h of the hypersingular integral operator $D : H^{1/4}(\Sigma) \to H^{-1/4}(\Sigma)$ we need to use a conforming ansatz space $Y_h = \operatorname{span}\{\psi_i\}_{i=1}^N \subset H^{1/4}(\Sigma)$ while the discretization of the single layer integral operator V is done with respect to $S_h^0(\Sigma)$. Since the boundary element space $S_h^0(\Sigma)$ of piecewise constant basis functions φ_k^0 also satisfies $S_h^0(\Sigma) \subset H^{1/4}(\Sigma)$ we can choose $Y_h = S_h^0(\Sigma)$. The inverse hypersingular operator D^{-1} is spectrally equivalent to the single layer operator V, therefore the approximation of the preconditioning operator corresponds to a mixed approximation scheme, and hence we need to assume a discrete stability condition to be satisfied.

Theorem 1 ([3, 10]). *Assume the discrete stability condition*

$$\sup_{0 \neq v_h \in Y_h} \frac{\langle \tau_h, v_h \rangle_{L^2(\Sigma)}}{\|v_h\|_{H^{1/4}(\Sigma)}} \geq c_1^M \|\tau_h\|_{H^{-1/4}(\Sigma)} \quad \text{for all } \tau_h \in S_h^0(\Sigma). \tag{5}$$

Then there exists a constant $c_\kappa > 1$ such that

$$\kappa\left(M_h^{-1} D_h M_h^{-\top} V_h\right) \leq c_\kappa$$

where, for $k, \ell = 1, \ldots, N$,

$$V_h[\ell,k] = \langle V\varphi_k^0, \varphi_\ell^0 \rangle_\Sigma, \quad D_h[\ell,k] = \langle D\psi_k, \psi_\ell \rangle_\Sigma, \quad M_h[\ell,k] = \langle \varphi_k^0, \psi_\ell \rangle_{L^2(\Sigma)}.$$

Thus we can use $C_V^{-1} = M_h^{-1} D_h M_h^{-\top}$ as a preconditioner for V_h. Since M_h is sparse and spectrally equivalent to a diagonal matrix, the inverse M_h^{-1} can be computed efficiently. It remains to define, for given $S_h^0(\Sigma)$, a suitable boundary element space Y_h such that the stability condition (5) is satisfied. In what follows we will discuss a possible choice.

If we choose $Y_h = S_h^0(\Sigma)$ for the discretization of the hypersingular operator D, then M_h becomes diagonal and is therefore easily invertible. In order to prove the stability condition (5) we need to establish the $H^{1/4}(\Sigma)$–stability of the $L^2(\Sigma)$–projection $Q_h^0 : L^2(\Sigma) \to S_h^0(\Sigma) \subset L^2(\Sigma)$ which is defined as

$$\langle Q_h^0 v, \tau_h \rangle_{L^2(\Sigma)} = \langle v, \tau_h \rangle_{L^2(\Sigma)} \quad \text{for all } \tau_h \in S_h^0(\Sigma).$$

Following [7], and when assuming local quasi-uniformity of the boundary element mesh Σ_N we are able to establish the stability of $Q_h^0 : H^{1/4}(\Sigma) \to H^{1/4}(\Sigma)$, see [2] for a more detailed discussion: For $\ell = 1, ..., N$ we define $I(\ell)$ to be the index set of the boundary element σ_ℓ and all its adjacent elements. We assume the boundary element mesh Σ_N to be locally quasi-uniform, i.e. there exists a constant $c_L \geq 1$ such that

$$\frac{1}{c_L} \leq \frac{h_\ell}{h_k} \leq c_L \quad \text{for all } k \in I(\ell) \text{ and } \ell = 1, ..., N.$$

In this case the operator $Q_h^0 : H^{1/4}(\Sigma) \to H^{1/4}(\Sigma)$ is bounded, i.e. there exists a constant $c_S^0 > 0$ such that

$$\left\| Q_h^0 v \right\|_{H^{1/4}(\Sigma)} \leq c_S^0 \, \|v\|_{H^{1/4}(\Sigma)} \quad \text{for all } v \in H^{1/4}(\Sigma). \tag{6}$$

By using the stability estimate (6) we can conclude

$$\frac{1}{c_S^0} \, \|\tau_h\|_{H^{-1/4}(\Sigma)} \leq \sup_{0 \neq v_h \in S_h^0(\Sigma)} \frac{\langle \tau_h, v_h \rangle_{L^2(\Sigma)}}{\|v_h\|_{H^{1/4}(\Sigma)}} \quad \text{for all } \tau_h \in S_h^0(\Sigma).$$

Hence the stability condition (5) holds and we can use $C_V^{-1} = M_h^{-1} D_h M_h^{-\top}$ as a preconditioner for V_h.

4 Numerical results

For the numerical experiments we choose $\Omega = (0,1)$, $T = 1$, and we consider the model problem (1) with homogeneous Dirichlet conditions $g = 0$, and some given initial datum u_0 satisfying the compatibility conditions $u_0(0) = u_0(1) = 0$. The Galerkin boundary element discretization of the variational formulation (3) is done by piecewise constant basis functions. The resulting system of linear equations $V_h \mathbf{w} = \mathbf{f}$ is solved by using the GMRES method. As a preconditioner we use the discretization $C_V^{-1} = M_h^{-1} D_h M_h^{-\top}$ of the hypersingular operator D with piecewise constant basis functions.

Uniform refinement

The first example corresponds to the initial datum $u_0(x) = \sin 2\pi x$ and a globally uniform boundary element mesh of mesh size $h = 2^{-L}$. Table 1 shows the $L^2(\Sigma)$–error $\|w - w_h\|_{L_2(\Sigma)}$ and the estimated order of convergence (eoc), which is linear as expected. Moreover, the condition numbers of the stiffness matrix V_h and of the preconditioned matrix $C_V^{-1}V_h$ as well as the number of iterations to reach a relative accuracy of 10^{-8} are given which confirm the theoretical estimates.

Table 1 Error, condition and iteration numbers in the case of uniform refinement

L	N	$\|w - w_h\|_{L_2(\Sigma)}$	eoc	$\kappa(V_h)$	It.	$\kappa(C_V^{-1}V_h)$	It.
0	2	2.249	-	1.001	1	1.002	1
1	4	1.311	0.778	2.808	2	1.279	2
2	8	0.658	0.996	4.905	4	1.422	4
3	16	0.324	1.021	7.548	8	1.486	8
4	32	0.160	1.017	11.140	16	1.541	14
5	64	0.079	1.010	16.724	31	1.563	13
6	128	0.040	1.006	13.470	41	1.590	13
7	256	0.020	1.003	22.053	50	1.615	12
8	512	0.010	1.001	32.043	59	1.636	12
9	1024	0.005	1.001	60.957	70	1.777	11
10	2048	0.002	1.000	88.488	82	1.762	11
11	4096	0.001	1.000	125.957	96	1.765	10

Adaptive refinement

For the second example we consider the initial datum $u_0(x) = 5e^{-10x}\sin \pi x$ which motivates the use of a locally quasi-uniform boundary element mesh resulting from some adaptive refinement strategy. The numerical results as given in Table 2 again confirm the theoretical findings, in particular the robustness of the proposed preconditioning strategy in the case of an adaptive refinement which is not the case when using none or only diagonal preconditioning $\widetilde{C}_V = \mathrm{diag} V_h$.

5 Conclusions and outlook

In this note we have described a space–time boundary element discretization of the spatially one-dimensional heat equation and an efficient and robust preconditioning strategy which is based on the use of boundary integral operators of opposite orders, but which requires a suitable stability condition for the boundary element spaces used for the discretization. In the particular case of the spatially one-dimensional

Table 2 Error, condition and iteration numbers in the case of adaptive refinement

L	N	$\|w-w_h\|_{L_2(\Sigma)}$	$\kappa(V_h)$	It.	$\kappa(\widetilde{C}_V^{-1}V_h)$	It.	$\kappa(C_V^{-1}V_h)$	It.
0	2	1.886	1.00	2	1.001	2	1.002	2
1	3	1.637	3.97	3	2.553	3	1.16	3
2	5	1.272	12.23	5	4.055	4	1.166	4
3	7	0.914	34.21	7	3.611	6	1.156	6
4	9	0.615	92.08	9	3.164	8	1.149	8
5	11	0.401	118.59	11	2.945	10	1.224	10
6	13	0.267	338.26	13	2.803	12	1.21	12
7	20	0.166	621.77	20	3.524	18	1.197	13
8	31	0.101	1608.08	31	4.457	27	1.252	12
9	47	0.063	2344.90	47	5.779	32	1.574	11
10	74	0.039	6141.47	74	8.348	37	1.692	11
11	114	0.024	8409.92	114	10.950	42	1.561	10
12	177	0.015	23007.60	173	14.324	47	1.716	10
13	278	0.010	27528.30	200	21.094	53	1.677	10

heat equation we can use the space $S_h^0(\Sigma)$ of piecewise constant basis functions to discretize both the single layer and the hypersingular boundary integral operator V and D, respectively. This is due to the inclusion $S_h^0(\Sigma) \subset H^{1/4}(\Sigma)$ where the latter is the Dirichlet trace space of the anisotropic Sobolev space $H^{1,1/2}(Q)$. In the case of a spatially two- or three-dimensional domain Ω a conformal approximation of the Dirichlet trace space $H^{1/2,1/4}(\Sigma)$ and therefore the discretization of the hypersingular integral operator D requires the use of continuous basis functions. Hence, to ensure the stability condition (5) we may use the space $S_h^1(\Sigma)$ of piecewise linear and continuous basis functions for the discretization of V and D, respectively, see [7, Theorem 3.2], and when assuming some appropriate mesh conditions locally [7, Section 4]. However, due to the approximation properties of $S_h^1(\Sigma)$ such an approach is restricted to spatial domains Ω with smooth boundary where the unknown flux is continuous.

When using the discontinuous boundary element space $S_h^0(\Sigma)$ for the approximation of the unknown flux we need to choose an appropriate boundary element space Y_h to ensure the stability condition (5). A possible approach is the use of a dual mesh using piecewise constant basis functions for the approximation of V, and piecewise linear and continuous basis functions for the approximation of D, see Fig. 2 for the situation in 1D. For a more detailed analysis of the proposed preconditioning strategy and suitable choices of stable boundary element spaces we refer to [2].

An efficient solution of local Dirichlet boundary value problems is an important tool when considering domain decomposition methods for the heat equation, see e.g. [9] in the case of the Laplace equation. Moreover, the preconditioning strategy of using operators of opposite order can also be used when considering related Schur complement systems on the skeleton, as they also appear in tearing and interconnecting domain decomposition methods, see, e.g., [4]. This also covers the coupling

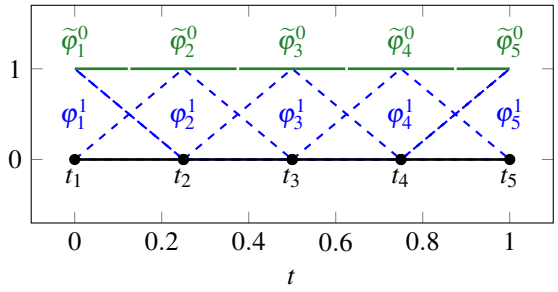


Fig. 2 Sample dual mesh. The piecewise linear and continuous functions φ_i^1 are used for the discretization of D. The piecewise constant basis functions $\tilde{\varphi}_i^0$ are used for the discretization of V

of space–time finite and boundary element methods. Related results on the stability and error analysis as well as on efficient solution strategies for space–time domain decomposition methods will be published elsewhere.

Acknowledgements This work was supported by the International Research Training Group 1754, funded by the German Research Foundation (DFG) and the Austrian Science Fund (FWF). Additionally, S. Dohr would like to acknowledge the financial support provided by the University of Bergen.

References

1. Costabel, M.: Boundary integral operators for the heat equation. Integral Equations Operator Theory, 13: 498–552, 1990.
2. Dohr, S., Niino, K., Steinbach, O.: Preconditioned space–time boundary element methods for the heat equation, in preparation.
3. Hiptmair, R.: Operator Preconditioning. Comput. Math. Appl., 52: 699–706, 2006.
4. Langer, U., Steinbach, O.: Boundary element tearing and interconnecting methods. Computing, 71: 205–228, 2003.
5. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications II. Springer, Berlin-Heidelberg, 1972.
6. Noon, P.J.: The Single Layer Heat Potential and Galerkin Boundary Element Methods for the Heat Equation. PhD thesis, University of Maryland, 1988.
7. Steinbach, O.: On the stability of the L_2 projection in fractional Sobolev spaces. Numer. Math., 88: 367–379, 2001.
8. Steinbach, O.: Space–time finite element methods for parabolic problems. Comput. Meth. Appl. Math., 15: 551-566, 2015.
9. Steinbach, O., Wendland, W.L.: Efficient preconditioners for boundary element methods and their use in domain decomposition. In: Domain Decomposition Methods in Sciences in Engineering: 8th International Conference, Beijing, China (R. Glowinski et. al. eds.), John Wiley, pp. 3–18, 1997.
10. Steinbach, O., Wendland, W.L.: The construction of some efficient preconditioners in the boundary element method. Adv. Comput. Math., 9: 191–216, 1998.

On high-order approximation and stability with conservative properties

Juan Galvis[1], Eduardo Abreu[2], Ciro Díaz[2], and Marcus Sarkis[3]

1 Summary

In this paper, we explore a method for the construction of locally conservative flux fields. The flux values are obtained through the use of a Ritz formulation in which we augment the resulting linear system of the continuous Galerkin (CG) formulation in a higher-order approximation space. These methodologies have been successfully applied to multi-phase flow models with heterogeneous permeability coefficients that have high-variation and discontinuities. The increase in accuracy associated with the high order approximation of the pressure solutions is inherited by the flux fields and saturation solutions. Our formulation allows us to use the saddle point problems analysis to study approximation and stability properties as well as iterative methods design for the resulting linear system. In particular, here we show that the low-order finite element problem preconditions well the high-order conservative discrete system. We present numerical evidence to support our findings.

2 Problem and conservative formulation

Consider the equation,

$$-\mathrm{div}(\Lambda(x)\nabla p) = q \quad \text{in} \quad \Omega \subset \Re^2, \tag{1}$$

$$p = 0 \quad \text{on} \quad \partial\Omega, \tag{2}$$

[1]Departamento de Matemáticas, Universidad Nacional de Colombia, Bogotá, Colombia.
[2]University of Campinas, Department of Applied Mathematics, 13.083-970, Campinas, SP, Brazil; supported by FAPESP Grant 2016/23374-1.
[3]Department of Mathematical Sciences, Worcester Polytechnic Institute Worcester USA; supported by NSF DMS-1522663

where Ω is a two-dimensional domain and Λ is a (smooth enough) positive definite symmetric matrix function. See [6] for the case of Λ being a multiscale coefficient with high-contrast. Our main interest is to obtain approximate solutions of the second order problem above[1] with: 1) high-order approximation (e.g., multiple basis per node), 2) local mass conservation properties and 3) stable-fast solver.

Our motivations come from the fact that in some applications it is **imperative** to have some conservative properties represented as conservations of total flux in control volumes. For instance, if $\mathbf{q}^h$ represents the approximation to the flux (in our case $\mathbf{q}^h = -\Lambda\nabla p^h$ where p^h is the approximation of the pressure), it is required that

$$\int_{\partial V} \mathbf{q}^h \cdot \mathbf{n} = \int_V q \quad \text{for each control volume } V.$$

For Dirichlet boundary condition, V is a control volume that does not cross $\partial\Omega$ from a set of control volumes of interest, and here and after $\mathbf{n}$ is the normal vector pointing out the control volume. We say that a discrete method is conservative if the total flux restriction such as the one written above holds.

We note that FV methods that use higher degree piecewise polynomials have been introduced in the literature; see [3, 4, 5]. We consider a Ritz formulation and construct a solution procedure that combines a continuous Galerkin-type formulation that concurrently satisfies mass conservation restrictions. We impose finite volume restrictions by using a scalar Lagrange multiplier for each restriction; see [1, 6].

The variational formulation of problem (1) is to find $p \in H_0^1(\Omega)$ such that

$$a(p, v) = F(v) \quad \text{for all } v \in H_0^1(\Omega), \tag{3}$$

where the bilinear form a is defined by

$$a(p, v) = \int_\Omega \Lambda(x)\nabla p(x)\nabla v(x)dx, \tag{4}$$

the functional F is defined by $F(v) = \int_\Omega q(x)v(x)dx$. The Problem (3) is equivalent to the minimization problem:

$$p = \arg \min_{v \in H_0^1(\Omega)} \mathcal{J}(v) \quad \text{where} \quad \mathcal{J}(v) = \frac{1}{2}a(v, v) - F(v). \tag{5}$$

Let the triangulation $\tau_h = \{R_k\}_{k=1}^{N_h}$ made of elements that are triangles or squares, where N_h is the number of elements. We also introduce the dual

[1] The use of second order formulation makes sense especially for cases where some form of high regularity holds. Usually in these cases the equality in the second order formulation is an equality in L^2 so that, in principle, there is no need to write the system of first order equations and weaken the equality by introducing less regular spaces for the pressure as it is done in mixed formulation with L^2 pressure.

mesh $\tau_h^* = \{V_k\}_{k=1}^{N_h^*}$ where the elements are called control volumes. In this paper we assume that each V_k is a subdomain of Ω with polygonal boundary. Let us introduce the space $H := \{v \in H_0^1(\Omega) : \Lambda\nabla v \in \mathrm{H}(\mathrm{div}, \Omega)\}$. If $q \in L^2$ we have that (3) is equivalent to: Find $p \in H_0^1$ such that

$$p = \arg\min_{v \in \mathcal{W}} \mathcal{J}(v), \tag{6}$$

where $\mathcal{W} = \left\{ v \in H : \int_{\partial T} -\Lambda\nabla v \cdot \mathbf{n} = \int_T q \quad \text{for all } T \in \tau_h^* \right\}$.

Problem (6) above can be view as Lagrange multipliers min-max optimization problem. See [2] and references therein. Let us denote $M_h = \mathbb{R}^{N_h^*}$.

The Lagrange multiplier formulation of problem (6) can be written as: Find $p \in H$ and $\lambda \in M_h$ that solves

$$(p, \lambda) = \arg \max_{\mu \in \mathbb{R}^{N_h^*}} \min_{v \in H,} \mathcal{J}(v) - (\overline{a}(v, \mu) - \overline{F}(\mu)). \tag{7}$$

Here, the total flux bilinear form $\overline{a} : H \times M_h \to \mathbb{R}$ is defined by

$$\overline{a}(v, \mu) = \sum_{k=1}^{N_h} \mu_k \int_{\partial V_k} \Lambda\nabla v \cdot \mathbf{n} \quad \text{for all } v \in H \text{ and } \mu \in M^h. \tag{8}$$

The functional $\overline{F} : M_h \to \mathbb{R}$ is defined by $\overline{F}(\mu) = \sum_{i=1}^{N_h} \mu_k \int_{V_k} q$, for all $\mu \in M^h$. The first order conditions of the min-max problem above give the following saddle point problem: Find $p \in H_0^1(\Omega)$ and $\lambda = 0 \in M_h$ that solves:

$$\begin{array}{ll} a(p, v) + \overline{a}(v, \lambda) = F(v) & \text{for all } v \in H, \\ \overline{a}(p, \mu) \qquad\quad = \overline{F}(\mu) & \text{for all } \mu \in M^h. \end{array} \tag{9}$$

3 Discretization and error

Let us consider $P^h = \mathbb{Q}^r(\tau_h) \cap H_0^1(\Omega)$. We also interpret M^h as $\mathbb{Q}^0(\tau_h^*)$, that is, the space of piecewise constant functions on the dual mesh τ_h^*. See for instance [6] where we consider GMsFEM spaces instead of piecewise polynomials.

The discrete version of (9) is to find $p^h \in P^h$ and $\lambda \in M^h$ such that

$$a(p^h, v^h) + \overline{a}(v^h, \lambda^h) = F(v^h) \qquad \text{for all } v^h \in P^h \tag{10}$$

$$\overline{a}(p^h, \mu^h) = \overline{F}(\mu^h) \qquad \text{for all } \mu^h \in M^h. \tag{11}$$

The equivalent matrix form is,

$$\begin{bmatrix} A & \overline{A}^T \\ \overline{A} & O \end{bmatrix} \begin{bmatrix} u^h \\ \lambda^h \end{bmatrix} = \begin{bmatrix} f \\ \overline{f} \end{bmatrix} \tag{12}$$

where A is the finite element stiffness matrix corresponding to finite element space $P^h = \text{span}\,\{\varphi_j\}$,

$$A = [a_{i,j}] \quad \text{where } a_{ij} = \int_\Omega \Lambda \nabla \varphi_i \cdot \nabla \varphi_j. \tag{13}$$

The restriction or finite volume matrix $\overline{A}$ is given by,

$$\overline{A} = [\overline{a}_{k,j}] \quad \text{where } \overline{a}_{kj} = \int_{\partial V_k} \Lambda \nabla \varphi_j \cdot \mathbf{n}. \tag{14}$$

Moreover, $f = [f_i]$ with $f_i = \int_\Omega q\,\varphi_i$ and $\overline{f} = [\overline{f}_k]_{k=1}^{N_h^*}$ with $\overline{f}_k = \int_{V_k} q$.

Note that matrix $\overline{A}$ is related to classical (low order) finite volume matrix. Matrix $\overline{A}$ is a rectangular matrix with more columns than rows. Several previous works on conservative high-order approximation of second order elliptic problem have been designed by "adding" rows using several constructions. See [1] for details.

We consider a particular case of a regular mesh made of squares. Our analysis is valid for high order finite element on regular meshes made of triangles since a similar analysis holds in this case. Define $\Gamma^* = \bigcup_{k=1}^{N_h^*}$ that is, Γ^* is the interior interface generated by the dual triangulation. For $\mu \in M^h$ define $[\mu]$ on Γ^* as the jump across element interfaces such that $[\mu]|_{\partial V_k \cap \partial V_{k'}} = \mu_k - \mu_{k'}$. Note that $\overline{a}(v, \overline{\mu}) = \sum_{k=1}^{N_h^*} \mu_k \int_{\partial V_k} \nabla v \cdot \mathbf{n} = \int_{\Gamma^*} \nabla v \cdot \mathbf{n}\,[\mu]$.

In our analysis we use the energy norm in the space that approximates the pressure and a discrete norm in the space of Lagrange multipliers. Denote $\|v\|_a^2 = \int_\Omega \Lambda \nabla v \cdot \nabla v$ for all $v \in H_0^1(\Omega)$. Let us recall the definition of space $H := \{v \in H_0^1(\Omega) : \Lambda \nabla v \in \text{H(div}, \Omega)\}$, and additional set $P_+^h = \text{Span}\{P^h, H\}$. We define the norm (that is motivated by the analysis)

$$\|v\|_{P_+^h}^2 = |v|_{H^1(\Omega)}^2 + h^2 \sum_{\ell=1}^{N_h^*} \|\Delta v\|_{L^2(R_\ell)}^2 \quad \text{for all} \quad v \in P_+^h. \tag{15}$$

Note that if $v \in \mathbb{Q}^r$, then $\|v\|_{P_+^h}^2 \asymp |v|_{H^1(\Omega)}^2$ using an inverse inequality. Also define the discrete norm for the spaces of Lagrange multipliers as

$$\|\mu\|_{M^h}^2 = \frac{1}{h} \int_{\Gamma^*} [\mu]^2. \tag{16}$$

It is possible to verify that ([1])

1. **Augmented norm**: $\|v\|_a \leq \|v\|_{P_+^h}$ for all $v \in P_+^h$.
2. **Continuity**: $|\bar{a}| \in \mathbb{R}$ such that $|\bar{a}(v,\mu^h)| \leq |\bar{a}|\|v\|_{P_+^h}\|\mu^h\|_{M^h}$ for all $v \in P_+^h$ and $\mu^h \in M^h$.
3. **Inf-Sup**: $\inf_{\mu^h \in M^h} \sup_{v \in P_+^h} \frac{\overline{a}(v,\mu^h)}{\|v\|_a \|\mu^h\|_{M^h}} \geq \alpha > 0$.

We also have established optimal approximation in energy norm ($\|p - p^h\|_a \preceq h|p|_{H^2(\Omega)}$) and using a duality argument it is possible to write the optimal L^2 approximation $\|p - (p^h + \lambda^h)\|_0 \preceq h^2|p|_{H^2(\Omega)}$; see [1] for details.

4 The case of highly anisotropic media

One issue with some cases of conservative methods is the lack of coerciveness under the presence of high-anisotropic coefficients. We can think our formulation as a stabilization for these cases (in the sense that we increase the space of the solution while keeping fixed the space for the Lagrange multipliers). Preliminary numerical studies suggest that our formulation is more robust (with respect to anisotropy) than the classical finite volume formulations.

A nice feature of our formulation is that the symmetric saddle point (12) is suitable for constructing robust preconditioners; see [2] for variety of solvers and iteration that can be used. Here we present a simple stationary iteration. Consider the iteration

$$\begin{aligned} Au_{k+1} &= f - \overline{A}^T\lambda_k \\ \lambda_{k+1} &= \lambda_k + \omega B^{-1}(\overline{A}u_{k+1} - \overline{f}). \end{aligned} \tag{17}$$

Here ω is a relaxation parameter and B a preconditioner to be defined. This iteration corresponds to a preconditioned Richardson iteration applied to the Schur complement problem (to solve for the Lagrange multiplier lambda equation). We have, by combining the two equations above,

$$\lambda_{k+1} = \lambda_k + \omega B^{-1}\left(g - S\lambda_k\right)$$

where $g = \overline{A}A^{-1}f - \overline{f}$ and S is the Schur complement $S = \overline{A}A^{-1}\overline{A}^T$. Note that the size of S is the number of interior vertices if the control volumes are constructed by joining the centers of the elements of the primal mesh. In the case of isotropic coefficients and square elements, we can take $B = M_h$ defined in (16); see [2]. In order to take into account the anisotropy, below in the numerical tests we consider B defined by

$$B = [b_{ij}] \text{ where } b_{ij} = \int_D \Lambda\nabla\varphi_i\nabla\varphi_j \text{ with } \varphi_i, \varphi_j \in \mathbb{Q}^1 \cap H_0^1(\Omega).$$

5 Numerical experiments

We consider the Dirichlet problem (1). Let $\Omega = (0,1) \times (0,1)$. We consider a regular mesh made of 4^L squares. The dual mesh is constructed by joining the centers of the elements of the primal mesh. We perform a series of numerical experiments to compare properties of FEM solutions with the solution of our high order FV formulation (to which we refer from now on as FV solution). We select the exact solution $p(x,y) = \sin(\pi x)\sin(\pi y)(-x+3y)$ and $f = -\Delta u$.

On Table 1 we compare our $\mathbb{Q}^1$ FV method with the classical $\mathbb{Q}^1$ finite element method. We compute L^2 and H^1 errors. We observe optimal convergence of both strategies however the FV is conservative. On Table 2 we consider $\mathbb{Q}^2$ elements and optimal higher convergence rates are confirmed.

L	$FEM,\ L^2\ Error$	$FV.\ L^2\ Error$	$FEM,\ H^1\ Error$	$FV.\ H^1\ Error$
1	1.5538×10^{-1}	1.5103×10^{-1}	1.1297×10^{0}	1.1338×10^{0}
2	3.6342×10^{-2}	3.1881×10^{-2}	5.3226×10^{-1}	5.3416×10^{-1}
3	8.9720×10^{-3}	$7.5.276 \times 10^{-3}$	2.6374×10^{-1}	2.6403×10^{-1}
4	2.2548×10^{-3}	1.9348×10^{-3}	1.3163×10^{-1}	1.3172×10^{-1}
5	5.5513×10^{-4}	4.6095×10^{-4}	6.5833×10^{-2}	6.5840×10^{-2}
6	1.3875×10^{-4}	1.1513×10^{-4}	3.2948×10^{-2}	3.2924×10^{-2}
7	3.4685×10^{-5}	2.8776×10^{-5}	1.6418×10^{-2}	1.6489×10^{-2}

Table 1 Table of **FEM** and **FV** L^2 and H^1 errors using $\mathbb{Q}^1$ elements.

L	$FEM\ L^2\ Error$	$FV.\ L^2\ Error$	$FEM\ H^1\ Error$	$FV.\ H^1\ Error$
1	1.4061×10^{-2}	2.4548×10^{-2}	1.9302×10^{-1}	2.2436×10^{-1}
2	2.1217×10^{-3}	4.9023×10^{-3}	5.4862×10^{-2}	7.2895×10^{-2}
3	2.6860×10^{-4}	6.4789×10^{-4}	1.4072×10^{-2}	1.8847×10^{-2}
4	3.3875×10^{-5}	8.1756×10^{-5}	3.5418×10^{-3}	4.7552×10^{-3}
5	4.2437×10^{-6}	1.0242×10^{-5}	8.3539×10^{-4}	1.2667×10^{-3}
6	5.3075×10^{-7}	1.2810×10^{-6}	2.2016×10^{-4}	2.9616×10^{-4}
7	6.6353×10^{-8}	1.6015×10^{-7}	5.5043×10^{-5}	7.4046×10^{-5}

Table 2 Table of **FEM** and **FV** L^2 and H^1 errors using $\mathbb{Q}^2$ elements.

We now move to symmetric anisotropic coefficients Λ. We now show in Tables 3-8 the smallest and the largest eigenvalues of $\lambda\max(B^{-1}S)/\lambda_{\min}(B^{-1}S)$ for different values of Λ, $h = 2^L$ and for $\mathbb{Q}^1, \mathbb{Q}^2$ and $\mathbb{Q}^3$ elements. The Λ has eigenvalues 1 and η and associate eigenvector $\eta = (\cos(\Theta), \sin(\Theta))^t$. From these results we see that the smallest eigenvalue is very stable, therefore, the discrete inf-sup is satisfied. This is a strong result since finite volume discretizations sometimes lack in coerciveness for highly anisotropic media. The proposed preconditioner performs well however has a mildly dependence with respect to the different configuration of anisotropy direction and anisotropy

ratio. This is somehow expected since the continuity given in (15) is with respect to the V_h-norm rather than a-norm, and further studies are on the way to eliminate this dependence. Recall that the application of the preconditioner requires the solution of a low-order ($\mathbb{Q}^1$) classical symmetric finite element problem. In practice, these solve can be replaced by a robust method for low-order finite element method and inexact Uzawa or Conjugated Gradient. Recall also that we obtain conservative solutions.

$L\backslash\eta$	1	10	100	1000	1	10	100	1000	1	10	100	1000
2	1.76	1.76	1.76	1.76	1.76	1.76	1.81	1.81	1.76	1.80	1.82	1.83
	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05
3	2.09	2.09	2.09	2.09	2.09	2.11	2.12	2.12	2.09	2.11	2.13	2.14
	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
4	2.20	2.20	2.20	2.20	2.20	2.21	2.22	2.22	2.20	2.21	2.22	2.22
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	2.24	2.24	2.24	2.24	2.23	2.24	2.24	2.24	2.24	2.24	2.24	2.24
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25	2.25
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3 Maximum and minimum eigenvalue $\frac{\lambda_{max}}{\lambda_{min}}$ for $\Theta = 1$(left), $\Theta = \frac{\pi}{6}$ (center), $\Theta = \frac{\pi}{4}$ (right) and $P^h = \mathbb{Q}^1$. The Λ has eigenvalues 1 and η and the eigenvector associated to η is $(\cos(\Theta), \sin(\Theta))^t$.

$L\backslash\eta$	1	10	100	1000	1	10	100	1000	1	10	100	1000
2	1.79	1.80	1.81	1.81	1.79	2.13	2.47	2.53	1.79	2.32	2.98	3.12
	1.05	1.05	1.06	1.06	1.05	1.09	1.11	1.11	1.05	1.10	1.12	1.12
3	2.10	2.10	2.11	2.11	2.10	2.50	2.99	3.18	2.10	2.77	4.03	4, 43
	1.01	1.01	1.01	1.01	1.01	1.02	1.03	1.03	1.01	1.02	1.03	1.03
4	2.21	2.21	2.21	2.21	2.21	2.61	3.27	3.92	2.21	2.91	4.40	5.24
	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.00	1.01	1.01	1.01
5	2.24	2.24	2.24	2.24	2.24	2.64	3.43	4.90	2.24	2.95	4.52	6.43
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	2.25	2.25	2.25	2.25	2.25	2.65	3.48	5.86	2.25	2.95	4.57	7.60
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4 Maximum and minimum eigenvalue $\frac{\lambda_{max}}{\lambda_{min}}$ for $\Theta = 1$ (left), $\Theta = \frac{\pi}{6}$ (center), $\Theta = \frac{\pi}{4}$ (right) and $P_h = \mathbb{Q}^2$. The Λ has eigenvalues 1 and η and the eigenvector associated to η is $(\cos(\Theta), \sin(\Theta))^t$.

6 Conclusions

In this paper we use a Ritz formulation with constraints to obtain locally conservative fluxes in the approximation of the Darcy equation. With this

$L\backslash\eta$	1	10	100	1000	1	10	100	1000	1	10	100	1000
2	4.52	4.54	4.57	4.58	4.52	4.72	6.92	7.21	4.52	4.80	5.01	5.05
	2.43	2.43	2.43	2.43	2.29	2.29	2.29	2.29	2.44	2.28	2.24	2.23
3	5.32	5.35	5.40	5.41	5.33	5.47	6.92	7.21	5.32	5.67	7.37	7.68
	2.29	2.29	2.29	2.29	2.29	1.95	1.89	1.90	2.29	1.92	1.86	1.64
4	5.59	5.62	5.68	5.69	5.60	5.89	10.8	12.2	5.59	6.48	12.2	13.7
	2.26	2.26	2.26	2.26	2.26	1.81	1.74	1.74	2.26	1.78	1.72	1.72
5	5.67	6.70	5.75	5.77	5.67	6.19	16.9	22.2	5.67	7.09	20.2	26.3
	2.25	2.25	2.25	2.25	2.25	1.75	1.68	1.67	2.25	1.73	1.67	1.66
6	5.69	5.72	5.77	5.79	5.68	5.68	24.7	41.4	5.68	7.46	30.8	50.7
	2.25	2.25	2.25	2.25	2.25	1.72	1.65	1.65	2.25	1.70	1.65	1.64

Table 5 Maximum and minimum eigenvalue $\frac{\lambda_{max}}{\lambda_{min}}$ for $\Theta = 1$ (left), $\Theta = \frac{\pi}{6}$ (center), $\Theta = \frac{\pi}{4}$ (right) and $P_h = \mathbb{Q}^3$.

formulation we obtain solution that have high-order approximation and still yield locally conservative fluxes with no post-processing. We show that the resulting linear system can be solve using a stationary iteration where the application of the preconditioner uses an approximation of a low-order finite element problem. We present numerical evidence to support our findings.

Acknowledgments E. Abreu thanks financial support FAPESP through grant No. 2016/23374-1.

References

[1] E. Abreu, C. Díaz, J. Galvis and M. Sarkis, On high-order conservative finite element methods, Computers & Mathematics with Applications. Online December 2017 (https://doi.org/10.1016/j.camwa.2017.10.020).
[2] M. Benzi, G. H. Golub and J. Liesen, Numerical solution of saddle point problems, Acta numerica 14 (2005) 1-137.
[3] L. Chen, A New Class of High Order Finite Volume Methods for Second Order Elliptic Equations, SIAM J. Numer. Anal., 47(6) (2010) 4021-4043.
[4] Z. Chen, J. Wu and Y. Xu, Higher-order finite volume methods for elliptic boundary value problems, 37(2) (2012) 191-253.
[5] Z. Chen, Y. Xu and Y. Zhang, A construction of higher-order finite volume methods, Math. Comp. 84 (2015) 599-628.
[6] M. Presho and J. Galvis, A mass conservative Generalized Multiscale Finite Element Method applied to two-phase flow in heterogeneous porous media, Journal of Computational and Applied Mathematics, 296 (2016) 376-388.

A Nonlinear ParaExp Algorithm

Martin J. Gander, Stefan Güttel, and Madalina Petcu

1 Derivation of the Nonlinear ParaExp Algorithm

Time parallelization has a long history, see [1] and references therein. The parallel speedup obtained is in general not as good as with space parallelization, especially for hyperbolic problems. A notable exception are waveform relaxation-type methods [3, 4], which in the hyperbolic case are related to the more recent tent-pitching approach [6], and the ParaExp algorithm [7, 9] based on Krylov methods, which is however restricted to linear problems. For an application in a nonlinear context, see [10], and for a different approach using Krylov information, see [8]. Here we propose and analyze a variant of the ParaExp algorithm for the nonlinear initial value problem

$$\mathbf{u}'(t) = A\mathbf{u}(t) + B\big(\mathbf{u}(t)\big) + \mathbf{g}(t), \quad t \in [0,T], \quad \mathbf{u}(0) = \mathbf{u}_0, \tag{1.1}$$

with $A \in \mathbb{C}^{m\times m}$, $B : \mathbb{C}^m \to \mathbb{C}^m$ a nonlinear operator, $\mathbf{g} : [0,T] \to \mathbb{C}^m$ a source function, and $\mathbf{u} : [0,T] \to \mathbb{C}^m$ the sought solution. Throughout this note we assume that all stated initial value problems have unique solutions. For the ParaExp algorithm, the

M. J. Gander
Université de Genève, 2-4 rue du Lièvre, 1211 Genève, Switzerland,
e-mail: Martin.Gander@unige.ch

S. Güttel
School of Mathematics, The University of Manchester, United Kingdom,
e-mail: stefan.guettel@manchester.ac.uk

M. Petcu
Laboratoire de Mathématiques, Université de Poitiers, France,
The Institute of Statistics and Applied Mathematics of the Romanian Academy,
The Institute of Mathematics of the Romanian Academy, Bucharest, Romania,
e-mail: Madalina.Petcu@math.univ-poitiers.fr

time interval $[0,T]$ is partitioned into N subintervals $[T_{n-1},T_n]$ with $n=1,\ldots,N$, and a direct application of this algorithm to the nonlinear problem (1.1) gives

Step 1: Solve for $n \geq 1$ in parallel the nonlinear problems with zero initial data

$$\begin{aligned} \mathbf{v}_n'(t) &= A\mathbf{v}_n(t) + B\big(\mathbf{v}_n(t)\big) + \mathbf{g}(t), \quad t \in [T_{n-1},T_n], \\ \mathbf{v}_n(T_{n-1}) &= \mathbf{0}. \end{aligned}$$

Step 2: Solve for $n \geq 1$ in parallel the linear non-homogeneous problems

$$\begin{aligned} \mathbf{w}_n'(t) &= A\mathbf{w}_n(t), & t \in [T_{n-1},T], \\ \mathbf{w}_n(T_{n-1}) &= \mathbf{v}_{n-1}(T_{n-1}), & \mathbf{v}_0(T_0) = \mathbf{u}_0. \end{aligned}$$

ParaExp then forms the linear combination $\mathbf{u}(t) = \mathbf{v}_n(t) + \sum_{j=1}^n \mathbf{w}_j(t)$, $t \in [T_{n-1},T_n)$, which still satisfies the initial condition, but not equation (1.1) since $\mathbf{u}'(t) = A\mathbf{u}(t) + B\big(\mathbf{v}_n(t)\big) + \mathbf{g}(t)$, $t \in [T_{n-1},T_n]$, except when B is not present. One can however naturally separate the solution into $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{w}(t)$, with $\mathbf{w}$ solving the linear problem $\mathbf{w}'(t) = A\mathbf{w}(t)$, $\mathbf{w}(t) = \mathbf{u}_0$, and $\mathbf{v}$ solving the nonlinear remaining part $\mathbf{v}'(t) = A\mathbf{v}(t) + B\big(\mathbf{v}(t) + \mathbf{w}(t)\big) + \mathbf{g}(t)$, $\mathbf{v}(0) = \mathbf{0}$. To apply this splitting on multiple time intervals $[T_{n-1},T_n]$ we need to iterate. Using the initialization $\mathbf{v}_n^0(T_n) = \mathbf{0}$ for $n = 1,\ldots,N$ (or some other approximation), we perform for $k = 1,2,\ldots$

Step 1: Solve for $n \geq 1$ in parallel the linear problems

$$\begin{aligned} \big(\mathbf{w}_n^k\big)'(t) &= A\mathbf{w}_n^k(t), & t \in [T_{n-1},T], \\ \mathbf{w}_n^k(T_{n-1}) &= \mathbf{v}_{n-1}^{k-1}(T_{n-1}), & \mathbf{w}_1^k(T_0) = \mathbf{u}_0. \end{aligned} \tag{1.2}$$

Step 2: Solve for $n \geq 1$ in parallel the nonlinear problems

$$\begin{aligned} \big(\mathbf{v}_n^k\big)'(t) &= A\mathbf{v}_n^k(t) + B\big(\mathbf{v}_n^k(t) + \sum_{j=1}^n \mathbf{w}_j^k(t)\big) + \mathbf{g}(t), \quad t \in [T_{n-1},T_n], \\ \mathbf{v}_n^k(T_{n-1}) &= \mathbf{0}. \end{aligned} \tag{1.3}$$

The new approximate solution is then defined by $\mathbf{u}^k(t) = \mathbf{v}_n^k(t) + \sum_{j=1}^n \mathbf{w}_j^k(t)$, $t \in [T_{n-1},T_n)$, which now satisfies equation (1.1) on each time interval $[T_{n-1},T_n)$, and $\mathbf{u}^k(0) = \mathbf{u}_0$. The solution of the linear part (1.2) can still be computed efficiently as in the ParaExp algorithm using Krylov techniques, but (1.3) requires the computation of $\sum_{j=1}^n \mathbf{w}_j^k$ on $[T_{n-1},T_n]$, and thus would need the Krylov approximation of $\mathbf{w}_j^k$ on the entire interval $[T_{n-1},T_n]$. To avoid this, we rewrite the algorithm in terms of $\mathbf{u}_n^k$ instead of $\mathbf{v}_n^k$, where $\mathbf{u}_n^k$ approximates $\mathbf{u}$: starting with $\mathbf{u}_n^0(T_n) = \mathbf{w}_j^0(T_n) = \mathbf{0}$ for all j and n, the nonlinear ParaExp algorithm performs for $k = 1,2,\ldots$

Step 1: Solve for $n \geq 1$ in parallel the linear problems

$$\begin{aligned} \left(\mathbf{w}_n^k\right)'(t) &= A\mathbf{w}_n^k(t), && t \in [T_{n-1}, T], \\ \mathbf{w}_n^k(T_{n-1}) &= \mathbf{u}_{n-1}^{k-1}(T_{n-1}) - \sum_{j=1}^{n-1} \mathbf{w}_j^{k-1}(T_{n-1}), && \mathbf{w}_1^k(T_0) = \mathbf{u}_0. \end{aligned} \tag{1.4}$$

Step 2: Solve for $n \geq 1$ in parallel the nonlinear problems

$$\begin{aligned} \left(\mathbf{u}_n^k\right)'(t) &= A\mathbf{u}_n^k(t) + B\left(\mathbf{u}_n^k(t)\right) + \mathbf{g}(t), \quad t \in [T_{n-1}, T_n], \\ \mathbf{u}_n^k(T_{n-1}) &= \sum_{j=1}^{n} \mathbf{w}_j^k(T_{n-1}), \end{aligned} \tag{1.5}$$

and form the new approximate solution as

$$\mathbf{u}^k(t) = \mathbf{u}_n^k(t), \quad t \in [T_{n-1}, T_n). \tag{1.6}$$

Remark 1. To avoid the computation of $\mathbf{u}_n^k$ as the solution of a nonlinear problem, one could linearize (1.5) by using in the nonlinear term $B(\mathbf{u}_n^{k-1})$ instead of $B(\mathbf{u}_n^k)$, where $\mathbf{u}_n^0 = \mathbf{0}$ or some other approximation of the solution. However, in what follows we focus on the fully nonlinear version, since then $\mathbf{u}^k$ is the solution of the nonlinear problem (1.1) on each time interval.

2 Analysis of the Nonlinear ParaExp Algorithm

We first show that the nonlinear ParaExp algorithm introduced in the previous section converges in a finite number of steps.

Theorem 1. *The approximate solution* $\mathbf{u}^k$ *obtained at iteration* k *and defined by* (1.6) *coincides with the exact solution* $\mathbf{u}$ *on the time interval* $[T_0, T_k)$.

Proof. Since $\mathbf{w}_1^k(T_0) = \mathbf{u}_0$ for all $k = 1, 2, \ldots$, $\mathbf{w}_1^k = \mathbf{w}_1^{k-1}$ on the time interval $[T_0, T]$ for all $k = 2, 3, \ldots$. Next, for $k = 1$ we have $\mathbf{u}^1(t) = \mathbf{u}_1^1(t)$ on $[T_0, T_1]$, and since $\mathbf{u}_1^1(T_0) = \mathbf{w}_1^1(T_0) = \mathbf{u}_0$ we get by the uniqueness of the solution of (1.5) that $\mathbf{u}_1^1$ coincides with the exact solution $\mathbf{u}$ on the time interval $[T_0, T_1]$.

We now prove by induction that for all $k = 2, 3 \ldots$ we have

$$\mathbf{u}_n^k = \mathbf{u} \ \text{ on } \ [T_{n-1}, T_n], \ \forall n \leq k, \qquad \mathbf{w}_n^k = \mathbf{w}_n^{k-1} \ \text{ on } \ [T_{n-1}, T], \ \forall\, n \leq k-1. \tag{2.1}$$

For $k = 2$, we only need to prove property (2.1) for $\mathbf{u}^2$, since for $\mathbf{w}_1^2$ it is ensured by the fact that $\mathbf{w}_1^k = \mathbf{w}_1^{k-1}$ for all $k \geq 2$. The initial condition for $\mathbf{u}_2^2$ is

$$\mathbf{u}_2^2(T_1) = \mathbf{w}_1^2(T_1) + \mathbf{w}_2^2(T_1) = \mathbf{w}_1^2(T_1) + \mathbf{u}_1^1(T_1) - \mathbf{w}_1^1(T_1) = \mathbf{u}_1^1(T_1) = \mathbf{u}(T_1),$$

where we used the fact that $\mathbf{w}_1^2 = \mathbf{w}_1^1$ and that $\mathbf{u}_1^1$ is the exact solution on the time interval $[T_0, T_1]$. Since $\mathbf{u}_2^2$ satisfies the same equation as $\mathbf{u}$ on the time interval $[T_1, T_2]$ and $\mathbf{u}_2^2(T_1) = \mathbf{u}(T_1)$, $\mathbf{u}_2^2$ must coincide with $\mathbf{u}$ on $[T_1, T_2]$. But we also know that

$\mathbf{u}_1^2(T_0) = \mathbf{w}_1^2(T_0) = \mathbf{u}_0$ and that $\mathbf{u}_1^2$ satisfies (1.5), which implies $\mathbf{u}_1^2 = \mathbf{u}$ on $[T_0, T_1]$, and hence $\mathbf{u}^2$ coincides with the exact solution of (1.1) on the time interval $[T_0, T_2)$.

We now suppose that (2.1) holds for all iterations up to an arbitrarily fixed index k and we prove (2.1) for $k+1$. To first check that $\mathbf{w}_n^{k+1} = \mathbf{w}_n^k$ on $[T_{n-1}, T]$ for all $n = 2, 3, \ldots, k$, we compute

$$\begin{aligned}\mathbf{w}_n^{k+1}(T_{n-1}) &= \mathbf{u}_{n-1}^k(T_{n-1}) - \sum_{j=1}^{n-1} \mathbf{w}_j^k(T_{n-1}) = \mathbf{u}(T_{n-1}) - \sum_{j=1}^{n-1} \mathbf{w}_j^{k-1}(T_{n-1}) \\ &= \mathbf{u}_{n-1}^{k-1}(T_{n-1}) - \sum_{j=1}^{n-1} \mathbf{w}_j^{k-1}(T_{n-1}) = \mathbf{w}_n^k(T_{n-1}),\end{aligned}$$

where we have used the recurrence hypothesis (2.1). Since $\mathbf{w}_n^{k+1}$ and $\mathbf{w}_n^k$ satisfy the same equation and have the same initial condition, the result follows. We next prove that $\mathbf{u}_n^{k+1} = \mathbf{u}$ on $[T_{n-1}, T_n]$ for all $n \le k+1$. Since we already know that $\mathbf{u}_n^{k+1}$ and $\mathbf{u}$ satisfy the same equation on the time interval $[T_{n-1}, T_n]$, we only need to check that the initial condition satisfied by $\mathbf{u}_n^{k+1}$,

$$\begin{aligned}\mathbf{u}_n^{k+1}(T_{n-1}) &= \sum_{j=1}^{n} \mathbf{w}_j^{k+1}(T_{n-1}) = \sum_{j=1}^{n-1} \mathbf{w}_j^{k+1}(T_{n-1}) + \mathbf{u}_{n-1}^k(T_{n-1}) - \sum_{j=1}^{n-1} \mathbf{w}_j^k(T_{n-1}) \\ &= \mathbf{u}_{n-1}^k(T_{n-1}),\end{aligned}$$

where we used the first result we just proved for $\mathbf{w}_n^{k+1}$ and that $\mathbf{w}_1^{k+1} = \mathbf{w}_1^k$ for all k. Now, using the recurrence hypothesis (2.1), we know that $\mathbf{u}_{n-1}^k$ coincides with the exact solution of (1.1) on $[T_{n-2}, T_{n-1}]$, which implies that $\mathbf{u}_n^{k+1}(T_{n-1}) = \mathbf{u}(T_{n-1})$. □

We now show that the nonlinear ParaExp algorithm can be interpreted in the context of the Parareal algorithm if written as a multiple shooting method (see [5, 2]). We will need the following result.

Lemma 1. *Let $(\mathbf{u}_n^k)_{k,n}$ be the sequence defined by the nonlinear ParaExp algorithm* (1.4)–(1.6)*. Defining $\widetilde{\mathbf{u}}_n^0(T_n) = \mathbf{0}$ and $\mathbf{C}_n^0(T_n) = \mathbf{0}$ for all $n \ge 0$, let $\left(\mathbf{C}_n^k\right)_{k,n}$ for all $k \ge 1$ and $n \ge 1$ be the solutions of the linear problems*

$$\begin{aligned}\left(\mathbf{C}_n^k\right)'(t) &= A\mathbf{C}_n^k(t), && t \in [T_{n-1}, T_n], \\ \mathbf{C}_n^k(T_{n-1}) &= \mathbf{C}_{n-1}^k(T_{n-1}) + \widetilde{\mathbf{u}}_{n-1}^{k-1}(T_{n-1}) - \mathbf{C}_{n-1}^{k-1}(T_{n-1}), && \mathbf{C}_1^k(T_0) = \mathbf{u}_0,\end{aligned}$$

and let $\left(\widetilde{\mathbf{u}}_n^k\right)_{k,n}$ be the solutions of the nonlinear problems

$$\begin{aligned}\left(\widetilde{\mathbf{u}}_n^k\right)'(t) &= A\widetilde{\mathbf{u}}_n^k(t) + B\left(\widetilde{\mathbf{u}}_n^k(t)\right) + \mathbf{g}(t), \quad t \in [T_{n-1}, T_n], \\ \widetilde{\mathbf{u}}_n^k(T_{n-1}) &= \mathbf{C}_n^k(T_{n-1}).\end{aligned}$$

Then $\mathbf{u}_n^k = \widetilde{\mathbf{u}}_n^k$ on $[T_{n-1}, T_n]$ for all $n \ge 0$ and $k \ge 1$.

Proof. At step $k = 1$ and for all $n \ge 1$, $\mathbf{C}_n^1$ is the solution of the linear problem

$$(\mathbf{C}_n^1)'(t) = A\mathbf{C}_n^1(t), \qquad t \in [T_{n-1}, T_n],$$
$$\mathbf{C}_n^1(T_{n-1}) = \mathbf{C}_{n-1}^1(T_{n-1}), \quad \mathbf{C}_1^1(T_0) = \mathbf{u}_0.$$

Hence $\mathbf{C}_n^1$ is the restriction of the solution of $\mathbf{u}' = A\mathbf{u}$, $\mathbf{u}(0) = \mathbf{u}_0$ on $[T_0, T]$ to the time interval $[T_{n-1}, T_n]$. Taking into account the definition (1.4) of $\mathbf{w}_n^1$, we notice that $\mathbf{w}_n^1 = \mathbf{0}$ for $n > 1$ and $\mathbf{w}_1^1$ is the solution of the linear problem $\mathbf{u}' = A\mathbf{u}$, $\mathbf{u}(0) = \mathbf{u}_0$ on $[T_0, T]$. Thus, $\mathbf{C}_n^1(t) = \sum_{j=1}^n \mathbf{w}_j^1(t)$ on $[T_{n-1}, T_n]$, and $\widetilde{\mathbf{u}}_n^1$ satisfies for $n \geq 1$

$$\big(\widetilde{\mathbf{u}}_n^1\big)'(t) = A\widetilde{\mathbf{u}}_n^1(t) + B\big(\widetilde{\mathbf{u}}_n^1(t)\big) + \mathbf{g}(t), \quad t \in [T_{n-1}, T_n],$$
$$\mathbf{u}_n^1(T_{n-1}) = \mathbf{C}_n^1(T_{n-1}) = \sum_{j=1}^n \mathbf{w}_j^1(T_{n-1}).$$

Comparing this with (1.5) and using the uniqueness of the solution for the nonlinear problem, we deduce that $\mathbf{u}_n^1(t) = \widetilde{\mathbf{u}}_n^1(t)$ on $[T_{n-1}, T_n]$ for all $n \geq 1$.

Assuming now that for all $n \geq 1$ and a given k we have $\mathbf{C}_n^k(t) = \sum_{j=1}^n \mathbf{w}_j^k(t)$, $\mathbf{u}_n^k(t) = \widetilde{\mathbf{u}}_n^k(t)$ on $[T_{n-1}, T_n]$,we need to show that this also holds for $k+1$. To do so, we prove by recurrence with respect to n that $\mathbf{C}_n^{k+1}(t) = \sum_{j=1}^n \mathbf{w}_j^{k+1}(t)$ on $[T_{n-1}, T_n]$. For $n = 1$, we have that $\mathbf{C}_1^{k+1}(T_0) = \mathbf{u}_0 = \mathbf{w}_1^{k+1}(T_0)$ and, since $\mathbf{C}_1^{k+1}$ and $\mathbf{w}_1^{k+1}$ satisfy the same equation and the same initial condition, we conclude that $\mathbf{C}_1^{k+1} = \mathbf{w}_1^{k+1}$ on $[T_0, T_1]$. Next, we suppose that $\mathbf{C}_n^{k+1}(t) = \sum_{j=1}^n \mathbf{w}_j^{k+1}(t)$ on $[T_{n-1}, T_n]$ and prove that $\mathbf{C}_{n+1}^{k+1}(t) = \sum_{j=1}^{n+1} \mathbf{w}_j^{k+1}(t)$ on $[T_n, T_{n+1}]$. By checking the initial condition of $\mathbf{C}_{n+1}^{k+1}$ at T_n and using the recurrence hypothesis, we find

$$\mathbf{C}_{n+1}^{k+1}(T_n) = \mathbf{C}_n^{k+1}(T_n) + \mathbf{u}_n^k(T_n) - \sum_{j=1}^n \mathbf{w}_j^k(T_n) = \mathbf{C}_n^{k+1}(T_n) + \mathbf{w}_{n+1}^{k+1}(T_n) = \sum_{j=1}^{n+1} \mathbf{w}_j^{k+1}(T_n).$$

Since $\mathbf{C}_{n+1}^{k+1}$ and $\sum_{j=1}^{n+1} \mathbf{w}_j^{k+1}$ solve the same linear problem on $[T_n, T_{n+1}]$ and satisfy the same initial condition at T_n, we obtain $\mathbf{C}_{n+1}^{k+1} = \sum_{j=1}^{n+1} \mathbf{w}_j^{k+1}$ on $[T_n, T_{n+1}]$. Further, for $n \geq 1$ we have

$$\big(\widetilde{\mathbf{u}}_n^{k+1}\big)'(t) = A\widetilde{\mathbf{u}}_n^{k+1}(t) + B\big(\widetilde{\mathbf{u}}_n^{k+1}(t)\big) + \mathbf{g}(t), \quad t \in [T_{n-1}, T_n],$$
$$\widetilde{\mathbf{u}}_n^{k+1}(T_{n-1}) = \mathbf{C}_n^{k+1}(T_{n-1}) = \sum_{j=1}^n \mathbf{w}_j^{k+1}(T_{n-1}).$$

Thus, $\widetilde{\mathbf{u}}_n^{k+1}$ and $\mathbf{u}_n^{k+1}$ solve the same equation with identical initial condition on $[T_{n-1}, T_n]$ and hence $\widetilde{\mathbf{u}}_n^{k+1} = \mathbf{u}_n^{k+1}$ on $[T_{n-1}, T_n]$. □

The following theorem is essentially a reformulation of Lemma 1 in the usual notation of the parareal algorithm in terms of a coarse and a fine integrator [11].

Theorem 2. *Let the coarse propagator $G(T_n, T_{n-1}, \mathbf{U})$ solve the linear problem*

$$\mathbf{u}'(t) = A\mathbf{u}(t) \ \textit{on} \ [T_{n-1}, T_n], \quad \mathbf{u}(T_{n-1}) = \mathbf{U},$$

and let the fine propagator $F(T_n, T_{n-1}, \mathbf{U})$ solve the nonlinear problem

$$\mathbf{u}'(t) = A\mathbf{u}(t) + B\big(\mathbf{u}(t)\big) + \mathbf{g}(t) \;\; on \;\; [T_{n-1}, T_n], \quad \mathbf{u}(T_{n-1}) = \mathbf{U}.$$

Then the solution $\mathbf{u}^k$ computed by the nonlinear ParaExp algorithm (1.4)–(1.6) *coincides at each time point T_n with the solution $\mathbf{U}_n^k$ computed by the parareal algorithm*

$$\mathbf{U}_n^k = F(T_n, T_{n-1}, \mathbf{U}_{n-1}^{k-1}) + G(T_n, T_{n-1}, \mathbf{U}_{n-1}^k) - G(T_n, T_{n-1}, \mathbf{U}_{n-1}^{k-1}). \tag{2.2}$$

Proof. Using the definition of $\mathbf{u}^k$ in (1.6) and the notation of Lemma 1, we have

$$\begin{aligned}
&\mathbf{u}^k(T_n) = \mathbf{u}_{n+1}^k(T_n) = \mathbf{C}_{n+1}^k(T_n) = \mathbf{C}_n^k(T_n) + \mathbf{u}_n^{k-1}(T_n) - \mathbf{C}_n^{k-1}(T_n) \\
&= G\big(T_n, T_{n-1}, \mathbf{C}_n^k(T_{n-1})\big) - G\big(T_n, T_{n-1}, \mathbf{C}_n^{k-1}(T_{n-1})\big) + \widetilde{\mathbf{u}}_n^{k-1}(T_n) \\
&= G\big(T_n, T_{n-1}, \mathbf{C}_n^k(T_{n-1})\big) - G\big(T_n, T_{n-1}, \mathbf{C}_n^{k-1}(T_{n-1})\big) + F\big(T_n, T_{n-1}, \mathbf{C}_n^{k-1}(T_{n-1})\big).
\end{aligned}$$

Thus $\mathbf{u}^k(T_n) = \mathbf{U}_n^k$ with $\mathbf{U}_n^k = \mathbf{C}_{n+1}^k(T_n)$. □

Theorem 2 shows that the nonlinear ParaExp algorithm is mathematically equivalent to the parareal algorithm (2.2) where the coarse integrator G is an exponential integrator for $\mathbf{w}' = A\mathbf{w}$. There is however an important computational difference: due to the linearity of G we can write

$$\begin{aligned}
&G(T_n, T_{n-1}, \mathbf{U}_{n-1}^{k+1}) \\
&= G\big(T_n, T_{n-1}, F(T_{n-1}, T_{n-2}, \mathbf{U}_{n-2}^k) - G(T_{n-1}, T_{n-2}, \mathbf{U}_{n-2}^k) + G(T_{n-1}, T_{n-2}, \mathbf{U}_{n-2}^{k+1})\big) \\
&= G\big(T_n, T_{n-1}, F(T_{n-1}, T_{n-2}, \mathbf{U}_{n-2}^k) - G(T_{n-1}, T_{n-2}, \mathbf{U}_{n-2}^k)\big) + G(T_n, T_{n-2}, \mathbf{U}_{n-2}^{k+1}),
\end{aligned}$$

which corresponds to the coarse propagation of a jump over $[T_{n-1}, T_n]$ plus the coarse propagation of $\mathbf{U}_{n-2}^{k+1}$ over a longer time interval $[T_{n-2}, T_n]$. Repeating a similar calculation for $G(T_n, T_{n-2}, \mathbf{U}_{n-2}^{k+1})$, we derive

$$\begin{aligned}
G(T_n, T_{n-2}, \mathbf{U}_{n-2}^{k+1}) \;=\; & G\big(T_n, T_{n-2}, F(T_{n-2}, T_{n-3}, \mathbf{U}_{n-3}^k) - G(T_{n-2}, T_{n-3}, \mathbf{U}_{n-3}^k)\big) \\
& + G(T_n, T_{n-3}, \mathbf{U}_{n-3}^{k+1}),
\end{aligned}$$

which again corresponds to the coarse propagation of a jump (over two intervals) plus a coarse propagation of $\mathbf{U}_{n-3}^{k+1}$ (over three intervals). This recursion can be repeated, and it will terminate as $\mathbf{U}_{n-n}^{k+1} = \mathbf{U}_0$ is known, leading to an alternative, more compact formulation of the nonlinear ParaExp algorithm:

$$\text{initialize} \quad \mathbf{U}_n^0 = G(T_n, T_0, \mathbf{U}_0) \quad \text{for } \; n = 0, 1, \ldots, N,$$

$$\mathbf{U}_n^{k+1} = G(T_n, T_0, \mathbf{U}_0) + \sum_{j=1}^{n} G\big(T_n, T_j, F(T_j, T_{j-1}, \mathbf{U}_{j-1}^k) - G(T_j, T_{j-1}, \mathbf{U}_{j-1}^k)\big).$$

Here the coarse integrator is applied in parallel, which is different from parareal. The price to pay is that the coarse integrations now span multiple overlapping time

intervals $[T_j, T_n]$. As in the original ParaExp algorithm, these linear homogeneous problems can be solved very efficiently using Krylov methods.

3 Numerical Illustration

We now investigate the nonlinear ParaExp algorithm numerically. We solve the nonlinear wave equation $u_{tt} = u_{xx} + \alpha u^2$ on the time-space domain $[0,4] \times [-1,1]$ with homogeneous Dirichlet boundary conditions and $u(0,x) = e^{-100x^2}$, $u'(0,x) = 0$, where the parameter $\alpha \geq 0$ controls the nonlinear character of the problem. The problem is discretized in space using finite differences with $m = 200$ equispaced interior grid points on $[-1,1]$. This gives rise to the ODE

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}' = \begin{bmatrix} O & I \\ L & O \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \alpha \mathbf{u}^2 \end{bmatrix},$$

where $L = \text{tridiag}(1,-2,1)/h^2$, $h = 2/(m+1)$, and the operation $\mathbf{u}^2$ has to be understood component-wise. We partition the time interval $[0,4]$ into $n = 20$ slices of equal length and use as fine integrator MATLAB's `ode15s` routine with a relative error tolerance of 10^{-6}. For the linear coarse integration we use MATLAB's `expm`.

Table 1 lists, for varying $\alpha \in \{0,2,4,6,8.2\}$, the number of iterations required by our nonlinear ParaExp algorithm to achieve an error of order $\approx 1e-6$ over all time slices. Figure 1 shows, again for varying α, the reference solutions $u(t,x)$ on the left, and on the right the error of the ParaExp solution at each time point t_j after $k = 1,2,\ldots$ iterations. Here a number of $k = 0$ iterations corresponds to the error of the ParaExp initialization with the coarse integrator.

The parameter $\alpha = 0$ gives rise to a linear problem. Note that for this case the error of the initialization is of order $\approx 10^{-6}$, and not of order machine precision as one would expect from the exponential integration using `expm`. This is because our reference solution has been computed via `ode15s` and is of lower accuracy.

For increasing values of α the nonlinear character of the wave equation becomes more pronounced and typically more ParaExp iterations are required. It depends on the efficiency of the coarse propagator (in this case `expm`) if any speed-up would be obtained in a parallel implementation. For large-scale problems the use of (rational) Krylov techniques as in [7] is recommended. The nonlinear ParaExp method becomes inefficient for highly nonlinear problems, with 14 iterations required for $\alpha = 8.2$. This is expected and we note that for $\alpha \approx 9$ the solution $u(t,x)$ even appears to have a singularity in the time-space domain of interest.

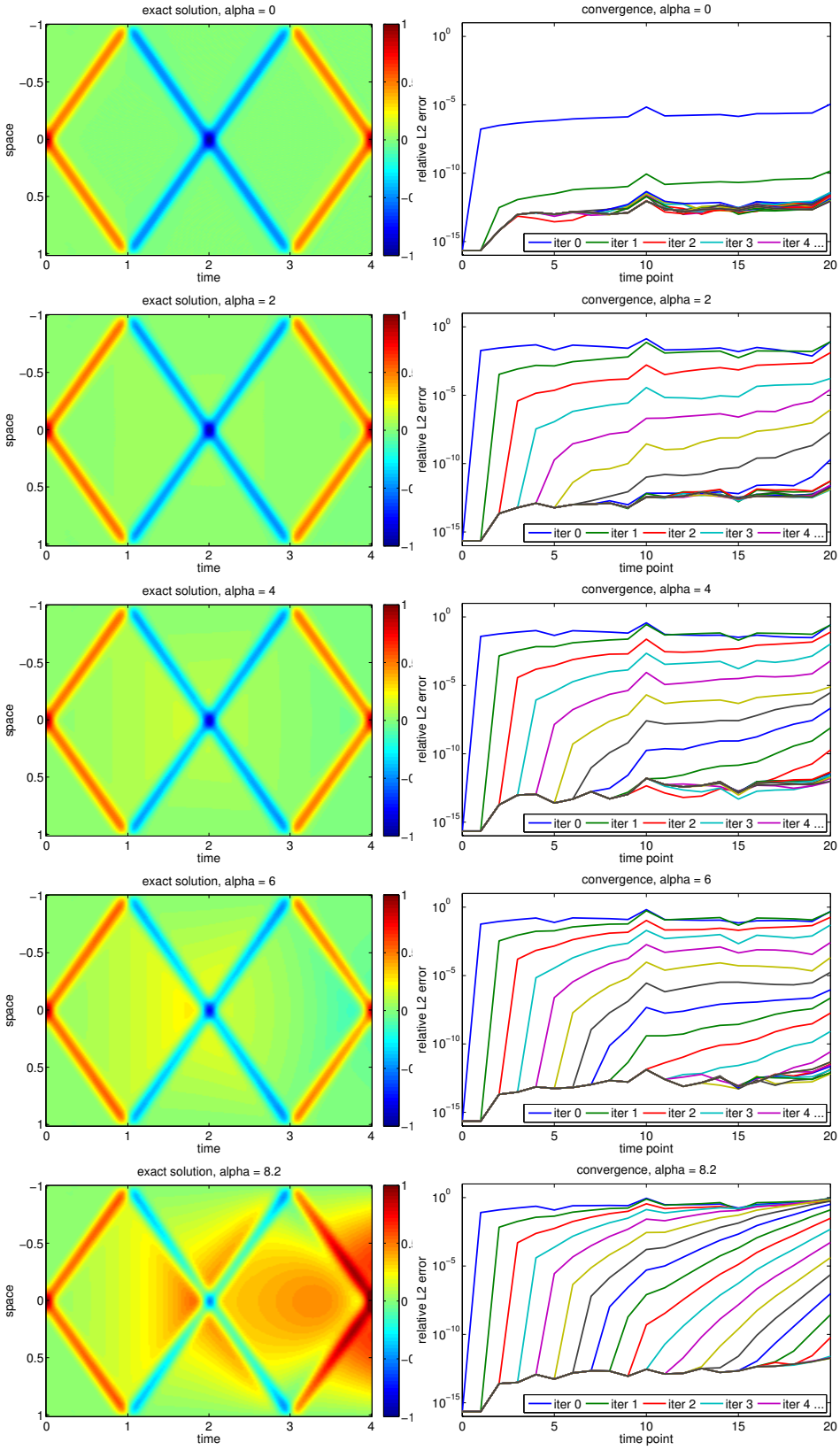


Fig. 1 Exact solutions (left) and convergence (right) of the nonlinear ParaExp algorithm applied to a nonlinear wave equation with varying parameter $\alpha \in \{0, 2, 4, 6, 8.2\}$ (top to bottom).

parameter α	0	2	4	6	8.2
# iterations	1	5	7	7	14

Table 1 Number of iterations required by the nonlinear ParaExp algorithm to solve a nonlinear wave equation to fixed accuracy uniformly over a time interval. The parameter α controls the nonlinearity of the problem.

References

1. M. J. GANDER, *50 years of time parallel time integration*, Multiple Shooting and Time Domain Decomposition Methods, Springer, pp. 69–113, 2015.
2. M. J. GANDER AND E. HAIRER, *Nonlinear convergence analysis for the parareal algorithm*, Domain Decomposition Methods in Science and Engineering XVII, Springer, pp. 45–56, 2007.
3. M. J. GANDER, L. HALPERN, AND F. NATAF, *Optimal Schwarz waveform relaxation for the one dimensional wave equation*, SIAM J. Numer. Anal., 41:1643–1681, 2003.
4. M. J. GANDER AND L. HALPERN, *Absorbing boundary conditions for the wave equation and parallel computing*, Math. Comp., 74:153–176, 2004.
5. M. J. GANDER AND S. VANDERWALLE, *Analysis of the parareal time-parallel time-integration method*, SIAM J. Sci. Comput., 29(2):556–578, 2007.
6. J. GOPALAKRISHNAN, J. SCHÖBERL, AND C. WINTERSTEIGER, *Mapped tent pitching schemes for hyperbolic systems*, SIAM J. Sci. Comput., 39(6), B1043–B1063, 2017.
7. M. J. GANDER AND S. GÜTTEL, *PARAEXP: A parallel integrator for linear initial-value problems*, SIAM J. Sci. Comput., 35(2):C123–C142, 2013.
8. M. J. GANDER AND M. PETCU, *Analysis of a Krylov subspace enhanced parareal algorithm*, ESAIM Proc., 25:45–56, 2008.
9. S. GÜTTEL, *A parallel overlapping time-domain decomposition method for ODEs*, Domain Decomposition Methods in Science and Engineering XX, Springer, pp. 483–490, 2013.
10. G. KOOIJ, M. BOTCHEV, AND B. GEURTS, *A block Krylov subspace implementation of the time-parallel ParaExp method and its extension for nonlinear partial differential equations*, J. Comput. Appl. Math., 316:229–246, 2017.
11. J.-L. LIONS, Y. MADAY, AND G. TURINICI, *A "parareal" in time discretization of PDE's*, C. R. Acad. Sci. Paris Sér. I Math., 332:661–668, 2001.

On Optimal Coarse Spaces for Domain Decomposition and Their Approximation

Martin J. Gander, Laurence Halpern, and Kévin Santugini

1 Definition of the Optimal Coarse Space

We consider a general second order elliptic model problem

$$\mathscr{L}u = f \quad \text{in } \Omega \tag{1}$$

with some given boundary conditions that make the problem well posed. We decompose the domain Ω first into non-overlapping subdomains $\tilde{\Omega}_j$, $j = 1,2,\ldots,J$, and to consider also overlapping domain decomposition methods, we construct overlapping subdomains Ω_j from $\tilde{\Omega}_j$ by simply enlarging them a bit. All domain decomposition methods provide at iteration n solutions u_j^n on the subdomains $\tilde{\Omega}_j$, $j = 1,2,\ldots,J$ (or on Ω_j in the case of overlapping methods, but then we just restrict those to the non-overlapping decomposition $\tilde{\Omega}_j$). We want to study here properties of the correction that needs to be added to these subdomain solutions in order to obtain the solution u of (1). This would be the best possible correction a coarse space can provide, independently of the domain decomposition method used, and it allows us to define an optimal coarse space, which we then approximate.

Since the u_j^n are subdomain solutions, they satisfy equation (1) on their corresponding subdomain,

$$\mathscr{L}u_j^n = f, \quad \text{in } \tilde{\Omega}_j. \tag{2}$$

Defining the error

Martin J. Gander
University of Geneva, Section of Mathematics, Rue du Lievre 2-4, CP 64, 1211 Geneva 4, e-mail: martin.gander@unige.ch

Laurence Halpern
LAGA, Université Paris 13, Avenue Jean Baptiste Clément, 93430 Villetaneuse, France e-mail: halpern@math.univ-paris13.fr

Kévin Santugini
Institut de Mathématiques de Bordeaux, France e-mail: kevin.santugini@math.u-bordeaux1.fr

$$e^n(x) := u(x) - u_j^n(x), \quad x \in \tilde{\Omega}_j,$$

we see that the error satisfies the homogeneous problem in each subdomain,

$$\mathscr{L} e^n = 0 \quad \text{in } \tilde{\Omega}_j. \tag{3}$$

At the interface between the non-overlapping subdomains $\tilde{\Omega}_j$ the error is in general not continuous, and also the normal derivative of the error is not continuous, since the subdomain solutions u_j^n in general do not have this property[1]. The best coarse space, which we call optimal coarse space, must thus contain piecewise harmonic functions on $\tilde{\Omega}_j$ to be able to represent the error.

2 Computing the Optimal Coarse Correction

Having identified the optimal coarse space, we need to explain a general method to determine the optimal coarse correction in it. While two different approaches for specific cases can be found in [8, 7], we present now a completely general approach: let us denote the interface between subdomain $\tilde{\Omega}_i$ and $\tilde{\Omega}_j$ by Γ_{ij}, and let the jumps in the Dirichlet and Neumann traces between subdomain solutions be denoted by

$$g_{ij}^n(x) := u_i^n(x) - u_j^n(x), \quad h_{ij}^n(x) := \partial_{n_i} u_i^n(x) + \partial_{n_j} u_j^n(x), \quad x \in \Gamma_{ij}, \tag{4}$$

where ∂_{n_j} denotes the outer normal derivative of subdomain $\tilde{\Omega}_j$. Then the error satisfies the transmission problem

$$\begin{aligned} \mathscr{L} e^n &= 0 && \text{in } \tilde{\Omega}_j, \\ e_i^n(x) - e_j^n(x) &= g_{ij}^n(x) && \text{on } \Gamma_{ij}, \\ \partial_{n_i} e_i^n(x) + \partial_{n_j} e_j^n(x) &= h_{ij}^n(x) && \text{on } \Gamma_{ij}. \end{aligned} \tag{5}$$

Its solution lies in the optimal coarse space, and when added to the iterates u_i^n, we obtain the solution: the domain decomposition method has become a direct solver, it is nilpotent, independently of the domain decomposition method and the problem we solve: no better coarse correction is possible!

We now give a weak formulation of the transmission problem (5). To simplify the exposition, we use the case of the Laplacian, $\mathscr{L} := -\Delta$. We multiply the partial differential equation from (5) in each subdomain $\tilde{\Omega}_j$ by a test function v_j and integrate by parts to obtain

$$\int_{\Omega_j} \nabla e_j \cdot \nabla v_j - \int_{\Gamma_j} \frac{\partial e_j}{\partial n_j} v_j = 0. \tag{6}$$

[1] For certain methods, continuity of the normal derivative is however assured, like in the FETI methods, or continuity of the Dirichlet traces, like in the Neumann-Neumann method or the alternating Schwarz method. This can be used to reduce the size of the optimal coarse space.

If we denote by $\tilde{e}$ and $\tilde{v}$ the functions defined on all of Ω by the piecewise definition $\tilde{e}|_{\Omega_j} := e_j$ and $\tilde{v}|_{\Omega_j} := v_j$, then we can combine (6) over all subdomains $\tilde{\Omega}_j$ to obtain

$$\int_{\Omega} \nabla\tilde{e}\cdot\nabla\tilde{v} - \sum_{\Gamma_{ij}} \int_{\Gamma_{ij}} \frac{\partial e_j}{\partial n_j} v_j + \frac{\partial e_i}{\partial n_i} v_i = 0. \tag{7}$$

If we impose now continuity on the test functions v_j, i.e. $\tilde{v}$ to be continuous, then (7) becomes

$$\int_{\Omega} \nabla\tilde{e}\cdot\nabla\tilde{v} - \sum_{\Gamma_{ij}} \int_{\Gamma_{ij}} \left(\frac{\partial e_j}{\partial n_j} + \frac{\partial e_i}{\partial n_i} \right) \tilde{v} = 0, \tag{8}$$

and we can use the data of the problem to remove the normal derivatives,

$$\int_{\Omega} \nabla\tilde{e}\cdot\nabla\tilde{v} - \sum_{\Gamma_{ij}} \int_{\Gamma_{ij}} h_{ij}\tilde{v} = 0. \tag{9}$$

It is therefore natural to choose a continuous test function $\tilde{v}$ to obtain a variational formulation of the transmission problem (5), a function in the space

$$V := \{v : v|_{\Omega_i} =: v_i \in H_1(\tilde{\Omega}_i),\ v_i = v_j \text{ on } \Gamma_{ij}\}. \tag{10}$$

Now the jump in the Dirichlet traces of the errors would in general be imposed on the trial function space,

$$U := \{u : u|_{\Omega_i} =: u_i \in H_1(\tilde{\Omega}_i),\ u_i - u_j = g_{ij} \text{ on } \Gamma_{ij}\}, \tag{11}$$

so the complete variational formulation for (5) is:

$$\text{find } \tilde{e} \in U, \text{ such that } \int_{\Omega} \nabla\tilde{e}\cdot\nabla\tilde{v} - \sum_{\Gamma_{ij}} \int_{\Gamma_{ij}} h_{ij}\tilde{v} = 0 \quad \forall \tilde{v} \in V. \tag{12}$$

To discretize the variational formulation (12), we have to choose approximations of the spaces V and U, and both spaces contain interior Dirichlet conditions. In a finite element setting, it is natural to enforce the homogeneous Dirichlet conditions in V_h strongly if the mesh is matching at the interfaces, i.e. we just impose the nodal values to be the same for V_h.

While at the continuous level, the optimal coarse correction lies in an infinite dimensional space except for 1d problems, see [5, 7], at the discrete level this space becomes finite dimensional. It is in principle then possible to use the optimal coarse space at the discrete level and to obtain a nilpotent method, i.e. a method which converges after the coarse correction, see for example [9, 8, 11, 10], and also [1] for conditions under which classical subdomain iterations can become nilpotent. It is however not very practical to use these high dimensional optimal coarse spaces, and we are thus interested in approximations.

3 Approximations of the Optimal Coarse Space

We have seen that the optimal coarse space contains functions which satisfy the homogeneous equation in each non-overlapping subdomain $\tilde{\Omega}_j$, i.e. they are harmonic in $\tilde{\Omega}_j$. To obtain an approximation of the optimal coarse space, it is therefore sufficient to define an approximation for the functions on the interfaces Γ_{ij}, which are then extended harmonically inside $\tilde{\Omega}_j$. A natural way to approximate the functions on the interfaces is to use a Sturm-Liuville eigenvalue problem, and then to select eigenfunctions which correspond to modes on which the subdomain iteration of the domain decomposition methods used is not effective. This can be done either for the entire subdomain, for example choosing eigenfunctions of the Dirichlet to Neumann operator of the subdomain, see [2], or any other eigenvalue problem along the entire boundary of the subdomain $\tilde{\Omega}_j$, or piecewise on each interface Γ_{ij}, in which case also basis functions relating cross points need be added [11, 10], see also the ACMS coarse space [12] and references therein. This can be done solving for example lower dimensional counterparts of the original problem along the interface Γ_{ij} with boundary conditions one at one end, and zero at the other, creating something like hat functions around the crosspoint. Doing this for example for a rectangular domain decomposed into rectangular subdomains for Laplaces equation, this would just generate Q1 functions on each subdomain. It is important however to not force these function to be continuous across subdomains, since they have to solve approximately the transmission problem (5) whose solution is not continuous, except for specific methods[2]. So the resulting coarse basis function is not a hat function with one degree of freedom, but it is a discontinuous hat function with e.g. four degrees of freedom if four subdomains meet at that cross point.

Different approaches not based on approximating an optimal coarse space, but also using eigenfunctions in the coarse space to improve specific inequalities in the convergence analysis of domain decomposition methods are GenEO [14], whose functions are also harmonic in the interior of subdomains, and [3, 4], where volume eigenfunctions are used which are thus not harmonic within subdomains. For a good overview, see [13].

4 Concrete Example: the Parallel Schwarz Method

We consider the high contrast diffusion problem $\nabla \cdot (a(x,y)\nabla u) = f$ in $\Omega = (0,1)^2$ with two subdomains $\Omega_1 = (0, \frac{1+\delta}{2}) \times (0,1)$ and $\Omega_2 = (\frac{1-\delta}{2}, 1) \times (0,1)$. The classical parallel Schwarz method is converging most slowly for low frequencies along the interface $x = \frac{1}{2}$, i.e. error components represented in the Laplacian case by $\sin(k\pi y)$, $k = 1, 2, \ldots, K$ for some small integer K, see for example [6]. These are precisely the eigenfunctions of the eigenvalue problem one obtains when using separation of variables, which in our high contrast case is

[2] see footnote 1

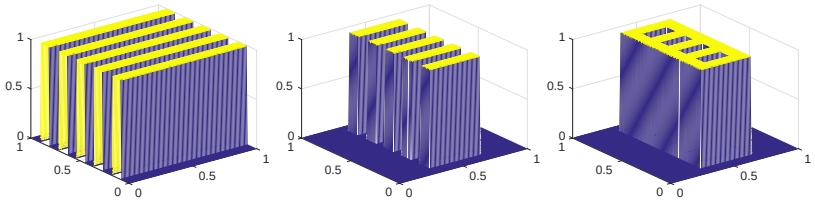


Fig. 1 An example with long channels, shortened channels, and closed shortened channels

$$\partial_y(a_\Gamma \phi_y) = \lambda a_\Gamma \phi, \tag{13}$$

where a_Γ denotes the trace of the high contrast parameter along the interface, in our simple example $a_\Gamma(y) := a(\frac{1}{2}, y)$. So already in the case of Laplaces equation, it would be good to enrich a classical $Q1$ coarse space aligned with the decomposition with harmonically extended eigenfunctions $\sin(k\pi y)$, $k = 1, 2, \ldots, K$ into the subdomains. We now illustrate why this is even more important in the case of high contrast channels, the $a(x,y)$ of which are shown in Figure 1. We show in Figure 2 the performance of a classical parallel Schwarz method with two subdomains for increasing overlap sizes. We see that for the case of the long channels increasing the overlap improves the performance of the classical Schwarz methods as for the Laplacian[3], and nothing special happens between overlap $41h$ and overlap $43h$. This is however completely different for the shortened channel case, independently if they are closed or not, were increasing the overlap does not help at all, until suddenly changing from overlap $41h$ and overlap $43h$, the method becomes fast. This can be easily understood by the maximum principle, and is illustrated in Figure 3 which shows the errors in the subdomains. We clearly see that due to the fast diffusion the error propagates rapidly from the interface into the subdomains, and the maximum principle indicates slow convergence, as long as the overlap does not contain the shortened channels. As soon as the overlap contains the shortened channels, convergence becomes rapid. This is very different for the long channels, as illus-

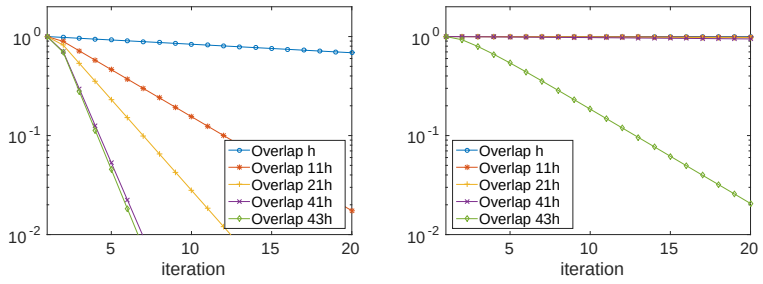


Fig. 2 Convergence behavior of a classical parallel Schwarz method for high contrast long and shortened channels

[3] the same happens if inclusions are only contained within the subdomains, outside the overlap

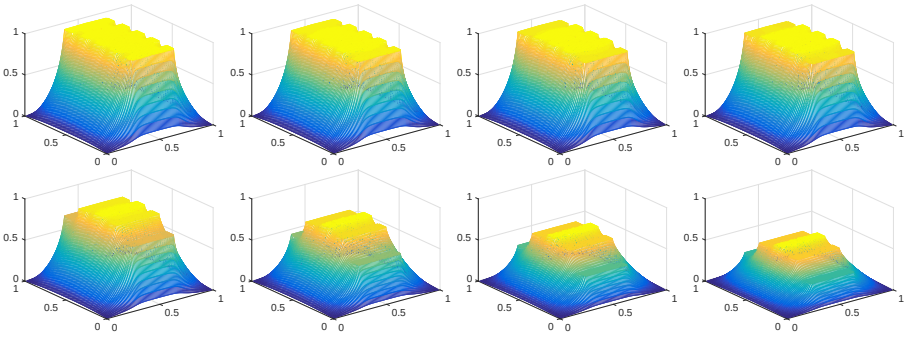


Fig. 3 Error for the first four iterations in the shortened channel case: Top overlap 41h, and bottom overlap 43h, and slightly more overlap suddenly leads to much more rapid convergence.

trated in Figure 4. Here the channels touch the outer boundary of the domain, and the maximum principle indicates rapid convergence.

The case of shortened channels is precisely the situation where the convergence mechanism of the underlying domain decomposition method has problems, and if one can not afford a large enough overlap, a well chosen coarse space can help. It suffices to add harmonically extended low frequency modes of the cheap, lower dimensional interface eigenvalues problem to the coarse space, leading to the so called Spectrally Harmonically Enriched Multiscale coarse space (SHEM), see [11, 10]. Figure 5 shows that the eigenfunctions of the cheap interface eigenvalue problem are almost identical to the eigenfunctions obtained from the expensive DtN eigenvalue problem on the shortened channels from [2, 12], and still very similar to the ones of the DtN eigenvalue problem on the shortened closed channels, except for the first one. We show in Figure 6 on the left the eigenvalues of the cheap interface eigenvalue problem, compared to the eigenvalues of the expensive DtN-operator on the shortened channels and the shortened closed channels. They all indicate via the smallest eigenvalues that there are five channels, and five coarse functions are

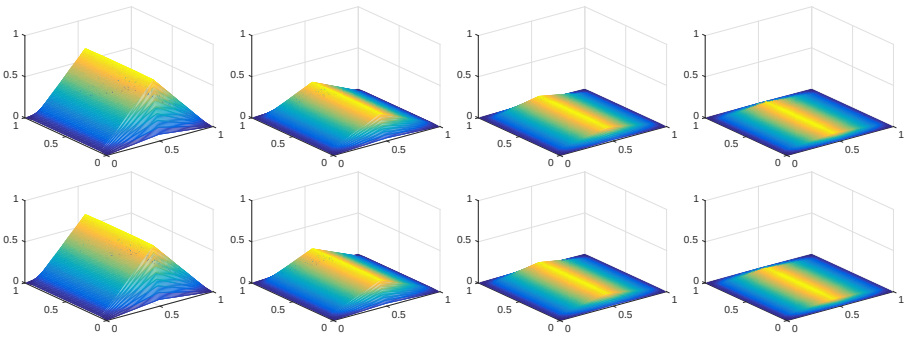


Fig. 4 Error for the first four iterations in the shortened channel case: Top overlap 41h, and bottom overlap 43h, and slightly more overlap leads to slightly more rapid converge.

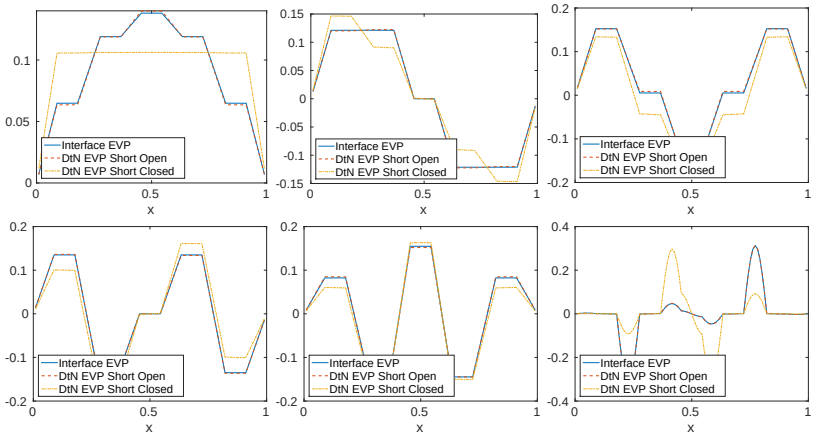


Fig. 5 Eigenfunctions of the different eigenvalue problems compared

needed for good convergence, see Figure 6 on the right. The DtN-eigenvalue problem for the shortened closed channels also indicates that there is only one eigenvalue going to zero when the contrast becomes large. To obtain good convergence, it is however also in the closed shortened channel case necessary to include five enrichment functions in the coarse space, see Figure 7. It thus suffices as in SHEM to use the inexpensive interface eigenvalue problem to construct an effective approximation of the optimal coarse space, see [10] for simulations in the more general case of many subdomains and contrast functions.

References

1. Faycal Chaouqui, Martin J. Gander, and Kévin Santugini-Repiquet. On nilpotent subdomain iterations. *Domain Decomposition Methods in Science and Engineering XXIII, Springer*, 2016.

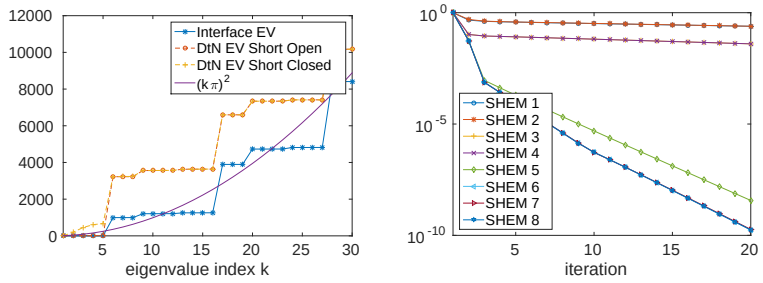


Fig. 6 Left: staircase behavior of the eigenvalues. Right: convergence with optimized coarse space on shortened channels

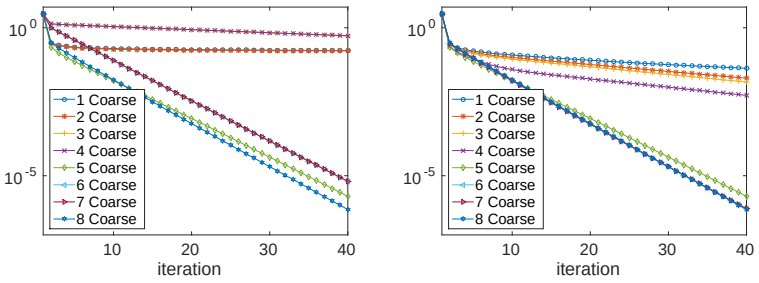


Fig. 7 Shortened closed channels. Left: coarse space based on the interface eigenvalue problem. Right: coarse space based on the DtN eigenvalue problem for the shortened closed channels.

2. Victorita Dolean, Frédéric Nataf, Robert Scheichl, and Nicole Spillane. Analysis of a two-level Schwarz method with coarse spaces based on local Dirichlet-to-Neumann maps. *Comput. Methods Appl. Math.*, 12(4):391–414, 2012.
3. Juan Galvis and Yalchin Efendiev. Domain decomposition preconditioners for multiscale flows in high-contrast media. *Multiscale Model. Simul.*, 8(4):1461–1483, 2010.
4. Juan Galvis and Yalchin Efendiev. Domain decomposition preconditioners for multiscale flows in high contrast media: reduced dimension coarse spaces. *Multiscale Model. Simul.*, 8(5):1621–1644, 2010.
5. M Gander and Laurence Halpern. Méthode de décomposition de domaine. *Encyclopédie électronique pour les ingénieurs*, 2012.
6. Martin J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44(2):699–731, 2006.
7. Martin J. Gander, Laurence Halpern, and Kévin Santugini-Repiquet. Discontinuous coarse spaces for DD-methods with discontinuous iterates. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 607–615. Springer, 2014.
8. Martin J. Gander, Laurence Halpern, and Kévin Santugini-Repiquet. A new coarse grid correction for RAS/AS. In *Domain Decomposition Methods in Science and Engineering XXI*, pages 275–283. Springer, 2014.
9. Martin J Gander and Felix Kwok. Optimal interface conditions for an arbitrary decomposition into subdomains. In *Domain Decomposition Methods in Science and Engineering XIX*, pages 101–108. Springer, 2011.
10. Martin J. Gander and Atle Loneland. SHEM: An optimal coarse space for RAS and its multiscale approximation. *Domain Decomposition Methods in Science and Engineering XXIII, Springer*, 2016.
11. Martin J. Gander, Atle Loneland, and Talal Rahman. Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods. *arXiv preprint arXiv:1512.05285*, 2015.
12. A. Heinlein, A. Klawonn, J. Knepper, and O. Rheinbach. Multiscale coarse spaces for overlapping Schwarz methods based on the ACMS space in 2d. *ETNA*, page submitted, 2017.
13. Robert Scheichl. Robust coarsening in multiscale PDEs. In Randolph Bank, Michael Holst, Olof Widlund, and Jinchao Xu, editors, *Domain Decomposition Methods in Science and Engineering XX*, volume 91 of *Lecture Notes in Computational Science and Engineering*, pages 51–62. Springer Berlin Heidelberg, 2013.
14. N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numer. Math.*, 126(4):741–770, 2014.

Analysis of Overlap in Waveform Relaxation Methods for RC Circuits

Martin J. Gander[1], Pratik M. Kumbhar[1], and Albert E. Ruehli[2]

1 Introduction

Classical Waveform Relaxation (WR) was introduced in 1981 for circuit solver applications [5]. In WR, large systems of differential equations modeling electric circuits are partitioned into small subcircuits, which are then solved separately, and an iteration is used to get better and better approximations to the overall solution of the underlying large circuit. For classical WR, smart partitioning is very important to enhance the convergence rate, while optimized WR uses more effective transmission conditions to enhance the convergence rate, and thus permits also partitioning at less suitable locations in the circuit without negatively affecting the convergence rate. We study here for the first time the influence of overlapping subcircuits in classical and optimized WR methods applied to RC circuits.

2 The RC Circuit Equations

Circuit equations are obtained from a given circuit using Modified Nodal Analysis (MNA), a major invention that led for circuits to a similar assembly procedure like the finite element method [4]. The MNA circuit equations for the RC circuit of length N shown in Figure 1 are

$$\dot{\mathbf{v}} = \begin{bmatrix} b_1 & c_1 & & & \\ a_1 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{N-2} & b_{N-1} & c_{N-1} \\ & & & a_{N-1} & b_N \end{bmatrix} \mathbf{v} + \mathbf{f}, \tag{1}$$

[1] Section de Mathématiques, Université de Genève, Switzerland, e-mail: martin.gander@unige.ch, pratik.kumbhar@unige.ch ·[2] EMC Laboratory, Missouri University of Science And Technology, U.S, e-mail: albert.ruehli@gmail.com.

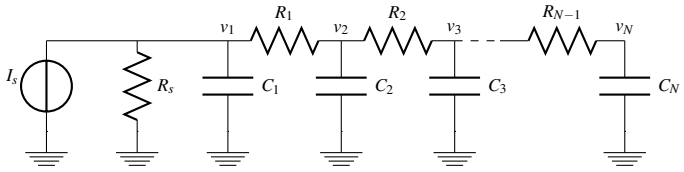


Fig. 1: Finite RC circuit of length N.

where the entries in the tridiagonal matrix are given by

$$\begin{cases} a_i = \frac{1}{R_i C_{i+1}}, \\ c_i = \frac{1}{R_i C_i}, \end{cases} \quad i = 1,2,..,N-1, \quad b_i = \begin{cases} -(\frac{1}{R_s} + \frac{1}{R_1})\frac{1}{C_1}, & i = 1, \\ -(\frac{1}{R_{i-1}} + \frac{1}{R_i})\frac{1}{C_i}, & i = 2,3,...,N-1, \\ -\frac{1}{R_{N-1}C_N}, & i = N. \end{cases}$$

The resistances R_i and capacitances C_i are strictly positive constants. The source term on the right-hand side is given by $\mathbf{f}(t)=(I_s(t)/C_1, 0, ...0)^T$ for some current function $I_s(t)$, and we need to specify initial voltage values $\mathbf{v}(0) = (v_1^0, v_2^0, .., v_N^0)^T$ at time $t = 0$ to solve this system.

3 The Classical WR Algorithm

To define the classical WR algorithm, we partition the circuit in Figure 1 with the voltages $\mathbf{v}$ to be determined into two sub-circuits with unknown voltages $\mathbf{u}$ and $\mathbf{w}$. For convenience in the analysis that will follow, we assume N to be even, and we renumber the nodes: instead of using the numbering from 1 to N, we use the numbering from $-\frac{N}{2}+1$ to $\frac{N}{2}$, see Figure 2. We thus have $\mathbf{v} := (v_{-\frac{N}{2}+1}, ..., v_{-1}, v_0, v_1, ...v_{N/2})^T$, which is still of length N, and

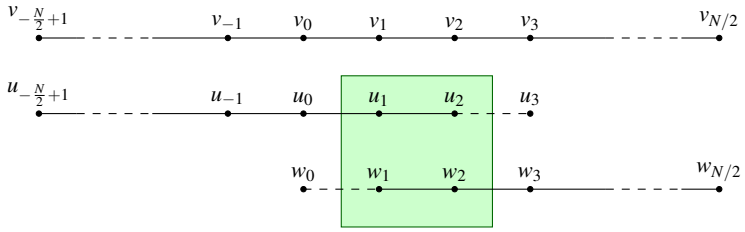


Fig. 2: Decomposition into two sub-circuits with two nodes overlap.

$$\mathbf{u} := (u_{-\frac{N}{2}+1}, ..., u_{n-2}, u_{n-1}, u_n)^T, \quad u_j = v_j \text{ for } j = -\tfrac{N}{2}+1, \ldots, n,$$
$$\mathbf{w} := (w_1, w_2, ..., w_{\frac{N}{2}})^T, \quad w_j = v_j \text{ for } j = 1, \ldots, \tfrac{N}{2},$$

which are of length $\frac{N}{2}+n$ and $\frac{N}{2}$, since we added n nodes to subcircuit $\mathbf{u}$ to have an overlap of n nodes. The classical WR algorithm applied to the two sub-systems is

$$\begin{aligned}
\dot{\mathbf{u}}^{k+1} &= \begin{bmatrix} b_{-\frac{N}{2}+1} & c_{-\frac{N}{2}+1} & & \\ \ddots & \ddots & \ddots & \\ & a_{n-2} & b_{n-1} & c_{n-1} \\ & & a_{n-1} & b_n \end{bmatrix} \begin{bmatrix} u_{-\frac{N}{2}+1} \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix}^{k+1} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ c_n u_{n+1}^{k+1} \end{bmatrix} + \begin{bmatrix} f_{-\frac{N}{2}+1} \\ \vdots \\ f_{n-1} \\ f_n \end{bmatrix}, \\
\dot{\mathbf{w}}^{k+1} &= \begin{bmatrix} b_1 & c_1 & & \\ a_1 & b_2 & c_2 & \\ & \ddots & \ddots & \ddots \\ & & a_{\frac{N}{2}-1} & b_{\frac{N}{2}} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{\frac{N}{2}} \end{bmatrix}^{k+1} + \begin{bmatrix} a_0 w_0^{k+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{\frac{N}{2}} \end{bmatrix},
\end{aligned} \tag{2}$$

where u_{n+1}^{k+1} and w_0^{k+1} are determined in classical WR by the transmission conditions

$$u_{n+1}^{k+1} = w_{n+1}^k \quad \text{and} \quad w_0^{k+1} = u_0^k. \tag{3}$$

Note that in these transmission conditions, we exchange voltages at the interfaces. The two subsystems are given the initial voltages $\mathbf{u}(0) = (v^0_{-\frac{N}{2}+1}, ..., v^0_{n-1}, v^0_n)^T$ and $\mathbf{w}(0) = (v_1^0, v_2^0, .., v^0_{\frac{N}{2}})^T$, and the initial waveforms u_0^0, w_{n+1}^0 are needed to start the WR algorithm.

To simplify our analysis of the convergence factor, we assume that all resistors and capacitors are the same, $R := R_i$ and $C := C_i$ for all $i \in \mathbb{Z}$, which implies

$$b := b_i \quad \text{and} \quad a := a_i = c_i \quad \text{for all } i \in \mathbb{Z}, \tag{4}$$

and for our RC circuit $b = -2a$. To further simplify the analysis, we also assume that the circuit is of infinite length, $N \to \infty$, and by linearity it suffices to analyze the homogeneous problem corresponding to the error equations, and to study convergence to the zero solution. Taking a Laplace transform in time with Laplace parameter $s \in \mathbb{C}$ of the WR algorithm (2), we get in the homogeneous case when $N \to \infty$

$$\begin{aligned}
s\hat{\mathbf{u}}^{k+1} &= \begin{bmatrix} \ddots & \ddots & \ddots & \\ & a & b & a \\ & & a & b \end{bmatrix} \begin{bmatrix} \vdots \\ \hat{u}_{n-1} \\ \hat{u}_n \end{bmatrix}^{k+1} + \begin{bmatrix} \vdots \\ 0 \\ a\hat{w}_{n+1}^k \end{bmatrix}, \\
s\hat{\mathbf{w}}^{k+1} &= \begin{bmatrix} b & a & & \\ a & b & a & \\ & \ddots & \ddots & \ddots \end{bmatrix} \begin{bmatrix} \hat{w}_1 \\ \hat{w}_2 \\ \vdots \end{bmatrix}^{k+1} + \begin{bmatrix} a\hat{u}_0^k \\ 0 \\ \vdots \end{bmatrix}.
\end{aligned} \tag{5}$$

Lemma 1. *Let $a > 0$, $b < 0$, $i = \sqrt{-1}$, and $s := \sigma + i\omega$, with $\sigma > 0$. If $-b \geq 2a$, then the roots $\lambda_{1,2} := \frac{s-b \pm \sqrt{(b-s)^2 - 4a^2}}{2a}$ of the characteristic equation $a\hat{u}_{j-1}^{k+1} + (b-s)\hat{u}_j^{k+1} + a\hat{u}_{j+1}^{k+1} = 0$ of the subsystems in (5) satisfy $|\lambda_2| < 1 < |\lambda_1|$.*

Proof. Since $a > 0$, $b < 0$ and $-b \geq 2a$, we can write $b = -(2+\varepsilon)a$ for some $\varepsilon \geq 0$. Let $p + iq := \sqrt{(b-s)^2 - 4a^2}$, for $p, q \in \mathbb{R}$, with $p > 0$. We then obtain with $\sigma > 0$ that

$$\begin{aligned} |\lambda_1| &= |\frac{s-b+\sqrt{(b-s)^2-4a^2}}{2a}| = |\frac{\sigma + i\cdot\omega + (2+\varepsilon)a}{2a} + \frac{1}{2a}(p + i\cdot q)| \\ &= |(1 + \frac{\varepsilon a + \sigma + p}{2a} + \frac{i}{2a}(\omega + q)| > 1. \end{aligned}$$

Now by Vieta's formulas, $\lambda_1\lambda_2 = 1$, which implies $|\lambda_2| < 1$ and thus completes the proof.

Theorem 1 (Convergence factor for Classical WR with Overlap). *The convergence factor of the classical WR algorithm* (5) *with n nodes overlap is*

$$\rho_{cla}(s,a,b) = \Big(\frac{1}{\lambda_1^2}\Big)^{n+1}. \tag{6}$$

Proof. The iterate $\mathbf{u}^{k+1}$ for the first subsystem satisfies the recurrence relation

$$a\,\hat{u}_{j-1}^{k+1} + (b-s)\,\hat{u}_j^{k+1} + a\,\hat{u}_{j+1}^{k+1} = 0 \qquad \text{for } j = \ldots, n-2, n-1, n, \tag{7}$$

whose solution is $\hat{u}_j^{k+1} = A^{k+1}\lambda_1^j + B^{k+1}\lambda_2^j$ for $j = \ldots, n-2, n-1, n$. Since the solution u_j^{k+1} must remain bounded for all j, we must have $B^{k+1} = 0$. Substituting $j = n$ into (7), we can determine A^{k+1} and obtain the general solution

$$\hat{u}_j^{k+1} = \Big(-\frac{a}{a\lambda_1^{-1} + (b-s)}\Big)\cdot\Big(\frac{1}{\lambda_1^n}\Big)\cdot\lambda_1^j\cdot\hat{w}_{n+1}^k \qquad \text{for } j = \ldots, n-2, n-1, n. \tag{8}$$

Similarly, we obtain for the second subsystem

$$\hat{w}_j^{k+1} = \Big(\frac{-a}{(b-s) + a\lambda_2}\Big)\cdot\lambda_2^{j-1}\cdot\hat{u}_0^k \qquad \text{for } j = 1, 2, \ldots. \tag{9}$$

Combining (8) and (9) and using Vieta's formulas $\lambda_1 + \lambda_2 = \frac{s-b}{a}$ and $\lambda_1\lambda_2 = 1$ then gives

$$\begin{aligned} \hat{u}_j^{k+1} &= \Big(\tfrac{-a}{a\lambda_1^{-1}+(b-s)}\Big)\cdot\Big(\tfrac{-a}{(b-s)+a\lambda_2}\Big)\cdot\Big(\tfrac{\lambda_2^n}{\lambda_1^n}\Big)\cdot\lambda_1^j\hat{u}_0^{k-1} \\ &= \Big(\tfrac{1}{\lambda_1^2}\Big)^{n+1}\hat{u}_j^{k-1} =: \rho_{cla}(s,a,b)\hat{u}_j^{k-1}, \end{aligned}$$

and similarly we find also for the second subsystem $\hat{w}_j^{k+1} = \rho_{cla}(s,a,b)\hat{w}_j^{k-1}$, which concludes the proof.

We see that the convergence factor $\rho_{cla}(s,a,b)$ is the same for all nodes in both subsystems, and since $|\lambda_1| > 1$, classical WR always converges, and convergence becomes faster when increasing the number of nodes the subsystems overlap. In the case $|b| = 2a$ however, $|\rho_{cla}(s,a,b)| \to 1$ when $s \to 0$, which indicates slow convergence for this case.

Remark 1. Theorem 1 implies $\hat{u}_j^{2k} = (\rho_{cla}(s,a,b))^k \hat{u}_j^0$ and $\hat{w}_j^{2k} = (\rho_{cla}(s,a,b))^k \hat{w}_j^0$. Using the Parseval-Plancherel identity, one can then obtain in the time domain

$$\|u_j^{2k}(t)\|_\sigma \leq \Big(\sup_{\omega\in\mathbb{R}} \rho_{cla}(s,a,b)\Big)^k \|u_j^0(t)\|_\sigma,\ \|w_j^{2k}(t)\|_\sigma \leq \Big(\sup_{\omega\in\mathbb{R}} \rho_{cla}(s,a,b)\Big)^k \|w_j^0(t)\|_\sigma,$$

where $\|x(t)\|_\sigma := \|e^{-\sigma t}x(t)\|_{L^2}$. For $\sigma = 0$, we thus obtain convergence in L^2.

4 The Optimized WR Algorithm

New transmission conditions were proposed in [1] for WR, namely

$$\begin{aligned} (u_{n+1}^{k+1} - u_n^{k+1}) + \alpha u_{n+1}^{k+1} &= (w_{n+1}^k - w_n^k) + \alpha w_{n+1}^k, \\ (w_1^{k+1} - w_0^{k+1}) + \beta w_0^{k+1} &= (u_1^k - u_0^k) + \beta u_0^k, \end{aligned} \tag{10}$$

where α and β are weighting factors that can be optimized to obtain more rapid convergence, leading to optimized waveform relaxation algorithms (OWR). If we divide the first equation in (10) by α and the second by β, we see that α and β represent resistances, and the new transmission conditions thus exchange both voltages and currents at the interfaces. Note also that the classical transmission conditions (3) become a special case when taking very large values of α and β.

Theorem 2 (Convergence factor for OWR with Overlap). *The convergence factor of the OWR algorithm with n nodes overlap is*

$$\rho_{opt}(s,a,b,\alpha,\beta) = \Big(\frac{1}{\lambda_1^2}\Big)^n \cdot \Big(\frac{\alpha+1-\lambda_1}{\lambda_1(1+\alpha)-1}\Big).\Big(\frac{\lambda_1+\beta-1}{1+(\beta-1)\lambda_1}\Big). \tag{11}$$

Proof. The transmission conditions (10) can we rewritten as

$$u_{n+1}^{k+1} = \frac{u_n^{k+1}}{1+\alpha} + w_{n+1}^k - \frac{w_n^k}{1+\alpha}, \quad w_0^{k+1} = -\frac{w_1^{k+1}}{\beta-1} + u_0^k + \frac{u_1^k}{\beta-1}.$$

Proceeding with these values as in the proof of Theorem 1 then leads to (11).

We see that OWR contains an extra term in its convergence factor, compared to classical WR, and with a good choice of α and β this term can be made smaller than one and thus leads to better convergence. To obtain the best possible convergence, we need to solve the min-max problem

$$\min_{\alpha,\beta}\Big(\max_{s}|\rho_{opt}(s,a,b,\alpha,\beta)|\Big). \tag{12}$$

To simplify this min-max problem in the complex plane, the following Lemma is useful:

Lemma 2. *Let $b<0$, $a>0$, $-b\geq 2a$, $\alpha>0$ and $\beta<0$. Then the convergence factor $\rho_{opt}(s,a,b,\alpha,\beta)$ is an analytic function in the right half of the complex plane.*

Proof. We need to show that the denominator of $\rho_{opt}(s,a,b,\alpha,\beta)$ does not have any zeros in the right half of the complex plane. We show this by contradiction. Assume there is a zero. Then $\lambda_1=0$ or $(1+\alpha)\lambda_1-1=0$ or $1+(\beta-1)\lambda_1=0$. The first case is not possible since under the given assumptions $|\lambda_1|>1$. Considering the second case we have $\lambda_1=\frac{1}{1+\alpha}$. Since $\alpha>0$, $|\lambda_1|=|\frac{1}{1+\alpha}|<1$ which is a contradiction. Similarly, the third case can not hold since $\beta<0$, which concludes the proof.

Since $\rho_{opt}(s,a,b,\alpha,\beta)$ is analytic in the right half of the complex plane, i.e for $s=\sigma+i\omega$, $\sigma\geq 0$, by the maximum principle for analytic functions, its maximum in modulus is attained on the boundary. Let $s=r\cdot e^{i\theta}$, where $r\in[0,\infty)$and $\theta\in[-\pi/2,\pi/2]$. From the definition of λ_1 given in Lemma 1, we observe that $\lim_{r\to\infty}\lambda_1=\infty$ and hence $\lim_{r\to\infty}\rho_{opt}(s,a,b,\alpha,\beta)=\lim_{r\to\infty}\left(\frac{-1}{(\alpha+1)(\beta-1)}\right)\cdot\left(\frac{1}{\lambda_1^2}\right)^n=0$. Thus the maximum lies on the boundary when $\theta=\pm\pi/2$ and $r<\infty$, i.e. when $\sigma=0$. For $\sigma=0$, one can show that $|\rho_{opt}(\omega,a,b,\alpha,\beta)|$ is symmetric in ω, and hence it is sufficient to optimize the convergence factor for $\omega\geq 0$. To simplify the min-max problem further, we use the fact that in our RC circuit, both sub-systems have very similar electrical properties. Since we assumed furthermore that all circuit elements have the same value, it makes sense to choose $\beta=-\alpha$, which can be interpreted as having the same current flow between the subsystems, just into opposite directions. Therefore, the min-max problem (12) simplifies to

$$\min_{\alpha}\Big(\max_{\omega\geq 0}|\rho_{opt}(\omega,a,b,\alpha)|\Big),\quad \rho_{opt}(\omega,a,b,\alpha)=\Big(\frac{\alpha+1-\lambda_1}{\lambda_1(1+\alpha)-1}\Big)^2\cdot\Big(\frac{1}{\lambda_1^2}\Big)^n. \tag{13}$$

Theorem 3 (Asymptotically optimized α). *For an RC circuit of infinite length with $b=-(2+\varepsilon)a$, where $\varepsilon\to 0$, the optimized parameter α^* for n nodes overlap is*

$$\alpha^*=\Big(\frac{\varepsilon}{n}\Big)^{1/3}. \tag{14}$$

Proof. This result can be proved using asymptotic analysis: one can show that the solution to the min-max problem (13) is given by equioscillation when $\varepsilon\to 0$, i.e α^* satisfies $|\rho_{opt}(\bar{\omega},a,b,\alpha^*)|=|\rho_{opt}(0,a,b,\alpha^*)|$ and $\frac{\partial}{\partial\omega}\rho_{opt}(\bar{\omega},a,b,\alpha^*)=0$ for some interior maximum point $\bar{\omega}>0$. The details are however too long and technical for this short paper, and will appear in [2].

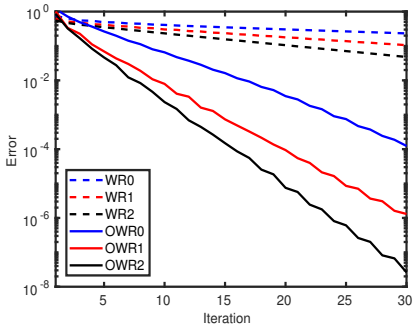


Fig. 3: Convergence for long time $T = 1000$.

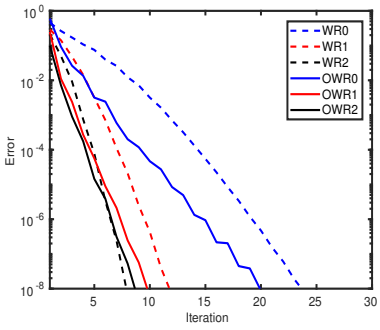


Fig. 4: Convergence for short time $T = 2$.

5 Numerical Results

We simulate an RC circuit of length $N = 80$ with $R = 0.5\,k\Omega$, $C = 0.63\mu F$, $a = \frac{1}{RC}$ and $b = -(2+\varepsilon)a$. We apply Backward Euler with $\Delta t = 0.1$, and simulate directly the error equations, starting with a random initial guess. In Figure 3, we show for $\varepsilon = 10^{-4}$ the influence of overlap on the convergence of classical and optimized WR (e.g. WR2 means WR with overlap 2) for a long time interval $(0,T)$, $T = 1000$. We see that OWR converges much faster than classical WR, see also Figure 5 for a theoretical comparison of the convergence factors. For a short time interval, $T = 2$, classical WR is already very fast, see Figure 4. We determined the optimal choice of α for these experiments solving the min-max problem (13) numerically. Next, we compare this min-max approach with the asymptotic optimization for $b = -(2+\varepsilon)a$ from Theorem 3, and also with running the algorithm for many choices of α numerically. Figure 6 shows that all three give similar results. Finally, we show in Figure 7 and 8 a comparison of the convergence factors for the differently optimized α for two choices of ε.

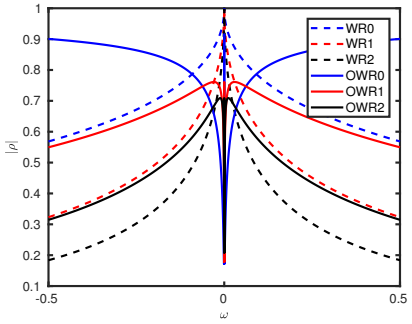


Fig. 5: Effect of overlap

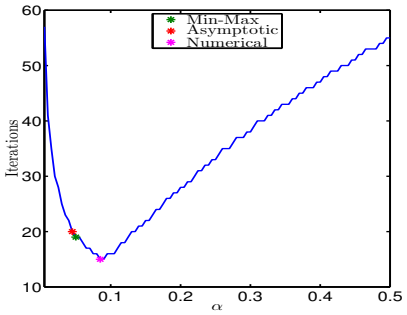


Fig. 6: Comparison of the optimized α

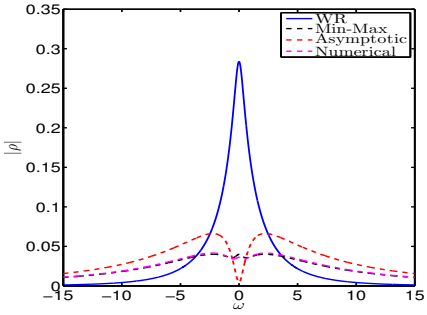


Fig. 7: Convergence factor for optimized α by different methods for $\varepsilon = 10^{-1}$.

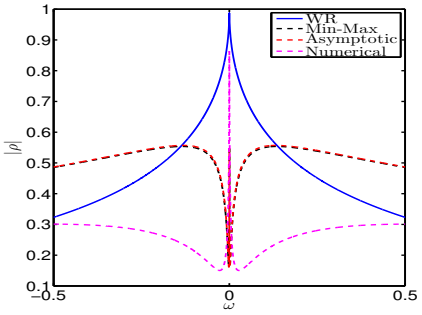


Fig. 8: Convergence factor for optimized α by different methods for $\varepsilon = 10^{-5}$.

6 Conclusion

We studied here for the first time the influence of overlap on the convergence of classical and optimized waveform relaxation algorithms for RC circuits. We defined an optimization problem which permits to obtain a theoretically optimized parameter leading to the fastest possible convergence of the optimized variant. Our analysis shows that overlap enhances the performance of both algorithm variants, which we also illustrated by numerical experiments. While the optimized variant converges much faster when used on long time intervals compared to the classical one, for short time intervals the optimization is less important. We finally compared numerically three different approaches to obtain the optimized parameter in the transmission conditions, and observed that the three methods give similar parameters.

References

1. Al-Khaleel, M.D., Gander, M.J., Ruehli, A.E.: Optimization of transmission conditions in waveform relaxation techniques for RC circuits. SIAM Journal on Numerical Analysis **52**(2), 1076–1101 (2014)
2. Gander, M.J., Kumbhar, P.M.: Asymptotic analysis of optimized waveform relaxation methods for RC type circuits. in preparation (2017)
3. Gander, M.J., Ruehli, A.E.: Optimized waveform relaxation methods for RC type circuits. IEEE Transactions on Circuits and Systems I: Regular Papers **51**(4), 755–768 (2004)
4. Ho, C.W., Ruehli, A., Brennan, P.: The modified nodal approach to network analysis. IEEE Transactions on Circuits and Systems **22**(6), 504–509 (1975)
5. Lelarasmee, E., Ruehli, A.E., Sangiovanni-Vincentelli, A.L.: The waveform relaxation method for time-domain analysis of large scale integrated circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **1**(3), 131–145 (1982)
6. Menkad, T., Dounavis, A.: Resistive coupling-based waveform relaxation algorithm for analysis of interconnect circuits. IEEE Transactions on Circuits and Systems I: Regular Papers **PP**(99), 1–14 (2017)

Convergence of Substructuring Methods for Elliptic Optimal Control Problems

Martin J. Gander[1], Felix Kwok[2], and Bankim C. Mandal[3]

1 Introduction

We are interested in an Optimal Control Problem (OCP) where the constraint is given by an elliptic partial differential equation (PDE):

$$\begin{aligned} -\nabla \cdot (\kappa(\boldsymbol{x})\nabla y(\boldsymbol{x})) &= u(\boldsymbol{x}) \quad && \boldsymbol{x} \in \Omega, \\ y(\boldsymbol{x}) &= 0 && \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{1}$$

The goal is to choose a control variable u from an admissible set U_{ad} to minimize the discrepancy between the solution and the desired state $\hat{y}(\boldsymbol{x})$, i.e. to minimize the objective functional

$$J(y,u) = \frac{1}{2}\int_\Omega |y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x})|^2 d\boldsymbol{x} + \frac{\lambda}{2}\int_\Omega |u(\boldsymbol{x})|^2 d\boldsymbol{x}. \tag{2}$$

We formulate and analyze substructuring algorithms for the model elliptic OCP (1)–(2), which originates from the optimal stationary heating example with controlled heat source, on a bounded domain $\Omega \subset \mathbb{R}^d$. In our setting, y denotes the temperature at a particular point, $\kappa(\boldsymbol{x})$ is the thermal conductivity of Ω, and $\lambda > 0$ is a regularization parameter. We assume $u, \hat{y} \in L^2(\Omega)$ to ensure a solution of the problem. For simplicity, we consider $U_{\text{ad}} = L^2(\Omega)$ as the set of all feasible controls. Then from the first-order optimality conditions (cf. [8]), we obtain the adjoint equation corresponding to the problem (1)–(2)

$$\begin{aligned} -\nabla \cdot (\kappa(\boldsymbol{x})\nabla p(\boldsymbol{x})) &= y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x}) \quad && \boldsymbol{x} \in \Omega, \\ p(\boldsymbol{x}) &= 0 && \boldsymbol{x} \in \partial\Omega, \end{aligned} \tag{3}$$

Department of Mathematics, University of Geneva, Switzerland. e-mail: martin.gander@unige.ch · Department of Mathematics, Hong Kong Baptist University. e-mail: felix_kwok@hkbu.edu.hk · School of Basic Sciences, Indian Institute of Technology Bhubaneswar. e-mail: bmandal@iitbbs.ac.in

together with the optimality condition

$$p(\boldsymbol{x}) + \lambda u(\boldsymbol{x}) = 0. \tag{4}$$

We apply Domain Decomposition (DD) methods, more specifically substructuring methods to solve the state and corresponding adjoint equations. For similar applications of substructuring methods to solve linear-quadratic elliptic OCPs, see [6]. Although our techniques can be extended to multiple subdomains, we only consider a decomposition into two non-overlapping subdomains for the sake of simplicity and compact presentation. For further details on DD methods applied to OCPs, see [1, 2]. We analyze the convergence of Dirichlet-Neumann (DN) [3] and Neumann-Neumann (NN) [4] DD methods for the underlying elliptic PDEs (1)–(3). For more details on DN and NN methods, see [7]. By linearity it suffices to consider the homogeneous problems, $\hat{y}(\boldsymbol{x}) = 0$, and to analyze convergence to zero, since the corresponding error equations coincide with these homogeneous equations.

2 Dirichlet-Neumann algorithm

We first apply the Dirichlet-Neumann algorithm to solve the PDEs (1) and (3), coupled through the condition (4). Suppose the domain Ω is decomposed into two non-overlapping subdomains, Ω_1 and Ω_2. We denote by y_i, u_i, p_i the restriction of y, u, p to Ω_i, and by n_i the unit outward normal for Ω_i on the interface $\Gamma := \partial\Omega_1 \cap \partial\Omega_2$. Then given two initial guesses $h_y^0(\boldsymbol{x})$ and $h_p^0(\boldsymbol{x})$ along the interface Γ, we write the DN algorithm for both state and adjoint equations (we do not write explicitly the homogeneous boundary conditions on the outer boundaries satisfied by the iterates): for $k = 1, 2, \ldots$ compute

$$\begin{aligned} -\nabla\cdot\big(\kappa(\boldsymbol{x})\nabla y_1^k\big) &= u_1^k \quad \text{in } \Omega_1, & -\nabla\cdot\big(\kappa(\boldsymbol{x})\nabla p_1^k\big) &= y_1^k \quad \text{in } \Omega_1, \\ y_1^k &= h_y^{k-1} \quad \text{on } \Gamma, & p_1^k &= h_p^{k-1} \quad \text{on } \Gamma, \end{aligned} \tag{5}$$

$$\begin{aligned} -\nabla\cdot\big(\kappa(\boldsymbol{x})\nabla y_2^k\big) &= u_2^k \quad \text{in } \Omega_2, & -\nabla\cdot\big(\kappa(\boldsymbol{x})\nabla p_2^k\big) &= y_2^k \quad \text{in } \Omega_2, \\ \partial_{n_2} y_2^k &= -\partial_{n_1} y_1^k \quad \text{on } \Gamma, & \partial_{n_2} p_2^k &= -\partial_{n_1} p_1^k \quad \text{on } \Gamma, \end{aligned} \tag{6}$$

together with the update conditions:

$$h_y^k(\boldsymbol{x}) = \theta_y y_2^k\,|_\Gamma + (1-\theta_y) h_y^{k-1}(\boldsymbol{x}),\ h_p^k(\boldsymbol{x}) = \theta_p p_2^k\,|_\Gamma + (1-\theta_p) h_p^{k-1}(\boldsymbol{x}), \tag{7}$$

where θ_y, θ_p are two relaxation parameters, one for the state variable and another for the adjoint variable. Note that the adjoint problem in (5) can be derived from the first order stationarity conditions for the modified objective function

$$J_1(y,u) = \frac{1}{2}\int_{\Omega_1} |y-\hat{y}|^2\,d\boldsymbol{x} + \frac{\lambda}{2}\int_{\Omega_1} |u|^2\,d\boldsymbol{x} - \int_\Gamma \kappa\frac{\partial y}{\partial n}\cdot h_p^{k-1}\,dS(\boldsymbol{x}).$$

The adjoint system for (6) can be interpreted similarly.

We analyze the convergence of the DN algorithm (5)-(6)-(7) for the 1d case with $\Omega_1 = (0, \alpha), \Omega_2 = (\alpha, 1)$ and $\kappa(x) = 1$. By the condition (4), we write $u_i^k = -p_i^k/\lambda$ for $i = 1, 2$. We denote by $D^{(m)} := \frac{d^m}{dx^m}$. Eliminating p_1^k, p_2^k from (5)–(6), we obtain

$$\begin{aligned} D^{(4)}y_1^k + \tfrac{1}{\lambda}y_1^k = 0, \qquad & D^{(4)}y_2^k + \tfrac{1}{\lambda}y_2^k = 0, \\ y_1^k(\alpha) = h_y^{k-1}, \qquad & D^{(1)}y_2^k(\alpha) = D^{(1)}y_1^k(\alpha), \\ D^{(2)}y_1^k(\alpha) = \tfrac{h_p^{k-1}}{\lambda}, \qquad & D^{(3)}y_2^k(\alpha) = D^{(3)}y_1^k(\alpha), \end{aligned} \tag{8}$$

with the homogenous boundary conditions $y_1^k(0) = 0$, $D^{(2)}y_1^k(0) = 0$, $y_2^k(1) = 0$, and $D^{(2)}y_2^k(1) = 0$ at the outer boundaries. Since $\lambda > 0$, we set $\mu^4 := 1/\lambda$. To simplify notation later, we set

$$\gamma_1 = \cosh\left(\tfrac{\mu\alpha}{\sqrt{2}}\right), \gamma_2 = \cosh\left(\tfrac{\mu(1-\alpha)}{\sqrt{2}}\right), \sigma_1 = \sinh\left(\tfrac{\mu\alpha}{\sqrt{2}}\right), \sigma_2 = \sinh\left(\tfrac{\mu(1-\alpha)}{\sqrt{2}}\right),$$
$$\eta_1 = \cos\left(\tfrac{\mu\alpha}{\sqrt{2}}\right), \quad \eta_2 = \cos\left(\tfrac{\mu(1-\alpha)}{\sqrt{2}}\right), \quad \rho_1 = \sin\left(\tfrac{\mu\alpha}{\sqrt{2}}\right), \quad \rho_2 = \sin\left(\tfrac{\mu(1-\alpha)}{\sqrt{2}}\right).$$

Then the general solution of (8) becomes

$$y_1^k(x) = A\sinh\left(\frac{\mu x}{\sqrt{2}}\right)\cos\left(\frac{\mu x}{\sqrt{2}}\right) + B\cosh\left(\frac{\mu x}{\sqrt{2}}\right)\sin\left(\frac{\mu x}{\sqrt{2}}\right), \tag{9}$$

where $A = \frac{h_y^{k-1}\sigma_1\eta_1 - \mu^2 h_p^{k-1}\gamma_1\rho_1}{\sigma_1^2+\rho_1^2}, B = \frac{h_y^{k-1}\gamma_1\rho_1 + \mu^2 h_p^{k-1}\sigma_1\eta_1}{\sigma_1^2+\rho_1^2}$, and

$$y_2^k(x) = C\sinh\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right)\cos\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right) + E\cosh\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right)\sin\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right), \tag{10}$$

with

$$\begin{aligned} C &= -A\frac{\sigma_1\sigma_2\rho_1\rho_2 + \gamma_1\gamma_2\eta_1\eta_2}{\eta_2^2+\sigma_2^2} + B\frac{\gamma_1\eta_1\sigma_2\rho_2 - \sigma_1\rho_1\gamma_2\eta_2}{\eta_2^2+\sigma_2^2}, \\ E &= -A\frac{\gamma_1\eta_1\sigma_2\rho_2 - \sigma_1\rho_1\gamma_2\eta_2}{\eta_2^2+\sigma_2^2} - B\frac{\sigma_1\sigma_2\rho_1\rho_2 + \gamma_1\gamma_2\eta_1\eta_2}{\eta_2^2+\sigma_2^2}. \end{aligned}$$

Using (9) and (10), the update conditions (7) are simplified to

$$\begin{aligned} h_y^k &= (1-\theta_y)\,h_y^{k-1} + \theta_y\left(h_y^{k-1}v - \mu^2 h_p^{k-1}w\right), \\ h_p^k &= (1-\theta_p)\,h_p^{k-1} + \theta_p\left(\tfrac{h_y^{k-1}}{\mu^2}w + h_p^{k-1}v\right), \end{aligned} \tag{11}$$

with the two functions

$$v(\alpha,\mu) = -\frac{\rho_1\rho_2\eta_1\eta_2 + \sigma_1\sigma_2\gamma_1\gamma_2}{(\sigma_1^2+\rho_1^2)(\eta_2^2+\sigma_2^2)}, \quad w(\alpha,\mu) = \frac{\gamma_1\sigma_1\rho_2\eta_2 - \rho_1\eta_1\gamma_2\sigma_2}{(\sigma_1^2+\rho_1^2)(\eta_2^2+\sigma_2^2)}, \tag{12}$$

and we obtain the following convergence results.

Theorem 1 (Convergence in the symmetric case). *For symmetric subdomains, $\alpha = 1/2$ in (5)-(6)-(7), the DN algorithm for the coupled PDEs converges linearly for $0 < \theta_y, \theta_p < 1$, $\theta_y \neq 1/2, \theta_p \neq 1/2$. For $\theta_y = 1/2 = \theta_p$, it converges in two iterations. Convergence is independent of the value of λ.*

Proof. For $\alpha = 1/2$, $v(\alpha, \mu) = -1, w(\alpha, \mu) = 0$. The expressions (11) become

$$h_y^k = (1 - 2\theta_y)\, h_y^{k-1} = (1 - 2\theta_y)^k\, h_y^0,\ h_p^k = (1 - 2\theta_p)\, h_p^{k-1} = (1 - 2\theta_p)^k\, h_p^0.$$

Therefore the convergence is linear for $0 < \theta_y, \theta_p < 1$, $\theta_y \neq 1/2, \theta_p \neq 1/2$. If $\theta_y = 1/2 = \theta_p$, we have $h_y^1 = 0 = h_p^1$, and hence the desired converged solution is achieved after one more iteration.

We now focus on the more interesting asymmetric subdomain case ($\alpha \neq 1/2$).

Theorem 2 (Convergence in the asymmetric case). *Suppose $\alpha \neq 1/2$. Then the DN algorithm (5)-(6)-(7) for the coupled PDEs converges in at most three iterations if and only if (θ_y, θ_p) equals either (Λ^+, Λ^-) or (Λ^-, Λ^+), where*

$$\Lambda^{\pm} := \frac{1}{(1-v)} \pm \frac{|w|}{(1-v)\sqrt{(1-v)^2 + w^2}}. \tag{13}$$

Proof. For $\alpha \neq 1/2$, we set $\bar{h}_p^k := \mu h_p^k, \bar{h}_y^k := \frac{h_y^k}{\mu}$. We rewrite the updating terms (11) in the matrix form

$$\begin{pmatrix} \bar{h}_y^k \\ \bar{h}_p^k \end{pmatrix} = \left[\begin{pmatrix} 1-\theta_y & 0 \\ 0 & 1-\theta_p \end{pmatrix} + \begin{pmatrix} \theta_y v(\alpha,\mu) & -\theta_y w(\alpha,\mu) \\ \theta_p w(\alpha,\mu) & \theta_p v(\alpha,\mu) \end{pmatrix} \right] \begin{pmatrix} \bar{h}_y^{k-1} \\ \bar{h}_p^{k-1} \end{pmatrix}.$$

Note that the matrix of the system on the right side (which we call S) is never zero for any particular set of values θ_y, θ_p. So we do not get two-step convergence for $\alpha \neq 1/2$, unlike in Theorem 1. We claim that there is some positive integer n, for which $S^n = 0$. This results in

$$\begin{pmatrix} \bar{h}_y^n \\ \bar{h}_p^n \end{pmatrix} = S^n \begin{pmatrix} \bar{h}_y^0 \\ \bar{h}_p^0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

so that the DN algorithm converges in $n+1$ iterations. The spectral radius of S is

$$\Upsilon(\theta_y, \theta_p, \alpha, \mu) := \max\left\{ \left| 1 - \frac{1}{2}(\theta_y + \theta_p)(1-v) \pm \frac{1}{2}\sqrt{(\theta_y - \theta_p)^2 (1-v)^2 - 4\theta_y \theta_p w^2} \right| \right\}.$$

For each $\alpha \in (0,1)$ and $\mu > 0$, we solve the system

$$1 - \frac{1}{2}(\theta_y + \theta_p)(1-v) = 0, \quad (\theta_y - \theta_p)^2 (1-v)^2 - 4\theta_y \theta_p w^2 = 0 \tag{14}$$

simultaneously for θ_y, θ_p to obtain (Λ^+, Λ^-), as in equation (13). Υ being symmetric with respect to θ_y, θ_p, (Λ^-, Λ^+) is also a solution of the system

(14). Therefore $\Upsilon\left(\Lambda^{\pm}, \Lambda^{\mp}, \alpha, \mu\right) = 0$, resulting in $S^2 = 0$ and hence three step convergence to the exact solution. For any other values of (θ_y, θ_p), the spectral radius of S is non-zero, so the algorithm cannot converge to the exact solution in a finite number of iterations.

Remark 1. Since $v(\alpha, \mu) \leq 0$ (which can be seen from (12) by noting that $\gamma_i \geq |\eta_i|, \sigma_i \geq |\rho_i|$ for all α, μ), equation (13) implies that $\Lambda^- \in (0, 1)$ and $\Lambda^+ \in (0, 2)$. Note that unlike the symmetric case $\alpha = 1/2$, it is possible to have convergence for $\theta_y > 1$; for $\alpha = 0.99$ and $\mu = \sqrt{8}$, convergence in three steps occurs for $(\theta_y, \theta_p) = (1.000685490, 0.9621364448)$.

Remark 2. For a symmetric decomposition of a rectangular domain in 2D into two equal subdomains, it can be shown that $\Lambda^{\pm} = 0.5$ still gives two-step convergence in the DN method. For an asymmetric decomposition, however, the optimal values may be different, see the last example in Section 4.

3 Neumann-Neumann algorithm

To write the NN algorithm for both state and adjoint equations (1)-(3), we again divide Ω into two non-overlapping subdomains, Ω_1 and Ω_2. We use the same notations as in Section 2. Given two initial guesses $g_y^0(\boldsymbol{x})$ and $g_p^0(\boldsymbol{x})$ along the interface Γ, the NN algorithm is (again we do not write explicitly the homogeneous boundary conditions on the outer boundaries satisfied by the iterates): for $k = 1, 2, \ldots$ compute the approximations

$$\begin{aligned} -\nabla \cdot \left(\kappa(\boldsymbol{x}) \nabla y_i^k\right) &= u_i^k && \text{in } \Omega_i, \\ y_i^k &= g_y^{k-1} && \text{on } \Gamma, \end{aligned} \tag{15}$$

followed by the correction step,

$$\begin{aligned} -\nabla \cdot \left(\kappa(\boldsymbol{x}) \nabla \psi_i^k\right) &= 0 && \text{in } \Omega_i, \\ \partial_{n_i} \psi_i^k &= \partial_{n_1} y_1^k + \partial_{n_2} y_2^k && \text{on } \Gamma, \end{aligned} \tag{16}$$

and similarly for the adjoint equation, we compute

$$\begin{aligned} -\nabla \cdot \left(\kappa(\boldsymbol{x}) \nabla p_i^k\right) &= y_i^k && \text{in } \Omega_i, \\ p_i^k &= g_p^{k-1} && \text{on } \Gamma, \end{aligned} \tag{17}$$

followed by the correction step,

$$\begin{aligned} -\nabla \cdot \left(\kappa(\boldsymbol{x}) \nabla \varphi_i^k\right) &= 0 && \text{in } \Omega_i, \\ \partial_{n_i} \varphi_i^k &= \partial_{n_1} p_1^k + \partial_{n_2} p_2^k && \text{on } \Gamma. \end{aligned} \tag{18}$$

The update conditions for g_y^k and g_p^k are

$$g_y^k(\boldsymbol{x}) = g_y^{k-1}(\boldsymbol{x}) - \theta_y \left(\psi_1^k|_\Gamma + \psi_2^k|_\Gamma\right),$$
$$g_p^k(\boldsymbol{x}) = g_p^{k-1}(\boldsymbol{x}) - \theta_p \left(\varphi_1^k|_\Gamma + \varphi_2^k|_\Gamma\right). \tag{19}$$

We again analyze the convergence for the NN algorithm (15)–(19) for $\Omega_1 = (0,\alpha), \Omega_2 = (\alpha,1)$ and $\kappa(x) = 1$. By (4), we have $u_i^k = -p_i^k/\lambda$ for $i = 1, 2$. Eliminating p_1^k, p_2^k from (15)-(17), we obtain

$$\begin{aligned} D^{(4)}y_1^k + \tfrac{1}{\lambda}y_1^k = 0, \qquad & D^{(4)}y_2^k + \tfrac{1}{\lambda}y_2^k = 0, \\ y_1^k(\alpha) = g_y^{k-1}, \qquad & y_2^k(\alpha) = g_y^{k-1}, \\ D^{(2)}y_1^k(\alpha) = \tfrac{g_p^{k-1}}{\lambda}, \qquad & D^{(2)}y_2^k(\alpha) = \tfrac{g_p^{k-1}}{\lambda}, \end{aligned} \tag{20}$$

with the homogenous boundary conditions $y_1^k(0) = 0$, $D^{(2)}y_1^k(0) = 0$, $y_2^k(1) = 0$, and $D^{(2)}y_2^k(1) = 0$ at the outer boundaries. With $\mu^4 := 1/\lambda$, the solutions of (20) become

$$y_1^k(x) = E_1 \sinh\left(\tfrac{\mu x}{\sqrt{2}}\right)\cos\left(\tfrac{\mu x}{\sqrt{2}}\right) + E_2 \cosh\left(\tfrac{\mu x}{\sqrt{2}}\right)\sin\left(\tfrac{\mu x}{\sqrt{2}}\right),$$
$$y_2^k(x) = F_1 \sinh\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right)\cos\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right) + F_2 \cosh\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right)\sin\left(\tfrac{\mu(1-x)}{\sqrt{2}}\right),$$

where

$$E_1 = \frac{g_y^{k-1}\sigma_1\eta_1 - \mu^2 g_p^{k-1}\gamma_1\rho_1}{\sigma_1^2+\rho_1^2}, \quad E_2 = \frac{g_y^{k-1}\gamma_1\rho_1 + \mu^2 g_p^{k-1}\sigma_1\eta_1}{\sigma_1^2+\rho_1^2},$$

$$F_1 = \frac{g_y^{k-1}\sigma_2\eta_2 - \mu^2 g_p^{k-1}\gamma_2\rho_2}{\sigma_2^2+\rho_2^2}, \quad F_2 = \frac{g_y^{k-1}\gamma_2\rho_2 + \mu^2 g_p^{k-1}\sigma_2\eta_2}{\sigma_2^2+\rho_2^2}.$$

Finally solving ψ_i^k, φ_i^k in (16)-(18) and replacing them in (19) we get the updating terms

$$\begin{aligned} g_y^k &= g_y^{k-1} - \theta_y(g_y^{k-1}z_1 + \mu^2 g_p^{k-1}z_2), \\ g_p^k &= g_p^{k-1} - \theta_p(g_p^{k-1}z_1 - \tfrac{1}{\mu^2}g_y^{k-1}z_2), \end{aligned} \tag{21}$$

with the functions $z_1(\alpha,\mu) = \frac{\mu}{\sqrt{2}}\left(\frac{\sigma_1\gamma_1+\rho_1\eta_1}{\sigma_1^2+\rho_1^2} + \frac{\sigma_2\gamma_2+\rho_2\eta_2}{\sigma_2^2+\rho_2^2}\right)$, and $z_2(\alpha,\mu) = \frac{\mu}{\sqrt{2}}\left(\frac{\sigma_1\gamma_1-\rho_1\eta_1}{\sigma_1^2+\rho_1^2} + \frac{\sigma_2\gamma_2-\rho_2\eta_2}{\sigma_2^2+\rho_2^2}\right)$.

Theorem 3 (Convergence of the NN algorithm). *The NN algorithm for the coupled PDEs* (15)–(19) *converges in at most three iterations if* (θ_y, θ_p) *is any of the pairs* $(\Theta^+, \Theta^-), (\Theta^-, \Theta^+)$, *where* $\Theta^\pm := \frac{1}{z_1} \pm \frac{|z_2|}{z_1\sqrt{z_1^2+z_2^2}}$.

Proof. Setting $\bar{g}_p^k := \mu g_p^k, \bar{g}_y^k := \frac{g_y^k}{\mu}$, we rewrite the updating terms (21) as:

$$\begin{pmatrix} \bar{g}_y^k \\ \bar{g}_p^k \end{pmatrix} = \begin{pmatrix} 1-\theta_y z_1 & -\theta_y z_2 \\ \theta_p z_2 & 1-\theta_p z_1 \end{pmatrix} \begin{pmatrix} \bar{g}_y^{k-1} \\ \bar{g}_p^{k-1} \end{pmatrix}.$$

The matrix on the right side (we call P) is never zero for any set of values θ_y, θ_p. But like in the DN method, if we have $P^n = 0$, for some n, then we get

$$\begin{pmatrix} \bar{g}_y^n \\ \bar{g}_p^n \end{pmatrix} = P^n \begin{pmatrix} \bar{g}_y^0 \\ \bar{g}_p^0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

resulting in convergence in $n+1$ iterations. The spectral radius of P is: $\Phi(\theta_y, \theta_p, \alpha, \mu) := \max\left\{\left|1 - \frac{1}{2}(\theta_y + \theta_p) z_1 \pm \frac{1}{2}\sqrt{(\theta_y - \theta_p)^2 z_1^2 - 4\theta_y\theta_p z_2^2}\right|\right\}$. We solve the system $1 - \frac{1}{2}(\theta_y + \theta_p) z_1 = 0, (\theta_y - \theta_p)^2 z_1^2 - 4\theta_y\theta_p z_2^2 = 0$ simultaneously for each $\alpha \in (0,1)$ and $\mu > 0$ to obtain a solution (Θ^+, Θ^-) as in the Theorem. Due to the symmetric nature of Φ with respect to θ_y, θ_p, (Θ^-, Θ^+) is another solution pair of the system of equations. Thus $\Phi(\Theta^\pm, \Theta^\mp, \alpha, \mu) = 0$, resulting in $P^2 = 0$ and therefore three step convergence to the exact solution.

4 Numerical Examples

We perform numerical experiments to verify the convergence rate of the DN and NN algorithms for the model problem (1)–(2) with $\lambda = 1/2$, $\hat{y}(x) = 0$. In the top two plots of Figure 1, we observe two-step convergence of the DN method for $\alpha = 1/2$ on the left, and three-step convergence for $\alpha = 0.6$ for the optimal choice of $(\Lambda^+, \Lambda^-) = (0.62, 0.57)$ on the right. The two bottom plots of Figure 1 show the convergence behavior of the DN algorithm for different choices of θ_y and θ_p. On the left panel, we get $\theta_y = \theta_p = 1/2$ to be the best parameters for the symmetric case, whereas on the right (Λ^+, Λ^-) yields the fastest convergence for $\alpha = 0.6$. For the NN experiment, we plot on the left panel of Figure 2 the first three iterates of the state variable for the optimal choice of $(\Theta^+, \Theta^-) = (0.30, 0.16)$, and on the right the convergence curves for various values of the parameters θ_y, θ_p. In Figure 3, we show convergence of the DN and NN methods for the 2D problem:

$$\begin{aligned} -\Delta y(\boldsymbol{x}) &= u(\boldsymbol{x}) \quad && \boldsymbol{x} \in \Omega = (0,1)^2, \\ y(\boldsymbol{x}) &= 0 && \boldsymbol{x} \in \partial\Omega, \end{aligned}$$

with an interface $\Gamma = \{0.6\} \times (0,1)$ and $\lambda = 1/2$. Note that the optimal parameters are different from the 1d case when the decomposition is non-symmetric, as the choice of $(0.5, 0.5)$ appears to perform better than (Λ^+, Λ^-) in the DN example. A full analysis of the 2D case will be the subject of a future paper. We are also working on the analysis of the case of multiple subdomains, where it is not clear if one can choose relaxation parameters to obtain finite termination of the algorithm; see [5] for the uncontrolled case.

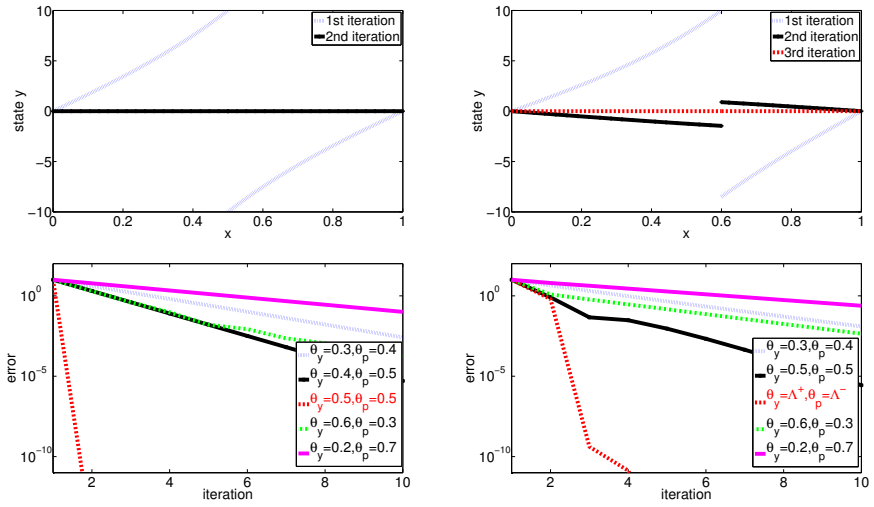


Fig. 1 Convergence of the iterative solution of the DN method: in two iterations for the symmetric case on the top left, and in three iterations for $\alpha = 0.6$ on the top right; error curves for various values of θ_y, θ_p for $\alpha = 1/2$ on the bottom left, and for $\alpha = 0.6$ on the bottom right

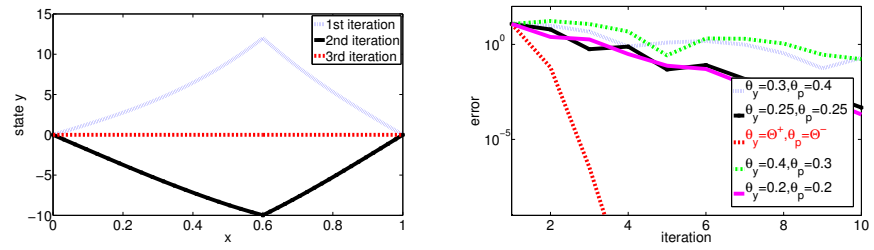


Fig. 2 Convergence of NN: convergence of the iterative solutions with optimal parameters in three iterations on the left, and convergence for various values of θ_y, θ_p on the right

References

1. Benamou, J.D.: A domain decomposition method with coupled transmission conditions for the optimal control of systems governed by elliptic partial differential equations. SIAM J. Numer. Anal. **33**(6), 2401–2416 (1996)
2. Bensoussan, A., Glowinski, R., Lions, J.: Méthode de décomposition appliquée au contrôle optimal de systèmes distribués. In: 5th Conference on Optimization Techniques Part I, pp. 141–151. Springer (1973)
3. Bjørstad, P., Widlund, O.: Iterative Methods for the Solution of Elliptic Problems on Regions Partitioned into Substructures. SIAM J. Numer. Anal. **23**(6), 1097–1120 (1986)
4. Bourgat, J.F., Glowinski, R., Tallec, P.L., Vidrascu, M.: Variational Formulation and Algorithm for Trace Operator in Domain Decomposition Calculations. In: Domain decomposition methods, pp. 3–16. SIAM, Philadelphia, PA (1989)

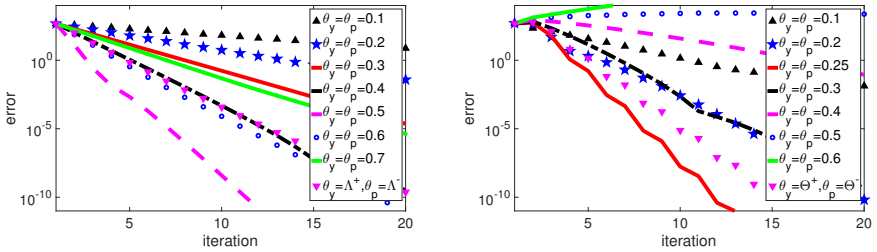


Fig. 3 Convergence in 2d: convergence of DN method in 2d on the left, and NN method in 2d on the right. $\Lambda^{\pm}, \Theta^{\pm}$ correspond to 1d optimal parameters

5. Chaouqui, F., Gander, M., Santugini-Repiquet, K.: On nilpotent subdomain iterations. In: Domain Decomposition Methods in Science and Engineering XXIII, LNCSE, pp. 113–120. Springer-Verlag (2016)
6. Heinkenschloss, M., Nguyen, H.: Neumann-Neumann domain decomposition preconditioners for linear-quadratic elliptic optimal control problems. SIAM J. Sci. Comput. **28**(3), 1001–1028 (2006)
7. Toselli, A., Widlund, O.B.: Domain Decomposition Methods, Algorithms and Theory, vol. 34. Springer-Verlag, Berlin (2005)
8. Tröltzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods and Applications, *Graduate Studies in Mathematics*, vol. 112. Amer. Math. Soc. (2010)

Complete, Optimal and Optimized Coarse Spaces for Additive Schwarz

Martin J. Gander[1] and Bo Song[2]

1 Introduction and Model Problem

Coarse spaces are needed to achieve scalability in domain decomposition methods, see [16] and references therein. More recently, new coarse corrections were also designed to improve convergence, for example in high contrast problems. Such enriched coarse spaces were first proposed in [4, 5], where volume eigenfunctions were combined with different types of partition of unity functions, and further developed in [3]. A coarse space using the eigenfunctions of the Dirichlet-to-Neumann maps on the boundary of each subdomain has been proposed and analyzed in [13, 2], and further development led to solving a generalized eigenvalue problem in the overlap (GenEO), see [14, 15].

A new, different idea is to first define an optimal coarse space, which leads to the best possible convergence and makes the method nilpotent [6][1] and then to approximate it [8, 9, 10, 11]. Following this principle, we design here for the first time an optimal coarse space for the additive Schwarz (AS) method with arbitrary sized overlaps, and then define an optimized approximation using a specific function in the overlap, combined with harmonic extensions of interface eigenfunctions. We compare our new coarse space to the GenEO coarse space [14, 15] and the local spectral multiscale coarse space (also with reduced energy) in [4].

Determining an optimal coarse space and then approximating it is a very general idea, see for example the SHEM coarse space [10, 11], but for simplicity we consider here

$$\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{1}$$

[1] Université de Genève, Section de Mathématiques, 2-4 rue du Lièvre, CP 64, CH-1211, Genève, Suisse, e-mail: martin.gander@unige.ch [2] Corresponding author. School of Science, Northwestern Polytechnical University, Xi'an 710072, China, e-mail: bosong@nwpu.edu.cn

[1] Classical one level domain decomposition methods can even be nilpotent in certain situations, see [1].

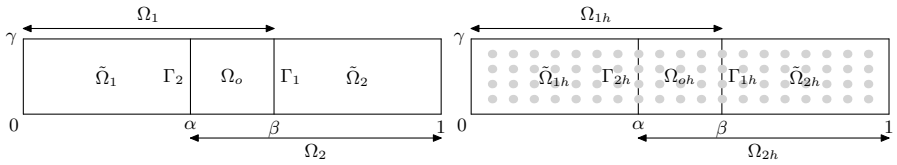


Fig. 1 Continuous (left) and discrete (right) partition of Ω into two overlapping subdomains.

where $\Omega := (0,1)\times(0,\gamma)$ is decomposed into two overlapping subdomains $\Omega_1 := (0,\beta)\times(0,\gamma)$ and $\Omega_2 := (\alpha,1)\times(0,\gamma)$, with overlap $\Omega_o := \Omega_1 \cap \Omega_2$, and interfaces $\Gamma_1 := \{(x,y)|x=\beta, 0<y<\gamma\}$ and $\Gamma_2 := \{(x,y)|x=\alpha, 0<y<\gamma\}$, which leads to the partition of the domain $\bar{\Omega} = \bar{\tilde{\Omega}}_1 \cup \bar{\Omega}_o \cup \bar{\tilde{\Omega}}_2$; see Figure 1.

Discretizing (1) by the classical five-point finite difference scheme, we obtain the linear system $A\mathbf{u} = \mathbf{f}$. Starting with an initial guess $\mathbf{u}^0$, the iterative two level additive Schwarz method with multiplicative (hybrid) coarse grid correction computes

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

where R_i are rectangular restriction matrices corresponding to Ω_i, $A_i = R_i A R_i^T$, $i = 1,2$, and R_c is a restriction matrix to a coarse space, $A_c = R_c A R_c^T$.

2 Complete, Optimal and Optimized Coarse Spaces

Definition 1 (Complete coarse space). A complete coarse space for the additive Schwarz method (2) is given by R_c such that (2) converges after one iteration for an arbitrary initial guess $\mathbf{u}^0$, i.e. the method is nilpotent and becomes a direct solver.

To give an example of a complete coarse space, and being able to write discrete problems using the same notation as continuous ones, we denote by Δ_h the discretized Laplacian, and by Ω_h, Ω_{ih}, $\tilde{\Omega}_{ih}$ Ω_{oh}, Γ_{ih} the corresponding discretized spaces, $i = 1,2$. Let N_{Γ_i} be the number of degrees of freedom (DOFs) on the interface Γ_{ih}, $i = 1,2$, and let $\phi_{i,\mathrm{cs}}^j$ be defined for each DOF on Γ_{ih} by harmonic extension,

$$\begin{aligned} \Delta_h \phi_{i,\mathrm{cs}}^j &= 0 \quad \text{in } \tilde{\Omega}_{ih}, \\ \phi_{i,\mathrm{cs}}^j &= 1 \quad \text{at DOF } j \text{ of } \Gamma_{3-i,h},\ j = 1,\ldots,N_{\Gamma_{3-i}}, \\ \phi_{i,\mathrm{cs}}^j &= 0 \quad \text{elsewhere in } \Omega_h. \end{aligned} \tag{3}$$

Denoting by N_o the number of DOFs in the overlap Ω_{oh}, we define for each of them the further basis function $\phi_{o,\mathrm{cs}}^j = 1$, extended by zero to the rest of Ω_h, $j = 1,\ldots,N_o$, and

$$V_{0,\mathrm{cs}} := \mathrm{span}\{\{\{\phi_{i,\mathrm{cs}}^j\}_{j=1}^{N_{\Gamma_{3-i}}}\}_{i=1}^2 \cup \{\phi_{o,\mathrm{cs}}^j\}_{j=1}^{N_o}\}. \tag{4}$$

Theorem 1. *A complete coarse space for the iterative two level additive Schwarz algorithm* (2) *is given by R_c containing in its columns the vectors of $V_{0,cs}$ from* (4).

Proof. The proof is technical, see [12], but the property is illustrated in Section 4.

The dimension of this complete coarse space depends on the number of DOFs in the overlap for AS which was designed to be symmetric [7], but it can be reduced[2].

Definition 2 (Optimal coarse space). An optimal coarse space for (2) is a complete coarse space such that its associated R_c has the smallest number of columns possible.

For an optimal coarse space, we define the restriction matrix in the overlap, $R_o := [0 \quad I_{\Omega_{oh}} \quad 0]$, where I_{Ω_o} is the identity matrix whose dimension equals the number of unknowns in Ω_{oh}, and the associated local solver in the overlap, $A_o := R_o A R_o^T$. We then construct *just one* specific basis function ϕ_o in the overlap Ω_{oh} for (2), based on the initial guess $\mathbf{u}^0$ by solving

$$A_o \phi_o = R_o(\mathbf{f} - A\mathbf{u}^0),$$

and then extending ϕ_o with zero to the rest of Ω_h. We also need the basis functions $\phi^j_{i,\text{opt}}$, $j = 1,\ldots,N_{\Gamma_{3-i}}, i = 1,2$, based on harmonic extensions,

$$\begin{aligned}
\Delta_h \phi^j_{i,\text{opt}} &= 0 \quad \text{in } \tilde{\Omega}_{ih}, & \Delta_h \phi^j_{i,\text{opt}} &= 0 \quad \text{in } \Omega_{oh},\\
\phi^j_{i,\text{opt}} &= 1 \quad \text{at DOF } j \text{ of } \Gamma_{3-i,h}, & \phi^j_{i,\text{opt}} &= 1 \quad \text{at DOF } j \text{ of } \Gamma_{3-i,h},\\
\phi^j_{i,\text{opt}} &= 0 \quad \text{elsewhere in } \Omega_h, & \phi^j_{i,\text{opt}} &= 0 \quad \text{elsewhere in } \Omega_h,
\end{aligned}$$

and define

$$V_{0,\text{opt}} := \text{span}\{\{\{\phi^j_{i,\text{opt}}\}_{j=1}^{N_{\Gamma_{3-i}}}\}_{i=1}^{2} \cup \{\phi_o\}\}. \tag{5}$$

Theorem 2. *An optimal coarse space for the iterative two level additive Schwarz algorithm* (2) *is given by R_c containing in its columns the vectors of $V_{0,opt}$ from* (5).

Proof. The proof is given in [12], but the property is again illustrated in Section 4.

At the continuous level, even the optimal coarse space would still be infinite dimensional, and we thus introduce now an approximation of the optimal coarse space based on SHEM (Spectral Harmonically Enriched Multiscale coarse space [10]) using an interface eigenvalue problem:

Definition 3 (Interface eigenvalue problem). Denoting by D_{yy} an approximation of the second derivative along the interface Γ_i, the interface eigenvalue problem is

$$-D_{yy}\psi_i = \lambda \psi_i \quad \text{on } \Gamma_{ih}, \tag{6}$$

with zero Dirichlet boundary conditions $\psi_i(0) = \psi_i(\gamma) = 0$, $i = 1,2$.

[2] This problem does not arise with overlap of one or two mesh sizes [11], or RAS [10].

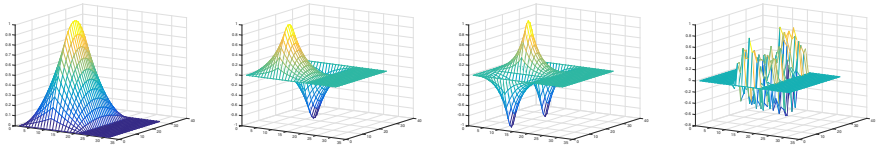


Fig. 2 First 3 basis functions used to approximate the optimal coarse space based on the interface eigenfunctions, and the single mode in the overlap for a random initial guess on the right.

In our example, the eigenvectors of the interface eigenvalue problem (6) are $\psi_i^j = \sin((j\pi/\gamma)y_m)$, $y_m = mh$. We can thus construct basis functions $\phi_{i,\text{app}}^j$ by the harmonic extensions of the sine functions for $i = 1,2$,

$$\begin{aligned} \Delta_h \phi_{i,\text{app}}^j &= 0 \quad \text{in } \tilde{\Omega}_{ih}, & \Delta_h \phi_{i,\text{app}}^j &= 0 \quad \text{in } \Omega_{oh}, \\ \phi_{i,\text{app}}^j &= \psi_{3-i}^j \quad \text{on } \Gamma_{3-i,h}, & \phi_{i,\text{app}}^j &= \psi_{3-i}^j \quad \text{on } \Gamma_{3-i,h} \\ \phi_{i,\text{app}}^j &= 0 \quad \text{elsewhere in } \Omega_h, & \phi_{i,\text{app}}^j &= 0 \quad \text{elsewhere in } \Omega_h, \end{aligned} \tag{7}$$

$j = 1,\ldots,\ell$, where ℓ is the number of the eigenvectors of the interface eigenvalue problem (6) selected; see Figure 2 for an illustration. We then define an optimized approximation of the optimal coarse space

$$V_{0,\text{cs-1}} = \text{span}\{\{\{\phi_{i,\text{app}}^j\}_{j=1}^{\ell}\}_{i=1}^{2} \cup \{\phi_o\}\}. \tag{8}$$

Theorem 3. *The iterative two level additive Schwarz algorithm* (2) *with R_c containing in its columns the vectors of $V_{0,cs\text{-}l}$ in* (8) *satisfies the error estimate*

$$\|\mathbf{u}^n - \mathbf{u}\|_{\infty,2} \leq \left(\frac{\cosh(\frac{(\ell+1)\pi}{\gamma}(\alpha+\beta-1)) - \cosh(\frac{(\ell+1)\pi}{\gamma}(\beta-\alpha-1))}{\cosh(\frac{(\ell+1)\pi}{\gamma}(\alpha+\beta-1)) - \cosh(\frac{(\ell+1)\pi}{\gamma}(\beta-\alpha+1))} \right)^{n/2} \|\mathbf{u}^0 - \mathbf{u}\|_{\infty,2},$$

and there is no other coarse space of this dimension that leads to faster convergence.

Proof. The proof can be obtained by a direct calculation using separation of variables for the residual after one additive Schwarz iteration, and will be given in [12].

3 Comparison to Two Other Coarse Spaces

We now compare our optimized coarse space to the GenEO coarse space from [14, 15], and the local spectral multiscale coarse space with reduced energy from [4]. The GenEO coarse space was designed for high contrast problems and is based on generalized eigenproblems "in the overlap": in our example it solves in Ω_i, $i = 1,2$,

$$\hat{A}_i \mathbf{p}_i^j = \lambda_i^j X_i \hat{A}_i^o X_i \mathbf{p}_i^j \tag{9}$$

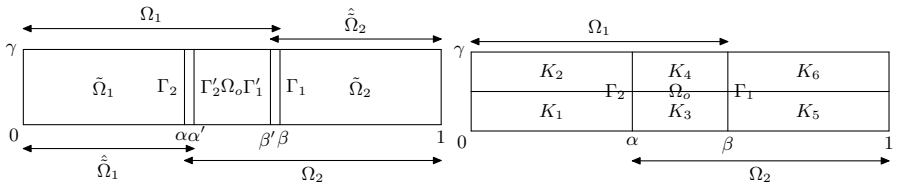


Fig. 3 Left: partition of the domain for the GenEO coarse space. Right: partition of the domain for the local multiscale coarse space with reduced energy.

for eigenvectors $\mathbf{p}_i^j \in \mathbb{R}^{\#\overline{\text{dof}}(\Omega_i)}$ assoicated with small eigenvalues $\lambda_i^j \in \mathbb{R}\bigcup\{+\infty\}$. In (9), X_j is a diagonal matrix indicating the partition of unity used to combine subdomain solutions, and $\hat{A}_i$, $\hat{A}_i^o$ are Neumann matrices for each subdomain. Selecting the ℓ eigenfunctions corresponding to the smallest eigenvalues then leads to the GenEO coarse space

$$V_{0,\text{GenEO}} = \text{span}\{R_i^T X_i \mathbf{p}_i^j, j = 1,\ldots,\ell,\ i = 1,2\}.$$

To understand how GenEO is related to our optimized coarse space, we first rewrite the eigenvalue problem (9) at the continuous level for $\lambda_i \neq 0$ and $\lambda_i \neq 4$,

$$\begin{aligned} \Delta \hat{p}_i(x,y) &= 0 \quad \text{in } \hat{\tilde{\Omega}}_i, & \Delta p_{io}(x,y) &= 0 \quad \text{in } \Omega_o, \\ \hat{p}_i &= p_{io} \quad \text{on } \hat{\Gamma}_{3-i}, & p_{io} &= \frac{4}{4-\lambda_i}\hat{p}_i \quad \text{on } \Gamma_{3-i}, \end{aligned} \tag{10}$$

with boundary conditions $\hat{p}_i = 0$ on $\partial\Omega \cap \hat{\bar{\tilde{\Omega}}}_i$, $p_{io} = 0$ on $(\partial\Omega \cap \hat{\bar{\tilde{\Omega}}}_o)\backslash\Gamma_i$ and $\partial_n p_{io} = 0$ on Γ_i, and then define $p_i := \hat{p}_i$ in $\hat{\tilde{\Omega}}_i$, $p_i := p_{io}$ in Ω_o, and $p_i := 0$ in the rest of Ω, $i = 1,2$. Here, Γ_i' are within one mesh size from the corresponding boundary Γ_i, $i = 1,2$, see Figure 3 on the left. Solving (10) with separation of variables for Ω_1, we find for our model problem

$$p_1^j(x,y) = \begin{cases} \dfrac{4}{4-\lambda_1^j}\dfrac{\sinh(\frac{j\pi}{\gamma}x)}{\sinh(\frac{j\pi}{\gamma}\alpha')}\dfrac{\cosh(\frac{j\pi}{\gamma}(\beta-\alpha'))}{\cosh(\frac{j\pi}{\gamma}(\beta-\alpha))}\sin(\frac{j\pi}{\gamma}y), & (x,y) \in (0,\alpha)\times(0,\gamma), \\ \dfrac{4}{4-\lambda_1^j}\dfrac{\cosh(\frac{j\pi}{\gamma}(\beta-x))}{\cosh(\frac{j\pi}{\gamma}(\beta-\alpha))}\sin(\frac{j\pi}{\gamma}y), & (x,y) \in (\alpha,\beta)\times(0,\gamma), \end{cases}$$

$$\lambda_1^j = 4 - 4\frac{\sinh(\frac{j\pi}{\gamma}\alpha)}{\sinh(\frac{j\pi}{\gamma}\alpha')}\frac{\cosh(\frac{j\pi}{\gamma}(\beta-\alpha'))}{\cosh(\frac{j\pi}{\gamma}(\beta-\alpha))}.$$

We show in Figure 4 the three types of GenEO eigenfunctions. The eigenfunctions corresponding to the smallest eigenvalues are like the ones in our optimized coarse space within the subdomains, but in the overlap they differ. Since GenEO uses an eigenvalue problem in the entire subdomain volume, it also contains many more eigenfunctions (which one avoids to compute in GenEO), like the overlap ones for

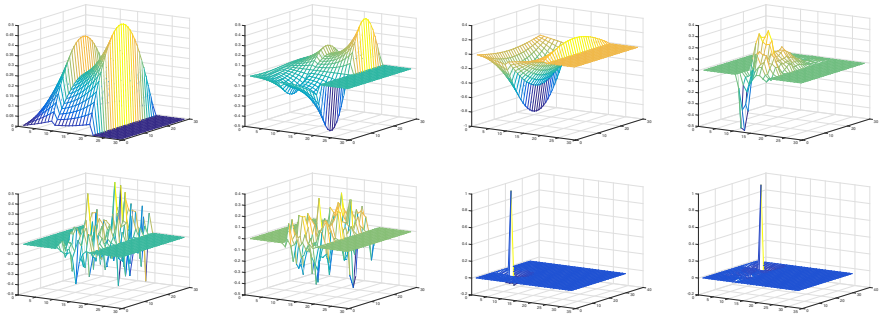


Fig. 4 First 3 basis functions on Ω_1 of the eigenvalue problem from GenEO for the smallest λ, followed by 3 of the randomly looking modes for $\lambda = 4$ and 2 of the modes for $\lambda = \infty$.

$\lambda = 4$ corresponding to $\phi^j_{o,\text{cs}}$ in our complete coarse space, plus the ones for $\lambda = \infty$ which do not contain relevant information for the coarse space.

We next compare our optimized coarse space with the local spectral multiscale coarse space (also with reduced energy) in [4]. The domain Ω is still decomposed into two overlapping subdomains Ω_1 and Ω_2, with six coarse blocks K_i, $i = 1,\ldots,6$, see Figure 3 on the right. Let $\mathbf{q}_i^j$ denote the jth eigenvector of the volume eigenvalue problem in subdomain Ω_{ih}, $i = 1,2$,

$$\begin{aligned} \Delta_h \mathbf{q}_i &= \lambda_i \mathbf{q}_i && \text{in } \Omega_{ih}, \\ \partial_n \mathbf{q}_i &= 0 && \text{on } \Gamma_{ih}, \\ \mathbf{q}_i &= 0 && \text{on } \partial\Omega_{ih}\backslash\Gamma_{ih}. \end{aligned} \tag{11}$$

With the partition of unity χ_i, $i = 1,2$, the local spectral multiscale coarse space of [4] using ℓ functions is defined by

$$V_{0,\text{mul}} = \text{span}\{R_i^T \chi_i \mathbf{q}_i^j, 1 \le j \le \ell, i = 1,2\}. \tag{12}$$

The local spectral multiscale coarse space with reduced energy of [4] is defined by

$$\tilde{V}_{0,\text{mul}} = \text{span}\{R_i^T \tilde{\mathbf{q}}_i^j, 1 \le j \le \ell, i = 1,2\},$$

where for each block $K_h \in \Omega_{ih}$, $i = 1,2$, $1 \le j \le \ell$, one still needs to solve $\Delta_h \tilde{\mathbf{q}}_i^j = 0$ in K_h, $\tilde{\mathbf{q}}_i^j = \chi_i \mathbf{q}_i^j$ on ∂K_h. Solving (11) using separation of variables, we find in Ω_1

$$\mathbf{q}_1^{jk}(x,y) = \sin(\frac{k\pi - \pi/2}{\alpha}x)\sin(\frac{j\pi}{\gamma}y), \quad \lambda_1^{jk} = (\frac{k\pi - \pi/2}{\alpha})^2 + (\frac{j\pi}{\gamma})^2.$$

We show in Figure 5 the first few of those modes. Note that these modes are different from the modes in our optimized coarse space and GenEO, and again one needs to

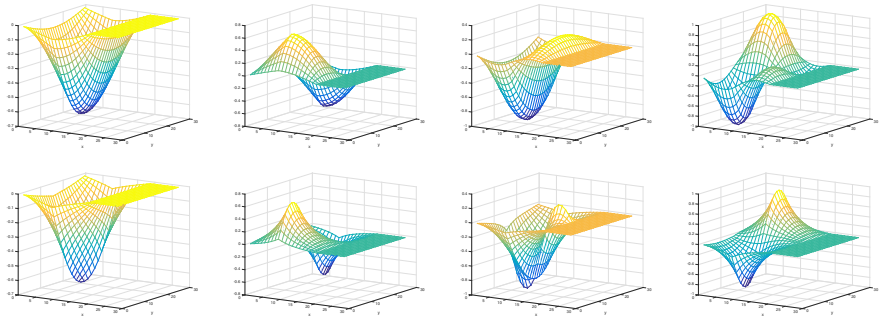


Fig. 5 Top: First 4 basis functions of the local spectral multiscale coarse space $V_{0,\text{mul}}$. Bottom: corresponding modes for the reduced energy case.

solve volume eigenvalue problems to construct the coarse spaces $V_{0,\text{mul}}$ and $\tilde{V}_{0,\text{mul}}$, which now also contain many redundant modes.

4 Numerical experiments

We solve (1) with $f = -3$ on $\Omega = (0,1) \times (0,1)$ discretized by centered finite differences and using an overlap of $4h$, h being the mesh parameter. We start with a random initial guess, and stop the iteration when the error in the iterative method or the residual in PCG reaches the tolerance $1e-8$. In Table 1 we show the dependence of the number of iterations on h for the complete coarse space (CS), the optimal coarse space (CS-opt), our optimized coarse space SHEM(ℓ), and GenEO(ℓ), GalvisI(ℓ) and GalvisII(ℓ) (reduced energy), using $\ell = 3$ enrichment functions for each subdomain. We see that CS and CS-opt are direct solvers, and only SHEM(ℓ) leads to a convergent stationary iteration; and SHEM(ℓ) also performs best with PCG.

In Table 2, we show the iteration numbers for 2×2, 4×4, and 8×8 subdomains using $h = 1/32, 1/64, 1/128$, i.e. keeping H/h fixed. We choose again $\ell = 3$ for

Table 1 Iteration number comparison in the two subdomain case for different mesh parameters h.

	h	AS	CS	CS-opt	SHEM(3)	GenEO(3)	GalvisI(3)	GalvisII(3)
Iterative	$\frac{1}{16}$	na	1	1	6	na	na	na
	$\frac{1}{32}$	na	1	1	11	na	na	na
	$\frac{1}{64}$	na	1	1	20	na	na	na
	$\frac{1}{128}$	na	1	1	38	na	na	na
PCG	$\frac{1}{16}$	9	1	1	4	8	8	7
	$\frac{1}{32}$	11	1	1	6	9	9	9
	$\frac{1}{64}$	14	1	1	9	11	12	11
	$\frac{1}{128}$	17	1	1	12	14	14	14

Table 2 Iteration number comparison for many subdomains.

	Subdomains	AS	CS	CS-opt	SHEM	GenEO	GalvisI	GalvisII
Iterative	2×2	na	1	1	8	na	na	na
	4×4	na	1	1	8	na	na	na
	8×8	na	1	1	8	na	na	na
PCG	2×2	15	1	1	5	13	12	10
	4×4	25	1	1	5	13	14	10
	8×8	42	1	1	5	13	14	10

SHEM, and approximately the same total number of coarse functions for the other coarse spaces. We see again that CS and CS-opt are direct solvers, and only SHEM leads to a convergent stationary method. When used with PCG, the methods are all scalable, but SHEM needs only half the number of iterations compared to the other methods.

References

1. Chaouqui, F., Gander, M.J., Santugini-Repiquet, K.: On nilpotent subdomain iterations. In: Domain Decomposition Methods in Science and Engineering XXIII. Springer (2016)
2. Dolean, V., Nataf, F., Scheichl, R., Spillane, N.: Analysis of a two-level Schwarz method with coarse spaces based on local Dirichlet-to-Neumann maps. Computational Methods in Applied Mathematics **12**(4), 391–414 (2012)
3. Efendiev, Y., Galvis, J., Lazarov, R., Willems, J.: Robust domain decomposition preconditioners for abstract symmetric positive definite bilinear forms. ESAIM: Mathematical Modelling and Numerical Analysis **46**(5), 1175–1199 (2012)
4. Galvis, J., Efendiev, Y.: Domain decomposition preconditioners for multiscale flows in high-contrast media. Multiscale Modeling & Simulation **8**(4), 1461–1483 (2010)
5. Galvis, J., Efendiev, Y.: Domain decomposition preconditioners for multiscale flows in high contrast media: reduced dimension coarse spaces. Multiscale Modeling & Simulation **8**(5), 1621–1644 (2010)
6. Gander, M., Halpern, L.: Méthode de décomposition de domaine. Encyclopédie électronique pour les ingénieurs (2012)
7. Gander, M.J.: Schwarz methods over the course of time. Electron. Trans. Numer. Anal **31**(5), 228–255 (2008)
8. Gander, M.J., Halpern, L., Repiquet, K.S.: Discontinuous coarse spaces for DD-methods with discontinuous iterates. In: Domain Decomposition Methods in Science and Engineering XXI, pp. 607–615. Springer (2014)
9. Gander, M.J., Halpern, L., Repiquet, K.S.: A new coarse grid correction for RAS/AS. In: Domain Decomposition Methods in Science and Engineering XXI, pp. 275–283. Springer (2014)
10. Gander, M.J., Loneland, A.: SHEM: An optimal coarse space for RAS and its multiscale approximation. In: Domain Decomposition Methods in Science and Engineering XXIII. Springer (2016)
11. Gander, M.J., Loneland, A., Rahman, T.: Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods. arXiv preprint arXiv:1512.05285 (2015)
12. Gander, M.J., Song, B.: Complete, optimal and optimized coarse spaces for Additive Schwarz. in preparation (2017)

13. Nataf, F., Xiang, H., Dolean, V., Spillane, N.: A coarse space construction based on local Dirichlet-to-Neumann maps. SIAM Journal on Scientific Computing **33**(4), 1623–1642 (2011)
14. Spillane, N., Dolean, V., Hauret, P., Nataf, F., Pechstein, C., Scheichl, R.: A robust two-level domain decomposition preconditioner for systems of PDEs. Comptes Rendus Mathematique **349**(23), 1255–1259 (2011)
15. Spillane, N., Dolean, V., Hauret, P., Nataf, F., Pechstein, C., Scheichl, R.: Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. Numerische Mathematik **126**(4), 741–770 (2014)
16. Toselli, A., Widlund, O.B.: Domain decomposition methods–algorithms and theory, volume 34 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin (2005)

Heterogeneous Optimized Schwarz Methods for Coupling Helmholtz and Laplace Equations

Martin J. Gander[1] and Tommaso Vanzan[1]

1 Introduction

Optimized Schwarz methods have increasingly drawn attention over the last two decades because of their improvements in terms of robustness and computational cost compared to the classical Schwarz method. Their optimized transmission conditions have been obtained through analytical or numerical procedures in many different situations, involving mostly the same partial differential equation on each subdomain, see [6, 3, 7] and references therein. When dealing with heterogeneous problems, a domain decomposition approach which allows one to exploit different solvers adapted to the different physical problems is important. Due to their favorable convergence properties in the absence of overlap, and their capability to take physical properties at the interfaces into account, optimized Schwarz methods are a natural framework for such heterogeneous domain decomposition methods, where the spatial decomposition is simply provided by the multi-physics of the problem.

We introduce and analyze here heterogeneous optimized Schwarz methods with zeroth order optimized transmission conditions for the coupling between the hard to solve Helmholtz equation [5] and the Laplace equation. It is a simplified instance of the coupling of parabolic and hyperbolic operators, which might arise in Maxwell equations. The Helmholtz equation is used in the time harmonic regime of a wave equation and the Laplace operator represents the parabolic part. We consider a bounded domain $\Omega \subset \mathbb{R}^2$, with sufficiently regular boundary, divided into two subdomains Ω_1 and Ω_2 such that $\Omega = \Omega_1 \cup \Omega_2$, $\Gamma = \overline{\Omega_1} \cap \overline{\Omega_2}$, and $\Sigma_j = \partial\Omega_j \setminus \Gamma$. Our model problem is

$$\begin{aligned} (-\Delta - q\omega^2)u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial n} + i\omega u &= 0 \quad \text{on } \Sigma_1, \end{aligned} \tag{1}$$

[1] Section de mathématiques, Université de Genève, 2-4 rue du Lièvre, Genève, e-mail: {martin.gander}{tommaso.vanzan}@unige.ch.

$$u = 0 \quad \text{on } \Sigma_2,$$

where $\omega > 0$ is the Helmholtz frequency, and $q \in L^\infty(\Omega)$ satisfies $q = 1$ in Ω_1 and $q = 0$ in Ω_2. Since the well-posedness of the problem is not straightforward due to the indefinite nature of the Helmholtz part, we first analyze it in more detail adapting arguments presented by Després in [2].

Lemma 1. *The norm* $||u||^2 = \int_\Omega |\nabla u|^2 + \omega \int_{\Sigma_1} |u|^2$ *is equivalent to the canonical norm on* $H^1(\Omega)$ *if* $|\Sigma_1| > 0$.

Proof. We first observe that $H^1(\Omega)$ is the direct sum of $\bar{V} = \left\{ v \in H^1(\Omega) : \int_\Omega v = 0 \right\}$ and $\tilde{V} = \left\{ v \in H^1(\Omega) : v \text{ is constant in } \Omega \right\}$, $H^1(\Omega) = \tilde{V} \oplus \bar{V}$. Then, on the one hand, it easy to see that for all $v \in \tilde{V}$, there exist a constant $C = \sqrt{\frac{\omega|\Sigma_1|}{|\Omega|}}$ such that

$$C||v||_{H^1(\Omega)} \leq ||v|| \leq C||v||_{H^1(\Omega)}. \tag{2}$$

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

$$||v||^2_{H^1(\Omega)} \leq (1+C) \int_\Omega |\nabla v|^2 \leq (1+C) \left(\int_\Omega |\nabla v|^2 + \omega \int_{\Sigma_1} |v|^2 \right) = (1+C)||v||^2. \tag{3}$$

Exploiting the continuity of the trace operator, we obtain

$$||v||^2 = \int_\Omega |\nabla v|^2 + \omega \int_{\Sigma_1} |v|^2 \leq \int_\Omega |\nabla v|^2 + \omega \int_{\partial\Omega} |v|^2 \leq \max(1, C_{\partial\Omega}\omega)(\int_\Omega |\nabla v|^2 + \int_\Omega |v|^2). \tag{4}$$

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

Let us define $V := \{v \in H^1(\Omega) : v = 0 \text{ on } \Sigma_2\}$, with $||\cdot||_V = ||\cdot||_{H^1(\Omega)}$, and consider problem (1) in the variational form

$$\text{Find } u \in V : \; a(u,v) - b(u,v) =_{V^{-1}} \langle f, v \rangle_V \quad \forall v \in V, \tag{5}$$

where $a(u,v) = \int_\Omega \nabla u \nabla \bar{v} + i\omega \int_{\Sigma_1} u\bar{v}$, $b(u,v) = \omega^2 \int_{\Omega_2} u\bar{v}$ and $f \in V^{-1}$. To use Fredholm theory, we now show that the bilinear form b is a compact pertubation of a.

Lemma 2. *Let* $\mathscr{B}$ *be an operator from* V *to* V *such that*

$$a(\mathscr{B}u, v) = b(u,v) \quad \forall v \in V, \tag{6}$$

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

Proof. We first prove continuity, i.e. $\exists C > 0 : \forall u \in V, ||\mathscr{B}u||_V \leq C||u||_V$. From the definition of $\mathscr{B}$, and applying Lax-Milgram to (6), we have $||\mathscr{B}u||_V \leq \frac{1}{\alpha}||b(u)||_{V^{-1}}$, where $b(u) : V \to \mathbb{R}$ is the functional defined by $_{V^{-1}} < b(u), v >_V := b(u,v)$. Then we have $\forall v \in V$

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

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

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

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

$$\left| \int_\Omega |\nabla(\mathscr{B}u_{n_j} - \mathscr{B}u)|^2 + i\omega \int_{\Sigma_1} |\mathscr{B}u_{n_j} - \mathscr{B}u|^2 \right| \leq \omega^2 ||u_{n_j} - u||_{L^2(\Omega_2)} ||\mathscr{B}u_{n_j} - \mathscr{B}u||_{L^2(\Omega_2)}. \tag{7}$$

We observe that $\mathscr{B}u_{n_j} \rightharpoonup \mathscr{B}u$ in V because $u_{n_j} \rightharpoonup u$ in V and $\mathscr{B}$ is a continuous operator [1]. Hence, both u_{n_j} and $\mathscr{B}u_{n_j}$ converge strongly in $L^2(\Omega)$. In particular we have that $a(\mathscr{B}u_{n_j} - \mathscr{B}u, \mathscr{B}u_{n_j} - \mathscr{B}u) \to 0$ which implies $||\mathscr{B}u_{n_j} - \mathscr{B}u|| \to 0$. With Lemma 1, we have that $\mathscr{B}u_{n_j} \to \mathscr{B}u$ in V and thus $\mathscr{B}$ is a compact operator.

Since $\mathscr{B}$ is a compact operator, thanks to Fredholm alternative, existence of the solution of problem (5) follows from uniqueness. We need two further Lemmas to prove uniqueness. We denote with $\gamma_j u$ and $S_j u$ the trace of u and the trace of the normal derivative on the j-th interface and we introduce the space $E(\Omega, \Delta) := \{u \in H^1(\Omega) : -\Delta u \in L^2(\Omega)\}$.

Lemma 3 (Grisvard, Theorem 1.5.3.11, page 61, [9]). *Let Ω be an open bounded subset of $\mathbb{R}^2$ whose boundary is a curvilinear polygon of class $C^{1,1}$ with interfaces $\Sigma_j, j = 1,..,N$. The mappings $u \to \gamma_j u$ and $u \to S_j u$ have a unique continuous extension from $E(\Omega, \Delta)$ to respectively $H^{\frac{1}{2}}(\Sigma_j)$ and $H^{-\frac{1}{2}}(\Sigma_j)$. Moreover for every $u \in E(\Omega, \Delta)$ and $v \in H^1(\Omega)$ with $\gamma_j v \in H^{\frac{1}{2}}(\Sigma_j)\ \forall j$, the Green's formula holds:*

$$(-\Delta u, v) = (\nabla u, \nabla v) - \sum_{j=1}^{N} \langle S_j u, \overline{\gamma_j v} \rangle. \tag{8}$$

Lemma 4 (Després, Corollary 2.1, page 22, [2]). *Let Ω be an open bounded arc-connected subset of $\mathbb{R}^2$ and assume that Γ is a nonempty open subset of $\partial\Omega$ of class C^2 and $q \in L^\infty(\Omega)$. If $u \in H^2(\Omega)$ satisfies*

$$(-\Delta - q\omega^2)u = 0 \text{ on } \Omega, \quad u|_\Gamma = \partial_n u|_\Gamma = 0, \tag{9}$$

then u=0 in Ω.

Theorem 1. *Under the hypotheses of Lemmas 3 and 4, $u \equiv 0$ is the only solution of the boundary value problem (1) with $f = 0$.*

Proof. Choosing $v \in D(\Omega)$, the space of $C^\infty(\Omega)$ functions with compact support, in the weak formulation of eq. (1) we obtain $-\Delta u - q\omega^2 u = 0$. Hence, since $u \in V$, $\Delta u \in L^2(\Omega)$ and $u \in E(\Omega, \Delta)$. Exploiting Green's formula and choosing $v = u$ we get

$$\int_\Omega |\nabla u|^2 - \omega^2 \int_{\Omega_1} |u|^2 + i\omega \int_{\Sigma_1} |u|^2 = 0. \tag{10}$$

Considering the imaginary part we have $\int_{\Sigma_1} |u|^2 = 0$, which implies $u = 0$ on Σ_1. We now have homogeneous Dirichlet data on the whole domain $\partial\Omega = \Sigma_1 \cup \Sigma_2$. Regularity results for Dirichlet problems in smooth domains state that $u \in H^2(\Omega)$. Exploiting again the Green's formula and $-\Delta u - q\omega^2 u = 0$ in Ω , we obtain

$$_{H^{-\frac{1}{2}}(\Sigma_1)}\langle \frac{\partial u}{\partial n}, v\rangle_{H^{\frac{1}{2}}(\Sigma_1)} + iw \int_{\Sigma_1} uv = 0. \tag{11}$$

Since $u = 0$ on Σ_1, we can conclude that $\partial_n u = 0$ on Σ_1 and by the unique continuation principle in Lemma 4, the result follows.

2 Heterogeneous Optimized Schwarz methods

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

$$\begin{aligned} (-\omega^2 - \Delta)u_1 &= f \ \text{ in } \Omega_1, \quad (\partial_x + S_1)(u_1^n)(0,\cdot) = (\partial_x + S_1)(u_2^{n-1})(0,\cdot), \\ -\Delta u_2 &= f \ \text{ in } \Omega_2, \quad (\partial_x + S_2)(u_2^n)(0,\cdot) = (\partial_x + S_2)(u_1^{n-1})(0,\cdot), \end{aligned} \tag{12}$$

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

$$\begin{aligned} (-\omega^2 - \partial_{xx} + k^2)(\hat{e}_1^n) &= 0 && k \in \mathbb{R},\ x < 0, \\ (\partial_x + \sigma_1(k))(\hat{e}_1^n)(0,k) &= (\partial_x + \sigma_1(k))(\hat{e}_2^{n-1})(0,k), && k \in \mathbb{R}, \\ (-\partial_{xx} + k^2)(\hat{e}_2^n) &= 0 && k \in \mathbb{R},\ x > 0, \\ (\partial_x + \sigma_2(k))(\hat{e}_2^n)(0,k) &= (\partial_x + \sigma_2(k))(\hat{e}_1^{n-1})(0,k), && k \in \mathbb{R}, \end{aligned} \tag{13}$$

where $\sigma_j(k)$ are the Fourier symbols of the operators S_j. Solving the equations in (13) and imposing the radiation/boundedness conditions, we get

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

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

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

where

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

Next, to approximate the optimal choice for $\sigma_1(k)$ and $\sigma_2(k)$ which would require non local operators, we set $\sigma_1 = -\sigma_2 = p(1+i)$. This choice is motivated by [4] where the single and double sided optimizations were studied and compared for the time harmonic Maxwell equations. Since both σ_j and $\lambda(k)$ contain complex numbers, we have to study the modulus of the convergence factor,

$$|\rho(k,p)|^2 = \begin{cases} \dfrac{((k-p)^2+p^2)}{((k+p)^2+p^2)} \dfrac{((\sqrt{k^2-\omega^2}-p)^2+p^2)}{((\sqrt{k^2-\omega^2}+p)^2+p^2)} & k \geq \omega, \\ \dfrac{((k-p)^2+p^2)}{((k+p)^2+p^2)} \dfrac{((\sqrt{\omega^2-k^2}-p)^2+p^2)}{((\sqrt{\omega^2-k^2}+p)^2+p^2)} & k < \omega. \end{cases} \tag{14}$$

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

$$\min_{p\geq 0} \max_{k\in[k_{\min},k_{\max}]} |\rho(k,p)|^2, \tag{15}$$

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

Theorem 2. *Assuming that $k_{\max} > 2\omega$, the solution of the minimax problem (15) is given by $p^* = \frac{\omega}{\sqrt{2}}$ if $|\rho(k_{\max}, p^* = \frac{\omega}{\sqrt{2}})|^2 \leq \frac{(\sqrt{2}-1)^2+1}{(\sqrt{2}+1)^2+1}$, and otherwise it is given by the unique p^* such that $|\rho(k=\omega,p^*)|^2 = |\rho(k_{\max},p^*)|^2$.*

Proof. We consider $p > 0$, because for $p = 0$ the convergence factor is equal to 1, and for $p < 0$ it is greater than one, while for values of $p > 0$, the convergence factor is always less than 1. We introduce a change of variables which will be useful in the computations, namely $x = \sqrt{k^2-\omega^2}$ if $k \geq \omega$ and $x = \sqrt{\omega^2-k^2}$ for $k < \omega$. Problem (15) then becomes

$$\min_{p>0} \max \left(\max_{[0,\sqrt{\omega^2-k_{\min}^2}]} G(x,p), \max_{[0,\sqrt{k_{\max}^2-\omega^2}]} F(x,p) \right), \tag{16}$$

where

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

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

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

$$x(4p^4+x^4)(2p^2+x^2-\omega^2) = ((\omega^2-x^2)^2+4p^2)(2p^2-x^2)\sqrt{\omega^2-x^2}. \quad (17)$$

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

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

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

$$F(\sqrt{k_{\max}^2-\omega^2}, \sqrt{\frac{k_{\max}^2-\omega^2}{2}}) = \left(\frac{(\sqrt{2}-1)^2+1}{(\sqrt{2}+1)^2+1}\right)^2 < F(0,\frac{\omega}{\sqrt{2}}) < F(0,\sqrt{\frac{k_{\max}^2-\omega^2}{2}}), \quad (18)$$

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

Remark 1. It is interesting to note that this problem is different from the ones already studied in the literature, because the convergence factor is immediately bounded

from below: it is not possible to get a better convergence factor than $\rho^2(k,p) = \frac{(\sqrt{2}-1)^2+1}{(\sqrt{2}+1)^2+1}$. We also did not have to exclude the resonnance frequency $k=\omega$ by introducing ω_- and ω_+, as in the Helmholtz case [8]; the optimized Schwarz method can benefit from the heterogeneity, leading to $|\rho(k=\omega,p)|^2 < 1$.

We now present two asymptotic results. First we want to study how our algorithm behaves when we take finer and finer meshes. Let $h \to 0$, h being the mesh size, and suppose that the maximum frequency supported by the numerical grid scales like $k_{\max} = \pi/h \to \infty$.

Theorem 3. *When the physical parameters ω and $k_{\min}$ are fixed, $k_{\max} = \frac{\pi}{h}$ and $h \to 0$, then the solution of problem (15) is given by*

$$p^* = \frac{\sqrt{\omega\pi}}{2} \cdot h^{-1/2} + o(h^{-1/2}), \quad |\rho(k,p^*)|^2 = 1 - \frac{4\sqrt{\omega}}{\sqrt{\pi}} h^{\frac{1}{2}} + o(h^{1/2}). \tag{19}$$

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

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

Theorem 4. *If $k_{\min}$ is fixed, $k_{\max} = \frac{\pi}{h}$, ω goes to infinity and $h = \frac{C_h}{\omega^\gamma}$, with $\gamma > 1$, then the solution of problem (15) is given by*

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

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

3 Numerical experiments

We implemented our heterogeneous optimized Schwarz method on a square domain $\Omega := (-1,1)\times(-1,1)$, with $\Omega_1 := (-1,0)\times(-1,1)$ and $\Omega_2 := (0,1)\times(-1,1)$. We used second order centered finite differences for the interior points and first order approximations for the boundary terms. In Figure 1 on the left, we show the modulus of the solution of problem (1) for $\omega^2 = 50$ and $f = 1$. On the right in Figure 1, we show a comparison between the optimal numerical value p and the theoretical estimation provided by Theorem 2. We see that our simplified analysis

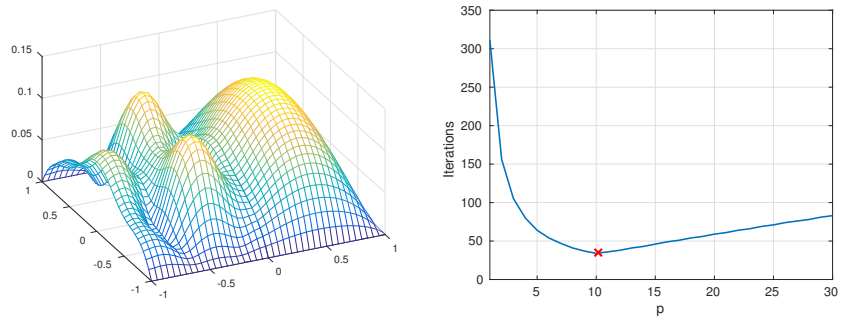


Fig. 1 Parameters $\omega^2 = 50$, $h = 0.05$. Left: Modulus of $u(x,y)$. Right: Parameter p vs number of iterations. The optimal p given by equioscillation is indicated by a star.

on unbounded domains is able to give quite a good approximation of the optimal parameter in the bounded domain context. Finally, we show in Table 1 the behavior of the algorithm when the mesh size h decreases and for large values of ω, with $h\omega^{\frac{3}{2}} = \text{const}$. In brackets, we show the number of iterations required for a non-

h	Optimal p^*	$\max_k \lvert\rho^2(p^*,k)\rvert$	iterations
$\frac{1}{50}$	16.52	0.4225	53 (810)
$\frac{1}{100}$	23.53	0.55043	73 (1614)
$\frac{1}{200}$	33.37	0.6543	104 (3284)
$\frac{1}{400}$	47.27	0.7403	148 (6554)

ω	Optimal p^*	$\max_k \lvert\rho^2(p^*,k)\rvert$	iterations
10π	34.8451	0.2119	31 (839)
20π	84.7084	0.2622	38 (2954)
40π	205.0570	0.3167	46 (8096)
60π	342.6739	0.3506	48 (>10000)

Table 1 The two tables show the behaviour of the heterogeneous optimized Schwarz method under mesh refinement and when ω increases with $h\omega^{\frac{3}{2}}$ held constant.

optimized case, i.e. using $p = 1$. We clearly see that the optimization leads to a much better algorithm, which deteriorates much more slowly when the mesh is refined, and ω increases.

4 Conclusions

We presented and analysed a heterogeneous optimized Schwarz method for the coupling of Helmholtz and Laplace equations. We proved the well-possedness of the coupled problem, and then introduced optimized Robin transmission conditions, giving asymptotic formulas for the optimized parameters and associated convergence factor. Our results indicate that a much weaker dependence on the mesh parameter can be achieved with optimized transmission conditions, and we are currently working on further improvement by studying second order optimized trans-

mission conditions.

Acknowledgements. The authors are grateful to L. Halpern for very useful remarks concerning the well posedness analysis.

References

1. Ciarlet, P.: Linear and Nonlinear Functional Analysis with Applications:. Applied mathematics. Society for Industrial and Applied Mathematics (2013)
2. Després, B.: Méthodes de décomposition de domaine pour les problèmes de propagation d'ondes en régimes harmoniques. Ph.D. thesis, Université Dauphine-Parix IX (1991)
3. Dolean, V., Gander, M.J., Gerardo-Giorda, L.: Optimized Schwarz methods for Maxwell's equations. SIAM Journal on Scientific Computing **31**(3), 2193–2213 (2009)
4. El Bouajaji, M., Dolean, V., Gander, M.J., Lanteri, S.: Comparison of a one and two parameter family of transmission conditions for maxwells equations with damping. In: Domain Decomposition Methods in Science and Engineering XX, pp. 271–278. Springer (2013)
5. Ernst, O.G., Gander, M.J.: Why it is difficult to solve Helmholtz problems with classical iterative methods. In: Numerical analysis of multiscale problems, pp. 325–363. Springer Berlin Heidelberg (2012)
6. Gander, M.J.: Optimized Schwarz methods. SIAM Journal on Numerical Analysis **44**(2), 699–731 (2006)
7. Gander, M.J., Halpern, L., Magoules, F.: An optimized Schwarz method with two-sided Robin transmission conditions for the Helmholtz equation. International journal for numerical methods in fluids **55**(2), 163–175 (2007)
8. Gander, M.J., Magoules, F., Nataf, F.: Optimized Schwarz methods without overlap for the Helmholtz equation. SIAM Journal on Scientific Computing **24**(1), 38–60 (2002)
9. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (1985)

Restrictions on the use of sweeping type preconditioners for Helmholtz problems

Martin J. Gander and Hui Zhang

1 Introduction

Helmholtz problems, and time harmonic problems in general like Maxwell's equations, are notoriously difficult to solve numerically. The first problem is that they require very fine discretizations to avoid the so called pollution effect [1], and then the discretized systems are so large that one needs to solve them iteratively, and none of the classical iterative methods are suitable for this task [10]. Over the past decade, several new ideas arrived for the iterative solution of Helmholtz problems, among them the shifted Laplace preconditioner [9]. Unfortunately in this preconditioner, one has to choose the shift small enough (at most $O(k)$ where k is the wave number) for the preconditioner to be close to the underlying operator to give provable wave number independent convergence [12], and large enough (at least $O(k^2)$) for the preconditioner to be easily invertible by multigrid independently of the wave number [5, 6]. In practice, a compromise has to be chosen, which can lead to a growth of up to $O(k^2)$ in the iteration numbers of preconditioned GMRES in the multigrid case [6]; for a rigorous analysis in the case of classical domain decomposition, see [20]. The best current preconditioners are based on domain decomposition methods using special transmission conditions, and have their roots in optimized Schwarz methods [14, 13] and the AILU preconditioner [15, 16]. These algorithms use transmission conditions adapted to the underlying Helmholtz nature of the problem, and this idea is so important that it has been rediscovered independently several times over the last few years, see the sweeping preconditioner [7, 8], the source transfer method, the methods based on single layer potentials [3, 4], and most recently the method

Martin J. Gander
University of Geneva, Section of Mathematics, Rue du Lievre 2-4, CP 64, 1211 Geneva 4, e-mail: martin.gander@unige.ch

Hui Zhang (corresponding author)
Zhejiang Ocean University, Key Laboratory of Oceanographic Big Data Mining & Application of Zhejiang Province, Zhoushan 316022, China, e-mail: huiz@zjou.edu.cn

of polarized traces [24, 23]. All these methods use the same underlying mathematical algorithm, which at the continuous level is the class of optimal Schwarz methods [11], and at the discrete level the block-LU factorization, and one can prove formally that they are all basically equivalent, see the review monograph [19]. The methods use a one way decomposition of the domain into a sequence of subdomains, and between subdomains they use as transmission condition an approximation of the Dirichlet to Neumann operator. An important technique advocated by these more recently proposed algorithms is the use of perfectly matched layers (PML) in the transmission conditions; for an earlier use of PML transmission conditions in a domain decomposition setting, see [22, 21], and [2] for high order Padé transmission conditions, with [17, 18] for their relation to PML transmission conditions. While one might think intuitively that the absorption at the interfaces is the most important property, and with PML one can reach as much absorption as one wants, the truly important property for the algorithm is not absorption, but approximation of the Dirichlet to Neumann operator, which is well known from optimized Schwarz theory [11]. For a constant wave number, these two coincide, and it was therefore possible to prove for the above methods that they can be made into arbitrarily good solvers by improving the PML, but this holds only for constant wave number. We show here that like all the other iterative Helmholtz solvers so far, the performance of these methods deteriorates as soon as the approximation of the Dirichlet to Neumann operator is not perfect any more in the case of wave propagation. To do so, we use a common algorithm formulation at the discrete level from [19], and provide the algorithm without any of the technicalities related to the various inventions, so that anybody can implement and check the method for themselves.

2 Common formulation of sweeping, source transfer, single layer, polarized traces and optimal/optimized Schwarz algorithms

To illustrate the limitations of these methods, it suffices to take the Helmholtz equation in a layered medium,

$$(\Delta + k(x)^2)u = f, \quad \text{in } \Omega := (0,1)^2, \tag{1}$$

with suitable boundary conditions for well posedness, such that after discretization by a standard five point finite difference method, the piecewise constant wave speeds are aligned with the block tridiagonal matrix structure

$$A\mathbf{u} := \begin{bmatrix} D_1 & L & & & \\ L & D_2 & L & & \\ & \ddots & \ddots & \ddots & \\ & & L & D_{J-1} & L \\ & & & L & D_J \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_{J-1} \\ \mathbf{u}_J \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_{J-1} \\ \mathbf{f}_J \end{bmatrix} =: \mathbf{f}. \tag{2}$$

The block LU factorization of the coefficient matrix in (2) is given by

$$A = \begin{bmatrix} T_1 & & & & \\ L & T_2 & & & \\ & \ddots & \ddots & & \\ & & L & T_{J-1} & \\ & & & L & T_J \end{bmatrix} \begin{bmatrix} I & T_1^{-1}L & & & \\ & I & T_2^{-1}L & & \\ & & \ddots & \ddots & \\ & & & I & T_{J-1}^{-1}L \\ & & & & I \end{bmatrix}, \tag{3}$$

where the T_j are the Schur complements that satisfy the recurrence relation

$$T_1 = D_1, \quad T_j = D_j - LT_{j-1}^{-1}L \text{ for } j \geq 2, \tag{4}$$

as one can verify by a direct calculation. The underlying system (2) can then be solved by a forward block substitution, followed by a backward block substitution, which corresponds to the sweeping over the domain back and forth, the source transfer from layer to layer, or the alternating solution over subdomains in the optimized Schwarz setting, see [19]. In the constant wave number case, the dense blocks T_j can be implemented using PML to arbitrary precision[1], and then all these sweeping type methods can be made arbitrarily close to being direct solvers, which explains their excellent performance in the constant wave number case. In the variable wave number case however, the best a PML can do is to be perfectly absorbing for the neighboring medium, assuming it to be constant up to infinity. To get such a perfect absorption for our model problem directly algebraically, without PML techniques, we consider for each wave number block D_i the constant coefficient problem

$$A^i\mathbf{u}^i := \begin{bmatrix} D_i & L & & & \\ L & D_i & L & & \\ & \ddots & \ddots & \ddots & \\ & & L & D_i & L \\ & & & L & D_i \end{bmatrix} \begin{bmatrix} \mathbf{u}_1^i \\ \mathbf{u}_2^i \\ \vdots \\ \mathbf{u}_{J-1}^i \\ \mathbf{u}_J \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_{J-1} \\ \mathbf{f}_J \end{bmatrix} =: \mathbf{f}, \tag{5}$$

with factorization

$$A^i = \begin{bmatrix} T_1^i & & & & \\ L & T_2^i & & & \\ & \ddots & \ddots & & \\ & & L & T_{J-1}^i & \\ & & & L & T_J^i \end{bmatrix} \begin{bmatrix} I & (T_1^i)^{-1}L & & & \\ & I & (T_2^i)^{-1}L & & \\ & & \ddots & \ddots & \\ & & & I & (T_{J-1}^i)^{-1}L \\ & & & & I \end{bmatrix}, \tag{6}$$

where T_j^i are the Schur complements that satisfy now the recurrence relation

$$T_1^i = D_i, \quad T_j^i = D_i - L(T_{j-1}^i)^{-1}L \text{ for } j \geq 2. \tag{7}$$

[1] provided the domain has indeed an open end or such a high order PML on the side where the sweeping begins.

Then the approximate factorization using this best possible approximation a PML technique could provide[2] is

$$\tilde{A} = \begin{bmatrix} \tilde{T}_1 & & & & \\ L & \tilde{T}_2 & & & \\ & \ddots & \ddots & & \\ & & L & \tilde{T}_{J-1} & \\ & & & L & \tilde{T}_J \end{bmatrix} \begin{bmatrix} I & \tilde{T}_1^{-1}L & & & \\ & I & \tilde{T}_2^{-1}L & & \\ & & \ddots & \ddots & \\ & & & I & \tilde{T}_{J-1}^{-1}L \\ & & & & I \end{bmatrix}, \tag{8}$$

where $\tilde{T}_j$ are the Schur complements using the exact Schur complements of the neighboring constant wave number case, namely

$$\tilde{T}_1 = D_1, \quad \tilde{T}_j = D_j - L(T_{j-1}^{j-1})^{-1}L \text{ for } j \geq 2. \tag{9}$$

Note that this best possible information a PML could provide is not necessarily a good approximation to the Dirichlet to Neumann operator which is represented by the exact blocks T_j, and thus contains information about all the reflections that will be created by all the layers outside the present subdomain. We will test now how much variation in the wave number this approximation can tolerate before the sweeping type algorithms loose their effectiveness, and how this depends on the source term and the boundary conditions of the underlying problem.

3 Numerical Study

We discretize the Helmholtz equation (1) using $n = 64$ interior mesh points, so that the mesh size is $h = 1/(n+1)$, and we use $p = 4, 8, 16$ layers. For the case of four layers, we use the wave numbers

$$k = [20\ 20\ 20\ 20] + \alpha[0\ 20\ 10\ -10], \tag{10}$$

where α is a contrast parameter, and for larger p we just repeat this structure. The resolution we chose guarantees at least ten points per wavelength resolution for this experiment. We start with the case of a wave guide in the x direction, where we used Robin radiation conditions on the left and right, and homogeneous Dirichlet conditions on top and bottom. We show in Figure 1 the solution[3] we obtain for $\alpha = 1$ with a point source at $x = 2h$, $y = \frac{1-h}{2}$ for the case of four and sixteen layers in the top row, and below for the constant source $f = 1$.

We now test the approximate factorization (8) both as an iterative solver and as a preconditioner for GMRES for varying contrast parameter α and right hand sides. We do this both for $n = 64$ interior meshpoints and the contrast profile (10), and on

[2] it is the exact Schur complement, including all boundary information, the only approximation is the constant wave number.

[3] The boundary points are not plotted, so one can not see the homogeneous Dirichlet condition.

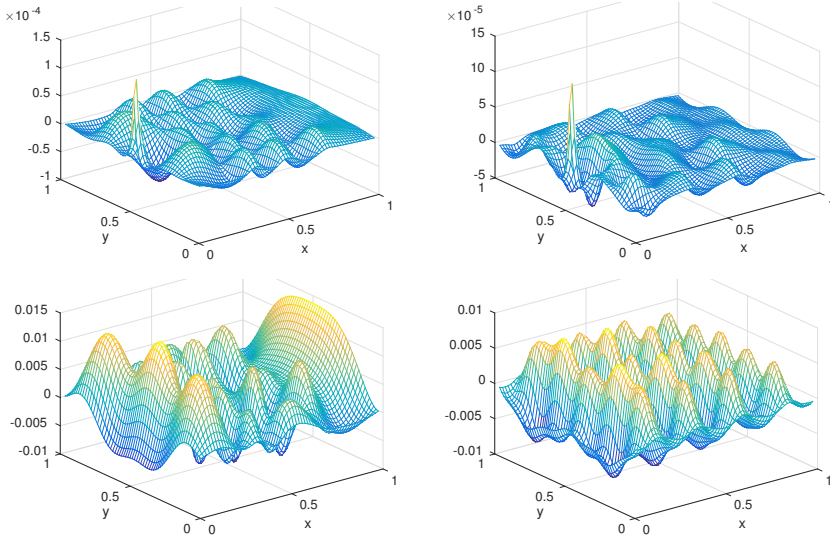


Fig. 1 Top: Solutions computed with a point source. Bottom: Solutions computed with $f = 1$. Left: 4 layers. Right: 16 layers.

a refined mesh with twice the number of interior meshpoints, $n = 128$, but also a profile with twice the size for the wave number, i.e.

$$k = [40\ 40\ 40\ 40] + \alpha[0\ 40\ 20\ -20], \tag{11}$$

so that we still have at least ten points per wavelength resolution. We show in Table 1 the number of iterations the methods took, where we stopped the iterative version of the algorithm at the relative error tolerance $1e-6$, and GMRES when the residual was reduced by $1e-6$, and we started with a zero initial guess. The three columns within each 'Iterative' or 'GMRES' column correspond to the point source f, constant $f = 1$ throughout the domain, and also a random f. The top part is for the smaller wave number experiment (10), and the bottom part is for the larger wave number experiment (11). We first see that for $\alpha = 0$, i.e. in the constant wave number case, the factorization is exact, both the iterative version and GMRES converge in one iteration step, and the contraction factor ρ (the spectral radius) of the iterative version equals numerically zero. As soon as we have however a non-constant wave number, already for $\alpha = 0.001$, the factorization is not exact any more. Nevertheless the methods still converge well, up to $\alpha = 0.01$ in the smaller wave number case in the top half of the table, i.e. a one percent variation in the wave number k. Here the contraction factor is $\rho = 0.2460$ for $p = 4$ subdomains, and grows when the number of subdomains p is increasing. For larger contrast, the iterative version of the algorithm can not be used any more, $\rho > 1$, and GMRES deteriorates now rapidly, for example if the contrast is at a factor of two, i.e. $\alpha = 1$, GMRES iteration numbers double when the number of subdomains doubles, the sweeping type methods are not

	$p=4$				$p=8$				$p=16$			
α	Iterative			ρ	GMRES			Iterative			ρ	GMRES		
0	1	1	1	5.8e-15	1	1	1	1	1	1	6.1e-15	1	1	1	1	1	1	4.6e-15	1	1	1
0.001	4	3	4	0.0250	3	3	4	5	4	5	0.0738	4	3	4	6	4	6	0.0979	5	4	5
0.005	6	4	6	0.1250	4	3	5	13	5	14	0.4031	7	5	7	12	8	11	0.3155	8	6	8
0.01	9	4	9	0.2460	5	4	5	32	7	34	0.6877	9	6	9	25	13	28	0.6244	11	7	11
0.05	-	7	-	1.6072	8	6	8	-	-	-	11.135	15	11	15	-	-	-	20.593	21	15	21
0.1	28	11	26	0.6887	9	7	9	-	-	-	3.0238	17	13	18	-	-	-	2.7604	25	17	26
1	-	-	-	2.4141	18	12	19	-	-	-	173.66	35	29	37	-	-	-	7.0979	62	44	67
0	1	1	1	5.9e-15	1	1	1	1	1	1	5.3e-15	1	1	1	1	1	1	6.5e-15	1	1	1
0.001	4	4	4	2.49e-2	4	3	4	5	5	5	0.1055	5	4	5	8	5	8	0.1991	6	5	6
0.005	7	7	7	0.1428	5	5	5	84	9	90	0.8824	9	6	9	26	11	27	0.6328	12	8	12
0.01	12	12	12	0.3300	6	5	6	-	18	-	1.9386	12	8	12	-	-	-	1.1614	19	11	19
0.05	-	-	-	4.5040	13	9	13	-	-	-	8.1397	23	17	22	-	-	-	1408.4	43	34	44
0.1	-	-	-	2.2412	14	11	15	-	-	-	20.614	20	14	19	-	-	-	2515.4	43	38	40
1	-	-	-	8.7091	31	20	33	-	-	-	6.9288	61	46	66	-	-	-	4.079e5	67	99	83

Table 1 Iteration numbers in the wave guide setting.

robust any more[4]. In the higher wave number case in the bottom part of the table, the methods start having problems already at $\alpha = 0.005$, variations of the wave speed of half a percent, and they deteriorate even more rapidly for higher contrast. We can also see comparing the last two lines of the top and bottom half of the table that doubling the wave number leads to twice the iteration numbers with GMRES as soon as the contrast is large enough, and GMRES failed to converge in less than hundred iterations at the bottom right. We also measured that in certain cases, the relative residual reduction of $1e-6$ for GMRES does not lead to a relative error of the same size. This is notably the case for $\alpha = 1$ in the smaller wave number case when $p = 8$ with point or random source (relative error $1.83e-4$ and $1.26e-4$ only), and in the larger wave number case when $p = 16$ with point or random source, (relative error 0.27317 and 0.52128 only !). So the corresponding GMRES iteration numbers (67 and 83) would need to be substantially higher to reach the same level of accuracy of $1e-6$ as for the other results in the table: we measured 129 instead of 67 to reach $1.8607e-6$ and 139 instead of 83 to reach $2.9641e-6$ respectively.

We next perform the same set of experiments, but now using Robin boundary conditions all around the domain, see Table 2. We see that the outer Robin boundary conditions are better than the wave guide setting for the sweeping type algorithms, they work now in the iterative version up to about a 10 percent variation of the wave number in this specific experiment. As soon as however there is a variation as large

[4] There are also two interesting apparent anomalies: in the smaller wavenumber case, for $p = 4$ and $\alpha = 0.05$ (and also one in the larger wave number case), the spectral radius is bigger than one, but for the source term $f = 1$ we observe convergence. We iterated in this case however further, and then the iterations also start to diverge, it is only that the divergent modes are not stimulated at the beginning by the source term $f = 1$ and zero initial guess, a typical phenomenon known from power iterations, which explains in the table the general observation that the problem with $f = 1$ is easier to solve than with the other sources, also for GMRES. For the same $p = 4$ and $\alpha = 0.1$, we then get surprisingly a spectral radius again smaller than 1, which is a lucky configuration and not observed for more subdomains or different α.

	$p=4$						$p=8$						$p=16$					
α	Iterative			ρ	GMRES		Iterative			ρ	GMRES		Iterative			ρ	GMRES	
0	1	1	1	3.6e-15	1	1	1	1	1	4.5e-15	1	1	1	1	1	2.8e-15	1	1	1
0.001	2	3	3	1.28e-3	2	2	2	3	3	3.40e-3	3	3	3	3	3	3.74e-3	3	3	3
0.005	3	3	3	6.58e-3	3	3	3	4	4	1.69e-2	4	3	3	4	4	1.92e-2	4	4	4
0.01	4	4	4	1.36e-2	3	3	3	4	4	3.35e-2	4	4	4	5	4	3.79e-2	4	4	4
0.05	6	6	6	8.25e-2	5	5	5	7	7	0.1446	6	6	6	10	9	0.2403	7	7	7
0.1	8	8	8	0.1677	6	5	6	9	9	0.2202	7	7	7	15	16	0.4182	9	9	10
1	80	80	80	0.8471	13	10	13	-	-	2.8446	24	19	25	-	-	3.1188	39	30	38
0	1	1	1	3.6e-15	1	1	1	1	1	4.0e-15	1	1	1	1	1	4.4e-15	1	1	1
0.001	3	3	3	1.91e-3	2	3	3	3	3	5.57e-3	3	3	3	4	4	1.29e-2	3	3	3
0.005	3	3	3	9.63e-3	3	3	3	4	4	2.73e-2	4	4	4	5	5	6.58e-2	5	5	5
0.01	4	4	4	1.97e-2	4	4	4	5	5	5.29e-2	5	5	5	7	7	0.1343	6	6	6
0.05	6	6	6	0.1006	5	5	5	11	11	0.2771	8	8	8	21	22	0.5287	10	10	11
0.1	10	9	9	0.2353	7	7	7	14	13	0.3796	9	9	9	41	44	0.7344	12	11	12
1	-	-	-	1.4684	19	14	19	-	-	2.9234	36	25	35	-	-	36.193	76	65	80

Table 2 Iteration numbers for a domain with Robin conditions all around.

		$p=4$		$p=8$		$p=16$		$p=4$			$p=8$			$p=16$		
L/h	C	GMRES		GMRES		GMRES		Iterative			GMRES			GMRES		
5	4π	18	13	19	36	29	38	62	42	63	20	24	28	12	11	12	23	24	25	32	32	34
10	8π	18	13	19	36	29	38	61	41	62	18	24	25	11	11	12	20	20	22	28	25	29
5	4π	28	17	21	61	46	64	86	86	95	-	-	-	14	13	13	25	24	24	58	60	61
10	8π	28	17	19	61	46	64	87	81	90	39	47	46	11	11	12	22	21	22	46	45	48

Table 3 Iteration numbers in the presence of outer PMLs. Left: waveguide. Right: PMLs all around. Bottom part has doubled wavenumber and half the mesh size like in Tables 1 and 2.

as a factor of two, the method is not an effective solver any more, the iterative version diverges because $\rho > 1$, and GMRES iteration numbers deteriorate when the number of subdomains increases, like in the previous case: we still observe a doubling of the GMRES iteration count when the number of subdomains doubles, and also when the wave number is multiplied by 2. With Robin conditions all around, there is less loss of accuracy compared to the residual tolerance than in the wave guide case: only in the high wave number case for $\alpha = 1$ and $p = 16$, the relative error reached $1.6463e-05$ for the point source and $1.2333e-05$ for the random source instead of the $1e-6$ asked for in the relative residual, all other results had the required level also in the relative error.

Finally, we use a complex stretching PML instead of the outer Robin boundary condition. For example, we extend the right boundary from 1 to $1+L$ and perform in the extended region in (1) the transform $\partial_x \to s\partial_x$, $s = \frac{1}{1-\mathbf{i}C(x-1)^2/(L^3k(1,y))}$, $\mathbf{i} = \sqrt{-1}$, and similarly on the other boundaries. We increase L and C to get more absorption in the PMLs, and check how this affects the results for $\alpha = 1$ in Table 1 and Table 2, see Table 3. The iterative version diverges in most cases except when $p = 4$ for the PML-all-around problem. Absorption helps GMRES marginally for the waveguide problem but remarkably for the PML-all-around problem. Note that, however, the iteration count still doubles along with the number of subdomains and when dou-

bling the wave number for many subdomains. We also tested the case of a fixed wave number profile, namely the one in Table 3 at the bottom right with 16 layers: for $p = 4$ we obtain for GMRES the iteration numbers $\boxed{16\ 20\ 21}$, and for $p = 8$ $\boxed{59\ 69\ 70}$. This indicates that also for a fixed difficulty, i.e. fixed number of layers, iteration numbers grow when subdomain numbers are increasing. We observe however also when comparing with $p = 16$ at the bottom right of Table 3 the interesting phenomenon that once layers are all aligned with subdomains, the problem becomes apparently a bit easier. We are currently studying this phenomenon theoretically. Note that if too many PMLs are used, the 2-norm of the residuals may be dominated by the residuals in the PMLs, and one should use a more reliable metric for the stopping criterion.

4 Conclusion

We presented the simplest common form of the fundamental algorithm underlying the new type Helmholtz (and Maxwell) solvers based on sweeping. These solvers are among the best currently available solvers for such type of problems, and they can be made robust in the wave number by increasing the accuracy of the PML, provided the wave number is constant. If the wave number is not constant however, the PML is not the right approximation of the Dirichlet to Neumann operator or the Schur complement any more, which is the essential ingredient for these algorithms to be effective. We showed by a simple set of numerical experiments which is easy to reproduce that in a layered medium with contrast of only one percent, these algorithms already perform substantially less well if the layers are not aligned with the sweeping direction, and when the contrast is as large as a factor of two, the methods do not work any more as stationary iterations, and preconditioned GMRES iteration numbers start to grow drastically: they increase linearly in the number of subdomains and the wave number in our experiments. One must therefore investigate an approximation different from PML for the Dirichlet to Neumann operator in the case of non-constant wave numbers.

References

1. Ivo M. Babuska and Stefan A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Journal on numerical analysis*, 34(6):2392–2423, 1997.
2. Yassine Boubendir, Xavier Antoine, and Christophe Geuzaine. A quasi-optimal non-overlapping domain decomposition algorithm for the Helmholtz equation. *Journal of Computational Physics*, 231(2):262–280, 2012.
3. Zhiming Chen and Xueshuang Xiang. A source transfer domain decomposition method for Helmholtz equations in unbounded domain. *SIAM J. Numer. Anal.*, 51:2331–2356, 2013.

4. Zhiming Chen and Xueshuang Xiang. A source transfer domain decomposition method for Helmholtz equations in unbounded domain Part II: Extensions. *Numer. Math. Theor. Meth. Appl.*, 6:538–555, 2013.
5. Pierre-Henri Cocquet and Martin J. Gander. On the minimal shift in the shifted laplacian preconditioner for multigrid to work. In *Domain Decomposition Methods in Science and Engineering XXII*, pages 137–145. Springer, 2016.
6. Pierre-Henri Cocquet and Martin J. Gander. How large a shift is needed in the shifted Helmholtz preconditioner for its effective inversion by multigrid? *SIAM Journal on Scientific Computing*, 39(2):A438–A478, 2017.
7. Bjorn Engquist and Lexing Ying. Sweeping preconditioner for the Helmholtz equation: Hierarchical matrix representation. *Comm. Pure Appl. Math.*, LXIV:0697–0735, 2011.
8. Bjorn Engquist and Lexing Ying. Sweeping preconditioner for the Helmholtz equation: Moving perfectly matched layers. *Multiscale Model. Sim.*, 9:686–710, 2011.
9. Yogi A. Erlangga, Cornelis Vuik, and Cornelis Willebrordus Oosterlee. On a class of preconditioners for solving the helmholtz equation. *Applied Numerical Mathematics*, 50(3-4):409–425, 2004.
10. Oliver G. Ernst and Martin J. Gander. Why it is difficult to solve Helmholtz problems with classical iterative methods. In *Numerical analysis of multiscale problems*, pages 325–363. Springer, 2012.
11. Martin J. Gander. Optimized Schwarz methods. *SIAM J. Numer. Anal.*, 44:699–731, 2006.
12. Martin J. Gander, Ivan G. Graham, and Euan A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed? *Numer. Math.*, 131:567–614, 2015.
13. Martin J. Gander, Laurence Halpern, and Frederic Magoules. An optimized Schwarz method with two-sided Robin transmission conditions for the Helmholtz equation. *Int. J. Numer. Meth. Fluids*, 55:163–175, 2007.
14. Martin J. Gander, Frederic Magoules, and Frederic Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24:38–60, 2002.
15. Martin J. Gander and Frederic Nataf. AILU: a preconditioner based on the analytic factorization of the elliptic operator. *Numer. Linear Algebra Appl.*, 7:505–526, 2000.
16. Martin J. Gander and Frederic Nataf. An incomplete LU preconditioner for problems in acoustics. *J. Comput. Acoust.*, 13:455–476, 2005.
17. Martin J. Gander and Achim Schädle. The pole condition: a Padé approximation of the Dirichlet to Neumann operator. In *Domain Decomposition Methods in Science and Engineering XIX*, pages 125–132. Springer, 2011.
18. Martin J. Gander and Achim Schädle. On the relationship between the pole condition, absorbing boundary conditions and perfectly matched layers. 2018. in preparation.
19. Martin J. Gander and Hui Zhang. A class of iterative solvers for the Helmholtz equation: factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods. *SIAM Review*, page to appear, 2018.
20. Ivan G. Graham, Euan A. Spence, and Eero Vainikko. Domain decomposition preconditioning for high-frequency Helmholtz problems using absorption. *ArXiv e-prints*, 2015.
21. Achim Schädle and Lin Zschiedrich. Additive Schwarz method for scattering problems using the PML method at interfaces. In *Domain Decomposition Methods in Science and Engineering XVI*, pages 205–212. Springer, 2007.
22. Andrea Toselli. Some results on overlapping Schwarz methods for the Helmholtz equation employing perfectly matched layers. In *Domain Decomposition Methods in Sciences and Engineering: Eleventh International Conference London, UK*, pages 539–545, 1998.
23. Leonardo Zepeda-Núñez and Laurent Demanet. The method of polarized traces for the 2D Helmholtz equation. *J. Comput. Phys.*, 308:347–388, 2016.
24. Leonardo Zepeda-Núñez, Russel J. Hewett, and Laurent Demanet. Preconditioning the 2D Helmholtz equation with polarized traces. In *SEG Technical Program Expanded Abstracts 2014*, pages 3465–3470. SEG, 2014.

Convergence of Asynchronous Optimized Schwarz Methods in the plane

José C. Garay, Frédéric Magoulès and Daniel B. Szyld

Abstract A convergence proof of Asynchronous Optimized Schwarz Methods applied to a shifted laplacian problem, with negative shift, in $\mathbb{R}^2$ is presented. Sufficient conditions for convergence involving initial values of the approximation of the solution are discussed.

1 Introduction

Optimized Schwarz Methods are Domain Decomposition methods in which the boundary conditions on the artificial interfaces are of Robin type, i.e., containing one or more parameters that can be optimized [1, 3, 4].

In our context, Asynchronous Schwarz methods are those where each subdomain solve is performed with whatever new information (to be used for the boundary conditions) has arrived from the neighboring subdomains since the last update, but without necessarily waiting for new information to arrive. For more details on asynchronous methods, see, e.g. [2] and references therein. See also Section 1.2 below.

In this paper we add more details to the convergence proof given in [5] of Asynchronous Optimized Schwarz (AOS) where it is used to solve Poisson's equation in $\mathbb{R}^2$. The results presented here complement those of [5].

José C. Garay
Temple University, Philadelphia, USA , e-mail: jose.garay@temple.edu.
Supported in part by the U.S. Deparment of Energy under grant DE-SC0016578.

Frédéric Magoulès
CentraleSupélec, Châtenay-Malabry, France, e-mail: frederic.magoules@centralesupelec.fr .

Daniel B. Szyld
Temple University, Philadelphia, USA, e-mail: szyld@temple.edu.
Supported in part by the the U.S. National Science Foundation under grant DMS-1418882 and the U.S. Deparment of Energy under grant DE-SC0016578.

1.1 Preliminaries

The aim is to provide a complete proof of the convergence of AOS for

$$\Delta u - \eta u = f \quad \text{in } \mathbb{R}^2, \tag{1}$$

with vanishing value of u at infinity, and $\eta > 0$. The space $\mathbb{R}^2$ is divided into p overlapping infinite vertical strips. This means we have $p-1$ vertical lines, say at coordinates $x = \ell_1, \ldots \ell_{p-1}$; and we assume for simplicity that we have the same overlap $2L$ between subdomains. We also assume, without loss of generality, that except for the subdomains at infinity, that each strip has the same width, i.e., $\ell_s - \ell_{s-1} = W$ for $s = 2, \ldots, p-1$, so that $\ell_s = \ell_1 + (s-1)W$. It follows then, that the overlap satisfies $2L < W$, and usually $L \ll W$. Thus, we have $\Omega^{(1)} =]-\infty; \ell_1 + L] \times \mathbb{R}$, $\Omega^{(s)} = [\ell_{s-1} - L; \ell_s + L] \times \mathbb{R}$, $s = 2, \ldots, p-1$, and $\Omega^{(p)} = [\ell_{p-1} - L; +\infty[\times \mathbb{R}$. In this context, the normal vector is in the x direction (with the appropriate sign).

Let $f^{(s)}$ and u_s^n denote the restriction of f and u^n, the approximation to the solution at the iteration n, to $\Omega^{(s)}$, $s = 1, \ldots, p$, respectively. Thus, $u_s^n \in V^{(s)}$, a space of functions defined on $\Omega^{(s)}$. We consider transmission conditions (on the artificial interfaces) composed of local operators. The *local problems* and the *synchronous* iteration process is described by the following equations

$$\begin{cases} (\Delta - \eta) u_1^{n+1} = f^{(1)} & \text{on } \Omega^{(1)}, \\ \frac{\partial u_1^{n+1}}{\partial x} + \Lambda u_1^{n+1} = \frac{\partial u_2^n}{\partial x} + \Lambda u_2^n & \text{for } x = \ell_1 + L, \\ \text{For } s = 2, \ldots, p-1, & \\ -\frac{\partial u_s^{n+1}}{\partial x} + \Lambda u_s^{n+1} = -\frac{\partial u_{s-1}^n}{\partial x} + \Lambda u_{s-1}^n & \text{for } x = \ell_{s-1} - L, \\ (\Delta - \eta) u_s^{n+1} = f^{(s)} & \text{on } \Omega^{(s)}, \\ \frac{\partial u_s^{n+1}}{\partial x} + \Lambda u_s^{n+1} = \frac{\partial u_{s+1}^n}{\partial x} + \Lambda u_{s+1}^n & \text{for } x = \ell_s + L, \\ -\frac{\partial u_p^{n+1}}{\partial x} + \Lambda u_p^{n+1} = -\frac{\partial u_{p-1}^n}{\partial x} + \Lambda u_{p-1}^n & \text{for } x = \ell_{p-1} - L, \\ (\Delta - \eta) u_p^{n+1} = f^{(p)} & \text{on } \Omega^{(p)}, \end{cases} \tag{2}$$

where Λ is a local approximation to the Poincaré-Steklov operator using differential operators (e.g., $\Lambda = \alpha$ and for artificial boundary conditions of OO0 family of artificial conditions, with α constant, and $\Lambda = \alpha + \beta \frac{\partial^2}{\partial \tau^2}$ for the OO2 family, where $\frac{\partial^2}{\partial \tau^2}$ is the tangential second derivative with respect to the boundary and β a constant; α and β are parameters whose values are chosen to optimize convergence properties and thus minimize convergence bounds).

Using linearity we obtain that the error of the synchronous iterative procedure is the solution of (2) with $f = 0$. The Fourier transform in the y direction of the error of the local problem s at iteration n then can be written as (see [5])

$$\hat{u}_n^s(x,k) = A_s^n(k) e^{-\theta(k)|x-(l_{s-1}-L)|} + B_s^n(k) e^{-\theta(k)|x-(l_s+L)|} \tag{3}$$

where $\theta(k) = \sqrt{\eta + k}$. Let $c(n)^T = ((c_1(n), c_2(n), \ldots, c_{p-1}(n), c_p(n)) = (B_1^n, A_2^n, B_2^n, \ldots, A_{p-1}^n, B_{p-1}^n, A_p^n)$, where $c_1 = B_1^n$ and $c_p = A_p^n$ are scalars, and

$c_s = (A_s^n, B_s^n)$ are ordered pairs for $s = 2, \ldots, p-1$. Plugging the expression (3) into (2) (with $f = 0$), we can write the iteration from $u(n)$ to $u(n+1)$ in terms of the coefficients $c(n)$ and $c(n+1)$ obtaining an $(2p-1) \times (2p-1)$ matrix $\hat{T}$ such that $c(n+1) = \hat{T}c(n)$; see [5] for more details. In that reference, it is shown that the operator $\hat{T}$ is contracting in max norm, and in this paper we continue the proof starting precisely from this result[1].

1.2 Mathematical model of asynchronous iterative methods

Let $X^{(1)}, ..., X^{(p)}$ be given sets and X be their Cartesian product, i.e., $X = X^{(1)} \times \cdots \times X^{(p)}$. Thus $x \in X$ implies $x = \left(x^{(1)}, ..., x^{(p)}\right)$ with $x^{(s)} \in X^{(s)}$ for $s \in \{1, ..., p\}$. Let $T^{(s)} : X \to X^{(s)}$ where $s \in \{1, ..., p\}$, and let $T : X \to X$ be a vector-valued map (iteration map) given by $T = (T^{(1)}, ..., T^{(p)})$ with a fixed point x^*, i.e., $x^* = T(x^*)$. Let $\{t_n\}_{n \in \mathbb{N}}$ be the sequence of time stamps at which at least one processor updates its associated component. Let $\{\sigma(n)\}_{n \in \mathbb{N}}$ be a sequence with $\sigma(n) \subset \{1, ..., p\}$ $\forall n \in \mathbb{N}$. The set $\sigma(n)$ consists of labels (numbers) of the processors that update their associated component at the n-th time stamp. Define for $s, q \in \{1, ..., p\}$, $\{\tau_q^s(n)\}_{n \in \mathbb{N}}$ a sequence of integers, representing the time-stamp index of the update of the data coming from processor q and available in processor s at the beginning of the computation of $x^{(s)}(n)$ which ends at the n-th time stamp. Let $x(0) = \left(x^{(1)}(0), ..., x^{(p)}(0)\right)$ be the initial approximation (of the fixed point x^*). Then, the new computed value updated by processor s at the nth time stamp is

$$x^{(s)}(n) = \begin{cases} T^{(s)}\left(x^{(1)}(\tau_1^s(n)), ..., x^{(p)}(\tau_p^s(n))\right), & s \in \sigma(n) \\ x^{(s)}(n-1), & s \notin \sigma(n). \end{cases}$$

It is assumed that the three following conditions (necessary for convergence) are satisfied

$$\forall s, q \in \{1, \ldots, p\}, \forall n \in \mathbb{N}^*, \tau_q^{(s)}(n) < n, \tag{4}$$

$$\forall s \in \{1, \ldots, p\}, \text{card}\{n \in \mathbb{N}^* | s \in \sigma(n)\} = +\infty, \tag{5}$$

$$\forall s, q \in \{1, \ldots, p\}, \lim_{n \to +\infty} \tau_q^{(s)}(n) = +\infty. \tag{6}$$

Condition (4) indicates that data used at the time t_n must have been produced before time t_n, i.e., time does not flow backward. Condition (5) means that no process will ever stop updating its components. Condition (6) corresponds to the fact that new data will always be provided to the process. In other words, no process will have a piece of data that is never updated.

[1] In [5] it is indicated that given $\hat{T}$ is contracting, then $T^n \to 0$, where T maps $u(n)$ to $u(n+1)$, but this implication may not always hold. This is why we need to complete the proof in a different manner. We do so by showing explicitly that (8) holds.

2 Convergence proof for the asynchronous case

We now present the convergence proof of the asynchronous implementation of Optimized Schwarz with transmission conditions composed of local operators (as described in section 1.1) when applied to (1). Note that the local problem of AOS is obtained by replacing, in (2), $n+1$ by t_{new} and n by the corresponding update times of the values of u received from the neighboring subdomains and available at the begining of the computation of the new update. Let us define a *time stamp* as the instant of time at which at least one processor finishes its computation and produces a new update. Let t_m be the $m-th$ time stamp and $u_s^{t_m}$ be the error of the local problem s at time $t = t_m$. Note then that the asynchronous method converges if for any monotonically increasing sequence of time stamps $\{t_m\}_{m\in\mathbb{N}}$ we have

$$\lim_{m\to\infty} u_s^{t_m} = 0 \tag{7}$$

Thus, in order to prove convergence of the asynchronous iterations, we just need to prove that (7) holds for any monotonically increasing sequence of time stamps $\{t_m\}_{m\in\mathbb{N}}$, which is what we prove next.

Theorem 1. *Let us define a time stamp t_m as the instant of time at which at least one processor finishes its computation and produces a new update. Let $u_s^{t_m}(x,y)$ be the error of the local problem s (of the asynchronous version of (2)), $s \in \{1,...,p\}$, and $\hat{u}_s^{t_m}(x,k)$ be its corresponding Fourier transform in the y direction. Let $S = \{l_{s-1} - L : s = 2,...,p\} \cup \{l_s + L : s = 1,...,p-1\}$ (i.e., S is the set of the $x-$coordinates of each of the artificial boundaries of each of the local problems. Then, if $\hat{u}_s^0(x,k)$ is uniformly bounded in $k \in \mathbb{N}$ and $x \in S$, we have, $\forall s \in \{1,...,p\}$, $\lim_{m\to\infty} u_s^{t_m}(x,y) = 0$ in $\Omega^{(s)}$ for any (monotonically increasing) sequence of time stamps $\{t_m\}_{m\in\mathbb{N}}$.*

Outline of the proof

Note first that all the derivatives of $u_s^{t_m}$ exist and are continuous. Then, if $u_s^{t_m}$ converges to zero uniformly in $[l-\varepsilon, l+\varepsilon]\times\mathbb{R}$ as $m\to\infty$ and the first and derivatives of other orders of $u_s^{t_m}$ contained in Λ are continuous, it can be shown that $\lim_{m\to\infty}\left(\frac{\partial u_s^{t_m}}{\partial x} + \Lambda u_s^{t_m}\right)(x,y) = 0$ uniformly in $\{l\}\times\mathbb{R}$.

We want to prove that for any sequence of time stamps $\{t_m\}_{m\in\mathbb{N}}$ and for every $s \in \{1,...,p\}$ we have $\lim_{m\to\infty} |u_s^{t_m}(x,y)| = 0$ in $\Omega^{(s)}$. Note that, to prove this statement, by the argument given in the previous paragraph, with $S_\varepsilon = \cup_{z\in S}[z-\varepsilon, z+\varepsilon]$, we just need to prove that for every $s \in \{1,...,p\}$ it holds $\lim_{m\to\infty} |u_s^{t_m}(x,y)| = 0$ uniformly in $S_\varepsilon \cap [\ell_{s-1}, \ell_s]\times\mathbb{R}$, since this implies that the values of the boundary conditions of each local problem will converge to zero, and consequently so will do the solution of each local problem in its interior domain.

Observe that, if $\lim_{m\to\infty} |\hat{u}_s^{t_m}(x,k)| = 0$ and

$$\lim_{m\to\infty} \int_{-\infty}^{\infty} |\hat{u}_s^{t_m}(x,k)|dk = \int_{-\infty}^{\infty} \lim_{m\to\infty} |\hat{u}_s^{t_m}(x,k)|dk, \tag{8}$$

we have

$$\lim_{m\to\infty}|u_s^{t_m}(x,y)| = \lim_{m\to\infty}\left|\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\hat{u}_s^{t_m}(x,k)e^{iyk}dk\right|$$
$$\leq \frac{1}{(2\pi)^2}\lim_{m\to\infty}\int_{-\infty}^{\infty}|\hat{u}_s^{t_m}(x,k)|dk = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\lim_{m\to\infty}|\hat{u}_s^{t_m}(x,k)|dk = 0.$$

Thus, in order to prove that $\lim_{m\to\infty}|u_s^{t_m}(x,y)| = 0$ in $\Omega^{(s)}$, it suffices to prove that, for every $s\in\{1,...,p\}$, the following three statements hold:

1. $\lim_{m\to\infty}|A_s^{t_m}(k)| = 0$ and $\lim_{m\to\infty}|B_s^{t_m}(k)| = 0$.
2. $\lim_{m\to\infty}|\hat{u}_s^{t_m}(x,y)| = 0$ for $\forall x\in S_\varepsilon\cap[\ell_{s-1},\ell_s]$ and $y\in\mathbb{R}$.
3. For all $x\in S_\varepsilon\cap[\ell_{s-1},\ell_s]$ and $y\in\mathbb{R}$, (8) holds.

Item 3. means, in other words, that if $|\hat{u}_s^{t_m}(x,.)|$ goes to zero as m goes to infinity, so will do its integral over $k\in\mathbb{R}$, and, in turn, the inverse Fourier transform of $\hat{u}_s^{t_m}(x,.)$.

Proof of the Theorem

We first prove that $||c(0)||_\infty < \infty$. For ease of notation, for each subdomain s, let the left artificial boundary condition be $p_s(k)$ and the right artificial boundary condition $q_s(k)$. Thus, it follows from the expression (3) that, at $x = l_{s-1}-L$,

$$\hat{u}_s^0(l_{s-1}-L,k) = A_s^0(k)+B_s^0(k)e^{\theta(k)(l_{s-1}-l_s-2L)} = A_s^0(k)+B_s^0(k)e^{-\theta(k)(W+2L)} = p_s(k) \tag{9}$$

and at $x = l_s+L$

$$\hat{u}_s^0(l_s+L,k) = A_s^0(k)e^{-\theta(k)(l_s-l_{s-1}+2L)}+B_s^0(k) = A_s^0(k)e^{-\theta(k)(W+2L)}+B_s^0(k) = q_s(k). \tag{10}$$

From (10) we have $B_s^0(k) = q_s(k)-A_s^0(k)e^{-\theta(k)(W+2L)}$. Then, plugging this expression of $B_s^0(k)$ into (9) gives

$$A_s^0(k)+\left[q_s(k)-A_s^0(k)e^{-\theta(k)(W+2L)}\right]e^{-\theta(k)(W+2L)} = p_s(k),$$
$$A_s^0(k)\left[1-e^{-2\theta(k)(W+2L)}\right] = p_s(k)-q_s(k)e^{-\theta(k)(W+2L)},$$
$$A_s^0(k) = \frac{p_s(k)-q_s(k)e^{-\theta(k)(W+2L)}}{1-e^{-2\theta(k)(W+2L)}},$$

$$|A_s^0(k)| = \frac{|p_s(k)-q_s(k)e^{-\theta(k)(W+2L)}|}{|1-e^{-2\theta(k)(W+2L)}|} \leq \frac{|p_s(k)|+|q_s(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}.$$

By a similar process we obtain

$$|B_s^0(k)| \leq \frac{|q_s(k)|+|p_s(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}.$$

Let $p^*(k)$ and $q^*(k)$ be such that

$$\max_{s\in\{1,\ldots,p\}}\left\{\max\left\{\frac{|p_s(k)|+|q_s(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}\quad,\quad\frac{|q_s(k)|+|p_s(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}\right\}\right\}$$
$$=\frac{|p^*(k)|+|q^*(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}$$

Then, we have that

$$||c(0)||_\infty\leq\frac{|p^*(k)|+|q^*(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}.$$

By hypothesis, $\hat{u}_s^0$ is uniformly bounded in $k\in\mathbb{N}$ and $x\in S$. Thus, there exists a number $M>0$ such that $\hat{u}^0(x,k)\leq M$ for any $k\in\mathbb{R}$ and $x\in S$. Then, we have that $|p_s(k)|,|q_s(k)|\leq M$ for any $k\in\mathbb{R}$ and $s\in\{1,\ldots p\}$. Then, necessarily,$|p^*(k)|,|q^*(k)|\leq M$, and consequently $||c(0)||_\infty<\infty$.

Let $\{t_m\}$ be a monotonically increasing sequence of time stamps. As mentioned previously, in [5] it is proven that $||\hat{T}(k)c(k)||_\infty\leq\rho||c(k)||_\infty$, with $\rho<1$. This implies that after one application of a local operator to an arbitrary vector $c_{\text{old}}(k)$ we have

$$|A_s^{\text{new}}|,|B_s^{\text{new}}|\leq||\hat{T}^s(k)c_{\text{old}}(k)||_\infty\leq\rho||c_{\text{old}}(k)||_\infty$$

and after all processes have updated their values at least once, say at time stamp t_*, we have at least $|A_s^{t_*}|,|B_s^{t_*}|\leq\rho||c_0(k)||_\infty$. This implies, in turn, that given a monotonically increasing sequence $\{t_m\}_{m\in\mathbb{N}}$, at time t_m we have

$$|A_s^{t_m}|,|B_s^{t_m}|\leq\rho^{\phi_s(m)}||c_0(k)||_\infty$$

where, for each $s\in\{1,\ldots,p\}$, $\phi_s:\mathbb{N}\to\mathbb{N}$ such that $\phi_s(m)\to\infty$ as $m\to\infty$. Then,

$$\lim_{m\to\infty}|A_s^{t_m}(k)|\leq\lim_{m\to\infty}\rho^{\phi_s(m)}||c_0(k)||_\infty=||c_0(k)||_\infty\lim_{m\to\infty}\rho^{\phi_s(m)}=||c_0(k)||_\infty 0=0.$$

Similarly, $\lim_{m\to\infty}|B_s^{t_m}(k)|=0$, and therefore

$$\begin{aligned}\lim_{m\to\infty}|\hat{u}_s^{t_m}(x,k)|&=\lim_{m\to\infty}\left|A_s^{t_m}(k)e^{-\theta(k)|x-(l_{s-1}-L)|}+B_s^{t_m}(k)e^{-\theta(k)|x-(l_s+L)|}\right|\\&\leq\lim_{m\to\infty}\left(\left|A_s^{t_m}(k)\right|e^{-\theta(k)|x-(l_{s-1}-L)|}+\left|B_s^{t_m}(k)\right|e^{-\theta(k)|x-(l_s+L)|}\right)\\&=\left(\lim_{m\to\infty}\left|A_s^{t_m}(k)\right|\right)e^{-\theta(k)|x-(l_{s-1}-L)|}+\left(\lim_{m\to\infty}\left|B_s^{t_m}(k)\right|\right)e^{-\theta(k)|x-(l_s+L)|}\\&=0.\end{aligned}$$

To complete the proof, we need to show that (8) holds for $x\in S_\varepsilon\cap[\ell_{s-1},\ell_s]$ and $y\in\mathbb{R}$. We show now that, for all $m\in\mathbb{N}$, $|\hat{u}_s^{t_m}(x,.)|$ is bounded by an $L^1(\mathbb{R})$ function. To that end, we have that,

$$
\begin{aligned}
|\hat{u}_s^{t_m}(x,k)| &= |A_s^{t_m}(k)e^{-\theta(k)|x-(l_{s-1}-L)|} + B_s^{t_m}(k)e^{-\theta(k)|x-(l_s+L)|}| \\
&\leq |A_s^{t_m}(k)|e^{-\theta(k)|x-(l_{s-1}-L)|} + |B_s^{t_m}(k)|e^{-\theta(k)|x-(l_s+L)|} \\
&\leq \rho^{\phi_s(m)}||c(0)||_\infty(k)\left(e^{-\theta(k)|x-(l_{s-1}-L)|} + e^{-\theta(k)|x-(l_s+L)|}\right) \\
&\leq \rho^{\phi_s(m)}\frac{|p^*(k)|+|q^*(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}\left(e^{-\theta(k)|x-(l_{s-1}-L)|} + e^{-\theta(k)|x-(l_s+L)|}\right).
\end{aligned} \tag{11}
$$

Let

$$
g(x,k) = \frac{|p^*(k)|+|q^*(k)|e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}\left(e^{-\theta(k)|x-(l_{s-1}-L)|} + e^{-\theta(k)|x-(l_s+L)|}\right). \tag{12}
$$

Thus, we have $|\hat{u}_s^{t_m}(x,k)| \leq g(x,k)$ for any $m \in \mathbb{N}$. We show next that $g(x,.) \in L^1(\mathbb{R})$. Since $|p^*(k)|, |q^*(k)| \leq M$, we have

$$
g(x,k) \leq M\frac{1+e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}\left(e^{-\theta(k)|x-(l_{s-1}-L)|} + e^{-\theta(k)|x-(l_s+L)|}\right) \tag{13}
$$

Thus,

$$
\begin{aligned}
\int_{-\infty}^{\infty}|g(x,k)|dk &\leq \int_{-\infty}^{\infty} M\frac{1+e^{-\theta(k)(W+2L)}}{1-e^{-2\sqrt{\eta}(W+2L)}}\left(e^{-\theta(k)|x-(l_{s-1}-L)|} + e^{-\theta(k)|x-(l_s+L)|}\right)dk \\
&= \frac{M}{1-e^{-2\sqrt{\eta}(W+2L)}}\int_{-\infty}^{\infty}\Big(e^{-\theta(k)|x-(l_{s-1}-L)|} + e^{-\theta(k)|x-(l_s+L)|} \\
&\quad + e^{-\theta(k)[W+2L+|x-(l_{s-1}-L)|]} + e^{-\theta(k)[W+2L+|x-(l_s+L)|]}\Big)dk \\
&\leq \frac{M}{1-e^{-2\sqrt{\eta}(W+2L)}}\left(\frac{2}{|x-(l_{s-1}-L)|} + \frac{2}{|x-(l_s+L)|}\right. \\
&\quad \left. + \frac{2}{W+2L+|x-(l_{s-1}-L)|} + \frac{2}{W+2L+|x-(l_s+L)|}\right).
\end{aligned} \tag{14}
$$

Note that for $x \in S_\varepsilon \cap [\ell_{s-1}, \ell_s]$ we have $|x-(l_{s-1}-L)|, |x-(l_{s-1}+L)| \geq 2L-\varepsilon$. Then, plugging these inequalities in (14), we obtain

$$
\int_{-\infty}^{\infty}|g(x,k)|dk \leq \frac{4M(W+6L-2\varepsilon)}{(1-e^{-2\sqrt{\eta}(W+2L)})(2L-\varepsilon)(W+4L-\varepsilon)}, \tag{15}
$$

i.e., $g(x,.) \in L^1(\mathbb{R})$. Consequently, for any $x \in S_\varepsilon \cap [\ell_{s-1}, \ell_s]$ there exists a $g(x,.) \in L^1(\mathbb{R})$ such that $|\hat{u}_s^{t_m}(x,k)| \leq g(x,k)$ for all $m \in \mathbb{N}$, and by the Lebesgue Dominated Convergence Theorem we have then that (8) holds.

The above argument was for $s = 2, ..., p-1$. Using the same argument but with $A_1^{t_m} = 0$ and $-\infty$ instead of $l_{s-1} - L$, we can see that (8) holds for $s = 1$; and, using the same argument but with $B_p^{t_m} = 0$ and ∞ instead of $l_s + L$, it can be shown that (8) holds for $s = p$.

Thus, from (11), (12), (15) we have $\forall x \in S_\varepsilon \cap [\ell_{s-1}, \ell_s]$ and $y \in \mathbb{R}$ that

$$|u_s^{t_m}(x,y)| \leq \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}|\hat{u}_s^{t_m}(x,k)|dk \leq \frac{\rho^{\phi_s(m)}}{\pi^2}\frac{M(W+6L-2\varepsilon)}{(1-e^{-2\sqrt{\eta}(W+2L)})(2L-\varepsilon)(W+4L-\varepsilon)}.$$

Consequently, $u_s^{t_m} \to 0$ uniformly in $S_\varepsilon \cap [\ell_{s-1}, \ell_s] \times \mathbb{R}$ as $m \to \infty$. Then, as explained in the outline of the proof, the values of the boundary conditions of each local problem go to zero as m goes to infinity, and therefore $\forall s \in \{1, ..., p\}$ we have $u_s^{t_m} \to 0$ in $\Omega^{(s)}$ as $m \to \infty$. Given that the sequence of time stamps was arbitrary, the theorem is proven. □

Remark 1: Note that the condition that $\hat{u}_s^0(x,k)$ is uniformly bounded in $k \in \mathbb{N}$ and $x \in S$ can be weakened to the condition that p^* and q^* be such that $g(x,.) \in L^1(\mathbb{R})$.

Remark 2: Note that, for synchronous and asynchronous iterations, for a given t_m, the value of $\phi_s(t_m)$ is, in general, different for each s, but they have a common lower bound, i.e., $\phi_s(t_m) \geq n_{\min}$, where $n_{\min} = \min_{s\in\{1,...,p\}}\{n_s\}$ and n_s is the local update number of process s. Also, for any s, the value of $\phi_s(t_m)$ can be much larger than n_s. For the synchronous case all the local update numbers are equal to the global iteration number, therefore, $n_{\min}$ is just the (global) iteration number.

3 Conclusion

In [5], it was shown that the operator $\hat{T}$ mapping the coefficients of the Fourier transform of the error at one iteration to those at the next iteration is contracting in max norm. In this paper, we use this result to complete a proof that, for the operator $\Delta - \eta$, the asynchronous optimized Schwarz method converges for any initial approximation u_0 that gives an initial error with Fourier Transform (along the y direction) uniformly bounded on each of the artificial interfaces.

References

1. Victorita Dolean, Pierre Jolivet, and Frédéric Nataf. An introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation. *SIAM*, Philadelphia, 2015.
2. Andreas Frommer and Daniel B. Szyld. On asynchronous iterations. *Journal of Computational and Applied Mathematics*, 123:201–216, 2000.
3. Martin J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44:699–731, 2006.
4. Frédéric Magoulés, Abal-Kassim Cheik Ahamed, and Roman Putanowicz. Optimized Schwarz method without overlap for the gravitational potential equation on cluster of graphics processing unit. *International Journal of Computer Mathematics*, 93: 955-980, 2016.
5. Frédéric Magoulès, Daniel B. Szyld, and Cédric Venet. Asynchronous Optimized Schwarz Methods with and without Overlap. *Numerische Mathematik*, 137:199–227, 2017.

INTERNODES for elliptic problems

Paola Gervasio and Alfio Quarteroni

1 Introduction

The INTERNODES (INTERpolation for NOnconforming DEcompositionS) method is an interpolation based approach to solve partial differential equations by means of non-overlapping domain decomposition methods featuring non-conforming discretizations at the interfaces [2, 4]. The non-conformity at a given interface is induced by independent discretizations (as, e.g., h-fem or hp-fem) on two adjacent subdomains.

For second order elliptic problems, the well known mortar method uses a single L^2-projection operator *per* interface to match the non-conforming local solutions. INTERNODES instead employs two interpolation operators: the first one is used to enforce the continuity of the traces, the second one to enforce the conservation of fluxes across the interface.

In this paper we sketch the formulation of INTERNODES when it is applied to second-order elliptic problems on two-domains decompositions. Then we apply it to two test problems: the Kellogg's problem with piece-wise constant diffusion coefficients, and a problem featuring an infinitely differentiable solution. In both cases, the numerical results show that INTERNODES attains optimal rate of convergence (i.e., that of the best approximation error in each subdomain), as predicted by the theoretical estimate proved in [4].

Let $\Omega \subset \mathbb{R}^d$, with $d = 2,3$, be an open domain with Lipschitz boundary $\partial\Omega$, Ω_1 and Ω_2 be two non-overlapping subdomains with Lipschitz boundary such that $\overline{\Omega} = \overline{\Omega_1 \cup \Omega_2}$, and $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ be their common interface.

Paola Gervasio
DICATAM, Università degli Studi di Brescia, via Branze 38, 25123 Brescia (Italy), e-mail: paola.gervasio@unibs.it

Alfio Quarteroni
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano (Italy) and École Polytechnique Fédérale de Lausanne (EPFL) (honorary professor), e-mail: alfio.quarteroni@epfl.ch

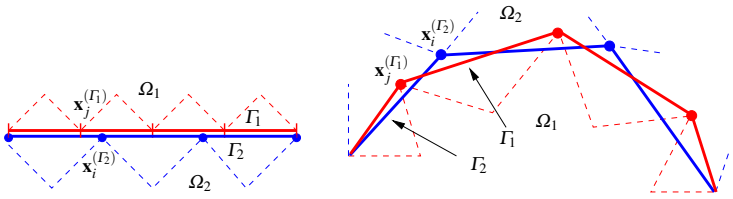


Fig. 1 Γ_1 and Γ_2 induced by the triangulations $\mathscr{T}_{1,h_1}$ and $\mathscr{T}_{2,h_2}$

Given $\alpha \in L^\infty(\Omega)$, $\mathbf{b} \in W^{1,\infty}(\Omega)$, $\gamma \in L^\infty(\Omega)$, and $f \in H^{-1}(\Omega)$, we look for u_1 in Ω_1 and u_2 in Ω_2 such that

$$\begin{cases} -\nabla\cdot(\alpha_k\nabla u_k) + \mathbf{b}\cdot\nabla u_k + \gamma u_k = f & \text{in } \Omega_k, \quad k = 1,2 \\ u_2 = u_1 & \text{on } \Gamma \\ \alpha_1 \frac{\partial u_1}{\partial \mathbf{n}_1} + \alpha_2 \frac{\partial u_2}{\partial \mathbf{n}_2} = 0 & \text{on } \Gamma \\ \text{boundary conditions} & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\mathbf{n}_k$ is the outward unit normal vector to $\partial\Omega_k$ and $\alpha_k = \alpha|_{\Omega_k}$. The transmission condition $(1)_2$ expresses the continuity of the solution across Γ, while $(1)_3$ enforces the conservation of normal fluxes across the interface, see [7].

2 Intergrid operators for non-conforming discretization

We consider two a-priori *independent families of triangulations*: $\mathscr{T}_{1,h_1}$ in Ω_1 and $\mathscr{T}_{2,h_2}$ in Ω_2, respectively. The meshes in Ω_1 and in Ω_2 can be non-conforming on Γ and characterized by different mesh-sizes h_1 and h_2. Moreover, different polynomial degrees p_1 and p_2 can be used to define the finite element spaces. Inside each subdomain Ω_k we assume that the triangulations $\mathscr{T}_{k,h_k}$ are affine, regular and quasi-uniform ([6, Ch.3]).

For $k = 1,2$, let $X_{k,h_k} = \{v \in C^0(\overline{\Omega_k}) : \ v_{|T} \in \mathbb{P}_{p_k}, \forall T \in \mathscr{T}_{k,h_k}\}$ be the usual Lagrangian finite element spaces associated with $\mathscr{T}_{k,h_h}$, while $Y_{k,h_k} = \{\lambda = v|_\Gamma, \ v \in X_{k,h_k}\}$ are the spaces of traces on Γ of functions in X_{k,h_k}, whose dimension is n_k.

We denote by Γ_1 and Γ_2 the internal boundaries of Ω_1 and Ω_2, respectively, induced by the triangulations $\mathscr{T}_{1,h_1}$ and $\mathscr{T}_{2,h_2}$. If Γ is a straight segment, then $\Gamma_1 = \Gamma_2 = \Gamma$, otherwise Γ_1 and Γ_2 can be different (see Fig. 1).

For $k = 1,2$, let $\{\mathbf{x}_1^{(\Gamma_k)}, \ldots, \mathbf{x}_{n_k}^{(\Gamma_k)}\} \in \overline{\Gamma}_k$ be the nodes induced by the mesh $\mathscr{T}_{k,h_k}$.

We introduce two independent operators that exchange information between the two independent grids on the interface Γ: $\Pi_{12} : Y_{2,h_2} \to Y_{1,h_1}$ and $\Pi_{21} : Y_{1,h_1} \to Y_{2,h_2}$.

If $\Gamma_1 = \Gamma_2$, Π_{12} and Π_{21} are the classical Lagrange interpolation operators defined by the relations:

$$(\Pi_{12}\mu_{2,h_2})(\mathbf{x}_i^{(\Gamma_1)}) = \mu_{2,h_2}(\mathbf{x}_i^{(\Gamma_1)}), \quad i = 1,\dots,n_1, \qquad \forall \mu_{2,h_2} \in Y_{2,h_2}, \tag{2}$$

$$(\Pi_{21}\mu_{1,h_1})(\mathbf{x}_i^{(\Gamma_2)}) = \mu_{1,h_1}(\mathbf{x}_i^{(\Gamma_2)}), \quad i = 1,\dots,n_2, \qquad \forall \mu_{1,h_1} \in Y_{1,h_1}. \tag{3}$$

If, instead, Γ_1 and Γ_2 are geometrical non-conforming, we define Π_{12} and Π_{21} as the Rescaled Localized Radial Basis Function (RL-RBF) interpolation operators introduced in formula (3.1) of [3]. More precisely, for $i = 1,\dots,n_k$ let $\tilde{\phi}_i^{(k)}(\mathbf{x}) = \phi(\|\mathbf{x}-\mathbf{x}_i^{(\Gamma_k)}\|, r) = \max\{0, (1-\|\mathbf{x}-\mathbf{x}_i^{(\Gamma_k)}\|/r)^4\}(1+4\|\mathbf{x}-\mathbf{x}_i^{(\Gamma_k)}\|/r)$ be the locally supported C^2 Wendland radial basis function [8] centered at $\mathbf{x}_i^{(\Gamma_k)}$ with radius $r > 0$. For any continuous function f on Ω, for $i = 1,\dots,n_k$ let $(\gamma_f^{(k)})_i \in \mathbb{R}$ be the solutions of the system

$$\sum_{i=1}^{n_k} (\gamma_f^{(k)})_i \tilde{\phi}_i^{(k)}(\mathbf{x}_j^{(\Gamma_k)}) = f(\mathbf{x}_j^{(\Gamma_k)}), \qquad j = 1,\dots,n_k$$

and set

$$(\Pi_{RBF}^{(k)} f)(\mathbf{x}) = \sum_{i=1}^{n_k} (\gamma_f^{(k)})_i \tilde{\phi}_i^{(k)}(\mathbf{x}).$$

Then, after setting $g(x) \equiv 1$, for any $\mu_{2,h_2} \in Y_{2,h_2}$ and $\mu_{1,h_1} \in Y_{1,h_1}$, the RL-RBF interpolation operators are defined by

$$(\Pi_{12}\mu_{2,h_2})(\mathbf{x}) = \frac{(\Pi_{RBF}^{(2)}\mu_{2,h_2})(\mathbf{x})}{(\Pi_{RBF}^{(2)} g)(\mathbf{x})}, \qquad (\Pi_{21}\mu_{1,h_1})(\mathbf{x}) = \frac{(\Pi_{RBF}^{(1)}\mu_{1,h_1})(\mathbf{x})}{(\Pi_{RBF}^{(1)} g)(\mathbf{x})}.$$

In both cases, the (rectangular) matrices associated with Π_{12} and Π_{21} are, respectively, $R_{12} \in \mathbb{R}^{n_1 \times n_2}$ and $R_{21} \in \mathbb{R}^{n_2 \times n_1}$ and they are defined by

$$\begin{aligned} (R_{12})_{ij} &= (\Pi_{12}\mu_j^{(2)})(\mathbf{x}_i^{(\Gamma_1)}) \quad i = 1,\dots,n_1,\ j = 1,\dots,n_2, \\ (R_{21})_{ij} &= (\Pi_{21}\mu_j^{(1)})(\mathbf{x}_i^{(\Gamma_2)}) \quad i = 1,\dots,n_2,\ j = 1,\dots,n_1, \end{aligned} \tag{4}$$

where $\{\mu_i^{(k)}\}$ are the Lagrange basis functions of Y_{k,h_k}, for $k = 1,2$ and $i = 1,\dots,n_k$.

Obviously, in the conforming case for which $\Gamma_1 = \Gamma_2$, $h_1 = h_2$ and $p_1 = p_2$, the interpolation operators Π_{12} and Π_{21} are the identity operator and $R_{12} = R_{21} = I$ (the identity matrix of size $n_1 = n_2$). Finally, let

$$(M_{\Gamma_k})_{ij} = (\mu_j^{(k)}, \mu_i^{(k)})_{L^2(\Gamma_k)}, \qquad k = 1,2, \tag{5}$$

the *interface mass matrices*. We notice that only information associated with the interface nodes (more precisely, the nodes coordinates) are needed to assemble both the interface mass matrices and the interpolation matrices for both the Lagrange and the RL-RBF interpolation approaches.

3 Mathematical foundation of INTERNODES for elliptic problems

Let us consider the transmission problem (1) and, for simplicity, we complete it with homogeneous Dirichlet boundary conditions on $\partial\Omega$. For $k=1,2$ we introduce the local spaces $V_k=\{v\in H^1(\Omega_k)\mid v=0 \text{ on } \partial\Omega\cap\partial\Omega_k\}$, $V_k^0=\{v\in V_k\mid v=0 \text{ on } \Gamma\}$, the bilinear forms $a_k: V_k\times V_k\to\mathbb{R}$: $a_k(u,v)=\displaystyle\int_{\Omega_k}(\alpha_k\nabla u\cdot\nabla v+(\mathbf{b}\cdot\nabla u)v+\gamma uv)d\Omega$, and the finite dimensional spaces $V_{k,h_k}=X_{k,h_k}\cap V_k$, $V_{k,h_k}^0=X_{k,h_k}\cap V_k^0$, and $\Lambda_{k,h_k}=\{\lambda=v|_\Gamma,\ v\in V_{k,h_k}\}$. Let $\mathscr{R}_k:\Lambda_{k,h_k}\to V_{k,h_k}$, s.t. $(\mathscr{R}_k\eta_{k,h_k})|_\Gamma=\eta_{k,h_k}$, $\forall\eta_{k,h_k}\in\Lambda_{k,h_k}$ be any linear and continuous discrete lifting from Γ_k to Ω_k (as, e.g., the finite element interpolant that is zero at all finite element nodes not lying on Γ_k). Finally, we denote by $\mathscr{I}_k$ the set of indices $i\in\{1,\dots,n_k\}$ of the nodes $\mathbf{x}_i^{(\Gamma_k)}$ of Γ_k.

In order to apply the INTERNODES method to problem (1), for any $v_{k,h_k}\in V_{k,h_k}$ and for $k=1,2$ we define the scalar quantities

$$\begin{aligned}&(r_v^{(k)})_i=a_k(v_{k,h_k},\mathscr{R}_k\mu_i^{(k)})-(f,\mathscr{R}_k\mu_i^{(k)})_{L^2(\Omega_k)}, && i\in\mathscr{I}_k,\\ &(z_v^{(k)})_j=\sum_{i\in\mathscr{I}_k}(M_{\Gamma_k}^{-1})_{ji}(r_v^{(k)})_i, && j\in\mathscr{I}_k,\end{aligned}\tag{6}$$

and the functions

$$(r_v)_{k,h_k}=\sum_{j\in\mathscr{I}_k}(z_v^{(k)})_j\mu_j^{(k)},\tag{7}$$

belonging to Λ_{k,h_k}. (The subscript v highlights the dependence of r on v.)

Remark 1. When non-homogeneous Dirichlet boundary conditions are assigned on $\partial\Omega$, we can recover the homogeneous case by a lifting of the Dirichlet data, so that only the right hand side has to be modified (see, e.g., [6]).

The weak form of INTERNODES applied to (1) reads: find $u_{1,h_1}\in V_{1,h_1}$ and $u_{2,h_2}\in V_{2,h_2}$ such that

$$\begin{cases}a_k(u_{k,h_k},v_{k,h_k})=(f,v_{k,h_k})_{L^2(\Omega_k)} & \forall v_{k,h_k}\in V_{k,h_k}^0, \quad k=1,2\\ u_{2,h_2}=\Pi_{21}u_{1,h_1} & \text{on } \Gamma_2,\\ (r_u)_{1,h_1}+\Pi_{12}(r_u)_{2,h_2}=0 & \text{on } \Gamma_1.\end{cases}\tag{8}$$

For $k=1,2$, $(r_u)_{k,h_k}\in Y_{k,h_k}$ are the so-called *residuals* at the interface Γ_k. In fact they are the discrete fluxes across the interface, i.e., they represent the approximations of $\alpha_k\partial u_k/\partial\mathbf{n}_k$ on Γ_k.

Remark 2. The values $(r_u^{(k)})_i$ are not the coefficients of $(r_u)_{k,h_k}$ w.r.t. the Lagrange basis $\{\mu_i^{(k)}\}$ (on which we can apply the interpolation). Rather, they are the coefficients of $(r_u)_{k,h_k}$ w.r.t. the dual basis $\{\psi_i^{(k)}\}_{i=1}^{n_k}$ of Y'_{k,h_k} defined by the relations $(\psi_i^{(k)},\mu_j^{(k)})_{L^2(\Gamma_k)}=\delta_{ij}$, for $i,j=1,\dots,n_k$ (δ_{ij} is the Kronecker delta), precisely,

$$(r_u)_{k,h_k} = \sum_{i \in \mathscr{I}_k} (r_u)_i^{(k)} \psi_i^{(k)}.$$

Y_{k,h_k} and Y'_{k,h_k} are identical linear spaces and it can be proved that $\psi_i^{(k)} = \sum_{j \in \mathscr{I}_k} (M_{\Gamma_k}^{-1})_{ji} \mu_j^{(k)}$ for any $i \in \mathscr{I}_k$, therefore (7) follows. The interface mass matrix M_{Γ_k} and its inverse play the role of transfer matrices from the Lagrange basis to the dual one and viceversa, respectively.

Denoting by z_k and r_k the arrays whose entries are the values $(z_u^{(k)})_j$ and $(r_u^{(k)})_i$, respectively, it follows that $\mathsf{z}_k = M_{\Gamma_k}^{-1} \mathsf{r}_k$.

Then, the algebraic form of the interface condition $(8)_3$ reads

$$M_{\Gamma_1}^{-1} \mathsf{r}_1 + R_{12} M_{\Gamma_2}^{-1} \mathsf{r}_2 = 0,$$

or, equivalently, $\mathsf{r}_1 + M_{\Gamma_1} R_{12} M_{\Gamma_2}^{-1} \mathsf{r}_2 = 0$.

For $k = 1,2$ let u_k denote the array of the Lagrange coefficients of u_{k,h_k} at the nodes of $\mathscr{T}_{k,h_k}$ and $\boldsymbol{\lambda}_k$ the array of the Lagrange coefficients of u_{k,h_k} at the nodes of $\mathscr{T}_{k,h_k} \cap \Gamma_k$. Denoting by A_k the finite element stiffness matrices associated with the discretization of $(8)_1$, the algebraic form of (8) reads:

$$\begin{cases} A_k \mathsf{u}_k = \mathsf{f}_k, & k = 1,2, \\ \boldsymbol{\lambda}_2 = R_{21} \boldsymbol{\lambda}_1, \\ \mathsf{r}_1 + M_{\Gamma_1} R_{12} M_{\Gamma_2}^{-1} \mathsf{r}_2 = 0, \end{cases} \tag{9}$$

with $\mathsf{u}_{k|\Gamma_k} = \boldsymbol{\lambda}_k$.

Under the assumptions that problem (1) is well posed (see, e.g., [6, 4]) the following convergence theorem, assessing the optimal error bound for the INTERNODES method, is proved in [4].

Theorem 1. *Assuming that $u \in H^s(\Omega)$, with $s > 3/2$, $\lambda = u_{|\Gamma} \in H^\sigma(\Gamma)$, with $\sigma > 1$, $(\alpha_k \partial u_2 / \partial \mathbf{n}_2) \in H^\nu(\Gamma)$, with $\nu > 0$, if $p_k \geq 1$ is the finite element polynomial degree in Ω_k, $k = 1,2$, and Lagrange interpolation is used to define Π_{12} and Π_{21}, there exist $\frac{1}{2} \leq q < 1$ and $\frac{3}{2} \leq z < 2$ s.t.*

$$\begin{aligned} \|u - u_h\|_* \lesssim\ & h_1^{\ell_1 - 1} \|u\|_{H^s(\Omega_1)} + h_2^{\ell_2 - 1} \|u\|_{H^s(\Omega_2)} \\ & + \left(h_1^{\rho_1 - 1/2} + h_2^{\rho_2 - 1/2} + h_1^{\rho_1 - 1/2} \left(\frac{h_2}{h_1} \right)^q \right) \|\lambda\|_{H^\sigma(\Gamma)} \\ & + \left(h_1^{\zeta_1 + 1/2} + h_2^{\zeta_2 + 1/2} + h_1^{\zeta_1 + 1/2} \left(\frac{h_1}{h_2} \right)^z \right) \|r_2\|_{H^\nu(\Gamma)}, \end{aligned}$$

with $\ell_k = \min(s, p_k + 1)$, $\rho_k = \min(\sigma, p_k + 1)$, $\zeta_k = \min(\nu, p_k + 1)$, and being $\|v\|_* = \{\|v\|^2_{H^1(\Omega_1)} + \|v\|^2_{H^1(\Omega_2)}\}^{1/2}$ the broken norm on Ω.

Remark 3. Π_{21} is used to match the traces, while Π_{12} is used to match the residuals, i.e. the fluxes.

Using instead only one intergrid interpolation operator would not guarantee an accurate non-conforming method; for example using only Π_{21} yields to the so-called *point wise matching* discussed, e.g., in [1]. At the algebraic level the latter approach uses only the matrix R_{21} and its transpose R_{21}^T, whereas INTERNODES uses both R_{21} and R_{12}.

Remark 4 (On the conservation of fluxes). The conservation of fluxes across the interface at the discrete level is enforced by the interface condition $(8)_3$. As this property depends on the interpolation operator Π_{12}, that in turns depends on the choice of the local subspaces, the flux jump vanishes, as h_1 and h_2 go to zero, with the same order of the broken norm of the error.

Remark 5. The INTERNODES method can be generalized to decompositions with more than two subdomains, possibly featuring internal cross-points (i.e., points shared almost among three subdomains). We refer to [4, Sect. 6] for a detailed description of the algorithm. What follows is a sketch of the generalization of INTERNODES when $\Omega \subset \mathbb{R}^2$. Let Ω_k and Ω_ℓ be two generic subdomains such that $\Gamma_{k\ell} = \partial\Omega_k \cap \partial\Omega_\ell$ is neither empty nor reduced to a vertex, while $\gamma_k^{(i)}$ and $\gamma_\ell^{(j)}$ denote the edges of $\partial\Omega_k$ and $\partial\Omega_\ell$, respectively, such that $\Gamma_{k\ell} = \gamma_k^{(i)} \cap \gamma_\ell^{(j)}$.

Two typical situations can occur: the end-points of $\gamma_k^{(i)}$ coincide with those of $\gamma_\ell^{(j)}$ (as in Fig. 2), or not (as in Fig. 3). In the first case, each interface $\Gamma_{k\ell}$ is handled as in the case of only two subdomains and we build couples of intergrid matrices $R_{\ell k}$ and $R_{k\ell}$ from $\gamma_k^{(i)}$ to $\gamma_\ell^{(j)}$ and viceversa, as done in Sect. 2. In the second case, let us suppose that the measure of $\gamma_k^{(i)}$ is larger than that of $\gamma_\ell^{(j)}$. Here all the basis functions living on $\gamma_k^{(i)}$ whose support has non-empty intersection with $\gamma_\ell^{(j)}$ must be taken into account when building the interpolation matrices $R_{\ell k}$ and $R_{k\ell}$ and the interface mass matrices $M_{\ell k}$ and $M_{k\ell}$. Alternatively, one can build both the interface mass matrices and the interpolation matrices on the larger interface $\gamma_k^{(i)}$ by assembling the contributions arising from all the shorter edges of the subdomains adjacent to Ω_k on the other side of $\gamma_k^{(i)}$.

Remark 6. Robin conditions could be used instead of Neumann ones. The formulation of INTERNODES would not change, provided the interface conditions are imposed weakly (as *natural* conditions). As a matter of fact, natural interface conditions are automatically accounted for when evaluating the discrete residuals of the differential problem as done in (6).

4 Numerical results: the Kellogg's test case

We test INTERNODES on a very challenging problem whose solution features low regularity. The so-called Kellogg's function (see, e.g., [5]) is an exact weak solution of the elliptic problem

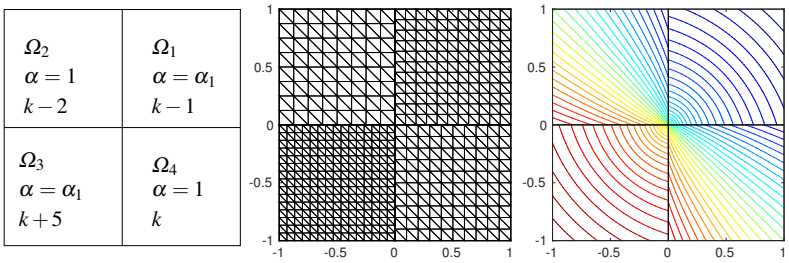


Fig. 2 At left, the decomposition of Ω into four subdomains. In the middle, the nonconforming $\mathbb{P}_1$ meshes for $k = 10$. At right, the Kellogg's solution with $\gamma = 0.4$ and $\alpha_1 = 9.472135954999585$ computed by INTERNODES and $\mathbb{P}_1$

$$\begin{cases} -\nabla \cdot (\alpha \nabla u) = 0 & \text{in } \Omega = (-1,1)^2 \\ \text{Dirichlet boundary conditions} & \text{on } \partial\Omega, \end{cases} \tag{10}$$

with piece-wise constant coefficient α: $\alpha = \alpha_1 > 0$ in the first and the third quadrants, and $\alpha = 1$ in the second and in the fourth ones. It can be written in terms of the polar coordinates r and θ as $u(r,\theta) = r^\gamma \mu(\theta)$, where $\gamma \in (0,2)$ is a given parameter, while $\mu(\theta)$ is a $2\pi-$periodic continuous function (more regular only when $\gamma = 1$). The case $\gamma = 1$ is trivial since the solution is a plane. The positive value α_1 depends on γ and on two other real parameters σ and ρ. The set $\{\alpha_1, \gamma, \sigma, \rho\}$ must satisfy a nonlinear system (see formula (5.1) of [5]). In particular we fixed $\rho = \pi/4$.

When $\gamma \neq 1$, $u \in H^{1+\gamma-\varepsilon}(\Omega)$, for any $\varepsilon > 0$; the solution features low regularity at the origin and its normal derivatives to the axis are discontinuous.

We solve problem (10) by applying INTERNODES to the 4-subdomains decomposition induced by the discontinuity of α and by using either $\mathbb{P}_1$ or $\mathbb{Q}_2$ finite elements in each subdomain (see the $\mathbb{P}_1$ mesh in Fig. 2). The meshes at the interfaces are non-conforming as shown in Figure 2, more precisely given $k \in \mathbb{N}$, the subdomains mesh-sizes are: $h_1 = 1/(k-1)$, $h_2 = 1/(k-2)$, $h_3 = 1/(k+5)$ and $h_4 = 1/k$.

By refining the meshes (we cycle on $k = 20,\ 40,\ 80,\ 160$), we measure the convergence order of INTERNODES on the Kellogg's solution for different values of the parameter γ. The results are shown in Table 1 and the convergence estimate provided by Theorem 1 for two subdomains is here confirmed, although this test case involves four subdomains instead of two.

We highlight that, although INTERNODES is based on interpolation operators rather than projections (as in the mortar methods), the best approximation error of the finite element discretization is preserved and not downgraded.

Table 1 Convergence orders of INTERNODES for the Kellogg's test solution. The case $\gamma = 0.4$ is not covered by the convergence Theorem 1 since $s < 3/2$ and $\sigma < 1$. $\min\{\ell-1, \rho-1/2, \zeta+1/2\}$ is the expected convergence order provided by Theorem 1, the measured convergence orders are shown in the last two columns

γ	s	σ	ν	$\min\{\ell-1, \rho-1/2, \zeta+1/2\}$	$\mathbb{P}_1$ order	$\mathbb{Q}_2$ order
0.4	$1.4-\varepsilon$	$0.9-\varepsilon$	$0.4-\varepsilon$	$0.4-\varepsilon$	0.363	0.429
0.6	$1.6-\varepsilon$	$1.1-\varepsilon$	$0.6-\varepsilon$	$0.6-\varepsilon$	0.574	0.651
1.4	$2.4-\varepsilon$	$1.9-\varepsilon$	$1.4-\varepsilon$	1 for $\mathbb{P}_1$, $1.4-\varepsilon$ for $\mathbb{Q}_2$	0.955	1.394
1.8	$2.8-\varepsilon$	$2.3-\varepsilon$	$1.8-\varepsilon$	1 for $\mathbb{P}_1$, $1.8-\varepsilon$ for $\mathbb{Q}_2$	0.949	1.615

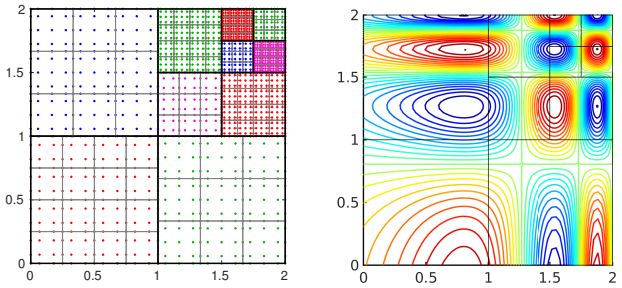


Fig. 3 At left, a partition of the computational domain into 10 subdomains; in each subdomain the quad hp-fem mesh is plotted, different colours refer to different subdomains. At right, the corresponding INTERNODES solution

5 Numerical results: infinite differentiable solution

Let us consider the problem (1) with $\alpha = 1$, $\mathbf{b} = [1,1]$, $\gamma = 1$ on $\Omega = (0,2)^2$. The boundary data and the function f are such that the exact solution is $u(x,y) = \sin(3\pi\exp(3(x-2)/2))\cos(3\pi\exp(3(y-2)/2))$.

A decomposition of $\Omega = (0,2)^2$ in 10 subdomains as in Fig. 3 is considered, and independent triangulations in each Ω_k are designed so that on each interface both polynomial non-conformity and geometric non-conformity may occur. Either $\mathbb{P}_1$ and quadrilateral hp-fem ($\mathbb{Q}_p$) are used to approximate the numerical solution. A non-conforming grid, obtained with $\mathbb{Q}_p$ discretizations in each subdomain, is shown in Fig. 3, left. In order to guarantee full non-conformity on each interface, we have set on two adjacent domains the polynomial degree equal to either $p = 3$ or $p = 4$ and the local mesh size equal to either $h = 1/4$ or $h = 1/3$. In Fig. 3, right, the corresponding numerical solution computed by INTERNODES is shown.

In order to measure the errors in broken norm, we take the same polynomial degree p in each subdomain and we consider only geometric non-conformity as in Fig. 3, left, but with a variable number k (or $k-1$) of elements (more precisely, $k = 4$ in Fig. 3, left). The reference parameter is the mesh size $h = 1/k$ of the left-bottom subdomain. In Fig. 4, the errors in broken norm are reported, w.r.t. to both h and p.

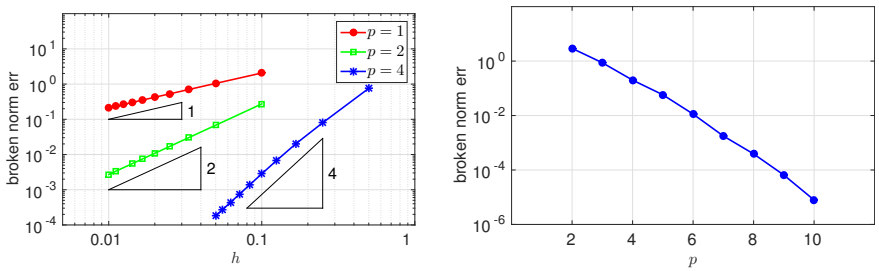


Fig. 4 At left, the broken norm error w.r.t. the mesh-size h of the bottom-left subdomain, p is fixed. At right, the broken norm error w.r.t. p, here the meshes sizes are fixed: that of the left-bottom subdomain is $h = 1/4$

The error behaviour versus h (see Fig. 4 left) agrees with the theoretical estimate of Theorem 1, for which we expect $\|u - u_h\|_* \leq c(u)h^p$ (in this case $p = 1, 2, 4$), as u is infinitely differentiable.

The convergence rate vs p shown in Fig. 4, right, is more than algebraic, as typical in hp-fem.

References

1. C. Bernardi, Y. Maday, and A.T. Patera. A new nonconforming approach to domain decomposition: the mortar element method. In *Nonlinear partial differential equations and their applications. Collège de France Seminar, Vol. XI (Paris, 1989–1991)*, volume 299 of *Pitman Res. Notes Math. Ser.*, pages 13–51. Longman Sci. Tech., Harlow, 1994.
2. S. Deparis, D. Forti, P. Gervasio, and A. Quarteroni. INTERNODES: an accurate interpolation-based method for coupling the Galerkin solutions of PDEs on subdomains featuring non-conforming interfaces. *Computers & Fluids*, 141:22–41, 2016.
3. S. Deparis, D. Forti, and A. Quarteroni. A rescaled localized radial basis function interpolation on non-Cartesian and nonconforming grids. *SIAM J. Sci. Comput.*, 36(6):A2745–A2762, 2014.
4. P. Gervasio and A. Quarteroni. Analysis of the INTERNODES method for non-conforming discretizations of elliptic equations. Technical report, MATHICSE, EPFL, Lausanne (Switzerland), 2016. Submitted.
5. P. Morin, R.H. Nochetto, and K.G. Siebert. Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2):466–488 (electronic), 2000.
6. A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer Verlag, Heidelberg, 1994.
7. A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford University Press, 1999.
8. H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.

A Nonlinear Elimination Preconditioned Newton Method with Applications in Arterial Wall Simulation

Shihua Gong and Xiao-Chuan Cai

Abstract Arterial wall can be modeled by a quasi-incompressible, anisotropic and hyperelastic equation that allows large deformation. Most existing nonlinear solvers for the steady hyperelastic problem are based on pseudo time stepping, which often requires a large number of time steps especially for the case of large deformation. It is also reported that the quasi-incompressibility and high anisotropy have negative effects on the convergence of both Newton's iteration and the linear Jacobian solver. In this paper, we propose and study a nonlinearly preconditioned Newton method based on nonlinear elimination to calculate the steady solution directly without pseudo time integration. We show numerically that the nonlinear elimination preconditioner accelerates Newton's convergence in cases with large deformation, quasi-incompressibility and high anisotropy.

1 Introduction

Some biological soft tissues, such as the arterial wall, are quasi-incompressible and are reinforced by collagen fibers, which induce the anisotropy in the mechanical response. Polyconvex hyperelastic models [2, 4], which are based on polyconvex energy-stored functions, provide a unified framework to describe the quasi-incompressibility, the anisotropy and the nonlinearly elastic behavior of arterial walls in the regime allowing large deformations. By using finite element discretizations [3] for these models and Newton-type nonlinear solvers, numerical simulation of arterial walls becomes a promising approach in clinical diagnosis and treatment

S. Gong
Beijing International Center for Mathematical Research, Peking University, Beijing 100871, P. R. China, e-mail: gongshihua@pku.edu.cn

X.-C. Cai
Department of Computer Science, University of Colorado Boulder, Boulder, CO 80309, USA, e-mail: cai@cs.colorado.edu

assistance. However, the design of robust nonlinear and linear solvers is a challenging problem due to the sophisticated mechanical properties of arterial walls.

In [5], the authors consider several material models for arterial walls in order to study the mechanical response and the influence on the nonlinear iteration as well as on the finite element tearing and interconnecting-dual primal (FETI-DP) iterative linear solver. The stagnation of Newton's method is observed for some parameter sets. In order to cope with the quasi-incompressible condition, an augmented Lagrange approach is proposed in [6]. The penalty parameter for the incompressibility can be chosen much smaller and therefore the resulting linear systems have better properties. Both nonlinear solvers mentioned above are based on pseudo time stepping, which often requires a large number of global nonlinear iterations especially for the case of large deformation.

To accelerate the convergence of the nonlinear iteration, we consider a nonlinearly preconditioned Newton method based on nonlinear elimination to calculate the solution directly without pseudo time integration. The nonlinear elimination method is first proposed and analyzed in [12] and then developed in [7, 11] for the problems with high local nonlinearity. For our cases of hyperelasticity, we numerically observe that the variables with stronger nonlinearity are not fixed, but change as the propagation of the elastic wave. Thus, we adaptively detect the variables and equations with stronger nonlinearity by the residuals. After eliminating these equations, the approximate solution is more accurate in some key locations of the elastic wave and therefore the global Newton's method converges better.

2 Modeling and Discretization

In this section, we discuss a hyperelastic model for arterial walls and its finite element discretization. First, we introduce some basic notations in continuum mechanics. The body of interest in the reference configuration is denoted by $\hat{\Omega} \in \mathbb{R}^3$, parameterized in $\hat{x}$, and the current configuration by $\Omega \in \mathbb{R}^3$, parametrized by x. The deformation map $\phi : \hat{\Omega} \mapsto \Omega$ is a differential isomorphism between the reference and current configuration. The deformation gradient F is defined by $F(\hat{x}) = \nabla\phi(\hat{x})$ with the Jacobian $J(\hat{x}) = \det F(\hat{x}) > 0$. The right Cauchy-Green tensor is defined as $C = F^T F$.

The hyperelastic materials postulate the existence of a so-called store-energy function ψ, defined per unit reference volume. According to the axiom of material frame-indifference [8], the energy functional depends on the Cauchy-Green tensor, i.e., $\psi = \psi(C)$. The first and second Piola-Kirchhoff stress tensor can be derived as $P = FS, \quad S = 2\partial_C \psi(C)$. And then the Cauchy stress is given by $\sigma = J^{-1} FSF^T$. The balance of the momentum is governed by the following partial differential equation

$$\mathrm{div} P = -f,$$

plus appropriate boundary condition. Here f is the body force vector.

We focus on the polyconvex energy functional proposed in [4],

$$\begin{aligned}\psi_A &= \psi^{isochoric} + \psi^{volumetric} + \psi^{ti} \\ &:= c_1\left(\frac{I_1}{I_3^{1/3}} - 3\right) + \varepsilon_1\left(I_3^{\varepsilon_2} + \frac{1}{I_3^{\varepsilon_2}} - 2\right) + \sum_{i=1}^{2}\alpha_1\left\langle I_1 J_4^{(i)} - J_5^{(i)} - 2\right\rangle^{\alpha_2},\end{aligned} \tag{1}$$

which models the quasi-incompressible and fibre-enforcing arterial wall. Here, $\langle b\rangle$ denotes the Macaulay brackets defined by $\langle b\rangle = (|b|+b)/2$, with $b \in R$. And I_1, I_2, I_3 are the principal invariants of C; i.e. $I_1 := \mathrm{tr}C, I_2 := \mathrm{tr}[\mathrm{cof}\ C], I_3 := \det C$, where $\mathrm{cof}\, C = (\det C)C^{-T}$. The additional mixed invariants $J_4^{(i)}, J_5^{(i)}$ characterize the anisotropic behavior of arterial wall and are defined as $J_4^{(i)} := \mathrm{tr}[CM^{(i)}], \quad J_5^{(i)} := \mathrm{tr}[C^2M^{(i)}], \quad$ for $i = 1:2$, where $M^{(i)} := a^{(i)} \otimes a^{(i)}, i = 1,2$ are the structural tensors with $a^{(i)}, i = 1,2$ denoting the direction fields of the embedded collagen fibers.

The polyconvexity condition in the sense of [2] is the essential condition to ensure the existence of energy minimizers. There are three parts in ψ_A:

- $\psi^{isochoric}$ is the isochoric part of the isotropic energy. Similar to the Neo-Hookean material, c_1 is stress-type coefficient with upper and lower bounds.
- $\psi^{volumetric}$ is the penalty function to account for the quasi-incompressibility. The coefficients $\varepsilon_1, \varepsilon_2$ would be very large for the incompressible material.
- ψ^{ti} is the transversely isotropic part. The anisotropy comes from the exponential stiffening of the fibers when increasing loads are applied. Relative large coefficients α_1, α_2 indicate large anisotropy.

According to [3], the lowest-order Lagrange finite element with linear shape functions is not sufficient to provide a good approximation for the arterial wall stresses, whereas for the Lagrange finite elements or F-bar formulations with quadratic shape functions, suitable results are obtained. Instead of concerning about the stress, we focus on the nonlinear solvers for the resulting system. Thus, for simplicity, we use the $\mathscr{P}_1$ Lagrange finite element to approximate the displacement.

3 Inexact Newton Method with Nonlinear Preconditioning

With a slight abuse of notation, we denote the nonlinear system after the discretization as described above

$$F(u^*) = 0$$

where $F : \mathrm{R}^n \mapsto \mathrm{R}^n$. Inexact Newton (IN) algorithms [9, 10] are commonly used for solving such system and can briefly be described here. Suppose $u^{(k)}$ is the current approximate solution, a new approximate solution $u^{(k+1)}$ can be computed through

$$u^{(k+1)} = u^{(k)} + \lambda^{(k)} p^{(k)},$$

where the inexact Newton direction $p^{(k)}$ satisfies

$$\|F(u^{(k)}) + F'(u^{(k)})p^{(k)}\| \leq \eta_k \|F(u^{(k)})\|.$$

Here $\eta_k \in [0,1)$ is a scalar that determines how accurately the Jacobian system needs to be solved, and $\lambda^{(k)}$ is another scalar that determines how far one should go in the selected direction.

3.1 Nonlinear Elimination

It is reported in [5, 6], the incompressibility and large anisotropy have a negative effect on the convergence of both Newton's iteration and the Jacobian solver. To accelerate Newton's convergence, we introduce a nonlinear elimination preconditioner [7, 11, 12], which balances the nonlinearity of the global problem by solving the subproblems defined in the subdomains or subspaces. Let $S = \{1, \cdots, n\}$ be an index set; i.e., one integer for each unknown u_i and residual F_i. We choose a subset $S_b \subset S$ of the indices corresponding to the "bad" degrees of freedom (d.o.f.), of which the nonlinearity is dominant. The corresponding subspace is denoted by

$$V_b = \{v \mid v = (v_1, \cdots, v_n)^T \in \mathrm{R}^n, v_k = 0, \text{ if } k \notin S_b\}.$$

The corresponding restriction operator is denoted by $R_b \in \mathrm{R}^{n \times n}$, whose kth column is either zero if $k \notin S_b$ or the kth column of the indentity matrix $I_{n\times n}$. Thus the subspace and the corresponding restriction for the "good" d.o.f. are denoted by V_g and $R_g = I_{n\times n} - R_b$.

Given an approximate solution u and a sub index set S_b, the nonlinear elimination algorithm finds the correction by approximately solving $u_b \in V_b$,

$$F_b(u_b) := R_b F(u_b + u) = 0. \tag{2}$$

The new approximate solution is then updated as $w = u_b + u$. It is easy to see that the Jacobian of the sub nonlinear problem (2) is $J_b(u_b) = R_b J(u_b + u) R_b^T$. Here $J = F' = \left(\frac{\partial F_i}{\partial u_j}\right)_{n\times n}$ is the Jacobian of F.

Suppose we are at the iteration k and $u^{(k)}$ is the current approximation, the inexact Newton algorithm with nonlinear elimination is described as below

---

Algorithm 1. (IN-NE)

Step 1. Compute the next approximate solution $u^{(k+1)}$ by solving the following nonlinear system

$$F(u) = 0$$

with one step of IN iteration using $u^{(k)}$ as the initial guess. If the global convergent condition is satisfied, stop. Otherwise, go to Step 2.

Step 2. (Nonlinearity checking)

2.1 If $\|F(u^{(k+1)})\| < \rho_1 \|F(u^{(k)})\|$, go to Step 1.
2.2 Finding "bad" d.o.f. by

$$S_b := \{ j \in S \mid |F_j(u^{(k+1)})| > \rho_2 \|F(u^{(k+1)})\|_\infty \}.$$

And extend S_b to S_b^δ by adding the neighbor d.o.f..
2.3 If $\#(S_b^\delta) < \rho_3 n$, go to Step 3. Otherwise, go to Step 1.

Here $\rho_1, \rho_2, \rho_3 \in (0,1)$ and $\delta \in Z_+$ are pre-chosen constants.

Step 3. Compute the correction $u_b^\delta \in V_b$ by solving the sub nonlinear system approximately

$$F_b^\delta(u_b^\delta) := R_b^\delta F(u_b^\delta + u^{(k+1)}) = 0,$$

with an initial guess $u_b^\delta = 0$ and a relative tolerance $\text{tol} = \max(\gamma_a, \gamma_r \|R_b^\delta F(u^{(k+1)})\|)$. If $\|F(u_b^\delta + u^{(k+1)})\| < \|F(u^{(k+1)})\|$, accept the correction and update $u^{(k+1)} \leftarrow u_b^\delta + u^{(k+1)}$. Go to Step 1.

There are three tolerance parameters in the nonlinear checking step: ρ_1 is the tolerance for the reduction of the residual norm, ρ_2 is the tolerance to pick up the bad d.o.f. and ρ_3 is the tolerance to limit the size of the subproblem. In Step 3, we only accept the correction by nonlinear elimination if the residual norm decreases. But in practice, if the norm of the corrected residual does not decrease for 3 successive steps, we choose to accept the correction without checking the residual.

Different to the nonlinear elimination method proposed in [12], where the authors fix for all steps the set of equations to eliminate, we construct adaptively the index set S_b by the residual $F(u^{(k+1)})$. Actually, the residual can be viewed as a measurement of the Hessian of F by the Taylor expansion,

$$\begin{aligned} F(u^{(k+1)}) &= F(u^{(k)}) + F'(u^{(k)})p^{(k)} + \langle F''(u^{(k)} + \theta p^{(k)})p^{(k)}, p^{(k)} \rangle \\ &\approx \langle F''(u^{(k)} + \theta p^{(k)})p^{(k)}, p^{(k)} \rangle, \end{aligned}$$

since the Jacobian system is solved approximately. From this perspective, eliminating the equations with large residual is a way to control the higher order terms of F such that it can be linearly approximated much better during the global Newton iteration. However, the nonlinear elimination just on the equations with indices in S_b could lead to thrashing (i.e., the norm of the residual $\|F\|$ could become larger due to the boundary effect). To ease this phenomenon, we extend the index set S_b to S_b^δ by adding the neighbor d.o.f, of which the distances to S_b are smaller than δ.

4 Numerical Results

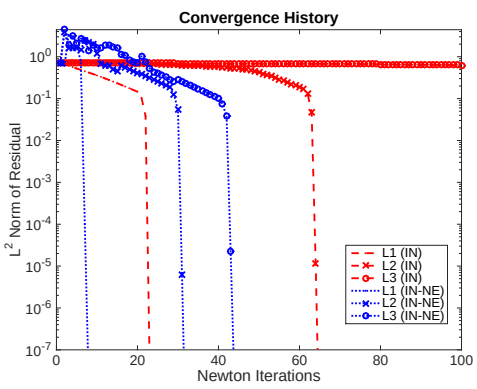
We implement the discretization for hyperelasticity and the nonlinear solvers described in the previous sections by using FEniCS [13] and PETSc [1], respectively.

Based on the parameter sets of the model ψ_A in Table. 1, we propose three test examples to investigate the performance of nonlinear elimination for the materials with large deformation, quasi-incompressibility and high anisotropy. In all of the tests, the backtracking line search strategy is used to determine the maximum amount to move along the search direction computed by a direct solver.

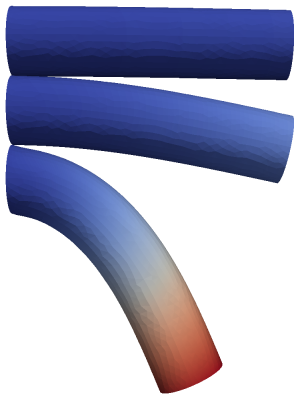
Set	Layer	c_1	ε_1	ε_2(-)	α_1	α_2	Purpose
L	–	1.e3	1.e3	1.0	0.0	0.0	Deformations by different pulls
C1	–	1.e3	1.e3	1.0	0.0	0.0	Different penalties for compressiblity
C2	–	1.e3	1.e4	1.0	0.0	0.0	
C3	–	1.e3	1.e5	1.0	0.0	0.0	
A1	Adv.	7.5	100.0	20.0	1.5e10	20.0	Anisotropic arterial walls
	Med.	17.5	100.0	50.0	5.0e5	7.0	
A2	Adv.	6.6	23.9	10	1503.0	6.3	
	Med.	17.5	499.8	2.4	30001.9	5.1	
A3	Adv.	7.8	70.0	8.5	1503.0	6.3	
	Med.	9.2	360.0	9.0	30001.9	5.1	

Table 1: Model parameter sets [5, 6] of ψ_A

Example 1. This example simulates the deformations of a cylindrical rod by different pulls. We fix one end of the rod and then pull it down from the other end. The material parameters are given in Set L of Table 1. It is an isotropic model since $\alpha_1 = 0.0$. The deformations by three different pulls $L_1 = 1.e1$ Pa, $L_2 = 1.e2$ Pa and $L_3 = 1.e3$ Pa are plotted in Fig. 1b. The convergence history of the Newton iteration with nonlinear elimination (IN-NE) is shown in Fig. 1a. We compare the results with those obtained by using a standard inexact Newton (IN) method. The blue lines are for the IN-NE algorithm while the red lines for the IN method. As indicated by Fig. 1a, the nonlinear elimination method accelerates the convergence of the Newton iteration even for the case of large deformation.



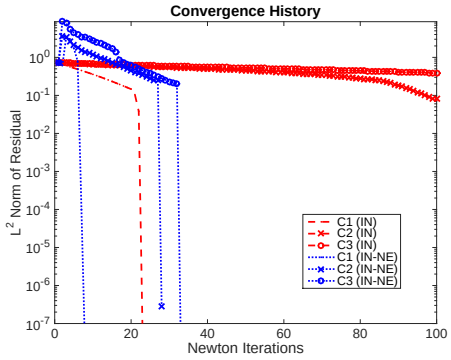
(a) Convergence history of IN and IN-NE



(b) Deformations by different pulls

Fig. 1: Numerical results of Example 1.

Example 2. This example studies the performance of nonlinear elimination for the cases of different compressibility. The parameters are given in the sets C1,C2 and C3 of Table 1. For consistency with linear elasticity, $C_1 = \frac{\mu}{2}, \varepsilon_1 = \frac{\kappa}{2}$, where μ, κ are the shear and bulk modulus. The Poisson ratio can be computed by $\nu = \frac{3\kappa - 2\mu}{2(3\kappa+\mu)}$; see Table 2. We use the same setting of the geometry and the boundary conditions with that of the previous example. Fig. 2 shows the superiority of the nonlinear elimination in the quasi-incompressible case.

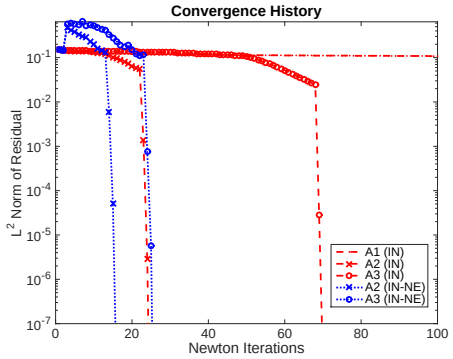


Set	Poisson's Ratio
C1	0.125
C2	0.452
C3	0.495

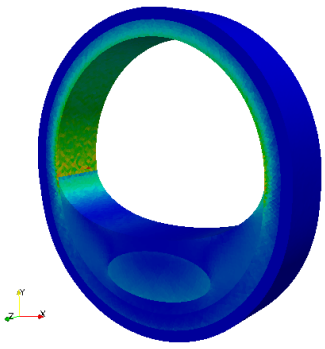
Table 2: Poisson's ratio of materials C1, C2 and C3.

Fig. 2: Convergence histories for Example 2.

Example 3. We consider an artificial arterial segment with a plaque and fibre-enforcing layers. The problem setting, including the geometry and boundary conditions, originates from [5]. More precisely, a pressure of up to 24 kPa ($<$ 180 mmHg) is applied to the interior of the arterial segment, of which the von Mises stress is shown in Fig. 3b. The parameter sets A1 and A2 of Table 1 are adjusted in [5] to fit the experiment data, and A3 in [6] with slight modification. The convergence histories of IN and IN-NE are shown in Fig. 3b. Similar to the previous examples, the nonlinear elimination increases the residual at the first few steps of Newton's iteration, but then the iteration converges faster.



(a) Convergence histories of IN and IN-NE



(b) von Mises stress

Fig. 3: Numerical result of Example 3.

5 Conclusions

The main contribution of this paper was to investigate the performance of a nonlinear elimination preconditioner with applications in computational hyperelasticity. A robust strategy of nonlinearity checking was adapted to capture the subregions with stronger nonlinearity, which coincide with the propagation of the elastic wave. Moreover, we found that the extension for the eliminating index set by adding the neighbor d.o.f. is an effective trick to ease the thrashing phenomenon of nonlinear elimination. As future work, we will use more feasible linear solvers for the Jacobian system and consider other arterial wall problems with patient-specific geometry.

References

1. S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, K. Rupp, B. F. Smith, S. Zampini, H. Zhang, and H. Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 3.7, Argonne National Laboratory, 2016.
2. J. M. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Arch. for Ration. Mech. Analysis*, 63:337–403, 1976.
3. D. Balzani, S. Deparis, S. Fausten, D. Forti, A. Heinlein, A. Klawonn, A. Quarteroni, O. Rheinbach, and J. Schröder. Numerical modeling of fluid–structure interaction in arteries with anisotropic polyconvex hyperelastic and anisotropic viscoelastic material models at finite strains. *Int. J. Numer. methods Biomed. Eng.*, 2015.
4. D. Balzani, P. Neff, J. Schröder, and G. A. Holzapfel. A polyconvex framework for soft biological tissues. Adjustment to experimental data. *Int. J. Solids Struct.*, 43:6052–6070, 2006.
5. D. Brands, A. Klawonn, O. Rheinbach, and J. Schröder. Modelling and convergence in arterial wall simulations using a parallel FETI solution strategy. *Comput. Methods Biomech. Biomed. Eng.*, 11:569–583, 2008.
6. S. Brinkhues, A. Klawonn, O. Rheinbach, and J. Schröder. Augmented Lagrange methods for quasi-incompressible materials–Applications to soft biological tissue. *Int. J. Numer. methods Biomed. Eng.*, 29:332–350, 2013.
7. X.-C. Cai and X. Li. Inexact Newton methods with restricted additive Schwarz based nonlinear elimination for problems with high local nonlinearity. *SIAM J. Sci. Comput.*, 33:746–762, 2011.
8. P. G. Ciarlet. *Mathematical Elasticity, Vol. I : Three-Dimensional Elasticity.* Series "Studies in Mathematics and its Applications". North-Holland, Amsterdam, 1988.
9. R. S. Dembo, S. C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Numer. Analysis*, 19:400–408, 1982.
10. S. C. Eisenstat and H. F. Walker. Globally convergent inexact Newton methods. *SIAM J. Optim.*, 4:393–422, 1994.
11. J. Huang, C. Yang, and X.-C. Cai. A nonlinearly preconditioned inexact Newton algorithm for steady state lattice Boltzmann equations. *SIAM J. Sci. Comput.*, 38:A1701–A1724, 2016.
12. P. J. Lanzkron, D. J. Rose, and J. T. Wilkes. An analysis of approximate nonlinear elimination. *SIAM J. Sci. Comput.*, 17:538–559, 1996.
13. A. Logg, K.-A. Mardal, and G. Wells. *Automated solution of differential equations by the finite element method: The FEniCS book*, Volume 84. Springer Science & Business Media, 2012.

Parallel-in-Time for Parabolic Optimal Control Problems Using PFASST

Sebastian Götschel and Michael L. Minion

Abstract In gradient-based methods for parabolic optimal control problems, it is necessary to solve both the state equation and a backward-in-time adjoint equation in each iteration of the optimization method. In order to facilitate fully parallel gradient-type and nonlinear conjugate gradient methods for the solution of such optimal control problems, we discuss the application of the parallel-in-time method PFASST to adjoint gradient computation. In addition to enabling time parallelism, PFASST provides high flexibility for handling nonlinear equations, as well as potential extra computational savings from reusing previous solutions in the optimization loop. The approach is demonstrated here for a model reaction-diffusion optimal control problem.

1 Introduction

Gradient-based methods for parabolic optimal control problems are computationally expensive due to the need to solve both a forward state equation and a backward-in-time adjoint equation to compute gradient information in each iteration of the optimization procedure. One potential way to reduce the overall computational time is to employ parallel-in-time (PinT) methods for solving state and adjoint equations. Attempts to construct PinT methods for the solution of differential equations date back more than 50 years and have gained increasing interest in the last 15 years [8]. More recently, the application of space-time parallel methods to the solution of optimization problems governed by PDEs has become an active research area, with approaches including multiple shooting (e.g. [11] and the references therein), Schwarz

Sebastian Götschel
Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany, e-mail: goetschel@zib.de

Michael Minion
Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, CA, 94720, USA, e-mail: mlminion@lbl.gov

methods [1, 9], the application of parareal preconditioners [14, 18], and space-time parallel multigrid methods [10].

Here we apply the PFASST method [7] ("Parallel Full Approximation Scheme in Space and Time") to both the state and adjoint equation to provide a fully time-parallel gradient- or nonlinear conjugate gradient method. This approach is somewhat related to the time-parallel gradient type method presented in [5]. There the time interval of interest is subdivided into time steps, which are solved in parallel using quantities from the previous optimization iteration as input. This leads to jumps in the solutions of state and adjoint equation such that these equations are not satisfied during optimization. While convergence is demonstrated in [5] if sufficiently small step sizes for updating the control are used, it is unclear how to automatically select such a step size. In our approach, the usual line search criteria, e.g. the (strong) Wolfe conditions, can be used to guarantee convergence.

2 Background

2.1 SDC, MLSDC, and PFASST

A distinguishing factor of the PFASST algorithm compared to other PinT methods is that, in each iteration, the solution on a given time step is improved using a deferred correction approach rather than being computed in full using a given ODE method. The correction sweeps are based on spectral deferred corrections (SDC) [6] and are applied on a hierarchy of space-time representations of the problem as in multi-level SDC (MLSDC) methods [17]. PFASST exposes parallelism in the time direction because MLSDC iterations are pipelined so that SDC sweeps are done concurrently on all but the coarsest level.

One advantage of SDC methods is the flexibility in choosing the type of substepping for the correction sweep. In the numerical example, we will use both a semi-implicit or IMEX approach [15] (wherein one component of the solution is treated explicitly and one implicitly) and a multi-implicit (MISDC) approach [2] (wherein one component of the solution is treated explicitly and two components implicitly but uncoupled). The motivation for using IMEX and MISDC variants are to replace the solution of coupled nonlinear systems in the time stepping by simpler linear equations (see Sect. 4.1 for further discussion).

Finally, PFASST is an iterative method, and the typical way in which the solution is initialized on each parallel time slice is by serial time stepping on the coarsest level. In optimal control problems, an alternative is to use the solution from the previous optimization iteration as the initial guess for the next state and adjoint equation solve. As the optimization procedure converges, the initial solutions improve in quality, and hence the number of PFASST iterations needed for convergence should decrease. We demonstrate this savings in Sect. 4.

2.2 Optimal Control of Parabolic PDEs

We consider optimal controls problems

$$\min_{y\in Y, u\in U} J(y,u) \text{ subject to } c(y,u)=0, \tag{1}$$

with $c: Y \times U \to Z^\star$ a semi-linear parabolic PDE on Banach spaces Y, Z and Hilbert space U over a spatial domain $\Omega \subset \mathbb{R}^d$. We assume that there exists a unique solution $y = y(u) \in Y$ of the state equation $c(y,u) = 0$ for each control $u \in U$. To avoid a full, typically 4D, discretization of this problem, methods working on the reduced functional

$$\min_{u\in U} j(u) := J(y(u),u) \tag{2}$$

are often employed. Under standard assumptions, the reduced gradient is given by

$$j'(u) = J_u(y(u),u) + c_u(y(u),u)^\star p(u),$$

where p solves the adjoint equation

$$J_y(y(u),u) + c_y(y(u),u)^\star p(u) = 0, \tag{3}$$

which is backward in time, see, e.g., [13] for details. Due to the occurrence of $-J_y(y(u),u)$ as a source term, and—in the nonlinear case—the dependence of $c_y(y(u),u)$ on the state solution $y(u)$, the adjoint gradient computation consists of three steps:

1. solve $c(y,u) = 0$ for $y \in Y$ and store the solution trajectory,
2. solve $c_y(y,u)^\star p = -J_y(y,u)$ for $p \in Z$,
3. set $j'(u) = J_u(y,u) + c_u(y,u)^\star p$.

In order to facilitate fully parallel algorithms to solve the optimal control problem (1), state and adjoint equations need to be solved using PinT methods.

3 PFASST for Optimal Control

Minimizing the objective function (2) is done via gradient- or nonlinear conjugate gradient (ncg) methods

$$\begin{aligned} u_{k+1} &= u_k + \alpha_k d_k \\ d_{k+1} &= -j'(u_{k+1}) + \beta_k d_k, \end{aligned}$$

where $d_0 = -j'(u_0)$, α_k denotes the step size, required to satisfy the (strong) Wolfe conditions [16], and the choice of β_k defines the type of method ($\beta_k = 0$ for the gradient method; various possibilities for β_k leading to different ncg methods, see [4] for a brief overview and the method used in the experiments). For the numerical

solution we apply a method of lines approach, discretizing first in space, then in time.

Parallelization in time for these methods requires three ingredients: time-parallel computation of inner products, step size selection, and the solution of state and adjoint equations. The first two ingredients are straightforward: on each time interval, local scalar products are computed, and then communicated to all other processors, summing them up. These scalar products are used to compute β_k, as well as to check sufficient decrease and curvature conditions during step size selection. For the time-parallel solution of state and adjoint equations we propose two different strategies using PFASST. In the first approach, the state and adjoint problems are solved separately using PFASST for both. The state solution at each time step and quadrature node is stored for subsequent use in the solution of the adjoint equation. Alternatively, PFASST could be used to solve the state and adjoint equation at the same time. Each SDC sweep of the state equation would be followed by a backward sweep of the adjoint equation on the same nodes, leading to more complicated communication patterns. In the numerical example, we focus on the first approach. Details and results for the second approach will be reported elsewhere.

4 Numerical Example

Here we consider the following optimal control problem ([3, 12]) governed by a semi-linear reaction-diffusion equation on $\Omega = (0,20)$:

$$\min_{y,u} \frac{1}{2}\int_0^T \int_\Omega (y-y_d)^2 \, dx\, dt + \frac{\lambda}{2}\int_0^T \int_\Omega u^2 \, dx\, dt,$$

where $T = 5$, and $y(x,t)$ is subject to

$$\begin{aligned} y_t - y_{xx} + y(\frac{1}{3}y^2 - 1) &= u(x,t) \quad \text{in } \Omega \times (0,T) \\ y(x,0) &= y_0(x) \quad \text{in } \Omega, \end{aligned} \tag{4}$$

with homogeneous Neumann boundary conditions. The initial condition and desired state are

$$y_0(x) = \begin{cases} 1.2\sqrt{3}, & x \in [9,11] \\ 0, & \text{elsewhere} \end{cases} \quad \text{and} \quad y_d(x,t) = \begin{cases} y_{\text{nat}}(x,t), & t \in [0,2.5] \\ y_{\text{nat}}(x,2.5), & t \in (2.5,T], \end{cases}$$

where y_{nat} denotes the solution to the PDE (4) for $u \equiv 0$. For $\lambda = 0$, an exact optimal control is known:

$$u_{\text{exact}} = \begin{cases} 0, & t \le 2.5 \\ \frac{1}{3}y_{\text{nat}}^3(x,2.5) - y_{\text{nat}}(x,2.5) - \frac{\partial^2}{\partial x^2} y_{\text{nat}}(x,2.5), & t > 2.5. \end{cases}$$

4.1 IMEX and MISDC Formulations

As mentioned in Sect. 2.1, there is great flexibility in how the substepping procedure in SDC is constructed since it need only be a first-order approximation. For our numerical example, two strategies are investigated, IMEX SDC and multi-implicit SDC. In both cases, the diffusion term in (4) is treated implicitly to avoid the severe time-step restriction inherent in explicit temporal methods. In the IMEX strategy, only the diffusion term is treated implicitly, while in the MISDC method, both diffusion and reaction are treated implicitly, but the implicit solutions are done independently as in operator splitting methods. In addition, we employ a lagging of nonlinear terms in MISDC iteration to turn the nonlinear solve into a linear problem.

Methods that employ operator splitting are desirable when the reduced cost of split implicit solvers compared to coupled solvers is significant. The overall accuracy of PFASST (assuming convergence of SDC iterations) does not depend on the form of the substepping, rather on the choice of number and type of integration nodes. Hence, the main concern in terms of efficiency is the computational cost of each SDC iteration and the number of iterations required for convergence.

The IMEX and MISDC approach are explained by examining a single substep of an SDC sweep. Letting k denote the SDC iteration, m the substep index, and D^2 the discretization of the second derivative term, then the correction equation for a single fully implicit, backward-Euler type discretization of the substep for (4) will take the form

$$y_{m+1}^{[k+1]} = y_m^{[k+1]} + \Delta t_m (D^2 y_{m+1}^{[k+1]} - y_{m+1}^{[k+1]} (\frac{1}{3}(y_{m+1}^{[k+1]})^2 - 1)) + S_j^{[k]}, \tag{5}$$

where the term $S_j^{[k]}$ contains terms that either depend on the previous iteration $[k]$ or values at iteration $[k+1]$ already computed at substep $j < m+1$, including the control terms arising from the discretization of $u(x,t)$. Note that the implicit equation couples nonlinear reaction and diffusion terms and hence would require a global nonlinear solver in each substep. For problems in which the reaction terms are nonstiff and can be treated explicitly, the reaction terms at node $m+1$ do not appear in the implicit equation, giving the form

$$y_{m+1}^{[k+1]} = y_m^{[k+1]} + \Delta t_m (D^2 y_{m+1}^{[k+1]} - y_m^{[k+1]} (\frac{1}{3}(y_m^{[k+1]})^2 - 1)) + S_j^{[k]}. \tag{6}$$

Each substep now requires only the solution of a linear implicit equation, and hence is computationally cheaper than a fully implicit approach, assuming that the explicit treatment of the reaction term does not impose an additional time step restriction.

When the reaction term is stiff, and hence it is advantageous to treat it implicitly, a standard MISDC approach applies an operator splitting between diffusion and reaction in the correction equation. For example,

$$y^* = y_m^{[k+1]} + \Delta t_m D^2 y^* + S_j^{*,[k]}, \tag{7}$$

$$y_{m+1}^{[k+1]} = y_m^{[k+1]} + \Delta t_m (D^2 y^* - y_{m+1}^{[k+1]} (\frac{1}{3}(y_{m+1}^{[k+1]})^2 - 1)) + S_j^{[k]}. \tag{8}$$

For the numerical methods here, the MISDC approach is further modified, so that the nonlinear solve for reaction in (8) is made linear by lagging terms in $[k]$:

$$y_{m+1}^{[k+1]} = y_m^{[k+1]} + \Delta t_m (D^2 y^* - y_{m+1}^{[k+1]} (\frac{1}{3}(y^*)^2 - 1)) + S_j^{[k]}. \tag{9}$$

This form creates an implicit solve with roughly the same cost as treating reaction explicitly but is more stable. In all the numerical examples presented here, a DIRK type approach [19] is used so that the generic form of $S_j^{[k]}$ contains both the usual SDC terms from iteration $[k]$ and a linear combination of previously computed right-hand-side terms from SDC iteration $[k+1]$.

4.2 Results

In this section we show the results using IMEX and MISDC approaches to solve the state and adjoint equation. In both cases, a method of lines is employed by using a spectral discretization in space with spatial derivatives computed with the fast Fourier transform. The PFASST iterations are stopped when the relative or absolute residual falls below 10^{-11}. For solving the optimization problem we set $\lambda = 10^{-6}$ and use the ncg method from [4], with initial control $u_0 = 0.5 u_{\text{exact}}$. As described in [3], the ncg method converges quite slowly for this particular problem; it was stopped after at most 200 iterations.

IMEX. Since the reaction terms in our example are not highly stiff, an IMEX approach can be used for the state and adjoint equations. PFASST is employed using three levels (32/64/128 spatial points and 3/5/9 LobattoIIIA nodes in time) with 20 parallel time intervals. Note the temporal method is formally 16th order. Running on 20 processors in parallel, the final objective function value after 200 ncg iterations is $2.4 \cdot 10^{-3}$, and the computed control has a relative L^2-error of 0.15 compared to u_{exact}. In contrast, the sequential version stops with a slightly worse objective function value of $3.2 \cdot 10^{-3}$, and a relative L^2-error of 0.15 in the computed control. A plot of the computed control, the error in the computed control, and the corresponding computed optimal state can be found in Fig. 1. By parallel execution, the overall runtime was reduced by a factor 3.8, yielding a parallel efficiency of 19%.

MISDC. For testing MISDC we used 20 parallel time intervals with two PFASST levels consisting of 64/128 spatial points and 5/9 LobattoIIIA nodes. After 200 ncg iterations, the sequential version reached an objective function value of $3.8 \cdot 10^{-3}$ and a relative L^2-error of 0.15 in the computed control. Running in parallel reduced the computation time by a rather small factor 2, but lead to improved results (objective: $1.8 \cdot 10^{-3}$, control: relative L^2-error 0.14). Initializing the state solution at the collocation nodes in optimization iteration k with their values from iteration $k-1$ ("warm start") reduced the required sweeps by 48% while reaching an objective function value of $1.5 \cdot 10^{-3}$ and relative L^2-error in the control of 0.13. The reduc-

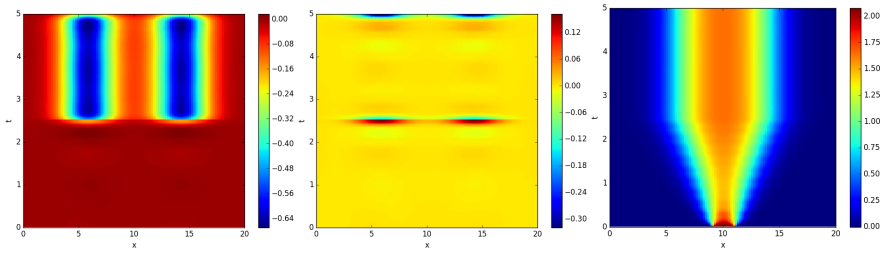


Fig. 1 Computed optimal control with $\lambda = 10^{-6}$ (left), difference to exact control for $\lambda = 0$ (middle) and optimal state (right) using IMEX.

tion in sweeps translates to a significant reduction of overall computation time by 45%. This is in contrast to the IMEX experiment, where for a reduction in sweeps by 39% the gain in overall speed was a mere 7%. Lagged linearization as in (9) increases the total number of PFASST sweeps for a state equation solve from 570 (on average 28.5 per time step) to 642 (avg. 32.1/time step). Using smaller time steps (40 parallel intervals), the average number of iterations is 28.3 in both cases.

For this example, it is unreasonable to attempt to compare the IMEX and MISDC approaches in terms of overall efficiency since MISDC is designed for problems where both diffusion and reaction components are stiff. The pertinent point here is that employing the MISDC procedure with a lagged linearization of reaction terms does not appear to increase the number of PFASST iterations needed for convergence substantially, thus offering the possibility of greatly reducing the cost of implicit substepping compared to fully implicit methods.

5 Discussion

An approach using PFASST for the time-parallel solution of PDE-constrained optimization problems has been presented, and non-trivial parallel speedup and efficiency have been obtained. It is important to note that the parallel efficiency of PFASST is improved when solutions on coarsest levels are much cheaper than in finer levels, and spatial coarsening has a larger effect in multiple dimensions compared to the one-dimensional example used here. In addition, applying PFASST simultaneously to state and adjoint equations with proper handling of communication offers further improved parallel speedup. The flexibility of SDC/PFASST has been used to reduce the cost of implicit solutions in the substepping and also to re-use information from previous optimization iterations. Future research will, for example, deal with adaptive control of the accuracy for inexact gradient computations, and different strategies for storing or recomputing state solutions for the adjoint solve.

Acknowledgements S.G. gratefully acknowledges partial funding by the Deutsche Forschungsgemeinschaft (DFG), Project WE 2937/6-1. The work of M.M. was supported by the Applied

Mathematics Program of the DOE Office of Advanced Scientific Computing Research under the U.S. Department of Energy under contract DE-AC02-05CH11231.

References

1. Barker, A. T., Stoll, M.: Domain decomposition in time for PDE-constrained optimization. Comput. Phys. Commun. **197**, 136–143 (2015)
2. Bourlioux, A., Layton, A. T., Minion, M. L.: High-order multi-implicit spectral deferred correction methods for problems of reactive flow. J. Comput. Phys. **189**(2), 651–675 (2003)
3. Buchholz, R., Engel, H., Kammann, E., Tröltzsch, F.: On the optimal control of the Schlögl-model. Comput. Optim. Appl. **56**(1), 153–185 (2013)
4. Dai, Y. H., Yuan, Y.: A nonlinear conjugate gradient method with a strong global convergence property. SIAM J. Optim. **10**(1), 177–182 (1999)
5. Deng, X., Heinkenschloss, M.: A parallel-in-time gradient-type method for discrete time optimal control problems. Preprint, Department of Computational and Applied Mathematics, Rice University (2016) Available from http://www.caam.rice.edu/∼heinken.
6. Dutt, A., Greengard, L., Rokhlin, V.: Spectral deferred correction methods for ordinary differential equations. BIT Numer. Math. **40**(2), 241–266 (2000)
7. Emmett, M., Minion, M. L.: Toward an efficient parallel in time method for partial differential equations. Comm. App. Math. Comp. Sci. **7**, 105–132 (2012)
8. Gander, M. J.: 50 years of time parallel time integration. In: Carraro, T. et al. (eds.), Multiple Shooting and Time Domain Decomposition Methods: MuS-TDD, Heidelberg, May 6-8, 2013, pp. 69–113. Springer, Cham (2015)
9. Gander, M. J., Kwok, F.: Schwarz methods for the time-parallel solution of parabolic control problems. In: Dickopf, T. et al. (eds.), Domain Decomposition Methods in Science and Engineering XXII, pp. 207–216. Springer, Cham (2016)
10. Günther, S., Gauger, N. R., Schroder, J. B.: A Non-Intrusive Parallel-in-Time Adjoint Solver with the XBraid Library. ArXiv e-prints (2017)
11. Heinkenschloss, M.: A time-domain decomposition iterative method for the solution of distributed linear quadratic optimal control problems. J. Comput. Appl. Math. **173**(1), 169–198 (2005)
12. Herzog, R. Rösch, A., Ulbrich, S., Wollner, W.: OPTPDE: A collection of problems in PDE-constrained optimization. In: Leugering, G. et al. (eds.), Trends in PDE Constrained Optimization, pp. 539–543. Springer, Cham (2014)
13. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE constraints. Springer (2009)
14. Mathew, T., Sarkis, M., Schaerer, C. E.: Analysis of Block Parareal Preconditioners for Parabolic Optimal Control Problems. SIAM J. Sci. Comput. **32**(3), 1180–1200 (2010)
15. Minion, M. L.: Semi-implicit spectral deferred correction methods for ordinary differential equations. Commun. Math. Sci. **1**(3), 471–500 (2003)
16. Nocedal, J., Wright, S. J.: Numerical Optimization. Springer, New York (2006)
17. Speck, R., Ruprecht, D., Emmett, M., Minion, M. L., Bolten, M., Krause, R.: A multi-level spectral deferred correction method. BIT Numer. Math. **55**, 843–867 (2015)
18. Ulbrich, S.: Preconditioners based on "parareal" time-domain decomposition for time-dependent PDE-constrained optimization. In: Carraro, T. et al. (eds.), Multiple Shooting and Time Domain Decomposition Methods: MuS-TDD, Heidelberg, May 6-8, 2013, pp. 203–232. Springer, Cham (2015)
19. Weiser, M.: Faster SDC convergence on non-equidistant grids by DIRK sweeps. BIT Numer. Math. **55**, 1219–1241 (2013)

An Adaptive GDSW Coarse Space for Two-Level Overlapping Schwarz Methods in Two Dimensions

Alexander Heinlein[1], Axel Klawonn[1], Jascha Knepper[1], and Oliver Rheinbach[2]

1 Introduction

We consider the second order elliptic problem in two dimensions

$$\begin{aligned} -\nabla \cdot (A(x)\nabla u(x)) &= f(x) \quad \text{in } \Omega \subset \mathbb{R}^2, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{1}$$

where the scalar coefficient function $A(x) > 0$ is highly heterogeneous, possibly with high jumps. While convergence of standard two-level Schwarz preconditioners depends on the contrast of the coefficient function, we propose a coarse space for two-level overlapping Schwarz methods which yields a condition number that is independent of the coefficient function. Our approach can be viewed as an extension to the GDSW (Generalized Dryja, Smith, Widlund) method [1, 2] since it always contains the standard GDSW coarse space. Originally, the method was inspired by the ACMS (Approximate Component Mode Synthesis) special finite element method [9, 6], which uses enrichment of the discretization space by local eigenfunctions. The ACMS space was first considered as a coarse space for overlapping Schwarz methods in [7].

Our new coarse space consists of simple nodal finite element functions and of energy minimizing extensions of solutions of generalized eigenvalue problems on the edges. Here, we restrict ourselves to the two-dimensional case. For the description of the three-dimensional case and the proof of the condition number bound, we refer to [8]. A related method is the SHEM (Spectral Harmonically Enriched Multiscale) coarse space, introduced in [5], however, our eigenvalue problems do not use mass matrices; see (5). Other related coarse spaces for overlapping Schwarz methods are,

[1]Mathematisches Institut, Universität zu Köln, Weyertal 86-90, 50931 Köln, Germany. e-mail: {alexander.heinlein,axel.klawonn,jascha.knepper}@uni-koeln.de
[2]Institut für Numerische Mathematik und Optimierung, Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg, Akademiestr. 6, 09596 Freiberg, Germany. e-mail: oliver.rheinbach@math.tu-freiberg.de

e.g., [3, 4]. In our new coarse space and in the one based on the ACMS discretization method, the construction of the generalized eigenvalue problems is computationally slightly more expensive than in the SHEM coarse space [5]. However, the dimension of the coarse space can be reduced significantly in certain cases.

The variational problem corresponding to (1) reads: find $u \in H_0^1(\Omega)$, such that

$$a_\Omega(u,v) = L(v) \qquad \forall v \in H_0^1(\Omega) \tag{2}$$

with the bilinear form and the linear functional

$$a_\Omega(u,v) := \int_\Omega (\nabla u(x))^T A(x) \nabla v(x)\, dx \quad \text{and} \quad L(v) := \int_\Omega f(x) v(x)\, dx,$$

respectively, where $f \in L^2(\Omega)$. We define the semi-norm corresponding to the bilinear form $a_\Omega(\cdot,\cdot)$ as $|u|^2_{a,\Omega} := a_\Omega(u,u)$. Let $Ku = f$ be the discretization of problem (2) by piecewise linear or bilinear finite elements on a family of triangulations $(\tau_h)_h$. We solve the discretized system using the conjugate gradient method preconditioned by a suitable two-level overlapping Schwarz preconditioner.

2 The GDSW Preconditioner

The GDSW preconditioner [1, 2] is a two-level additive overlapping Schwarz preconditioner with exact solvers; cf. [10]. It can therefore be written in the form

$$M_{\mathrm{GDSW}}^{-1} = \Phi K_0^{-1} \Phi^T + \sum_{i=1}^{N} R_i^T \tilde{K}_i^{-1} R_i, \tag{3}$$

where $K_0 = \Phi^T K \Phi$ and $\tilde{K}_i = R_i^T K R_i$. The matrices R_i are the restriction operators to the overlapping subdomains $\tilde{\Omega}_i$, $i = 1, \ldots, N$, and the columns of Φ are the coarse basis functions. The coarse basis functions are discrete harmonic extensions of interface functions into the interior degrees of freedom of the corresponding nonoverlapping subdomains. On the interface, the values are defined as the restrictions of the nullspace of the operator to the edges and vertices of the nonoverlapping domain decomposition.

The condition number estimate for the GDSW Schwarz operator in case of a constant coefficient function A is

$$\kappa\left(M_{\mathrm{GDSW}}^{-1} K\right) \leq C \left(1 + \frac{H}{\delta}\right) \left(1 + \log\left(\frac{H}{h}\right)\right)^2; \tag{4}$$

cf. [1, 2]. If A is not constant, the constant C also depends the contrast of A.

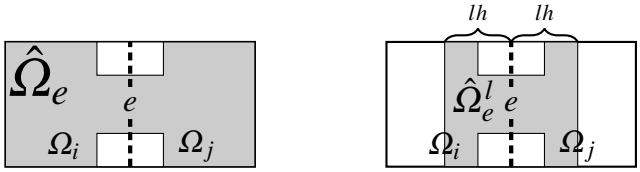


Fig. 1 (**left**) Graphical representation of $\Omega_e = \Omega_i \cup \Omega_j$ and $\hat{\Omega}_e$. The set $\hat{\Omega}_e$ is obtained by removing from Ω_e all elements which are adjacent to the coarse nodes. From this, we also obtain the interior edge $\hat{e} := e \cap \hat{\Omega}_e$. (**right**) Graphical representation of the slab $\hat{\Omega}_e^l$ corresponding to the edge e.

3 Adaptive GDSW in 2D

The adaptive GDSW coarse space is an extension to the standard GDSW coarse space since it automatically includes the standard GDSW coarse space. However, if necessary due to coefficient jumps, additional coarse constraints are selected. These additional coarse constraints are constructed from solving local generalized eigenvalue problems. Let the interface Γ be partitioned into edges $\mathscr{E}$ and vertices $\mathscr{V}$, i.e., $\Gamma = (\bigcup_{e\in\mathscr{E}} e) \cup (\bigcup_{v\in\mathscr{V}} v)$. For each edge e, we define the sets Ω_e and $\hat{\Omega}_e$ as depicted in Fig. 1 (left) and the following extension operator:

$$w_e : V_0^h(e) \to V_0^h(\Omega_e), \quad v \mapsto w_e(v) := \begin{cases} v & \text{in all interior nodes of } e, \\ 0 & \text{on all other nodes in } \Omega_e, \end{cases}$$

where $V_0^h(e) := \{ v|_e : v \in V, v = 0 \text{ on } \partial e\}$. Then, we consider on each edge $e \in \mathscr{E}$ the generalized eigenvalue problem: find $\tau_{*,e} \in V_0^h(e)$ such that

$$a_{\hat{\Omega}_e}\left(\mathscr{H}_{\hat{e}\to\hat{\Omega}_e}(\tau_{*,e}), \mathscr{H}_{\hat{e}\to\hat{\Omega}_e}(\theta)\right) = \lambda_{*,e}\, a_{\Omega_e}(w_e(\tau_{*,e}), w_e(\theta)) \quad \forall \theta \in V_0^h(e). \quad (5)$$

Here, $\mathscr{H}_{\hat{e}\to\hat{\Omega}_e}$ denotes the discrete harmonic extension from the interior edge $\hat{e}$ into $\hat{\Omega}_e$ with respect to the bilinear form $a_{\hat{\Omega}_e}(\cdot,\cdot)$. Let the corresponding eigenvalues be sorted in non-descending order, i.e., $\lambda_{1,e} \le \lambda_{2,e} \le \ldots \le \lambda_{m,e}$, and the eigenmodes accordingly, where $m = \dim\left(V_0^h(e)\right)$. We select all eigenmodes $\tau_{*,e}$ where the eigenvalues are below a certain tolerance, i.e., $\lambda_{*,e} \le tol_{\mathscr{E}}$. Then we extend the selected eigenfunctions by zero to $\Gamma \setminus e$, denoted by $\tilde{\tau}_{*,e}$, and subsequently compute the discrete harmonic extension into the interior of the subdomains, i.e., $v_{*,e} := \mathscr{H}_{\Gamma\to\Omega}(\tilde{\tau}_{*,e})$.

Note that for every edge e, the left hand side of the eigenvalue problem (5) is singular. Therefore, since $tol_{\mathscr{E}} \ge 0$, eigenmodes which span the nullspace are always selected and added to the coarse space. Therefore, the standard GDSW coarse space is always a subspace of our automatic coarse space.

In addition to the edge basis functions, we use the nodal coarse basis functions from the GDSW coarse space, which span the space $V_{\mathscr{V}}$. We denote the resulting coarse space as the AGDSW (Adaptive GDSW) coarse space:

$$V_{\mathrm{AGDSW}}^{tol_{\mathscr{E}}} = V_{\mathscr{V}} \oplus \left(\bigoplus_{e \in \mathscr{E}} \mathrm{span}\left\{ v_{k,e} : \lambda_{k,e} \leq tol_{\mathscr{E}} \right\} \right)$$

Remark 1: For $tol_{\mathscr{E}} \geq 0$, we obtain $V_{\mathrm{GDSW}} = V_{\mathrm{AGDSW}}^{0} \subseteq V_{\mathrm{AGDSW}}^{tol_{\mathscr{E}}}$.
Remark 2: The right hand side of the eigenvalue problem (5) can be extracted from the fully assembled global stiffness matrix K.
Remark 3: The condition number of the AGDSW Schwarz operator is bounded by

$$\kappa\left(M_{\mathrm{AGDSW}}^{-1} K\right) \leq C \left(1 + \frac{1}{tol_{\mathscr{E}}}\right); \tag{6}$$

see [8]. The constant C is independent of H, h, and the contrast of the coefficient function A. In [8], the three-dimensional case is also covered including the theory.

4 Variants of Adaptive GDSW

Here, we will briefly discuss some possible variants of the AGDSW method.

Mass matrix As in other adaptive coarse spaces where a spectral estimate is used to replace a Poincaré type inequality, cf., e.g., [3, 4, 5, 7], we can use a (scaled) mass matrix on the right hand side of the eigenvalue problems. The scaled mass matrix corresponding to an edge $e \subset (\bar{\Omega}_i \cap \bar{\Omega}_j)$ arises from the discretization of the scaled L^2-inner product

$$b_e(u,v) := \frac{1}{h^2} (A \cdot w_e(u), w_e(v))_{L^2(\Omega_e)}. \tag{7}$$

Therefore, we obtain for each edge the modified generalized eigenvalue problem: find $\tau_{*,e} \in V_0^h(e)$ such that

$$a_{\hat{\Omega}_e}\left(\mathscr{H}_{\hat{e}\to\hat{\Omega}_e}(\tau_{*,e}), \mathscr{H}_{\hat{e}\to\hat{\Omega}_e}(\theta)\right) = \lambda_{*,e} b_e(\tau_{*,e}, \theta) \quad \forall \theta \in V_0^h(e). \tag{8}$$

The condition number bound (6) can also be proven for this variant; see [8].

Slabs In order to reduce the computational cost of constructing the generalized eigenvalue problems, the set $\hat{\Omega}_e$ can be replaced by a slab of width l elements around the edge e in (5); cf. Fig. 1 (right) for the graphical representation of a slab corresponding to the edge e. We denote the slab by $\hat{\Omega}_e^l$. This idea, to use slabs in the eigenvalue problems, has already been introduced in [7] for related multiscale coarse spaces based on the ACMS space and is also common in FETI-DP and BDDC domain decomposition methods with adaptive coarse spaces.

The modified generalized eigenvalue problem reads: find $\tau_{*,e} \in V_0^h(e)$ such that

$$a_{\hat{\Omega}_e^l}\left(\mathscr{H}_{\hat{e}\to\hat{\Omega}_e^l}(\tau_{*,e}), \mathscr{H}_{\hat{e}\to\hat{\Omega}_e^l}(\theta)\right) = \lambda_{*,e} a_{\Omega_e}(w_e(\tau_{*,e}), w_e(\theta)) \quad \forall \theta \in V_0^h(e). \tag{9}$$

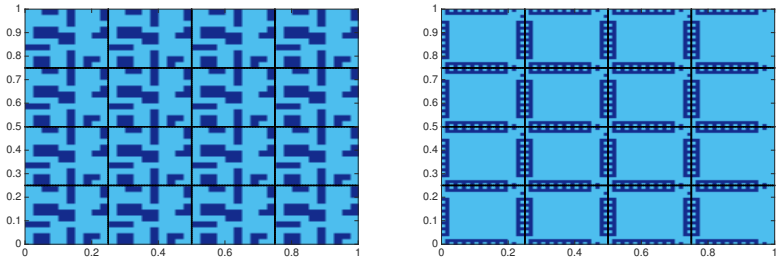


Fig. 2 Discontinuous coefficient functions A with different types of channels and inclusions intersecting the interface. Maximum coefficient (dark blue color): $A_{\max} = 10^6$ (left), $A_{\max} = 10^8$ (right); $1/H = 4$; $H/h = 30$ (left); $H/h = 40$ (right); $\delta = 2h$.

	Coeff. function A from Fig. 2 (left)				Coeff. function A from Fig. 2 (right)			
V_0	$tol_{\mathcal{E}}$	it.	κ	$\dim V_0$	$tol_{\mathcal{E}}$	it.	κ	$\dim V_0$
V_{GDSW}		264	$1.04 \cdot 10^6$	33		45	26.18	33
V_{AGDSW}	10^{-1}	29	7.15	93	10^{-1}	34	10.06	81
	10^{-2}	29	7.15	93	10^{-2}	44	26.20	57
$V_{\text{AGDSW-M}}$	10^{-1}	29	7.15	93	10^{-1}	44	26.20	57
	10^{-2}	29	7.15	93	10^{-2}	44	26.20	57
V_{SHEM}	10^{-3}	20	4.33	69	10^{-3}	23	5.03	213
	10^{-6}	20	4.33	69	10^{-6}	23	5.03	213

Table 1 Results for the coefficient functions in Fig. 2: tolerance for the selection of the eigenfunctions, iterations counts, condition numbers, and resulting coarse space dimension for different coarse space variants; $1/H = 4$, $H/h = 30$ (left), $H/h = 40$ (right), and $\delta = 2h$; maximum coefficient $A_{\max} = 10^6$ (left) and $A_{\max} = 10^8$ (right).

The slab variant is computationally cheaper and the bound can be proven analogously to the standard version with no modifications. However, the coarse space dimension can increase due to the use of this variant (if $\hat{\Omega}_e^l \subset \hat{\Omega}_e$).

5 Numerical Results

We present numerical results for model problem (1) for $f \equiv 1$ and various coefficient functions, comparing the different AGDSW approaches with the standard GDSW as well as the SHEM coarse space, recently introduced by Gander, Loneland, and Rahman in [5]. Finally, we show results using slabs of varying widths.

In all figures, the light and dark blue colors correspond to the minimum coefficient ($A_{\min} = 1.0$) and maximum coefficient ($A_{\max} = 10^6$ or $A_{\max} = 10^8$), respectively. We use piecewise bilinear finite elements, and solve the discrete linear system using the conjugate gradient method with a relative stopping criterion, $||r^{(k)}||_2/||r^{(0)}||_2 \leq 10^{-8}$, where $r^{(0)}$ and $r^{(k)}$ are the initial and the k-th unpreconditioned residual, respectively. By V_{GDSW}, we denote the standard GDSW space and by V_{AGDSW}^{tol} the new adaptive GDSW coarse space. The variant which uses a scaled

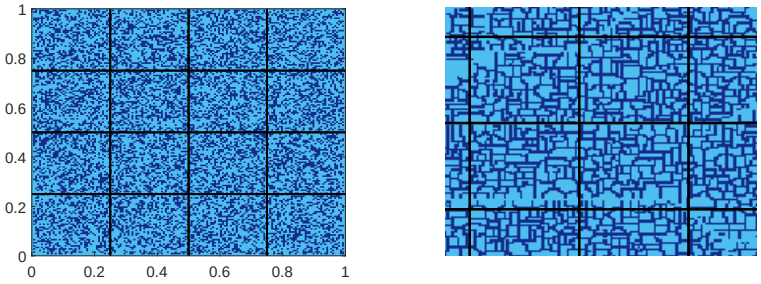


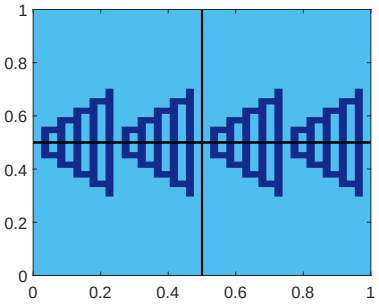
Fig. 3 (**Left:**) Sample random coefficient function with a density of approximately 40% high coefficients $A_{\max} = 10^6$ (dark blue color). $1/H = 4$; $H/h = 40$; $\delta = 1h$. (**Right:**) Detailed view of a coefficient function with $A_{\max} = 10^8$ (dark blue color) and $1/H = 20$, $H/h = 40$, $\delta = 1h$.

	Random coeff. function A from Fig. 3 (left)				Coeff. fn. A from Fig. 3 (right)			
V_0	$tol_{\mathcal{E}}$	it.	κ	dim V_0	$tol_{\mathcal{E}}$	it.	κ	dim V_0
V_{GDSW}		> 500 (-)	$2.8 \cdot 10^5$ ($6.9 \cdot 10^4$)	33 (0.0)		3 042	$4.9 \cdot 10^7$	1 121
V_{AGDSW}	10^{-1}	34.3 (1.7)	11.8 (2.0)	185.1 (7.0)	10^{-1}	47	16.2	3 087
$V_{\text{AGDSW-M}}$	10^{-1}	51.6 (3.7)	22.6 (7.6)	148.4 (8.5)	10^{-1}	75	40.1	1 862
V_{AGDSW}					$5 \cdot 10^{-2}$	62	28.5	2 257
$V_{\text{AGDSW-M}}$					$5 \cdot 10^{-2}$	85	59.4	1 706
V_{AGDSW}	10^{-2}	78.9 (6.4)	81.7 (25.1)	127.7 (9.5)	10^{-2}	92	97.2	1 702
$V_{\text{AGDSW-M}}$	10^{-2}	112.2 (11.6)	119.5 (44.8)	181.1 (10.7)	10^{-2}	92	97.2	1 702
V_{SHEM}	10^{-3}	36.6 (3.3)	18.2 (6.8)	215.0 (8.4)	10^{-2}	48	19.9	4 450
V_{SHEM}	10^{-6}	80.1 (28.2)	14 283.8 (15 740.5)	189.2 (8.1)	10^{-4}	60	32.3	4 324

Table 2 Results for the coefficient functions in Fig. 3: tolerance for the selection of the eigenfunctions, iteration counts, condition numbers, and resulting coarse space dimension for different coarse space variants. (**Left:**) Averaged results for 100 random coefficient functions ($\approx 40\%$ density); standard deviation in brackets. GDSW never converged within the maximum iteration number of 500. $1/H = 4$, $H/h = 40$, and $\delta = 1h$; maximum coefficient $A_{\max} = 10^6$. (**Right:**) $1/H = 20$; $H/h = 40$; $\delta = 1h$; maximum coefficient $A_{\max} = 10^8$.

mass matrix in the right hand side of the eigenvalue problem, cf. section 4, is denoted by $V^{tol}_{\text{AGDSW-M}}$, the variant using a slab of width $w = lh$ is denoted by $V^{tol}_{\text{AGDSW-E}(l)}$, and the SHEM coarse space by V^{tol}_{SHEM}; cf. [5].

In Table 1, we compare the different coarse spaces for the two coefficient functions illustrated in Fig. 2. It is evident that, for the coefficient function from Fig. 2 (left), the GDSW coarse space is not sufficient to yield a low condition number and a small number of iterations; see Table 1 (left). This is due to multiple disconnected, high coefficient channels and inclusions intersecting the interface. However, the GDSW coarse space is sufficient for the coefficient function from Fig. 2 (right); see Table 1 (right). Here, only one connected high coefficient component exists per edge, all other high coefficient components are entirely contained in the overlap. Let us remark that a reduction of the overlap to one element, i.e., $\delta = 1h$, and only using the standard GDSW coarse space leads to 207 iterations and a condition number of $8.97 \cdot 10^7$. In Table 1, all adaptive methods achieve low condition numbers and



	Coeff. function A from Fig. 4			
V_0	slab width (lh)	it.	κ	dim V_0
V_{GDSW}	-	55	761 497.6	5
$V_{\text{AGDSW-E}(l)}$	$1h$	26	10.8	23
$V_{\text{AGDSW-E}(l)}$	$2h$	26	10.8	23
$V_{\text{AGDSW-E}(l)}$	$3h$	26	10.8	23
$V_{\text{AGDSW-E}(l)}$	$4h$	26	10.8	19
$V_{\text{AGDSW-E}(l)}$	$7h$	28	10.8	15
$V_{\text{AGDSW-E}(l)}$	$10h$	30	15.0	11
$V_{\text{AGDSW-E}(l)}$	$13h$	32	19.9	7
$V_{\text{AGDSW-E}(l)}$	$42h$	31	19.9	7
V_{SHEM}	-	24	8.3	21

Fig. 4 Coefficient function with many connected channels intersecting the interface. Maximum coefficient $A_{\text{max}} = 10^6$ (dark blue); $1/H = 2$; $H/h = 42$; $\delta = 2h$.

Table 3 Results for the coefficient function in Fig. 4: slab width, iterations counts, condition numbers, and resulting coarse space dimension for different coarse space variants. A tolerance for the selection of the eigenfunctions of 10^{-3} was used for $V_{\text{AGDSW-E}(l)}$ and V_{SHEM}; $1/H = 2$, $H/h = 42$, and $\delta = 2h$; maximum coefficient $A_{\text{max}} = 10^6$.

converge in few iterations for both coefficient functions. For the coefficient function from Fig. 2 (left), both adaptive GDSW coarse spaces have higher coarse space dimensions compared to the SHEM coarse spaces. This can be explained as follows: first, the entire GDSW coarse space is always included in the AGDSW coarse space and second, all high coefficient components intersecting the interface are disconnected. For the coefficient function from Fig. 2 (right), many channels of high coefficients intersecting the interface are connected. Here, the coarse space $V_{\text{SHEM}}^{10^{-6}}$ has a dimension of 213, where both AGDSW approaches lead to a significantly lower coarse space dimension of 57 using a tolerance of 10^{-2}.

In Fig. 3 (left), a randomly generated coefficient function is displayed. Averaged results for 100 random coefficient functions are listed in Table 2 (left). The coefficient functions are constructed as follows: uniformly distributed numbers are randomly generated in the interval $[0,1]$. A value above 0.6 corresponds to a high coefficient $A_{\text{max}} = 10^6$ in a finite element. Otherwise the coefficient is set to $A_{\text{min}} = 1.0$. The coefficient of an element that touches the global domain boundary is always set to A_{min}.

The results in Table 2 (left) show that all adaptive coarse spaces (AGDSW and SHEM) yield low condition numbers and numbers of iterations. On average, compared to the SHEM coarse space, for these problems, the adaptive GDSW approaches have lower coarse space dimensions. For example, $V_{\text{SHEM}}^{10^{-6}}$ and $V_{\text{AGDSW}}^{10^{-2}}$ converge in approximately the same number of iterations, i.e., 80.1 and 78.9, respectively. However, $V_{\text{SHEM}}^{10^{-6}}$ has a coarse space dimension of 189.2, whereas the dimension of $V_{\text{AGDSW}}^{10^{-2}}$ is 127.7. This corresponds to a reduction by 33 percent.

We also consider a foam-like coefficient function, as depicted in Fig. 3 (right). The results in Table 2 (right) show that a robust preconditioner, with additional coarse constraints, is needed as V_{GDSW} requires over 3 000 iterations to converge. The adaptive GDSW variants and V_{SHEM} need few iterations to converge. However, $V_{\mathrm{SHEM}}^{10^{-4}}$ requires a much larger coarse space, of dimension 4 324, compared to $V_{\mathrm{AGDSW}}^{5\cdot 10^{-2}}$, dimension 2 257, while requiring approximately the same number of iterations to converge. This corresponds to a reduction by 48 percent.

We now investigate the use of different slab widths in the variant $V_{\mathrm{AGDSW-E}(l)}$; cf. section 4. We are able to reduce the computational cost by using small slabs. However, when the detection of connected high coefficient components is weakened, we may enlarge the coarse space. This can be observed clearly for the coefficient function in Fig. 4. Increasing the slab width decreases the resulting coarse space dimension for $V_{\mathrm{AGDSW-E}(l)}$; also cf. Table 3. In this particular example, a slab width of 13 is sufficient to achieve the same result as with the maximum slab width of 42 since the slab then contains only two high coefficient components per edge.

References

1. Clark R. Dohrmann, Axel Klawonn, and Olof B. Widlund. Domain decomposition for less regular subdomains: overlapping Schwarz in two dimensions. SIAM J. Numer. Anal., 46(4):2153–2168, 2008.
2. Clark R. Dohrmann, Axel Klawonn, and Olof B. Widlund. A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners. In Domain decomposition methods in science and engineering XVII, volume 60 of Lect. Notes Comput. Sci. Eng., pages 247–254. Springer, Berlin, 2008.
3. Victorita Dolean, Frédéric Nataf, Robert Scheichl, and Nicole Spillane. Analysis of a two-level Schwarz method with coarse spaces based on local Dirichlet-to-Neumann maps. Comput. Methods Appl. Math., 12(4):391–414, 2012.
4. Juan Galvis and Yalchin Efendiev. Domain decomposition preconditioners for multiscale flows in high-contrast media. Multiscale Modeling & Simulation, 8(4):1461–1483, 2010.
5. Martin J. Gander, Atle Loneland, and Talal Rahman. Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods. Technical report, arxiv.org. 16 Dec 2015.
6. Alexander Heinlein, Ulrich Hetmaniuk, Axel Klawonn, and Oliver Rheinbach. The approximate component mode synthesis special finite element method in two dimensions: Parallel implementation and numerical results. J. Computat. Appl. Math., Vol. 289:116–133, 2015.
7. Alexander Heinlein, Axel Klawonn, Jascha Knepper, and Oliver Rheinbach. Multiscale coarse spaces for overlapping Schwarz methods based on the ACMS space in 2D. Technical Report Preprint 2016-09 at http://tu-freiberg.de/fakult1/forschung/preprints, Technische Universität Bergakademie Freiberg, Fakultät für Mathematik und Informatik, 2016. Submitted 08/2016 to ETNA.
8. Alexander Heinlein, Axel Klawonn, Jascha Knepper, and Oliver Rheinbach. Adaptive GDSW. Technical report, 2017. In Preparation.
9. Ulrich Hetmaniuk and Richard B. Lehoucq. A special finite element method based on component mode synthesis. ESAIM: Mathematical Modelling and Numerical Analysis, 44(3):401–420, 4 2010.
10. Andrea Toselli and Olof Widlund. Domain decomposition methods—algorithms and theory, volume 34 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2005.

Improving the Parallel Performance of Overlapping Schwarz Methods by Using a Smaller Energy Minimizing Coarse Space

Alexander Heinlein[1], Axel Klawonn[1], Oliver Rheinbach[2], and Olof Widlund[3]

1 Introduction

The GDSW preconditioner (Generalized Dryja, Smith, Widlund; see also [8]) is a two-level additive overlapping Schwarz preconditioner with exact local solvers (cf. [16]) using a coarse space constructed from energy-minimizing functions. It can be written in the form

$$M_{\mathrm{GDSW}}^{-1} = \Phi K_0^{-1} \Phi^T + \sum_{i=1}^{N} R_i^T \widetilde{K}_i^{-1} R_i, \tag{1}$$

where $K_0 = \Phi^T K \Phi$ is the coarse space matrix and the $\widetilde{K}_i = R_i K R_i^T$ represent the overlapping local problems; cf. [4]. The matrix Φ is the essential ingredient of the GDSW preconditioner. It is composed of coarse space functions which are discrete harmonic extensions from the interface into the interior degrees of freedom of nonoverlapping subdomains. The values on the interface are restrictions of the elements of the nullspace of the operator to the edges, vertices, and faces of the decomposition. Therefore, for a scalar elliptic problem, the coarse basis functions form a partition of unity on all subdomains which do not touch the Dirichlet boundary.

For $\Omega \subset \mathbb{R}^2$ being decomposed into John subdomains, the condition number of the GDSW preconditioner is bounded by

$$\kappa\left(M_{\mathrm{GDSW}}^{-1} A\right) \leq C \left(1 + \frac{H}{\delta}\right) \left(1 + \log\left(\frac{H}{h}\right)\right)^2, \tag{2}$$

[1]Mathematisches Institut, Universität zu Köln, Weyertal 86-90, 50931 Köln, Germany. e-mail: {alexander.heinlein,axel.klawonn}@uni-koeln.de
[2]Institut für Numerische Mathematik und Optimierung, Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg, Akademiestr. 6, 09596 Freiberg, Germany. e-mail: oliver.rheinbach@math.tu-freiberg.de
[3]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012, USA. e-mail: widlund@cims.nyu.edu

cf. [3, 4]. Here, H is the size of a subdomain, h the size of a finite element, and δ is the overlap.

GDSW-type preconditioners have succesfully been developed for almost incompressible elasticty, e.g., [5] and problems in $H(\mathrm{curl})$ [2]. An efficient parallel implementation of the GDSW preconditioner based on Trilinos [13] was recently introduced by the authors in [10]. Although the preconditioner can use geometric information, a focus in [10] was to make use of the Trilinos infrastructure to construct the preconditioner algebraically from the assembled sparse stiffness matrix.

A coarse space for Overlapping Schwarz methods in two dimensions related to but smaller than the standard GDSW coarse space has been considered in [6]. Following [7], in this paper, we consider two reduced versions of the GDSW coarse space in three dimensions denoted by *Option 1* and *Option 2.2* in [7]. These spaces are also smaller than the standard GDSW coarse space. In the following, we will denote this reduced GDSW coarse space as RGDSW. Our reduced coarse spaces have a relation to discretization methods such as Multiscale Finite Element Methods (MsFEM), which also use harmonic extensions; see, e.g., [14, 17].

2 A Reduced GDSW Coarse Space

We have implemented the RGDSW coarse space in our parallel preconditioner [10] since, among the proposed options in [7], it is the most algebraic. As in the standard version, we introduce coarse basis functions that form a partition of unity on the interface of the domain decomposition. Again, we extend the values on the interface as discrete harmonic functions into the interior of the nonoverlapping subdomains.

Let $\mathscr{S}_n$ be the index set of all subdomains which share the node n. A node n_i is called an ancestor of n_j if $\mathscr{S}_{n_j} \subset \mathscr{S}_{n_i}$. If no other node is an ancestor of a node n_j, it is called a coarse node. Using this definition, we can construct for each coarse node n_i a coarse basis function φ_i such that

$$\sum_{n_i \text{ coarse node}} \varphi_i = 1$$

on all subdomains which do not touch the Dirichlet boundary. A coarse basis function φ_i is constructed as follows:

$$\varphi_i(n) = \begin{cases} \frac{1}{|\mathscr{C}_n|} & \text{if } n_i \in \mathscr{C}_n, \\ 0 & \text{otherwise,} \end{cases}$$

with $\mathscr{C}_n$ being the set of all ancestors of the interface node n; cf. Fig. 1 (top). On the Dirichlet boundary, we set all coarse basis functions to zero.

Another option to define a reduced coarse space, using basis function based on an inverse distance weighting approach, has been introduced in [7, eq. (5)]. In particular, according to [7, eq. (5)], the values of the coarse basis function on the interface are chosen as

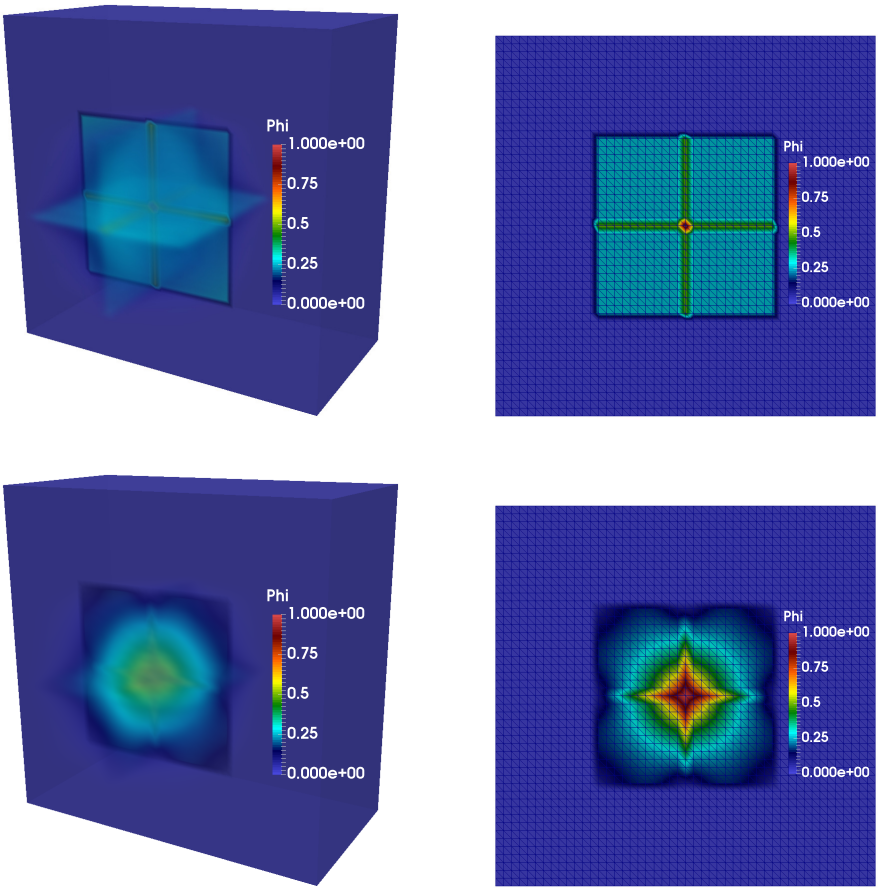


Fig. 1 Plot of the coarse basis function corresponding to the center node for the reduced GDSW coarse spaces, denoted *Option 1* (top) and *Option 2.2* (bottom) in [7]. Here, we assume the structured decomposition of a cube into 4x4x4 cubic subdomains.

$$\varphi_i(n) = \begin{cases} \frac{1/d_i(n)}{1/d_1(n)+1/d_2(n)+1/d_3(n)+1/d_4(n)} & \text{if } n_i \in \mathscr{C}_n, i \in \{1,2,3,4\} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

for components with four coarse nodes. Here, $d_i(n), i = 1,\ldots,4$ is the distance to the coarse node n_i. For components with any number of coarse nodes, we set

$$\varphi_i(n) = \begin{cases} \frac{1/d_i(n)}{\sum_{n_j \in \mathscr{C}_n} 1/d_j(n)} & \text{if } n_i \in \mathscr{C}_n, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

on the interface; cf. Fig. 1 (bottom). This construction is denoted as *Option 2.2* in [7]. As we will observe in section 4, this choice leads to a better convergence,

	Dimension of the Coarse Space		
# subdomains	Standard GDSW	RGDSW (Option 1&2.2)	Reduction
2^3	19	1	94.74 %
4^3	279	27	90.32 %
8^3	2 863	343	88.02 %
16^3	25 695	3 375	86.87 %
24^3	89 999	12 167	86.48 %
32^3	217 279	29 791	86.29 %
40^3	429 039	59 319	86.17 %
80^3	3 507 679	493 039	85.94 %
100^3	6 880 599	970 299	85.90 %
$1\,000^3$	$7.0 \cdot 10^9$	$1.0 \cdot 10^9$	85.73 %
$10\,000^3$	$7.0 \cdot 10^{12}$	$1.0 \cdot 10^{12}$	85.72 %

Table 1 Dimension of the coarse spaces and the reduction due to the use of the reduced coarse spaces in percent. We use one subdomain for each processor core.

	scalar elliptic		compressible linear elasticity
	face paths	edge paths	face paths
Option 1	$\alpha = 1$	$\alpha = 2$	$\alpha = 1$
Option 2.2	$\alpha = 0$	$\alpha = 1$	$\alpha = 0$

Table 2 Values of α in the condition number bound (5). For the definition of quasi-monotone paths, see [7].

however, it relies on additional geometric information to allow for the computation of the distance between interface nodes and the relevant coarse nodes. Therefore, it can be regarded as less algebraic compared to *Option 1*.

The advantage of these two reduced GDSW coarse problems over the classical GDSW coarse problem is their smaller size; cf. Fig. 2. Indeed, in 3D, for structured decompositions, they are smaller by more than 85 percent; cf. Table 1. This can be a significant advantage when striving for better parallel scalability on larger supercomputers.

For the reduced coarse spaces, for scalar elliptic problems in 3D as well as elasticity, the condition number of the preconditioned operator satisfies

$$\kappa(M_{\mathrm{RGDSW}}^{-1}A) \leq C \left(1 + \frac{H}{\delta}\right) \left(1 + \log\left(\frac{H}{h}\right)\right)^{\alpha}, \tag{5}$$

where α is given in Table 2; cf. [7] and also see Fig. 4.

3 Implementation

Our parallel implementation of the GDSW preconditioner and its more recent version with a reduced coarse space size (here denoted by RGDSW) is based on the implementation described in [10, 9, 12, 11]. We use Trilinos version 12.0; cf. [13].

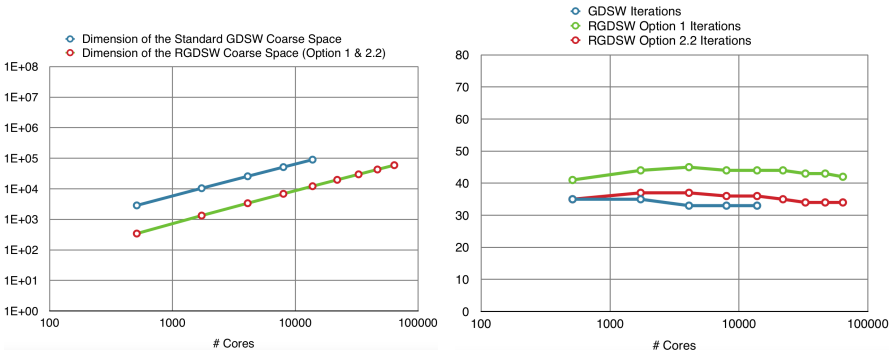


Fig. 2 We compare for a Laplace model problem in three dimensions: dimension of the coarse spaces (left) and corresponding numbers of iterations for the standard and the reduced GDSW coarse space (right); we use $H/h = 30$ and two layers of overlap. Computations run on the JUQUEEN supercomputer.

In our experiments presented here, for simplicity, we use a structured decomposition of our cubic computational domain into cubic subdomains. The overlapping subdomain problems and the coarse problem are solved using Mumps 4.10.0 (cf. [1]) in sequential mode. On the JUQUEEN BG/Q supercomputer, we use the IBM XL compilers 12.1 and the ESSL 5.1 when compiling Trilinos and the GDSW preconditioner. On the magnitUDE supercomputer at Universität Duisburg-Essen, we use the Intel compiler and the Intel MKL 2017.1.132.

4 Numerical Results

Based on the infrastructure given by our parallel implementation [10], we compare the reduced coarse space (denoted by RGDSW) to the standard coarse space (denoted by GDSW) for a scalar elliptic problem in 3D. Our numerical results in Fig. 2, and 3 show that the smaller dimension of the new coarse spaces *Option 1* and *Option 2.2* proposed in [7] indeed help to increase the parallel efficiency of the method significantly; see also Tables 3 and 4. By "Total Time", we denote the total time to solution including the assembly of the problem. The "Setup Time" includes the assembly of the problem and the setup of the preconditioner. This includes the factorization of the subdomain matrices. Finally, "Solver Time" only denotes the time spent in the GMRES iteration. The number of Krylov iterations for the new methods increases but only slightly in comparison with the standard GDSW preconditioner (cf. Fig. 2, right), as also demonstrated in [7]; the increase is too small to be reflected in the computation times. Indeed, as shown in Fig. 3, the total time to solution is always smaller for the new coarse spaces.

Acknowledgements
This work was supported in part by the German Research Foundation (DFG) through the Priority

	GDSW					RGDSW Option 1					RGDSW Option 2.2				
# cores = # subdomains	# it.	Setup Time	Solver Time	Total Time	Effic.	# it.	Setup Time	Solver Time	Total Time	Effic.	# it.	Setup Time	Solver Time	Total Time	Effic.
512	35	45.94 s	16.75 s	62.69 s	96%	41	43.84 s	18.82 s	62.66 s	96%	35	43.94 s	15.97 s	59.91 s	100%
1 728	35	47.47 s	19.49 s	67.23 s	89%	44	44.96 s	20.98 s	65.94 s	91%	37	45.15 s	17.55 s	62.70 s	96%
4 096	33	52.46 s	23.49 s	75.95 s	79%	45	45.66 s	22.56 s	68.22 s	88%	37	45.78 s	18.35 s	64.13 s	93%
8 000	33	63.55 s	31.73 s	92.28 s	65%	44	48.91 s	23.89 s	72.80 s	83%	36	49.35 s	19.35 s	68.70 s	87%
13 824	33	81.61 s	46.77 s	128.38 s	47%	44	51.08 s	26.28 s	77.36 s	77%	36	50.92 s	21.30 s	72.22 s	83%
21 952	Out of memory					44	53.98 s	30.26 s	84.24 s	71%	35	53.88 s	23.89 s	77.77 s	77%
32 768	Out of memory					43	57.71 s	34.27 s	91.98 s	65%	34	57.93 s	27.12 s	85.05 s	70%
46 656	Out of memory					43	64.48 s	41.60 s	106.08 s	56%	34	65.41 s	32.59 s	98.00 s	61%
64 000	Out of memory					42	74.18 s	48.17 s	122.35 s	49%	34	76.26 s	38.90 s	115.16 s	52%

Table 3 Detailed timers corresponding to Fig. 3 (left). Base line for the efficiency ("Effic.") is the fastest "Total Time" for 512 processor cores.

	GDSW				RGDSW Option 1				RGDSW Option 2.2			
# cores =	First Level		Second Level		First Level		Second Level		First Level		Second Level	
# subdomains	Setup	Apply	Setup	Apply	Setup	Apply	Setup	Apply	Setup	Apply	Setup	Apply
512	13.54 s	11.36 s	12.95 s	0.30 s	14.69 s	13.19 s	9.78 s	2.52 s	14.68 s	11.36 s	9.81 s	2.15 s
1 728	13.71 s	11.57 s	14.51 s	0.55 s	15.45 s	14.44 s	10.01 s	3.12 s	15.61 s	12.20 s	10.04 s	2.64 s
4 096	14.71 s	10.97 s	18.22 s	10.50 s	15.79 s	14.89 s	10.34 s	4.24 s	15.90 s	12.25 s	10.35 s	3.48 s
8 000	16.97 s	11.16 s	26.62 s	18.71 s	18.14 s	14.81 s	11.06 s	5.74 s	18.37 s	12.04 s	11.13 s	4.74 s
13 824	17.21 s	11.15 s	44.58 s	34.22 s	18.53 s	14.77 s	12.71 s	8.14 s	18.40 s	12.13 s	12.60 s	6.69 s
21 952	Out of memory				18.61 s	14.98 s	15.40 s	12.05 s	18.57 s	11.97 s	15.35 s	9.62 s
32 768	Out of memory				18.48 s	14.39 s	19.39 s	16.89 s	18.72 s	11.42 s	19.43 s	13.52 s
46 656	Out of memory				18.70 s	14.66 s	25.56 s	23.98 s	18.73 s	11.66 s	26.61 s	18.92 s
64 000	Out of memory				18.96 s	14.36 s	34.94 s	31.16 s	18.92 s	11.71 s	37.13 s	25.34 s

Table 4 Detailed timers corresponding to Fig. 3 (right).

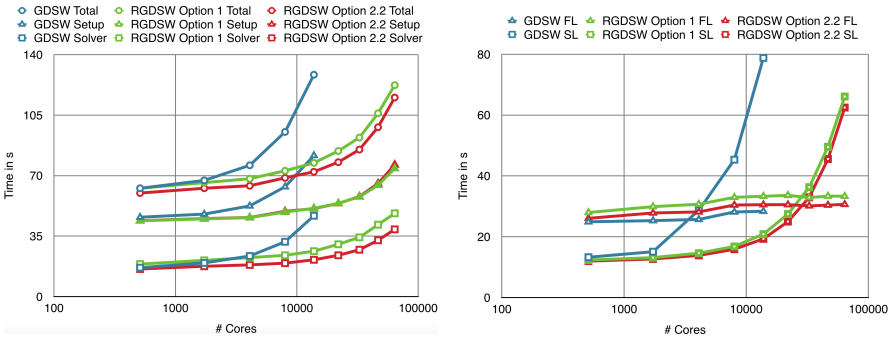
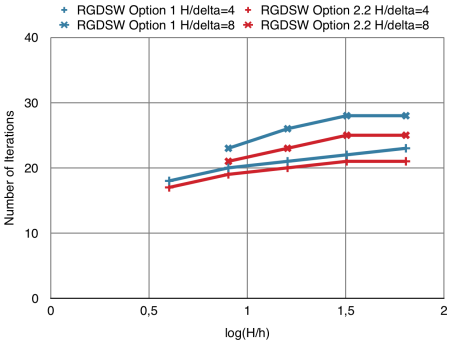


Fig. 3 Detailed times for the computations of a Laplace model problem in three dimensions using the standard GDSW coarse space and the reduced GDSW coarse space; we use $H/h = 30$ and two layers of overlap. Computations run on the JUQUEEN supercomputer.



H/h	Option 1		Option 2.2	
	$H/\delta = 4$	$H/\delta = 8$	$H/\delta = 4$	$H/\delta = 8$
4	18	-	17	-
8	20	23	19	21
16	21	26	20	23
32	22	28	21	25
64	23	28	21	25

Fig. 4 Numbers of iterations versus $\log(H/h)$ for the reduced GDSW coarse space and $1/H = 4$. Computations run on the magnitUDE supercomputer.

Programme 1648 “Software for Exascale Computing” (**SPPEXA**) under grants KL 2094/4-2 and RH 122/3-2. The work of the fourth author was supported by the National Science Foundation Grant DMS-1522736

The authors gratefully acknowledge the computing time granted by the Center for Computational Sciences and Simulation (CCSS) at Universität Duisburg-Essen and provided on the supercomputer magnitUDE (DFG grants INST 20876/209-1 FUGG, INST 20876/243-1 FUGG) at Zentrum für Informations- und Mediendienste (ZIM).

The authors also gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for providing computing time on the GCS Supercomputer JUQUEEN BG/Q supercomputer ([15]) at JSC Jülich. GCS is the alliance of the three national supercomputing centres HLRS (Universität Stuttgart), JSC (Forschungszentrum Jülich), and LRZ (Bayerische Akademie der Wissenschaften), funded by the German Federal Ministry of Education and Research (BMBF) and the German State Ministries for Research of Baden-Württemberg (MWK), Bayern (StMWFK) and Nordrhein-Westfalen (MIWF).

References

1. Patrick R. Amestoy, Iain S. Duff, Jean-Yves L'Excellent, and Jacko Koster. A fully asynchronous multifrontal solver using distributed dynamic scheduling. SIAM J. Matrix Anal. Appl., 23(1):15–41, January 2001.
2. Juan G. Calvo. A two-level overlapping Schwarz method for $H(\mathrm{curl})$ in two dimensions with irregular subdomains. Electron. Trans. Numer. Anal., 44:497–521, 2015.
3. Clark R. Dohrmann, Axel Klawonn, and Olof B. Widlund. Domain decomposition for less regular subdomains: overlapping Schwarz in two dimensions. SIAM J. Numer. Anal., 46(4):2153–2168, 2008.
4. Clark R. Dohrmann, Axel Klawonn, and Olof B. Widlund. A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners. In Domain decomposition methods in science and engineering XVII, volume 60 of Lect. Notes Comput. Sci. Eng., pages 247–254. Springer, Berlin, 2008.
5. Clark R. Dohrmann and Olof B. Widlund. Hybrid domain decomposition algorithms for compressible and almost incompressible elasticity. Internat. J. Numer. Methods Engrg., 82(2):157–183, 2010.
6. Clark R. Dohrmann and Olof B. Widlund. An alternative coarse space for irregular subdomains and an overlapping Schwarz algorithm for scalar elliptic problems in the plane. SIAM J. Numer. Anal., 50(5):2522–2537, 2012.
7. Clark R. Dohrmann and Olof B. Widlund. On the design of small coarse spaces for domain decomposition algorithms. SIAM Journal on Scientific Computing, 39(4):A1466–A1488, 2017.
8. Maksymilian Dryja, Barry F. Smith, and Olof B. Widlund. Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions. SIAM J. Numer. Anal., 31(6):1662–1694, December 1994.
9. Alexander Heinlein. Parallel Overlapping Schwarz Preconditioners and Multiscale Discretizations with Applications to Fluid-Structure Interaction and Highly Heterogeneous Problems. PhD thesis, Universität zu Köln, Germany, 2016.
10. Alexander Heinlein, Axel Klawonn, and Oliver Rheinbach. A parallel implementation of a two-level overlapping Schwarz method with energy-minimizing coarse space based on Trilinos. SIAM J. Sci. Comput., 38(6):C713–C747, 2016.
11. Alexander Heinlein, Axel Klawonn, and Oliver Rheinbach. Parallel two-level overlapping Schwarz methods in fluid-structure interaction. In Bülent Karasözen, Murat Manguoğlu, Münevver Tezer-Sezgin, Serdar Göktepe, and Ömür Uğur, editors, Numerical Mathematics and Advanced Applications ENUMATH 2015, pages 521–530, Cham, 2016. Springer International Publishing. Also Preprint 2015-15 at http://tu-freiberg.de/fakult1/forschung/preprints.
12. Alexander Heinlein, Axel Klawonn, and Oliver Rheinbach. Parallel overlapping Schwarz with an energy-minimizing coarse space. In Chang-Ock Lee, Xiao-Chuan Cai, David E. Keyes, Hyea Hyun Kim, Axel Klawonn, Eun-Jae Park, and Olof B. Widlund, editors, Domain Decomposition Methods in Science and Engineering XXIII, volume 116 of Lect. Notes Comput. Sci. Eng., pages 353–360, Cham, 2017. Springer International Publishing.
13. Michael A Heroux, Roscoe A Bartlett, Vicki E Howle, Robert J Hoekstra, Jonathan J Hu, Tamara G Kolda, Richard B Lehoucq, Kevin R Long, Roger P Pawlowski, Eric T Phipps, Andrew G Salinger, Heidi K Thornquist, Ray S Tuminaro, James M Willenbring, Alan Williams, and Kendall S Stanley. An overview of the Trilinos project. ACM Trans. Math. Softw., 31(3):397–423, 2005.
14. Thomas Y. Hou and Xiao-Hui Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. J. Comput. Phys., 134(1):169 – 189, 1997.
15. Michael Stephan and Jutta Docter. JUQUEEN: IBM Blue Gene/Q Supercomputer System at the Jülich Supercomputing Centre. Journal of large-scale research facilities, 1:A1, 2015.
16. Andrea Toselli and Olof Widlund. Domain decomposition methods—algorithms and theory, volume 34 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2005.
17. Jan Van lent, Robert Scheichl, and Ivan G. Graham. Energy-minimizing coarse spaces for two-level Schwarz methods for multiscale PDEs. Numer. Linear Algebra Appl., 16(10):775–799, 2009.

Inexact Dual-Primal Isogeometric Tearing and Interconnecting Methods

Christoph Hofer, Ulrich Langer, and Stefan Takacs

1 Introduction

Isogeometric Analysis (IgA), cf. Hughes et al. [2005], Beirão da Veiga et al. [2014], is a variant of the Galerkin method where both the geometry of the computational domain and the solution of the partial differential equation (PDE) are represented by B-splines or Non Uniform Rational B-splines (NURBS). One of the strengths of IgA consists in its capability of creating high-order smooth function spaces, while keeping the number of degrees of freedom relatively small. Originally, IgA was formulated by means of one *global geometry mapping*, which restricts the method to simple domains being topologically equivalent to the unit square or the unit cube. More complicated domains are represented as a non-overlapping composition of such simple domains, called *patches*. In such a *multi-patch* setting, each of the patches has its own geometry mapping, and all of the patches are discretized separately.

We are interested in fast solvers for linear systems arising from the discretization of elliptic PDEs in such a multi-patch setting. The local discretization on each patch has typically tensor-product structure.

We use a non-overlapping domain decomposition (DD) method to couple the problem across the patches, namely the dual-primal IsogEometric Tearing and Interconnecting (IETI-DP) method, a variant of the FETI-DP method, see Kleiss et al. [2012]. In general, the geometry mapping does not exhibit more than C^0-continuity across the interfaces. Thus, we only aim to guarantee C^0-continuity of the solution across the interfaces. Moreover, also for

Christoph Hofer, Ulrich Langer
Johannes Kepler University (JKU), Altenbergerstr. 69, A-4040 Linz, Austria, e-mail: {christoph.hofer,ulrich.langer}@jku.at

Ulrich Langer, Stefan Takacs
Austrian Academy of Sciences, RICAM, Altenbergerstr. 69, A-4040 Linz, Austria, e-mail: {ulrich.langer,stefan.takacs}@ricam.oeaw.ac.at

a decomposition of the patches into smaller subpatches, e.g., for parallelization, the choice of C^0 continuity is reasonable if the number of inner dofs stays large enough, cf. Hofer [2017]. The IETI method is closely related to the BDDC method, see Toselli and Widlund [2005], Beirão da Veiga et al. [2013, 2017] and references therein.

So far, the local problems have been solved using direct solvers. Since we want to choose the given patches also as subdomains of the DD-method, the local problems become large if the discretization is refined. In this case, inexact solvers for the local subproblems, as introduced in Klawonn and Rheinbach [2007], could be superior to direct solvers. The aim of this work is to investigate such approaches in combination with the p-robust multigrid solvers, which were proposed by Hofreither and Takacs [2017], as inexact solvers.

In the present paper, we consider the Poisson problem on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, with $d \in \{2, 3\}$, as model problem: For a given, sufficiently smooth f, find $u \in V_0 := H^1_0(\Omega)$ such that

$$a(u, v) := (\nabla u, \nabla v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)} =: \langle F, v \rangle \quad \forall v \in V_0. \tag{1}$$

2 Isogeometric Analysis and IETI-DP

On the unit interval, for any spline degree p and number of basis functions M, we define the basis $(\widehat{N}_{i,p})_{i=1}^M$ of univariate B-splines of maximum smoothness C^{p-1} via Cox-de Boor's algorithm. A basis for the *parameter domain* $\widehat{\Omega} := (0, 1)^d$, is realized by the tensor product of such basis functions, again denoted by $\widehat{N}_{i,p}$, where $i = (i_1, \dots, i_d) \in \mathcal{I} := \{1, \dots, M_1\} \times \dots \times \{1, \dots, M_d\}$ and $p = (p_1, \dots, p_d)$ are multi-indices.

In standard (single-patch) IgA, the *physical domain* Ω is given as the image of the parameter domain under the *geometry mapping* $G : \widehat{\Omega} \to \mathbb{R}^d$, defined by $G(\xi) := \sum_{i \in \mathcal{I}} P_i \widehat{N}_{i,p}(\xi)$, with the control points $P_i \in \mathbb{R}^d$, $i \in \mathcal{I}$.

In a multi-patch setting, the domain Ω (*multipatch domain*) is composed of non-overlapping patches $\Omega^{(k)}$, $k = 1, \dots, N$, such that $\overline{\Omega} := \bigcup_{k=1}^N \overline{\Omega}^{(k)}$. Each patch $\Omega^{(k)} := G^{(k)}(\widehat{\Omega})$ is represented by its own geometry mapping. We call $\Gamma := \bigcup_{k>l} \partial\Omega^{(k)} \cap \partial\Omega^{(l)}$ the *interface*, and denote its restriction to one of the patches $\Omega^{(k)}$ by $\Gamma^{(k)} := \Gamma \cap \partial\Omega^{(k)}$. Throughout the paper, the superscript (k) denotes the restriction of the underlying symbol to $\Omega^{(k)}$.

We use B-splines not only for defining the geometry, but also for representing the approximate solution of (1). Once the basis functions are defined on the parameter domain $\widehat{\Omega}$, we define the bases on the patches $\Omega^{(k)}$ via the pull-back principle, and obtain the basis functions $N_{i,p} := \widehat{N}_{i,p} \circ G^{-1}$.

The main idea of IETI-DP is to decouple the patches by tearing the interface unknowns which introduces additional degrees of freedom (dofs). We denote the resulting space by V_h. Then, continuity is again enforced using

Lagrange multipliers λ. Hence, the local subproblems on each patch are essentially pure Neumann problems (at least for interior patches). Due to the presence of a kernel, a straight-forward Schur complement formulation is not possible. In order to overcome this problem, certain continuity conditions are enforced by incorporating them into the space V_h, (*strongly enforced continuity conditions*) which yields the smaller space $\widetilde{V}_h$. There, we formulate the following problem. Find $(u, \lambda) \in \widetilde{V}_h \times \Lambda$ such that

$$\begin{bmatrix} \widetilde{K} & \widetilde{B}^T \\ \widetilde{B} & 0 \end{bmatrix} \begin{bmatrix} u \\ \lambda \end{bmatrix} = \begin{bmatrix} \widetilde{f} \\ 0 \end{bmatrix}, \tag{2}$$

where $\widetilde{K}$ is the stiffness matrix, $\widetilde{B}$ the jump operator, and $\widetilde{f}$ the right hand side. Here and in what follows, we do not distinguish between the IgA functions and their vector representation with respect to the chosen basis.

Now, we split V_h into interior dofs and interface dofs, which yields an interface space W. By splitting $\widetilde{V}_h$ analogously, we obtain the space $\widetilde{W}$. Based on this splitting, we formulate the problem using the Schur complement of the stiffness matrix K in V_h with respect to the interface dofs: $S := K_{BB} - K_{BI}K_{II}^{-1}K_{IB}$, where the subindices B and I denote the boundary and interior dofs, respectively. The restriction of S to $\widetilde{W}$ is denoted by $\widetilde{S}$, which yields the following saddle-point formulation: Find $(w, \lambda) \in \widetilde{W} \times \Lambda$ such that

$$\begin{bmatrix} \widetilde{S} & \widetilde{B}^T \\ \widetilde{B} & 0 \end{bmatrix} \begin{bmatrix} w \\ \lambda \end{bmatrix} = \begin{bmatrix} \widetilde{g} \\ 0 \end{bmatrix}, \tag{3}$$

where $\widetilde{g} := \widetilde{I}^T(f_B - K_{BI}K_{II}^{-1}f_I)$ and $\widetilde{I} : \widetilde{W} \to W$ is the canonical embedding. We denote the subspace of $\widetilde{W}$ satisfying the strongly enforced continuity conditions homogeneously by W_Δ and the S-orthogonal complement by W_Π. In the literature, our choice of W_Π is often called *energy minimizing primal subspace*. Finally, we can define the Schur complement F of the saddle-point problem (3), and obtain the problem: Find $\lambda \in \Lambda$ such that

$$F\lambda := (\widetilde{B}\widetilde{S}^{-1}\widetilde{B}^T)\lambda = \widetilde{B}\widetilde{S}^{-1}\widetilde{g} := d. \tag{4}$$

Equation (4) is solved by means of the conjugate gradient (CG) method using the scaled Dirichlet preconditioner $M_{sD}^{-1} := B_D S B_D^T$, where B_D is a scaled version of the jump operator B on V_h. Note that we can approximate $\widetilde{S}^{-1}$ because $\widetilde{S}$ can be represented (by reordering of the dofs) as a block diagonal matrix of matrices $S_{\Delta\Delta}^{(k)}$ for each patch and the matrix $S_{\Pi\Pi}$. For a summary of the algorithm and a more detailed explanation, we refer, e.g., to Toselli and Widlund [2005], Hofer and Langer [2017] and references therein.

3 Incorporating Multigrid in IETI-DP

We investigate different possibilities to incorporate a multigrid solver into the IETI-DP algorithm. The application of the IETI-DP algorithm requires the solution of local Neumann and Dirichlet problems.

3.1 Local Dirichlet problems

We have to solve linear systems with the system matrix $K_{II}^{(k)}$ in the application of S in the preconditioner and when calculating the right hand side $\widetilde{g}$. These linear systems are Dirichlet problems (up to boundary conditions). The right hand side $\widetilde{g}$ has to be computed very accurately, i.e., at least up to discretization error. However, for the preconditioner, a few MG V-cycles are usually enough, since we only have to ensure the spectral equivalence of the inexact scaled Dirichlet preconditioner to the exact one, cf. Klawonn et al. [2016] and references therein.

3.2 Local Neumann problems

Local Neumann problems appear in the construction of the S-orthogonal basis for W_Π and in the application of $S_{\Delta\Delta}$. In order to construct the nodal and S-orthogonal basis $\{\phi_j^{(k)}\}_j$ of $W_\Pi^{(k)}$, we have to solve

$$\begin{bmatrix} S^{(k)} & C^{(k)^T} \\ C^{(k)} & 0 \end{bmatrix} \begin{bmatrix} \phi_j^{(k)} \\ \mu_j^{(k)} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{e}_j^{(k)} \end{bmatrix}, \quad \forall j \in \{1, \dots, n_\Pi^{(k)}\}, \tag{5}$$

where $\boldsymbol{e}_j^{(k)} \in \mathbb{R}^{n_\Pi^{(k)}}$ is the j-th unit vector, and the matrix $C^{(k)}$ realizes the $n_\Pi^{(k)}$ strongly enforced continuity conditions contributing to the patch $\Omega^{(k)}$. Instead of solving (5) directly, we solve

$$\begin{bmatrix} K^{(k)} & C^{(k)^T} \\ C^{(k)} & 0 \end{bmatrix} \begin{bmatrix} \overline{\phi}_j^{(k)} \\ \mu_j^{(k)} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{e}_j^{(k)} \end{bmatrix}, \quad \forall j \in \{1, \dots, n_\Pi^{(k)}\}, \tag{6}$$

and obtain the desired basis functions by $\phi_j = \overline{\phi}_j|_{\Gamma^{(k)}}$. Note that $\{\overline{\phi}_j^{(k)}\}_j$ is a K-orthogonal basis. The system is solved with the *Schöberl-Zulehner* (SZ) preconditioner, see Schöberl and Zulehner [2007].

The SZ preconditioner for (6) requires preconditioners $\hat{K}^{(k)}$ and $\hat{H}^{(k)}$ for the upper left block $K^{(k)}$ and its inexact Schur complement $H^{(k)} := C^{(k)}(\hat{K}^{(k)})^{-1}C^{(k)^T}$, respectively. The preconditioner $K^{(k)}$ is realized by a few

MG V-cycles. It is required that $\hat{K}^{(k)} > K^{(k)}$, which implies that $\hat{K}^{(k)}$ has to be positive definite. In order to handle also the case where $K^{(k)}$ is singular, we need to set up MG based on a regularized matrix $K_M^{(k)} := K^{(k)} + \alpha \widehat{M}^{(k)}$, where α is chosen to be 10^{-2} and $\widehat{M}^{(k)}$ is the mass matrix on the parameter domain. Note that we can exploit the tensor product structure to efficiently assemble the mass matrix $\widehat{M}^{(k)}$. Secondly, the SZ preconditioner requires that $\hat{H}^{(k)} < H^{(k)}$. Since in our case the number of rows of $C^{(k)}$ is given by $n_\Pi^{(k)}$, a small number that does not change during refinement, we calculate the inexact Schur complement exactly. This can be performed by applying $\left(\hat{K}^{(k)}\right)^{-1}$ to $n_\Pi^{(k)}$ vectors. Finally, by a suitable scaling, e.g., $\hat{H}^{(k)} := 0.99\, H^{(k)}$, we obtain the desired matrix inequality.

The second type of Neumann problem appears in the application of F. We look for a solution of the system $S_{\Delta\Delta}^{(k)} w_\Delta^{(k)} = f_\Delta^{(k)}$, which can be written as

$$\begin{bmatrix} S^{(k)} & {C^{(k)}}^T \\ C^{(k)} & 0 \end{bmatrix} \begin{bmatrix} w_\Delta^{(k)} \\ \mu^{(k)} \end{bmatrix} = \begin{bmatrix} f^{(k)} \\ 0 \end{bmatrix}. \tag{7}$$

Certainly, one can use the same method as above. However, we can utilize the fact that we search for a minimizer of $\frac{1}{2}(S^{(k)} w^{(k)}, w^{(k)}) - (w^{(k)}, f^{(k)})$ in the subspace given by $C^{(k)} w^{(k)} = 0$. This solution can be computed by first solving the unconstrained problem, and then projecting the minimizer into the subspace using a energy-minimizing projection. The projection is trivial because the decomposition of $\widetilde{W}$ into W_Π and W_Δ is S-orthogonal.

Note that the CG algorithm, when applied to a positive semidefinite matrix, stays in the factor space with respect to the kernel and computes one of the minimizers. The solution of the constrained minimization problem is, as outlined above, obtained by applying the projection. As long as the number of CG iterations is not too large, numerical instabilities are not observed when applying CG to a positive semidefinite problem.

The S-orthogonal basis has to be computed very accurately in order to maintain the orthogonality. Since the equation $S_{\Delta\Delta}^{(k)} w_\Delta^{(k)} = f_\Delta^{(k)}$ appears in the system matrix F, its solution also requires an accuracy of at least the discretization error.

3.3 Variants of inexact formulations

From the discussion above, we deduce four (reasonable) versions:

(D-D) The classical IETI-DP method, using direct solvers everywhere.

(D-MG) We use MG in the scaled preconditioner for the solution of the local Dirichlet problems and the transformation of the right hand side, see Section 3.1. As already mentioned, the required accuracy for computing $\widetilde{g}$ has

to be of the order of discretization error, whereas a few V-cycles are enough for the preconditioner.

(MG-MG) We use MG for all patch-local problems, i.e., the local Dirichlet and Neumann problems. This implies that also the calculation of the basis for W_Δ is performed by means of MG, which turns out to be very costly. Moreover, for each application of F, we have to solve a local Neumann problem in W_Δ with the accuracy in the order of the discretization error.

(MG-MG-S) To overcome the efficiency problem of the requirement of solving a linear system with MG very accurately, we use the saddle point formulation instead of F. On the one hand, at each iteration step, we only have to apply a given matrix instead of solving a linear system. On the other hand, we now have to deal with a saddle point problem. Moreover, the iteration is not only applied to the interface dofs, but also to the dofs in the whole domain.

We will always assume that the considered multipatch domain has only a moderate number of patches, such that the coarse problem can still be handled by a direct solver. For extensions to inexact version for the coarse problem, we refer to Klawonn and Rheinbach [2007].

For the first three methods, we use the CG method to solve $F\lambda = d$ as outer iteration. For (MG-MG-S), we have to deal with the saddle point problem (2), which we solve using the Bramble-Pasciak CG (BPCG) method, cf. Bramble and Pasciak [1988]. The building blocks for this method are a preconditioner $\widehat{\widetilde{K}}$ for $\widetilde{K}$ and $\hat{F}$ for the Schur complement F. The construction of $\widehat{\widetilde{K}}$ follows the same steps as in the previous section, but we only apply a few MG V-cycles. Concerning $\hat{F}$, a good choice is the scaled Dirichlet preconditioner M_{sD}^{-1}, cf. Klawonn and Rheinbach [2007].

4 Numerical Experiments

We solve the model problem (1) on a two and a three dimensional computational domain. In the two dimensional case, we use the quarter annulus divided into $32 = 8 \times 4$ patches, as illustrated in Fig. 1(left). The three dimensional domain is the twisted quarter annulus, decomposed into $128 = 4 \times 4 \times 8$ patches as presented in Fig. 1(right). We use B-splines of maximal smoothness inside a patch and C^0-coupling across the patch interfaces.

We have chosen the continuity of the vertex values and the edge averages for the two dimensional example, and the continuity of the edge averages for the three dimensional example as strongly enforced continuity conditions.

For the examples with polynomial degree $p = 2$, we use a standard MG method based on a hierarchy of nested grids keeping p fixed and use a standard Gauss Seidel (GS) smoother. For the examples with higher polynomial degree ($p = 4$ or 7), we have used $p = 1$ on all grid levels but the finest grid.

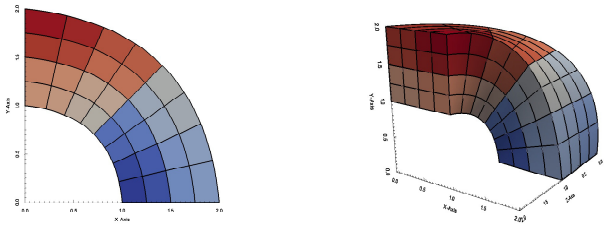


Fig. 1 Quarter annulus in 2d (left), twisted quarter annulus in 3d (right).

This does not yield nested spaces. Thus, we cannot use the canonical embedding and restriction. Instead, we use L^2-projections to realize them. On the finest grid, we use a MG smoother suitable for high-order IgA, namely a variant of the subspace-corrected mass smoother proposed and analyzed in Hofreither and Takacs [2017]. For this smoother, it was shown that a resulting MG method is robust with respect to both the grid size and the polynomial degree. However, for $p = 1$ or 2, standard approaches are more efficient. Thus, we again use this smoother only for the finest level, while for all other grid levels we use standard GS smoothers. To archive better results, we have modified the subspace-corrected mass smoother by incorporating a rank-one approximation of the geometry transformation.

For the outer CG or BPCG iteration, we use a zero initial guess, and the reduction of the initial residual by the factor 10^{-6} as stopping criterion. The local problems related to the calculation of the S-orthogonal basis are solved up to a tolerance of 10^{-12}. In case of the (MG-MG) version, the local Neumann problems (7) in W_Δ are solved up to a relative error of 10^{-10}. The number of MG cycles in the preconditioner is fixed. For the local Dirichlet problems in the scaled Dirichlet preconditioner, we use 2 V-cycles. The local Neumann problems, which appear in the preconditioner of the (MG-MG-S) version, are approximately solved by 3 V-cycles. In the following, we report on the number of CG iterations to solve (4) and BPCG iterations for (2) and the total time in seconds, which includes the assembling, the IETI-DP setup and solving phase. For the weak scalability tests in Table 1 and Table 2, we observe in all cases a polylogarithmic growth of the outer iterations and a quasi-optimal behavior of the computation time.

The algorithm is realized with the open source C++ library G+Smo[1] We utilize the PARDISO 5.0.0 Solver, cf. Kuzmin et al. [2013], for performing the LU factorizations. To allow a better comparison of the different variants, we only perform serial computations.[2]

[1] G+Smo (Geometry plus Simulation modules) v0.8.1, `http://gs.jku.at/gismo`.

[2] Our code is compiled with the `gcc 4.8.3` compiler with optimization flag `-O3`. The results are obtain on the RADON1 cluster at Linz. We use a single core of a node, equipped with 2x Xeon E5-2630v3 "Haswell" CPU (8 Cores, 2.4Ghz, 20MB Cache) and 128 GB RAM.

	D-D		**MG-D**		**MG-MG**		**MG-MG-S**	
$p = 2$\Dofs	It.	Time	It.	Time	It.	Time	It.	Time
134421	9	10	9	8	9	13	14	14
530965	10	45	10	37	10	54	15	90
2110485	11	224	11	172	11	272	16	568
8415253	11	1005	11	762	11	1181	15	3394
33607701	OoM		OoM		13	5070	OoM	
$p = 7$\Dofs	It.	Time	It.	Time	It.	Time	It.	Time
45753	10	26	10	27	10	57	14	54
155961	11	108	11	110	11	225	15	211
572985	12	498	12	495	12	1048	17	1013
2193465	13	2384	13	2265	14	4427	18	4344
8580153	OoM		OoM		15	18484	20	19958

Table 1 Numerical results for the quarter annulus in 2d.

	D-D		**MG-D**		**MG-MG**		**MG-MG-S**	
$p = 2$\Dofs	It.	Time	It.	Time	It.	Time	It.	Time
14079	11	3	11	3	11	8	25	7
86975	12	19	12	19	12	59	26	59
606015	14	213	14	197	14	484	30	616
4513343	OoM		16	2764	16	5244	35	11657
$p = 4$\Dofs	It.	Time	It.	Time	It.	Time	It.	Time
40095	13	30	13	33	13	112	23	104
160863	15	234	15	254	15	659	28	633
849375	16	2237	17	2356	17	5403	32	5298
5390559	OoM		OoM		19	45243	37	52831

Table 2 Numerical results for the twisted quarter annulus in 3d.

In Table 1, we summarize the results for the two dimensional domain for $p = 2$ and 7. The size of the coarse space W_{Π} is 73. We observe that replacing the direct solver in the preconditioner with two MG V-cycles does not change the number of outer iterations. Moreover, going from the Schur complement to the saddle point formulation and using BPCG there, leads only to a minor increase in the number of outer iterations. In all cases, the logarithmic dependence of the condition number on h is preserved. The advantage of the formulation using only MG, especially (MG-MG), is its smaller memory footprint, therefore, the possibility of solving larger systems. However, the setting with the best performance is (MG-D). Concluding, for small polynomial degrees and using the GS smoother, (MG-MG) gives reasonable trade off between performance and memory usage and for larger polynomial degrees, this setting can be still recommended if memory consumption is an issue.

In the case $p = 2$, for the inner iterations, we have observed that the CG needed on average 8 iterations to compute $\widetilde{g}$, the calculation of the S-orthogonal basis needed on average 14 iterations, and the solution of (7)

required on average 10 iterations. For the second case, $p = 7$, we needed 9 iterations to compute $\widetilde{g}$, 13 iterations for the calculation of the S-orthogonal basis and 10 iterations for the solutions of (7). Here and in what follows, we have taken the average over the patches, the individual levels and the individual steps of the outer iteration. We mention that the number of inner iterations was only varying slightly.

In Table 2, we summarize the results for the three dimensional domain and for $p = 2$ and 4. The size of the coarse space W_Π is 240. We observe that replacing the direct solver in the preconditioner with two MG V-cycles does not change the number of outer iterations. We further observe that the results are similar to the one of the two dimensional case. However, the number of iterations almost doubled when using BPCG for (MG-MG-S). In all cases, the logarithmic dependence of the condition number on h is preserved. The advantage of the formulation using only MG, especially (MG-MG), is its smaller memory footprint, therefore the possibility of solving larger systems. The best performance is obtained sometimes by (D-D) and sometimes by (MG-D), where both approaches are comparable in all cases.

Concerning the inner iterations, for $p = 2$, we need on average 15 CG iterations to compute $\widetilde{g}$, 22 CG iterations to build up each S-orthogonal basis function, and 18 CG iterations to solve (7). In the case of $p = 4$, we needed on average only 10 iterations to compute $\widetilde{g}$, 14 iterations for the construction of the S-orthogonal basis functions, and 11 iterations for solving (7).

The last test deals with the weak scalability of the method, where we only investigate the two dimensional setting for $p = 7$. We fix the ratio H/h and increase the number of patches. We expect constant number of iterations and a linear increase of the computation time. In Table 3, beside the Dofs, we report the size of the coarse space n_Π and the number of patches N. For each method, we provide the number of iterations and the computation time in seconds. We observe that the number of iterations and computation time behave as expected.

$p = 7$			**D-D**		**MG-D**		**MG-MG**		**MG-MG-S**	
n_Π	N	Dofs	It.	Time	It.	Time	It.	Time	It.	Time
73	32	45753	10	27	10	27	10	62	20	60
337	128	183921	11	111	11	108	11	268	15	234
1441	512	737505	11	446	11	438	11	1111	13	943
5953	2048	2953665	10	1777	10	1729	10	4468	12	3821
24193	8192	11821953	OoM		OoM		10	19691	11	15392

Table 3 Weak scalability of the methods with respect to the number of patches.

Acknowledgements This work was supported by the Austrian Science Fund (FWF) under the grant W1214, project DK4. This support is gratefully acknowledged.

References

L. Beirão da Veiga, D. Cho, L. F. Pavarino, and S. Scacchi. BDDC preconditioners for isogeometric analysis. *M³AS*, 23(6):1099–1142, 2013.

L. Beirão da Veiga, A. Buffa, G. Sangalli, and R. Vázquez. Mathematical analysis of variational isogeometric methods. *Acta Numer.*, 23:157–287, 2014.

L. Beirão da Veiga, L. F. Pavarino, S. Scacchi, O. B. Widlund, and S. Zampini. Adaptive selection of primal constraints for isogeometric BDDC deluxe preconditioners. *SISC*, 39(1):A281–A302, 2017.

J. H. Bramble and J. E. Pasciak. A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. *Math. Comput.*, 50(181):1–17, 1988.

C. Hofer. Parallelization of continuous and discontinuous Galerkin dual–primal isogeometric tearing and interconnecting methods. *CAMWA*, 74(7):1607 – 1625, 2017.

C. Hofer and U. Langer. Dual-primal isogeometric tearing and interconnecting solvers for multipatch dG-IgA equations. *CMAME*, 316:2 – 21, 2017.

C. Hofreither and S. Takacs. Robust multigrid for isogeometric analysis based on stable splittings of spline spaces. *SINUM*, 55(4):2004–2024, 2017.

T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs. Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *CMAME*, 194:4135–4195, 2005.

A. Klawonn and O. Rheinbach. Inexact FETI-DP methods. *Int. J. Numer. Methods Eng.*, 69(2):284–307, 2007.

A. Klawonn, M. Lanser, and O. Rheinbach. *A Highly Scalable Implementation of Inexact Nonlinear FETI-DP Without Sparse Direct Solvers*, pages 255–264. Springer International Publishing, Cham, 2016.

S. Kleiss, C. Pechstein, B. Jüttler, and S. Tomar. IETI–isogeometric tearing and interconnecting. *CMAME*, 247:201–215, 2012.

A. Kuzmin, M. Luisier, and O. Schenk. Fast methods for computing selected elements of the Greens function in massively parallel nanoelectronic device simulations. In F. Wolf, B. Mohr, and D. Mey, editors, *Euro-Par 2013*, volume 8097 of *LNCS*, pages 533–544. Springer Berlin Heidelberg, 2013.

J. Schöberl and W. Zulehner. Symmetric Indefinite Preconditioners for Saddle Point Problems with Applications to PDE-Constrained Optimization Problems. *SIMAX*, 29(3):752–773, 2007.

A. Toselli and O. B. Widlund. *Domain decomposition methods – algorithms and theory.* Berlin: Springer, 2005.

Abstract In Isogeometric Analysis (IgA), non-trivial computational domains are often composed of volumetric patches where each of them is discretized by means of tensor-product B-splines or NURBS. In such a setting, the dual-primal IsogEometric Tearing and Interconnecting (IETI-DP) method, that is nothing but the generalization of the FETI-DP method to IgA, has proven to be a very efficient solver for huge systems of IgA equations. Using IETI-DP, basically any patch-local solver can be extended to the global problem. So far, only direct solvers have been considered as patch-local solvers. In the present paper, we compare them with the option of using robust multigrid as patch-local solver. This is of special interest for large-scale patch-local systems or / and for large spline degrees, because the convergence of standard smoothers deteriorates with large spline degrees and the robust multigrid smoother chosen is only available on tensor-product discretizations.

Coupling Parareal and Dirichlet-Neumann/Neumann-Neumann Waveform Relaxation Methods for the Heat Equation

Yao-Lin Jiang[1] and Bo Song[2]

1 Introduction

We introduce two new space-time Waveform Relaxation (WR) methods based on the parareal algorithms and Dirichlet-Neumann waveform relaxation (DNWR) and Neumann-Neumann waveform relaxation (NNWR). The WR method was first introduced by Lelaramee, Ruehli and Sangiovanni-Vincentelli [15], which has been applied to analyze for many different kinds of problems, such as differential algebraic equations[11], fractional differential equations [13] , reaction diffusion equations [17]; for further details, see [12]. Domain decomposition methods for time-dependent partial differential equations (PDEs) can also lead to WR methods, i.e. Schwarz waveform relaxation (SWR) algorithm [8, 10], optimized Schwarz waveform relaxation (OSWR) algorithm [2, 3], and Dirichlet-Neumann and Neumann-Neumann waveform relaxation methods [6, 7, 22].

The parareal algorithm is a time-parallel method that was proposed by Lions, Maday, and Turinici in the context of virtual control to solve evolution problems in parallel [16]. In this algorithm, initial value problems are solved on subintervals in time, and through iterations the initial values on each subinterval are corrected to converge to the correct values of the overall solution [1, 9, 5]. The parareal algorithm has also been combined with waveform relaxation methods [18].

Parallel algorithms based on the decomposition of both time and space domain have been also studied [21, 19]. However, there was no parallel mechanism in the time direction. In [20], it was the first time that the combination of Schwarz waveform relaxation and parareal for PDEs had been introduced. Further, in [4], a new parallel algorithm where there is no order between the Schwarz waveform relaxation algorithm and the parareal algorithm was introduce.

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China, e-mail: yljiang@xjtu.edu.cn .[2] Corresponding author. School of Science, Northwestern Polytechnical University, Xi'an 710072, China, e-mail: bosong@nwpu.edu.cn

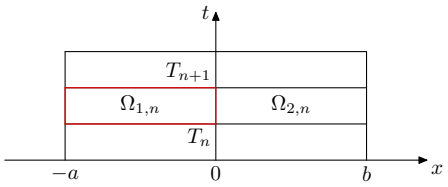


Fig. 1 Space time decomposition on which the proposed algorithms are based.

In this paper, we propose the parareal Dirichlet-Neumann waveform relaxation (PA-DNWR) and the parareal Neumann-Neumann waveform relaxation (PA-NNWR) methods for the time-dependent problem. For ease of presentation for the new algorithms, we derive our results for two subdomains in one spatial dimension.

We consider the following initial-value problem of heat equation on bounded $\Omega \subset \mathbb{R}$

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = f(x,t), & x \in \Omega,\ 0 < t < T, \\ u(x,0) = u_0(x), & x \in \Omega, \\ u(x,t) = g(x,t), & x \in \partial\Omega,\ 0 < t < T. \end{cases} \tag{1}$$

2 Parareal Dirichlet-Neumann/Neumann-Neumann waveform relaxation algorithms

We define the new algorithms for the model problem (1) on the space-time domain $\Omega \times (0,T) = (-a,b) \times (0,T)$. We assume that Ω is decomposed into two nonoverlapping subdomains, i.e. $\Omega_1 = (-a,0)$ and $\Omega_2 = (0,b)$, and the time interval $(0,T)$ is decomposed into N equal time subintervals (T_n, T_{n+1}) with $\Delta T = T_{n+1} - T_n = T/N$, $n = 0,1,\ldots,N-1$. We then can define the non-overlapping space-time subdomain $\Omega_{i,n} = \Omega_i \times (T_n, T_{n+1})$, $i = 1,2, n = 0,1,\ldots,N-1$; see Figure 1.

In order to introduce the parareal Dirichlet-Neumann waveform relaxation algorithm for the model problem (1), we first introduce several propagators. We define two propagator $F_{1,n}(U(x),\omega(t))$ and $G_{1,n}(U(x),\omega(t))$ to solve the following Dirichlet problem in $\Omega_{1,n}$

$$\begin{cases} \dfrac{\partial u_{1,n}}{\partial t} = \dfrac{\partial^2 u_{1,n}}{\partial x^2} + f(x,t), & (x,t) \in \Omega_{1,n}, \\ u_{1,n}(-a,t) = g(-a,t), & t \in (T_n, T_{n+1}), \\ u_{1,n}(0,t) = \omega(t), & t \in (T_n, T_{n+1}), \\ u_{1,n}(x,T_n) = U(x), & x \in \Omega_1, \end{cases} \tag{2}$$

using an accurate approximation and a rough approximation, where $U(x)$ and $\omega(t)$ are given data. Furthermore, two propagators $F_{2,n}(U(x),\omega(x,t))$ and $G_{2,n}(U(x),\omega(x,t))$ are defined to solve the following Neumann problem in $\Omega_{2,n}$

$$\begin{cases} \dfrac{\partial u_{2,n}}{\partial t} = \dfrac{\partial^2 u_{2,n}}{\partial x^2} + f(x,t), & (x,t) \in \Omega_{2,n}, \\ \partial_x u_{2,n}(0,t) = \partial_x \omega(0,t), & t \in (T_n, T_{n+1}), \\ u_{2,n}(b,t) = g(b,t), & t \in (T_n, T_{n+1}), \\ u_{2,n}(x,T_n) = U(x), & x \in \Omega_2, \end{cases} \tag{3}$$

using an accurate approximation and a rough approximation. Therefore the parareal Dirichlet-Neumann waveform relaxation algorithm for the model problem (1) consists of the following steps: Given an initial guess $\omega_n^0(t)$ along the interface $\Gamma = \{x=0\} \times (T_n, T_{n+1})$, and an initial guess $U_{i,n}^0(x,t)$, and for $k = 0,1,2,\ldots$, Step I: use the more accurate evolution operator from (2) and (3) to calculate

$$\begin{aligned} u_{1,n}^{k+1}(x,t) &:= F_{1,n}(U_{1,n}^k(x), \omega_n^k(t)), \\ u_{2,n}^{k+1}(x,t) &:= F_{2,n}(U_{2,n}^k(x), u_{1,n}^{k+1}(x,t)); \end{aligned}$$

Step II: update interface information

$$\omega_n^{k+1}(t) = \theta u_{2,n}^{k+1}(0,t) + (1-\theta)\omega_n^k(t);$$

Step III: update new initial conditions using a parareal step both in space and time for $n = 0,1,\ldots,N-1$ by

$$\begin{aligned} U_{1,n+1}^{k+1} &= u_{1,n}^{k+1}(\cdot, T_{n+1}) + G_{1,n}(U_{1,n}^{k+1}(x), \omega_n^{k+1}(t)) - G_{1,n}(U_{1,n}^k(x), \omega_n^k(t)), \\ U_{2,n+1}^{k+1} &= u_{2,n}^{k+1}(\cdot, T_{n+1}) + G_{2,n}(U_{2,n}^{k+1}(x), U_{1,n+1}^{k+1}(x,t)) - G_{2,n}(U_{2,n}^k(x), U_{1,n+1}^k(x,t)). \end{aligned} \tag{4}$$

Next we will introduce the parareal Neumann-Neumann waveform relaxation algorithm. Similar, we first introduce two propagators $FD_{i,n}(U(x), h(t))$ and $GD_{i,n}(U(x), h(t))$ to solve the following Dirichlet problem in $\Omega_{i,n}$

$$\begin{cases} \dfrac{\partial u_{i,n}}{\partial t} = \dfrac{\partial^2 u_{i,n}}{\partial x^2} + f(x,t), & (x,t) \in \Omega_{i,n}, \\ u_{i,n}(x,t) = g(x,t), & x \in \partial\Omega \cap \Omega_i,\ t \in (T_n, T_{n+1}), \\ u_{i,n}(0,t) = h(t), & t \in (T_n, T_{n+1}), \\ u_{i,n}(x,T_n) = U(x), & x \in \Omega_i, \end{cases} \tag{5}$$

and two propagators $FN_{i,n}(u_{1,n}(x,t), u_{2,n}(x,t))$ and $GN_{i,n}(u_{1,n}(x,t), u_{2,n}(x,t))$, $i = 1,2$ to solve the following Neumann problem in $\Omega_{i,n}$

$$\begin{cases} \dfrac{\partial \Psi_{i,n}}{\partial t} = \dfrac{\partial^2 \Psi_{i,n}}{\partial x^2}, & (x,t) \in \Omega_{i,n}, \\ \Psi_{i,n}(x,t) = 0, & x \in \partial\Omega \cap \Omega_i,\ t \in (T_n, T_{n+1}), \\ \partial_{n_i} \Psi_{i,n}(0,t) = \displaystyle\sum_j \partial_{n_j} u_{j,n}(0,t), & x \in \Gamma,\ t \in (T_n, T_{n+1}), \\ \Psi_{i,n}(x,T_n) = 0, & x \in \Omega_i, \end{cases} \tag{6}$$

using an accurate approximation and a rough approximation.

Therefore the parareal Neumann-Neumann waveform relaxation algorithm for the model problem (1) consists of the following steps: Given an initial guess $h_n^0(t)$ along the interface $\Gamma = \{x = 0\} \times (T_n, T_{n+1})$, and an initial guess $U_{i,n}^0(x,t)$, and for $k = 0, 1, 2, \ldots$, Step I: use the more accurate evolution operator from (5) to calculate the Dirichlet problem

$$u_{i,n}^{k+1}(x,t) := FD_{i,n}(U_{i,n}^k(x), h_n^k(t)), i = 1,2;$$

Step II: use the more accurate evolution operator from (6) to calculate the Neumann problem

$$\Psi_{i,n}^{k+1}(x,t) := FN_{i,n}(u_{1,n}^{k+1}(x,t), u_{2,n}^{k+1}(x,t))), i = 1,2;$$

Step III: update interface information

$$h_n^{k+1}(t) = h_n^k(t) - \theta(\Psi_{1,n}^{k+1}(0,t) + \Psi_{2,n}^{k+1}(0,t));$$

Step IV: update the new initial conditions using a parareal step both in space and time for $n = 0, 1, \ldots, N-1$ by

$$\begin{aligned} U_{1,n+1}^{k+1} &= u_{1,n}^{k+1}(\cdot, T_{n+1}) + GD_{1,n}(U_{1,n}^{k+1}(x), h_n^{k+1}(t)) - GD_{1,n}(U_{1,n}^k(x), h_n^k(t)), \\ U_{2,n+1}^{k+1} &= u_{2,n}^{k+1}(\cdot, T_{n+1}) + GD_{2,n}(U_{2,n}^{k+1}(x), h_n^{k+1}(t)) - GD_{2,n}(U_{2,n}^k(x), h_n^k(t)). \end{aligned} \tag{7}$$

Different from regular DNWR/NNWR and using parareal to solve the subproblems, our new methods are in parallel both in space and time, and there is no order between DNWR/NNWR and parareal. Meanwhile, we don't need to using parareal to achieve the convergence for each subproblem for each DNWR/NNWR iteration.

Theorem 1 (Convergence for parareal DNWR). *Assuming that the F-propagator is an exact solver and G-propagator is chosen as backward Euler method, if $a = b$, then $\theta = 1/2$ is the optimal parameter and fixed $T > 0$, and the parareal DNWR algorithm is convergent in finite steps; if $a \neq b$, for $\theta = 1/2$ and fixed $T > 0$, the parareal DNWR algorithm is convergent.*

Theorem 2 (Convergence for parareal NNWR). *Assuming that the F-propagator is an exact solver and G-propagator is chosen as backward Euler method, if $a = b$, then $\theta = 1/4$ is the optimal parameter and fixed $T > 0$, and the parareal DNWR algorithm is convergent in finite steps; if $a \neq b$, for $\theta = 1/4$ and and fixed $T > 0$, the parareal DNWR algorithm is convergent*

Proof. The first parts of both theorems can be directly obtained by the convergence results of parareal in [9], and DNWR and NNWR in [6]; and the proves of the second parts are technical and will in [14], a detailed numerical study of how the algorithm depends on the various parameters in Section 3.

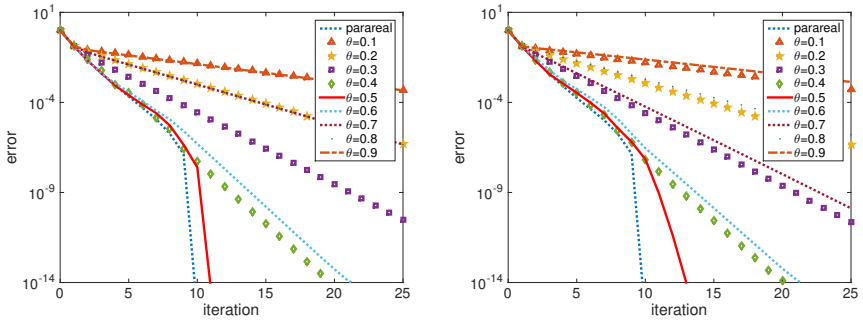


Fig. 2 Convergence of parareal DNWR for various values of the parameter θ with $T = 2$ and $\Delta T = 1/5$ for $a = b = 3$ on the left and $a = 2$, $b = 3$ on the right.

3 Numerical experiments

The numerical experiments in this section were performed for the model problem (1) on the domain $(-a,b) \times (0,T)$ with $f = 0$, $u_0(x) = x(x+1)(x+3)(x-2)\exp(-x)$, $g(-a,t) = t$ and $g(b,t) = t\exp(t)$. The diffusion problem is discretized using a centered finite differences with mesh size $h = \Delta x = 2 \times 10^{-2}$ in space and backward Euler with $\Delta t = 4 \times 10^{-3}$ in time. The domain is decomposed into the space-time subdomains $\Omega_{i,n}$ as described in Section 2. We test the algorithms by choosing $h_n^0(t) = t^2$, $t \in (T_n, T_{n+1})$ as an initial guess.

We first test the parareal DNWR algorithm. Figure 2 shows the convergence behavior for different values of θ with $T = 2$ and $\Delta T = 1/5$ for the case $a = b = 3$ on the left, and for the case $a = 2$, $b = 3$ on the right. Note that $\theta = 1/2$ is the best parameter in both cases as sated in Theorem 1, and the performance of the parareal DNWR algorithm is similar when compared to the parareal algorithm, especially when chose the parameter $\theta = 1/2$. Then we show the convergence behavior for the best parameters $\theta = 1/2$ for different numbers of the time subintervals N with $T = 2$ for both cases in Figure 3, and for different time window length T with $\Delta T = 1/5$ in Figure 4. We observe that the convergence of the parareal DNWR slows down when the number of time intervals N is increased and time interval T is increased, which is similar to the performance of the parareal algorithm; see [9].

For the parareal DNWR algorithm, Figure 5 shows the convergence behavior for different values of θ with $T = 2$ and $\Delta T = 1/5$ for the case $a = b = 3$ on the left, and for the case $a = 2$, $b = 3$ on the right. Note that $\theta = 1/4$ is the best parameter in both cases. Then we show the convergence behavior for the best parameters $\theta = 1/4$ for different numbers of the time subintervals N with $T = 2$ for both cases in Figure 6, and for different time window length T with $\Delta T = 1/5$ in Figure 7. We observe that parareal NNWR also has the similar perfomance as that of the parareal algorithm and parareal DNWR. However, compared to parareal DNWR, the parareal NNWR needs almost double numbers of iterations to achieve convergence in the same cases.

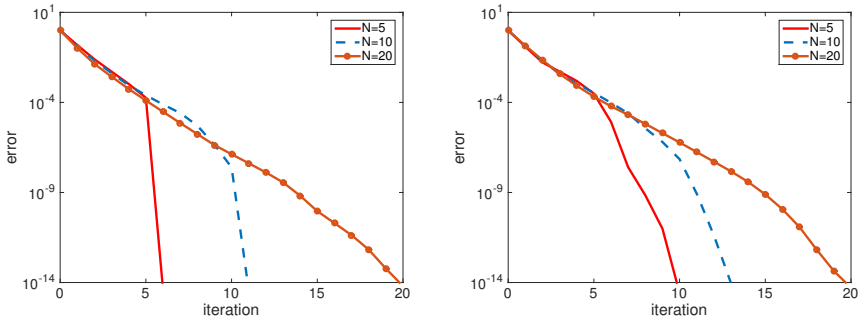


Fig. 3 Convergence of parareal DNWR for various values of the number of time subintervals N with $T = 2$ and $\theta = 1/2$ for $a = b = 3$ on the left and $a = 2$, $b = 3$ on the right.

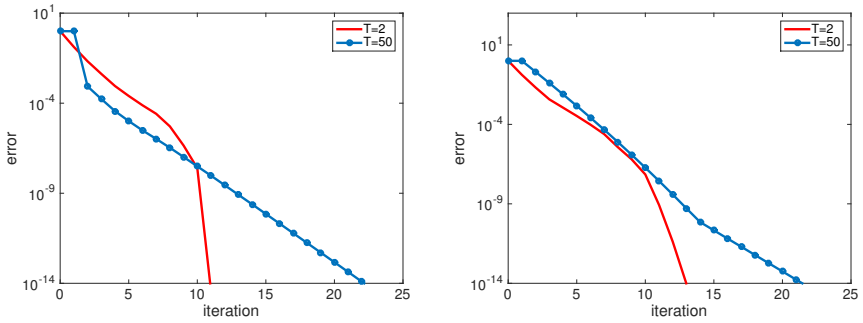


Fig. 4 Convergence of parareal DNWR for various values of the time window length T with $\Delta T = 1/5$ and $\theta = 1/2$ for $a = b = 3$ on the left and $a = 2$, $b = 3$ on the right.

4 Conclusions

We introduced the parareal DNWR and parareal NNWR algorithms for the heat equation, and provide their convergence properties for the two subdomain decomposition in one spatial dimension case. We showed that the convergence can be achieved in a finite number of iterations when choosing a proper relaxation parameter as chose for the DNWR and NNWR algorithms. Numerical results illustrate our analysis, which also indicate that the performance of parareal DNWR is better than that of parareal NNWR. We will further find the possible way to improve the performance parareal NNWR.

References

1. Gander, M.J., Hairer, E.: Nonlinear convergence analysis for the parareal algorithm. In: Domain decomposition methods in science and engineering XVII, pp. 45–56. Springer (2008)

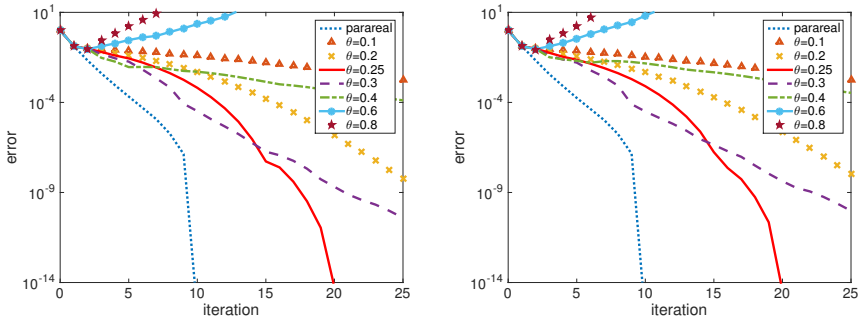


Fig. 5 Convergence of parareal NNWR for various values of the parameter θ with $T = 2$ and $\Delta T = 1/5$ for $a = b = 3$ on the left and $a = 2$, $b = 3$ on the right.

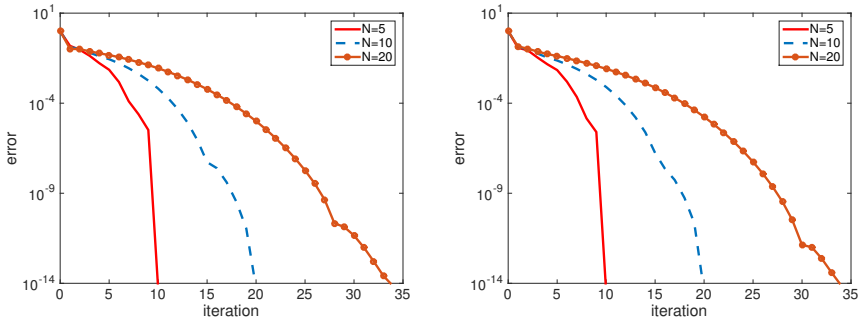


Fig. 6 Convergence of parareal NNWR for various values of the number of time subintervals N with $T = 2$ and $\theta = 1/4$ for $a = b = 3$ on the left and $a = 2$, $b = 3$ on the right.

2. Gander, M.J., Halpern, L.: Optimized Schwarz waveform relaxation methods for advection reaction diffusion problems. SIAM Journal on Numerical Analysis **45**(2), 666–697 (2007)
3. Gander, M.J., Halpern, L., Nataf, F.: Optimal Schwarz waveform relaxation for the one dimensional wave equation. SIAM Journal on Numerical Analysis **41**(5), 1643–1681 (2003)
4. Gander, M.J., Jiang, Y.L., Li, R.J.: Parareal Schwarz waveform relaxation methods. In: Domain Decomposition Methods in Science and Engineering XX, pp. 451–458. Springer (2013)
5. Gander, M.J., Jiang, Y.L., Song, B., Zhang, H.: Analysis of two parareal algorithms for time-periodic problems. SIAM Journal on Scientific Computing **35**(5), A2393–A2415 (2013)
6. Gander, M.J., Kwok, F., Mandal, B.C.: Dirichlet-Neumann and Neumann-Neumann waveform relaxation algorithms for parabolic problems. Electronic Transactions on Numerical Analysis **45**, 424–456 (2016)
7. Gander, M.J., Kwok, F., Mandal, B.C.: Dirichlet-Neumann and Neumann-Neumann waveform relaxation for the wave equation. In: Domain decomposition methods in science and engineering XXII, pp. 501–509. Springer (2016)
8. Gander, M.J., Stuart, A.M.: Space-time continuous analysis of waveform relaxation for the heat equation. SIAM Journal on Scientific Computing **19**(6), 2014–2031 (1998)
9. Gander, M.J., Vandewalle, S.: Analysis of the parareal time-parallel time-integration method. SIAM Journal on Scientific Computing **29**(2), 556–578 (2007)
10. Giladi, E., Keller, H.B.: Space-time domain decomposition for parabolic problems. Numerische Mathematik **93**(2), 279–313 (2002)

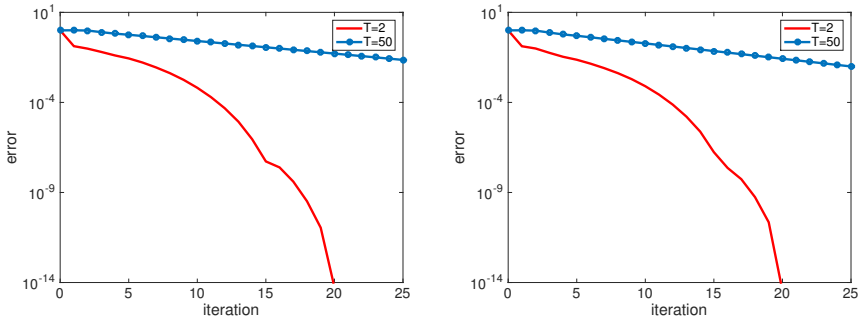


Fig. 7 Convergence of parareal NNWR for various values of the time window length T with $\Delta T = 1/5$ and $\theta = 1/4$ for $a = b = 3$ on the left and $a = 2$, $b = 3$ on the right.

11. Jiang, Y.L.: A general approach to waveform relaxation solutions of nonlinear differential-algebraic equations: the continuous-time and discrete-time cases. IEEE Transactions on Circuits and Systems I: Regular Papers **51**(9), 1770–1780 (2004)
12. Jiang, Y.L.: Waveform Relaxation Methods. Scientific Press, Beijing (2010)
13. Jiang, Y.L., Ding, X.L.: Waveform relaxation methods for fractional differential equations with the caputo derivatives. Journal of Computational and Applied Mathematics **238**, 51–67 (2013)
14. Jiang, Y.L., Song, B.: Parareal substructuring waveform relaxation methods for parabolic problems. in preparation (2017)
15. Lelarasmee, E., Ruehli, A.E., Sangiovanni-Vincentelli, A.L.: The waveform relaxation method for time-domain analysis of large scale integrated circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **1**(3), 131–145 (1982)
16. Lions, J.L., Maday, Y., Turinici, G.: A parareal in time discretization of PDEs. Comptes Rendus de l'Académie des Sciences-Series I-Mathematics **332**(7), 661–668 (2001)
17. Liu, J., Jiang, Y.L.: Waveform relaxation for reaction-diffusion equations. Journal of computational and applied mathematics **235**(17), 5040–5055 (2011)
18. Liu, J., Jiang, Y.L.: A parareal waveform relaxation algorithm for semi-linear parabolic partial differential equations. Journal of Computational and Applied Mathematics **236**(17), 4245–4263 (2012)
19. Lubich, C., Ostermann, A.: Multi-grid dynamic iteration for parabolic equations. BIT Numerical Mathematics **27**(2), 216–234 (1987)
20. Maday, Y., Turinici, G.: The parareal in time iterative solver: a further direction to parallel implementation. In: Domain decomposition methods in science and engineering, pp. 441–448. Springer (2005)
21. Vandewalle, S., Van de Velde, E.: Space-time concurrent multigrid waveform relaxation. Annals of Numer. Math **1**, 347–363 (1994)
22. Wu, S.L., Al-Khaleel, M.: Convergence analysis of the Neumann-Neumann waveform relaxation method for time-fractional RC circuits. Simulation Modelling Practice and Theory **64**, 43–56 (2016)

Preconditioning of Iterative Eigenvalue Problem Solvers in Adaptive FETI-DP

Axel Klawonn[1], Martin Kühn[1], and Oliver Rheinbach[2]

Abstract Adaptive FETI-DP and BDDC methods are robust methods that can be used for highly heterogeneous problems when standard approaches fail. In these approaches, local generalized eigenvalue problems are solved approximately, and the eigenvectors are used to enhance the coarse problem. Here, a few iterations of an approximate eigensolver are usually sufficient. Different preconditioning options for the iterative LOBPCG eigenvalue problem solver are considered. Numerical results are presented for linear elasticity problems with heterogeneous coefficients.

1 Introduction

Adaptive coarse spaces for FETI-DP or BDDC methods make use of locally computed (approximate) eigenvectors to enhance the coarse problem for faster Krylov convergence; for different approaches to domain decomposition methods with adaptive coarse spaces, see, e.g., [13, 5, 3, 17, 10, 6, 2, 14, 1, 15]. Of course, the solution of the corresponding local generalized eigenvalue problems in all these approaches adds a certain computational overhead to the setup of the method which then needs to be amortized in the iteration phase. It has been observed that an approximation of the eigenvectors already yields good convergence behavior; see [7]. In this paper, we consider different types of preconditioners for the iterative eigensolvers to obtain good approximate eigenvectors in a few steps.

We will give numerical results for the adaptive method of [8] for the equations of linear elasticity on a bounded polyhedral domain $\Omega \subset \mathbb{R}^3$, i.e., we search for $u \in \{v \in H^1(\Omega)^d : v = 0 \text{ on } \partial\Omega_D\}$ such that

[1]Mathematisches Institut, Universität zu Köln, Weyertal 86-90, 50931 Köln, Germany.
e-mail:{axel.klawonn,martin.kuehn}@uni-koeln.de [2]Institut für Numerische Mathematik und Optimierung, Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg, Akademiestr. 6, 09596 Freiberg, Germany. e-mail: oliver.rheinbach@math.tu-freiberg.de

$$\int_\Omega 2\mu\varepsilon(u):\varepsilon(v)\mathrm{d}x+\int_\Omega \lambda\,\mathrm{div}(u)\mathrm{div}(v)\mathrm{d}x=\int_\Omega f\cdot v\mathrm{d}x+\int_{\partial\Omega_N} g\cdot v\mathrm{d}s. \qquad (1)$$

Here, $\partial\Omega_D\subset\partial\Omega$ is a subset with positive surface measure where Dirichlet boundary conditions are prescribed. Moreover, $\partial\Omega_N:=\partial\Omega\setminus\partial\Omega_D$ is the Neumann boundary, and λ,μ are the Lamé constants.

2 FETI-DP with a Generalized Transformation of Basis

For an introduction of FETI-DP; see, e.g., [4, 18]. Given a polyhedral domain $\Omega\subset\mathbb{R}^3$, we subdivide Ω into N nonoverlapping subdomains $\Omega_1,\dots,\Omega_N$ such that $\overline{\Omega}=\bigcup_{i=1}^N\overline{\Omega}_i$. The FETI-DP system is given by $F\lambda=d$, where

$$F=B_BK_{BB}^{-1}B_B^T+B_BK_{BB}^{-1}\widetilde{K}_{\Pi B}^T\widetilde{S}_{\Pi\Pi}^{-1}\widetilde{K}_{\Pi B}K_{BB}^{-1}B_B^T=B_\Gamma\widetilde{S}^{-1}B_\Gamma^T.$$

Here, $\widetilde{S}_{\Pi\Pi}$ constitutes the a priori coarse space where all vertex variables are chosen to be primal.

We then use the generalized transformation-of-basis approach, as presented in [9] and applied to the adaptive context in [8], to enforce additional, adaptively computed constraints, which we also denote as a posteriori constraints. The idea of the transformation-of-basis approach is to make a constraint vector c corresponding to a (generally local) constraint on the displacements u, i.e., $c^Tu=0$ an explicit basis vector and enforce the constraint by partial subassembly at the degree of freedom where the new basis vector is introduced. Given these (orthogonal) transformations $T^{(i)}$, $i=1,\dots,N$, we therefore solve systems with transformed stiffness matrices $\overline{K}^{(i)}=T^{(i)T}K^{(i)}T^{(i)}$, transformed displacements $\overline{u}^{(i)}=T^{(i)T}u^{(i)}$, and transformed right hand sides $\overline{f}^{(i)}=T^{(i)T}f^{(i)}$, $i=1,\dots,N$. In the standard approach, constraints in the jump operator B corresponding to these a posteriori primal constraints are removed. In the generalized approach, we do not remove these rows but assemble the a posteriori primal variables and directly redistribute the continuous values subsequently to all connected subdomains. That means, in contrast to the standard transformation-of-basis approach, we also allow for scalings of a posteriori primal variables, e.g., obtained from the adaptive approach in the next section. For more details, see [9, 8].

3 Adaptive FETI-DP with a Generalized Transformation of Basis

3.1 Generalized Local Eigenvalue Problems and Constraints for a Transformation of Basis

We now present briefly the adaptive approach introduced in [7, 8]. Given a domain decomposition $\overline{\Omega}=\bigcup_{i=1}^N\overline{\Omega}_i$, we define as an *edge* $\mathscr{E}^{il}$ the interior of $\partial\Omega_i\cap\partial\Omega_j\cap$

$\partial\Omega_l$, i.e., excluding the end points, and as a face $\mathscr{F}^{ij}$ the interior of $\partial\Omega_i \cap \partial\Omega_j$. Subsequently, we will use the index $s \in \{j,l\}$ to describe eigenvalue problems and their operators defined on faces ($s = j$) and edges ($s = l$), respectively. Let us note that eigenvalue problems on faces are defined on the closure of the face.

Let $\mathscr{Z}$ be either a face or an edge shared by two subdomains Ω_i and Ω_s. We introduce $B_{\overline{\mathscr{Z}}^{is}} = [B^{(i)}_{\overline{\mathscr{Z}}^{is}} B^{(s)}_{\overline{\mathscr{Z}}^{is}}]$ consisting of all the rows of $[B^{(i)}B^{(s)}]$ that contain exactly one $+1$ and one -1. Analogously, we introduce the scaled jump operator $B_{D,\overline{\mathscr{Z}}^{is}} = [B^{(i)}_{D,\overline{\mathscr{Z}}^{is}} B^{(s)}_{D,\overline{\mathscr{Z}}^{is}}]$ as the submatrix of $[B^{(i)}_D B^{(s)}_D]$. We need the local operators $S_{is} := \text{blockdiag}(S^{(i)}, S^{(s)})$ and $P_{D,\overline{\mathscr{Z}}^{is}} := B^T_{D,\overline{\mathscr{Z}}^{is}} B_{\overline{\mathscr{Z}}^{is}}$.

We now want to solve generalized eigenvalue problems on a subspace where S_{is} is positive definite since S_{is} is in general only semidefinite. We therefore study the problem of finding $w^k_{is} \in (\ker S_{is})^\perp$ with $\mu^k_{is} \geq \text{TOL}$, such that

$$s_{is}(P_{D,\overline{\mathscr{Z}}^{is}} v_{is}, P_{D,\overline{\mathscr{Z}}^{is}} w^k_{is}) = \mu^k_{is} s_{is}(v_{is}, w^k_{is}) \quad \forall v_{is} \in (\ker S_{is})^\perp. \tag{2}$$

There, $s_{is}(\cdot,\cdot) := (\cdot, S_{is}\cdot)$ for $u_{is} \times v_{is}$ with $u_{is}, v_{is} \in W_i \times W_s$ and W_i, W_s are the local finite element spaces on Ω_i and Ω_s. In practice, this is achieved by implementing projections Π_{is} and $\overline{\Pi}_{is}$ and making the computation numerically stable; cf. [13].

The constraint vectors $q^k_{is} := P^T_{D,\overline{\mathscr{Z}}^{is}} S_{is} P_{D,\overline{\mathscr{Z}}^{is}} w^k_{is}$ computed from the eigenvalue problems are either defined on edges or on closed faces. The constraints on closed faces are then split into (additional) edge constraints and constraints on the open face. This also enables an edge by edge and face by face orthogonalization.

In our approach, an edge constraint resulting from the eigenvalue problem of two subdomains sharing this edge will always be enforced for all subdomains sharing this edge. This does not increase the size of the coarse problem.

All the adaptive constraints are stored in an (orthogonalized) transformation matrix T which is block diagonal with respect to the subdomains and with respect to blocks corresponding to the faces and edges. The operator R^T performs the finite element assembly in the a posteriori primal variables, i.e., in all degrees of freedom which belong to an adaptively computed new basis vector. The transposed operator R then redistributes the values to the individual subdomains. We define the operator $R^T_\mu := (R^T R)^{-1} R^T$. For more details, see [9, 8].

In contrast to the standard transformation-of-basis approach, we use the same jump operator B as in the original FETI-DP master system. As a result, as in deflation, the preconditioned system has at least one zero eigenvalue for each adaptively computed constraint, i.e., for the a posteriori constraints.

The adaptive FETI-DP system using a generalized transformation of basis writes

$$\begin{aligned} \widehat{M}_T^{-1}\widehat{F}\lambda &:= (\widehat{B}_D \widehat{\widetilde{S}} \widehat{B}_D^T)\,(\widehat{\widetilde{S}}^{-1} \widehat{B}^T)\lambda \\ &:= (B_D T R_\mu (R^T T^T \widetilde{S} T R) R_\mu^T T^T B_D^T)(B T R (R^T T^T \widetilde{S} T R)^{-1} R^T T^T B^T)\lambda = d, \end{aligned} \tag{3}$$

where $\widehat{F}$ is the transformed FETI-DP operator and $\widehat{M}_T^{-1}$ is the transformed Dirichlet preconditioner. For this system, we now give, without proof, the condition number bound. For more details, see [8].

Theorem 1. *Let $N_{\mathscr{F}}$ denote the maximum number of faces of a subdomain, $N_{\mathscr{E}}$ the maximum number of edges of a subdomain, $M_{\mathscr{E}}$ the maximum multiplicity of an edge and* TOL *a given tolerance for solving the local generalized eigenvalue problems. If all vertices are chosen to be primal, the condition number $\kappa(\widehat{M}_T^{-1}\widehat{F})$ of the FETI-DP algorithm with adaptive constraints enforced by the generalized transformation-of-basis approach satisfies*

$$\kappa(\widehat{M}_T^{-1}\widehat{F}) \leq 4\max\{N_{\mathscr{F}}, N_{\mathscr{E}}M_{\mathscr{E}}\}^2\text{TOL}.$$

3.2 Solving the Local Generalized Eigenvalue Problems

Adaptive methods are most suitable for hard problems that are not solvable by standard techniques, e.g., as a result of strong heterogeneities present in the problem. However, as a result of these heterogeneities the local generalized eigenvalue problems can also be expected to be ill-conditioned, and unpreconditioned iterative eigensolvers may also struggle; see, e.g., [16]. As in [16], we use the iterative LOBPCG eigenvalue problem solver; see [12]. In practice, when using two projections Π_{is} and $\overline{\Pi}_{is}$ to remove the rigid body modes from S_{is}, the right hand side of the eigenvalue problems writes

$$\overline{\Pi}_{is}(\Pi_{is}S_{is}\Pi_{is} + \sigma_{is}(I-\Pi_{is}))\overline{\Pi}_{is} + \sigma_{is}(I-\overline{\Pi}_{is}) \tag{4}$$

where σ_{is} is chosen as $\sigma_{is} = \max(\text{diag}(S_{is}))$. The projection $I-\overline{\Pi}_{is}$ consists of the sum of several rank one matrices, and we usually avoid to building the matrix explicitly. The operator $\Pi_{is}S_{is}\Pi_{is} + \sigma_{is}(I-\Pi_{is})$ can be built cheaply by only scaling a few rows and columns of the Schur complements and adding some constants; see Figure 1 for the nonzero pattern of S_{is} and $\Pi_{is}S_{is}\Pi_{is} + \sigma_{is}(I-\Pi_{is})$.

We test five different preconditioners for the iterative eigensolver. First, we take a Cholesky decomposition of the fully assembled right hand side (4) as the (expensive) base line to compare against. We also test an LU and ILU(0) decomposition of $\Pi_{is}S_{is}\Pi_{is} + \sigma_{is}(I-\Pi_{is})$ and use the projection $\overline{\Pi}_{is}$ to remove the corresponding kernel from the preconditioner, i.e., we, e.g., use

$$\overline{\Pi}_{is}\mathbf{LU}\Big(\Pi_{is}S_{is}\Pi_{is} + \sigma_{is}(I-\Pi_{is})\Big)\overline{\Pi}_{is},$$

where $\mathbf{LU}(\cdot)$ denotes the computation of the LU decomposition of the argument. Finally, we also test two different local lumped versions, i.e., an LU and a ILU(0) decomposition of $K_{\Gamma\Gamma,is} = \text{blockdiag}(K_{\Gamma\Gamma}^{(i)}, K_{\Gamma\Gamma}^{(s)})$, so for the LU decomposition, we implement the preconditioner

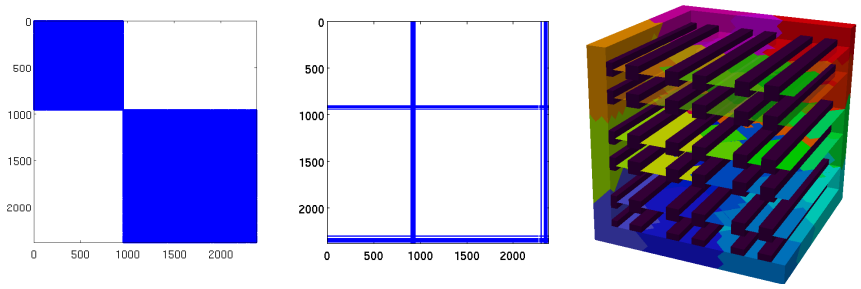


Fig. 1 Representative nonzero pattern of the matrices S_{is} (left) and $S_{is} - \left[\Pi_{is} S_{is} \Pi_{is} + \sigma_{is}(I - \Pi_{is})\right]$ (center) for two randomly chosen subdomains Ω_i, Ω_s. Composite material with irregular decomposition (right; visualization for $N = 27$ and $1/h = 10N^{1/3}$). In the right picture, large coefficients $\mathrm{E}_2 = 1e+06$ are shown in dark purple in the picture and low coefficients are not shown; subdomains are shown in different colors in the background and by half-transparent slices.

$$\overline{\Pi}_{is} \Pi_{is} \mathbf{LU}\Big(K_{\Gamma\Gamma,is}\Big) \Pi_{is} \overline{\Pi}_{is}.$$

3.3 Heuristic Modifications

As in [7], we will introduce two heuristic variants (denoted *Algorithm Ib* and *Ic*). The original algorithm is denoted *Algorithm Ia*.

Algorithm Ib: Reducing the number of edge eigenvalue problems We discard edge eigenvalue problems for edges that do not have high coefficient jumps in their neighborhood of one finite element.

Algorithm Ic: Reducing the number of edge constraints In addition, we also discard all edge constraints from face eigenvalue problems if there are no coefficient jumps in the neighborhood of the edge.

The condition number bound derived for *Algorithm Ia* will, in general, not hold for the two variants, however, it is likely that a modified theory, using slab techniques as in [10], can be derived for *Algorithm Ib*.

4 Numerical Results

We present numerical results for Algorithms Ia, Ib, and Ic. We have a soft matrix material with $\mathrm{E}_1 = 1$ with $4N^{2/3}$ stiff beams with $\mathrm{E}_2 = 1e+06$; see Fig. 1. We consider $\Omega = [0,1]^3$ with Dirichlet boundary conditions for the face with $x = 0$ and zero Neumann boundary conditions elsewhere; we have $f = [0.1, 0.1, 0.1]^T$ and $\mathrm{E}(x) \in \{1, 1e+6\}$. For the domain decomposition, the METIS graph partitioner

with options `-ncommon=3` and `-contig` is used. Each local eigenvalue problem is solved using LOBPCG with a block size 10, a given number of maximum iterations from $\{5, 25, 100\}$, and a preconditioner; see Section 3.2. Our a priori coarse space uses at least three primal vertices on each edge in order to remove local hinge modes; see [13, 7]. We also set edge nodes primal that belong to ~~an~~ single noded edges. The corresponding edge eigenvalue problem becomes superfluous. We assume the Young modulus $\mathrm{E}(x)$ to be constant on each finite element, and we use ρ-scaling in the form of patch-ρ-scaling. The coefficient $(\mathrm{E}(x^*))$ at a node x^* will be set as the maximum coefficient on the support of the corresponding nodal basis function φ_{x^*}; cf. [11]. In the tables, "κ" denotes the condition number of the adaptively preconditioned FETI-DP operator, "*its*" the number pcg iterations, "$|\Pi'|$" the size of the initial vertex coarse space and "$|\Pi|$" the size of the corresponding a posteriori coarse space; the number of subdomains is "N". The pcg algorithm is stopped after a relative reduction of the starting residual by 10^{-10} or when 500 iterations are reached.

5 Conclusion

We have presented results for different preconditioniers of the local generalized eigenvalue problems. Obviously, the most expensive algorithm, the Cholesky decomposition of the assembled right hand side of the eigenvalue problem yields the best results with respect to the condition numbers and the iteration counts of the FETI-DP algorithm. In this case, only a few iterations (e.g., 1-5) of the LOBPCG solver are sufficient; cf. also our results in [7, 8]. However, an LU or ILU(0)-factorization of $\Pi_{is} S_{is} \Pi_{is} + \sigma_{is}(I - \Pi_{is})$ with a few more iterations can suffice. To choose an LU or ILU decomposition of $\Pi_{is} S_{is} \Pi_{is} + \sigma_{is}(I - \Pi_{is})$ is a reasonable choice since this matrix can be built easily but just manipulating a few rows and columns of S_{is}; see Figure 1. Note that the slight differences in the condition numbers and iteration counts result from a small difference in the coarse space size. The results for the lumped preconditioner, an LU or ILU decomposition of $K_{\Gamma\Gamma,is}$ are given for completeness and to show that the results were not as satisfactory as expected. Eventually, note from [8] that also too many iterations (e.g., 200) of the local solver might not be helpful if the local scheme diverges without notice. A heuristic strategy for an (almost) optimal a priori choice of the maximum LOBPCG iteration number is still under development.

References

1. Juan G. Calvo and Olof B. Widlund. An adaptive choice of primal constraints for BDDC domain decomposition algorithms. Electronic Transactions on Numerical Analysis, 45:524–544, 2016.

Local Preconditioner: Chol$\left(\overline{\Pi}_{is}(\Pi_{is}S_{is}\Pi_{is}+\sigma_{is}(I-\Pi_{is}))\overline{\Pi}_{is}+\sigma_{is}(I-\overline{\Pi}_{is})\right)$.											
			Algorithm Ia			*Algorithm Ib*			*Algorithm Ic*		
N	$\lvert\Pi'\rvert$	LOBPCG max its	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$
3^3	168	5	3.35	16	1905	3.35	16	1905	3.53	19	594
		25	8.89	18	2025	8.89	18	2025	9.12	21	684
		100	10.59	18	2013	10.59	18	2013	10.78	21	672
4^3	351	5	3.34	16	5259	3.34	16	5259	3.56	19	1674
		25	14.95	24	5535	14.95	24	5535	15.33	25	1869
		100	5.07	18	5496	5.07	18	5496	5.08	21	1848
Local Preconditioner: $\overline{\Pi}_{is}$**LU**$\left(\Pi_{is}S_{is}\Pi_{is}+\sigma_{is}(I-\Pi_{is})\right)\overline{\Pi}_{is}$.											
			Algorithm Ia			*Algorithm Ib*			*Algorithm Ic*		
N	$\lvert\Pi'\rvert$	LOBPCG max its	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$
3^3	168	5	110.84	38	1872	110.84	38	1872	163.73	43	603
		25	3.84	18	1926	3.84	18	1926	3.84	20	660
		100	3.84	18	1938	3.84	18	1938	3.85	21	666
4^3	351	5	471.97	62	5074	471.97	62	5074	521.66	67	1647
		25	54.34	30	5259	54.34	30	5259	90.89	33	1830
		100	56.50	30	5328	56.50	30	5328	99.32	32	1884
Local Preconditioner: $\overline{\Pi}_{is}$**ILU(0)**$\left(\Pi_{is}S_{is}\Pi_{is}+\sigma_{is}(I-\Pi_{is})\right)\overline{\Pi}_{is}$.											
			Algorithm Ia			*Algorithm Ib*			*Algorithm Ic*		
N	$\lvert\Pi'\rvert$	LOBPCG max its	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$
3^3	168	5	5.36	17	2088	5.36	17	2088	5.45	21	711
		25	3.82	20	1995	3.82	20	1995	3.84	21	678
		100	3.35	17	1998	3.35	17	1998	3.52	20	675
4^3	351	5	24.35	26	6225	24.35	26	6225	26.50	30	2394
		25	3.82	20	5964	3.82	20	5964	3.83	22	2277
		100	4.37	20	5850	4.37	20	5850	4.42	22	2181
Local Preconditioner: $\overline{\Pi}_{is}\Pi_{is}$**LU**$\left(K_{\Gamma\Gamma,is}\right)\Pi_{is}\overline{\Pi}_{is}$.											
			Algorithm Ia			*Algorithm Ib*			*Algorithm Ic*		
N	$\lvert\Pi'\rvert$	LOBPCG max its	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$
3^3	168	5	1.81e+06	500	0	1.81e+06	500	0	1.81e+06	500	0
		25	3.83e+04	500	441	3.83e+04	500	441	1.56e+05	500	102
		100	452.95	126	442	452.95	126	442	468.46	129	81
4^3	351	5	1.06e+06	500	0	1.06e+06	500	0	1.06e+06	500	0
		25	5.97e+04	500	1254	5.97e+04	500	1254	1.72e+05	500	273
		100	677.56	181	936	677.56	181	936	685.30	183	213
Local Preconditioner: $\overline{\Pi}_{is}\Pi_{is}$**ILU(0)**$\left(K_{\Gamma\Gamma,is}\right)\Pi_{is}\overline{\Pi}_{is}$.											
			Algorithm Ia			*Algorithm Ib*			*Algorithm Ic*		
N	$\lvert\Pi'\rvert$	LOBPCG max its	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$	κ	its	$\lvert\Pi\rvert$
3^3	168	5	1.81e+06	500	0	1.81e+06	500	0	1.81e+06	500	0
		25	3.26e+04	500	462	3.26e+04	500	462	8.40e+04	500	111
		100	197.47	108	324	197.47	108	324	200.09	110	75
4^3	351	5	1.06e+06	500	0	1.06e+06	500	0	1.06e+06	500	0
		25	4.56e+04	500	1236	4.56e+04	500	1236	8.51e+04	500	282
		100	2.54e+04	316	978	2.54e+04	316	978	6.15e+04	329	222

Table 1 Compressible linear elasticity on an irregular decomposition of $\overline{\Omega}=[0,1]^3$ with N subdomains, $1/h=10N^{1/3}$ and composite material with Young's modulus $\mathrm{E}_1=1$ and $\mathrm{E}_2=1e+06$. Coarse spaces for $\mathrm{TOL}=10$ for all generalized eigenvalue problems.

2. L. Beirao da Veiga, L. F. Pavarino, S. Scacchi, O. B. Widlund, and S. Zampini. Adaptive selection of primal constraints for isogeometric BDDC deluxe preconditioners. SIAM Journal on Scientific Computing, 39(1):A281–A302, 2017.
3. Victorita Dolean, Frédéric Nataf, Robert Scheichl, and Nicole Spillane. Analysis of a two-level Schwarz method with coarse spaces based on local Dirichlet-to-Neumann maps. Comput. Methods Appl. Math., 12(4):391–414, 2012.
4. Charbel Farhat, Michael Lesoinne, and Kendall Pierson. A scalable dual-primal domain decomposition method. Numer. Linear Algebra Appl., 7(7-8):687–714, 2000. Preconditioning techniques for large sparse matrix problems in industrial applications (Minneapolis, MN, 1999).
5. Juan Galvis and Yalchin Efendiev. Domain decomposition preconditioners for multiscale flows in high-contrast media. Multiscale Model. Simul., 8(4):1461–1483, 2010.
6. Hyea Hyun Kim and Eric T. Chung. A BDDC algorithm with enriched coarse spaces for two-dimensional elliptic problems with oscillatory and high contrast coefficients. Multiscale Model. Simul., 13(2):571–593, 2015.
7. Axel Klawonn, Martin Kühn, and Oliver Rheinbach. Adaptive coarse spaces for FETI-DP in three dimensions. SIAM J. Sci. Comput., 38(5):A2880–A2911, 2016.
8. Axel Klawonn, Martin Kühn, and Oliver Rheinbach. Adaptive FETI-DP and BDDC methods with a transformation of basis for heterogeneous problems. Technical report, Technische Universität Bergakademie Freiberg, Fakultät für Mathematik und Informatik, Preprint 2017-04, 2017. http://tu-freiberg.de/fakult1/forschung/preprints. Submitted.
9. Axel Klawonn, Martin Kühn, and Oliver Rheinbach. FETI-DP and BDDC methods with a transformation of basis for heterogeneous problems: Connections to deflation. Technical report, Technische Universität Bergakademie Freiberg, Fakultät für Mathematik und Informatik, Preprint 2017-01, 2017. http://tu-freiberg.de/fakult1/forschung/preprints. Submitted.
10. Axel Klawonn, Patrick Radtke, and Oliver Rheinbach. FETI-DP methods with an adaptive coarse space. SIAM J. Numer. Anal., 53(1):297–320, 2015.
11. Axel Klawonn and Oliver Rheinbach. Robust FETI-DP methods for heterogeneous three dimensional elasticity problems. Comput. Methods Appl. Mech. Engrg., 196(8):1400–1414, 2007.
12. Andrew V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. SIAM J. Sci. Comput., 23(2):517–541, 2001.
13. Jan Mandel and Bedřich Sousedík. Adaptive selection of face coarse degrees of freedom in the BDDC and the FETI-DP iterative substructuring methods. Comput. Methods Appl. Mech. Engrg., 196(8):1389–1399, 2007.
14. Duk-Soon Oh, Olof B. Widlund, Stefano Zampini, and Clark R. Dohrmann. BDDC algorithms with deluxe scaling and adaptive selection of primal constraints for Raviart-Thomas vector fields. Math. Comp., 2017. Published electronically June 21, 2017.
15. Clemens Pechstein and Clark R. Dohrmann. A unified framework for adaptive BDDC. Electron. Trans. Numer. Anal., 46:273–336, 2017.
16. Bedřich Sousedík. Adaptive-Multilevel BDDC. PhD thesis, University of Colorado Denver, 2010.
17. Nicole Spillane and Daniel J. Rixen. Automatic spectral coarse spaces for robust finite element tearing and interconnecting and balanced domain decomposition algorithms. Internat. J. Numer. Methods Engrg., 95(11):953–990, 2013.
18. Andrea Toselli and Olof B. Widlund. Domain Decomposition Methods - Algorithms and Theory, volume 34 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin Heidelberg New York, 2005.

Using Algebraic Multigrid in Inexact BDDC Domain Decomposition Methods

Axel Klawonn, Martin Lanser, and Oliver Rheinbach

1 Introduction

Traditionally, domain decomposition methods use sparse direct solvers as building blocks, i.e., to solve local subdomain problems and/or the coarse problem. Often, the sparse direct solvers can be replaced by spectrally equivalent preconditioners without loss of convergence speed. In FETI-DP and BDDC domain decomposition methods, such approaches have first been introduced in [9, 8, 4], and have since then successfully been used in large parallel codes [6, 1].

2 An Inexact BDDC Method

2.1 A BDDC Preconditioner for the Assembled System

Let us briefly describe the BDDC preconditioner which can directly be applied to a linear system

$$Au = b \tag{1}$$

arising from a finite element discretization of a partial differential equation on a computational domain $\Omega \subset \mathbb{R}^d$, $d = 2,3$. The variant discussed here was first introduced in [9]. Let Ω_i, $i = 1,\dots,N$, be a nonoverlapping domain decomposition of Ω such that $\overline{\Omega} = \bigcup_{i=1}^{N} \overline{\Omega}_i$. Each subdomain Ω_i is discretized using finite elements,

Axel Klawonn, Martin Lanser
Mathematisches Institut, Universität zu Köln, Weyertal 86-90, 50931 Köln, Germany, e-mail: {axel.klawonn,martin.lanser}@uni-koeln.de

Oliver Rheinbach
Institut für Numerische Mathematik und Optimierung, Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg, Akademiestr. 6, 09596 Freiberg, e-mail: oliver.rheinbach@math.tu-freiberg.de

the corresponding local finite element spaces are denoted by W_i, $i = 1,\dots,N$, and the product space is defined by $W = W_1 \times \dots \times W_N$. Let us also introduce the global finite element space V^h corresponding to the discretization of Ω and a restriction $R : V^h \to W$. We obtain local problems in the spaces W_i

$$K_i u_i = f_i,\ i = 1,\cdots,N.$$

Introducing the block operators

$$K = \begin{pmatrix} K_1 & & \\ & \ddots & \\ & & K_N \end{pmatrix},\ f = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix},$$

we can write $A := R^T K R$ and $b := R^T f$. Finally, the interface between the subdomains is $\Gamma := \bigcup_{i=1}^{N} \partial\Omega_i \setminus \partial\Omega$. Let us assume that the degrees of freedom (d.o.f.) on the Dirichlet boundary $\partial\Omega_D \subset \partial\Omega$ are eliminated.

We use the index Γ for degrees of freedom on Γ. For degrees of freedom in the interior of the subdomains and on the Neumann boundary $\partial\Omega_N \subset \partial\Omega$, we use the index I. For the construction of a BDDC preconditioner directly applicable to the assembled linear system $Au = b$, we subdivide, as usual in BDDC and FETI-DP methods, the interface Γ into primal (Π) and the remaining dual (Δ) degrees of freedom. As primal variables usually subdomain vertices or averages over edges or faces are chosen.

Let us introduce the space $\widetilde{W} \subset W$ of functions, which are continuous in all primal variables and the restriction operator $\bar{R} : \widetilde{W} \to W$. We can now define a partially assembled system matrix

$$\widetilde{K} := \bar{R}^T K \bar{R} \tag{2}$$

and the corresponding right hand side $\tilde{f} := \bar{R}^T f$. Using a scaled restriction operator $\widetilde{R}_D : V^h \to \widetilde{W}$, we define the BDDC preconditioner by

$$M^{-1}_{BDDC} := \left(\widetilde{R}_D^T - \mathscr{H} P_D\right) \widetilde{K}^{-1} \left(\widetilde{R}_D - P_D^T \mathscr{H}^T\right); \tag{3}$$

see [9]. Here, $\mathscr{H} : \widetilde{W} \to V^h$ is a discrete harmonic extension operator defined by

$$\mathscr{H} := \begin{pmatrix} 0 & -(K_{II})^{-1} \widetilde{K}_{\Gamma I}^T \\ 0 & 0 \end{pmatrix}, \tag{4}$$

where K_{II} and $\widetilde{K}_{\Gamma I}$ are blocks of the partially assembled stiffness matrix

$$\widetilde{K} = \begin{pmatrix} K_{II} & \widetilde{K}_{\Gamma I}^T \\ \widetilde{K}_{\Gamma I} & \widetilde{K}_{\Gamma\Gamma} \end{pmatrix}, \tag{5}$$

which are common to both, BDDC and FETI-DP methods. The matrix K_{II} is block-diagonal and applications of K_{II}^{-1} only require local solves on the interior parts of the subdomains and are thus easily parallelizable.

Finally, let $P_D : \widetilde{W} \to \widetilde{W}$ be a scaled jump operator defined by

$$P_D = I - E_D := I - \widetilde{R}\widetilde{R}_D^T. \tag{6}$$

In the FETI-DP literature this operator is often defined as $P_D = B_D^T B$; see [12, Chapter 6] and [9] for more details. There, B is the standard jump matrix used in FETI-type methods. Let us remark that the preconditioned system $M_{BDDC}^{-1}A$ has, except for some eigenvalues equal to 0 and 1, the same spectrum as the standard BDDC preconditioner formulated on the Schur complement; see [9, Theorem 1]. Therefore, under sufficient assumptions (see [9, Assumption 1]), the condition number of the preconditioned system is bounded by

$$\kappa(M_{BDDC}^{-1}A) \leq \Phi(H,h). \tag{7}$$

For a homogeneous linear elasticity problem, if appropriate primal constraints are chosen, we obtain the well known BDDC (and FETI-DP) condition number bound with $\Phi(H,h) = C(1+\log(H/h))^2$. Here, H always denotes the maximal diameter of all subdomains and h the minimal diameter of all finite elements.

2.2 Using Inexact Solvers and Implementation Remarks

In this paragraph, we describe the use of inexact solvers in the preconditioner M_{BDDC}^{-1} as suggested in [9] and also provide some remarks on our implementation. We assume that $\widehat{K}^{-1}$ and $\widehat{K}_{II}^{-1}$ are spectrally equivalent preconditioners for $\widetilde{K}$ and K_{II}, respectively. In this paper, we always choose a fixed number of V-cycles of an AMG method for solving problems including $\widetilde{K}^{-1}$ and K_{II}^{-1} for those preconditioners. While $\widehat{K}^{-1}$ requires an MPI parallel implementation of an AMG method, an application of $\widehat{K}_{II}^{-1}$ requires only a sequential AMG, due to the block diagonal structure of K_{II}. Using $\widehat{K}_{II}^{-1}$, we define an approximate discrete harmonic extension $\widehat{\mathscr{H}}$ by

$$\widehat{\mathscr{H}} := \begin{pmatrix} 0 & -\widehat{K}_{II}^{-1}\widetilde{K}_{\Gamma I}^T \\ 0 & 0 \end{pmatrix}. \tag{8}$$

We investigate two different variants of the inexact BDDC preconditioner in this paper, namely

$$\widehat{M}_{BDDC,1}^{-1} := \left(\widetilde{R}_D^T - \mathscr{H}P_D\right)\widehat{K}^{-1}\left(\widetilde{R}_D - P_D^T\mathscr{H}^T\right) \tag{9}$$

and

$$\widehat{M}_{BDDC,2}^{-1} := \left(\widetilde{R}_D^T - \widehat{\mathscr{H}}P_D\right)\widehat{K}^{-1}\left(\widetilde{R}_D - P_D^T\widehat{\mathscr{H}}^T\right). \tag{10}$$

Let us remark that in $M_{BDDC,1}^{-1}$ the discrete harmonic extension is applied exactly using a direct solver, while in $M_{BDDC,2}^{-1}$ the approximate discrete harmonic extension

$\widehat{\mathscr{H}}$ is used. Assuming that we have chosen an appropriate $\widehat{K}$, i.e., satisfying

$$\tilde{c}u^T\widetilde{K}u \leq u^T\widehat{K}u \leq \widetilde{C}u^T\widetilde{K}u,\ \forall u \in \widetilde{W}, \tag{11}$$

a condition number bound of the same quality as (7) is valid,

$$\kappa(M^{-1}_{BDDC,2}A) \leq \frac{\widetilde{C}C}{\tilde{c}}(1+\Phi(H,h));$$

see [9, Theorem 4]).

Our parallel implementation uses C/C++ and PETSc version 3.6.4 [3]. While the matrix $\widetilde{K}$ is an MPI parallel matrix, all other matrices are completely local to the computational cores. All restrictions and prolongations are performed using PETSc *VecScatter* and *VecGather* operations. More details on the implementation of the linear BDDC preconditioner can be found in [7], where a parallel implementation of an nonlinear inexact BDDC method is applied to hyperelasticity and elasto-plasticity problems.

2.3 The GM (Global Matrix) Interpolation

Good constants $\tilde{c},\widetilde{C}$ in equation (11) are important for fast convergence. It is well known, that for scalability of multigrid methods the preconditioner should preserve nullspace or near-nullspace vectors of the operator. This is especially important for $\widetilde{K}$. It is a bit less important for the blocks $K_{II}^{(i)}$ in K_{II}, where a large portion of the boundary has Dirichlet data. In this latter case, standard methods can also work well.

Since the AMG method should preserve the nullspace of the operator on all levels, these nullspace vectors have to be in the range of the AMG interpolation. While classical AMG guarantees this property only for constant vectors, the global matrix approach (GM), introduced in [2], allows the user to specify certain near-nullspace vectors, which are interpolated exactly from the coarsest to the finest level; details on the method and its scalability can be found in [10, 2]. Since we are interested in linear elasticity problems, we choose the rotations of the body in $\widetilde{W}$ for the exact interpolation. All translations of the body are already interpolated exactly in classical AMG approaches for systems of PDEs since they use classical interpolation applied component-by-component. We partially assemble the rotations of the subdomains Ω_i in the primal variables. In our implementation, we always use BoomerAMG from the hypre package [5], where a highly scalable implementation of the GM2 approach is integrated; see [2]. We will compare the use of the GM2 approach with a hybrid AMG approach for systems of PDEs. By hybrid AMG approaches, we refer to methods, where the coarsening is based on the physical nodes (nodal coarsening) but the interpolation is based on the unknowns. In general, a nodal coarsening approach is beneficial for the solution of systems of PDEs, and all degrees of freedom belonging to the same physical node are either all coarse or fine on a certain level.

The latter fact is also mandatory for the GM approach. Therefore, GM is based on the same nodal coarsening and can also be considered as a hybrid approach.

3 Numerical Results

As model problems, we choose linear elasticity problems in two and in three dimensions. In two dimensions, we consider a beam $\Omega = [0,8] \times [0,1]$ with a homogeneous Dirichlet boundary condition on the left; see also Fig. 1. A constant volume force is applied in y-direction and the material is chosen to be homogeneous with $E = 210$ and $\nu = 0.3$.

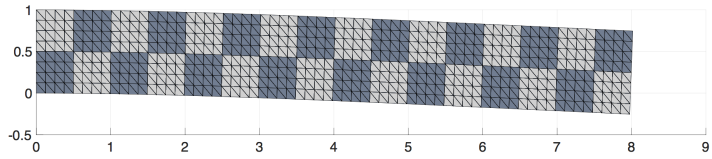


Fig. 1 Beam problem in two dimensions; exemplary decomposition in 32 subdomains depicted.

We first provide a comparison of the preconditioners $M^{-1}_{BDDC,1}$ and $M^{-1}_{BDDC,2}$ using a hybrid AMG approach or the GM2 approach for $\widehat{K}$, respectively; see Fig. 2 for the results. Let us remark that we always use the standard hybrid approach for the approximation of the discrete harmonic extension $\widehat{\mathscr{H}}$ in the case of $M^{-1}_{BDDC,2}$, since this appears to be sufficient so far; also see the remark above on the large Dirichlet boundary. We always use an *HMIS* coarsening, $extended+i$ interpolation, and a threshold of 0.375 for the detection of strong coupling. The interpolation operators of the AMG method are truncated to a maximum of P_{max} entries per row, to keep the operator complexity low and to obtain sufficient weak scalability. We always choose P_{max} such that the operator complexity of the hybrid approach and GM2 approach are similar, to provide a fair comparison. We always use preconditioned GMRES with a relative stopping criteria of 10^{-8}.

In Fig. 2, we present results for the two dimensional beam which is decomposed into 512 subdomains. We increase the problem size by increasing the subdomain size. As primal constraints, we only consider subdomain vertices. We use piecewise quadratic finite elements and thus, the smallest problem carries 882 and the largest problem 136K degrees of freedom per subdomain. We always use one MPI rank per subdomain but use two MPI ranks for each core of the JUQUEEN BlueGene/Q at Forschungszentrum Jülich, Germany, to make use of the hardware threads. Therefore, we have 500 MB of memory available for each subdomain. Using direct solvers for the discrete harmonic extension (i.e., $M^{-1}_{BDDC,1}$), we always have slightly lower GMRES iteration counts and faster runtimes compared to $M^{-1}_{BDDC,2}$, but $M^{-1}_{BDDC,2}$ is more memory efficient. The largest problem, which can be solved with $M^{-1}_{BDDC,1}$

carries 81K d.o.f. per subdomain ($H/h = 100$), while $M^{-1}_{BDDC,2}$ can handle problems twice as large, with 136K d.o.f. per subdomain ($H/h = 130$).

As expected, BDDC using the GM2 approach clearly outperforms the hybrid approach. While the iteration count grows with H/h for the hybrid approach, it stays nearly constant for the GM2 approach. For the problem with $H/h = 120$, $M^{-1}_{BDDC,2}$ with GM2 is six times faster than $M^{-1}_{BDDC,2}$ combined with the hybrid approach, and for $H/h = 130$, $M^{-1}_{BDDC,2}$ with the hybrid approach does not fit in the memory. Choosing $P_{max} = 2$ solves this problem, but the number of iterations is even higher.

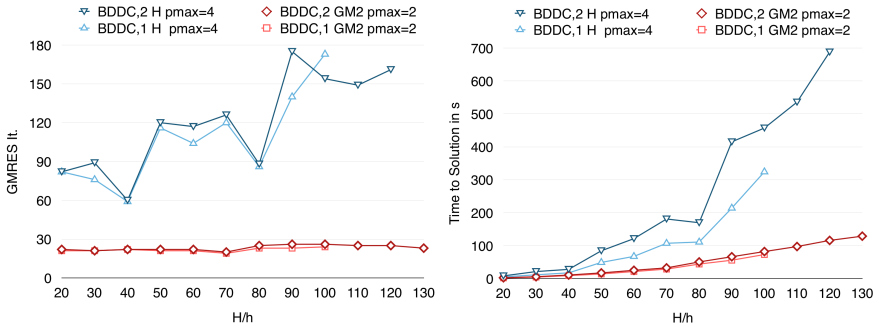


Fig. 2 Comparison for growing H/h and 512 subdomains of the different preconditioners $M^{-1}_{BDDC,1}$ using direct solvers (UMFPACK) for the discrete harmonic extension and $M^{-1}_{BDDC,2}$ using an inexact discrete harmonic extension. Both variants are equipped with hybrid AMG (marked with an H) or GM2, respectively. *pmax* denotes the truncation of the interpolation matrices. **Left:** GMRES iterations. **Right:** Time to solution. Computation performed on JUQUEEN BlueGene/Q at FZ Jülich, Gemany.

We also present a weak scaling study for the best performing combination of $M^{-1}_{BDDC,2}$ and the GM2 approach using $H/h = 80$ and $H/h = 100$; see Fig. 3. While a radical truncation of $P_{max} = 2$ works fine for up to 8 192 subdomains, $P_{max} = 4$ is necessary for the larger configurations. All in all, the parallel efficiency of 91% on 131K MPI ranks and 65K cores and a total problem size of 10 billion degrees of freedom is satisfying.

Finally, we present a weak scaling study in three dimensions. We again consider a linear elastic material and deform a heterogeneous cube. We have a single spherical stiff inclusion ($E = 21\,000$, $nu = 0.3$) in each subdomain. The remaining material is softer with $E = 210$, $nu = 0.3$. This time, we choose piecewise linear finite elements, $H/h = 20$, and, as primal constraints, we enforce continuity in all subdomain vertices and in the midpoints of all edges. We use the same AMG settings as before. In Fig. 4, we again observe a sufficient weak scaling behavior using $M^{-1}_{BDDC,2}$ with the GM2 approach, while the hybrid approach cannot deliver satisfying convergence behavior, since it cannot fulfill (11) with good bounds.

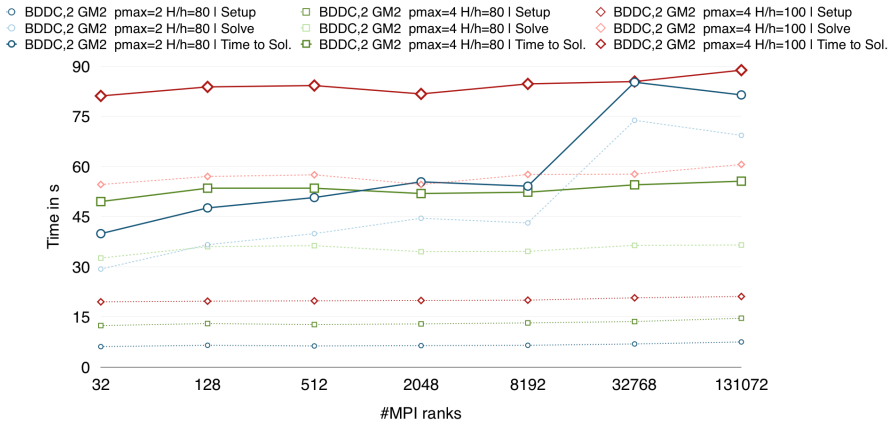


Fig. 3 Weak scalability for $H/h = 80$ and $H/h = 100$ and different truncations *pmax*. *Setup* denotes the BDDC setup time, including all AMG setup times and *Solve* the time spent in the GMRES iteration. Computation performed on JUQUEEN BlueGene/Q at FZ Jülich, Gemany.

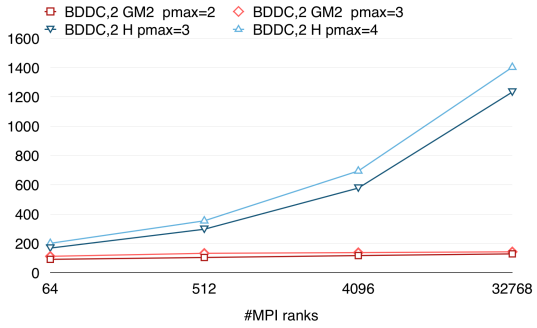


Fig. 4 Heterogeneous and linear elastic material in three dimensions; $H/h = 20$. See Fig. 2 for the remaining notation. Good scalability is achieved using the GM2 interpolation. Computation performed on JUQUEEN BlueGene/Q at FZ Jülich, Gemany.

4 Conclusion

We have shown that a classical AMG approach based on nodal coarsening for systems of PDEs is not sufficient as a preconditioner of the partially coupled matrix in the inexact BDDC approach introduced in Li and Widlund [9], since, for elasticity, it does not fulfill (11) with good bounds. This can be resolved using the GM2 approach, which preserves the nullspace of the partially assembled stiffness matrix in the inexact BDDC method [9]. Our results show that the inexact BDDC approach from [9] using a classical AMG preconditioner with GM2 interpolation is highly parallel scalable and memory efficient.

Acknowledgements This work was supported in part by the German Research Foundation (DFG) through the Priority Programme 1648 "Software for Exascale Computing" (**SPPEXA**) under grants KL 2094/4-1, KL 2094/4-2, RH 122/2-1, and RH 122/3-2. The authors also gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for providing computing time on the GCS Supercomputer SuperMUC at Leibniz Supercomputing Centre (LRZ, www.lrz.de) and **JUQUEEN** [11] at Jülich Supercomputing Centre (JSC, www.fz-juelich.de/ias/jsc). GCS is the alliance of the three national supercomputing centres HLRS (Universität Stuttgart), JSC (Forschungszentrum Jülich), and LRZ (Bayerische Akademie der Wissenschaften), funded by the German Federal Ministry of Education and Research (BMBF) and the German State Ministries for Research of Baden-Württemberg (MWK), Bayern (StMWFK) and Nordrhein-Westfalen (MIWF).

References

1. Santiago Badia, Alberto F. Martn, and Javier Principe. On the scalability of inexact balancing domain decomposition by constraints with overlapped coarse/fine corrections. Parallel Computing, 50:1 – 24, 2015.
2. Allison H. Baker, Axel Klawonn, Tzanio Kolev, Martin Lanser, Oliver Rheinbach, and Ulrike M. Yang. Scalability of classical algebraic multigrid for elasticity to half a million parallel tasks. In Hans-Joachim Bungartz, Philipp Neumann, and Wolfgang E. Nagel, editors, Software for Exascale Computing - SPPEXA 2013-2015, pages 113–140, Cham, 2016. Springer International Publishing.
3. Satish Balay, William D. Gropp, Lois Curfman McInnes, and Barry F. Smith. Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, Modern Software Tools in Scientific Computing, pages 163–202. Birkhauser Press, 1997.
4. Clark R. Dohrmann. An approximate BDDC preconditioner. Numer. Linear Algebra Appl., 14(2):149–168, 2007.
5. Van E. Henson and Ulrike M. Yang. Boomeramg: A parallel algebraic multigrid solver and preconditioner. Applied Numerical Mathematics, 41:155–177, 2002.
6. Axel Klawonn, Martin Lanser, and Oliver Rheinbach. Toward extremely scalable nonlinear domain decomposition methods for elliptic partial differential equations. SIAM Journal on Scientific Computing, 37(6):C667–C696, 2015.
7. Axel Klawonn, Martin Lanser, and Oliver Rheinbach. Nonlinear BDDC methods with inexact solvers. Technical report, 2017. In preparation.
8. Axel Klawonn and Oliver Rheinbach. Inexact FETI-DP methods. Internat. J. Numer. Methods Engrg., 69(2):284–307, 2007.
9. Jing Li and Olof B. Widlund. On the use of inexact subdomain solvers for BDDC algorithms. Comput. Meth. Appl. Mech. Engrg., 196:1415–1428, 2007.
10. John Ruge and Klaus Stüben. Efficient solution of finite difference and finite element equations by algebraic multigrid (AMG). In J.D. Paddon and H. Holstein, editors, The Institute of Mathematics and its Applications Conference Series, volume 3, pages 169–212. Clarenden Press, Oxford, 1985.
11. Michael Stephan and Jutta Docter. JUQUEEN: IBM Blue Gene/Q Supercomputer System at the Jülich Supercomputing Centre. Journal of large-scale research facilities, 1:A1, 2015.
12. Andrea Toselli and Olof Widlund. Domain Decomposition Methods - Algorithms and Theory, volume 34 of Springer Series in Computational Mathematics. Springer, 2004.

On the Accuracy of the Inner Newton Iteration in Nonlinear Domain Decomposition

Axel Klawonn, Martin Lanser, Oliver Rheinbach, and Matthias Uran

1 Introduction

Nonlinear FETI-DP methods [4, 5, 6, 7] belong to the family of nonoverlapping nonlinear domain decomposition methods and can be used to solve discrete nonlinear problems $A(u) = 0$ arising from the discretization of nonlinear partial differential equations. They can be characterized by decomposition before linearization, and they can be interpreted as nonlinearly right-preconditioned Newton-Krylov methods; see [6]. These methods localize work and have shown to be highly scalable to more than 131 072 cores [6].

We decompose the computational domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, into N nonoverlapping subdomains Ω_i, $i = 1, \ldots, N$, such that $\Omega = \bigcup_i^N \Omega_i$. The associated local finite element spaces are denoted by $W^{(i)}$ and the product space by $W = W^{(1)} \times \cdots \times W^{(N)}$. We introduce $\widetilde{W} \subset W$ as the space of all finite element functions from W which are continuous in certain primal variables, e.g., subdomain vertices.

The fully assembled original finite element problem is equivalent to the nonlinear FETI-DP saddle point system

$$A(\tilde{u}, \lambda) = \begin{bmatrix} \widetilde{K}(\tilde{u}) + B^T \lambda - \tilde{f} \\ B\tilde{u} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \tilde{u},\ \tilde{f},\ \widetilde{K}(\tilde{u}) \ \in \widetilde{W}; \tag{1}$$

see [4]. Nonlinear FETI-DP methods are based on solving (1). Here, Lagrange multipliers $\lambda \in V$ are used to decompose the nonlinear problem into parallel local problems on subdomains, and the linear constraint $B\tilde{u} = 0$ enforces the continuity of

Axel Klawonn, Martin Lanser, Matthias Uran
Mathematisches Institut, Universität zu Köln, Weyertal 86-90, 50931 Köln, Germany, e-mail: {axel.klawonn, martin.lanser, m.uran}@uni-koeln.de

Oliver Rheinbach
Institut für Numerische Mathematik und Optimierung, Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg, Akademiestr. 6, 09596 Freiberg, e-mail: oliver.rheinbach@math.tu-freiberg.de

1. Mapping: $M : \widetilde{W} \times V \to \widetilde{W} \times V$.
2. M puts the current iterate into the neighborhood of the solution; see also [1].
3. $M(\tilde{u}, \lambda)$ is easily computable compared to the inverse action of $A(\tilde{u}, \lambda)$.

Fig. 1 Properties on the nonlinear preconditioner M for nonlinear FETI-DP methods.

the solution across the interface for nonprimal variables. Here, B is the standard finite element jump operator and the space of Lagrange multipliers is defined as $V := \text{range}(B)$.

Instead of solving $A(\tilde{u}, \lambda) = 0$ directly with Newton's method, which was denoted Nonlinear-FETI-DP-1 in [4, 6], we introduce a nonlinear right-preconditioner $M(\tilde{u}, \lambda)$; see Figure 1 for some desirable properties the preconditioner should fulfill. The resulting nonlinear equation

$$A(M(\tilde{u}, \lambda)) = 0 \tag{2}$$

is solved by a Newton-Krylov method. In each Newton iteration the evaluation of the preconditioner $g^{(k)} = M(\tilde{u}^{(k)}, \lambda^{(k)})$ is computed. The nonlinear right-preconditioner can be used to describe a (partial) nonlinear elimination of variables [3]. We introduce the index sets E and L, where E is the set of variables which will be eliminated nonlinearly by the application of M and L is the set of variables which will be linearized. According to these two index sets, we split the variables $\tilde{u}$, and the jump operator B, $\tilde{u} = (\tilde{u}_E, \tilde{u}_L)$, $B = \begin{bmatrix} B_E & B_L \end{bmatrix}$. Using this splitting, the nonlinear system (1) writes

$$A(\tilde{u}_E, \tilde{u}_L, \lambda) = \begin{bmatrix} A_E(\tilde{u}_E, \tilde{u}_L, \lambda) \\ A_L(\tilde{u}_E, \tilde{u}_L, \lambda) \\ B_E\tilde{u}_E + B_L\tilde{u}_L \end{bmatrix} = \begin{bmatrix} \widetilde{K}_E(\tilde{u}_E, \tilde{u}_L) + B_E^T\lambda - \tilde{f}_E \\ \widetilde{K}_L(\tilde{u}_E, \tilde{u}_L) + B_L^T\lambda - \tilde{f}_L \\ B_E\tilde{u}_E + B_L\tilde{u}_L \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{3}$$

Since the nonlinear elimination process is restricted to the variables $\tilde{u}_E$, the nonlinear preconditioner $M(\tilde{u}, \lambda)$ is linear in $\tilde{u}_L$ and λ. Therefore, we introduce the following notation

$$M(\tilde{u}, \lambda) = M(\tilde{u}_E, \tilde{u}_L, \lambda) := (M_{\tilde{u}_E}(\tilde{u}_E, \tilde{u}_L, \lambda), \tilde{u}_L, \lambda) = (M_{\tilde{u}_E}(\tilde{u}_L, \lambda), \tilde{u}_L, \lambda) \tag{4}$$

and $M_{\tilde{u}_E}(\tilde{u}_E, \tilde{u}_L, \lambda)$ is defined implicitly by

$$\widetilde{K}_E(M_{\tilde{u}_E}(\tilde{u}_E, \tilde{u}_L, \lambda), \tilde{u}_L) + B_E^T\lambda - \tilde{f}_E = 0. \tag{5}$$

Hence, for the evaluation of $g^{(k)} := M(\tilde{u}_E^{(k)}, \tilde{u}_L^{(k)}, \lambda^{(k)})$, the nonlinear system

$$A_E(g^{(k)}) = 0 \tag{6}$$

has to be solved for fixed $\tilde{u}_L^{(k)}$ and $\lambda^{(k)}$ until a sufficient tolerance ε_I is reached, e.g., by Newton's method with the partial update

$g_0^{(k)} = (\tilde{u}^{(k)}, \lambda^{(k)})$ and $l = 0$
while $||A_E(g_l^{(k)})|| > \varepsilon_I$ **do**
  Newton update to $g_{l+1}^{(k)}$
  $l = l + 1$
  $g^{(k)} = g_l^{(k)}$
end while

$g_0^{(k)} = (\tilde{u}^{(k)}, \lambda^{(k)})$, $l = 0$, $J_{\text{old}} = \frac{1}{2}||A(g_0^{(k)})||^2$
while $||A_E(g_l^{(k)})|| > \varepsilon_I$ **do**
  Newton update to $g_{l+1}^{(k)}$
  Compute: $J_{\text{new}} = \frac{1}{2}||A(g_{l+1}^{(k)})||^2$
  if $J_{\text{new}} > \tau J_{\text{old}}$ **then**
    $g^{(k)} = g_l^{(k)}$
    break while
  else
    $J_{\text{old}} = J_{\text{new}}$
  end if
  $l = l + 1$
  $g^{(k)} = g_l^{(k)}$
end while

Fig. 2 Left: Computation of M. **Right:** Computation of $\mathscr{M}$.

$$g_{E,l+1}^{(k)} = g_{E,l}^{(k)} - (D_{\tilde{u}_E} A_E(g_l^{(k)}))^{-1} A_E(g_l^{(k)}); \tag{7}$$

see also Figure 2 on the left. Thus, the application of the nonlinear right-preconditioner is nothing else than minimizing the energy $J_E(\tilde{u}, \lambda) := \frac{1}{2}||A_E(\tilde{u}, \lambda)||^2$.

Replacing $\tilde{u}_E$ in the second and third line of (3) by $M_{\tilde{u}_E}(\tilde{u}_L, \lambda)$ yields the nonlinear Schur complement

$$S_L(\tilde{u}, \lambda) := \begin{bmatrix} \widetilde{K}_L(M_{\tilde{u}_E}(\tilde{u}_L, \lambda), \tilde{u}_L) + B_L^T \lambda - \tilde{f}_L \\ B_E M_{\tilde{u}_E}(\tilde{u}_L, \lambda) + B_L \tilde{u}_L \end{bmatrix}. \tag{8}$$

Finally, we can solve the resulting nonlinear Schur complement system $S_L(\tilde{u}, \lambda) = 0$ with standard Newton-Krylov-FETI-DP (see [4]). For more details, we also refer to [6].

2 Nonlinear FETI-DP Methods Using Energy Reducing Nonlinear Preconditioning

It is possible that the nonlinear elimination presented above leads to an increase in the global energy $J(\tilde{u}, \lambda) = \frac{1}{2}||A(\tilde{u}, \lambda)||^2$, e.g., if the strong nonlinearities are not contained in the index set E. In this case, our nonlinear FETI-DP methods can show a loss of robustness and performance compared to the traditional Newton-Krylov-FETI-DP approach; see Section 3. It can also happen, that our nonlinear FETI-DP methods do not converge to a solution due to an inappropriate coarse space.

To increase the convergence radius for Newton type methods it is standard to enforce a sufficient decrease in the global energy J in each Newton step [9]. This can be achieved by controlling the Newton update. If the Newton update does not result in a sufficient decrease of the energy, the Newton step is rejected and replaced, e.g., by

a steepest descent step. To prove global convergence properties, usually additional assumptions about the step length have to be fulfilled, which can be controlled by a line search approach enforcing certain conditions of, e.g., Armijo or Wolfe type [8]. For the use of line search in nonlinear FETI-DP methods, see [4].

Analogously to classical Newton-Krylov approaches, it is also possible to apply these strategies to nonlinear right-preconditioned Newton-Krylov methods, which is not considered in this paper. Nevertheless, we additionally have to control the application of the nonlinear preconditioner to enforce an energy decrease in each step, or, at least, to avoid an increase with respect to J.

To enlarge the convergence radius of our nonlinear FETI-DP methods we therefore have to compute $g^{(k)}$ not only with respect to $J_E = \frac{1}{2}||A_E||^2$ but also $J = \frac{1}{2}||A||^2$; cf. (3).

As described above, the application of the nonlinear preconditioner M in our nonlinear FETI-DP methods leads to a minimization of $\frac{1}{2}||A_E(\tilde{u},\lambda)||^2$, but we do not control how the global energy J evolves during this update process. To do so, we introduce an approximation $\mathscr{M}(\tilde{u},\lambda)$ of $M(\tilde{u},\lambda)$, which at least does not increase the global energy J. The idea is, to stop the Newton iteration and choose $\mathscr{M}(\tilde{u},\lambda) = g_l$ whenever the updated g_{l+1} does not fulfill the simple decrease property $J(g_{l+1}) \leq \tau J(g_l)$ for the global energy functional. We thus avoid oversolving in the inner Newton iteration, somewhat analogously to inexact Newton methods with carefully chosen forcing terms [2]. To make this property a robust decrease condition, we choose $0 < \tau \leq 1$ and, if not noted otherwise, we use $\tau = 0.8$ in our experiments. For more details see Figure 2 on the right.

It is obvious that this approach never leads to an increased number of inner Newton iterations but it can end up with two extreme cases. First, if the decrease property is fulfilled for all inner Newton steps we have $\mathscr{M}(\tilde{u},\lambda) = M(\tilde{u},\lambda)$. Second, if the decrease condition is not fulfilled for the first inner Newton step, we obtain $\mathscr{M}(\tilde{u},\lambda) = (\tilde{u},\lambda)$ and the application of $\mathscr{M}$ reduces to the identity. The latter case is identical to a single step of Nonlinear-FETI-DP-1, regardless which set of variables E is chosen. Let us briefly recall the definition of Nonlinear-FETI-DP-1 from [6], where the variable set E is chosen to be the empty set. Let us also remark that in the second case all factorizations from the inner Newton iteration can be recycled for the subsequent outer Newton iteration and therefore no additional work compared with a Nonlinear-FETI-DP-1 step is necessary.

Let us remark that we handle the very first computation of $\mathscr{M}$ in a slightly different way, since we do not want to rely on the initial value $\tilde{u}^{(0)}$. We do not stop the Newton iteration if $J(g_1) > \tau J(g_0)$ but we also compute g_l until $J(g_l) \leq \tau J(g_{l-1})$, $l \geq 2$, is not fulfilled. In a similar way we can control the computation of the initial value $\widetilde{K}(\tilde{u}^{(0)}) = \tilde{f} - B^T\lambda^{(0)}$ (see [4]) in the Nonlinear-FETI-DP-1 approach.

In each outer Newton iteration, we now have to solve the linear system

$$DA\left(\mathscr{M}(\tilde{u}^{(k)},\lambda^{(k)})\right)\left(\delta\tilde{u}^{(k)},\lambda^{(k)}\right)^T = A\left(\mathscr{M}(\tilde{u}^{(k)},\lambda^{(k)})\right). \tag{9}$$

Here, the entries in the right hand side belonging to the index set E can not be guaranteed to be zero due to the fact that $\mathscr{M}$ might just be an approximation to M.

3 Numerical Results

In this section, we present numerical results for nonlinear FETI-DP methods using the newly introduced energy reducing and robust preconditioner and compare them to the nonlinear FETI-DP methods introduced in [4, 5, 6, 7] and to the traditional Newton-Krylov-FETI-DP approach. To provide a fair comparison, we choose for all methods the same initial values $u^{(0)}(x_1,x_2) = x_1 \cdot x_2 \cdot (1-x_1) \cdot (1-x_2)$, $\lambda^{(0)} = 0$, and the same tolerances ε_I and ε_O. Inner Newton iterations are stopped if $\frac{1}{2}||A_E||^2 \leq \varepsilon_I = 1e-12$ or the decrease condition is not fulfilled and the global Newton iteration is stopped if $\frac{1}{2}||A||^2 \leq \varepsilon_O = 1e-12$.

We refer to the nonlinear FETI-DP methods als NL-i, $i = 1,\ldots,4$, and to the nonlinear FETI-DP methods using the new nonlinear preconditioner as NL-ane-i, $i = 1,\ldots,4$. The traditional Newton-Krylov-FETI-DP method is denoted NK. Let us briefly recall the different nonlinear variants from [6] by specifying the nonlinear elimination sets. We choose $E = \emptyset$ in NL-1, $E = [I,\Delta,\Pi]$ in NL-2, $E = [I,\Delta]$ in NL-3, and $E = I$ in NL-4, where I denotes the set of variables inside subdomains, Π denotes the set of primal variables, and Δ denotes the set of all remaining interface variables.

As a model problem, we choose a two dimensional problem based on the scaled p-Laplace operator for $p = 4$

$$\alpha\Delta_p u := \operatorname{div}(\alpha|\nabla u|^{p-2}\nabla u).$$

We consider

$$\begin{aligned} -\alpha\Delta_4 u - \beta\Delta_2 u &= 1 \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

with the computational domain $\Omega = (0,1)^2$ and the coefficients $\alpha = 1e5$ and $\beta = 1$.

The computational domain is decomposed into square subdomains and discretized by piecewise linear finite elements. We choose a problem, where the nonlinearities have a nonlocal character. Here, columns of subdomains are intersected by channels of width $H/2$ from the upper to the lower boundary of Ω, where H is the width of a subdomain; see the left picture in Figure 3. To simulate a less structured domain decomposition, we also consider subdomains with ragged edges; see the right picture in Figure 3 for details. For all our tests we used a sequential MATLAB implementation and we exclusively consider subdomain vertices as primal constraints. Due to our sequential implementation, we choose and evaluate different metrics or indicators to obtain a good estimation of the parallel potential of the different nonlinear FETI-DP preconditioner variants. As a metric for the global communication, we count the number of Krylov iterations (denoted # Krylov It.). For the local work, we count the number of factorizations of $D\widetilde{K}_{BB}$ or $D\widetilde{K}_{II}$ (denoted by "Local Fact."), and we also count the factorizations of the FETI-DP coarse problem (denoted by "Coarse Fact."). Factorizations of the coarse problem are necessary in the computation of the initial value for NL-1 and in the evaluation of the

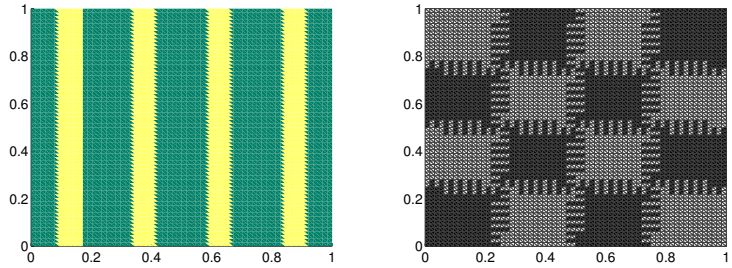


Fig. 3 Left: Channels with a width of $H/3$, where H is the width of a subdomain; $\alpha = 1e5$. **Right:** Domain decomposition with ragged edges, $H/h = 16$.

nonlinear preconditioner for NL-2, while the evaluation of the preconditioner for NL-3 and NL-4 does not include factorizations of the coarse problem. Therefore, we subdivide the section "Coarse Fact." into factorizations of the coarse problem in the first/inner loop (denoted by "in.") and in the main loop (denoted by "out."). For all methods the number of outer coarse factorizations is equal to the number of Newton steps.

For our model problem, the index set E does not contain the nonlinearities for the NL-4 and NL-ane-4 method. As a result the performance of NL-4 is worse than the performance of the traditional NK approach and the number of local factorizations of NL-ane-4 is equal to the number of Newton steps plus one. This shows that the elimination of the interior variables is inappropriate for this problem, but NL-ane-4 detects this and avoids spending time in the evaluation of the inappropriate nonlinear preconditioner. As a consequence, NL-ane-4 is nearly equivalent to NL-1 without the computation of the initial value or to NK and thus superior compared to NL-4. The difference of one factorization results from the additional step in the inner loop in the very first Newton step.

For the structured decomposition into square subdomains NL-2, NL-3, NL-ane-2, and NL-ane-3 perform quite similar. The number of local solves for NL-ane-2 and NL-ane-3 is half as large as for NL-2 and NL-3, but the number of Krylov iterations is slightly higher.

For the less structured decomposition with ragged edges the chosen coarse space (subdomain vertices) is insufficient for NL-2 and NL-3, so these methods do not converge, but using the new approach leads to convergence and saves about 50% of Newton steps and Krylov iterations compared to the traditional NK approach. The new strategy thus increases the convergence radius for NL-2 and NL-3.

4 Conclusion

We have introduced a strategy to automatically decide on the computational effort to be spent in the inner Newton iteration in nonlinear domain decomposition. The

Table 1 Model problem "Nonlocal Nonlinearities"; comparison of standard nonlinear FETI-DP methods and nonlinear FETI-DP methods using the new approach ("NL-ane-*"); channels with a width of $H/2$; $\alpha = 1e5$ inside channels and $\beta = 1$ elsewhere; see also Figure 3; domain $\Omega = (0,1)^2$; decomposed into square subdomains; $H/h = 16$; $\varepsilon_I = 1e{-}12$; $\varepsilon_O = 1e{-}12$; $\tau = 0.8$; computed on Schwarz.

Channels 2D $H/h = 16$**; exact FETI-DP; computed on Schwarz,** $\alpha = 1e5$											
				Normal Edges				**Ragged Edges**			
N	**Problem Size**	$\lvert E\rvert$	**Nonlinear Solver**	**Local Factor. in.**	**Coarse Factor. out.**		**Krylov It.**	**Local Factor. in.**	**Coarse Factor. out.**		**Krylov It.**
16	4225		NK	**13**	-	13	173	**13**	-	13	2108
		0	NL-1 no Init	**13**	-	13	188	**13**	-	13	2227
		0	NL-ane-1	15	6	9	124	14	5	9	1211
		0	NL-1	22	12	10	150	26	17	9	1170
		4225	NL-ane-2	16	11	**5**	**68**	16	10	**6**	**794**
		4225	NL-2	30	24	6	86	div	div	div	div
		4216	NL-ane-3	23	**0**	7	95	25	**0**	9	1134
		4216	NL-3	30	**0**	6	86	div	div	div	div
		3856	NL-ane-4	17	**0**	13	254	14	**0**	13	2227
		3856	NL-4	47	**0**	13	284	43	**0**	13	2277
256	66049		NK	15	-	15	1391	15	-	15	3064
		0	NL-1 no Init	**14**	-	14	1471	**14**	-	14	3139
		0	NL-ane-1	16	6	10	741	**15**	5	10	2149
		0	NL-1	23	13	10	730	36	25	11	2387
		66049	NL-ane-2	16	10	**6**	447	19	11	**8**	**1664**
		66049	NL-2	31	25	**6**	395	div	div	div	div
		65824	NL-ane-3	17	**0**	**6**	429	18	**0**	**8**	1683
		65824	NL-3	35	**0**	**6**	**379**	div	div	div	div
		58624	NL-ane-4	19	**0**	14	1647	**15**	**0**	14	3139
		58624	NL-4	54	**0**	14	1681	50	**0**	14	3156

strategy considers the reduction of the global energy resulting from performing local Newton steps on the subdomains. The Newton iteration performed for the local elimination is stopped (and the step is discarded) when the resulting decrease in the global energy is not satisfactory. This can also be interpreted as an inexact nonlinear elimination. We have shown, that the local work can be significantly reduced compared to standard nonlinear FETI-DP methods while the number of Newton steps and Krylov iterations remains nearly constant. We have also shown, that the dependency on the coarse space is reduced for nonlinear FETI-DP methods and that the robustness of the resulting methods is dramatically increased.

Acknowledgements This work was supported in part by the German Research Foundation (DFG) through the Priority Programme 1648 "Software for Exascale Computing" (**SPPEXA**) under grants KL 2094/4-1, KL 2094/4-2, RH 122/2-1, and RH 122/3-2.

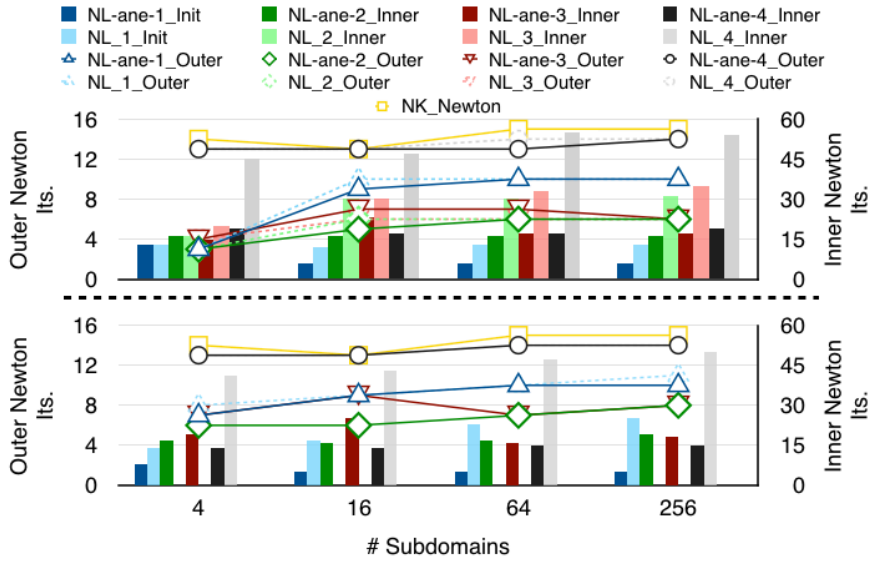


Fig. 4 Model problem "Nonlocal Nonlinearities"; comparison of nonlinear FETI-DP methods and nonlinear FETI-DP methods using energy minimizing preconditioning; channels with a width of $H/3$; $p = 4$ and $\alpha = 1e5$ in the channels and $\beta = 1$ elsewhere; see also Figure 3; domain $\Omega = (0,1)^2$; decomposed into square subdomains; $H/h = 16$; $\varepsilon_I = 1e{-}12$; $\varepsilon_O = 1e{-}12$; $\tau = 0.8$; computed on Schwarz. **Top:** Normal edges; **Bottom:** Ragged edges

References

1. Brune, P. R., Knepley, M. G., Smith, B. F., Tu, X.: Composing scalable nonlinear algebraic solvers. In: SIAM Rev., 57, pp. 535-565. (2015)
2. Eisenstat, S. C., Walker, S. F.: Choosing the forcing terms in an inexact Newton method. In: SIAM J. Sci. Comput., 17(1), pp. 16-32. (1996)
3. Lanzkorn, P. J., Rose, D. J., Wilkes, J. T.: An analysis of approximate nonlinear elimination. In: SIAM J. Sci. Comput., 17, pp. 538-559. (1996)
4. Klawonn, A., Lanser, M., Rheinbach, O.: Nonlinear FETI-DP and BDDC Methods. In: SIAM J. Sci. Comput., 36, pp. A737-A765. (2014)
5. Klawonn, A., Lanser, M., Rheinbach, O., Uran, M.: New nonlinear FETI-DP methods based on partial nonlinear elimination of variables: In: Lect. Notes Comp. Sci. Eng., vol 116, Proceedings of the 23rd International Conference on Domain Decomposition, Springer-Verlag, pp. 207-215. (2017)
6. Klawonn, A., Lanser, M., Rheinbach, O., Uran, M.: Nonlinear FETI-DP and BDDC Methods: A Unified Framework and Parallel Results. In: SIAM J. Sci. Comput., 39, pp. C417-C451. (2017)
7. Klawonn, A., Lanser, M., Radtke, P., Rheinbach, O.: On an adaptive coarse space and on nonlinear domain decomposition. In: Domain Decomposition Methods in Science and Engineering XXI, Erhel, J. and Gander, M. J. and Halpern, L. and Pichot, G. and Sassi, T. and Widlund, O., eds., vol. 98 of Lecture Notes in Computational Science and Engineering, Springer International Publishing, pp. 71-83. (2014)
8. Nocedal, J., Wright, S. J.: Numerical Optimization, 2nd Edition, Springer, Berlin. (2006)
9. Ulbrich, M., Ulbrich, S.: Nichtlineare Optimierung. Springer Basel, (2012).

Adaptive BDDC and FETI-DP methods with change of basis formulation

Hyea Hyun Kim[1], Eric T. Chung[2], and Junxian Wang[3]

1 Introduction

In this paper, BDDC (Balancing Domain Decomposition by Constraints) and FETI-DP (Dual-Primal Finite Element Tearing and Interconnecting) algorithms with a change of basis for adaptive primal constraints are analyzed. In our formulation, adaptive primal constraints are introduced from appropriate generalized eigenvalue problems. In the authors previous study Kim et al. [2017a], for the FETI-DP algorithm the adaptive primal constraints are enforced by using a projection and it was shown that the condition numbers are controlled by the user-defined tolerance value, which is used to select the adaptive primal constraints from generalized eigenvalue problems on each equivalence classes, edges and faces. The analysis in Kim et al. [2017a] could not be extended to the FETI-DP algorithm with a change of basis formulation on the adaptive primal constraints. In the change of basis formulation, each primal constraint is transformed into a single unknown and treated just like unknowns at subdomain vertices as in the standard FETI-DP algorithm. It is often observed that the change of basis formulation is numerically more stable than the projection approach.

Here we will propose a more general form of the FETI-DP preconditioner and extend the analysis to the change of basis formulation. For the proposed preconditioner, we can obtain the identity $E_D + P_D = I$ for the averaging and jump operators , see (8) for their definitions, and thus show that the condition numbers of the adaptive BDDC and FETI-DP algorithms with the change of basis formulation are identical. Unlike in the standard FETI-DP preconditioners, the blocks of subdomain matrices and scaling matrices corresponding to the adaptive primal unknowns appear in the proposed preconditioner. We

[1]Department of Applied Mathematics and Institute of Natural Sciences, Kyung Hee University, Korea. hhkim@khu.ac.kr .[2]Department of Mathematics, The Chinese University of Hong Kong, Hong Kong SAR. tschung@math.cuhk.edu.hk .[3]School of Mathematics and Computational Science, Xiangtan University, China. wangjunxian@xtu.edu.cn

note that in the same mini-symposium an adaptive FETI-DP algorithm with a change of basis formulation was presented in the talk by Axel Klawonn, where different generalized eigenvalue problems are introduced and different tools are used in the analysis of condition numbers.

We note that adaptive primal constraints are often required to obtain robustness of domain decomposition preconditioners with respect to coefficient variations in the model problem. For related works, we refer to Galvis and Efendiev [2010] and Dolean et al. [2012] for two-level additive Schwarz methods, and Spillane et al. [2013] and Spillane and Rixen [2013] for FETI/BDD methods. In a pioneering work by Mandel et al. [2012], adaptive BDDC algorithms are developed and tested for $3D$ problems, where the adaptive primal constraints are selected from generalized eigenvalue problems on each face. For $3D$ problems, more advanced FETI-DP/BDDC algorithms are developed and analyzed in more recent works, see Klawonn et al. [2016],Calvo and Widlund [2016], and Kim et al. [2017b]. In Klawonn et al. [2016], Kim et al. [2017b], and Kim et al. [2017a], the adaptive primal constraints are enforced by using a projection in the FETI-DP algorithm.

2 BDDC and FETI-DP algorithms

For the presentation of BDDC and FETI-DP algorithms, we introduce a finite element space X for a given domain Ω, where the model elliptic problem is define as

$$-\nabla \cdot (\rho(x)\nabla u(x)) = f(x) \tag{1}$$

with a zero boundary condition on $u(x)$ and with $\rho(x)$ being highly varying and heterogeneous. The domain Ω is then partitioned into non-overlapping subdomains $\{\Omega_i\}$. We assume that the subdomain boundaries do not cut the triangles in the finite element space X. We use the notation X_i to denote the restriction of X to Ω_i. Each subdomain is then equipped with the finite element space X_i.

We further introduce W_i as the restriction of X_i to the subdomain interface unknowns, W, and X as the product of local finite element spaces W_i and X_i, respectively. We note that functions in W or X are decoupled across the subdomain interfaces. We then select some primal unknowns among the decoupled unknowns on the interfaces and enforce continuity on them and denote the corresponding spaces $\widetilde{W}$ and $\widetilde{X}$.

The preconditioners in BDDC and FETI-DP algorithms will be developed based on the partially coupled space $\widetilde{W}$ and appropriate scaling matrices. In our adaptive methods, we will select primal unknowns on each nodal equivalence classes of subdomain interfaces. In more detail, edges in $2D$ and faces in $3D$ are nodal equivalence classes shared by two subdomains, edges in $3D$ are

nodal equivalence classes shared by more than two subdomains, and vertices are end points of edges in both $2D$ and $3D$.

In our approach, we first include the unknowns at subdomain vertices to the set of primal unknowns. Adaptive primal constraints will be selected from eigenvectors of generalized eigenvalue problems on faces and edges using a given tolerance value. The associated adaptive primal unknowns are then obtained by applying change of basis on the adaptively selected primal constraints and these explicit unknowns can then be assembled strongly just like unknowns at subdomain vertices.

We introduce notations K_i and S_i. The matrices K_i are obtained from Galerkin approximation of

$$a(u,v) = \int_{\Omega_i} \rho(x) \nabla u \cdot \nabla v \, dx$$

by using finite element spaces X_i and S_i are Schur complements of K_i, which are obtained after eliminating unknowns interior to Ω_i. Let $\widetilde{R}_i : \widetilde{W} \to W_i$ be the restriction into $\partial\Omega_i$ and let $\widetilde{S}$ be a partially coupled matrix defined by

$$\widetilde{S} = \sum_{i=1}^{N} \widetilde{R}_i^T S_i \widetilde{R}_i. \tag{2}$$

We note that $\widetilde{S}$ is then coupled at the unknowns on subdomain vertices and the adaptive primal unknowns. Let $\widetilde{R}$ be the restriction from $\widehat{W}$ to $\widetilde{W}$, where the subspace $\widehat{W}$ of $\widetilde{W}$ has unknowns continuous on the subdomain interface. The discrete problem of (1) is then written as

$$\widetilde{R}^T \widetilde{S} \widetilde{R} = \widetilde{R}^T \widetilde{g},$$

where $\widetilde{g}$ is the vector related to the right hand side $f(x)$.

In the BDDC algorithm the above matrix equation is solved iteratively by using the following preconditioner,

$$M_{BDDC}^{-1} = \widetilde{R}^T \widetilde{D} \widetilde{S}^{-1} \widetilde{D}^T \widetilde{R}, \tag{3}$$

where $\widetilde{D}$ is a scaling matrix of the form

$$\widetilde{D} = \sum_{i=1}^{N} \widetilde{R}_i^T D_i \widetilde{R}_i.$$

Here the matrices D_i are defined for unknowns in W_i and they are introduced to resolve heterogeneity in $\rho(x)$ across the subdomain interface. In a more detail, D_i consists of blocks $D_F^{(i)}$, $D_E^{(i)}$, $D_V^{(i)}$, where F denotes corresponding blocks to faces, E to edges, and V to vertices, respectively. We note that those

blocks satisfy the partition of unity for a given F, E, and V, respectively. We refer to Klawonn and Widlund [2006] for these definitions.

The FETI-DP algorithm is a dual form of the BDDC algorithm. After the change of unknowns on the adaptively selected constraints, we obtain the resulting FETI-DP algebraic system

$$B\widetilde{S}^{-1}B^T\lambda = d, \tag{4}$$

where $\widetilde{S}$ is the partially coupled matrix defined in (2), and B is the matrix with entries 0, -1, and 1, which is used to enforce continuity at the remaining decoupled interface unknowns, i.e., dual unknowns. We introduce the notation M for the set of Lagrange multipliers λ, of which dimension is identical to the number of continuity constraints enforced on the remaining decoupled interface unknowns. The above algebraic system is then solved by an iterative method with the following preconditioner

$$M^{-1}_{FETI} = \sum_{i=1}^{N} B^{(i)}_{D,\Delta} S_i (B^{(i)}_{D,\Delta})^T \tag{5}$$

where $(B^{(i)}_{D,\Delta})^T : M \to W_i$ is defined by

$$(B^{(i)}_{D,\Delta})^T\lambda|_F = D^{(j)}_{F,\Delta}\lambda_{ij} \text{ on each } F \in F(i) \tag{6}$$

and

$$(B^{(i)}_{D,\Delta})^T\lambda|_E = \sum_{l\in n(E,i)} D^{(l)}_{E,\Delta}\lambda_{il} \text{ on each } E \in E(i). \tag{7}$$

Here $F(i)$ and $E(i)$ denote the set of faces and edges of subdomain Ω_i, respectively, $n(E,i)$ denotes the set of neighboring subdomain indices sharing the edge E with Ω_i, and λ_{ij} denotes the part of Lagrange multipliers λ used to enforce continuity on the decoupled unknowns across Ω_i and Ω_j. The matrices $D^{(j)}_{F,\Delta}$ and $D^{(l)}_{E,\Delta}$ are given by blocks of $D^{(j)}_F$ and $D^{(l)}_E$ as follows,

$$D^{(j)}_{F,\Delta} = \begin{pmatrix} D^{(j)}_{F,\Delta\Delta} \\ D^{(j)}_{F,\Pi\Delta} \end{pmatrix}, \quad D^{(l)}_{E,\Delta} = \begin{pmatrix} D^{(l)}_{E,\Delta\Delta} \\ D^{(l)}_{E,\Pi\Delta} \end{pmatrix},$$

where the subscripts Δ and Π denote blocks of matrix $D^{(j)}_F$ and $D^{(l)}_E$ corresponding to the decoupled unknowns and the adaptive primal unknowns, respectively. For the unknowns at subdomain vertices, which belong to the initial set of primal unknowns, the values of $(B^{(i)}_{D,\Delta})^T\lambda$ are defined as zero. Differently from the standard FETI-DP preconditioner, the proposed preconditioner contains the scaling matrices involving the adaptive primal unknowns. With this new form of the FETI-DP preconditioner, we can show that the adaptive FETI-DP algorithm with the change of basis formulation

has the same spectra except the values zero and one and thus can obtain the same condition number bound as that of the BDDC algorithm. When no adaptive primal unknowns are chosen, the preconditioner is identical to that considered in the standard FETI-DP algorithm.

3 Adaptively enriched coarse spaces

The adaptive constraints will be selected by considering generalized eigenvalue problems on each equivalence class. The idea is originated from the upper bound estimate of BDDC and FETI-DP preconditioner. In the estimate of condition numbers of BDDC and FETI-DP preconditioners, the average and jump operators are defined as

$$E_D = \widetilde{R}\widetilde{R}^T\widetilde{D}, \quad P_D = B_D^T B, \tag{8}$$

where $B = (B_\Delta \ \ 0)$ and $B_D^T = (B_{D,\Delta}^{(1)} \ \cdots \ B_{D,\Delta}^{(N)})^T$. We note that $B : \widetilde{W} \to M$ and $B_D^T : M \to W$, see the definition of $(B_{D,\Delta}^{(i)})^T$ in (6) and (7).

The adaptive constraints are then treated just like unknowns at subdomain vertices after change of basis formulation in both BDDC and FETI-DP algorithms, i.e., the continuity on them can be strongly enforced. We note that in our previous work one can not get $E_D + P_D = I$ when the standard FETI-DP preconditioner is considered for the change of basis formulation, i.e., without the blocks from the adaptive primal unknowns in the definition of the scaled jump operator B_D^T.

We will now introduce generalized eigenvalue problems for each face and each edge. For a face F, the following generalized eigenvalue problem is considered

$$A_F v_F = \lambda \widetilde{A}_F v_F, \tag{9}$$

where

$$A_F = (D_F^{(j)})^T S_F^{(i)} D_F^{(j)} + (D_F^{(i)})^T S_F^{(j)} D_F^{(i)}, \ \widetilde{A}_F = \widetilde{S}_F^{(i)} : \widetilde{S}_F^{(j)}.$$

In the above $S_F^{(i)}$ denote block matrix of S_i to the unknowns interior to F and $\widetilde{S}_F^{(i)}$ are Schur complements of S_i obtained by eliminating unknowns except those interior to F. The matrices then satisfy the following minimal energy property,

$$v_F^T \widetilde{S}_F^{(i)} v_F \le v^T S_i v, \text{ for any } v|_F = v_F, \tag{10}$$

where $v|_F$ denotes the restriction of v to the unknowns interior to F. The notation $A : B$ is a parallel sum defined as, see Anderson and Duffin [1969],

$$A : B = A(A + B)^+ B,$$

where $(A+B)^+$ denotes a pseudo inverse. The parallel sum satisfies the following properties

$$A:B = B:A, \quad A:B \leq A, \quad A:B \leq B, \tag{11}$$

and it was first used in forming generalized eigenvalues problems by Dohrmann and Pechstein [2013], of which idea was originated from the energy estimate of the average operator in the BDDC algorithm.

In (9), the eigenvalues are all positive and we select eigenvectors $v_{F,l},\ l \in N(F)$ with associated eigenvalues λ_l larger than the given λ_{TOL}. The following constraints will then be enforced on the unknowns in F,

$$(A_F v_{F,l})^T (w_F^{(i)} - w_F^{(j)}) = 0,\ l \in N(F).$$

After a change of basis, the above constraints can be transformed into explicit unknowns.

In $3D$, we can have an edge, a nodal equivalence class shared by more than two subdomains, and for an edge E we introduce the following generalized eigenvalue problem,

$$A_E v_E = \lambda \widetilde{A}_E v_E,$$

where

$$A_E = \sum_{m \in I(E)} \sum_{l \in I(E)\setminus\{m\}} (D_E^{(l)})^T S_E^{(m)} D_E^{(l)}, \quad \widetilde{S}_E = \prod_{m \in I(E)} \widetilde{S}_E^{(m)},$$

and $I(E)$ denotes the set of subdomain indices sharing E in common, and $\prod_{m\in I(E)} \widetilde{S}_E^{(m)}$ is the parallel sum of matrices $\widetilde{S}_E^{(m)}$. We note that $S_E^{(m)}$ and $\widetilde{S}_E^{(m)}$ are defined similarly as $S_F^{(m)}$ and $\widetilde{S}_F^{(m)}$. For a given λ_{TOL}, the eigenvectors with their eigenvalues larger than λ_{TOL} will be selected and denoted by $v_{E,l},\ l \in N(E)$. The following constraints will then enforced on the unknowns in E,

$$(A_E v_{E,l})^T (w_E^{(i)} - w_E^{(m)}) = 0,\ l \in N(E),\ m \in I(E) \setminus \{i\}.$$

Similarly to the face case, the above constraints can be transformed into explicit unknowns after the change of basis.

By using the adaptively selected primal unknowns on each face F and edge E as above, we can obtain the following estimate

$$\langle \widetilde{S}(I - E_D)\widetilde{w}, (I - E_D)\widetilde{w}\rangle \leq C\lambda_{TOL} \langle \widetilde{S}\widetilde{w}, \widetilde{w}\rangle, \tag{12}$$

where C is a constant depending on the maximum number of edges and faces per subdomain, and the maximum number of subdomains sharing an edge but independent of the coefficient $\rho(x)$. We note that the above inequality is the key estimate in the analysis of the BDDC algorithm.

4 Condition number estimate and numerical results

Using the adaptively enriched primal unknowns described in Section 3 and the estimate in (12), we can obtain the following estimate of condition numbers for the given λ_{TOL}:

Theorem 1. *The BDDC algorithm with the change of basis formulation for the adaptively chosen set of primal unknowns with a given tolerance λ_{TOL} has the following bound of condition numbers,*

$$\kappa(M_{BDDC}^{-1}\widetilde{R}^T\widetilde{S}\widetilde{R}) \leq C\lambda_{TOL},$$

and the FETI-DP algorithm with the change of basis formulation for the same set of adaptively chosen set of primal unknowns has the bound

$$\kappa(M_{FETI}^{-1}B\widetilde{S}^{-1}B^T) \leq C\lambda_{TOL},$$

where C is a constant depending only on $N_{F(i)}$, $N_{E(i)}$, $N_{I(E)}$, which are the number of faces per subdomain, the number of edges per subdomain, and the number of subdomains sharing an edge E, respectively. In fact, the two algorithms share the same set of eigenvalues except zero and one.

The proof of the above theorem and some numerical examples can be found in a complete version of this paper Kim et al. [2017c]. In Table 1, we present some numerical experiments for a 3D model problem. In particular, we consider a random coefficient with value varying between 10^{-3} to 10^3, and show the number of iterations and the number of primal unknowns with various choice of coarse partition N_d. We observe a very robust performance.

Table 1 Performance of adaptive BDDC and FETI-DP with $\lambda_{TOL}^F = 10, \lambda_{TOL}^E = 10^3$ for highly varying and random $\rho(x)$ in $(10^{-3}, 10^3)$ by increasing N_d and with a fixed $H/h = 12$: $\lambda_{\min}$ (minimum eigenvalues), $\lambda_{\max}$ (maximum eigenvalues), Iter (number of iterations), pnumF (total number of adaptive primal unknowns on faces), and pnumE (total number of adaptive primal unknowns on edges). pF and pE are the number of adaptive primal unknowns per face and per edge, respectively.

N_d	method	$\lambda_{\min}$	$\lambda_{\max}$	Iter	pnumF	pnumE	pF	pE
2^3	Bddc	1.00	5.29	18	21	18	1.75	3.00
	Fdp	1.00	5.29	18	21	18	1.75	3.00
3^3	Bddc	1.01	6.97	26	71	115	1.31	3.19
	Fdp	1.00	6.97	27	71	115	1.31	3.19
4^3	Bddc	1.01	9.45	29	205	320	1.42	2.96
	Fdp	1.00	9.45	30	205	320	1.42	2.96

Acknowledgements The first author was supported by the National Research Foundation of Korea(NRF) grants funded by NRF20151009350, the second author was supported by the Hong Kong RGC General Research Fund (Project 14317516) and the CUHK Direct

Grant for Research 2016-17, and the third author was supported by the National Natural Science Foundation of China (Grant No. 11201398) and Open Foundation of Guangdong Provincial Engineering Technology Research Center for Data Science(2016KF07).

References

W. N. Anderson, Jr. and R. J. Duffin. Series and parallel addition of matrices. *J. Math. Anal. Appl.*, 26:576–594, 1969.

Juan G. Calvo and Olof B. Widlund. An adaptive choice of primal constraints for BDDC domain decomposition algorithms. *Electron. Trans. Numer. Anal.*, 45:524–544, 2016.

Clark R. Dohrmann and Clemens Pechstein. Modern domain decomposition solvers: BDDC, deluxe scaling, and an algebraic approach, http://people.ricam.oeaw.ac.at/c.pechstein/pechstein-bddc2013.pdf. 2013.

Victorita Dolean, Frédéric Nataf, Robert Scheichl, and Nicole Spillane. Analysis of a two-level Schwarz method with coarse spaces based on local Dirichlet-to-Neumann maps. *Comput. Methods Appl. Math.*, 12(4):391–414, 2012.

Juan Galvis and Yalchin Efendiev. Domain decomposition preconditioners for multiscale flows in high-contrast media. *Multiscale Model. Simul.*, 8(4): 1461–1483, 2010.

Hyea Hyun Kim, Eric Chung, and Junxian Wang. BDDC and FETI-DP mathods with enriched coarse spaces for elliptic problems with oscillatory and high contrast coefficients. In *Domain decomposition methods in science and engineering XXIII*, volume 116 of *Lect. Notes Comput. Sci. Eng.*, pages 179–186. Springer, Heidelberg, 2017a.

Hyea Hyun Kim, Eric Chung, and Junxian Wang. BDDC and FETI-DP preconditioners with adaptive coarse spaces for three-dimensional elliptic problems with oscillatory and high contrast coefficients. *J. Comput. Phys.*, 349:191–214, 2017b.

Hyea Hyun Kim, Eric Chung, and Junxian Wang. Adaptive BDDC and FETI-DP algorithms with change of basis formulation. *Submitted.*, 2017c.

Axel Klawonn and Olof B Widlund. Dual-primal FETI methods for linear elasticity. *Comm. Pure Appl. Math.*, 59(11):1523–1572, 2006.

Axel Klawonn, Martin Kühn, and Oliver Rheinbach. Adaptive coarse spaces for FETI-DP in three dimensions. *SIAM J. Sci. Comput.*, 38(5):A2880–A2911, 2016.

Jan Mandel, Bedřich Sousedík, and Jakub Šístek. Adaptive BDDC in three dimensions. *Math. Comput. Simulation*, 82(10):1812–1831, 2012.

N. Spillane and D. J. Rixen. Automatic spectral coarse spaces for robust finite element tearing and interconnecting and balanced domain decomposition algorithms. *Internat. J. Numer. Methods Engrg.*, 95(11):953–990, 2013.

Nicole Spillane, Victorita Dolean, Patrice Hauret, Frédéric Nataf, and Daniel J. Rixen. Solving generalized eigenvalue problems on the interfaces to build a robust two-level FETI method. *C. R. Math. Acad. Sci. Paris*, 351(5-6):197–201, 2013.

Nonoverlapping three grid Additive Schwarz for hp-DGFEM with discontinuous coefficients

Piotr Krzyżanowski

Abstract We discuss a nonoverlapping additive Schwarz method for an h-p DGFEM discretization of an elliptic PDE with discontinuous coefficients, where the fine grid is decomposed into subdomains of size H and the coarse grid consists of cells size $\mathscr{H}$ such that $h \leq H \leq \mathscr{H}$. We prove the condition number is $O(p^2/q) \cdot O(\mathscr{H}^2/Hh)$ and is independent from the jumps of the coefficient if the discontinuities are aligned with the coarse grid.

1 Introduction

Let us consider a second order elliptic equation

$$-\operatorname{div}(\rho \nabla u) = f, \tag{1}$$

with homogeneous Dirichlet boundary condition. The problem is discretized by an h-p symmetric weighted interior penalty discontinuous Galerkin finite element method. A nonoverlapping additive Schwarz method (see [3], [1]) is applied to precondition the discrete equations. For $\rho \equiv 1$, Antonietti and Houston [1] conjectured on the basis of numerical experiments that if the coarse space contains piecewise polynomial functions up to degree p, the condition number is $O(p\,\mathscr{H}/h)$. This conjecture has recently been proved in [5] and independently by Antonietti, Houston and Smears in [2], using slightly different techniques. In the former paper, a general framework for the analysis of problems with discontinuous coefficients and varying polynomial degrees across finite elements has been developed; however, a technical assumption that the basis functions are continuous inside subdomains was made when the coefficient was allowed discontinous in Ω. On the other hand, [2] made use of approximation ideas of [6], allowing for more flexibility in the choice of the finite element spaces. In this note, we extend the analysis to the case when fully

University of Warsaw, Poland, e-mail: p.krzyzanowski@mimuw.edu.pl

discontinuous finite elements are employed, under additional assumption that the coefficient is constant inside coarse grid cells and an H^2–regularity assumption on (1) holds.

For more flexibility and enhanced parallelism, we formulate our results addressing the case when the subdomains (where the local problems are solved in parallel) are potentially smaller than the coarse grid cells [4]. By allowing small subdomains of diameter $H \leq \mathscr{H}$, local problems are cheaper to solve and the amount of concurrency of the method is substantially increased, which can be an advantage e.g. on multi-threaded processors. Moreover, small subdomains give more flexibility in assigning them to processors for load balancing in coarse grain parallel processing. In this way, an additional level of domain partitioning gives the user more parameters to fine tune the actual parallel performance, and thus overall efficiency, of the preconditioner for a given hardware architecture.

The paper is organized as follows. In Section 2, the differential problem and its discontinuous Galerkin discretization are formulated. In Section 3, a nonoverlapping two-level, three-grid additive ASM for solving the discrete problem is designed and analyzed under assumption that the coarse mesh resolves the discontinuities of the coefficient, the variation of the mesh size and of the polynomial degree are locally bounded, and the original problem satisfies some regularity assumption. Section 4 presents some numerical experiments.

For nonnegative scalars x, y, we shall write $x \lesssim y$ if there exists a positive constant C, independent of: x, y, the fine, subdomain and coarse mesh parameters $h, H, \mathscr{H}$, the orders of the finite element spaces p, q, and of jumps of the diffusion coefficient ρ as well, such that $x \leq Cy$. If both $x \lesssim y$ and $y \lesssim x$, we shall write $x \simeq y$.

The norm of a function f from the Sobolev space $H^k(S)$ will be denoted by $||f||_{k,S}$, while the seminorm of f will be denoted by $|f|_{k,S}$. For short, the L^2-norm of f will then be denoted by $|f|_{0,S}$.

2 Differential problem and its *h-p* discontinuous Galerkin discretization

Let Ω be a bounded open convex polyhedral domain in R^d, $d \in \{2,3\}$, with Lipschitz boundary $\partial\Omega$. We consider the following problem for given $f \in L^2(\Omega)$ and $\rho \in L^\infty(\Omega)$:

Find $U^* \in H_0^1(\Omega)$ such that

$$a(U^*, v) = (f, v)_\Omega, \qquad \forall v \in H_0^1(\Omega), \tag{2}$$

where

$$a(u,v) = \int_\Omega \rho \nabla u \cdot \nabla v dx, \qquad (f,v)_\Omega = \int_\Omega f v dx.$$

We assume that there exist constants α_0 and α_1 such that $0 < \alpha_0 \leq \rho \leq \alpha_1$ a.e. in Ω so that (2) is well–posed. Without loss of generality we shall additionally

suppose that $\alpha_0 \geq 1$ and $\mathrm{diam}(\Omega) = 1$, which can always be guaranteed by simple scaling. We also assume that ρ is piecewise constant, i.e. Ω can be partitioned into nonoverlapping polyhedral subregions with the property that ρ restricted to any of these subregions is some positive constant.

Let $\mathscr{T}_h = \{K_1, \ldots, K_{N_h}\}$ denote an affine nonconforming partition of Ω, where K_i are either triangles in 2-D or tetrahedrons in 3-D. For $K \in \mathscr{T}_h$ we set $h_K = \mathrm{diam}(K)$. By $\mathscr{E}_h^{\mathrm{in}}$ we denote the set of all common (internal) faces (edges in 2-D) of elements in $\mathscr{T}_h$, so that $e \in \mathscr{E}_h^{\mathrm{in}}$ iff $e = \partial K_i \cap \partial K_j$ is of positive measure. We will use symbol $\mathscr{E}_h$ to denote the set of all faces (edges in 2-D) of fine mesh $\mathscr{T}_h$, that is those either in $\mathscr{E}_h^{\mathrm{in}}$ or on the boundary $\partial\Omega$. For $e \in \mathscr{E}_h$ we set $h_e = \mathrm{diam}(e)$. We assume that $\mathscr{T}_h$ is shape- and contact–regular, that is, it admits a matching submesh $\mathscr{T}_{\hat{h}}$ which is shape–regular and such that for any $K \in \mathscr{T}_h$ the ratios of h_K to diameters of simplices in $\mathscr{T}_{\hat{h}}$ covering K are uniformly bounded by an absolute constant. In consequence, if $e = \partial K_i \cap \partial K_j$ is of positive measure, then $h_e \simeq h_{K_i} \simeq h_{K_j}$. We shall refer to $\mathscr{T}_h$ as the "fine mesh". Throughout the paper we will assume that the fine mesh is chosen in such a way that $\rho_{|_K}$ is already constant for all $K \in \mathscr{T}_h$.

We define the finite element space V_h^p in which problem (2) is approximated,

$$V_h^p = \{v \in L^2(\Omega) : v_{|_K} \in \mathbb{P}_{p_K} \text{ for } K \in \mathscr{T}_h\} \tag{3}$$

where $\mathbb{P}_{p_K}$ denotes the set of polynomials of degree not greater than p_K. We shall assume that $1 \leq p_K$ and that polynomial degrees have bounded local variation, that is, if $e = \partial K_i \cap \partial K_j \in \mathscr{E}_h^{\mathrm{in}}$, then $p_{K_i} \simeq p_{K_j}$.

Next, we discretize (2) by the symmetric weighted interior penalty discontinuous Galerkin method, see for example [3], [1]:

Find $u^* \in V_h^p$ such that

$$\mathscr{A}_h^p(u^*, v) = (f, v)_\Omega, \qquad \forall v \in V_h^p, \tag{4}$$

where

$$\mathscr{A}_h^p(u, v) = A_h^p(u, v) - F_h^p(u, v) - F_h^p(v, u)$$

and

$$A_h^p(u, v) = \sum_{K \in \mathscr{T}_h} (\rho \nabla u, \nabla v)_K + \sum_{e \in \mathscr{E}_h} \langle \gamma [u], [v] \rangle_e, \qquad F_h^p(u, v) = \sum_{e \in \mathscr{E}_h} \langle \{\rho \nabla u\}, [v] \rangle_e.$$

Here for $K \in \mathscr{T}_h$ and $e \in \mathscr{E}_h$ we use standard notation: $(u, v)_K = \int_K u\, v\, dx$ and $\langle u, v \rangle_e = \int_e u\, v\, d\sigma$. On $e \in \mathscr{E}_h^{\mathrm{in}}$ such that $e = \partial K_i \cap \partial K_j$ we set

$$\{\rho \nabla u\} = \overline{\rho}(\nabla u_{|_{K_i}} + \nabla u_{|_{K_j}}), \qquad [u] = u_{|_{K_i}} n_{|_{K_i}} + u_{|_{K_j}} n_{|_{K_j}},$$

with

$$\overline{\rho} = \frac{\rho_{|_{K_i}} \rho_{|_{K_j}}}{\rho_{|_{K_i}} + \rho_{|_{K_j}}}, \qquad \underline{h} = \min\{h_{K_i}, h_{K_j}\}, \qquad \overline{p} = \max\{p_{K_i}, p_{K_j}\}, \qquad \gamma = \frac{\overline{\rho}\, \overline{p}^2}{\underline{h}}\, \delta,$$

where $\delta > 0$ is a prescribed constant. The unit normal vector pointing outward K_i is denoted by $n_{|K_i}$. On e which lies on the boundary of Ω and belongs to a face of K_i, we set $\{\rho\nabla u\} = \rho_{|K_i}\nabla u_{|K_i}$, $[u] = u_{|K_i}n_{|K_i}$ and $\gamma = \rho_{|K_i}p_{K_i}\delta/h_{K_i}$.

For sufficiently large penalty constant δ the discrete problem (4) is well–defined, therefore we can define a norm $|||u|||_\Omega$ by the identity $|||u|||^2_\Omega = A^p_h(u,u)$.

3 Nonoverlapping two-level, three-grid additive Schwarz method

Let us introduce the subdomain grid $\mathscr{T}_H$ as a partition of Ω into N_H disjoint open polygons (polyhedrons in 3-D) Ω_i, $i = 1,\ldots,N_H$, such that $\bar{\Omega} = \bigcup_{i=1,\ldots,N_H}\bar{\Omega}_i$ and that each Ω_i is a union of certain elements from the fine mesh $\mathscr{T}_h$. We shall retain the common notion of "subdomains" while referring to elements of $\mathscr{T}_H$. We set $H_i = \operatorname{diam}(\Omega_i)$ and $H = (H_1,\ldots,H_{N_H})$. We assume that there exists a reference simply-connected polygonal (polyhedral in 3-D) domain $\hat{\Omega} \subset R^d$ with Lipschitz boundary, such that every Ω_i is affinely homeomorphic to $\hat{\Omega}$ and the aspect ratios of Ω_i are bounded independently of h and H. Moreover, we assume that the number of neighboring regions in $\mathscr{T}_H$ is uniformly bounded by an absolute constant $\mathscr{N}$.

Next, let $\mathscr{T}_\mathscr{H}$ be a shape-regular affine triangulation by triangles in 2-D or tetrahedrons in 3-D, with diameter $\mathscr{H}$. We denote the elements of $\mathscr{T}_\mathscr{H}$ by D_n, $n = 1,\ldots,N_\mathscr{H}$. We shall call this partition the "coarse grid" and assume:

$$\rho_{|D_n} = \rho_n \text{ is a constant for each } D_n \in \mathscr{T}_\mathscr{H}.$$

We clearly have $N_\mathscr{H} \le N_H \le N_h$ and $\mathscr{T}_\mathscr{H} \subseteq \mathscr{T}_H \subseteq \mathscr{T}_h$ (inclusions understood in the sense of subsequent refinements of the coarsest partitioning), and $\max h \le \max H \le \mathscr{H}$. We define the additive Schwarz method following [1] and [4], by introducing the following decomposition of V^p_h:

$$V^p_h = V_0 + \sum_{i=1}^{N_H} V_i, \tag{5}$$

where the coarse space consists of functions which are polynomials inside each element of the coarse grid:

$$V_0 = \{v \in V^p_h : v_{|D_n} \in \mathbb{P}_q \text{ for all } n = 1,\ldots,N_\mathscr{H}\} \tag{6}$$

where $1 \le q \le \min\{p_K : K \in \mathscr{T}_h\}$. Next, for $i = 1,\ldots,N_H$ we define

$$V_i = \{v \in V^p_h : v_{|\Omega_j} = 0 \text{ for all } j \ne i\}.$$

One can view V_0 as a rough approximation to V^p_h (using coarser grid and lower order polynomials), cf. condition (9), while V_i can be thought of as V^p_h restricted to Ω_i, extended by zero elsewhere. Note that V^p_h already is a direct sum of spaces $V_1,\ldots,N_H$

and when $\mathscr{T}_{\mathscr{H}} = \mathscr{T}_H$, this decomposition coincides with [1]. Using decomposition (5) we define, for $i = 1, \ldots, N_H$, subdomain solvers $T_i : V_h^p \to V_i$, by

$$A_h^p(T_i u, v) = \mathscr{A}_h^p(u, v) \qquad \forall v \in V_i,$$

so that on each subdomain one has to solve only a relatively small system of linear equations (a "local problem") for $u_i = T_i u|_{\Omega_i}$. These problems are independent one from another, so can be solved in parallel. The coarse solve operator is $T_0 : V_h^p \to V_0$ defined analogously as $A_h^p(T_0 u, v_0) = \mathscr{A}_h^p(u, v_0)$ for all $v_0 \in V_0$. The preconditioned operator is

$$T = T_0 + \sum_{i=1}^{N_H} T_i. \tag{7}$$

Obviously, T is symmetric with respect to $\mathscr{A}_h^p(\cdot,\cdot)$. For D_n in $\mathscr{T}_{\mathscr{H}}$ let us define an auxiliary seminorm

$$|||u|||_{D_n,\mathrm{in}}^2 = \sum_{K \in \mathscr{T}_h(D_n)} \rho |\nabla u|_{0,K}^2 + \sum_{e \in \mathscr{E}_h^{\mathrm{in}}(D_n)} \gamma |[u]|_{0,e}^2, \tag{8}$$

where $\mathscr{E}_h^{\mathrm{in}}(D_n) = \{e \in \mathscr{E}_h : e \subset \bar{D}_n \setminus \partial D_n\}$.

Lemma 1 (see [5]). *Assume that V_0 has the following approximation property:*

$$\forall u \in V_h^p \quad \exists u^{(0)} \in V_0 : \quad \sum_{n=1}^{N_{\mathscr{H}}} \left(\frac{\rho_n q^2}{\mathscr{H}^2} |u - u^{(0)}|_{0,D_n}^2 + |||u - u^{(0)}|||_{D_n,\mathrm{in}}^2 \right) \lesssim \mathscr{A}_h^p(u,u). \tag{9}$$

Then the operator T defined in (7) *satisfies the inequalities*

$$\beta^{-1} \mathscr{A}_h^p(u,u) \lesssim \mathscr{A}_h^p(Tu,u) \lesssim \mathscr{A}_h^p(u,u) \qquad \forall u \in V_h^p,$$

where

$$\beta = \frac{\mathscr{H}^2}{q} \max_{n=1,\ldots,N_H} \left\{ \frac{\overline{p}_i^2}{\underline{h}_i H_i} \right\} \tag{10}$$

with $\underline{h}_i = \min\{h_K : K \in \mathscr{T}_h(\Omega_i)\}$ and $\overline{p}_i = \max\{p_K : K \in \mathscr{T}_h(\Omega_i)\}$.

Theorem 1. *Let us assume that there holds the following H^2–stability property: for every $g \in L^2$ the solution $z \in H_0^1(\Omega)$ of the problem*

$$-\operatorname{div}(\rho \nabla z) = \rho g \tag{11}$$

belongs to $H^2(\Omega)$ and $\sum_{n=1}^{N_{\mathscr{H}}} \rho_n ||z||_{2,D_n}^2 \lesssim \sum_{n=1}^{N_{\mathscr{H}}} \rho_n |g|_{0,D_n}^2$ with constant independent of g. Then $\operatorname{cond}(T) = O(\beta)$ where β is as in (10).

Proof. We will show that the assumptions of Lemma 1 are satisfied. The proof will extend the tools from [2] to the case of discontinuous coefficient; see also [6]. Let us define the lifting operator $R : L^2(\mathscr{E}_h) \to V_h^p$ by

$$(\rho R(\phi), w) = \sum_{e\in\mathscr{E}_h} \langle \{\rho w\}, \phi\rangle_e \qquad \forall w \in V_h^p$$

and the discrete gradient of $u \in V_h^p$ as $G(u) = \nabla_h u - R([u])$. Note that

$$(\rho R([u]), R([u])) = \sum_{e\in\mathscr{E}_h} \langle \{\rho R([u])\}, [u]\rangle_e \lesssim \sum_{e\in\mathscr{E}_h} \frac{\underline{h}^{1/2}}{\overline{p}} |\rho^{1/2} R([u])|_{0,e} \cdot \frac{\overline{p}}{\underline{h}^{1/2}} |\overline{\rho}^{1/2}[u]|_{0,e},$$

so by trace inequality $(\rho R([u]), R([u])) \lesssim |\rho^{1/2} R([u])|_{0,\Omega} \cdot \sum_{e\in\mathscr{E}_h} \langle \gamma[u], [u]\rangle_e$, from which we conclude stability estimate

$$|\rho^{1/2} R([u])|^2_{0,\Omega} \lesssim \sum_{e\in\mathscr{E}_h} \langle \gamma[u], [u]\rangle_e \qquad \forall u \in V_h^p. \tag{12}$$

Let $U \in H^1_0(\Omega)$ solve the problem

$$(\rho \nabla U, \nabla w)_\Omega = (\rho G(u), \nabla w)_\Omega \qquad \forall w \in H^1_0(\Omega).$$

From the definition of U and mentioned above property of the lifting operator R it directly follows that

$$|\rho^{1/2} \nabla U|_{0,\Omega} \lesssim |||u|||. \tag{13}$$

In order to prove (9) we estimate separately

$$\sum_{n=1}^{N_{\mathscr{H}}} |||u - u^{(0)}|||^2_{D_n,\mathrm{in}} \lesssim \sum_{n=1}^{N_{\mathscr{H}}} |||u - U|||^2_{D_n,\mathrm{in}} + \sum_{n=1}^{N_{\mathscr{H}}} |||U - u^{(0)}|||^2_{D_n,\mathrm{in}} = I_1 + I_2$$

and

$$\sum_{n=1}^{N_{\mathscr{H}}} \rho_n |u - u^{(0)}|^2_{0,D_n} \lesssim \sum_{n=1}^{N_{\mathscr{H}}} \rho_n |u - U|^2_{0,D_n} + \sum_{n=1}^{N_{\mathscr{H}}} \rho_n |U - u^{(0)}|^2_{0,D_n} = I_3 + I_4.$$

Clearly, $I_1 \lesssim |||u|||^2 + |||U|||^2 = |||u|||^2 + |\rho^{1/2}\nabla U|^2_{0,\Omega} \lesssim |||u|||^2$ by (13). In order to bound I_3, we use a variant of Aubin–Nitsche trick [2], which is the reason for our H^2–stability assumption. Let us define $z \in H^1_0(\Omega)$ as in (11) with $g = u - U$. After multiplying (11) by $(u - U)$ and integrating by parts on each fine grid element K, we sum over all $K \in \mathscr{T}_h$; using the definition of R we arrive after some calculations at

$$I_3 = |\rho^{1/2}(u-U)|^2_{0,\Omega} = \sum_{e\in\mathscr{E}_h} \langle \{\rho\nabla(z_h - z)\}, [u]\rangle_e + (\rho\nabla(z - z_h), R([u]))_\Omega = I_5 + I_6$$

for any $z_h \in V_h^p$. Applying Schwarz inequality first and then choosing z_h as the approximation to z in V_h^p we have, by the approximation property of V_h^p (cf. e.g. [2, eq. (13)],

$$I_6 \lesssim |\rho^{1/2} R([u])|_{0,\Omega} \cdot |\rho^{1/2} \nabla(z - z_h)|_{0,\Omega}$$
$$\lesssim |||u|||(\sum_{K\in\mathscr{T}_h} \rho \frac{h_K^2}{p_K^2} ||z||_{2,K}^2)^{1/2} \lesssim |||u||| \frac{\mathscr{H}}{q} (\sum_{n=1}^{N_{\mathscr{H}}} \rho_n ||z||_{2,D_n}^2)^{1/2},$$

so from H^2–stability assumption we conclude that $I_6 \lesssim |||u||| \cdot \frac{\mathscr{H}}{q} |\rho^{1/2}(u-U)|_{0,\Omega}$. In a similar way we obtain $I_5 \lesssim |||u||| \cdot \frac{\mathscr{H}}{q} |\rho^{1/2}(u-U)|_{0,\Omega}$, whence $I_3 \lesssim \frac{\mathscr{H}}{q} |||u|||$.

Finally, we bound the terms I_2 and I_4 in a standard way, by choosing $u^{(0)}$ on each D_n as the q–th order polynomial interpolant of $U_{|_{D_n}}$. See [5, Corollary 2] for details.

4 Numerical experiments

The H^2–stability requirement in Theorem 1 is quite limiting. As the following experimental results indicate, the preconditioner works well for checkerboard distribution of the coefficient, so there is room to relax assumptions Theorem 1.

Let us choose $\Omega = (0,1)^2$. We divide Ω into $N_{\mathscr{H}} = 2^{\mathscr{M}} \times 2^{\mathscr{M}}$ squares D_n ($n = 1,\ldots,N_{\mathscr{H}}$) of equal size. Let ρ be constant on a 2×2 grid with checkerboard distribution: $\rho = 1$ in "white" squares and $\rho = \rho_R$ (specified later) in "red" squares. For simplicity we choose $\mathscr{T}_H = \mathscr{T}_{\mathscr{H}}$, refined into a uniform fine triangulation $\mathscr{T}_h$ based on a square $2^m \times 2^m$ grid, with each square split into two triangles of identical shape. We discretize problem (2) on the fine mesh $\mathscr{T}_h$ using (4) with equal polynomial degree p across all elements in $\mathscr{T}_h$ and with $\delta = 7$. For the coarse problem, we use polynomials of degree q.

We report the number of Preconditioned Conjugate Gradient iterations (with zero as the initial guess) for operator T, required to reduce the initial norm of the preconditioned residual by a factor of 10^8 and (in parentheses) the condition number of T estimated from the PCG convergence history. We set the coefficients of the discrete solution u^* as random numbers from uniform distribution and construct f such that (4) holds.

$q \rightarrow$ $\rho_R \downarrow$	1	2	3	4	5
10^0	90 (166)	72 (96)	64 (70)	57 (56)	54 (47)
10^8	89 (155)	69 (94)	63 (71)	57 (55)	53 (48)

Table 1 Dependence of the number of iterations and the condition number (in parentheses) on the contrast ratio ρ_R and the coarse space polynomial degree q. Fixed $p = 6$, $\mathscr{M} = 2$, $m = 4$.

From Table 1 it is clear the converegence rate is independent from the jump of the coefficient and the improvement of the condition number due to increase of q is diminishing roughly like $O(1/q)$. Table 2 confirms that the condition number

$p \rightarrow$ $m \downarrow$	2	3	4	5	6
3	26 (11)	37 (22)	47 (37)	58 (57)	67 (78)
4	36 (20)	50 (42)	62 (72)	75 (112)	83 (149)
5	48 (38)	65 (79)	81 (140)	98 (219)	113 (303)

Table 2 Dependence of the number of iterations and the condition number (in parentheses) on the fine mesh size $h = 2^{-m}$ and polynomial degree p. Fixed $q = 1$, $\mathscr{M} = 2$ and $\rho_R = 10^4$.

$\mathscr{M} \rightarrow$ $m \downarrow$	2	3	4	5
3	47 (37)	38 (20)		
4	62 (72)	49 (39)	38 (20)	
5	81 (140)	65 (75)	50 (39)	38 (20)

Table 3 Dependence of the number of iterations and the condition number (in parentheses) on $\mathscr{H} = H = 2^{-\mathscr{M}}$ and $h = 2^{-m}$. Fixed $p = 4$, $q = 1$, $\rho_R = 10^4$.

dependence on p and h behaves approximately like $O(p^2/h)$. For varying h and $\mathscr{H} = H$, an $O(\mathscr{H}/h)$ dependence of the condition number is verified in Table 3. See [5] for more experimental results.

Acknowledgement

The author wishes to thank two anonymous referees whose comments and remarks helped to improve the paper substantially. This research has been partially supported by the Polish National Science Centre grant 2016/21/B/ST1/00350.

References

1. Paola F. Antonietti and Paul Houston. A class of domain decomposition preconditioners for *hp*-discontinuous Galerkin finite element methods. *J. Sci. Comput.*, 46(1):124–149, 2011.
2. Paola F. Antonietti, Paul Houston, and Iain Smears. A note on optimal spectral bounds for nonoverlapping domain decomposition preconditioners for *hp*-version Discontinuous Galerkin method. *Int. J. Numer. Anal. Model.*, 13(4):513–524, 2016.
3. Maksymilian Dryja. On discontinuous Galerkin methods for elliptic problems with discontinuous coefficients. *Comput. Methods Appl. Math.*, 3(1):76–85 (electronic), 2003.
4. Maksymilian Dryja and Piotr Krzyżanowski. A massively parallel nonoverlapping additive Schwarz method for discontinuous Galerkin discretization of elliptic problems. *Num. Math.*, 132(2):347–367, 2015.
5. Piotr Krzyżanowski. On a nonoverlapping additive Schwarz method for *h*-*p* discontinuous Galerkin discretization of elliptic problems. *Num. Meth. PDEs*, 32(6):15721590, 2016.
6. Iain Smears. Nonoverlapping domain decomposition preconditioners for discontinuous galerkin approximations of hamilton–jacobi–bellman equations. *Journal of Scientific Computing*, pages 1–30, 2017.

Adaptive deluxe BDDC Mixed and Hybrid Primal Discretizations

Alexandre Madureira[1] and Marcus Sarkis[2]

1 Summary

Major progress has been made recently to make FETI-DP and BDDC preconditioners robust with respect to any variation of coefficients inside and/or across the subdomains. A reason for this success is the adaptive selection of primal constraints technique based on local generalized eigenvalue problems. Here we introduce a mathematical framework to transfer this technique to the field of discretizations. We design discretizations where the number of degrees of freedom is the number of primal constraints on the coarse triangulation and associated basis functions are built on the fine mesh and with a priori energy error estimates independent of the contrast of the coefficients.

2 Hybrid Primal Formulation

Consider the problem of finding the weak solution $u : \Omega \to \mathbb{R}$ of

$$\begin{aligned} -\operatorname{div} \rho \boldsymbol{\nabla} u = \rho g = f \quad \text{in } \Omega, \\ u = 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ for $d = 2$ or 3 is an open bounded connected domain with polyhedral boundary $\partial\Omega$, the coefficient ρ satisfies $0 < \rho_{\min} \leq \rho(x) \leq \rho_{\max}$ and g is a given forcing data. Define the ρ-weighted $L^2(\Omega)$-norm by $\|g\|_{L^2_\rho(\Omega)} = \|\rho^{1/2} g\|_{L^2(\Omega)}$ and the energy norm by $\|v\|_{H^1_\rho(\Omega)} = \|\rho^{1/2} \boldsymbol{\nabla} v\|_{L^2(\Omega)}$. We obtain the following stability result:

[1]Laboratório Nacional de Computação Científica, Brazil; supported by CNPq/Brazil
[2]Department of Mathematical Sciences, Worcester Polytechnic Institute, MA 01609; this work was supported by the National Science Foundation Grant DMS-1522663

$$\|u\|_{H^1_\rho(\Omega)} \leq C_P \|g\|_{L^2_\rho(\Omega)},$$

where C_P is the weighted Poincaré constant of $\|v\|_{L^2_\rho(\Omega)} \leq C_P |v|_{H^1_\rho(\Omega)}$ for all $v \in H^1_\rho(\Omega)$ vanishing on $\partial\Omega$.

We start by recasting the continuous problem in a weak formulation that depends on a polyhedral and regular mesh $\mathcal{T}_H$, which can be based on different geometries. Without loss of generality, we adopt above and in the remainder of the text, the terminology of three-dimensional domains, denoting for instance the boundaries of the elements by faces. For a given element $\tau \in \mathcal{T}_H$ let $\partial\tau$ denote its boundary and $\boldsymbol{n}^\tau$ the unit size normal vector that points outward τ. We denote by $\boldsymbol{n}$ the outward normal vector on $\partial\Omega$. Consider now the following spaces:

$$\begin{aligned} H^1(\mathcal{T}_H) &= \{v \in L^2(\Omega) : v|_\tau \in H^1(\tau),\, \tau \in \mathcal{T}_H\}, \\ \Lambda(\mathcal{T}_H) &= \left\{ \prod_{\tau\in\mathcal{T}_H} \boldsymbol{\tau}\cdot\boldsymbol{n}^\tau|_{\partial\tau} : \boldsymbol{\tau} \in H(\mathrm{div};\Omega) \right\} \subsetneq \prod_{\tau\in\mathcal{T}_H} H^{-1/2}(\partial\tau). \end{aligned} \tag{2}$$

For w, $v \in H^1(\mathcal{T}_H)$ and $\mu \in \Lambda(\mathcal{T}_H)$ define

$$(w,v)_{\mathcal{T}_H} = \sum_{\tau\in\mathcal{T}_H} \int_\tau wv \, d\boldsymbol{x} \qquad (\mu,v)_{\partial\mathcal{T}_H} = \sum_{\tau\in\mathcal{T}_H} (\mu,v)_{\partial\tau}, \tag{3}$$

where $(\cdot,\cdot)_{\partial\tau}$ is the dual product involving $H^{-1/2}(\partial\tau)$ and $H^{1/2}(\partial\tau)$. Then

$$(\mu,v)_{\partial\tau} = \int_\tau \mathrm{div}\,\boldsymbol{\sigma} v \, d\boldsymbol{x} + \int_\tau \boldsymbol{\sigma}\cdot\boldsymbol{\nabla} v \, d\boldsymbol{x}$$

for all $\boldsymbol{\sigma} \in H(\mathrm{div};\tau)$ such that $\boldsymbol{\sigma}\cdot\boldsymbol{n}^\tau = \mu$. We also define the norms

$$\begin{aligned} \|\boldsymbol{\sigma}\|^2_{H_\rho(\mathrm{div};\Omega)} &= \|\rho^{-1/2}\boldsymbol{\sigma}\|^2_{0,\Omega} + \|\rho^{-1/2}\,\mathrm{div}\,\boldsymbol{\sigma}\|^2_{0,\Omega}, \\ \|\mu\|_{H^{-1/2}_\rho(\mathcal{T}_H)} &= \inf_{\substack{\boldsymbol{\sigma}\in H(\mathrm{div};\Omega) \\ \boldsymbol{\sigma}\cdot\boldsymbol{n}^\tau = \mu \text{ on } \partial\tau,\, \tau\in\mathcal{T}_H}} \|\boldsymbol{\sigma}\|_{H_\rho(\mathrm{div};\Omega)}, \\ |v|^2_{H^1_\rho(\mathcal{T}_H)} &= \sum_{\tau\in\mathcal{T}_H} \|\rho^{1/2}\boldsymbol{\nabla} v\|^2_{0,\tau}. \end{aligned} \tag{4}$$

We use analogous definitions on subsets of $\mathcal{T}_H$, in particular when the subset consists of a single element τ (and in this case we write τ instead of $\{\tau\}$). We note that since $0 < \rho_{\min} \leq \rho(x) \leq \rho_{\max}$, the space $H_\rho(\mathrm{div};\Omega)$ and $H^1_\rho(\mathcal{T}_H)$ are equal to the spaces $H(\mathrm{div};\Omega)$ and $H^1(\mathcal{T}_H)$, respectively.

In the primal hybrid formulation [11], $u \in H^1(\mathcal{T}_H)$ and $\lambda \in \Lambda(\mathcal{T}_H)$ are such that

$$\begin{aligned} (\rho\boldsymbol{\nabla} u, \boldsymbol{\nabla} v)_{\mathcal{T}_H} - (\lambda, v)_{\partial\mathcal{T}_H} &= (\rho g, v)_{\mathcal{T}_H} \quad \text{for all } v \in H^1(\mathcal{T}_H), \\ (\mu, u)_{\partial\mathcal{T}_H} &= 0 \quad \text{for all } \mu \in \Lambda(\mathcal{T}_H). \end{aligned} \tag{5}$$

Following Theorem 1 of [11], it is possible to show that the solution (u, λ) of (5) is such that $u \in H^1(\Omega)$ and vanishing on $\partial\Omega$ satisfies (1) in the weak sense and $\lambda = \rho \boldsymbol{\nabla} u \cdot \boldsymbol{n}^\tau$ for all elements τ.

In the spirit of [11, 3] we consider the decomposition

$$H^1(\mathcal{T}_H) = \mathbb{P}^0(\mathcal{T}_H) \oplus \widetilde{H}^1(\mathcal{T}_H),$$

where $\mathbb{P}^0(\mathcal{T}_H)$ is the space of piecewise constants, and $\widetilde{H}^1(\mathcal{T}_H)$ is its $L^2_\rho(\tau)$ orthogonal complement, i.e., the space of functions with zero ρ-weighted average within each element $\tau \in \mathcal{T}_H$

$$\begin{aligned} \mathbb{P}^0(\mathcal{T}_H) &= \{v \in H^1(\mathcal{T}_H) : v|_\tau \text{ is constant, } \tau \in \mathcal{T}_H\}, \\ \widetilde{H}^1(\mathcal{T}_H) &= \{\tilde{v} \in H^1(\mathcal{T}_H) : \int_\tau \rho \tilde{v}\, d\boldsymbol{x} = 0, \tau \in \mathcal{T}_H\}. \end{aligned} \tag{6}$$

We then write $u = u^0 + \tilde{u}$, where $u^0 \in \mathbb{P}^0(\mathcal{T}_H)$ and $\tilde{u} \in \widetilde{H}^1(\mathcal{T}_H)$, and find from (5) that

$$\begin{aligned} (\rho \boldsymbol{\nabla} \tilde{u}, \boldsymbol{\nabla} \tilde{v})_{\mathcal{T}_H} - (\lambda, \tilde{v})_{\partial\mathcal{T}_H} &= (\rho g, \tilde{v})_{\mathcal{T}_H} && \text{for all } \tilde{v} \in \widetilde{H}^1(\mathcal{T}_H), \\ (\lambda, v^0)_{\partial\mathcal{T}_H} &= -(\rho g, v^0)_{\mathcal{T}_H} && \text{for all } v^0 \in \mathbb{P}^0(\mathcal{T}_H), \\ (\mu, u^0 + \tilde{u})_{\partial\mathcal{T}_H} &= 0 && \text{for all } \mu \in \Lambda(\mathcal{T}_H). \end{aligned} \tag{7}$$

Let $T : \Lambda(\mathcal{T}_H) \to \widetilde{H}^1(\mathcal{T}_H)$ and $\tilde{T} : L^2(\Omega) \to \widetilde{H}^1(\mathcal{T}_H)$ be such that, given $\tau \in \mathcal{T}_H$, $\mu \in \Lambda(\mathcal{T}_H)$ and $g \in L^2_\rho(\Omega)$, for all $\tilde{v} \in \widetilde{H}^1(\mathcal{T}_H)$ we have

$$\int_\tau \rho \boldsymbol{\nabla}(T\mu) \cdot \boldsymbol{\nabla} \tilde{v}\, d\boldsymbol{x} = (\mu, \tilde{v})_{\partial\tau}, \qquad \int_\tau \rho \boldsymbol{\nabla}(\tilde{T}g) \cdot \boldsymbol{\nabla} \tilde{v}\, d\boldsymbol{x} = (\rho g, \tilde{v})_\tau. \tag{8}$$

Note from the first equation of (7) that $\tilde{u} = T\lambda + \tilde{T}g$, and substituting in the other two equations of (7), we have that $u^0 \in \mathbb{P}^0(\mathcal{T}_H)$ and $\lambda \in \Lambda(\mathcal{T}_H)$ solve

$$\begin{aligned} (\mu, \gamma T\lambda)_{\partial\mathcal{T}_H} + (\mu, u^0)_{\partial\mathcal{T}_H} &= -(\mu, \gamma\tilde{T}g)_{\partial\mathcal{T}_H} && \text{for all } \mu \in \Lambda(\mathcal{T}_H), \\ (\lambda, v^0)_{\partial\mathcal{T}_H} &= -(\rho g, v^0)_{\mathcal{T}_H} && \text{for all } v^0 \in \mathbb{P}^0(\mathcal{T}_H). \end{aligned} \tag{9}$$

From now on we drop the trace operator γ.

We use the unknowns u^0 and λ to reconstruct the u as follows:

$$u = u^0 + \tilde{u} = u^0 + T\lambda + \tilde{T}g. \tag{10}$$

Unlike the HMM [3] and DEM [1], the methods we describe below approximate $\Lambda(\mathcal{T}_H)$ by multiscale basis functions with larger support and with the lowest global energy property which decay exponentially, achieving optimal energy approximation without requiring regularity of the problem.

3 Primal Hybrid Finite Element Methods

Let $\mathcal{F}_h$ be a partition of the faces of elements in $\mathcal{T}_H$, refining them in the sense that every (coarse) face of the elements in $\mathcal{T}_H$ can be written as a union of faces of $\mathcal{F}_h$. Let $\Lambda_h \subset \Lambda(\mathcal{T}_H)$ be the space of piecewise constants on $\mathcal{F}_h$, i.e.,

$$\Lambda_h = \{\mu_h \in \Lambda(\mathcal{T}_H) : \mu_h|_{F_h} \text{ is constant on each face } F_h \in \mathcal{F}_h\}.$$

For simplicity, we do not discretize $H^1(\tau)$ and $H(\mathrm{div};\tau)$ for $\tau \in \mathcal{T}_H$. We remark that the methods develop here extend easily when we discretize $H(\mathrm{div};\tau)$ by simplices or cubical elements with lowest order Raviart–Thomas spaces or discretize $H^1(\tau)$ fine enough to resolve the heterogeneities of $\rho(x)$ and to satisfy inf-sup conditions with respect to the space Λ_h.

We then pose the problem of finding $u_h^0 \in \mathbb{P}^0(\mathcal{T}_H)$ and $\lambda_h \in \Lambda_h$ such that

$$\begin{aligned} (\mu_h, T\lambda_h)_{\partial\mathcal{T}_H} + (\mu_h, u_h^0)_{\partial\mathcal{T}_H} &= -(\mu_h, \tilde{T}g)_{\partial\mathcal{T}_H} \quad \text{for all } \mu_h \in \Lambda_h, \\ (\lambda_h, v^0)_{\partial\mathcal{T}_H} &= -(\rho g, v_h^0)_{\mathcal{T}_H} \quad \text{for all } v_h^0 \in \mathbb{P}^0(\mathcal{T}_H). \end{aligned} \tag{11}$$

We note that T restricted to τ, denoted by $T^\tau : \Lambda_h^\tau \to \widetilde{H}^1(\tau)$ solves

$$(\rho\, \boldsymbol{\nabla}(T^\tau \mu_h^\tau), \boldsymbol{\nabla}\, v)_\tau = (\mu_h^\tau, v)_{\partial\tau} \quad \text{for all } v \in \widetilde{H}^1(\tau),$$

and note that $\rho\, \boldsymbol{\nabla}(T^\tau \mu_h^\tau) \cdot \boldsymbol{n}^\tau = \mu_h$ on $\partial\tau$. Note also that $(\mu_h, T\mu_h)_{\partial\mathcal{T}_H} = 0$ implies $T\mu_h = 0$ and $\mu_h = 0$. As (11) is finite dimensional, it is well-posed since it is injective. We define our approximation as in (10), by

$$u_h = u_h^0 + T\lambda_h + \tilde{T}g. \tag{12}$$

Simple substitutions yield u_h, λ_h solve (5) if $\Lambda(\mathcal{T}_H)$ is replaced by Λ_h, i.e.,

$$\begin{aligned} (\rho\, \boldsymbol{\nabla}\, u_h, \boldsymbol{\nabla}\, v)_{\mathcal{T}_H} - (\lambda_h, v)_{\partial\mathcal{T}_H} &= (g, v)_{\mathcal{T}_H} \quad \text{for all } v \in H^1(\mathcal{T}_H), \\ (\mu_h, u_h)_{\partial\mathcal{T}_H} &= 0 \quad \text{for all } \mu_h \in \Lambda_h. \end{aligned}$$

We also assume that Λ_h is chosen fine enough so that

$$|u - u_h|^2_{H^1_\rho(\mathcal{T}_H)} = \big(\lambda - \lambda_h, T(\lambda - \lambda_h)\big)_{\mathcal{T}_H} \leq \widetilde{\mathcal{H}}^2 \|g\|^2_{L^2_\rho(\Omega)},$$

where $\widetilde{\mathcal{H}}$ represents a "target precision" the method should achieve. For instance, one could choose $\widetilde{\mathcal{H}} = H$ or $\widetilde{\mathcal{H}} = h^s$ for some $0 < s \leq 1$. It must be mentioned that λ_h is never computed, only an approximation of order $\widetilde{\mathcal{H}}$.

Above, and in what follows, c denotes an arbitrary constant that does not depend on H, $\widetilde{\mathcal{H}}$, h, ρ. For details and proofs, see [6]. See also [7] for a related multiscale conforming method.

4 Adaptive BDDC Spectral Decomposition I

Let $\tau \in \mathcal{T}_H$, F a face of $\partial\tau$, and let $F_\tau^c = \partial\tau \backslash F$. Define

$$\Lambda_h^\tau = \{\mu_h|_{\partial\tau} : \mu_h \in \Lambda_h\}, \Lambda_h^F = \{\mu_h|_F : \mu_h \in \Lambda_h^\tau\}, \Lambda_h^{F_\tau^c} = \{\mu_h|_{F_\tau^c} : \mu_h \in \Lambda_h^\tau\}.$$

Denote $\mu_h^\tau = \{\mu_h^F, \mu_h^{F_\tau^c}\}$ with $\mu_h^\tau \in \Lambda_h^\tau$, $\mu_h^F \in \Lambda_h^F$ and $\mu_h^{F_\tau^c} \in \Lambda_h^{F_\tau^c}$, and define

$$T_{FF}^\tau : \Lambda_h^F \to (\Lambda_h^F)', \qquad T_{F^cF}^\tau : \Lambda_h^F \to (\Lambda_h^{F_\tau^c})'$$
$$T_{FF^c}^\tau : \Lambda_h^{F_\tau^c} \to (\Lambda_h^F)', \qquad T_{F^cF^c}^\tau : \Lambda_h^{F_\tau^c} \to (\Lambda_h^{F_\tau^c})',$$

and note that $(\mu_h, T^\tau \mu_h)_{\partial\tau} = (\mu_h^F, T_{FF}^\tau \mu_h^F)_F +$
$(\mu_h^F, T_{FF^c}^\tau \mu_h^{F_\tau^c})_F + (\mu_h^{F_\tau^c}, T_{F^cF}^\tau \mu_h^F)_{F_\tau^c} + (\mu_h^{F_\tau^c}, T_{F^cF^c}^\tau \mu_h^{F_\tau^c})_{F_\tau^c}$.

It follows from the properties of T^τ that T_{FF}^τ and $T_{F^cF^c}^\tau$ are symmetric and positive definite matrices, and follows by Schur complement arguments that

$$\begin{aligned}(\mu_h^F, T_{FF}^\tau \mu_h^F)_F &= (\{\mu_h^F, 0\}, T^\tau\{\mu_h^F, 0\})_{\partial\tau} \\ &\geq \min_{\nu_h^{F_\tau^c} \in \Lambda_h^{F_\tau^c}} (\{\mu_h^F, \nu_h^{F_\tau^c}\}, T^\tau\{\mu_h^F, \nu_h^{F_\tau^c}\})_{\partial\tau} = (\mu_h^F, \hat{T}_{FF}^\tau \mu_h^F)_F, \quad (13)\end{aligned}$$

$$\hat{T}_{FF}^\tau = T_{FF}^\tau - T_{FF^c}^\tau (T_{F^cF^c}^\tau)^{-1} T_{F^cF}^\tau$$

and the minimum is attained at $\nu_h^{F_\tau^c} = -(T_{F^cF^c}^\tau)^{-1} T_{F^cF}^\tau \mu_h^F$.

To take into account high-contrast coefficients, we consider the following generalized eigenvalue problem: Find $(\alpha_i^F, \mu_{i,h}^F) \in (\mathbb{R}, \Lambda_h^F)$ such that:

1. If the face F is shared by elements τ and τ' we solve

$$(\nu_h^F, (T_{FF}^\tau + T_{FF}^{\tau'})\mu_{h,i}^F)_F = \alpha_i^F (\nu_F, (\hat{T}_{FF}^\tau + \hat{T}_{FF}^{\tau'})\mu_{h,i}^F)_F, \quad \forall \nu_h^F \in \Lambda_h^F.$$

2. If the face F is on the boundary $\partial\Omega$ we solve

$$(\nu_h^F, T_{FF}^\tau \mu_{h,i}^F)_F = \alpha_i^F (\nu_h^F, \hat{T}_{FF}^\tau \mu_{h,i}^F)_F, \quad \forall \nu_h^F \in \Lambda_h^F.$$

The use of such generalized eigenvalue problems is known in the domain decomposition community as "adaptive selection of primal constraints". It is used to make preconditioners robust with respect to coefficients; see [9, 12] for RT_0 and BDM_1 where only face eigenvalue problems for two- as well as for three-dimensional problems. Here, we apply this technique to design robust discretizations; see [4, 7] on related work for classical FEM discretizations.

Now we decompose $\Lambda_h^F := \Lambda_h^{F,\triangle} \oplus \Lambda_h^{F,\Pi}$ where

$$\Lambda_h^{F,\triangle} := \mathrm{span}\{\mu_{h,i}^F : \alpha_i^F < \alpha_*\}, \qquad \Lambda_h^{F,\Pi} := \mathrm{span}\{\mu_{h,i}^F : \alpha_i^F \geq \alpha_*\}.$$

From (13) we know that $\alpha_i^F \geq 1$. The parameter α_* is defined by the user and it controls how fast is the exponential decay of the multiscale basis functions. We point out that the dimension of the space $\Lambda_h^{F,\Pi}$ is related to the number of connected subregions on $\bar{\tau} \cup \bar{\tau}'$ with large coefficients surrounded by regions with small coefficients. Finally, let $\Lambda_h = \Lambda_h^{\Pi} \oplus \Lambda_h^{\triangle}$, where

$$\begin{aligned} \Lambda_h^{\Pi} &:= \{\mu_h \in \Lambda_h : \mu_h|_F \in \Lambda_h^{F,\Pi} \text{ for all } F \in \partial\mathcal{T}_H\}, \\ \Lambda_h^{\triangle} &:= \{\mu_h \in \Lambda_h : \mu_h|_F \in \Lambda_h^{F,\triangle} \text{ for all } F \in \partial\mathcal{T}_H\}. \end{aligned} \tag{14}$$

5 NLSD-Nonlocalized Spectral Decomposition Method I

Define the operator $P : H^1(\Omega) \to \Lambda_h^{\triangle}$ such that for $w \in H^1(\mathcal{T}_H)$,

$$(\mu_h^{\triangle}, TPw)_{\partial\mathcal{T}_H} = (\mu_h^{\triangle}, w)_{\partial\mathcal{T}_H} \quad \text{for all } \mu_h^{\triangle} \in \Lambda_h^{\triangle}. \tag{15}$$

Let us decompose $\lambda_h = \lambda_h^{\Pi} + \lambda_h^{\triangle}$. We first eliminate $\lambda_h^{\triangle}$ from the first equation of (11) to obtain

$$\lambda_h^{\triangle} = -P(u_h^0 + T\lambda_h^{\Pi} + \tilde{T}g), \tag{16}$$

hence

$$u_h = (I - TP)u_h^0 + T(I - PT)\lambda_h^{\Pi} + (I - TP)\tilde{T}g). \tag{17}$$

Then using algebraic manipulations with (11) and (15) we find $u_h^0 \in \mathbb{P}^0(\mathcal{T}_H)$ and $\lambda_h^{\Pi} \in \Lambda_h^{\Pi}$ satisfy:

$$\begin{aligned} &(\hat{\mu}_h^{\Pi}, T\hat{\lambda}_h^{\Pi})_{\partial\mathcal{T}_H} + (\hat{\mu}_h^{\Pi}, \hat{u}_h^0)_{\partial\mathcal{T}_H} = -(\hat{\mu}_h^{\Pi}, \widehat{\tilde{T}g})_{\partial\mathcal{T}_H} \quad \text{for all } \mu_h^{\Pi} \in \Lambda_h^{\Pi} \\ &(\hat{\lambda}_h^{\Pi}, \hat{v}_h^0)_{\partial\mathcal{T}_H} - (Pu_0^h, v_0^h)_{\partial\mathcal{T}_H} \;\; = -(\rho\hat{g}, \hat{v}_h^0)_{\mathcal{T}_H} \qquad \text{for all } v_h^0 \in \mathbb{P}^0(\mathcal{T}_H), \end{aligned} \tag{18}$$

where the hat functions are non-local multiscale functions defined by

$$\hat{\lambda}_h^{\Pi} = (I - PT)\lambda_h^{\Pi}, \qquad \hat{\mu}_h^{\Pi} = (I - PT)\mu_h^{\Pi}, \qquad \hat{u}_h^0 = (I - TP)u_h^0,$$

$$\hat{v}_h^0 = (I - TP)v_h^0, \qquad \widehat{\tilde{T}g} = (I - TP)\tilde{T}g \quad \text{and} \quad \hat{g} = (I - P\tilde{T})g.$$

We note that the idea of performing global static condensation goes back to the Multiscale Variational Finite Element Method [5]. Recent variations of this method called Localized Orthogonal Decomposition Methods were introduced and analyzed in [10] and references therein. Some theoretical progresses for high-contrast were made in [5] for a class of coefficients and by using overlapping spectral decomposition introduced in [2]. Here in this paper no condition on the coefficient is imposed and the theoretical results are based on non-overlapping decomposition techniques.

5.1 NLSD Method II

In the splitting (17), the non-local term TPu_h^0 adds theoretical difficulties and more complexity on the implementation. We now introduce the Adaptive BDDC Spectral Decomposition II such that $Pu_h^0 = 0$. Indeed, first decompose $\Lambda_h = \Lambda_h^{RT} \oplus \tilde{\Lambda}_h^f$, where Λ_h^{RT} ($\tilde{\Lambda}_h^f$) is the space of constant (average zero) functions on each face F of $\mathcal{T}_H$. Further decompose $\tilde{\Lambda}_h^f = \tilde{\Lambda}_h^{f,\Pi} \oplus \tilde{\Lambda}_h^{f,\triangle}$ by solving the same generalized eigenvalue problem before however on $\tilde{\Lambda}_h^{f,F}$ rather than on Λ_h^F. Denote $\Lambda_h^\Pi = \Lambda_h^{RT} \oplus \tilde{\Lambda}_h^{f,\Pi}$ and $\Lambda_h^\triangle = \tilde{\Lambda}_h^{f,\triangle}$. Repeat the same algebraic steps as in Section 5 and use that $(\mu_h^\triangle, v_h^0)_{\partial\mathcal{T}_H} = 0$. This method is analyzed in [6].

6 LSD-Localized Spectral Decomposition Method II

We next show that the exponential decay of the multiscale basis functions is independently of the coefficient contrast. Hence, instead of building global multiscale basis functions we actually build local basis functions. Lemma 1 implies exponential decay of functions, such as $PT\mu_h^\Pi$ and Pv_h^0 when μ_h^Π and v_h^0 has local support, and Lemma 2 shows $T(P-P^j)v$ decreases exponentially.

For $K \in \mathcal{T}_H$, define $\mathcal{T}_0(K) = \emptyset$, $\mathcal{T}_1(K) = \{K\}$, and for $j = 1, 2, \ldots$ let

$$\mathcal{T}_{j+1}(K) = \{\tau \in \mathcal{T}_H : \overline{\tau} \cap \overline{\tau}_j \neq \emptyset \text{ for some } \tau_j \in \mathcal{T}_j(K)\}.$$

Lemma 1. *Let $v \in H^1(\mathcal{T}_H)$ such that* $\operatorname{supp} v \subset K$*, and $\mu_h^\triangle = Pv$. Then*

$$|T\mu_h^\triangle|^2_{H^1_\rho(\mathcal{T}_H \setminus \mathcal{T}_{j+1}(K))} \leq e^{-\frac{[(j+1)/2]}{1+d^2\alpha_*}} |T\mu_h^\triangle|^2_{H^1_\rho(\mathcal{T}_H)}.$$

We now localize Pv since it decays exponentially when v has local support. For each fixed K, j, let $\Lambda_h^{\triangle,K,j} \subset \Lambda_h^\triangle$ be the set of functions of $\Lambda_h^\triangle$ which vanish on faces of elements in $\mathcal{T}_H \setminus \mathcal{T}_j(K)$. We introduce the operator $P^{K,j} : H^1(\mathcal{T}_H) \to \Lambda_h^{\triangle,K,j}$ such that, for $v \in H^1(\mathcal{T}_H)$,

$$(\mu_h^\triangle, TP^{K,j}v)_{\partial\mathcal{T}_H} = (\mu_h^\triangle, v)_{\partial\mathcal{T}_H} \quad \text{for all } \mu_h^\triangle \in \Lambda_h^{\triangle,K,j}.$$

For $v \in H^1(\mathcal{T}_H)$ let v_K be equal to v on K and zero otherwise. We define then $P^j v \in \Lambda_h^\triangle$ by

$$P^j v = \sum_{K \in \mathcal{T}_H} P^{K,j} v_K. \tag{19}$$

Lemma 2. *Let $v \in H^1(\mathcal{T}_H)$ and P defined by* (15) *and P^j by* (19)*. Then*

$$|T(P - P^j)v|^2_{H^1_\rho(\mathcal{T}_H)} \leq cj^{2d}d^4\alpha_*^2 e^{-\frac{[(j-3)/2]}{1+d^2\alpha_*}} |v|^2_{H^1_\rho(\mathcal{T}_H)}.$$

We define the LSD methods by (18), (16) and (17) with P_j instead of P. Denote the solution by u_h^j. The follow lemma shows the localization error.

Theorem 1. *For the LSD II method, if* $j = c\left(4d^2\alpha_* \log(C_P/\widetilde{\mathcal{H}})\right)$ *then*

$$|u_h - u_h^j|_{H^1_\rho(\mathcal{T}_H)} \leq c\widetilde{\mathcal{H}}\|g\|_{L^2_\rho(\Omega)}.$$

References

1. C. Farhat, I. Harari and L. P. Franca, The discontinuous enrichment method, Comput. Methods Appl. Mech. Engrg. 190(48), (2001), 6455-6479.
2. J. Galvis and Y. Efendiev, Domain decomposition preconditioners for multiscale flows in high-contrast media, Multiscale Model. Simul. 8(4) (2010), 1461–1483.
3. C. Harder, A. Madureira and F. Valentin, A hybrid-mixed method for elasticity, ESAIM Math. Model. Numer. Anal. 50(2), (2016) 311–336.
4. A. Heinlein, U. Hetmaniuk, A. Klawonn and O. Rheinbach, The approximate component mode synthesis special finite element method in two dimensions: parallel implementation and numerical results. J. Comput. Appl. Math. 289 (2015), 116133.
5. T. J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods, Comput. Methods Appl. Mech. Engrg. 127 (1995), no. 1-4, 387401.
6. A. Madureira and M. Sarkis, Hybrid localized spectral decomposotion for multiscale problems, https://arxiv.org/pdf/1706.08941.pdf.
7. A. Madureira and M. Sarkis, Adaptive ACMS: A robust localized approximated component mode synthesis method, https://arxiv.org/pdf/1709.04044.pdf.
8. A. Målqvist, P. Henning and F. Hellman, Multiscale mixed finite elements, Discrete Contin. Dyn. Syst. Ser. S. 9(5), (2016) 1269–1298.
9. D-S Oh, O. B. Widlund, S. Zampini and C. R. Dohrmann, BDDC Algorithms with deluxe scaling and adaptive selection of primal constraints for Raviart-Thomas vector fields, Math. Comp., 87 (2018), 659-692.
10. D. Peterseim and R. Scheichl, Robust numerical upscaling of elliptic multiscale problems at high contrast, Comput. Methods Appl. Math. 16(4), (2016) 579–603.
11. P.-A. Raviart and J. M. Thomas, Primal hybrid finite element methods for 2nd order elliptic equations, Math. Comp. 31(138), (1977) 391–413.
12. S. Zampini, X. Tu, Multilevel balancing domain decomposition by constraints deluxe algorithms with adaptive coarse spaces for flow in porous media, SIAM J. Sci. Comput., 39(4), (2017) 1389–1415.

Additive Schwarz with vertex based adaptive coarse space for multiscale problems in 3D

Leszek Marcinkowski*[1] and Talal Rahman[2]

1 Introduction

The choice of coarse spaces play an important role in the design of fast and robust Schwarz methods for problems of multiscale nature. Standard methods with standard coarse spaces have often difficulties to solve such problems, and even fail to converge due to computing in the finite precision arithmetic. The purpose of this paper is to propose a robust coarse space, adaptively enriched, for solving second order elliptic problems in three dimensions with highly varying coefficients, using the standard finite element for the discretization and the overlapping additive Schwarz method as the preconditioner. The coefficient may have discontinuities both inside and across subdomains. The convergence of the proposed method, as presented in the paper, is independent of the distribution of the coefficient, as well as the jumps in the coefficients, when the coarse space is chosen large enough. For similar works on domain decomposition methods addressing such problems, we refer to Galvis and Efendiev (2010), Spillane et al (2014) and the references therein.

Additive Schwarz methods for solving elliptic problems discretized by the finite element, which was proposed over thirty years ago, have been studied extensively over the past decades, see Smith et al (1996), Toselli and Widlund (2005) for an overview. It is known in general that if the coefficients are discontinuous across subdomains but are varying moderately with in each subdomain, then the standard coarse spaces are enough to generate additive Schwarz methods which are robust with respect to those jumps, cf. e.g. Smith et al (1996); Toselli and Widlund (2005). This is however not true in

Faculty of Mathematics, University of Warsaw, Banacha 2, 02-097 Warszawa, Poland Leszek.Marcinkowski@mimuw.edu.pl · Faculty of Engineering and Science, Western Norway University of Applied Sciences, Inndalsveien 28, 5063 Bergen, Norway Talal.Rahman@hvl.no

* This work was partially supported by Polish Scientific Grant 2016/21/B/ST1/00350.

the case when the coefficients may be highly varying and discontinuous almost everywhere, the fact which has in recent years drawn several researchers' attraction, cf. e.g. Chartier et al (2003); Mandel and Sousedík (2007); Klawonn et al (2015, 2016b,a); Galvis and Efendiev (2010); Efendiev et al (2012a,b); Nataf et al (2010, 2011); Spillane et al (2014); Dolean et al (2012); Kim and Chung (2015); Kim et al (2017); Calvo and Widlund (2016).

In the present work, we extend some of the ideas presented in those papers, and propose to construct a coarse space based on the vertices of the subdomains and a two fold enrichment of the coarse space, which is done through solving two specially designed lower dimensional eigenvalue problems, one on each face common to two neighboring subdomains and one on each interior edge of the subdomains, and chosing the first few eigenfunctions corresponding to the bad eigenmodes. The analysis show that the condition number bound of the resulting system depends only on the threshold used to choose the bad eigenvalues.

The remainder of the paper is organized as follows: in Section 2 we introduce our differential problem, and its finite element discretization. In Section 3 a classical overlapping Additive Schwarz method is presented. Section 4 is devoted to the construction of our adaptive coarse space and Section 5 gives the theoretical bound for the condition number of the resulting system.

2 Discrete Problem

We consider the following elliptic boundary value problem: Find $u^* \in H_0^1(\Omega)$

$$\int_\Omega \alpha(x)\nabla u^* \nabla v \, dx = \int_\Omega fv \, dx, \qquad \forall v \in H_0^1(\Omega), \tag{1}$$

where $\alpha(x) \geq \alpha_0 > 0$ is the coefficient, Ω is a polyhedral domain in $\mathbb{R}^3$ and $f \in L^2(\Omega)$. Let $\mathcal{T}_h$ be the quasi-uniform triangulation of Ω consisting of closed tetrahedra such that $\bar{\Omega} = \bigcup_{K\in\mathcal{T}_h} K$. Let h_K denote the diameter of K, and $h = \max_{K\in\mathcal{T}_h} h_K$ the mesh parameter for the triangulation.

We will further assume that α is piecewise constant on T_h without any loss of generality. We assume that there exists a coarse nonoverlapping partitioning of Ω into open connected Lipschitz polytopes Ω_i, called structures, such that $\overline{\Omega} = \bigcup_{i=1}^N \overline{\Omega}_i$ and they are aligned with the fine triangulation, in other words a fine triangle of $\mathcal{T}_h$ can be contained in only one of the coarse substructures. For the simplicity of presentation, we further assume that these substructures form a coarse triangulation of the domain which is shape regular in the sense of Brenner and Sung (1999).

Let $\mathcal{F}_{ij}$ denote the open face common to subdomains Ω_i and Ω_j, and let $\mathcal{E}$ denote an open edge of a substructure, not in $\partial\Omega$. We denote with Ω_h,

$\partial\Omega_h$, Ω_{ih}, $\partial\Omega_{ih}$, $\mathcal{F}_{ij,h}$, and $\mathcal{E}_h$, the sets of vertices of the elements of $\mathcal{T}_h$, corresponding to Ω, $\partial\Omega$, Ω_i, $\partial\Omega_i$, $\mathcal{F}_{ij}$, and $\mathcal{E}$, respectively

Let S_h be the standard linear conforming finite element space defined on the triangulation $\mathcal{T}_h$,

$$S_h = S_h(\Omega) := \{u \in C(\overline{\Omega}) \cap H_0^1(\Omega) : \ v_{|K} \in P_1, \ K \in \mathcal{T}_h\}.$$

The finite element approximation u_h^* of (1) is then defined as the solution to the following problem: Find $u_h^* \in S_h$ such that

$$a(u_h^*, v) = (f, v), \quad \forall v \in S_h. \tag{2}$$

Note that α can be scaled without influencing the solution, hence we can easily assume that $\alpha(x) \geq 1$. As ∇u_h^* is piecewise constant over the fine elements, we can further assume that α is piecewise constants over the elements of $\mathcal{T}_h$, since $\int_K \alpha \nabla u \nabla v \, dx = (\nabla u)_{|K}(\nabla v)_{|K} \int_K \alpha(x) \, dx$.

Since each subdomain inherits a local triangulation $\mathcal{T}_h(\Omega_k)$ from $\mathcal{T}_h(\Omega)$, two local subspaces can be defined as the following,

$$S_h(\Omega_i) := \{u_{|\overline{\Omega}_i} : u \in S_h\} \ \text{ and } \ S_{h,0}(\Omega_i) := S_h(\Omega_i) \cap H_0^1(\Omega_i),$$

along with a local projection operator $\mathcal{P}_i : S_h \to S_{h,0}(\Omega_i)$ as the following, find $\mathcal{P}_i u \in S_{h,0}(\Omega_i)$ such that

$$a_i(\mathcal{P}_i u, v) = a_i(u, v), \quad \forall v \in S_{h,0}(\Omega_i),$$

where $a_i(u, v) := a_{|\Omega_i}(u, v) = \int_{\Omega_i} \alpha(x) \nabla u \nabla v \, dx$.

The discrete harmonic part of $u \in S_h(\Omega_i)$ is defined as $\mathcal{H}_i u := u - \mathcal{P}_i u$, or equivalently as $\mathcal{H}_i u \in S_h(\Omega_i)$ which satisfies the following,

$$\begin{cases} a_i(\mathcal{H}_i u, v) = 0, \ \forall v \in S_{h,0}(\Omega_i), \\ \mathcal{H}_i u(s) = u(s), \quad \forall s \in \partial\Omega_{ih}. \end{cases} \tag{3}$$

We say that a function $u \in S_h$ is discrete harmonic if it is discrete harmonic in each subdomain, i.e. $u_{|\Omega_i} = \mathcal{H}_i u_{|\Omega_i} \ \forall i$.

3 Additive Schwarz Method

In this section, we present the overlapping additive Schwarz method for the discrete problem (2). We refer to Smith et al (1996); Toselli and Widlund (2005) for a more general discussion of the method.

Decomposition of S_h

The space S_h is decomposed into the local subspaces $\{V_i\}_i$, and the global coarse space V_0, as follows.

$$V_i = \{u \in S_h : \ v(x) = 0 \ \ \forall x \in \Omega_h \setminus \overline{\Omega}_i\}, \quad i = 1, \ldots, N,$$

where $u \in V_i$ can take nonzero values at the nodes that are in Ω_i and on $\partial\Omega_i$ only, giving $\{V_i\}_i$ as subspaces with minimal overlap. The global coarse space V_0 is defined in Section 4. For $i = 0, \ldots, N$, the projection like operators $T_i \colon S_h \to V_i$ are defined as

$$a(T_i u, v) = a(u, v), \quad \forall v \in V_i. \tag{4}$$

Now, introducing the additive Schwarz operator as $T := T_0 + \sum_{i=1}^{N} T_i$, the original problem (2) can be replaced with the following equivalent problem: Find u_h^* such that

$$T u_h^* = g, \tag{5}$$

where $g = \sum_{i=0}^{N} g_i$ and $g_i = T_i u$. Note that g_i may be computed without knowing the solution u_h^* of (2): $a(g_i, v) = (f, v)$ for all $v \in V_i$.

4 Adaptive vertex coarse space

We introduce our adaptive vertex based coarse space in this section. Each edge $\mathcal{E}$ inherits a 1D triangulation $\mathcal{T}_h(\mathcal{E})$ from $\mathcal{T}_h$. For each edge $\mathcal{E}_h$, let $S_h(\mathcal{E})$ be the space of traces of functions of S_h on the edge, that is the space of continuous piecewise linear functions on $\mathcal{T}_h(\mathcal{E})$, let $S_{h,0}(\mathcal{E}) = S_h(\mathcal{E}) \cap H_0^1(\mathcal{E})$ be its subspace with compact support, and let the edge bilinear form $a_{\mathcal{E}}(u, v) : S_{h,0}(\mathcal{E}) \times S_{h,0}(\mathcal{E}) \to \mathbb{R}$ be defined as

$$a_{\mathcal{E}}(u, v) = \sum_{e \in \mathcal{T}_h(\mathcal{E})} \int_e \overline{\alpha}_e u' v' \, ds, \tag{6}$$

where $\overline{\alpha}_e = \max_{e \subset \partial K} \alpha_K$ is the maximum value of the coefficient over the tetrahedra sharing the fine edge $e \in \mathcal{T}_h(\mathcal{E})$. Here u', v' are the weak derivatives of $u, v \in S_{h,0}(\mathcal{E})$. The definition of the form $a_{\mathcal{E}}(u, v)$, in particular the definition of $\overline{\alpha}$, is introduced in a way which enables us to estimate this form from above by the sum of energy norms over all subdomains which share this edge.

4.1 Vertex based interpolation operator

We introduce the vertex interpolation operator $I_V : S_h(\Omega) \to S_h(\Omega)$ as follows. For $u \in S_h(\Omega)$

- $I_V u(x) = u(x)$ where x is a crosspoint (a subdomain vertex inside Ω),
- $I_V u$ on each edge $\mathcal{E}$ satisfies, cf. (6):

$$a_{\mathcal{E}}(I_V u, v) = 0, \quad \forall v \in S_{h,0}(\mathcal{E}). \tag{7}$$

- $I_V u(x) = 0$ at all $x \in \mathcal{F}_{ij,h}$ for each face $\mathcal{F}_{ij}$,
- $I_V u$ is discrete harmonic in the sense as described in Section 2.

Note that $I_V u$ is uniquely determined by the values of u at the crosspoints, as (7) uniquely determines $I_V u$ at the edge interior nodes, $I_V u$ is equal to zero at all face interior nodes, and then extended as discrete harmonic to the subdomain interior nodes, cf. (3). The auxiliary coarse space $\hat{V}_0$ is then defined as the image of this interpolation operator I_V, that is $\hat{V}_0 := Im(I_V) = I_V S_h$.The coarse space V_0 is the algebraic sum of $\hat{V}_0$ and a sequence of small subspaces built with functions that are extensions of certain eigenfunctions of the two particular classes of eigenvalue problems presented below.

4.2 Eigenvalue problems

We start by introducing the two classes of local eigenvalue problems, one on the subdomain edges or the edge interfaces, and one on the subdomain faces or the face interfaces.

Eigenvalue problem on edge interface

Find the eigen pairs $(\lambda_j^{\mathcal{E}}, \psi_j^{\mathcal{E}}) \in \mathbb{R}_+ \times S_{h,0}(\mathcal{E})$

$$a_{\mathcal{E}}(\psi_j^{\mathcal{E}}, v) = \lambda_j^{\mathcal{E}} b_{\mathcal{E}}(\psi_j^{\mathcal{E}}, v), \qquad \forall v \in S_{h,0}(\mathcal{E}), \tag{8}$$

where $a_{\mathcal{E}}(u, v)$ is as defined in (6), and

$$b_{\mathcal{E}}(u, v) = h^{-4} \int_{G_{\mathcal{E}}} \alpha \hat{u}\, \hat{v}\, dx, \tag{9}$$

and $G_{\mathcal{E}}$ is a 3D layer around and along the edge $\mathcal{E}$, defined as the sum of all fine tetrahedra of $\mathcal{T}_h$ those touching $\mathcal{E}$ by a fine edge or a vertex, and $\hat{u}, \hat{v} \in S_h$ are the discrete zero extensions of $u, v \in S_{h,0}(\mathcal{E})$. The scaling in the form $b_{\mathcal{E}}(u, v)$, and in the form $b_{kl}(u, v)$ in (11) below, comes from an inverse inequality and the lines of the proof of Theorem 1, which will be provided in a

full version of this paper published elsewhere. The functions $\psi_j^{\mathcal{E}}$ are extended inside as follows, taking zero values at the nodal points of all remaining edges and faces, and then extending further inside as discrete harmonic in the sense as described in Section 2. The extension is denoted by the same symbol. Writing the eigenvalues in the increasing order, i.e. $0 < \lambda_1^{\mathcal{E}} \leq \lambda_2^{\mathcal{E}} \leq \ldots \lambda_{M_{\mathcal{E}}}^{\mathcal{E}}$ for $M_{\mathcal{E}} = \dim(S_{h,0}(\mathcal{E}))$, we define the local edge spectral component of the coarse space as follows. Let $V_{\mathcal{E}} = \mathrm{Span}(\psi_j^{\mathcal{E}})_{j=1}^{n_{\mathcal{E}}}$,where $n_{\mathcal{E}} \leq M_{\mathcal{E}}$ is the number of eigenfunctions $\psi_j^{\mathcal{E}}$, whose eigenvalues $\lambda_j^{\mathcal{E}}$ are less then a given threshold prescribed for each subdomain by the user.

Eigenvalue problem on face interface

Each face $\mathcal{F}_{kl}$ inherits a 2D triangulation consisting of triangles $\mathcal{T}_h(\mathcal{F}_{kl})$, and a local face finite element space $S_h(\mathcal{F}_{kl})$ being the space of traces of S_h onto $\mathcal{F}_{kl}$, and $S_{h,0}(\mathcal{F}_{kl}) = S_h(\mathcal{F}_{kl}) \cap H_0^1(\mathcal{F}_{kl})$. We introduce $\overline{\mathcal{F}}_{I,ij}$ as the sum of closed triangles of $\mathcal{T}_h(\mathcal{F}_{kl})$ such that all their nodes are not in $\partial\mathcal{F}_{kl}$.

The face eigenvalue problem is then to find the eigen pairs $(\lambda_j^{kl}, \psi_j^{kl}) \in \mathbb{R}_+ \times S_{h,0}(\mathcal{F}_{kl})$ such that

$$a_{kl}(\psi_j^{kl}, v) = \lambda_j^{\mathcal{F}_{kl}} b_{kl}(\psi_j^{kl}, v), \qquad \forall v \in S_{h,0}(\mathcal{F}_{kl}), \tag{10}$$

where

$$a_{kl}(u,v) = \sum_{\tau \subset \mathcal{F}_{I,kl}} \int_\tau \underline{\alpha}_\tau \nabla u(x) \nabla v(x), \quad b_{kl}(u,v) = h^{-3} \int_{G_{\mathcal{F}_{kl}}} \alpha \hat{u}\, \hat{v}\, dx, \tag{11}$$

and $\underline{\alpha}_\tau = \max_{\tau \subset \partial K} \alpha_K$ is the maximum value of the coefficient over the tetrahedra sharing the fine face $\tau \in \mathcal{T}_h(F_{I,kl})$, $G_{\mathcal{F}_{kl}}$ is a 3D layer of tetrahedra around and along the face $\mathcal{F}_{kl}$, defined as sum of all fine tetrahedra of T_h those touching $\mathcal{F}_{kl}$ by a fine face, a fine edge or a vertex, and $\hat{u}, \hat{v} \in S_h$ are the discrete zero extensions of $u, v \in S_{h,0}(\mathcal{F}_{kl})$. The functions ψ_j^{kl} are extended inside as follows, taking zero values at the nodal points of all remaining faces and edges, and then extending further inside as discrete harmonic in the same sense as in Section 2. The extension is denoted by the same symbol.

Again, by writing the eigenvalues in the increasing order as $0 \leq \lambda_1^{kl} \leq \lambda_2^{kl} \leq \ldots \lambda_{M_{kl}}^{kl}$ for $M_{kl} = \dim(S_{h,0}(\mathcal{F}_{kl}))$, we can define the local face spectral component of the coarse space as follows. Let $V_{kl} = \mathrm{Span}(\psi_j^{kl})_{j=1}^{n_{kl}}$,where $n_{kl} \leq M_{kl}$ is the number of eigenfunctions ψ_j^{kl} whose eigenvalues λ_j^{kl} are less than a given threshold provided by an user.

Finally, The coarse space V_0, after the enrichment takes the following form:

$$V_0 = \hat{V}_0 + \sum_{\mathcal{F}_{kl} \subset \Gamma} V_{kl} + \sum_{\mathcal{E} \subset \Gamma} V_{\mathcal{E}}. \tag{12}$$

Note that $\hat{V}_0 = I_V S_h$, as defined in Section 4.1.

Remark 1. The bilinear forms $b_{\mathcal{E}}(u,v)$, cf. (9), and $b_{kl}(u,v)$, cf. (11), can be defined in other ways. For instance, we can consider larger layers $G_{\mathcal{E}}$ or $G_{\mathcal{F}_{kl}}$, or even consider nonzero extensions of $u \in S_{h,0}(\mathcal{E})$ and $u \in S_{h,0}(\mathcal{F}_{kl})$, but with minimal energy. We can also take the bilinear forms to be equal to the restrictions of the scaled original energy form to their respective layers or to the whole substructures, that is following the ideas of Klawonn et al (2015, 2016b,a). In all cases, we will have similar estimates as in Theorem 1 in the next section.

5 Condition number

Following the abstract Schwarz framework, cf. Smith et al (1996); Toselli and Widlund (2005), and the classical theory of eigenvalue problems, we can show the following theoretical bound on the condition number for the preconditioned system of our method.

Theorem 1. *For all $u \in S_h$, the following holds,*

$$c\left(1 + \max_{\mathcal{E}} \frac{1}{\lambda_{n_{\mathcal{E}}+1}} + \max_{\mathcal{F}_{kl}} \frac{1}{\lambda_{n_{kl}+1}}\right) a(u,u) \leq a(Tu,u) \leq C\, a(u,u),$$

where C, c are positive constants independent of the coefficient α, the mesh parameter h and the sudomain size H.

References

Brenner SC, Sung LY (1999) Balancing domain decomposition for nonconforming plate elements. Numer Math 83(1):25–52

Calvo JG, Widlund OB (2016) An adaptive choice of primal constraints for BDDC domain decomposition algorithms. Electron Trans Numer Anal 45:524–544

Chartier T, Falgout RD, Henson VE, Jones J, Manteuffel T, McCormick S, Ruge J, Vassilevski PS (2003) Spectral AMGe (ρAMGe). SIAM J Sci Comput 25(1):1–26, DOI 10.1137/S106482750139892X, URL `http://dx.doi.org/10.1137/S106482750139892X`

Dolean V, Nataf F, Scheichl R, Spillane N (2012) Analysis of a two-level schwarz method with coarse spaces based on local Dirichlet-to-Neumann maps. Comput Methods Appl Math 12:391–414

Efendiev Y, Galvis J, Lazarov R, Margenov S, Ren J (2012a) Robust two-level domain decomposition preconditioners for high-contrast anisotropic

flows in multiscale media. Comput Methods Appl Math 12(4):415–436, DOI 10.2478/cmam-2012-0031, URL http://dx.doi.org/10.2478/cmam-2012-0031

Efendiev Y, Galvis J, Lazarov R, Willems J (2012b) Robust domain decomposition preconditioners for abstract symmetric positive definite bilinear forms. ESAIM Math Mod Num Anal 46:1175–1199

Galvis J, Efendiev Y (2010) Domain decomposition preconditioners for multiscale flows in high-contrast media. Multiscale Model Simul 8(4):1461–1483, DOI 10.1137/090751190, URL http://dx.doi.org/10.1137/090751190

Kim HH, Chung ET (2015) A BDDC algorithm with enriched coarse spaces for two-dimensional elliptic problems with oscillatory and high contrast coefficients. Multiscale Modeling & Simulation 13(2):571–593

Kim HH, Chung E, Wang J (2017) BDDC and FETI-DP preconditioners with adaptive coarse spaces for three-dimensional elliptic problems with oscillatory and high contrast coefficients. J Comput Phys 349:191–214, URL https://doi.org/10.1016/j.jcp.2017.08.003

Klawonn A, Radtke P, Rheinbach O (2015) FETI-DP methods with an adaptive coarse space. SIAM J Numer Anal 53(1):297–320

Klawonn A, Kuhn M, Rheinbach O (2016a) Adaptive coarse spaces for FETI-DP in three dimensions. SIAM Journal on Scientific Computing 38(5):A2880–A2911

Klawonn A, Radtke P, Rheinbach O (2016b) A comparison of adaptive coarse spaces for iterative substructuring in two dimensions. Electronic Transactions on Numerical Analysis 45:75–106

Mandel J, 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

Nataf F, Xiang H, Dolean V (2010) A two level domain decomposition preconditioner based on local Dirichlet-to-Neumann maps. C r mathématique 348(21-22):1163–1167

Nataf F, Xiang H, Dolean V, Spillane N (2011) A coarse space construction based on local Dirichlet-to-Neumann maps. SIAM J Sci Comput 33(4):1623–1642

Smith BF, Bjørstad PE, Gropp WD (1996) Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations. Cambridge University Press, Cambridge

Spillane N, Dolean V, Hauret P, Nataf F, Pechstein C, Scheichl R (2014) Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. Numer Math 126:741–770, DOI 10.1007/s00211-013-0576-y, URL http://dx.doi.org/10.1007/s00211-013-0576-y

Toselli A, Widlund O (2005) Domain decomposition methods—algorithms and theory, Springer Series in Computational Mathematics, vol 34. Springer-Verlag, Berlin

An immersed boundary method based on the L^2-projection approach

M.G.C. Nestola, B. Becsek, H. Zolfaghari, P. Zulian, D. Obrist and R. Krause

Abstract In this paper we present a framework for Fluid-Structure Interaction simulations. Taking inspiration from the Immersed Boundary technique introduced by Peskin [1] we employ the finite element method for discretizing the equations of the solid structure and the finite difference method for discretizing the fluid flow. The two discretizations are coupled by using a volume based L^2–projection approach to transfer elastic forces and velocities between the fluid and the solid domain. We present results for a Fluid–Structure Intercation benchmark which describes self-induced oscillating deformations of an elastic beam in a flow channel.

Dr. Maria Giuseppina Chiara Nestola
Institute of Computational Science, Center for Computational Medicine in Cardiology (CCMC), Università della Svizzera italiana,Via Giuseppe Buffi 13, 9600 Lugano, Switzerland e-mail: nestom@usi.ch

Barna Becsek
ARTORG Center for Biomedical Engineering Research, University of Bern, Murtenstrasse 50 3008 Bern, Switzerland e-mail: barna.becsek@artorg.unibe.ch

Hadi Zolfaghari
ARTORG Center for Biomedical Engineering Research, University of Bern, Murtenstrasse 50 3008 Bern, Switzerland e-mail: hadi.zolfaghari@artorg.unibe.ch

Prof. Dr. Dominik Obrist
ARTORG Center for Biomedical Engineering Research, University of Bern, Murtenstrasse 50 3008 Bern, Switzerland e-mail: dominik.obrist@artorg.unibe.ch

Dr. Patrick Zulian
Institute of Computational Science, CCMC, Università della Svizzera italiana, Via Giuseppe Buffi 13, 9600 Lugano, Switzerland e-mail: patrick.zulian@usi.ch

Prof. Dr. Rolf Krause
Institute of Computational Science, CCMC, Università della Svizzera italiana, Via Giuseppe Buffi 13, 9600 Lugano, Switzerland e-mail: rolf.krause@usi.ch

1 Introduction

During the last decades, Fluid–Structure Interaction (FSI) [1, 2] has received considerable attention due to various applications where a fluid and a solid interact with each other (such as in aeronautics, turbomachinery, and biomedical applications).

Several approaches have been developed in order to reproduce the interaction between a fluid and a surrounding solid structure, which can be classified in boundary-fitted and embedded boundary methods. In the boundary-fitted methods, the fluid problem is resolved in a moving spatial domain over which the incompressible Navier-Stokes equations are formulated in an Arbitrary Lagrange Eulerian (ALE) framework [3] while the solid structure is usually described in a Lagrangian fashion. Although this approach is known to allow for accurate results at the interface between solid and fluid, for scenarios that involve large displacements and/or rotations, the fluid grid may become severely distorted, thus affecting both the numerical stability of the problem and the accuracy of the solution.

In order to circumvent those difficulties, embedded boundary approaches such as the Immersed Boundary Method (IBM), have been introduced to model the fluid-structure interaction on a stationary fluid grid analyzed in a Eulerian fashion. The main aspect of this technique is the representation of the immersed solid material as a force density in the Navier-Stokes equations.

In the IBM, the volume of the solid is commonly described by systems of fibres that resist extension, compression, or bending [1, 2, 4]. Some alternative approaches have been proposed on the basis of the finite element method for the spatial approximation of the Lagrangian quantities (force densities, displacement field, etc.). In all these approaches the reaction force exerted by the solid on the fluid is computed *explicitly* by using the fluid velocity field to get the corresponding displacement of the solid structure [5, 6, 7].

We describe an alternative framework for FSI simulations, where we employ the finite difference method for simulating the fluid flow and couple it with a finite element method for the structural problem. The main novelties of this work are (I) the description of the solid body motion obtained by solving *implicitly* the elastodynamic equations and (II) the treatment of the Lagrangian-Eulerian interaction which is achieved by means of the L^2– projection. Such approach allows for the transfer of data between non-matching structured (Cartesian) and unstructured meshes arbitrarily distributed among different processors.

All the modules of the FSI computational frameworks are integrated into the multi-physics simulation framework MOOSE (`mooseframework.org`). The code is optimised for modern hybrid high-performance computing platforms such as the Cray XC50 system at the Swiss National Supercomputing Centre CSCS.

2 Strong Formulation of the FSI Problem

In this section we provide a brief description of the methodology adopted in our framework to solve the FSI problem. Since the proposed approach follows the main

principle of the IBM, we employ the standard Eulerian formulation for the Navier-Stokes equations for incompressible flows, whereas the elastic response of the embedded structure is described in a Lagrangian fashion.

Let $\Omega \subset \mathbb{R}^d$ (with $d = 1,2,3$) be a bounded Lipschitz domain denoting the physical region occupied by the coupled fluid-structure system. We label $\mathbf{x} \in \Omega$ as the spatial point, and $\widehat{\mathbf{x}} \in \widehat{\Omega}_s$ as the material (or reference) point, with $\widehat{\Omega}_s \subset \mathbb{R}^d$ denoting the material (reference) configuration of the solid domain (Fig. 1).

We assume that the map $\widehat{\boldsymbol{\chi}} : \widehat{\Omega}_s \times I \to \mathbb{R}^d$ is a one-to-one correspondence between the material $\widehat{\mathbf{x}}$ and the actual $\mathbf{x}$ positions occupied by the elastic structure during the time interval $I = [0\ \mathrm{T}]$, s. t. $(\widehat{\mathbf{x}}, t) \to \mathbf{x} = \widehat{\boldsymbol{\chi}}(\widehat{\mathbf{x}}, t)$, $\forall t \in I$. Additionally, we denote with Γ_{fsi} the physical interface between the fluid and the solid mesh.

The strong formulation of the complete FSI problem reads as follows:

$$
\begin{aligned}
\widehat{\rho}_{s0} \frac{\partial^2 \widehat{\mathbf{u}}_s}{\partial t^2} - \widehat{\nabla}_{\widehat{\mathbf{x}}} \cdot \widehat{\mathbf{P}} &= \mathbf{d} && \text{on} \quad \widehat{\Omega}_s \quad (a) \\
\rho_f \frac{\partial \mathbf{v}_f}{\partial t} + \rho_f \left(\mathbf{v}_f \cdot \nabla\right) \mathbf{v}_f + \nabla p_f - \mu \Delta \mathbf{v}_f &= \mathbf{f}_{\mathrm{fsi}} && \text{on} \quad \Omega \quad (b) \\
\nabla \cdot \mathbf{v}_f &= 0 && \text{on} \quad \Omega \quad (c) \\
\mathbf{v}_f &= \frac{\partial \mathbf{u}_s}{\partial t} && \text{on} \quad \Gamma_{\mathrm{fsi}} \quad (d)
\end{aligned}
\tag{1}
$$

Here Eq. 1(a) is the equation of the elastodynamics where $\widehat{\rho}_{s0}$ is the mass density per unit undeformed volume of the elastic structure, $\widehat{\mathbf{u}}_s = \widehat{\mathbf{u}}_s(\widehat{\mathbf{x}}, t)$ is the related displacement field, $\widehat{\mathbf{P}} = \widehat{\mathbf{P}}(\widehat{\mathbf{x}}, t)$ is the first Piola-Kirchhoff stress tensor, $\mathbf{d}$ is a prescribed external body force, and $\widehat{\nabla}_{\widehat{\mathbf{x}}} \cdot$ is the divergence operator computed in the reference configuration. For an hyperelastic material, the first Piola-Kirchhoff stress tensor $\widehat{\mathbf{P}}$ is related to the deformation through a constitutive equation derived from a given scalar valued energy function Ψ, i. e. $\widehat{\mathbf{P}} = \widehat{\mathbf{F}} \frac{\partial \psi(\widehat{\mathbf{E}})}{\partial \widehat{\mathbf{E}}}$, where $\widehat{\mathbf{E}} := 1/2(\widehat{\mathbf{F}}^T \widehat{\mathbf{F}} - \mathbf{I})$ is the Lagrangian-Green strain tensor and $\widehat{\mathbf{F}}$ is the deformation gradient tensor defined as $\widehat{\mathbf{F}} = \nabla_{\widehat{\mathbf{x}}} \mathbf{x}$.

Eq.s 1(b-c) represent the standard Navier-Stokes equations where ρ_f is the fluid density, $\mathbf{v}_f$ is the velocity field of the fluid, p_f is the pressure, $\nabla_{\mathbf{x}}$ is the gradient operator, $\Delta_{\mathbf{x}}$ is the Laplacian operator computed in the current configuration and $\mathbf{f}_{\mathrm{fsi}}$

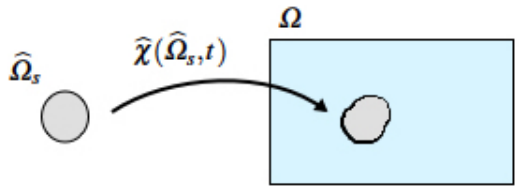


Fig. 1 Lagrangian (left) and Eulerian (right) coordinate systems adopted in the Immersed Boundary method.

is the force density generated by the embedded solid structure as we will describe in Section 3.1.

Remark In the equation of the elastodynamics, i. e. Eq. 1(a), the evaluation of the inertial term must take care of the fluid in which it is embedded. This can be done by subtracting the density of the fluid phase from the solid one (i.e. $\widehat{\rho}_{s0} - \rho_f$) [14]. It is worth to pointing out that, since in our case the fluid velocity field is used to recover the displacement of the FSI interface, this difference is restricted only to Γ_{fsi}.

3 Discretization of the FSI problem

In this section, we provide some details about the discretization in time and in space of the solid and the fluid sub-problem.

3.1 Solid Problem

For the time discretization of the solid problem, we adopt the classical Newmark scheme. This scheme is based on a Taylor expansion of the displacements and the velocities:

$$\widehat{\mathbf{u}}_{s,n+1} = \widehat{\mathbf{u}}_{s,n} + \Delta t\ \mathbf{v}_{s,n} + \tfrac{\Delta t^2}{2}((1-2\beta)\mathbf{a}_{s,n} + 2\beta\mathbf{a}_{s,n+1})$$

$$\widehat{\mathbf{v}}_{s,n+1} = \widehat{\mathbf{v}}_{s,n} + \Delta t\ ((1-\alpha)\widehat{\mathbf{u}}_{s,n} + \alpha\widehat{\mathbf{a}}_{s,n+1})$$

where Δt is the time step size, $\mathbf{a}_s := \frac{\partial^2 \widehat{\mathbf{u}}_s}{\partial t^2}$ and $\mathbf{v}_s := \frac{\partial \widehat{\mathbf{u}}_s}{\partial t}$ are the the acceleration and the velocity of the solid, respectively, and the parameters α and 2β are chosen such that $\alpha = 2\beta = 1/2$.

For the spatial discretization of the structure problem, we assume that the solid domain $\widehat{\Omega}_s$ can be approximated by a discrete domain $\widehat{\Omega}_s^h$ and the associated mesh $\widehat{T}_s^h = \{\widehat{E}_s \subseteq \widehat{\Omega}_s^h | \bigcup \widehat{E}_s = \widehat{\Omega}_s^h\}$, where its elements $\widehat{E}_s$ form a partition. The Galerkin formulation of the elastodynamics equation reads:

For every $t \in (0;T]$ find $\widehat{\mathbf{u}}_s^h(\cdot,t) \in \widehat{\mathbf{V}}_s^h := [\widehat{V}_s^h(\widehat{T}_s^h)]^d \subset [H_0^1(\widehat{\Omega}_s)]^d$ so that:

$$(\widehat{\rho}_{s0}\widehat{\mathbf{a}}_s^h, \delta\mathbf{u}_s^h) + a(\mathbf{u}_s^h, \delta\mathbf{u}_s^h) - (\mathbf{d}_s^h, \delta\mathbf{u}_s^h) = \mathbf{0} \tag{2}$$

By defining $(\mathbf{F}^h, \delta\mathbf{u}_s^h) = a(\mathbf{u}_s^h, \delta\mathbf{u}_s^h) - (\mathbf{d}_s^h, \delta\mathbf{u}_s^h)$ and using the Green's formula we get:

$$(\widehat{\rho}_{s0}\widehat{\mathbf{a}}_s^h, \delta\mathbf{u}_s^h) + (\mathbf{F}^h, \delta\mathbf{u}_s^h) = (\mathbf{f}_{\text{fsi}}^h, \delta\mathbf{u}_s^h)_{L^2(\Gamma_{\text{fsi}}^h)} \tag{3}$$

where $\mathbf{f}_{\text{fsi}}^h$ represents the reaction force exerted by the solid structure on the fluid.

3.2 Fluid Problem

The time integration of the fluid problem is carried out by a 3rd order low-storage Runge-Kutta scheme for both the advective and the diffusion terms [8].

For the discretization of Eq. 1(b), the usage of high-order (sixth) explicit-finite differences leads to a linear system of equations of the form:

$$\begin{bmatrix} \mathbf{H} & \mathbf{G} \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_f \\ \mathbf{p}_f \end{bmatrix} = \begin{bmatrix} \mathbf{z} \\ \mathbf{0} \end{bmatrix}$$

Here the matrices $\mathbf{D}$ and $\mathbf{G}$ are the spatial discretization of the divergence and the gradient operators, $\mathbf{z}$ is the discrete representation of the right hand side, whereas $\mathbf{H}$ is the Helmholtz operator which coincides with the identity matrix (except for the boundary conditions) due to the usage of a purely explicit time integration scheme. By applying $\mathbf{D}$ to the equation $\mathbf{H}\mathbf{v}_f + \mathbf{G}\mathbf{p}_f = \mathbf{0}$, one may derive the following equation for the pressure:

$$\mathbf{D}\mathbf{H}^{-1}\mathbf{G}\mathbf{p}_f = \mathbf{D}\mathbf{H}^{-1}\mathbf{z} \tag{4}$$

In order to guarantee the gradient of the pressure to be unique, the Schur complement $\mathbf{D}\mathbf{H}^{-1}\mathbf{G}$ must be h-elliptic (i.e. must have only one zero eigenvalue). To this aim Arakawa-C grids are adopted which combine several types of nodal points located in different geometrical positions.

4 L^2 - projection

For coupling the two sub-problems we adopt a volume L^2–projection which allows for the transfer of discrete fields between non conforming meshes arbitrarily distributed among several processors. Such an approach ensures convergence, efficiency, flexibility and accuracy without requiring a priori information on the relation between the different meshes. To this aim, we attach Lagrangian basis functions to the finite difference discretization [9], define the corresponding finite element space as $\mathbf{V}_f^h = \mathbf{V}_f^h(T_f^h) \subset [H_0^1(\Omega)]^d$ and introduce the vector of Lagrange multipliers $\boldsymbol{\lambda}_{\text{fsi}}^h$ with the related virtual variations, $\delta\boldsymbol{\lambda}_{\text{fsi}}^h \in \mathbf{M}_{\text{fsi}}^h(\widehat{T}_s^h \cap T_f^h) \subset [H^1(\widehat{\Omega}_s \cap \Omega)]^d$, where T_f^h represents the fluid grid.

In the following, the projection operator $\mathbb{P} : V_f^h \to V_s^h$ is defined by focusing on the scalar case, which means that for each component of the velocity $v_{f,i}^h \in V_f^h$, we may find $w_{s,i}^h = \mathbb{P}(v_{f,i}^h) \in \widehat{V}_s^h$, such that the following weak-equality condition holds:

$$\int_{\widehat{T}_s^h \cap T_f^h} (v_{f,i}^h - \mathbb{P}(v_{f,i}^h))\delta\lambda_{fsi}^h \, dV = \int_{\widehat{T}_s^h \cap T_f^h} (v_{f,i}^h - w_{s,i}^h)\delta\lambda_{\text{fsi}}^h \, dV = 0 \qquad \forall\ \delta\lambda_{\text{fsi}}^h \in M_{\text{fsi}}^h \tag{5}$$

By writing v_f^h, w_s^h and $\delta\lambda_{\text{fsi}}^h$ in term of basis functions (here the index i is omitted for a simpler notation), i.e. $v_f^h = \sum_{l \in J_f} v_f^l N_f^l$, $w_s^h = \sum_{j \in J_s} w_s^j N_s^j$ and $\delta\lambda_{\text{fsi}}^h = \sum_{k \in J_{\text{fsi}}} \delta\lambda_{\text{fsi}}^k N_{\text{fsi}}^k$ (where J_s, J_f and J_{fsi} are index sets), we get the so called mortar integrals: $B_{k,l} = \int_{I_h} N_f^l N_{\text{fsi}}^k \, dV$ and $S_{k,j} = \int_{I_h} N_s^j N_{\text{fsi}}^k \, dV$. Equation 5 can be then written in the following algebraic form:

$$\mathbf{w}_s = \mathbf{S}^{-1}\mathbf{B}\mathbf{v}_f = \mathbf{T}\mathbf{v}_f \tag{6}$$

The transpose of $\mathbf{T}$ is used to transfer the reaction force from the solid to the fluid grid.

In order to reduce the computational cost required to compute the inverse of the matrix $\mathbf{S}$, we adopt dual basis functions for the function space $\mathbf{M}^h_{\text{fsi}}$. In this case this function space is spanned by a set of functions which are biorthogonal to the basis functions of $\widehat{\mathbf{V}}^h_s$ with respect to the L^2-inner product:

$$(N^k_{\text{fsi}}, N^j_s)_{L^2(I_h)} = \delta^{k,j}(N^j_s, \mathbf{1})_{L^2(I_h)} \quad \forall k, j \tag{7}$$

The usage of the dual basis functions corresponds to replacing the standard L^2–projection with a Pseudo–L^2–projection, which allows for a more efficient evaluation of the transfer operator $\mathbf{T}$ since the matrix $\mathbf{S}$ becomes diagonal. The assembly of the transfer operator is done in several steps [10]: (a) we compute the overlapping region by means of a tree search algorithm, (b) generate the quadrature points for integrating in the intersecting region, (c) compute the local element-wise contributions for the operators $\mathbf{B}$ and $\mathbf{S}$ by means of numerical quadrature, and (d) assemble the two mortar matrices.

5 Overview of the FSI algorithm

In our framework a segregated approach is adopted to solve the fully coupled FSI problem. More specifically, we use a fixed point (Picard) iteration scheme for solving the arising coupled non-linear discrete system.

For a given time step n and given a starting solution at the Picard iteration l, the following steps are performed within iteration $l+1$:

Step 1: Velocity values are transferred from the fluid grid to the solid mesh.
Step 2: The elastodynamic equation (Eq. 1(a)) is solved with the Dirichlet boundary conditions (Eq. 1(d)).
Step 3: The reaction force $\mathbf{f}_{\text{fsi}}$ is computed and transferred from the solid mesh to the fluid grid.
Step 4: The Navier-Stokes problem (Eq. 1(b)-(c)) is solved by using the force $\mathbf{f}_{\text{fsi}}$ as source term.
Step 5: Suitable residual norms are computed between the FSI interaction force terms evaluated at iterations l and $l+1$, i. e. $\|\mathbf{f}^{l+1}_{\text{fsi}} - \mathbf{f}^{l}_{\text{fsi}}\|_\infty / \|\mathbf{f}^{0}_{\text{fsi}}\|$ for the relative convergence criterion and $\|\mathbf{f}^{l+1}_{\text{fsi}} - \mathbf{f}^{l}_{\text{fsi}}\|_\infty$ for the absolute convergence criterion [6], and compared with given threshold values. This ensures the satisfaction of the coupling between the two sub-problems, thus leading to either a new Picard iteration or a new time step $n+1$ otherwise.

We employ the numerical solver IMPACT (**I**ncompressible (Turbulent) flows on **M**assively **PA**rallel **C**ompu**T**ers) for solving the non-dimensional Navier-Stokes equations [8]. The solid problem and the assembly of the transfer operator are implemented in the finite-element framework MOOSE (`www.mooseframework.org`), whereas the library MOONoLiTH (`https://bitbucket.org/zulianp/par_moonolith`) is used for detecting the overlapping region between the fluid and the solid grids and computing the corresponding intesections.

6 Numerical Results

In this section we present results related to the Turek-Hron FSI benchmark which considers the incompressible flow of a Newtonian fluid around an elastic solid structure composed of a disk and a rectangular trailing beam.

The dimensions of the fluid channel are (Fig. 2 (a)): length $L_f = 3.0\,m$ and height $H_f = 0.41\,m$. The disk center is positioned at $C = (0.2\,m, 0.2\,m)$ (measured from the left bottom corner of the channel) and the radius is $r = 0.05\,m$. The elastic structure bar has length $L_s = 0.35\,m$ and height $H_s = 0.02\,m$; the right bottom corner is positioned at $(0.6\,m, 0.19\,m)$, and the left end is fully attached to the circle. The fluid properties are $\rho_f = 1000\,kg/m^3$ and $\mu = 1\,Pa \cdot s$ which lead to a Reynolds number of 200. The density of the solid structure is the same as the fluid phase, and a Saint-Venant Kirchhoff model is adopted as constitutive law, for which the first Piola-Kirchhoff stress tensor is defined as: $\widehat{\mathbf{P}} = \widehat{\mathbf{F}}(\lambda\ \mathrm{tr}(\widehat{\mathbf{E}})\mathbf{I} + 2\mu\widehat{\mathbf{E}})$ with $\mu = 2.0\,MPa$ and $\lambda = 4.7\,MPa$. Periodic boundary conditions are imposed along the inlet and the outlet of the fluid channel together with no-slip boundary conditions on the top and the bottom. Moreover at the inlet a Poiseuille flow with a centerline velocity of $1.5\,m/s$ is enforced by a fringe region appended downstream.

In Fig. 2 (b) we show the displacements in x and y direction of a control point P located at the end of the elastic beam ($A \equiv (0.6\,m,\ 0.2\,m)$, Fig. 2 (a)). The amplitude of the last period of oscillation is in the range of $0.03\,m$ for the vertical displacement and of $0.0025\,m$ for the horizontal displacement; the frequency of the y-displacement is about $6\,s^{-1}$, and the frequency for the x-displacement is about $11\,s^{-1}$. All values are in good agreement with the original benchmark results [11]. In Fig. 2 (c) we also show the forces exerted by the lift and drag forces acting on the cylinder and the beam structure together. Again the values agree well with the results obtained by other numerical methods applied to the same problem [12]. Finally, the fluid

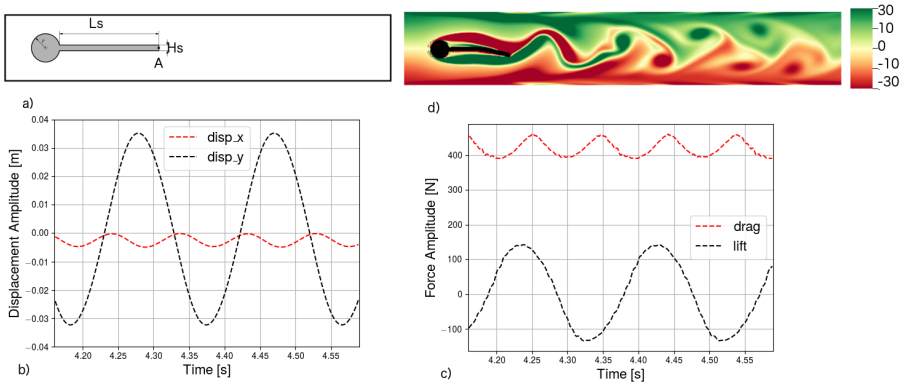


Fig. 2 (a) Geometry of the Turek-Hron benchmark. (b) Amplitude displacement in x and y direction of a control point A located at the end of the elastic beam. (c) Lift and drag forces. (d) Fluid vorticity.

vorticity is depicted in Fig. 2 (d) ranging from $-30\,s^{-1}$ to $30\,s^{-1}$, in agreement with numerical values reported in Griffith [13] .

7 Conclusion

In this article we present a novel FSI framework based on the IMB. The description of the solid motion, obtained by solving *implicitly* the elastodynamic equations, ensures to yield extra stability and robustness. Moreover, the use of the fluid solver IMPACT and of the software MOONoLith for the L^2-projection allows for a completely parallel framework suitable for the simulation of complex and large simulations such the blood flow in human arteries and through heart valves.

References

1. Peskin, Charles S."Flow patterns around heart valves: a numerical method." Journal of computational physics 10.2 (1972): 252-271.
2. Liu, Wing Kam, et al."Immersed finite element method and its applications to biological systems." Computer methods in applied mechanics and engineering 195.13 (2006): 1722-1749.
3. Nestola, Maria GC, et al. "Three-band decomposition analysis in multiscale FSI models of abdominal aortic aneurysms." International Journal of Modern Physics C 27.02 (2016): 1650017.
4. Devendran, Dharshi, and Charles S. Peskin."An immersed boundary energy-based method for incompressible viscoelasticity." Journal of Computational Physics 231.14 (2012): 4613-4642.
5. Griffith, B. E., and Luo, X. (2017). "Hybrid finite difference/finite element immersed boundary method". International journal for numerical methods in biomedical engineering. 33.12 (2017).
6. Gil, Antonio J., et al. The immersed structural potential method for haemodynamic applications. Journal of Computational Physics 229.22 (2010): 8613-8641.
7. Boffi, D., et al. On the hyper-elastic formulation of the immersed boundary method. Computer Methods in Applied Mechanics and Engineering 197.25 (2008): 2210-2231.
8. Henniger, Rolf, Dominik Obrist, and Leonhard Kleiser."High-order accurate iterative solution of the Navier-Stokes equations for incompressible flows." PAMM 7.1 (2007): 4100009-4100010.
9. Fackeldey, K., et al. "Coupling molecular dynamics and continua with weak constraints". Multiscale Modeling and Simulation, 9.4 (2011) 1459-1494.
10. Krause, Rolf, and Patrick Zulian."A parallel approach to the variational transfer of discrete fields between arbitrarily distributed unstructured finite element meshes." SIAM Journal on Scientific Computing 38.3 (2016): C307-C333.
11. Turek, Stefan, and Jaroslav Hron."Proposal for numerical benchmarking of fluid-structure interaction between an elastic object and laminar incompressible flow." Springer Berlin Heidelberg, 2006. 371-385.
12. Turek, Stefan, et al."Numerical benchmarking of fluid-structure interaction: A comparison of different discretization and solution approaches." Fluid Structure Interaction II. Springer, Berlin, Heidelberg, 2011. 413-424.
13. Griffith, B. E., and Xiaoyu Luo."Hybrid finite difference/finite element version of the immersed boundary method." International Journal for Numererical Methods in Engineering 1-26 (Submitted in revised form, 2012) DOI: 10.1002/nme.
14. Hesch, C., et al."A mortar approach for fluid-structure interaction problems: Immersed strategies for deformable and rigid bodies." Computer Methods in Applied Mechanics and Engineering 278 (2014): 853-882.

Combining space-time multigrid techniques with multilevel Monte Carlo methods for SDEs

Martin Neumüller and Andreas Thalhammer

Abstract *In this work we combine multilevel Monte Carlo methods for time-dependent stochastic differential equations with a space-time multigrid method. The idea is to use the space-time hierarchy from the multilevel Monte Carlo method also for the solution process of the arising linear systems. This symbiosis leads to a robust and parallel method with respect to space, time and probability. We show the performance of this approach by several numerical experiments which demonstrate the advantages of this approach.*

1 Introduction

Stochastic differential equations (SDEs) have become an invaluable tool for modelling time-dependent problems that are perturbed by random influences. Since the importance of such models increases constantly, there is a high demand on improving the efficiency of numerical algorithms for SDEs, especially, if one is interested in the approximation of $\mathbb{E}[\varphi(X(T))]$, where $X(T)$ denotes the (mild) solution of an SDE evaluated at time T and $\mathbb{E}$ denotes the expectation, where φ is a mapping determining the statistical quantity of interest.

In this work we focus on approximating $\mathbb{E}[\varphi(X(T))]$ for the solution process of linear SDEs driven by additive noise. For this we combine space-time multigrid methods for approximating solutions of time-dependent deterministic differential equations, see [4] and the references therein, and multilevel Monte Carlo (MLMC) methods, see e.g. [5, 6]. Both methods as such are well-known to be parallelizable,

Martin Neumüller
Johannes Kepler University, Institute of Computational Mathematics,
Altenberger Straße 69, A-4040 Linz, Austria, e-mail: neumueller@numa.uni-linz.ac.at.

Andreas Thalhammer
Johannes Kepler University, Doctoral Program "Computational Mathematics" and Institute for Stochastics, Altenberger Straße 69, A-4040 Linz, Austria, e-mail: andreas.thalhammer@jku.at.

however, the combination of both methods is a completely new approach that enables the full parallelization of the problem in space, time and probability.

The outline of this article is as follows: In the remainder of this section, we introduce two model problems (the Ornstein-Uhlenbeck process and the stochastic heat equation) together with discretization techniques for these model problems with respect to space and time. Afterwards, we consider the multilevel Monte Carlo (MLMC) method for approximating the expectation in Section 2 and we discuss parallelizable space-time multigrid methods based on the inherited space-time hierarchy of the MLMC estimator in Section 3. Finally, we conclude by presenting numerical experiments in Section 4.

1.1 Model problems

Let $T > 0$ and let $(\Omega, \{\mathscr{F}_t\}_{t\in[0,T]}, \mathscr{F}, \mathbb{P})$ be a complete filtered probability space. At first, we consider a one-dimensional model problem given by the stochastic ordinary differential equation (SODE)

$$\begin{aligned} \mathrm{d}u(t) + \lambda u(t)\,\mathrm{d}t &= \sigma\,\mathrm{d}\beta(t) \qquad \text{for } t \in (0,T], \\ u(0) &= u_0, \end{aligned} \tag{1}$$

where $\lambda \in \mathbb{R}_0^+, \sigma, u_0 \in \mathbb{R}$ and $\beta = (\beta(t), t \in [0,T])$ is a standard Brownian motion. The solution of this SODE is a *Ornstein-Uhlenbeck* process defined by

$$u(t) = u_0 e^{-\lambda t} + \sigma \int_0^t e^{-\lambda(t-s)}\,\mathrm{d}\beta(s), \qquad t \in [0,T]. \tag{2}$$

As second model problem we consider the *stochastic heat equation* on a bounded and convex domain $D \subset \mathbb{R}^d, d = 1,2,3$, with homogeneous Dirichlet boundary conditions. We rewrite the corresponding stochastic partial differential equation (SPDE) as a stochastic evolution equation on the Hilbert space $H = L^2(D)$

$$\begin{aligned} \mathrm{d}U(t) &= \Delta U(t)\,\mathrm{d}t + \mathrm{d}W(t) \qquad \text{for } t \in (0,T], \\ U(0) &= U_0 \in H^2(D) \cap H_0^1(D). \end{aligned} \tag{3}$$

Subsequently, we denote by $(e_j, j \in \mathbb{N})$ the set of eigenfunctions of the Laplace operator $-\Delta$, which forms an orthonormal basis of H. Furthermore, let $W = (W(t), t \in [0,T])$ be an H-valued Q-Wiener process with a linear, positive definite, symmetric, trace class covariance operator Q. Then W can be represented as (see e.g. [3, 7])

$$W(t) = \sum_{j=1}^{\infty} \sqrt{\mu_j} e_j \beta_j(t), \tag{4}$$

where $(\mu_j, j \in \mathbb{N})$ denotes the set of eigenvalues of Q satisfying $Qe_j = \mu_j e_j$ and $(\beta_j, j \in \mathbb{N})$ is a sequence of independent standard Brownian motions.

Then, by [3], there exists a unique, square-integrable mild solution to SPDE (3)

$$U(t) = S(t)U_0 + \int_0^t S(t-s)\,\mathrm{d}W(s) \qquad \text{for } t \in [0,T], \tag{5}$$

where $S(t), t \in [0,T]$, denotes the semigroup generated by the Laplace operator.

1.2 Discretization of model problems

In this section, we present fully discrete schemes for approximating the solution processes from Eq. (2) and Eq. (5). For this we fix an equidistant partition Θ_K of the time interval $[0,T]$ given by $\Theta_K = \{0 = t_0 < t_1 < \cdots < t_K = T\}$, where for $0 \le j \le K$ we choose $t_j = j\Delta t$ with time step size $\Delta t = T/K$.

For approximating the solution of the Ornstein-Uhlenbeck process (2), we consider the backward Euler–Maruyama scheme given by the recursion

$$(1+\lambda\Delta t)\mathbf{u}_j = \mathbf{u}_{j-1} + \sigma\Delta\beta^j, \qquad \text{for } 1 \le j \le K, \tag{6}$$

where $\mathbf{u}_0 = u_0$ and $\Delta\beta^j = \beta(t_j) - \beta(t_{j-1})$. Rewriting the recursion (6) in a matrix-vector representation yields

$$\begin{pmatrix} (1+\lambda\Delta t) & & & \\ -1 & (1+\lambda\Delta t) & & \\ & \ddots & \ddots & \\ & & -1 & (1+\lambda\Delta t) \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_K \end{pmatrix} = \begin{pmatrix} \sigma\Delta\beta^1(\omega) + \mathbf{u}_0 \\ \sigma\Delta\beta^2(\omega) \\ \vdots \\ \sigma\Delta\beta^K(\omega) \end{pmatrix}. \tag{7}$$

In this article, we abbreviate this linear system by $\mathscr{L}_\tau \mathbf{u} = \mathbf{f}(\omega)$, where we use the ω-dependency in $\mathbf{f}(\omega)$ to indicate that the right hand side is a random vector.

For the stochastic heat equation we want to obtain a fully discrete approximation U_h^j of the mild solution $U(t_j), t_j \in \Theta_K$, where U_h^j attains values in a finite-dimensional subspace $V_h \subset H_0^1(D)$. Besides an appropriate time integration method, we apply a discretization scheme in space. For this we consider a standard Galerkin finite element (FE) discretization based on a regular family $(\mathscr{T}_h, h \in (0,1])$ of triangulations of D with maximal mesh size h. Then V_h denotes the space of globally continuous and on $\mathscr{T}_h$ piecewise linear functions. Furthermore, we denote by N_h the dimension of V_h. By using the nodal basis functions $(\phi_i, 1 \le i \le N_h) \subset H_0^1(D)$, the fully discrete approximation scheme based on Galerkin finite elements in space and on the backward Euler–Maruyama scheme in time is given by (see e.g. [2])

$$(M_h + \Delta t K_h)\mathbf{U}_j = M_h\mathbf{U}_{j-1} + \Delta\mathbf{W}^j \qquad \text{for } 1 \le j \le K, \tag{8}$$

where $\Delta\mathbf{W}^j$ denotes the vector representation of the FE approximation of the Q-Wiener increments $\Delta W^j(\mathbf{x}) = W(t_j,\mathbf{x}) - W(t_{j-1},\mathbf{x}), \mathbf{x} \in D$, and for $j = 0,\ldots,K$,

$$U_h^j = \sum_{i=1}^{N_h} \mathbf{U}_j[i]\phi_i,$$

where $\mathbf{U}_j[i]$ denotes the ith component of the vector $\mathbf{U}_j \in \mathbb{R}^{N_h}$. Here, we denote by M_h the standard *mass* matrix and K_h the standard *stiffness* matrix defined by

$$M_h[i,j] := \int_D \phi_j(\mathbf{x})\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}, \quad K_h[i,j] := \int_D \nabla\phi_j(\mathbf{x})\cdot\nabla\phi_i(\mathbf{x})\,\mathrm{d}\mathbf{x}, \quad \text{for } i,j = 1,\dots,N_h.$$

Finally, by rewriting the numerical scheme (8) in terms of a matrix-vector formulation we obtain the large linear system

$$\begin{pmatrix} M_h + \Delta t K_h & & & \\ -M_h & M_h + \Delta t K_h & & \\ & \ddots & \ddots & \\ & & -M_h & M_h + \Delta t K_h \end{pmatrix} \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \vdots \\ \mathbf{U}_K \end{pmatrix} = \begin{pmatrix} \Delta\mathbf{W}^1(\omega) + M_h\mathbf{U}_0 \\ \Delta\mathbf{W}^2(\omega) \\ \vdots \\ \Delta\mathbf{W}^K(\omega) \end{pmatrix} \tag{9}$$

that is subsequently abbreviated by $\mathscr{L}_{h,\tau}\mathbf{U} = \mathbf{F}(\omega)$.

2 Multilevel Monte Carlo methods

The goal is to approximate $\mathbb{E}[\varphi(u(T))]$ or $\mathbb{E}[\varphi(U(T))]$ for a sufficiently smooth mapping $\varphi : H \to B$, where B is a separable Hilbert space, by using suitable estimators. For $Y \in L^2(\Omega;B)$ a common way to approximate $\mathbb{E}[Y]$ is to use a standard *Monte Carlo* (MC) estimator defined by

$$E_M[Y] := \frac{1}{M}\sum_{i=1}^{M} Y^{(i)},$$

where $(Y^{(i)}, i = 1,\dots,M)$ are independent realizations of Y. Here, $L^2(\Omega;B)$ denotes the space of strongly measurable random variables Y that satisfy

$$\|Y\|^2_{L^2(\Omega;B)} := \mathbb{E}[\|Y\|_B^2] < \infty.$$

Due to the rather slow convergence of the MC estimator of order $M^{-1/2}$ in the $L^2(\Omega;B)$-sense, the efficient *multilevel Monte Carlo* (MLMC) estimator has been proposed in [5]. For its definition we consider a sequence $(Y_\ell, \ell \in \mathbb{N}_0)$ of approximations of $Y \in L^2(\Omega;B)$ based on different refinement levels $\ell \in \mathbb{N}_0$. The MLMC estimator is then given by

$$E^L[Y_L] := \sum_{\ell=0}^{L} E_{M_\ell}[Y_\ell - Y_{\ell-1}],$$

where $Y_{-1} = 0$. The $L^2(\Omega;B)$-error of the MLMC estimator satisfies (see [6])

$$\|\mathbb{E}[Y] - E^L[Y_L]\|_{L^2(\Omega;B)} \leq \|\mathbb{E}[Y - Y_L]\|_B + \left(\sum_{\ell=0}^{L} M_\ell^{-1} \operatorname{Var}[Y_\ell - Y_{\ell-1}] \right)^{1/2} \tag{10}$$

where $\operatorname{Var}[Y] = \mathbb{E}[\|Y - \mathbb{E}[Y]\|_B^2]$ for $Y \in L^2(\Omega;B)$.

In the following two subsections, we discuss how to choose the number of samples $(M_\ell, \ell \in \mathbb{N}_0)$ and the refinement parameter h and Δt in order to guarantee the convergence of the MLMC estimator.

2.1 Ornstein-Uhlenbeck process

Let u be given in Eq. (2) and for $\ell \in \mathbb{N}_0$ let $\mathbf{u}_{K_\ell}$ be the numerical approximation of $u(T)$ based on the backward Euler–Maruyama scheme (6) with respect to the partition Θ_{K_ℓ} with time step size Δt_ℓ. Furthermore, let $\varphi \in C_b^2(\mathbb{R},\mathbb{R})$, i.e., $\varphi : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable with bounded first and second derivatives. Due to the additive noise structure of SDE (1) we obtain by results from [8] that

$$|\mathbb{E}[\varphi(u(T)) - \varphi(\mathbf{u}_{K_L})]| \leq C\Delta t_L, \quad \operatorname{Var}[\varphi(\mathbf{u}_{K_\ell}) - \varphi(\mathbf{u}_{K_{\ell-1}})]^{1/2} \leq C\Delta t_\ell.$$

Thus, by similar arguments as in [6], if we choose for any $\varepsilon, C_M > 0$,

$$M_0 = \lceil C_M \Delta t_L^{-2} \rceil, \qquad M_\ell = \lceil C_M \Delta t_\ell^2 \Delta t_L^{-2} \ell^{1+\varepsilon} \rceil \quad \text{for } \ell = 1, \dots, L, \tag{11}$$

then $\|\mathbb{E}[\varphi(u(T))] - E^L[\varphi(\mathbf{u}_{K_L})]\|_{L^2(\Omega;\mathbb{R})} = \mathcal{O}(\Delta t_L)$.

2.2 Stochastic heat equation

Let U be given in Eq. (5) and for $\ell \in \mathbb{N}_0$ let $U_{h_\ell}^{K_\ell}$ be an approximation of $U(T)$ based on the FE backward Euler–Maruyama scheme (8) with respect to the partition Θ_{K_ℓ} and the FE space V_{h_ℓ}. Furthermore, let $\varphi \in \mathscr{C}_b^2(H,B)$, i.e., $\varphi : H \to B$ is twice Fréchet differentiable with bounded first and second Fréchet derivatives. Then by using the results from [1], we get by choosing $\Delta t_\ell = h_\ell^2$ for any $\gamma \in [0,1)$

$$\|\mathbb{E}[\varphi(U(T)) - \varphi(U_{h_L}^{K_L})]\|_B \leq C h_L^{2\gamma}, \ \operatorname{Var}[\varphi(U_{h_\ell}^{K_\ell}) - \varphi(U_{h_{\ell-1}}^{K_{\ell-1}})] \leq C h_\ell^{2\gamma}.$$

Thus, by [6], if we choose $\Delta t_\ell = h_\ell^2$ and for any

$$M_0 = \lceil C_M h_L^{-2\gamma} \rceil, \qquad M_\ell = \lceil C_M h_\ell^{2\gamma} h_L^{-2\gamma} \ell^{1+\varepsilon} \rceil \quad \text{for } \ell = 1, \dots, L. \tag{12}$$

then $\|\mathbb{E}[\varphi(U(T))] - E^L[\varphi(U_{h_L}^{K_L})]\|_{L^2(\Omega;B)} = \mathcal{O}(h_L^\gamma)$.

3 Space-time multigrid methods

The idea is to use the space-time hierarchy from the MLMC methods discussed in Sections 2.1 and 2.2 also for a space-time multigrid approach. In detail we use the space-time multigrid method presented in [4] to solve the linear system (7) and (9) at once. The advantage is that we can also add parallelization in time direction and also with respect to the space dimension. So using the space-time hierarchy coming from the MLMC method for the linear solver we obtain an algorithm which can be applied in parallel with respect to space, time and probability. For the space-time multigrid method we use a (inexact) damped block Jacobi smoother, see also [4], i.e. for the problem (7) we use

$$\mathbf{u}^{(n+1)} = \mathbf{u}^{(n)} + \alpha \mathscr{D}_{\tau}^{-1} \left[\mathbf{f}(\omega) - \mathscr{L}_{\tau} \mathbf{u}^{(n)} \right] \qquad \text{for } n = 0, 1, \dots,$$

with the diagonal matrix $\mathscr{D}_{\tau} := \operatorname{diag}(1 + \lambda \Delta t)$. Whereas, for the problem (9) we use the smoothing iteration

$$\mathbf{U}^{(n+1)} = \mathbf{U}^{(n)} + \alpha \mathscr{D}_{h,\tau}^{-1} \left[\mathbf{F}(\omega) - \mathscr{L}_{h,\tau} \mathbf{U}^{(n)} \right] \qquad \text{for } n = 0, 1, \dots,$$

with the block diagonal matrix $\mathscr{D}_{h,\tau} := \operatorname{diag}(M_h + \Delta t K_h)$. To speed up the application of the smoothing procedure we replace the exact inverse of $\mathscr{D}_{h,\tau}$ by applying one iteration of a multigrid V-cycle with respect to the matrix $M_h + \Delta t K_h$. Moreover we always set the damping parameter to $\alpha = \frac{1}{2}$, see [4] for more details. Choosing $\Delta t \approx h^2$ leads – in combination with the space-time hierarchy coming from the MLMC method – to a robust solver which is independent of the number of time steps K the time step size Δt and the randomness ω.

4 Numerical experiments

4.1 Ornstein-Uhlenbeck process

We consider the SODE (1) with $\lambda = 1, \sigma = 1$ and $u_0 = 1$. By choosing $T = 1$ and $\varphi(x) = x$ for all $x \in \mathbb{R}$ we are interested in approximating $\mathbb{E}[u(T)] = e^{-T}$.

For the numerical approximation we consider the backward Euler–Maruyama scheme from Eq. (6) in the matrix-vector representation $\mathscr{L}_{\tau} \mathbf{u} = \mathbf{f}(\omega)$, which is solved by the time multigrid method described in Section 3. For the approximation of the expectation we consider a multilevel Monte Carlo estimator based on the sample size selection from Eq. (11) with $\varepsilon = \frac{1}{2}$ and $C_M = 10$.

In Table 1, $\|\mathbb{E}[u(T)] - E^L[\mathbf{u}_{K_L}]\|_{L^2(\Omega;\mathbb{R})}$ is approximated by a standard Monte Carlo estimator given by

$$\text{MS-err} = \left(\frac{1}{M} \sum_{i=1}^{M} \left| e^{-T} - E^L[\mathbf{u}_{K_L}]^{(i)} \right|^2 \right)^{1/2},$$

where $(E^L[\mathbf{u}_{K_L}]^{(i)}, 1 \le i \le M)$ are independent realizations of the MLMC estimator $E^L[\mathbf{u}_{K_L}]$. For this we choose $M = 100$ in the numerical experiments from Table 1 and we observe the right convergence rates as predicted by the theory.

Table 1 Numerical test for SODE (1) (Ornstein-Uhlenbeck process).

L	time steps	realizations level 0	realizations level L	MS-err	EOC
0	1	10	10	2.61915E-01	-
1	2	40	20	1.39399E-01	0.91
2	4	160	50	6.73215E-02	1.05
3	8	640	80	3.92162E-02	0.78
4	16	2560	110	2.02307E-02	0.95
5	32	10240	140	1.00032E-02	1.02
6	64	40960	180	4.80065E-03	1.06
7	128	163840	220	2.31171E-03	1.05
8	256	655360	270	1.13875E-03	1.02
9	512	2621440	310	5.29684E-04	1.10
10	1024	10485760	360	2.62618E-04	1.01

4.2 Stochastic heat equation

For the stochastic heat equation (3) we consider the one-dimensional case $D = (0,1)$ with initial value $U_0(\mathbf{x}) = \sin(\pi x)$. By choosing $T = 0.2$ and $\varphi(v) = v$ for all $v \in L^2(D)$, we are interested in approximating $\mathbb{E}[U(T,\mathbf{x})] = \exp(-\pi^2 T)\sin(\pi\mathbf{x}), \mathbf{x} \in D$.

The eigenvalues of the Q-Wiener process are $\mu_j = j^{-(2r+1+\varepsilon)}$ with $r = 2$ and any $\varepsilon > 0$, see e.g. [7] for details. For approximating paths of the Q-Wiener process we truncate the series representation (4) after the first $J_h = N_h$ summands, see e.g. [2].

For the numerical approximation in space and time, we consider the FE Euler–Maruyama scheme from Eq. (8) on an equidistant mesh with grid width $h_\ell = 2^{-\ell-1}$ in the matrix-vector formulation $\mathscr{L}_{h,\tau}\mathbf{U} = \mathbf{F}(\omega)$, which is again solved by the space-time multigrid method described in Section 3. For the approximation of the expectation we consider the MLMC method based on the sample size selection (12) with $\varepsilon = 0.5$ and $C_M = 10$.

In numerical experiments $\|\mathbb{E}[U(T)] - E^L[U_{h_L}^{K_L}]\|_{L^2(\Omega;B)}$ is approximated by a standard Monte Carlo estimator, i.e., we consider

$$\text{MS-err} = \left(\frac{1}{M} \sum_{i=1}^{M} \left\| \mathbb{E}[U(T)] - E^L[U_{h_L}^{K_L}]^{(i)} \right\|_{L^2(D)}^2 \right)^{1/2},$$

where $(E^L[U_{h_L}^{K_L}]^{(i)}, 1 \le i \le M)$ are independent realizations of the estimator $E^L[U_{h_L}^{K_L}]$ and

$$\|\mathbb{E}[U(T)] - E^L[U_{h_L}^{K_L}]^{(i)}\|_{L^2(D)}^2 = \int_0^1 \left(\exp(-\pi^2 T)\sin(\pi \mathbf{x}) - E^L[U_{h_L}^{K_L}(\mathbf{x})]^{(i)}\right)^2 \mathrm{d}\mathbf{x}.$$

In Table 2 we use $M = 100$ independent realizations of the MLMC estimator and we observe the optimal convergence rates as predicted by the theory. Moreover we give in Table 3 the averaged solving times for one signle MLMC run for different levels and different distributions of 512 cores. Here we observe that the best possible setting is given by a balanced distribution of cores between parallelization in time and parallelization of the Monte Carlo estimators. For example for level $L = 7$ the best possible setting is given by 8 cores for time parallelization and 64 cores for the Monte Carlo parallelization.

Table 2 Numerical test for SPDE (1) (stochastic heat equation) – convergence.

L	time steps	# elements	realizations level 0	realizations level L	MS-err	EOC
0	1	2	10	10	7.83487E-02	-
1	4	4	40	20	3.39860E-02	1.20
2	16	8	160	30	1.29145E-02	1.40
3	64	16	640	60	5.99035E-03	1.11
4	256	32	2560	90	2.71909E-03	1.14
5	1024	64	10240	120	1.39772E-03	0.96
6	4096	128	40960	150	6.89668E-04	1.02
7	16384	256	163840	190	3.41996E-04	1.01

Table 3 Numerical test for SPDE (1) (stochastic heat equation) – computation time for one MLMC run with respect to different distributions of 512 cores (in sec).

	cores time / cores Monte Carlo							
L	1 / 512	2 / 256	4 / 128	8 / 64	16 / 32	32 / 16	64 / 8	128 / 4
3	0.04	0.02	0.02	0.02	0.03	0.06	0.1	0.14
4	0.27	0.17	0.12	0.13	0.16	0.26	0.47	0.93
5	2.64	1.51	0.95	1.01	1.17	1.64	2.47	4.41
6	24.12	13.92	13.64	11.47	10.76	12.53	15.88	23.5
7	282.46	157.97	153.41	125.56	127.84	133.6	146.81	178.76

References

1. A. Andersson, R. Kruse and S. Larsson: Duality in refined Sobolev-Malliavin spaces and weak approximation of SPDE. Stoch. Partial Differ. Equ. Anal. Comput. 4, 113-149 (2016)
2. A. Barth and A. Lang: Simulation of stochastic partial differential equations using finite element methods. Stochastics 84(2-3), 217-231 (2012)
3. G. Da Prato and J. Zabczyk: Stochastic Equations in Infinite Dimensions. Encyclopedia of Mathematics and its Applications, 44. Cambridge University Press, Cambridge, 1992
4. M. Gander and M. Neumüller: Analysis of a new space-time parallel multigrid algorithm for parabolic problems. SIAM J. Sci. Comput., 38(4) : A2173-A2208 (2016)
5. M. B. Giles: Multilevel Monte Carlo path simulation. Oper. Res. 56, 607-617 (2008)
6. A. Lang: A note on the importance of weak convergence rates for SPDE approximations in multilevel Monte Carlo schemes. In Ronald Cools and Dirk Nuyens, editors, Monte Carlo and Quasi-Monte Carlo Methods, MCQMC, Leuven, Belgium, pp. 489-505, Springer, 2016
7. G. J. Lord, C. E. Powell and T. Shardlow: An Introduction to Computational Stochastic PDEs Cambridge. Texts in Applied Mathematics. Cambridge University Press, New York, 2014
8. G. N. Milstein and M. V. Tretyakov: Stochastic Numerics for Mathematical Physics. Scientific Computation. Springer-Verlag, Berlin, 2004

On Block Triangular Preconditioners for the Interior Point Solution of PDE-Constrained Optimization Problems

John W. Pearson and Jacek Gondzio

Abstract We consider the numerical solution of saddle point systems of equations resulting from the discretization of PDE-constrained optimization problems, with additional bound constraints on the state and control variables, using an interior point method. In particular, we derive a Bramble–Pasciak Conjugate Gradient method and a tailored block triangular preconditioner which may be applied within it. Crucial to the usage of the preconditioner are carefully chosen approximations of the $(1,1)$-block and Schur complement of the saddle point system. To apply the inverse of the Schur complement approximation, which is computationally the most expensive part of the preconditioner, one may then utilize methods such as multigrid or domain decomposition to handle individual sub-blocks of the matrix system.

1 Introduction

A key application of domain decomposition methods, alongside a range of other numerical techniques, is within preconditioned iterative methods for linear systems of equations. In this paper, we examine such systems arising from optimization problems constrained by PDEs—in particular we wish to consider the application of interior point methods to formulations with additional bound constraints. The crucial computational element of such solvers is the development of a fast and robust method for the Newton systems that arise at each interior point iteration. We refer to [1, 3, 8, 13], and the references therein, for previous research on such iterative methods, as well as to [5] for the development of a multigrid scheme.

The key component of the authors' previous work [13] was the consideration of saddle point solvers for these linear systems. It was found that iterative methods

John W. Pearson
School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom, e-mail: j.pearson@ed.ac.uk

Jacek Gondzio
School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom, e-mail: j.gondzio@ed.ac.uk

accelerated by block triangular preconditioners are highly effective for the solution of such systems, often more so than those incorporating analogous block diagonal matrices; however, in general it is difficult to robustly predict the convergence rate of the iterative scheme when using block triangular preconditioners. In this work, we present a new Bramble–Pasciak Conjugate Gradient method which allows one to employ an efficient block triangular approximation, for which the preconditioned system is self-adjoint and positive definite in some non-standard inner product. This also enables one to predict the convergence of the algorithm based on the eigenvalues of the preconditioned system. Such guarantees are not available if one uses more standard Krylov subspace methods for non-symmetric systems, for instance GMRES or BICG. This also provides a framework for domain decomposition techniques, multigrid methods, or other tailored schemes to tackle the individual portions of the block matrix systems at hand. The main contribution of this paper is therefore the presentation of a new solver with the shared advantages of both its faster computational performance, due to the favourable properties of block triangular preconditioners, and the theoretical guarantees of convergence which it provides.

This paper is structured as follows. In Section 2 we describe the PDE-constrained optimization problem of which we wish to consider the numerical solution. In Section 3 we outline the Bramble–Pasciak Conjugate Gradient method, as well as the block triangular preconditioner that we apply within it. In Section 4 we ascertain the effectiveness of our methodology when applied to a number of practical problems.

2 PDE-Constrained Optimization Problem

The problem of which we consider the numerical solution in this paper is given as follows:

$$\begin{aligned} \min_{y,u} \quad & \frac{1}{2}\|y-\widehat{y}\|^2_{L_2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L_2(\Omega)} \\ \text{s.t.} \quad & \mathscr{D}y = u, \quad \text{in } \Omega, \\ & y = f, \quad \text{on } \partial\Omega, \\ & y_a \leq y \leq y_b, \quad \text{a.e. in } \Omega, \\ & u_a \leq u \leq u_b, \quad \text{a.e. in } \Omega. \end{aligned}$$

This problem is solved on a domain $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, with boundary $\partial\Omega$. Here, y, $\widehat{y}$ and u represent the *state*, *desired state* and *control variables*, with $\mathscr{D}$ some given PDE operator. Further, β is a (positive) *regularization parameter*, with f, y_a, y_b, u_a, u_b given functions. The key to this problem is that we wish to find functions y and u which solve the minimization problem constrained by a system of PDEs, while also placing upper and lower bounds on the values that these functions may take.

As illustrated in [13], we may solve this problem using a *discretize-then-optimize* strategy, where a Lagrangian is built on the discrete level and optimality conditions

are subsequently derived from it. The Lagrangian of which we wish to find the stationary point(s), when a finite element method is applied to tackle the barrier optimization problem, is given as follows:

$$\begin{aligned}\mathscr{L}(\mathbf{y},\mathbf{u},\boldsymbol{\lambda}) = \frac{1}{2}\mathbf{y}^T M\mathbf{y} - \mathbf{y}_d^T\mathbf{y} + \frac{\beta}{2}\mathbf{u}^T M\mathbf{u} + \boldsymbol{\lambda}^T(K\mathbf{y} - M\mathbf{u} - \mathbf{f}) \\ - \mu\sum_j \log\left(y_j - y_{a,j}\right) - \mu\sum_j \log\left(y_{b,j} - y_j\right) \\ - \mu\sum_j \log\left(u_j - u_{a,j}\right) - \mu\sum_j \log\left(u_{b,j} - u_j\right),\end{aligned}$$

where $\mathbf{y}$ and $\mathbf{u}$ are the discrete state and control variables, and y_j, $y_{a,j}$, $y_{b,j}$, u_j, $u_{a,j}$, $u_{b,j}$ denote the values of y, y_a, y_b, u, u_a, u_b at the j-th finite element node. The vector $\boldsymbol{\lambda}$ is the discrete *adjoint variable*, enforcing the PDE constraint (which in discretized form is given by $K\mathbf{y} - M\mathbf{u} = \mathbf{f}$). The matrix M is the well known finite element *mass matrix*, with entries defined by $[M]_{ij} = \int_\Omega \phi_i\phi_j \,\mathrm{d}\Omega$, where ϕ_i denote the finite element basis functions used. The matrix K relates to the weak form of the PDE operator $\mathscr{D}$. The vectors $\mathbf{y}_d$ and $\mathbf{f}$ correspond to the functions $\widehat{y}$ and f on the discrete level, and contain entries of the form $\int_\Omega \widehat{y}\phi_i \,\mathrm{d}\Omega$ and $\int_\Omega f\phi_i \,\mathrm{d}\Omega$ respectively. The (positive) *barrier parameter* μ precedes a sum of logarithmic terms which help to enforce the bound constraints on the state and control variables.

The essence of our interior point method is that at each step we wish to find the stationary point of the Lagrangian $\mathscr{L}$, with y_j and u_j updated to take account of the previous iterate, and with μ reduced at each iteration by a factor which is chosen in advance. The algorithm applied is stated in [13]—it is then shown that the main computational bottleneck is the solution of the Newton system

$$\begin{bmatrix} M+D_y & 0 & K^T \\ 0 & \beta M + D_u & -M \\ K & -M & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}\mathbf{y} \\ \boldsymbol{\delta}\mathbf{u} \\ \boldsymbol{\delta}\boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mu(Y-Y_a)^{-1}\mathbf{e} - \mu(Y_b-Y)^{-1}\mathbf{e} + \mathbf{y}_d - M\mathbf{y}^* - K^T\boldsymbol{\lambda}^* \\ \mu(U-U_a)^{-1}\mathbf{e} - \mu(U_b-U)^{-1}\mathbf{e} - \beta M\mathbf{u}^* + M\boldsymbol{\lambda}^* \\ \mathbf{f} - K\mathbf{y}^* + M\mathbf{u}^* \end{bmatrix} \tag{1}$$

at each interior point step. The (diagonal) matrices D_y and D_u are given by

$$\begin{aligned} D_y &= (Y-Y_a)^{-1}Z_{y,a} + (Y_b-Y)^{-1}Z_{y,b}, \\ D_u &= (U-U_a)^{-1}Z_{u,a} + (U_b-U)^{-1}Z_{u,b}. \end{aligned}$$

Here, Y, U, Y_a, Y_b, U_a, U_b are diagonal matrices containing the entries of y, u (at the previous Newton step), y_a, y_b, u_a, u_b; further, $Z_{y,a}$, $Z_{y,b}$, $Z_{u,a}$, $Z_{u,b}$ denote diagonal matrices with entries defined by Lagrange multipliers associated with bounds y_a, y_b, u_a, u_b, respectively. At each iteration, an interior point algorithm attempts to approximately satisfy the following centrality condition:

$$(Z_{y,a})_{jj} = \frac{\mu}{y_j - y_{a,j}}, \quad (Z_{y,b})_{jj} = \frac{\mu}{y_{b,j} - y_j},$$
$$(Z_{u,a})_{jj} = \frac{\mu}{u_j - u_{a,j}}, \quad (Z_{u,b})_{jj} = \frac{\mu}{u_{b,j} - u_j}.$$

The vector $\mathbf{e}$ contains a one at each entry, and the vectors $\mathbf{y}^*$, $\mathbf{u}^*$, $\boldsymbol{\lambda}^*$ contain the previous iterates for $\mathbf{y}$, $\mathbf{u}$, $\boldsymbol{\lambda}$. We wish to solve the matrix system (1) for $\boldsymbol{\delta}\mathbf{y}$, $\boldsymbol{\delta}\mathbf{u}$, $\boldsymbol{\delta}\boldsymbol{\lambda}$, the Newton updates of $\mathbf{y}$, $\mathbf{u}$, $\boldsymbol{\lambda}$, at each interior point iteration.

3 Bramble–Pasciak Conjugate Gradients and Preconditioning

We now wish to approach the main computational challenge within the interior point algorithm, namely the fast and efficient solution of the matrix system (1). This is an example of a *saddle point system*, which is defined in general as a system of equations of the form

$$\underbrace{\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}}_{\mathscr{A}} \underbrace{\begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \end{bmatrix}}_{\mathbf{b}}.$$

There has been a great deal of research on the subject of the numerical solution of such systems, and we refer to [2] for a comprehensive survey. However, in the setting of interior point methods, we face the additional challenge that the $(1,1)$-block A is severely ill-conditioned, due to the presence of diagonal scaling matrices (defined as D_y and D_u in Section 2 for our problem).

In [13], a block diagonal preconditioner was presented, involving approximations $\widehat{A}$ and $\widehat{S}$ for the $(1,1)$-block and the (negative) Schur complement $S := BA^{-1}B^T$, respectively. These approximations were carefully chosen such that the preconditioned system $\mathscr{P}^{-1}\mathscr{A}$ had clustered eigenvalues, and also such that $\widehat{A}^{-1}$ and $\widehat{S}^{-1}$ could be applied cheaply. In this work, we wish to apply a suitable block triangular preconditioner

$$\mathscr{P} = \begin{bmatrix} \widehat{A} & 0 \\ B & -\widehat{S} \end{bmatrix}$$

within a non-standard Conjugate Gradient method. By doing so, we are able to exploit the often superior convergence properties of block triangular preconditioners, alongside the theoretical guarantees of convergence that Conjugate Gradient type methods provide. In particular, we may predict a certain rate of convergence of the iterative method by examining the eigenvalues of the preconditioned system.

The idea of the Bramble–Pasciak Conjugate Gradient method [4] is that we apply this method using an inner product within which the preconditioned system is self-adjoint and positive definite. A suitable inner product is given by $\langle \cdot, \cdot \rangle_{\mathscr{H}}$, with

$$\mathscr{H} = \begin{bmatrix} A - \widehat{A} & 0 \\ 0 & \widehat{S} \end{bmatrix}.$$

The structure of the algorithm is presented below, and we refer to [4, 17, 18] for further details.

Algorithm: Bramble–Pasciak Method for $\mathscr{A}\mathbf{x} = \mathbf{b}$ with Preconditioner $\mathscr{P}$

Initial vectors

Given $\mathbf{x}_0$, set $\mathbf{r}_0 = \mathscr{P}^{-1}(\mathbf{b} - \mathscr{A}\mathbf{x}_0)$, $\mathbf{p}_0 = \mathbf{r}_0$

Conjugate Gradient loop

for $k = 0, 1, \ldots$

$$\alpha_k = \frac{\langle \mathbf{r}_k, \mathbf{r}_k \rangle_{\mathscr{H}}}{\langle \mathscr{P}^{-1}\mathscr{A}\mathbf{p}_k, \mathbf{p}_k \rangle_{\mathscr{H}}}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$$

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k \mathscr{P}^{-1}\mathscr{A}\mathbf{p}_k$$

$$\beta_k = \frac{\langle \mathbf{r}_{k+1}, \mathbf{r}_{k+1} \rangle_{\mathscr{H}}}{\langle \mathbf{r}_k, \mathbf{r}_k \rangle_{\mathscr{H}}}$$

$$\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_k \mathbf{p}_k$$

end

The key components within the algorithm involve computing terms of the form $\mathscr{P}^{-1}\mathbf{v}$ and $\mathscr{H}\mathscr{P}^{-1}\mathbf{v}$, where we write $\mathbf{v} = \left[\mathbf{v}_1^T,\ \mathbf{v}_2^T\right]^T$. The first of these tasks may be accomplished by applying $\widehat{A}^{-1}$ and $\widehat{S}^{-1}$ efficiently, whenever the inverse of the preconditioner is required. For the application of $\mathscr{H}\mathscr{P}^{-1}\mathbf{v}$, which is needed to compute terms of the form $\langle \mathscr{P}^{-1}\mathscr{A}\mathbf{p}_k, \mathbf{p}_k \rangle_{\mathscr{H}}$ and $\langle \mathbf{r}_k, \mathbf{r}_k \rangle_{\mathscr{H}}$ within the Bramble–Pasciak algorithm, we observe that

$$\mathscr{H}\mathscr{P}^{-1}\mathbf{v} = \begin{bmatrix} A - \widehat{A} & 0 \\ 0 & \widehat{S} \end{bmatrix} \begin{bmatrix} \widehat{A}^{-1}\mathbf{v}_1 \\ \widehat{S}^{-1}B\widehat{A}^{-1}\mathbf{v}_1 - \widehat{S}^{-1}\mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} A\widehat{A}^{-1}\mathbf{v}_1 - \mathbf{v}_1 \\ B\widehat{A}^{-1}\mathbf{v}_1 - \mathbf{v}_2 \end{bmatrix}.$$

Therefore, we are only required to apply $\widehat{A}^{-1}$ once in order to compute this term.

We therefore require efficient approximations for the $(1,1)$-block and Schur complement of the matrix system (1) under consideration. For this matrix,

$$A = \begin{bmatrix} M + D_y & 0 \\ 0 & \beta M + D_u \end{bmatrix}, \qquad B = \begin{bmatrix} K & -M \end{bmatrix},$$
$$S = BA^{-1}B^T = K(M + D_y)^{-1}K^T + M(\beta M + D_u)^{-1}M.$$

To approximate the $(1,1)$-block, we apply a Chebyshev semi-iteration method [6, 7] to the diagonally dominant matrices $M + D_y$ and $\beta M + D_u$. As it is necessary to ensure that $A - \widehat{A}$ is positive definite, in turn to guarantee that the inner product

matrix $\mathscr{H}$ is positive definite, we pre-multiply this approximation by a constant $0 \ll \gamma < 1$, which is chosen a priori such that this property holds (see [17]).

In order to approximate the Schur complement, we employ a 'matching strategy', which was derived in [14, 15, 16], and was demonstrated to be highly effective in the context of interior point methods in [13]. We write

$$\widehat{S} = \big(K + \widehat{M}\big)(M + D_y)^{-1}\big(K + \widehat{M}\big)^T,$$

where $\widehat{M} = M\big[\mathrm{diag}(\beta M + D_u)\big]^{-1/2}\big[\mathrm{diag}(M + D_y)\big]^{1/2}$, with the aim of capturing both terms of the exact Schur complement S within our approximation. The inverses of $K + \widehat{M}$ and its transpose may be efficiently approximated using multigrid, domain decomposition, or other methods.

Making use of our approximations of A and S, we may then compile our preconditioner

$$\mathscr{P} = \begin{bmatrix} \gamma(M + D_y)_{\mathrm{Cheb}} & 0 & 0 \\ 0 & \gamma(\beta M + D_u)_{\mathrm{Cheb}} & 0 \\ K & -M & -\widehat{S} \end{bmatrix},$$

which may be readily inverted, giving rise to a computationally efficient algorithm within the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$. Eigenvalue estimates for $\widehat{A}^{-1}A$ and $\widehat{S}^{-1}S$ are discussed in detail in [13]; applying these estimates within the Bramble–Pasciak method leads to robust estimates of convergence rates for the iterative solver, using previous research on this method for PDE-constrained optimization problems without additional bound constraints [17].

4 Numerical Experiments

To test the practical effectiveness of our method we implement an interior point scheme, within which we apply the Bramble–Pasciak Conjugate Gradient method with the preconditioner stated in Section 3. For each problem, we discretize the state, control and adjoint variables using $Q1$ finite elements. The Bramble–Pasciak method is run to a tolerance of 10^{-8} at each interior point step, with the outer (interior point) solver run to a tolerance of 10^{-6}. We measure the average number of Bramble–Pasciak iterations required per outer iteration, until convergence of the interior point method is achieved. The $(1,1)$-block of the matrix system (1) is approximated using 20 steps of Chebyshev semi-iteration, with parameter $\gamma = 0.95$ chosen to ensure positive definiteness of $\mathscr{H}$; the matrices $K + \widehat{M}$ and its transpose, within the Schur complement approximation, are approximately inverted using the Aggregation-based Algebraic Multigrid (AGMG) software [9, 10, 11, 12]. All tests are carried out using MATLAB R2017b, on a quad-core 3.2 GHz processor.

For our first test problem, we consider the Poisson operator $\mathscr{D} = -\nabla^2$, take $\widehat{y} = \sin(\pi x_1)\sin(\pi x_2)$, where $\mathbf{x} = [x_1, x_2]^T \in \Omega = [0,1]^2$, and set $y = 0$ on the bound-

Table 1 Results for the Poisson control example with state constraints, for a range of values of h and β. Presented are the average number of Bramble–Pasciak Conjugate Gradient iterations required, per interior point step.

		$\beta = 1$ $0 \le y \le 0.002$	$\beta = 10^{-1}$ $0 \le y \le 0.02$	$\beta = 10^{-2}$ $0 \le y \le 0.15$	$\beta = 10^{-3}$ $0 \le y \le 0.5$	$\beta = 10^{-4}$ $0 \le y \le 0.8$	$\beta = 10^{-5}$ $0 \le y \le 0.9$
h	2^{-2}	8.5	8.4	7.7	7.4	7.9	8.1
	2^{-3}	12.4	12.6	11.3	13.1	14.0	18.3
	2^{-4}	14.6	14.5	14.2	16.2	18.1	19.9
	2^{-5}	15.8	15.9	16.2	18.3	20.3	22.7
	2^{-6}	16.6	17.1	17.4	20.7	30.0	25.9
	2^{-7}	17.3	17.8	18.5	30.2	26.2	27.8

Table 2 Results for the convection–diffusion control example with control constraints, for a range of values of h and β. Presented are the average number of Bramble–Pasciak Conjugate Gradient iterations required, per interior point step.

		$\beta = 1$ $0 \le u \le 0.1$	$\beta = 10^{-1}$ $0 \le u \le 0.5$	$\beta = 10^{-2}$ $0 \le u \le 2$	$\beta = 10^{-3}$ $0 \le u \le 5$	$\beta = 10^{-4}$ $0 \le u \le 6$	$\beta = 10^{-5}$ $0 \le u \le 6$
h	2^{-2}	8.3	9.8	11.8	14.3	15.4	16.0
	2^{-3}	8.4	10.9	14.8	16.9	20.6	24.4
	2^{-4}	8.2	10.4	13.6	26.6	33.8	35.2
	2^{-5}	8.1	10.1	12.4	16.9	29.9	33.5
	2^{-6}	8.1	9.9	12.2	15.3	25.9	24.6
	2^{-7}	8.3	9.9	12.1	15.3	18.3	18.9

ary $\partial\Omega$ of Ω. We prescribe bound constraints on the state variable y, based on the physical properties of the problem. We solve the matrix systems for a range of step-sizes h and values of β, and present the results obtained in Table 1. We observe very low iteration numbers, considering the complexity of the problem and the accuracy to which we solve the matrix systems, with only moderate increases as h is decreased (i.e. as the dimensions of the matrix systems are increased). We also observe a benign increase in Bramble–Pasciak iterations as β is decreased.

Our second test problem involves a convection–diffusion operator $\mathscr{D} = -0.01\nabla^2 + \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T \cdot \nabla$, a desired state $\widehat{y} = e^{-64\left((x_1-0.5)^2+(x_2-0.5)^2\right)}$, and the boundary condition $y = \widehat{y}$. On this occasion we provide bound constraints for the control variable u, as stated in Table 2. Once again, strong robustness of the Bramble–Pasciak method is observed when either h or β is altered, illustrating that our strategy may be applied to more varied differential operators and types of bound constraints.

We thus establish that the new Bramble–Pasciak Conjugate Gradient algorithm presented for this class of problems provides both enjoyable theoretical properties, and the fast, robust numerical solution of a range of practical examples. It may be concluded that this is therefore a suitable and effective technique for the interior point solution of a number of PDE-constrained optimization problems.

Acknowledgements JWP gratefully acknowledges financial support from the Engineering and Physical Sciences Research Council (EPSRC) Fellowship EP/M018857/2. JG gratefully acknowledges support from the EPSRC Grant EP/N019652/1. The authors thank an anonymous referee for their helpful comments.

References

1. A. Battermann and M. Heinkenschloss, *Preconditioners for Karush–Kuhn–Tucker matrices arising in the optimal control of distributed systems*. In: W. Desch, F. Kappel, and K. Kunisch (eds), *Control and Estimation of Distributed Parameter Systems*, pp.15–32, 1998.
2. M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14, pp.1–137, 2005.
3. M. Benzi, E. Haber, and L. Taralli, *Multilevel algorithms for large-scale interior point methods*, SIAM Journal on Scientific Computing, 31, pp.4152–4175, 2009.
4. J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Mathematics of Computation, 50(181), pp.1–17, 1988.
5. A. Drăgănescu and C. Petra, *Multigrid preconditioning of linear systems for interior point methods applied to a class of box-constrained optimal control problems*, SIAM Journal on Numerical Analysis, 50(1), pp.328–353, 2012.
6. G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second order Richardson iterative methods. I*, Numerische Mathematik, 3, pp.147–156, 1961.
7. G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second order Richardson iterative methods. II*, Numerische Mathematik, 3, pp.157–168, 1961.
8. M. J. Grotte, J. Huber, D. Kourounis, and O. Schenk, *Inexact interior-point method for PDE-constrained nonlinear optimization*, SIAM Journal on Scientific Computing, 36, pp.A1251–A1276, 2014.
9. A. Napov and Y. Notay, *An algebraic multigrid method with guaranteed convergence rate*, SIAM Journal on Scientific Computing, 34(2), pp.A1079–A1109, 2012.
10. Y. Notay, *Aggregation-based algebraic multigrid for convection–diffusion equations*, SIAM Journal on Scientific Computing, 34(4), pp.A2288–A2316, 2012.
11. Y. Notay, *AGMG software and documentation*; see `http://homepages.ulb.ac.be/~ynotay/AGMG`.
12. Y. Notay, *An aggregation-based algebraic multigrid method*, Electronic Transactions on Numerical Analysis, 37, pp.123–146, 2010.
13. J. W. Pearson and J. Gondzio, *Fast interior point solution of quadratic programming problems arising from PDE-constrained optimization*, Numerische Mathematik, 137(4), pp.959–999, 2017.
14. J. W. Pearson, M. Stoll, and A. J. Wathen, *Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems*, SIAM Journal on Matrix Analysis and Applications, 33(4), pp.1126–1152, 2012.
15. J. W. Pearson and A. J. Wathen, *A new approximation of the Schur complement in preconditioners for PDE-constrained optimization*, Numerical Linear Algebra with Applications, 19, pp.816–829, 2012.
16. J. W. Pearson and A. J. Wathen, *Fast iterative solvers for convection–diffusion control problems*, Electronic Transactions on Numerical Analysis, 40, pp.294–310, 2013.
17. T. Rees and M. Stoll, *Block triangular preconditioners for PDE-constrained optimization*, Numerical Linear Algebra with Applications, 17(6), pp.977–996, 2010.
18. M. Stoll, *Solving Linear Systems Using the Adjoint*, DPhil Thesis, University of Oxford, 2008.

Robust multigrid methods for isogeometric discretizations of the Stokes equations

Stefan Takacs[1]

[1] Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
stefan.takacs@ricam.oeaw.ac.at

Abstract. In recent publications, the author and his coworkers have proposed a multigrid method for solving linear systems arizing from the discretization of partial differential equations in isogeometric analysis and have proven that the convergence rates are robust in both the grid size and the polynomial degree. So far the method has only been discussed for the Poisson problem. In the present paper, we discuss the extension the of these results to the Stokes equations.

Keywords: Isogeometric analysis · Multigrid methods · Stokes problem

1 Introduction

Isogeometric analysis (IgA) was introduced in [12], aiming to improve the connection between computer aided design (CAD) and finite element (FEM) simulation. In IgA, as in CAD software, B-splines and non-uniform rational B-splines (NURBS) are used for representing both the geometrical objects of interest and the solution of the partial differential equation (PDE) to be solved.

In IgA, mostly B-splines or NURBS of maximum smoothness are used, i.e., having a spline degree of p, the functions are $p-1$ times continuously differentiable. Using such a function space, one obtains on the one hand the approximation power of high order functions, while on the other hand, unlike in standard high-order FEM, one does not suffer from a growth of the number of degrees of freedom.

From the computational point of view, the treatment of the linear systems arizing from the discretization with high spline degrees is still challenging as the condition number both of mass and stiffness matrices grows exponentially with the spline degree. In the early IgA literature, finite element solvers have often been transferred to IgA with only minimal adaptations. Numerical experiments indicate that such approaches result in methods that work well for small spline degrees, but their performance deteriorates as the degree is increased, often dramatically. In [11,10], the author and his coworkers have proposed multigrid methods which are provable robust in the polynomial degree and the grid size. Numerical experiments indicated that the proposed approach of *subspace corrected mass smoothers* seems to pay off (compared to multigrid methods with a standard Gauss-Seidel smoother) for polynomial degrees of four or five.

In the present paper, we discuss the extension of the subspace corrected mass smoothers beyond the case of the Poisson problem to the Stokes flow problem. Unlike for the Poisson problem, for the Stokes problem already the setup of a stable isogeometric discretization is non standard. As there have already been results in the

literature, we refer to [1], which serves as a basis of the present paper. Alternative approaches can be found in [8,3,6] and others. After introducing discretizations, we discuss the setup of the preconditioner.

For the Poisson problem, the multigrid solver has been applied directly and as a preconditioner for the conjugate gradient method. For the case of a non-trivial geometry transformation, in [10] a conjugate gradient method, preconditioned with the multigrid method for the parameter domain, has been used. It has been shown that in this case the resulting method is robust both with respect to the grid size and the polynomial degree, but not in the geometry transformation.

There are a few approaches how to carry this over to the Stokes equations. The first possibility is to apply the multigrid method directly to the problem of interest (all-at-once multigrid method), cf. [16] for a particularly popular method in standard FEM or [13] for a survey. As the results for the Poisson problem have indicated that a direct application of the multigrid method in the presence of a non-trivial geometry transformation is not optimal, we do not concentrate on that case.

We therefore consider a Krylov space method with an appropriate preconditioner, living on the parameter domain. In principle, this could be the Stokes problem on the physical domain, but such a choice (an indefinite preconditioner for an indefinite problem) typically requires the use of a GMRES method, the convergence of which is less well understood than that of the minimal residual algorithm, cf. [9]. So, we consider elliptic preconditioners, particularly block-diagonal preconditioners. As the Stokes equations are well-posed in H^1 (velocity) and L_2 (pressure), we just setup preconditioners for those spaces (operator preconditioning). Since the subspace corrected mass smoothers suffer significantly from the geometry transformation, we propose a variant (by incorporating an approximation to the geometry transformation) which led to a significant speedup in several experiments.

As alternative Stokes solvers in IgA, we want to mention overlapping Schwarz approaches, cf. [2], and BDDC approaches, cf. [15], which also yield robustness in the spline degree for certain configurations (like generous overlap or the choice of C^0 regularity across the subdomain interfaces).

This paper is organized as follows. We will introduce the particular model problem in section 2 and discuss three kinds of discretizations for the mixed system in section 3. As a next step, in section 4, we propose a preconditioner. Finally, in section 5, we give the results of the numerical examples and draw some conclusions.

2 Model problem

Let $\Omega \subseteq \mathbb{R}^2$ be a simply connected domain with Lipschitz boundary $\partial\Omega$ and assume a force field f given on Ω and boundary data given on $\partial\Omega$. The *Stokes flow model problem* reads as follows. Find the velocity field u and the pressure p such that

$$-\Delta u + \nabla p = f \qquad \text{and} \qquad \nabla \cdot u = 0 \tag{1}$$

hold on Ω and Dirichlet boundary conditions hold on $\partial\Omega$. After homogenization, we obtain a mixed variational form, which reads as follows. Find $u \in V := H_0^1(\Omega)$

and $p \in Q := L_2(\Omega)$ such that

$$\underbrace{(\nabla u, \nabla v)}_{a(u,v)} + \underbrace{(\nabla \cdot v, p)}_{b(v,p)} = (f, v) \quad \forall v \in V, \qquad \underbrace{(\nabla \cdot u, q)}_{b(u,q)} = 0 \quad \forall q \in Q.$$

Here, and in what follows $L_2(\Omega)$, $H^1(\Omega)$ and $H^1_0(\Omega)$ are the standard Lebesgue and Sobolev spaces, and $(\cdot, \cdot)$ is the standard norm on $L_2(\Omega)$.

Existence and uniqueness of the solution and its dependence of the data follows from Brezzi's theorem [4], which requires besides boundedness and H^1-coercivity of a the inf-sup stability

$$\inf_{q \in L_2(\Omega)} \sup_{v \in H^1(\Omega)} \frac{(\nabla \cdot v, q)}{\|v\|_{H^1(\Omega)} \|q\|_{L_2(\Omega)}} \geq C,$$

which is known to be satisfied for the Stokes problem, cf. [5].

3 Discretization

The discretization is done using a standard Galerkin approach, i.e., we replace the spaces V and Q by finite-dimensional subspaces V_h and Q_h. As for the continuous problem, existence and uniqueness of the solution can be shown by Brezzi's theorem. Boundedness and H^1-coercivity of a follow directly from the continuous problem, but the inf-sup stability for the discrete problem does not. Therefore, we have to guarantee that the *discrete inf-sup condition*

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{(\nabla \cdot v_h, q_h)}{\|v_h\|_{H^1(\Omega)} \|q_h\|_{L_2(\Omega)}} \geq C$$

is satisfied, which is actually a condition on the discretization. In the subsection 3.2, we will discuss discretizations satisfying this condition.

Assuming a particular discretization and a basis for the chosen space, one ends up with a linear system to be solved: For a given $\underline{f}_h$, find $\underline{x}_h$ such that

$$A_h \, \underline{x}_h = \underline{f}_h, \qquad \text{where} \qquad A_h = \begin{pmatrix} K_h & D_h^T \\ D_h & 0 \end{pmatrix} \qquad \text{and} \qquad \underline{x}_h = \begin{pmatrix} \underline{u}_h \\ \underline{p}_h \end{pmatrix} \tag{2}$$

and K_h is a standard stiffness matrix and D_h is a matrix representing the divergence.

3.1 Discretization in isogeometric analysis

Let $S^q_{p,h}$ be the space of all q times continuously differentiable functions on $(0,1)$, which are piecewise polynomials of degree p on a (uniform) grid of size $h = 1/n$. As a basis for $S^q_{p,h}$ we choose the classical basis of B-splines, see, e.g., [7].

For the computational domains $\Omega \subset \mathbb{R}^2$, we first define the spline spaces for the *parameter domain* $\hat{\Omega} = (0,1)^2$. On the parameter domain, we introduce the space of tensor-product splines, $S^{q_1,q_2}_{p_1,p_2,h} := S^{q_1}_{p_1,h} \otimes S^{q_2}_{p_2,h}$, where $A \otimes B$ denotes the linear span of all functions $(x,y) \mapsto u(x)v(y)$, where $u \in A$ and $v \in B$. Note that the

restriction to two dimensions and to a uniform grid is only for ease of notation. The extension to three and more dimensions or to non-uniform grids is completely straight-forward. Assuming that *physical domain* Ω is the image of a B-spline or NURBS mapping

$$G : \hat{\Omega} = (0,1)^2 \to \Omega,$$

we define the spline spaces on the physical domain typically using a classical pull-back principle. More complicated domains are represented patch-wise, where for each patch a separate geometry transformation G exists. For simplicity, we do not discuss that in the present paper.

3.2 Stable discretizations for the Stokes problem

As mentioned above, we are required to set up the discretization such that the discrete inf-sup condition holds. We discuss this first for the parameter domain. Here, we follow the outline of the paper [1], where three spline space configurations have been proposed, which are variants of known stable spaces from standard finite elements: Taylor-Hood like splines $\hat{X}_h^{(\mathrm{TH})}$, Nédélec like splines $\hat{X}_h^{(\mathrm{NE})}$ and Raviart-Thomas like splines $\hat{X}_h^{(\mathrm{RT})}$. All of them utilize the same grid for both the velocity and the pressure, which makes the implementation significantly easier compared to approaches that are based on setting up two different grids (like IgA-variants of the macro elements as proposed in [3]). All of these discretizations follow the spirit of IgA, allowing to freely choose the underlying polynomial degree p. For all of them, the smoothness is on the order of the polynomial degree, which preserves the feature that the number of degrees of freedom is basically not increased when the polynomial degree is increased. For the case of two dimensions, the spaces are given by

$$\begin{aligned}
\hat{X}_h^{(\mathrm{TH})} &:= \hat{V}_h^{(\mathrm{TH})} \times \hat{Q}_h, \quad & \hat{V}_h^{(\mathrm{TH})} &:= S^{p-1,p-1}_{p+1,p+1} \times S^{p-1,p-1}_{p+1,p+1}, \\
\hat{X}_h^{(\mathrm{NE})} &:= \hat{V}_h^{(\mathrm{NE})} \times \hat{Q}_h, \quad & \hat{V}_h^{(\mathrm{NE})} &:= S^{p,p-1}_{p+1,p+1} \times S^{p-1,p}_{p+1,p+1}, \\
\hat{X}_h^{(\mathrm{RT})} &:= \hat{V}_h^{(\mathrm{RT})} \times \hat{Q}_h, \quad & \hat{V}_h^{(\mathrm{RT})} &:= S^{p,p-1}_{p+1,p} \times S^{p-1,p}_{p,p+1}, \qquad & \hat{Q}_h &:= S^{p-1,p-1}_{p,p},
\end{aligned}$$

where $A \times B := \{(a,b) : a \in A,\ b \in B\}$. Observe that these spline spaces are nested, i.e., we have $\hat{V}_h^{(\mathrm{RT})} \subset \hat{V}_h^{(\mathrm{NE})} \subset \hat{V}_h^{(\mathrm{TH})}$ and (for $n \gg p$) a ratio of $9 : 5 : 3$ for the number of degrees of freedom. The extension of these definitions to three dimensions is straight-forward, cf. [1].

For all of these settings, the discrete inf-sup condition has been shown in [1]. For the Raviart-Thomas like splines, the discrete inf-sup condition has not been proven if Dirichlet boundary conditions are present. As the method still seems to work well in practice, we include also the Raviart-Thomas discretization in our experiments.

The next step is to introduce the discretization on the physical domain. As outlined in the beginning of this section, the discretization, once introduced on the parameter domain, is typically defined on the physical domain just by *direct composition*:

$$V_h^{(X,\mathrm{D})} := \{v_h \mid v_h \circ G \in \hat{V}_h^{(X)}\}, \qquad X \in \{\mathrm{TH}, \mathrm{NE}, \mathrm{RT}\}.$$

For the Stokes problem, as an alternative, the divergence preserving *Piola transform* has been proposed:

$$V_h^{(X,\mathrm{P})} := \left\{ v_h \;\middle|\; \frac{1}{\det J_G} J_G\, v_h \circ G \in \hat{V}_h^{(X)} \right\}, \qquad X \in \{\mathrm{TH}, \mathrm{NE}, \mathrm{RT}\},$$

where J_G is the Jacobi matrix of G. The pressure distribution, which is a scalar quantity, is always mapped directly, i.e., in all cases we choose the direct composition

$$Q_h := \{q_h \mid q_h \circ G \in \hat{Q}_h\}.$$

In [1], the inf-sup stability has been shown if the Piola transform is used and for the Taylor-Hood like splines also if the direct composition is used. Again, we report also on the numerical results for the cases that are not covered by the convergence theory (direct composition for the Nédélec like and the Raviart-Thomas like splines).

4 Robust multigrid solvers

As outlined in the introduction, the multigrid preconditioner aims to represent the theoretical block-diagonal preconditioner $Q_h := \mathrm{diag}(K_h, \beta^{-1} M_h)$ where K_h is the stiffness matrix, M_h is the mass matrix and $\beta > 0$ is an scaling parameter, accordingly chosen. As mentioned above and as discussed in detail in [10], we use as preconditioner for the problem on the physical domain the corresponding preconditioner, say $\hat{Q}_h$, on the parameter domain. There, the matrices M_h and K_h are replaced by $\hat{M}_h$ and $\hat{K}_h$, their counterparts on the parameter domain. Note that the stiffness matrix acts on the velocity variable, a vector-valued quantity, and that this matrix is block-diagonal on the parameter domain and, iff the direct composition is used, on the physical domain. In all cases, K_h and $\hat{K}_h$ are spectrally equivalent.

Instead of an exact inverse of the matrix $\hat{Q}_h$, we only need to realize an approximation to the application of $\hat{K}_h^{-1}$ and $\hat{M}_h^{-1}$ to any given vector. The approximation of $\hat{K}_h^{-1}$ is realized using one multigrid V-cycle with one pre- and one post-smoothing step of the subspace corrected mass smoother, as proposed in [10]. There, the algorithm was analyzed only for the case of splines of maximum smoothness, however it can be applied for any spline space and robustness in the polynomial degree can be guaranteed by a slight extension of the presented theory as long as the smoothness is on the order of the polynomial degree. As in the previous publications [11,10], the grid hierarchy is set up for a fixed polynomial degree and a fixed smoothness by just uniformly refining the grid. Using this approach, one obtains nested spaces, so the setup of the coarse-grid correction is trivial.

One of the key observations which was leading to the results in [11,10] was that the spectral equivalence of the mass matrix and its diagonal deteriorates if p is increased. This has also to be taken into account when constructing the preconditioner for the pressure variable. Analogously to the smoother, we realize the application of $\hat{M}_h^{-1}$ exactly, based on the tensor-product structure of the mass matrix.

The preconditioner is symmetric and positive definite and can therefore be applied in the framework of a MINRES iteration.

5 Numerical results

The numerical experiments have been performed using the C++ library G+SMO, see [14], both for the unit square, i.e., for a problem without geometry transformation, and for a quarter annulus $\{(x, y) \in \mathbb{R}^2_+ : 1 < x^2 + y^2 < 4\}$. For both

problems, the problem has been constructed (with inhomogeneous right-hand-side and inhomogeneous Dirichlet boundary conditions) such that the exact solution is

$$u_h(x,y) = \begin{pmatrix} \cos(5x+5y) + \sin(5x-5y) \\ -1 - \cos(5x+5y) + \sin(5x-5y) \end{pmatrix},$$

and $p_h(x,y) = -(1+x)(1+y) + c$, where c is chosen such that $\int_\Omega p_h \,\mathrm{d}x = 0$.

In Table 1, we report on the number of MINRES steps required for reducing the initial error (measured in the ℓ^2-norm of the solution vector) by a factor of 10^{-6}; cases where the memory was not enough are indicated with OoM. We report on all discretization schemes proposed. The need of the discussion of p-robust methods is easily observed when looking at the results for a standard preconditioner: We display the results if one multigrid V-cycle with Gauss-Seidel smoother is used for the velocity and one symmetric Gauss-Seidel sweep is used for the pressure (*GS-MG*). There, the number of iterations increases drastically if p is increased. As the approach is perfectly robust in the grid size $h = 2^\ell$, we omit the numbers for finer grids. Compared to that approach, the preconditioner proposed in Section 4 (*SCMS-MG*) led to results which are robust both in the grid size and the polynomial degree and which works well for all discretizations. Although the iteration numbers are smaller than for the GS-MG preconditioner, one has to consider that the costs of the SCMS-MG preconditioner are significantly higher than those of the GS-MG preconditioner, so the proposed method only pays off if higher polynomial degrees (starting from 4 or 5) are considered. We have chosen $\beta = 0.05$ and as damping parameter σ of the underlying smoother, cf. [10], either $\sigma^{-1} = 0.04\,\hat{h}^2$ (for Taylor-Hood and Nédélec) or $0.16\,\hat{h}^2$ (for Raviart-Thomas), where $\hat{h}$ is the grid size on the parameter domain. While some of the numbers might be improved by fine-tuning the parameters, the given tables for reasonable uniform choices show what one can expect for each of the methods.

In Table 2 we see how well the computed solution approximates the exact solution in the L^2-norm. Here, we have used the abovementioned solver, where the stopping criterion has been chosen to reach either a relative error of 10^{-10} or 100 iterations. We present the error between the computed solution and the known exact solution (for the pressure after projecting into the space of functions with vanishing mean). We observe that, for the same choice of the polynomial order p and the same grid size, the Taylor-Hood discretization yields the smallest errors, at the cost of the largest number of degrees of freedom. For the Raviart-Thomas discretization (where the inf-sup condition cannot be shown for the chosen Dirichlet boundary conditions), we observe that the error for the velocity converges, while the error of the pressure stagnates at around 10^{-2}. Observe moreover that for $p = 5$, the approximation on the coarsest grid was fine enough such that the approximation error could not be improved by refinement.

For the case of the quarter annulus, we distinguish between the results obtained by the direct composition (Table 3) and for the Piola transform (Table 4). Again, we obtain first that GS-MG is robust in h, but that the convergence deteriorates if the polynomial degree grows. As it leads to better results, we have set up the GS-MG on the physical domain. For the proposed SCMS-MG preconditioner, observe that the results behave similar to those for the unit square, however the iteration counts are much larger, particularly if the Piola transform is used. For the direct composition,

	Taylor-Hood				Nédélec				Raviart-Thomas			
$\ell \diagdown p$	2	3	5	8	2	3	5	8	2	3	5	8
MINRES, preconditioned with SCMS-MG												
5	55	54	49	46	80	74	68	55	44	36	35	29
6	54	58	53	51	76	76	70	63	44	37	36	32
7	54	54	54	53	76	76	71	65	45	37	33	29
8	50	51	55	OoM	71	71	67	65	41	37	33	29
MINRES, preconditioned with standard GS-MG												
5	64	167	>1k	>1k	84	213	>1k	>1k	124	219	>1k	>1k

Table 1: Iteration counts for the unit square

		Taylor-Hood			Nédélec			Raviart-Thomas		
p	ℓ	dof	v	p	dof	v	p	dof	v	p
2	4	2372	2e-5	1e-5	1637	2e-5	4e-5	869	3e-4	3e-2
	5	9348	1e-6	6e-7	6341	1e-6	4e-5	3269	3e-5	2e-2
	6	37124	7e-8	4e-6	24965	7e-7	9e-5	12677	7e-6	2e-2
	7	147972	2e-8	7e-7	99077	8e-7	1e-4	49925	3e-6	2e-2
5	4	2891	2e-9	4e-8	2066	6e-8	2e-6	1202	9e-7	6e-3
	5	10347	2e-9	1e-7	7154	8e-8	5e-6	3890	1e-6	3e-3
	6	39083	3e-9	2e-7	26546	9e-7	2e-4	13876	6e-7	3e-3
	7	151951	7e-9	4e-7	102194	2e-6	3e-4	52274	6e-7	4e-3

Table 2: Problem size and L_2-errors for the unit square

	Taylor-Hood				Nédélec				Raviart-Thomas			
$\ell \diagdown p$	2	3	5	8	2	3	5	8	2	3	5	8
MINRES, preconditioned with SCMS-MG												
5	195	190	185	172	257	246	244	206	244	139	128	116
6	208	217	213	199	295	296	280	241	192	170	142	129
7	220	222	232	219	329	330	314	281	213	195	158	140
8	231	239	244	OoM	333	342	333	306	223	200	168	149
MINRES, preconditioned with SCMS-MG-geo												
5	72	69	68	72	69	69	65	63	73	62	53	56
6	77	75	73	79	76	74	64	70	71	69	59	63
7	72	71	70	84	79	70	68	74	75	74	64	69
8	74	73	72	OoM	73	73	71	78	71	70	68	74
MINRES, preconditioned with standard GS-MG												
5	70	173	>1k	>1k	110	225	>1k	>1k	182	220	>1k	>1k

Table 3: Iteration counts for the quarter annulus (direct composition)

	Taylor-Hood				Nédélec				Raviart-Thomas			
$\ell \diagdown p$	2	3	5	8	2	3	5	8	2	3	5	8
MINRES, preconditioned with SCMS-MG												
5	331	331	338	317	288	313	332	305	480	309	295	300
6	407	400	402	371	361	387	405	374	368	344	323	299
7	452	455	455	450	413	450	476	476	418	395	367	341
8	487	485	500	OoM	458	494	556	568	441	438	411	361
MINRES, preconditioned with standard GS-MG												
5	70	165	>1k	>1k	69	164	>1k	>1k	206	199	>1k	>1k

Table 4: Iteration counts for the quarter annulus (Piola transform)

it is possible to improve the convergence significantly by replacing the mass and stiffness matrix on the parameter domain by a simple tensor-rank-one approximation of those matrices on the physical domain (*SCMS-MG-geo*). Note that the tensor-rank-one approximation does not lead to any additional computational costs after the assembling phase. The extension of such a rank-one geometry approximation to the Piola transform is not yet known. For the original SCMS-MG preconditioner, we have chosen β and σ as for the first model problem. Just for the Raviart-Thomas smoother for the case with Piola transformation, we have chosen $\beta = 0.0025$. For the rank-one corrected version, we have chosen $\beta = 0.01$; the damping has been chosen based on approximations of constants of the inverse inequality.

As in the case of standard finite elements, there are several possibilities to discretize the mixed formulation of the Stokes equations. Our experiments indicate that it might pay off to use the (in terms of degrees of freedom) more expensive variant of Taylor Hood discretizations than the other variants, particularly because it is known that that discretization also works for direct composition. The p-robust smoothers which we have proposed for the Poisson problem can be carried over also to the Stokes flow problem, however it seems that a further study is necessary concerning its application in the framework of non-trivial geometry transformations.

References

1. G. S. A. Buffa, C. de Falco. Isogeometric Analysis: new stable elements for the Stokes equation. *Int. J. Num. Meth. Fluids*, 2010.
2. L. Beirão da Veiga, D. Cho, L. Pavarino, and S. Scacchi. Isogeometric Schwarz preconditioners for linear elasticity systems. *Comp. Meth. App. Mech. Eng.*, 253:439 – 454, 2013.
3. A. Bressan and G. Sangalli. Isogeometric discretizations of the Stokes problem: stability analysis by the macroelement technique. *IMA J. Numer. Anal.*, 33(2):629–651, 2013.
4. F. Brezzi. On the Existence, Uniqueness and Approximation of Saddle Point Problems Arising from Lagrangian Multipliers. *RAIRO Anal. Numér.*, 8(2):129 – 151, 1974.
5. F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods.* Springer-Verlag, 1991.
6. A. Buffa, C. de Falco, and G. Sangalli. IsoGeometric Analysis: Stable elements for the 2D Stokes equation. *Int. J. Numer. Methods Fluids*, 65(11-12):1407–1422, 2011.
7. C. de Boor. On calculating with B-splines. *J. on Approximation Theory*, 6(1):50–62, 1972.
8. J. A. Evans and T. J. R. Hughes. Isogeometric divergence-conforming B-splines for the steady Navier–Stokes equations. *Math. Mod. Meth. Appl. Sci.*, 23(08):1421–1478, 2013.
9. A. Greenbaum, V. Pták, and Z. Strakoš. Any Nonincreasing Convergence Curve is Possible for GMRES. *SIAM J. Matrix Anal. Appl.*, 17(3):465–469, 1996.
10. C. Hofreither and S. Takacs. Robust Multigrid for Isogeometric Analysis Based on Stable Splittings of Spline Spaces. *SIAM J. Numer. Anal.*, 55(4):2004–2024, 2017.
11. C. Hofreither, S. Takacs, and W. Zulehner. A robust multigrid method for Isogeometric Analysis in two dimensions using boundary correction. *Comp. Meth. App. Mech. Eng.*, 316:22–42, 2017.
12. T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs. Isogeometric Analysis: CAD, Finite Elements, NURBS, Exact Geometry and Mesh Refinement. *Comp. Meth. App. Mech. Eng.*, 194(39-41):4135–4195, 2005.
13. M. Larin and A. Reusken. A comparative study of efficient iterative solvers for generalized Stokes problem. *Numer. Linear Algebra Appl.*, 15:13–34, 2008.
14. A. Mantzaflaris, S. Takacs, et al. G+Smo. `http://gs.jku.at/gismo`, 2017.
15. L. Pavarino and S. Scacchi. Isogeometric block FETI-DP preconditioners for the Stokes and mixed linear elasticity systems. *Comp. Meth. App. Mech. Eng.*, 310:694 – 710, 2016.
16. S. P. Vanka. Block-implicit multigrid solution of Navier-Stokes equations in primitive variables. *Math. Comp.*, 65:138 – 158, 1986.

A Smoother Based on Nonoverlapping Domain Decomposition Methods for $H(\mathrm{div})$ Problems: A Numerical Study

Susanne C. Brenner and Duk-Soon Oh

Abstract The purpose of this paper is to introduce a V-cycle multigrid method for vector field problems discretized by the lowest order Raviart-Thomas hexahedral element. Our method is connected with a smoother based on a nonoverlapping domain decomposition method. We present numerical experiments to show the effectiveness of our method.

1 Introduction

Let Ω be a bounded domain in $\mathbb{R}^3$ and $H_0(\mathrm{div};\Omega)$ be the space of square integrable vector fields on Ω that have square integrable divergence in Ω and vanishing normal components on $\partial\Omega$ (cf. [7]). In this paper we consider a multigrid method for the following problem: Find $\boldsymbol{u} \in H_0(\mathrm{div};\Omega)$ such that

$$a(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{f},\boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in H_0(\mathrm{div};\Omega), \tag{1}$$

where

$$a(\boldsymbol{w},\boldsymbol{v}) = \alpha(\mathrm{div}\,\boldsymbol{w},\mathrm{div}\,\boldsymbol{v}) + \beta(\boldsymbol{w},\boldsymbol{v}), \tag{2}$$

and $(\cdot,\cdot)$ is the inner product on $L_2(\Omega)$ (or $[L_2(\Omega)]^3$). We assume that $\boldsymbol{f} \in [L_2(\Omega)]^3$ and α and β are positive. Unlike the scalar elliptic equation case, multigrid methods for the problem (1) with simple smoothers do not work. We need a special treatment

Susanne C. Brenner
Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803, USA
e-mail: brenner@math.lsu.edu

Duk-Soon Oh
Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA
e-mail: duksoonoh@gmail.com, duksoon@math.rutgers.edu

for the smoother. In [2–4, 9], an overlapping domain decomposition preconditioner was employed in the construction of the smoother.

Our goal is to develop multigrid methods in the same spirit but using nonoverlapping domain decomposition preconditioners instead, which reduce the dimensions of the subproblems that have to be solved. We note that other multigrid methods for $H(\mathrm{div})$ were investigated in [8, 10].

Applications of fast solvers for $H(\mathrm{div})$ problems are discussed for example in [2, 11–13, 16]. In particular the multigrid method in this paper can be applied to a mixed method for second order partial differential equations based on a first-order system least-squares formulation [2, 6], which is equivalent to our model problem. It can also be used as an effective preconditioner for $H(\mathrm{div})$ problems with variable coefficients. The model problem also arise in Reissner-Mindlin plates [1] and Brinkman equations [15].

In [5], there are similar ingredients and convergence analysis for the convex domain and the constant coefficient case. In this paper, we mainly focus on the numerical study that is not covered by the theory in [5].

The rest of this paper is organized as follows. We present the standard discretization of (1) by the lowest order Raviart-Thomas hexahedral element in Section 2. We next introduce the V-cycle multigrid method in Section 3. Finally, numerical experiments are presented in Section 4.

2 The Discrete Problem

Let $\mathscr{T}_h$ be a hexahedral triangulation of Ω. The lowest order Raviart-Thomas $H(\mathrm{div})$ conforming finite element space [14] is denoted by V_h. A vector field $\boldsymbol{v}$ belongs to V_h if and only if it belongs to $H_0(\mathrm{div};\Omega)$ and takes the form

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} + \begin{bmatrix} b_1 x_1 \\ b_2 x_2 \\ b_3 x_3 \end{bmatrix}$$

on each hexahedral element, where the a_i's and b_i's are constants. On each hexahedral element T the vector field $\boldsymbol{v}$ is determined by the six degrees of freedom defined by the average of the normal component on each face. The discrete problem for (1) is to find $\boldsymbol{u}_h \in V_h$ such that

$$a(\boldsymbol{u}_h, \boldsymbol{v}) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx \qquad \forall\, \boldsymbol{v} \in V_h. \tag{3}$$

In the multigrid approach we solve (3) on a sequence of triangulations $\mathscr{T}_0, \mathscr{T}_1, \ldots$, where $\mathscr{T}_0$ is an initial triangulation of Ω by hexahedral elements and $\mathscr{T}_k$ ($k \geq 1$) is obtained from $\mathscr{T}_{k-1}$ by uniform subdivision. We will denote the lowest order Raviart-Thomas finite element space associated with $\mathscr{T}_k$ by V_k. The k-th level discrete problem is to find $\boldsymbol{u}_k \in V_k$ such that

$$a(\boldsymbol{u}_k, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in V_k.$$

Let $A_k : V_k \longrightarrow V_k'$ be defined by

$$\langle A_k \boldsymbol{w}, \boldsymbol{v} \rangle = a(\boldsymbol{w}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v}, \boldsymbol{w} \in V_k, \tag{4}$$

where $\langle \cdot, \cdot, \rangle$ is the canonical bilinear form on $V_k' \times V_k$. We can then rewrite the k-th level discrete problem as

$$A_k \boldsymbol{u}_k = f_k, \tag{5}$$

where $f_k \in V_k'$ is defined by

$$\langle f_k, \boldsymbol{v} \rangle = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in V_k.$$

Multigrid methods are optimal order iterative methods for equations of the form

$$A_k \boldsymbol{z} = g \tag{6}$$

that includes (5) as a special case.

3 A V-Cycle Multigrid Method

Since the finite element spaces are nested, we can take the coarse-to-fine operator $I_{k-1}^k : V_{k-1} \longrightarrow V_k$ to be the natural injection. The fine-to-coarse operator $I_k^{k-1} : V_k' \longrightarrow V_{k-1}'$ is then defined by

$$\langle I_k^{k-1} \ell, \boldsymbol{v} \rangle = \langle \ell, I_{k-1}^k \boldsymbol{v} \rangle \qquad \forall\, \ell \in V_k',\, \boldsymbol{v} \in V_{k-1}. \tag{7}$$

We will use a smoother of the form

$$z_{\mathrm{new}} = z_{\mathrm{old}} + M_k^{-1}(g - A_k z_{\mathrm{old}}) \tag{8}$$

for the equation (6), where $M_k^{-1} : V_k' \longrightarrow V_k$ is a nonoverlapping domain decomposition preconditioner defined below.

3.1 A Nonoverlapping Domain Decomposition Preconditioner

To conform with standard terminology in domain decomposition, in this subsection we will denote $\mathscr{T}_{k-1}$ by $\mathscr{T}_H$ and $\mathscr{T}_k$ by $\mathscr{T}_h$. (Thus each element in $\mathscr{T}_H$ is partitioned into eight elements in $\mathscr{T}_h$). The spaces V_{k-1} and V_k are denoted by V_H and V_h respectively. The preconditioner M_k^{-1} in (8) is denoted by M_h^{-1} here. It is constructed by substructuring.

For each element $T \in \mathscr{T}_H$, we define the twelve dimensional subspace V_h^T of V_h by

$$V_h^T = \{\boldsymbol{v} \in V_h : \boldsymbol{v} = 0 \text{ on } \Omega \setminus T\}. \tag{9}$$

The natural injection from V_h^T into V_h is denoted by J_T and the operator $A_T : V_h^T \longrightarrow (V_h^T)'$ is defined by

$$\langle A_T \boldsymbol{w}, \boldsymbol{v} \rangle = a(\boldsymbol{w}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v}, \boldsymbol{w} \in V_h^T. \tag{10}$$

Let $\mathscr{F}_H$ be the set of the interior faces of the triangulation $\mathscr{T}_H$. Given any $F \in \mathscr{F}_H$ that is the common face of two elements T_F^+ and T_F^- in $\mathscr{T}_H$, we define the four dimensional subspace V_h^F of V_h by

$$V_h^F = \{\boldsymbol{v} \in V_h : \boldsymbol{v} = \boldsymbol{0} \text{ on } \Omega \setminus (T_F^- \cup T_F^+) \quad \text{and} \quad a(\boldsymbol{v}, \boldsymbol{w}) = 0 \quad \forall\, \boldsymbol{w} \in (V_h^{T_F^-} + V_h^{T_F^+})\}. \tag{11}$$

The natural injection from V_h^F into V_h is denoted by J_F and the operator $A_F : V_h^F \longrightarrow (V_h^F)'$ is defined by

$$\langle A_F \boldsymbol{w}, \boldsymbol{v} \rangle = a(\boldsymbol{w}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v}, \boldsymbol{w} \in V_h^F. \tag{12}$$

if $\boldsymbol{w} \in V_h$ has the same degrees of freedom as $\boldsymbol{v}$ on $\partial T_F^+ \cup \partial T_F^-$.

The subspaces associated with the elements and interior faces of $\mathscr{T}_H$ form a direct sum decomposition of V_h:

$$V_h = \sum_{T \in \mathscr{T}_H} V_h^T + \sum_{F \in \mathscr{F}_H} V_h^F, \tag{13}$$

and the preconditioner M_h^{-1} is given by

$$M_h^{-1} = \eta_F \Big(\sum_{T \in \mathscr{T}_H} J_T A_T^{-1} J_T^t + \sum_{F \in \mathscr{F}_H} J_F A_F^{-1} J_F^t \Big), \tag{14}$$

where η_F is a damping factor and $J_T^t : V_h' \longrightarrow (V_h^T)'$ (resp. $J_F^t : V_h' \longrightarrow (V_h^F)'$) is the transpose of J_T (resp. J_F) with respect to the canonical bilinear forms.

3.2 The k^{th} Level V-Cycle Multigrid Algorithm

The output $MG(k, g, \boldsymbol{z}_0, m)$ of the k^{th} level (symmetric) multigrid V-cycle algorithm for (6), with initial guess $\boldsymbol{z}_0 \in V_k$ and m smoothing steps, is defined by the following recursive steps:

For $k = 0$, the output is obtained from a direct method:

$$MG(0, g, \boldsymbol{z}_0, m) = A_0^{-1} g.$$

For $k \geq 1$, we set

$$
\begin{aligned}
z_l &= z_{l-1} + M_k^{-1}(g - A_k z_{l-1}) \qquad \text{for } 1 \le l \le m,\\
\overline{g} &= I_k^{k-1}(g - A_k z_m),\\
z_{m+1} &= z_m + I_{k-1}^{k} MG(k-1, \overline{g}, 0, m),\\
z_l &= z_{l-1} + M_k^{-1}(g - A_k z_{l-1}) \qquad \text{for } m+2 \le l \le 2m+1.
\end{aligned}
$$

The output of $MG(k, g, z_0, m)$ is z_{2m+1}.

Remark 1. Given $\ell \in V_k'$, the cost of computing $M_k^{-1}\ell$ is $O(n_k)$, where n_k is the dimension of V_k. Therefore the overall cost for computing $MG(k, g, z_0, m)$ is also $O(n_k)$.

If the domain Ω is convex, we have the following convergence theorem:

Theorem 1. *If $z \in V_k$ and $g \in V_k'$ satisfy $A_k z = g$, then we have*

$$
\|z - MG(k, g, z_0, m)\|_a \le \frac{C}{C + 2m} \|z - z_0\|_a \qquad \forall k \ge 1,
$$

where $\|\cdot\|_a^2 = a(\cdot,\cdot)$.

Due to space restriction, a detailed analysis will not be reported here. Further details are provided in [5].

4 Numerical Results

4.1 Jump Coefficient

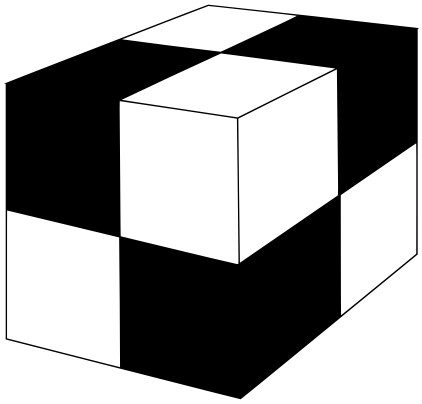


Fig. 1: Checkerboard distribution of the coefficients

In the first experiment we consider (1) on the unit cube $\Omega = (0,1)^3$. We apply multigrid algorithms with smoothers introduced in Section 3.1. The damping factor η_F is taken to be $1/11$. The initial triangulation $\mathscr{T}_0$ consists of eight identical cubes and we use the coefficients α and β that have jumps across the interface between the sub-cubes with a checkerboard pattern as in Fig. 1. We estimate the contraction numbers of the k^{th} level V-cycle multigrid method for $k = 1,\ldots,5$ and for m smoothing steps, where $m = 1,\ldots,6$. We report the contraction numbers obtained by computing the largest eigenvalue of the error propagation operators. The results are presented in Table 1. The uniform convergence of the V-cycle multigrid methods for $m \geq 1$ is clearly observed and the method is not sensitive to the jumps of coefficients.

Table 1: Contraction numbers of the V-cycle multigrid method for the unit cube. α_b and β_b for the black subregions and α_w and β_w for the white subregions as indicated in a checkerboard pattern as in Fig. 1

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
	$\alpha_b = 0.01, \beta_b = 100, \alpha_w = 1, \beta_w = 1$					
$k=1$	8.3e-1	6.8e-1	4.7e-1	2.2e-1	5.1e-2	4.7e-3
$k=2$	9.0e-1	8.2e-1	7.1e-1	5.1e-1	3.2e-1	2.7e-1
$k=3$	9.3e-1	8.8e-1	7.9e-1	6.4e-1	5.2e-1	4.7e-1
$k=4$	9.3e-1	9.0e-1	8.4e-1	7.2e-1	6.4e-1	6.0e-1
$k=5$	9.3e-1	9.0e-1	8.6e-1	7.8e-1	6.9e-1	6.9e-1
	$\alpha_b = 0.1, \beta_b = 10, \alpha_w = 1, \beta_w = 1$					
$k=1$	8.7e-1	7.7e-1	6.0e-1	3.8e-1	2.1e-1	8.1e-2
$k=2$	9.1e-1	8.4e-1	7.1e-1	5.4e-1	3.6e-1	2.8e-1
$k=3$	9.2e-1	8.7e-1	7.8e-1	6.4e-1	5.2e-1	4.7e-1
$k=4$	9.3e-1	9.0e-1	8.4e-1	7.4e-1	6.5e-1	6.0e-1
$k=5$	9.4e-1	9.1e-1	8.7e-1	8.0e-1	7.2e-1	6.9e-1
	$\alpha_b = 1, \beta_b = 1, \alpha_w = 1, \beta_w = 1$					
$k=1$	9.1e-1	8.3e-1	7.1e-1	5.0e-1	3.1e-1	2.3e-1
$k=2$	9.2e-1	8.7e-1	7.9e-1	6.3e-1	5.0e-1	4.3e-1
$k=3$	9.3e-1	9.0e-1	8.4e-1	7.4e-1	6.3e-1	5.8e-1
$k=4$	9.4e-1	9.1e-1	8.7e-1	8.0e-1	7.1e-1	6.7e-1
$k=5$	9.4e-1	9.2e-1	8.8e-1	8.2e-1	7.5e-1	7.2e-1
	$\alpha_b = 10, \beta_b = 0.1, \alpha_w = 1, \beta_w = 1$					
$k=1$	9.0e-1	8.4e-1	7.0e-1	4.9e-1	3.3e-1	2.8e-1
$k=2$	9.2e-1	8.9e-1	7.9e-1	6.4e-1	5.2e-1	4.7e-1
$k=3$	9.4e-1	9.1e-1	8.4e-1	7.4e-1	6.4e-1	6.0e-1
$k=4$	9.4e-1	9.1e-1	8.6e-1	8.0e-1	7.3e-1	6.8e-1
$k=5$	9.4e-1	9.2e-1	8.9e-1	8.2e-1	7.6e-1	7.4e-1
	$\alpha_b = 100, \beta_b = 0.01, \alpha_w = 1, \beta_w = 1$					
$k=1$	9.1e-1	8.4e-1	7.1e-1	5.1e-1	3.3e-1	2.9e-1
$k=2$	9.3e-1	8.9e-1	7.9e-1	6.5e-1	5.2e-1	4.8e-1
$k=3$	9.3e-1	9.1e-1	8.5e-1	7.4e-1	6.4e-1	6.0e-1
$k=4$	9.4e-1	9.2e-1	8.8e-1	8.0e-1	7.1e-1	6.9e-1
$k=5$	9.4e-1	9.3e-1	9.0e-1	8.4e-1	7.7e-1	7.5e-1

4.2 Nonconvex Domain

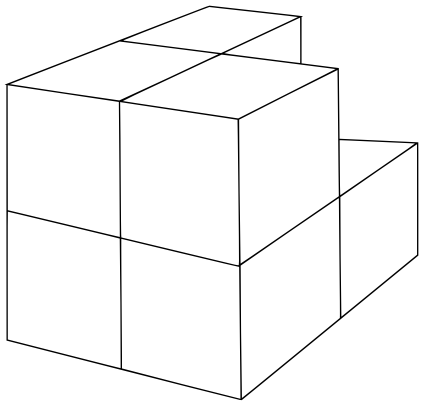


Fig. 2: Nonconvex domain

In the second numerical experiment we report the results for our model problem (1) on the nonconvex domain $\Omega = (0,1)^3 \setminus ([1/2,1]^3)$. We use the constant coefficients $\alpha = 1$ and $\beta = 1$ and other general settings are quite similar to those of Section 4.1. The results are presented in Table 2. It is observed that the method provides a uniform convergence of the V cycle multigrid. However, the contraction numbers are generally larger that those of the convex domain.

Table 2: Contraction numbers of the V-cycle multigrid method for the non-convex domain as in Fig. 2 with $\alpha = 1, \beta = 1$

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
$k=1$	9.3e-1	9.0e-1	8.2e-1	6.9e-1	5.5e-1	4.6e-1
$k=2$	9.5e-1	9.2e-1	8.5e-1	7.7e-1	6.8e-1	6.3e-1
$k=3$	9.6e-1	9.2e-1	8.8e-1	8.2e-1	7.7e-1	7.3e-1
$k=4$	9.6e-1	9.3e-1	8.9e-1	8.5e-1	8.0e-1	7.8e-1
$k=5$	9.6e-1	9.3e-1	9.0e-1	8.7e-1	8.4e-1	8.2e-1

References

1. D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning discrete approximations of the Reissner-Mindlin plate model. *RAIRO Modél. Math. Anal. Numér.*, 31(4):517–557, 1997.
2. D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning in $H(\mathrm{div})$ and applications. *Math. Comp.*, 66(219):957–984, 1997.

3. D. N. Arnold, R. S. Falk, and R. Winther. Multigrid preconditioning in $H(\mathrm{div})$ on non-convex polygons. *Comput. Appl. Math.*, 17(3):303–315, 1998.
4. D. N. Arnold, R. S. Falk, and R. Winther. Multigrid in $H(\mathrm{div})$ and $H(\mathrm{curl})$. *Numer. Math.*, 85(2):197–217, 2000.
5. S. C. Brenner and D.-S. Oh. Multigrid methods for $H(\mathrm{div})$ in three dimensions with nonoverlapping domain decomposition smoothers. submitted.
6. Z. Cai, R. D. Lazarov, T. A. Manteuffel, and S. F. McCormick. First-order system least squares for second-order partial differential equations: Part I. *SIAM J. Numer. Anal.*, 31(6):1785–1799, 1994.
7. V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.
8. R. Hiptmair. Multigrid method for $H(\mathrm{div})$ in three dimensions. *Electron. Trans. Numer. Anal.*, 6(Dec.):133–152, 1997. Special issue on multilevel methods (Copper Mountain, CO, 1997).
9. R. Hiptmair and A. Toselli. Overlapping and multilevel Schwarz methods for vector valued elliptic problems in three dimensions. In *Parallel solution of partial differential equations (Minneapolis, MN, 1997)*, volume 120 of *IMA Vol. Math. Appl.*, pages 181–208. Springer, New York, 2000.
10. T. V. Kolev and P. S. Vassilevski. Parallel auxiliary space AMG solver for $H(\mathrm{div})$ problems. *SIAM J. Sci. Comput.*, 34(6):A3079–A3098, 2012.
11. K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebra Appl.*, 18(1):1–40, 2011.
12. D.-S. Oh. An overlapping Schwarz algorithm for Raviart-Thomas vector fields with discontinuous coefficients. *SIAM J. Numer. Anal.*, 51(1):297–321, 2013.
13. D.-S. Oh, O. B. Widlund, S. Zampini, and C. R. Dohrmann. BDDC algorithms with deluxe scaling and adaptive selection of primal constraints for Raviart–Thomas vector fields. *Math. Comp.* published online, available at https://doi.org/10.1090/mcom/3254.
14. P.-A. Raviart and J. M. Thomas. A mixed finite element method for 2nd order elliptic problems. In *Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975)*, pages 292–315. Lecture Notes in Math., Vol. 606. Springer, Berlin, 1977.
15. P. S. Vassilevski and U. Villa. A block-diagonal algebraic multigrid preconditioner for the Brinkman problem. *SIAM J. Sci. Comput.*, 35(5):S3–S17, 2013.
16. B. I. Wohlmuth, A. Toselli, and O. B. Widlund. An iterative substructuring method for Raviart-Thomas vector fields in three dimensions. *SIAM J. Numer. Anal.*, 37(5):1657–1676 (electronic), 2000.

Optimized Schwarz Method for Poisson's Equation in Rectangular Domains

José C. Garay, Frédéric Magoulès and Daniel B. Szyld

Abstract An analysis of the convergence properties of Optimized Schwarz methods applied as solvers for Poisson's Equation in a bounded rectangular domain with Dirichlet (physical) boundary conditions and Robin transmission conditions on the artificial boundaries is presented. To our knowledge this is the first time that this is done for multiple subdomains forming a 2D array in a bounded domain.

1 Introduction

Classical Schwarz methods are Domain Decomposition (DD) methods in which the transmission conditions between subdomains are Dirichlet boundary conditions. Optimized Schwarz methods are DD methods in which the transmission conditions are chosen in such a way as to improve the convergence rate with respect to the classical method [2, 3, 5]. These transmission conditions are optimized approximations of the optimal transmission conditions, which are obtained by approximating the global Poincaré-Steklov operator by local differential operators. There is more than one family of transmission conditions that can be used for a given PDE , each of these families consisting of a particular approximation of the optimal transmission conditions. For example, for the problem involving Poisson's equation, we have $OO0$ and $OO2$ family of transmission conditions. The $OO0$ family of transmission

José C. Garay
Temple University, Philadelphia, USA , e-mail: jose.garay@temple.edu. Corresponding author.
Supported in part by the U.S. Department of Energy under grant DE-SC0016578.

Frédéric Magoulès
CentraleSupélec, Châtenay-Malabry, France, e-mail: frederic.magoules@centralesupelec.fr .

Daniel B. Szyld
Temple University, Philadelphia, USA, e-mail: szyld@temple.edu.
Supported in part by the the U.S. National Science Foundation under grant DMS-1418882 and the U.S. Deparment of Energy under grant DE-SC0016578.

conditions is obtained by using the zero-th order approximation of the Poincaré-Steklov operator, i.e., it is approximated by a constant α, which leads to have Robin boundary conditions on the artificial boundaries. The $OO2$ family of boundary conditions involves the use of a differential operator that is a linear combination of the normal derivative and tangential second derivatives.

Optimized Schwarz methods (OSM) are fast methods in terms of iteration count when they are used as outer solvers. In [1] it is shown that OSM (as outer solvers) are faster than GMRES preconditioned with a classical Schwarz preconditioner. Also, in parallel computations, OSM requires much less communications between processes in comparison to Krylov methods. Given that communication dominates the execution time of solvers in current supercomputer architectures and will also do so in the upcoming exascale supercomputers, OSM has the potential to be a very good method for solving problems arising from the discretization of PDEs.

In this paper we analyze the convergence properties of OSM applied as solvers for Poisson's Equation in a bounded rectangular domain with Dirichlet (physical) boundary conditions and Robin transmission conditions. To our knowledge, this is the first time an analysis of convergence of Optimized Schwarz applied to a problem defined in a bounded domain and with arbitrary number of subdomains forming a 2D array (i.e., containing cross points) is presented.

2 Equations of OSM for Poisson's in rectangular domain for the $OO0$ case

We want to solve Poisson's equation in a rectangular domain subject to nonhomogeneous Dirichlet boundary conditions, i.e,

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = g & \text{on } \partial\Omega. \end{cases} \tag{1}$$

where $\Omega = [0, L_1] \times [0, L_2]$.

We divide the physical domain into $p \times q$ overlapping rectangular subdomains. To simplify the presentation, we consider square subdomains where each side is of length H and the same overlap on each side, but the analysis presented here is also valid for arbitrary rectangles and arbitrary ovelaps. Each of these subdomains is represented by a pair of indexes, (s,r), with $s \in \{1,...,p\}$ and $r \in \{1,...,q\}$. Let h be the length of the side of each subdomain as if it were a partition with no overlap. Let us now displace (outward) each of the boundaries of the nonoverlapping subdomains by a γ amount. We have then overlapping square subdomains with side $H = h + 2\gamma$ and can use γ as a parameter to quantify the amount of overlap between subdomains. The Optimized Schwarz iteration process associated with problem (1) and with $OO0$ transmission conditions is defined, for an interior subdomain (i.e., for $1 < s < p$, $1 < r < q$), by

$$\begin{cases} \Delta u_{n+1}^{(s,r)} = f & \text{in } \Omega^{(s,r)} \\ -\frac{\partial u_{n+1}^{(s,r)}}{\partial x} + \alpha u_{n+1}^{(s,r)} = -\frac{\partial u_n^{(s-1,r)}}{\partial x} + \alpha u_n^{(s-1,r)} & \text{for } x = (s-1)h - \gamma \\ \frac{\partial u_{n+1}^{(s,r)}}{\partial x} + \alpha u_{n+1}^{(s,r)} = \frac{\partial u_n^{(s+1,r)}}{\partial x} + \alpha u_n^{(s+1,r)} & \text{for } x = sh + \gamma \\ -\frac{\partial u_{n+1}^{(s,r)}}{\partial y} + \alpha u_{n+1}^{(s,r)} = -\frac{\partial u_n^{(s,r-1)}}{\partial y} + \alpha u_n^{(s,r-1)} & \text{for } y = (r-1)h - \gamma \\ \frac{\partial u_{n+1}^{(s,r)}}{\partial y} + \alpha u_{n+1}^{(s,r)} = \frac{\partial u_n^{(s,r+1)}}{\partial y} + \alpha u_n^{(s,r+1)} & \text{for } x = rh + \gamma. \end{cases} \tag{2}$$

where $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ are, in this instance, normal derivatives and $u_{n+1}^{(s,r)}$ is the solution of the local problem (2) at the $(n+1)$ iteration in $\Omega^{(s,r)}$. The parameter α is the one which we want to tune to optimize the convergence rate of the method. Note that $\alpha = 0$ would reduce the problem to pure Neuman boundary conditions and therefore this case is not allowed. The subdomains touching the boundary have one or two boundaries that are actually physical (not artificial) boundaries. The equations for the subdomains touching the boundary are similar to (2) with the exception that one or two of the boundary conditions are Dirichlet, namely, the ones associated to the physical boundaries.

3 Recasting equations as an equivalent fixed point iteration

By linearity, we can see that the local error (of interior subdomains) of the iteration process is described by (2) with $f = 0$. Similar equations can be obtained for subdomains touching the boundary. Using separation of variables, Sturm-Liouville theory and superposition principle, we can write the local errors in the form of a series [4]. Then, using the non-homogeneous boundary conditions in each local problem, we obtain a relationship between the error series coefficients at iteration $(n+1)$ and the ones at iteration n.

Fourier Analysis of solution of PDEs defining the local error

We analyze the local error of an interior subdomain, but the same analysis holds for subdomains touching the boundary. Let $\eta_n^{(s,r)}$ be the local error in $\Omega^{(s,r)}$ at the iteration n. By superposition principle, we can write $\eta_n^{(s,r)} = \eta_{n,1}^{(s,r)} + \eta_{n,2}^{(s,r)} + \eta_{n,3}^{(s,r)} + \eta_{n,4}^{(s,r)}$, where $\eta_{n,i}^{(s,r)}$, $i = 1, \ldots, 4$, is the solution of (2) with $f = 0$, and with one non-homogeneous boundary condition and the rest homogeneous. Then, each part of the local error $\eta_n^{(s,r)}$ can be written as:

$$\eta_{n,1}^{(s,r)}(x_\ell, y_\ell) = \sum_{m=1}^{\infty} A_{n,m,1}^{(s,r)} \phi_m(x_\ell) \psi_m(H - y_\ell) \tag{3}$$

$$\eta_{n,2}^{(s,r)}(x_\ell, y_\ell) = \sum_{m=1}^{\infty} A_{n,m,2}^{(s,r)} \phi_m(y_\ell) \psi_m(x_\ell) \tag{4}$$

$$\eta_{n,3}^{(s,r)}(x_\ell, y_\ell) = \sum_{m=1}^{\infty} A_{n,m,3}^{(s,r)} \phi_m(x_\ell) \psi_m(y_\ell) \tag{5}$$

$$\eta_{n,4}^{(s,r)}(x_\ell, y_\ell) = \sum_{m=1}^{\infty} A_{n,m,4}^{(s,r)} \phi_m(y_\ell) \psi_m(H - x_\ell), \tag{6}$$

where $\phi_m(x_\ell) = \frac{\bar{\alpha}}{z_m} \sin\left(\frac{z_m x_\ell}{H}\right) + \cos\left(\frac{z_m x_\ell}{H}\right)$ and $\psi_m(x_\ell) = \frac{\bar{\alpha}}{z_m} \sinh\left(\frac{z_m x_\ell}{H}\right) + \cosh\left(\frac{z_m x_\ell}{H}\right)$ with z_m satisfying the transcendental equation

$$\tan(z) = \frac{2z\bar{\alpha}}{\bar{\alpha}^2 - z^2},$$

$\bar{\alpha} = \alpha H$, and x_ℓ and y_ℓ are local coordinates related to the global coordinates x and y by

$$\begin{aligned} x_\ell &= x - (s-1)h + \gamma \\ y_\ell &= y - (r-1)h + \gamma. \end{aligned} \tag{7}$$

Note that $\{\phi_m\}_{m\in\mathbb{N}}$ is a complete orthogonal set in $[0, H]$. Therefore, equations (3) and (5) can be seen as Generalized Fourier series in x_ℓ and equations (4) and (6) as Generalized Fourier series in y_ℓ. Then, we have that

$$A_{n,m,1}^{(s,r)} = \frac{\int_0^H \eta_{n,1}^{(s,r)}(x_\ell, y_l) \left[\frac{\bar{\alpha}}{z_m} \sin\left(\frac{z_m x_\ell}{H}\right) + \cos\left(\frac{z_m x_\ell}{H}\right)\right] dx_\ell}{\left[\frac{-\bar{\alpha}}{z_m} \sinh\left(\frac{z_m (y_\ell - H)}{H}\right) + \cosh\left(\frac{z_m (y_\ell - H)}{H}\right)\right] \int_0^H \left[\frac{\bar{\alpha}}{z_m} \sin\left(\frac{z_m x_\ell}{H}\right) + \cos\left(\frac{z_m x_\ell}{H}\right)\right]^2 dx_\ell} \tag{8}$$

Let $\beta : \mathbb{N} \times \mathbb{R} \to \{-1\} \cup [0, 1]$ such that

$$\beta(m, \bar{\alpha}) = \begin{cases} -1, & \text{if } z_m < 1 \\ \frac{1}{2}, & \text{if } z_m \geq 1 \end{cases}.$$

Then, with $y_\ell = 0$ and using integration by parts in (8) we can write

$$A_{n,m,1}^{(s,r)} = \frac{B_{n,m,1}^{(s,r)}}{z_m^{1+\beta(m,\bar{\alpha})} \left[\frac{\bar{\alpha}}{z_m} \sinh\left(\frac{z_m}{H}\right) + \cosh\left(\frac{z_m}{H}\right)\right]},$$

where $B_{n,m,1}^{(s,r)}$ is uniformly bounded for all $m \in \mathbb{N}$. The same relationship holds between $A_{n,m,i}^{(s,r)}$ and uniformly bounded quantities $B_{n,m,i}^{(s,r)}$ for $i \in \{2, 3, 4\}$. Plugging these equalities in (3)-(6) and applying the nonhomogeneous boundary conditions, we obtain the expression of the coefficients at iteration $(n+1)$ in terms of those at iteration n. For example, with a normalized overlap $\bar{\gamma} = \gamma/H$, we have for a specific index k,

$$
\begin{aligned}
B^{(s,r)}_{n+1,k,1} = & \frac{\left(z_k+\frac{\bar\alpha^2}{z_k}\right)\sinh(2\bar\gamma z_k)+2\bar\alpha\cosh(2\bar\gamma z_k)}{\left(z_k+\frac{\bar\alpha^2}{z_k}\right)\sinh(z_k)+2\bar\alpha\cosh(z_k)}B^{(s,r-1)}_{n,k,1} \\
& + \sum_{m=1}^{\infty}\left\{\frac{4z_k^{4+\beta(k,\bar\alpha)}\left[\frac{\bar\alpha}{z_k}\tanh(z_k)+1\right]\left(z_m+\frac{\bar\alpha^2}{z_m}\right)\sin((1-2\bar\gamma)z_m)}{\left[\left(z_k+\frac{\bar\alpha^2}{z_k}\right)\tanh(z_k)+2\bar\alpha\right]z_m^{1+\beta(m,\bar\alpha)}\left(z_mz_k^3+z_kz_m^3\right)}\right. \\
& \left.\frac{\left\{\tanh(z_m)\left[\bar\alpha(z_k^2+z_m^2)\sin(z_k)-z_k(\bar\alpha^2-z_m^2)\cos(z_k)\right]+z_m(\bar\alpha^2+z_k^2)\sin(z_k)\right\}}{\left[\frac{\bar\alpha}{z_m}\tanh(z_m)+1\right]\left[(z_k^2-\bar\alpha)^2\sin(2z_k)+2z_k(\bar\alpha^2+z_k^2+\bar\alpha)-2\bar\alpha z_k\cos(2z_k)\right]}B^{(s,r-1)}_{n,m,2}\right\} \\
& + \frac{\left(-z_k+\frac{\bar\alpha^2}{z_k}\right)\sinh((1-2\bar\gamma)z_k)}{\left(z_k+\frac{\bar\alpha^2}{z_k}\right)\sinh(z_k)+2\bar\alpha\cosh(z_k)}B^{(s,r-1)}_{n,k,3} \\
& + \sum_{m=1}^{\infty}\left\{\frac{4z_k^{4+\beta(k,\bar\alpha)}\left[\frac{\bar\alpha}{z_k}\tanh(z_k)+1\right]\left(z_m+\frac{\bar\alpha^2}{z_m}\right)\sin((1-2\bar\gamma)z_m)}{\left[\left(z_k+\frac{\bar\alpha^2}{z_k}\right)\tanh(z_k)+2\bar\alpha\right]z_m^{1+\beta(m,\bar\alpha)}\left(z_mz_k^3+z_kz_m^3\right)}\right. \\
& \left.\frac{\left\{\tanh(z_m)z_k(\bar\alpha^2+z_m^2)-z_m\left[-2\bar\alpha z_k+\frac{(\bar\alpha^2-z_k^2)\sin(z_k)+2\bar\alpha z_k\cos(z_k)}{\cosh(z_m)}\right]\right\}}{\left[\frac{\bar\alpha}{z_m}\tanh(z_m)+1\right]\left[(z_k^2-\bar\alpha)^2\sin(2z_k)+2z_k(\bar\alpha^2+z_k^2+\bar\alpha)-2\bar\alpha z_k\cos(2z_k)\right]}B^{(s,r-1)}_{n,m,4}\right\}.
\end{aligned}
\tag{9}
$$

Let B_n be the infinite vector containing all the error series coefficients at iteration n, i.e., $B_n = (b_{n_1}, b_{n_2}, \ldots)$ with $b_{n_j} \in \left\{B^{(s,r)}_{n,k,i} : s \in \{1,\ldots,p\}, r \in \{1,\ldots,q\}, k \in \mathbb{N}, i \in \{1,\ldots,4\}\right\}$. Then the relation between coefficients can be written as $B_{n+1} = \hat{T}B_n$, where $\hat{T} : \mathbb{R}^\infty \to \mathbb{R}^\infty$ is an infinite matrix. Note that $\hat{T} = (\hat{T}^{1,1}, \ldots, \hat{T}^{p,q})$, where $\hat{T}^{(s,r)}$ is a local operator such that $B^{(s,r)}_{n+1} = \hat{T}^{(s,r)}B_n$ with $B^{(s,r)}_{n+1}$ being a vector containing all the error coefficients of the local problem (s,r) at iteration $(n+1)$.

Our main result is the following.

Theorem 1. *For any positive value of the normalized overlap $\bar\gamma$ there exist a computable range of values of the normalized boundary parameter $\bar\alpha$ for which the OSM iteration given by (2) converges.*

For its proof it suffices to show that each of the series in (3)-(6) converge uniformly and that the error series coefficients tend to zero as the number of iteration goes to infinity.

4 Approximation of the infinite operator $\hat{T}$ by a matrix of finite dimensions

Note that the following statements hold

1. In the r.h.s. of (9), the terms containing the coefficients $B^{(s,r-1)}_{n,k,i}$, $i = 1,3$, decrease with k.

2. For a given $n \in \mathbb{N}_0$, $B_{n,m,i}^{(s,r-1)}$ is uniformly bounded in $m \in \mathbb{N}$ and $i = 1,...,4$. Moreover, $B_{n,m,i}^{(s,r-1)} \leq M/z_m$ for all $m \in \mathbb{N}$ and some $M > 0$.
3. For any number $\delta > 0$ there exists a number k_δ, such that for $k > k_\delta$, the sum of the absolute values of the terms in the r.h.s. of (9) is less than δ.
4. For any number $\delta > 0$ there exists a number m_δ, such that for every $k \in \mathbb{N}$ the sum of the absolute values of the terms in the r.h.s. of (9) corresponding to for $m > m_\delta$ is less than δ.

Let $(B_n)_{|k\leq k_\delta}$ denote the vector resulting after discarding all the entries of B_n corresponding to $k > k_\delta$. Then, based on the above three facts, we can write

$$(B_{n+1})_{|k\leq k_\delta} = \left(\hat{T}\,(B_n)\right)_{|k\leq k_\delta} = \hat{T}_\delta\left((B_n)_{|k\leq k_\delta}\right) + \xi_{n+1,k_\delta}((B_n)_{|k>k_\delta}), \qquad (10)$$

where $\hat{T}_\delta$ is a finite matrix obtained by discarding the rows and columns of $\hat{T}$ related to the coefficients pertaining to $k > k_\delta$, and $\xi_{n+1,k_\delta}((B_n)_{|k>k_\delta})$ is the error obtained by approximating $(B_{n+1})_{|k\leq k_\delta}$ by $\hat{T}_\delta\left((B_n)_{|k\leq k_\delta}\right)$.

We will discuss in the next section situations in which $\rho(\hat{T}_\delta) < 1$, i.e., the spectral radius of $\hat{T}_\delta$ is less than one. In the rest of this section we show that in addition the error $\xi_{n+1,k_\delta}((B_n)_{|k>k_\delta})$ tends to zero as $n \to \infty$, and consequently $B_n \to 0$ as $n \to \infty$.

A necessary condition for convergence of Optimized Schwarz is that $B_n \to 0$ as $n \to \infty$. Note that each entry of $\xi_{n+1,k_\delta}((B_n)_{|k>k_\delta})$ is the truncation error that results after truncating the series in the formulas of the coefficients $B_{n+1,k,i}^{(s,r)}$, by keeping only the terms corresponding to $k \leq k_\delta$. Thus, as it can be seen in (9), each entry of $\xi_{n+1,k_\delta}((B_n)_{|k>k_\delta})$ is just a linear combination of the entries of $(B_n)_{|k>k_\delta}$. Note also that the entries of $(B_n)_{|k>k_\delta}$ are linear combinations of the entries of B_{n-1}. Hence, based on the four facts from above, we can choose a large enough k_δ so that the entries of $(B_{n+1})_{|k>k_\delta}$ and $\xi_{n+1,k_\delta}((B_n)_{|k>k_\delta})$ are as small as desired.

Using equation (10) recursively, we obtain the following equation

$$(B_{n+1})_{|k\leq k_\delta} = \hat{T}_\delta^{n+1}((B_0)_{|k\leq k_\delta}) + \sum_{j=1}^{n+1} \hat{T}_\delta^{n+1-j}(\xi_{j,k_\delta}((B_{j-1})_{|k>k_\delta})). \qquad (11)$$

Using (11), the four facts from above, and assuming that the spectral radius of $\hat{T}_\delta$ is less than one and that remains practically constant for large values of k_δ, it can be shown that given a $0 < \varepsilon < 1$ there exists a n_ε such that $||B_n||_\infty \leq \varepsilon||B_0||_\infty$ for all $n \geq n_\varepsilon$. Repeating this argument, we can then show that $\lim_{n\to\infty} B_n = 0$. Hence in order to prove that $B_n \to 0$ as $n \to \infty$, it suffices to show that $\rho(\hat{T}_\delta) < 1$ and that it remains practically constant for large values of k_δ. We show this in the next section.

It can be shown that the series describing the local errors converge uniformly in $\Omega^{(s,r)}$. This implies that if each term of the error series goes to zero as n goes to infinity, so will do the series. Thus, given that $B_n \to 0$ as $n \to \infty$, i.e., the coefficients of the error series go to zero as n goes to infinity, the error of the iterative process con-

verges to zero as n goes to infinity, which means that Optimized Schwarz converges for the given Poisson's problem for any initial error.

5 Spectral Radius of $\hat{T}_\delta$

The spectral radius of $\hat{T}_\delta$ describes the convergence rate of the Optimized Schwarz method. Thus, we define the optimal normalized boundary parameter $\bar{\alpha} = \alpha H$ as the one which minimizes the spectral radius of $\hat{T}_\delta$ and thus gives the optimal asymptotic convergence rate.

The values of the entries of the matrix $\hat{T}_\delta$ depend on the normalized overlap $\bar{\gamma}$, $\bar{\alpha}$ and the truncation parameter k_δ. The structure of the matrix depends on k_δ, p, q and the way we order the entries of B_n, i.e., the way we order each coefficient $B_{n,k,i}^{(s,r)}$ based on its values of s, r, k and i. For the ordering we have chosen, we computed the spectral radius of the resulting matrix $\hat{T}_\delta$, for $\bar{\gamma} \in \{0, 0.001, 0.01, 0.04, 0.08\}$, a set of values of $\bar{\alpha}$ in the range $[0.1, 500]$, $k_\delta \in \{3, 5, 10, 20, 50, 100\}$, and $p, q \in \{4, 5, 10, 20, 30\}$. In these computations we have observed the following.

1. There exist values of $\bar{\alpha}$ for which the spectral radius of $\hat{T}_\delta$ is less than one.
2. For a given $\bar{\gamma}$ and the range of $\bar{\alpha}$ considered in the experiments, $\rho(\hat{T}_\delta)$ has a local minimum, and it approaches a constant less than one for large values of $\bar{\alpha}$.
3. Given $\bar{\gamma}$, $\bar{\alpha}$, p and q, the value of $\rho(\hat{T}_\delta)$ remains practically constant for large enough k_δ (see Figure 2).
4. For a given $\bar{\gamma}$, the optimal spectral radius of $\hat{T}_\delta$ remains practically constant as p and q increase.

In Figure 1, the results for the cases $\bar{\gamma} = 0.001$ and $\bar{\gamma} = 0.01$, with $p, q = 10$, $k_\delta = 20$, $\bar{\alpha} \in [1, 100]$, are shown.

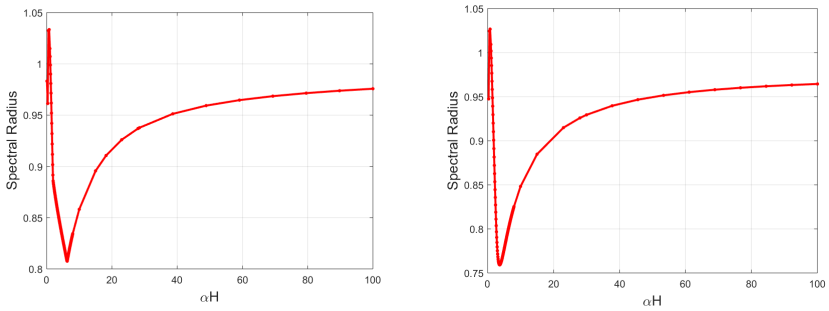


Fig. 1 (a) Spectral Radius of $\hat{T}_\delta$ vs. $\bar{\alpha}$ for $\bar{\gamma} = 0.001$, $p, q = 10$, $k_\delta = 20$ and $\bar{\alpha} \in [0.1, 100]$. (b) Spectral radius of $\hat{T}_\delta$ vs. $\bar{\alpha}$ for $p, q = 10$, $k_\delta = 20$, $\bar{\gamma} = 0.01$ and $\bar{\alpha} \in [0.1, 100]$

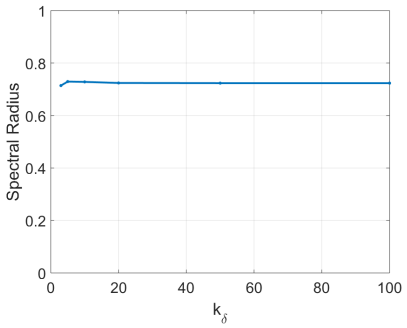


Fig. 2 Spectral radius of $\hat{T}_\delta$ vs. k_δ for $p,q = 10$, $\bar{\gamma} = 0.01$ and $\bar{\alpha} = 3.9697$

6 Further comments and conclusion

In the case of elliptic problems with varying coefficients, the same procedure can be applied to obtain an operator $\hat{T}$ such that $B_{n+1} = \hat{T}B_n$ as long as the coefficients are separable as products of one-variable functions. In this case, as well as in the constant coefficients case, the entries of the operator $\hat{T}$ depend on values and first derivatives of ϕ_m and ψ_m with $m \in \mathbb{N}$ at specific points. Note that in the constant coefficient case an explicit formula can be obtained for ϕ_m and ψ_m. In the varying coefficients case, an explicit formula for ϕ_m and ψ_m may not always be available. However, we can still compute values of ϕ_m and ψ_m and their first derivatives at specific points using numerical methods and then use these values to compute $\rho(\hat{T}_\delta)$.

In conclusion, we analyzed the convergence of the Optimized Schwarz method applied to Poisson's equation in a bounded rectangular domain subject to nonhomogeneous Dirichlet boundary conditions and transmission conditions of the family $OO0$. The spectral radius of $\hat{T}_\delta$ can be less than one for any positive amount of overlap. One can obtain the optimal boundary parameter that minimizes this spectral radius. We outlined a proof showing that this bound on the spectral radius, together with other results, can guarantee convergence of OSM for the problem studied.

References

1. Daniel Bennequin, Martin J. Gander, Loic Gouarin, Laurence Halpern. Optimized Schwarz Waveform Relaxation for Advection Reaction Diffusion Equations in Two Dimensions. *Numerische Mathematik*, 134:513–567,2016.
2. Victorita Dolean, Pierre Jolivet, and Frédéric Nataf. An introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation. *SIAM*, Philadelphia, 2015.
3. Martin J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44:699–731, 2006.
4. Richard Haberman *Applied Partial Differential Equations with Fourier Series and Boundary Value Problems.* Prentice Hall, fourth edition, Englewood Cliffs, NJ, 2003.
5. Frédéric Magoulés, Abal-Kassim Cheik Ahamed, and Roman Putanowicz. Optimized Schwarz method without overlap for the gravitational potential equation on cluster of graphics processing unit. *International Journal of Computer Mathematics*, 93: 955-980 (2016)

The HTFETI method variant gluing cluster subdomains by kernel matrices representing the rigid body motions

Alexandros Markopoulos, Lubomír Říha, Tomáš Brzobohatý, Ondřej Meca, Radek Kučera, and Tomáš Kozubek

IT4Innovations National Supercomputing Center, VŠB - Technical University of Ostrava, 17. listopadu 15/2172, Ostrava, Czech Republic
{alexandros.markopoulos,lubomir.riha}@vsb.cz

Abstract. The proposed algorithm called the Hybrid Total Finite Element Tearing and Interconnecting method (HTFETI) is a variant of the TFETI domain decomposition method suitable for large-scale problems with hundreds of thousands of subdomains. The floating subdomains are gathered into several groups belonging to individual clusters. We use the new idea consisting in gluing the cluster subdomains using kernel matrices defined by the rigid body motions. This technique reduces the size of the coarse problem. While the size of the coarse problem depends linearly on the number of subdomains in the classical TFETI method, it depends linearly on the number of clusters in the HTFETI method. The zero weighted averages across the interfaces of neighbouring subdomains (an alternative to the constraints enforcing the continuity across the corners used, e.g., in the FETI-DP method) improve conditioning of the resulting system of linear equations.

1 Introduction

The history of the FETI (Finite Element Tearing and Interconnecting) method [6] is longer than twenty years and over the years, numerous variants have been developed (FETI-DP method [2, 5], T(otal)FETI method [4] etc). The important impulse for development of new FETI variants was given by the implementation on more sophisticated computer architectures, where parallel processors are grouped into clusters. From the point of view of minimal communications, it is reasonable to copy the computer architecture into the FETI method that lead to the hybrid (two-level) FETI methods. The FETI–FETI-DP method proposed in [7, 8] combines the classical FETI method used on the global level with the FETI-DP method used on the clusters. In this paper, we deal with the TFETI–TFETI method that uses the TFETI method on both levels [3, 10]. It will be denoted as the H(ybrid)TFETI method. The new approach presented in this paper is called HTFETI$_{ker}$. In this method the gluing of subdomains (belonging to one cluster) is done using kernels of the local subdomains. This technique accelerates iterations like in the case of the transformation of basis discussed in [7].

In the numerical experiments we compare HTFETI$_{ker}$ with HTFETI$_{cor}$ method where the subdomains belonging to one cluster are glued by the Lagrange multipliers (LMs) corresponding to the corner nodes. Such method is similar to the FETI–FETI-DP method. The basic idea of the two-level FETI method is graphically explained in the following benchmark, in which we introduce also respective notation.

In order to simplify the presentation of the method, we use a simple cube benchmark with a hierarchical decomposition and discretization depicted in Fig. 1.

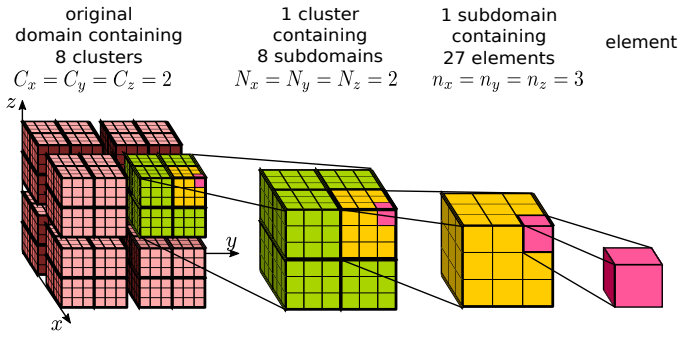


Fig. 1. Two levels of decomposition: 2 clusters ($C = 2$), 2 subdomains ($N = 2$), 3 elements ($n = 3$) in each space dimension

This hierarchical decomposition and discretization consists of three levels:

- Level 1 - decomposition into clusters is controlled by parameters C_x, C_y, and C_z (numbers of clusters in x, y, and z direction). Each cluster occupies one computational node.
- Level 2 - each cluster is decomposed into the subdomains controlled by parameters N_x, N_y, and N_z (numbers of subdomains in x, y, and z direction).
- Level 3 - each subdomain is discretized uniformly by hexahedral finite elements handled by parameters n_x, n_y, n_z (numbers of elements in x, y, and z direction).

If, for example, the number of clusters in all directions is the same $C_x = C_y = C_z = 2$, the description in the text is simplified to $C = 2$. This simplified notation is also applied to subdomains N and elements n.

2 Cluster constraints

2.1 Types of subdomains-gluing

In the following part we are going to focus on the constraints among subdomains in the cluster. All the details of the HTFETI method and also the derivation of

the algorithm can be found in [10]. The notation used in this section relates to the same paper.

Compared to the FETI method, in the described hybrid variant the neighboring subdomains are grouped into clusters using additional constraints. Together, with the commonly used joining of subdomains via corner nodes known in the FETI-DP method, we present a new technique based on the kernels of stiffness matrices. Such an approach requires a robust algorithm for factorizing singular matrices but, on the other hand, it simplifies implementation of the HTFETI method. Implicitly, it enforces zero averages across the faces between the neighbouring subdomain.

For simplification, let us use the cluster consisting of two subdomains Ω_j and Ω_k (see Fig. 2). The stiffness matrix of I-th cluster then will be

$$\tilde{\mathbf{K}}_I = \begin{pmatrix} \mathbf{K}_{j:k} & \mathbf{B}_{c,j:k}^\top \\ \mathbf{B}_{c,j:k} & \mathbf{O} \end{pmatrix} = \begin{pmatrix} \mathbf{K}_j & \mathbf{O} & \mathbf{B}_{c,j}^\top \\ \mathbf{O} & \mathbf{K}_k & \mathbf{B}_{c,k}^\top \\ \mathbf{B}_{c,j} & \mathbf{B}_{c,k} & \mathbf{O} \end{pmatrix} \tag{1}$$

which corresponds to Eq. (14) in [10] if interval $j : k$ consists of j and k only. Here, $\mathbf{K}_j$ and $\mathbf{K}_k$ are stiffness matrices, $\mathbf{B}_{c,j}$, $\mathbf{B}_{c,k}$ are linear constraints keeping both subdomains together, and $\mathbf{O}$ is a zero matrix with the appropriate size. In the next subsections, let us explain how to choose the blocks $\mathbf{B}_{c,j}$, $\mathbf{B}_{c,k}$.

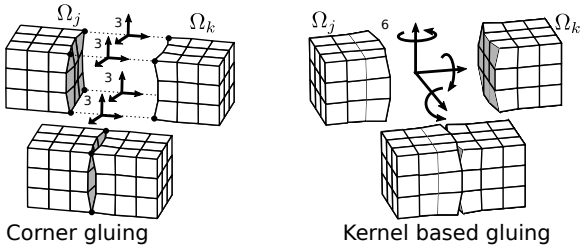


Fig. 2. Two-subdomain bonding, corners versus kernels: 3 forces per corner node (12 LM in total) / 3 forces and 3 moments per interface (6 LM in total).

Corner strategy - HTFETI$_{cor}$ method Using this method, $\mathbf{B}_{c,j}$, $\mathbf{B}_{c,k}$ are signed booleans matrices that enforce the connectivity across the corner nodes (see Fig. 2 left). The structure is similar as matrix of constraints in the FETI method (commonly denoted as $\mathbf{B}$).

Kernel strategy - HTFETI$_{ker}$ method The kernel strategy glues the domains Ω_j and Ω_k in a weaker sense using the kernel $\mathbf{R}_j$ of matrix $\mathbf{K}_j$. Instead of enforcing relative zero displacements in particular nodes belonging to the interface $\Gamma_{jk} = \overline{\Omega}_j \cap \overline{\Omega}_k$, we prescribe constraints acting onto all DOFs belonging to the face Γ_{jk}. The number of these constraints is determined by the defect d of $\mathbf{K}_j$ (and $\mathbf{K}_k$) that is $d = 6$ for the three-dimensional linear elasticity problems.

Let $\mathbf{Q}_j$ be an appropriate permutation matrix separating $\mathbf{R}_j$ into two parts:

$$\mathbf{Q}_j\mathbf{R}_j = \begin{pmatrix} \mathbf{R}_{j,\Omega_j\setminus\Gamma_{jk}} \\ \mathbf{R}_{j,\Gamma_{jk}} \end{pmatrix},$$

where $\mathbf{R}_{j,\Gamma_{jk}}$ is given by the rows of $\mathbf{R}_j$ belonging to the interface $\Gamma_{jk} = \overline{\Omega}_j \cap \overline{\Omega}_k$ and $\mathbf{R}_{j,\Omega_j\setminus\Gamma_{jk}}$ contains by the remaining rows. It is required that $\mathbf{R}_{j,\Gamma_{jk}} \in \mathcal{R}^{m\times d}$, where $m \geq d$. In the case of a three-dimensional linear elasticity problem, this requirement is always satisfied if the common interface between two neighboring subdomains is given by at least three nodes not lying in one line. The parameter m is then equal to 9 (number of all degrees of freedom belonging to this set of nodes). Then we define $\mathbf{B}_{c,j}$ and $\mathbf{B}_{c,k}$ as follows:

$$\mathbf{B}_{c,j}^\top = \mathbf{Q}_j^\top \begin{pmatrix} \mathbf{O} \\ \mathbf{R}_{j,\Gamma_{jk}} \end{pmatrix}, \quad \mathbf{B}_{c,k}^\top = \mathbf{Q}_k^\top \begin{pmatrix} \mathbf{O} \\ -\mathbf{R}_{j,\Gamma_{jk}} \end{pmatrix}, \tag{2}$$

where the permutation matrix $\mathbf{Q}_k$ maps the rows of $-\mathbf{R}_{j,\Gamma_{jk}}$ (in Ω_k) onto the corresponding rows of $\mathbf{R}_{j,\Gamma_{jk}}$ (in Ω_j). The non-sigularity of the matrices $\mathbf{B}_{c,j}\mathbf{R}_j$ and $\mathbf{B}_{c,k}\mathbf{R}_k$ guarantees that the subdomains $\boldsymbol{\Omega}_j$ and $\boldsymbol{\Omega}_k$ are properly glued together [9]. The gluing condition is schematically depicted in the right Fig. 2. The presented idea can be simply extended to the clusters with more than two subdomains. The approach is also applicable to non-singular matrices (transient problems).

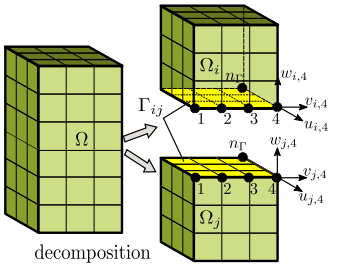


Fig. 3. Domain decomposition of Ω body into 2 subdomains Ω_i and Ω_j.

Example: Constraints assembled from the analytically computed kernel. Let us explain some general ideas regarding the analytical form available for kernels in linear elasticity . Let the nodes shared by $\boldsymbol{\Omega}_i$ and $\boldsymbol{\Omega}_j$ lying on the interface $\Gamma_{i,j} = \overline{\boldsymbol{\Omega}}_i \cap \overline{\boldsymbol{\Omega}}_j$ depicted in Fig. 3 be indexed by the set $\mathcal{G} = \{1, 2, \cdots, n_{\Gamma_{i,j}}\}$. Let the displacement vector of the g-th node $\mathbf{x}_g = (x_g,\ y_g,\ z_g) \in \boldsymbol{\Gamma}_{i,j}$ be denoted $\mathbf{u}_{i,g} = (u_{i,g},\ v_{i,g},\ w_{i,g})$ with respect to Ω_i and $\mathbf{u}_{j,g} = (u_{j,g},\ v_{j,g},\ w_{j,g})$ with respect to Ω_j. It follows from the mechanical arguments that two subdomains are kept together by 3 forces and 3 moments acting across the whole interface Γ_{ij} that avoids mutual movements and rotations. It can be achieved by zero averages

of displacements

$$\sum_{g=1}^{n_\Gamma}(u_{i,g}-u_{j,g})=0,\ \sum_{g=1}^{n_\Gamma}(v_{i,g}-v_{j,g})=0,\ \sum_{g=1}^{n_\Gamma}(w_{i,g}-w_{j,g})=0,$$

and rotations

$$\sum_{g=1}^{n_\Gamma}((u_{i,g}-u_{j,g})\cdot y_g-(v_{i,g}-v_{j,g})\cdot x_g)=0,$$

$$\sum_{g=1}^{n_\Gamma}((u_{i,g}-u_{j,g})\cdot z_g-(w_{i,g}-w_{j,g})\cdot x_g)=0,$$

$$\sum_{g=1}^{n_\Gamma}((v_{i,g}-v_{j,g})\cdot z_g-(w_{i,g}-w_{j,g})\cdot y_g)=0$$

across $\Gamma_{i,j}$. It also guarantees that the subdomains are sufficiently and optimally bonded together with the minimal number of constraints.

Apart from the natural *accelerating property*, there is also another significant feature of kernel-based $\mathbf{B}_{c,i}$. Since its constraints enforce the equality across the interface on average, the Dirichlet preconditioner acts on the whole interface as well and it is completely adopted from (T)FETI method in an unchanged form.

2.2 Rank of the cluster constraint matrix $\mathbf{B}_{c,j:k}$

Sufficient mutual gluing of all cluster subdomains realized by kernels requires 6 constraints per interface between two neighboring subdomains. The comparison with the corner strategy will be shown on the cube problem. Since the matrix $\mathbf{B}_{c,j:k}$ is always assembled without linearly dependent constraints, the rank and number of rows are equal.

Academic problem For the sake of clarity, the cube problem is uniformly decomposed into subdomains by setting: $C=1$ and $N=2,3,\cdots,10$. Thanks to a simple cube geometry and the uniform discretization and decomposition, we can derive the dependency between the number of subdomains N and the rank of the cluster matrix $\mathbf{B}_{c,j:k}$. If the corner strategy is used, three following situations can occur. The node is shared by two subdomains (then it produces $3\cdot 1$ LM), by four subdomains ($3\cdot 3$ LM) or by eight subdomains ($3\cdot 7$ LM). In the first case, the subdomains are glued using corner nodes, the dimension of $\mathbf{B}_{c,j:k}$ is

$$\operatorname{rank}\left(\mathbf{B}^{cor}_{c,j:k}\right)=21(N-1)^3+54(N-1)^2+36(N-1). \tag{3}$$

In the case $\mathbf{B}_{c,j:k}$ is assembled via parts of the kernels, each common interface generates 6 LM and the dimension is

$$\operatorname{rank}\left(\mathbf{B}^{ker}_{c,j:k}\right)=18N^2(N-1). \tag{4}$$

The ratio between "corner" and "kernel" case for $N \to \infty$ is

$$\lim_{N\to\infty} \frac{\operatorname{rank}\left(\mathbf{B}^{cor}_{c,j:k}\right)}{\operatorname{rank}\left(\mathbf{B}^{ker}_{c,j:k}\right)} = \frac{7}{6} \approx 1.1667. \tag{5}$$

In the numerical tests presented later we have used a variant with 1000 subdomains ($N = 10$) for each cluster. The kernel strategy exhibits an interesting property because it provides fewer iterations, although in the corner strategy (in this particular case) the matrix $\mathbf{B}^{cor}_{c,j:k}$ contains 23.5% more constraints.

3 Numerical test

The described algorithms were implemented into our ESPRESO (ExaScale PaRallel FETI SOlver) package developed at IT4Innovations National Supercomputing Center in Ostrava, the Czech Republic [11, 1].

For these computations we used facilities of IT4Innovations Czech national supercomputing center (www.it4i.cz), namely Salomon cluster. The Salomon cluster consists of 1008 compute nodes. Each node contains 24 core Intel Xeon E5-2680v3 processors and 128 GB RAM. The interconnect is a 7D Enhanced hypercube InfiniBand.

We varied the decomposition and discretization parameters on a cube benchmark test in order to demonstrate the scalability of our method. The cube (30 mm) is made of steel with the following parameters: Young's modulus $E = 2.1 \cdot 10^5$ MPa, Poisson's ratio $\mu = 0.3$, density $\rho = 7850$ kg/m^3, and gravity constant $g_{x_1} = 9.81$ m/s^2. The cube is fixed on the plane $x = 0$, and loaded by its own weight in the x direction.

The problem is solved by the projected preconditioned conjugate gradient method. The iterations are stopped after the relative preconditioned residual is reduced by stopping criterion to preconditioned residual $\varepsilon = 1 \cdot 10^{-4}$. The first test shows weak scalability for the benchmark depicted in Fig. 1 with one cluster, a fixed number of DOFs on each subdomain, and a variable number of subdomains. The considered parameters are: $C = 1$, $N = 2,\ 3,\ \cdots,\ 12$ and $n = 10$. The initial and last variant contain $27,783$ DOFs and $5,314,683$ DOFs, respectively. The linear system is preconditioned by the Dirichlet preconditioner. In Fig. 4 left, the problem is decomposed uniformly. Naturally, the TFETI method provides the best results. For the HTFETI$_{ker}$ method, the number of iterations slightly increases with the increasing number of subdomains N^3. The hybrid variant with corners (the HTFETI$_{cor}$ method) exhibits the worst results of all three methods. On the other hand when METIS is used as the decomposer (Fig. 4 right), the TFETI method can lose the scalability due to the irregular interface. The HTFETI$_{cor}$ method is also influenced by the decomposition, but the HTFETI$_{ker}$ method keeps the relatively same performance (a slightly increasing number of iterations) as in the uniform decomposition case.

Result of similar tests with a larger number of DOFs per subdomain (parameters: $C = 1$, $N = 2,\ 3,\ \cdots,\ 6$, $n = 20$, DOFs ranging from $206,763$ to

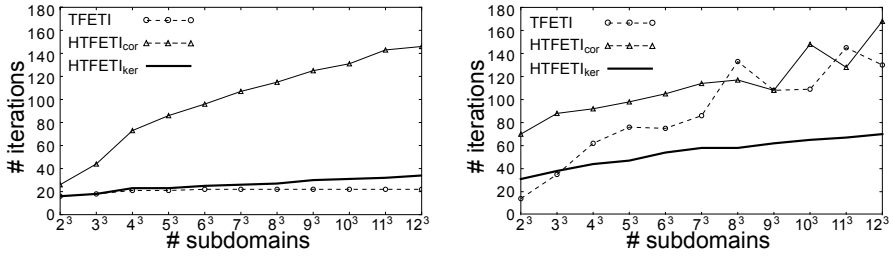


Fig. 4. Decomposition: uniform-left, METIS-right; $C = 1,\ N = 2, 3, \cdots, 12,\ n = 10$. Number of unknowns ranges from $27,783$ to $5,314,683$.

$5,314,683$) are displayed in Fig. 5. For a uniform decomposition, the TFETI and HTFETI$_{ker}$ method exhibit an equal number of iterations. It implies that if the interface is large enough (in this case 20×20 nodes versus 10×10), the TFETI method can be replaced by the HTFETI$_{ker}$ method containing one cluster. However, the HTFETI$_{ker}$ method is more expensive in preprocessing and partially also during the iterations. On the other hand, as it was already observed, when METIS is used, the TFETI method loses scalability faster, and therefore the utilization of the HTFETI method can be meaningful.

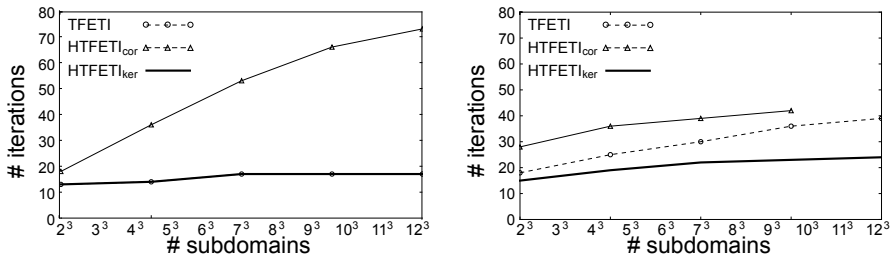


Fig. 5. Decomposition: uniform-left, METIS-right; $C = 1,\ N = 2, 3, \cdots, 6,\ n = 20$. Number of unknowns from $206,763$ to $5,314,683$.

The next set of numerical experiments in Fig. 6 shows weak scalability with the lumped preconditioner (the number of iterations on the left, solver time on the right) up to $1,259,712$ subdomains and 10.4 billion unknowns. Because of the very large number of subdomains, the TFETI method cannot be used for all the settings, and for this reason, it is not included in this comparison. However, both diagrams show weak scalability of the HTFETI method. It is also seen that the variant based on kernels requires three times fewer iterations compared to the case with corners.

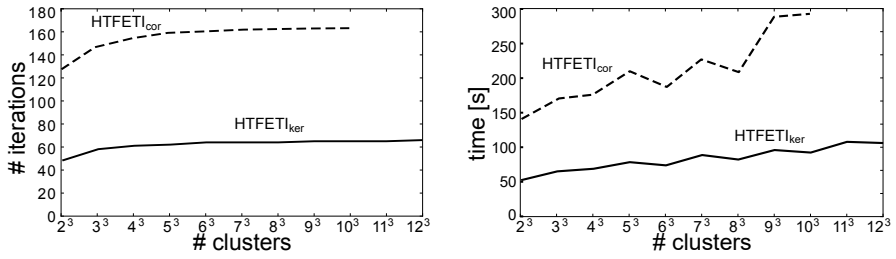


Fig. 6. HTFETI, uniform decomposition; C = 2,3,...,12, N = 9, n = 14. Number of unknowns from $4,858,2831$ to $10,390,538,091$.

4 Conclusion

This work presents the Hybrid variant of the Total FETI method.The main idea stems from the work published in [7], where the FETI and FETI-DP method are combined. Here, the presented version is the TFETI-TFETI method that uses the TFETI method on both levels. In the newly proposed variant, the subdomains are not glued together by corners but through the whole interface between each neighboring pair of subdomains via the kernels of the stiffness matrices. The numerical tests show efficiency of our algorithm. The very promising results were obtained for non-uniform decompositions. The Hybrid TFETI method based on kernels exhibits better weak scalability compared to the TFETI method.

5 Acknowledgment

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project IT4Innovations excellence in science - LQ1602 and from the Large Infrastructures for Research, Experimental Development and Innovations project IT4Innovations National Supercomputing Center LM2015070.

References

1. ESPRESO - Exascale Parallel FETI Solver, http://espreso.it4i.cz `http://espreso.it4i.cz`
2. A. Klawonn, O.B.W., Dryja, M.: Dual-primal feti methods for three-dimensional elliptic problems with heterogeneous coefficients. SIAM Journal on Numerical Analysis 40, 159–179 (2002)
3. Brzobohatý, T., Jarošová, M., Kozubek, T., Menšík, M., Markopoulos, A.: The hybrid total FETI method. In: Proceedings of the Third International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering. Civil-Comp, Ltd. (2011)
4. Dostál, Z., Horák, D., Kučera, R.: Total FETI - an easier implementable variant of the FETI method for numerical solution of elliptic PDE. Commun Numer Meth En 196, 1155–1162 (2006)

5. Farhat, C., Lesoinne, M., LeTallec, P., Pierson, K., Rixen, D.: FETI-DP: a dual-primal unified FETI method, part i: A faster alternative to the two-level FETI method. International Journal for Numerical Methods in Engineering 50(7), 1523–1544 (2001)
6. Farhat, C., Roux, F.X., Oden, J.T.: Implicit parallel processing in structural mechanics. Elsevier Science SA (1994)
7. Klawonn, A., Rheinbach, R.: Highly scalable parallel domain decomposition methods with an application to biomechanics. ZAMM 1, 5–32 (2010)
8. Klawonn, A., Lanser, M., Rheinbach, O.: Toward extremely scalable nonlinear domain decomposition methods for elliptic partial differential equations. SIAM J. Sci. Comput. 37(6), C667–C696 (jan 2015), http://dx.doi.org/10.1137/140997907
9. Kučera, R., Kozubek, T., Markopoulos, A.: On large-scale generalized inverses in solving two-by-two block linear systems. Linear Algebra and its Applications 438(7), 3011–3029 (apr 2013)
10. Merta, M., Riha, L., Meca, O., Markopoulos, A., Brzobohaty, T., Kozubek, T., Vondrak, V.: Intel xeon phi acceleration of hybrid total feti solver. Advances in Engineering Software 112(Supplement C), 124 – 135 (2017), http://www.sciencedirect.com/science/article/pii/S0965997816302745
11. Říha, L., Brzobohatý, T., Markopoulos, A., Meca, O., Kozubek, T.: Massively parallel hybrid total feti (htfeti) solver. In: Proceedings of the Platform for Advanced Scientific Computing Conference. pp. 7:1–7:11. PASC '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2929908.2929909

Small coarse spaces for overlapping Schwarz algorithms with irregular subdomains

Olof B. Widlund and Clark R. Dohrmann

1 Introduction

Coarse spaces are at the heart of many domain decomposition algorithms. Building on the foundation laid in [10], we have an ongoing interest in the development of coarse spaces based on energy minimization concepts; see [2, 4, 5, 6].

This paper is a short report on a project which substantially extends results in a DD21 conference paper, [7], and which now has resulted in an archival publication [8]. Our work primarily concerns two-level overlapping Schwarz methods and is exclusively for low order, conforming finite element approximations of three–dimensional elliptic problems. What is new in this paper are some variants of the algorithms reported in [8]. The focus of this study is the development of smaller coarse spaces which, to the extent possible, will give us similar rates of convergence as for those developed in the past. Extensive large scale experiments show that this is possible and important; see e.g. [11] in these proceedings.

The domain of a scalar elliptic or elasticity operator is partitioned into non-overlapping subdomains Ω_i each of which is the union of elements. We use *nodal equivalence classes* of finite element nodes on the interface, i.e., the nodes that belong to more than one subdomain boundary, in the construction of our coarse spaces. Two such nodes belong to the same equivalence class if they belong to the same set of subdomain boundaries. The *coarse nodes* are associated with those equivalence classes which are maximal in the sense that they are not subsets of any other. In many cases, the coarse nodes are simply the vertices of the subdomains but there are also other cases which are identified automatically by our algorithm. Each in-

Olof B. Widlund
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, e-mail: widlund@cims.nyu.edu

Clark R.Dohrmann
Computational Solid Mechanics and Structural Dynamics Department, Sandia National Laboratories, Albuquerque, New Mexico, 87185, e-mail: crdohrm@sandia.gov

terface node n is thereby associated with a set of coarse nodes $\mathscr{C}_n$. A coarse node c is included in $\mathscr{C}_n$ if the equivalence class of n is a subset of that of c. For each coarse node, we will construct one coarse basis function for scalar elliptic and six for elasticity problems, which span the coarse space.

2 Elliptic Problems and the Coarse Basis Functions

In our study, we consider scalar elliptic problems defined in terms of a bilinear form

$$\int_\Omega \rho \nabla u \cdot \nabla v \, dx$$

where $\rho(x) > 0$ and constant $= \rho_i$ in each subdomain Ω_i into which Ω has been partitioned. The functions u and v belong to a subspace of $H^1(\Omega)$ subject to a Dirichlet condition on $\partial\Omega$ or a subset thereof. We also consider linear, compressible elasticity defined by a bilinear form

$$2\int_\Omega \mu \ \varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v}) \, dx + \int_\Omega \lambda \ \text{div} \ \boldsymbol{u} \ \text{div} \ \boldsymbol{v} \, dx,$$

where $\mu(x)$ and $\lambda(x)$ are the positive Lamé parameters, $\varepsilon_{ij}(\mathbf{u}) = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})$, and $\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v}) := \sum_{i=1}^3 \sum_{j=1}^3 \varepsilon_{ij}(\mathbf{u})\varepsilon_{ij}(\mathbf{v})$. The Lamé parameters are also assumed constant, μ_i and $\lambda_i = (2\mu_i\nu_i)/(1-2\nu_i)$, in Ω_i, with $0 < \nu_i < 1/2$. This variational problem is posed in a subspace of $(H^1(\Omega))^3$ determined by a Dirichlet condition. The energy of these systems are defined by these bilinear forms.

Three recipes for the construction of coarse space elements have been developed in [8], each defined in terms of a partition of unity for each interface node. The simplest one, referred to as Option 1, is given by

$$p_{nc} := 1/N_c, \tag{1}$$

where $N_c := |\mathscr{C}_n|$. Of the two other recipes, the one relevant for this paper is the third defined in terms of $d_i(n), i = 1, \ldots, N_c$, the distances between an interface node n and the $c_i \in \mathscr{C}_n$, and given by

$$p_{nc_i} := \frac{1/d_i(n)}{1/d_1(n) + 1/d_2(n) + \ldots + 1/d_{N_c}(n)}. \tag{2}$$

This Option 2 is the only one used in the experiments reported in this paper.

The values of these functions are used as Dirichlet data and extended into the interior of the subdomains, minimizing the energy, and resulting in continuous coarse basis functions for scalar elliptic problems. The support of a coarse basis function associated with the coarse node c is the union of the closure of all Ω_j with c on their boundaries. For elasticity, we multiply the scalar function p_{nc} by a 3×6 matrix with

columns forming a basis for the space of rigid body modes prior to extending the resulting values on the interface into the interiors of the subdomains. We note that the resulting finite element functions will all be continuous given that there are no jumps in the Dirichlet data across the interface.

The choice of minimal energy extensions results in coarse basis functions which sum to 1, in any subdomain that does not touch the Dirichlet boundary, for the scalar elliptic problems and rigid body modes for elasticity. This fact shows that the *null space* condition will be satisfied, a condition necessary to obtain a convergence rate bounded independently of the number of subdomains for any domain decomposition algorithm. This approach works well even for subdomains with irregular boundaries such as those obtained from mesh partitioners. The crucial part of the analysis of an overlapping Schwarz algorithm requires a bound on the sum of the energy of the components of the coarse space and of the local spaces of finite element functions, supported in the overlapping subdomains chosen to define the local problems, in terms of the energy of the sum of these functions; see, e.g., [15, Chapter 2]. The choice of minimal energy extensions therefore makes sense also for this reason.

The development of domain decomposition theory has often focused on the effect of large discontinuities of the coefficient. Thus, for iterative substructuring algorithms, based on non-overlapping subdomains, a number of strong results have been developed for elliptic problems where the coefficients are constant or vary slowly inside the subdomains but without any restrictions on their variation across the interface between the subdomains; see, e.g., [15, Chapters 4–6] and [14]. Many of these algorithms are well-defined for arbitrary subdomains although the theory has been fully developed mostly for subdomains that are tetrahedral or unions of a few large tetrahedra; we note that some of the standard tools now have been extended to Lipschitz subdomains, see [8]. In contrast, the theory for two–level additive Schwarz methods is developed only for constant coefficients in [15, Section 3.2]. However, the classical coarse spaces for these Schwarz algorithms have been shown to be stable for *quasi–monotone* coefficients in [9]; for a related condition, see Assumption 1 of this paper. The results in [9] considerably expanded the class of subdomain coefficients for which results quite similar to those for constant coefficients became possible.

To derive the final bounds for our overlapping Schwarz algorithms, we also need to consider the components associated with local problems on overlapping subdomains, which are often constructed by extending the nonoverlapping subdomains, Ω_i, into which the given domain Ω has been decomposed, by adding one or a few layers of elements. Observing that we need solvers for Dirichlet problems on the original subdomains Ω_i to construct the coarse basis functions, we will in some of our numerical experiments instead use the Ω_i as part of the covering. In addition, to cover all of the domain Ω, we can then use *boundary layers*, which are unions of elements which include all points within a minimal distance δ_i to the boundary of an individual Ω_i. As an alternative, we also use sets created by adding one or more element layers to the closure of the individual subdomain faces.

We note that no new ideas are required to complete the part of the analysis related to these locally supported subspaces; cf. [15, Subsection 3.2] and the discus-

sion in [6, Section 3]. Therefore, we have been able to focus on developing the coarse spaces and bounds for the coarse component, which are always required in the analysis of any Schwarz algorithm; see [15, Subsection 2.3].

The subdomains, Ω_i, are unions of elements, that form quasi–uniform meshes for each subdomain, and often have irregular boundaries, in particular, if they have been generated by a mesh partitioner. Some of the tools used in our analysis, such as a trace theorem, will require that the subdomains are Lipschitz. We note that in our previous studies of two-dimensional problems, [3, 6], we have been able to extend our analysis even to subdomains with fractal boundaries assuming only that they are uniform in the sense of Jones [12].

We note from formula (1) that p_{nc} is the same for all n in a nodal equivalence class and for a particular $c \in \mathscr{C}_n$; in the case of tetrahedral subdomains, the basis functions constructed will be built from the face and edge functions, $\theta_{\mathscr{F}}$ and $\theta_{\mathscr{E}}$, used extensively in the development of iterative substructuring algorithms as in [15, Chapters 4-6]. The fact that these functions are piecewise constant causes large changes in the coarse basis functions across boundaries between equivalence classes, resulting in logarithmic factors, $(1+\log(H_i/h_i))$, in our estimate of the energy of the coarse basis functions; cf. [15, Lemma 4.25] for a bound on the energy of the classical face function $\theta_{\mathscr{F}}$. Here, H_i is the diameter of Ω_i and h_i the diameter of its smallest element. In [8], we have obtained the same quality bound for Lipschitz subdomains by generalizing bounds for the face and edge functions for subdomains to the Lipschitz case. By using the alternative (2), we obtain smoother coarse basis functions and improved bounds.

3 Assumptions and Major Results

We will now consider two different assumptions on the coefficient μ of the elasticity problem. The same assumptions are also used for the coefficient, ρ, of the scalar elliptic problems.

Assumption 1 *(Quasi-monotone face-connected paths) Let c be any coarse node of Ω_i and $\mathscr{S}_c$ be the index set of all subdomains containing c on their boundaries. Select $j_c \in \mathscr{S}_c$ such that $\mu_{j_c} \geq \mu_j$ for all $j \in \mathscr{S}_c$. Assume that there exists a constant C and for any $i \in \mathscr{S}_c$ a sequence $\{i = j_c^0, j_c^1, ..., j_c^p = j_c\}$, all in $\mathscr{S}_c$, such that $\mu_i \leq C\mu_{j_c^\ell}$ and that $\Omega_{j_c^{\ell-1}}$ and $\Omega_{j_c^\ell}$ have a subdomain face $\mathscr{F}_{j_c^{\ell-1},j_c^\ell}$ in common for all $\ell = 1,\ldots,p$ and $i = 1,\ldots,N$. In the case that $c \in \partial\Omega$, we also assume that $\partial\Omega_{j_c} \cap \partial\Omega$ contains at least one subdomain face.*

In other words, Assumption 1 means that there is a face connected path between Ω_i and Ω_{j_c} such that the Lamé parameter μ_i is no greater than a constant times the Lamé parameter of any subdomain along the path. This assumption is similar to the quasi-monotonicity assumption of [9]. We will also work with an additional assumption.

Assumption 2 *(Quasi-monotone edge-connected paths) Using the same notation as in Assumption 1, assume that there exists a sequence* $\{i = j_c^0, j_c^1, ..., j_c^p = j_c\}$, *all in* $\mathscr{S}_c$, *such that* $\rho_i \leq C\rho_{j_c^\ell}$ *and* $\Omega_{j_c^{\ell-1}}$ *and* $\Omega_{j_c^\ell}$ *have at least a subdomain edge in common for all* $\ell = 1,\ldots,p$ *and* $i = 1,\ldots,N$. *In the case that* $c \in \partial\Omega_i$ *also assume that* $\partial\Omega_{j_c} \cap \partial\Omega$ *contains at least one subdomain edge.*

We note that Assumption 2 is weaker than Assumption 1 since we have more options of continuing at every step in the construction of a path. We note that in our proof for linear elasticity, we have had to use the more restricted Assumption 1. The need for this has also been demonstrated by experiments reported in [8].

Our analysis can closely follow the theory as developed in [15, Section 2.3]; a main effort is directed to constructing a coarse component u_0, for any u, with a good bound on the energy $E(u_0)$ in terms of $E(u)$, the energy of the function u..

With estimates for our coarse interpolants in hand, we can then perform a local analysis for an overlapping additive Schwarz algorithm using basically the same approach as in [3] or [6]. This involves a set of partition of unity functions $\{\vartheta_j\}_{j=1}^N$ with $0 \leq \vartheta_j \leq 1$, $|\nabla\vartheta_j| \leq C/\delta_i$, and with ϑ_j supported in the closure of a subdomain which is part of the covering of Ω. Here, δ_j is the thickness of the part of subdomain which is common to its neighbors. Given an estimate of the form

$$E(u_0) \leq C\Theta(H/h)E(u),$$

where $H/h := \max_i H_i/h_i$, the resulting condition number estimate for the preconditioned operator is given by

$$\kappa(M^{-1}A) \leq C\Theta(H/h)(1+H/\delta), \tag{3}$$

where $H/\delta := \max_i H_i/\delta_i$. For Option 2, we can prove a uniform bound of $\Theta(H/h)$ if Assumption 1 is satisfied. In addition, we have a bound $\Theta(H/h) \leq (1+\log(H/h))$ for the scalar case if Assumption 2 holds.

We note that our coarse spaces could alternatively be combined with local spaces previously developed for iterative substructuring algorithms such as those of [10]; see also [15, Chapter 5].

4 Numerical Results

Numerical results are presented in this section to help confirm the theory and to demonstrate some advantages of the face-based local spaces. We note that large-scale experiments with closely related algorithms are also reported in [11]. Our results are for a unit cube domain with homogeneous essential boundary conditions applied to one of its faces. Condition numbers (cond) of the preconditioned operator and the number of iterations (iter) needed to achieve a relative residual tolerance of 10^{-8} for the solution of the linear system of equations, $Ax = b$, with random right-hand-side vectors b are reported. The domain is decomposed into smaller cubic

subdomains, and formula (2) is used to construct the coarse space. We note that the interface preconditioner is of a hybrid type which employs overlapping Schwarz local spaces as in [5]. We also note that at the end of each step of the iteration, the residual will vanish at all interior nodes of the subdomains. We use the lowest order hexahedral nodal elements and Matlab.

Three different local spaces are considered. The *standard* one starts with all the nodes of a non-overlapping subdomain and adds to them nodes from an integer number of layers of elements adjacent to the original nodes. The *boundary layer* local spaces are identical to the standard ones with the exception that the starting nodes only include those on the subdomain interfaces. We note these local spaces were considered previously in [5]. Finally, the *face* local spaces of this study start with nodes in the closure of each subdomain face and add layers of elements just as for the other two local spaces. These spaces locally precondition an interface problem.

Our example has the overlap parameter $H/\delta = 3$ fixed while the number of elements (H/h) in each subdomain direction increases. In addition to condition numbers and iteration counts, we also report in Table 1 the number of non-zeros in the sparse Cholesky factorizations for the local spaces. Specifically, r_{nnz} denotes these numbers normalized by the number for the standard local space. Further, we report estimates of the maximum eigenvalue $\lambda_{\max}$ of the preconditioned operator.

Consistent with the theory, condition numbers for the face local spaces exhibit sub-linear growth with respect to H/h. Although the number of iterations and condition numbers are noticeably larger compared with the standard and boundary layer spaces, the number of non-zeros in the local factorizations are considerably smaller for the face spaces. One reason for the larger condition numbers of the face spaces are the larger values of $\lambda_{\max}$ shown in Table 1. The larger values can be explained using a coloring argument. For instance, there are 12 different faces which include each subdomain vertex in their closures. In contrast, each subdomain vertex is included in only 8 of the standard or boundary layer spaces.

Table 1 Results for a unit cube decomposed into 64 smaller cubic subdomains with overlap $H/\delta = 3$ for three different local spaces. The material properties are constant with $\rho = 1$ for scalar problems and $\mu = .385$, $\lambda = 1.54$ for elasticity problems.

	standard			boundary layer				face			
H/h	iter	cond	$\lambda_{\max}$	iter	cond	r_{nnz}	$\lambda_{\max}$	iter	cond	r_{nnz}	$\lambda_{\max}$
scalar problem results											
3	26	14.1	8.2	27	14.8	0.47	8.2	39	32.7	0.18	12.1
6	28	17.7	8.2	31	18.9	0.77	8.2	50	40.9	0.25	12.0
9	30	19.7	8.2	33	21.1	0.80	8.2	55	45.9	0.28	12.0
12	30	30.0	8.2	33	22.5	0.87	8.2	58	49.6	0.30	12.0
elasticity problem results											
3	33	13.6	8.2	34	14.4	0.47	8.2	47	30.9	0.17	12.1
6	36	15.8	8.2	38	16.8	0.69	8.2	59	35.3	0.23	12.0
9	37	17.1	8.2	40	18.2	0.78	8.2	62	38.7	0.26	11.9
12	38	18.0	8.2	41	19.1	0.78	8.2	64	41.2	0.27	11.9

Normalized solution times for the preconditioned conjugate gradient algorithm applied to an elasticity problem (see bottom half of Table 1) are shown in Figure 1 for the boundary layer and face local spaces. Notice for all values of H/h that the normalized times are less than 1 for the boundary layer local spaces. Remarkably, the smallest times are achieved using the face local spaces for $H/h > 5$ even though the number of iterations are larger than those for the other two local spaces. The improved performance here can be attributed to the much smaller factorization sizes for the face spaces.

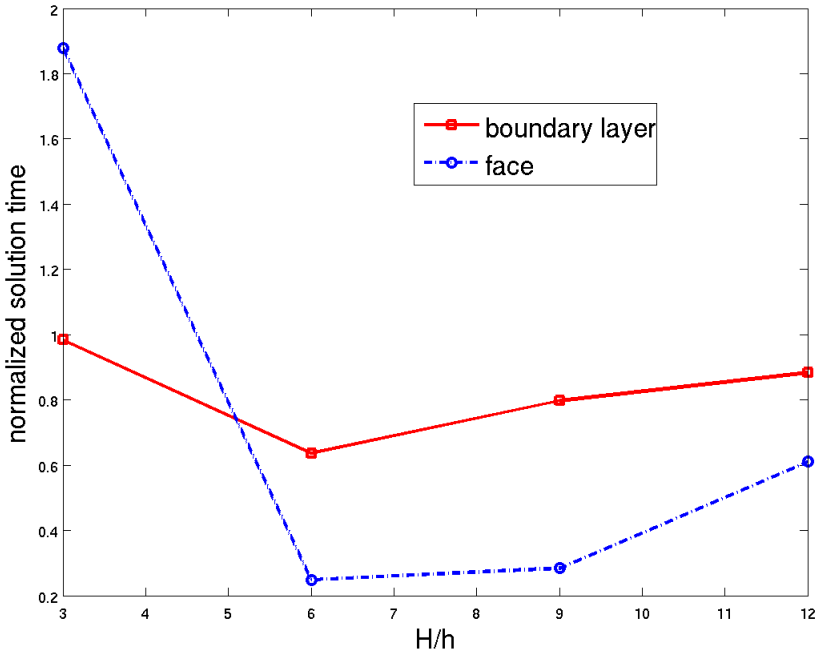


Fig. 1 Elasticity problem solution times for the preconditioned conjugate gradient algorithm normalized with respect to solution times for the standard local space.

As found in [5], the number of iterations can be reduced significantly, for all three local spaces, by dividing each element in the right-hand-side vectors for the local solvers by the number of local spaces which share this element. Although this results in a non-symmetric preconditioner, reduced solution times can be achieved as for restricted additive Schwarz preconditioners [1]. As a final note, for parallel computations, it makes sense to assign the work for each face to just one of the two subdomains, i.e. processors, which contain it. To achieve good load balance, an assignment algorithm can be used to approximately minimize the maximum work for any one processor.

Acknowledgements The work of the first author was supported in part by the National Science Foundation Grant DMS-1522736.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

1. Cai, X.-C. and Sarkis, M.: A restricted additive Schwarz preconditioner for general sparse linear systems, SIAM J. Sci. Comput.**21** (2), 792–797 (1999).
2. Dohrmann, C.R., Klawonn, A., and Widlund, O.B.: A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners. In: U. Langer, M. Discacciati, D. Keyes, O. Widlund, and W. Zulehner (eds.) Proceedings of the 17th International Conference on Domain Decomposition Methods in Science and Engineering. no. 60 in Springer-Verlag, Lecture Notes in Computational Science and Engineering, 247–254 (2007).
3. Dohrmann, C.R., Klawonn, A., and Widlund, O.B.: Domain decomposition for less regular subdomains: Overlapping Schwarz in two dimensions. SIAM J. Numer. Anal. **46** (4), 2153–2168 (2008).
4. Dohrmann, C.R. and Widlund, O.B.: An overlapping Schwarz algorithm for almost incompressible elasticity. SIAM J. Numer. Anal. **47** (4), 2897–2923 (2009).
5. Dohrmann, C.R. and Widlund, O.B.: Hybrid domain decomposition algorithms for compressible and almost incompressible elasticity. Internat. J. Numer. Meth. Engrg. **82**, 157–183 (2010).
6. Dohrmann, C.R. and Widlund, O.B.: An alternative coarse space for irregular subdomains and an overlapping Schwarz algorithm for scalar elliptic problems in the plane. SIAM J. Numer. Anal. **50** (5), 2522–2537 (2012).
7. Dohrmann, C.R. and Widlund, O.B.: Lower Dimensional Coarse Spaces for Domain Decomposition. In J. Erhel, M. Gander, L. Halpern, G. Pichot, T. Sassi, and O.B. Widlund (eds.) Proceedings of the 21th International Conference on Domain Decomposition Methods in Science and Engineering, no. 98 in Springer-Verlag, Lecture Notes in Computational Science and Engineering, 527–535 (2014).
8. Dohrmann, C.R. and Widlund, O.B.: On the design of small coarse spaces for domain decomposition algorithms. SIAM J. Sci. Comput. **39 (4)**, A1466–A1488, (2017).
9. Dryja, M., Sarkis, M.V., and Widlund, O.B.: Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions. Numer. Math. **72**, 313–348 (1996).
10. Dryja, M., Smith, B.F., and Widlund, O.B.: Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions. SIAM J. Numer. Anal. **31** (6), 1662–1694 (1994).
11. Heinlein, A., Klawonn, A., Rheinbach, O., and Widlund, O.: Improving the Parallel Performance of Overlapping Schwarz Methods by Using a Smaller Energy Minimizing Coarse Space. These proceedings.
12. Jones, P.W.: Quasiconformal mappings and extendability of functions in Sobolev space. Acta Math. **147** (1–2), 71–88 (1981).
13. Klawonn, A., Rheinbach, O., and Widlund, O.B.: An analysis of a FETI–DP algorithm on irregular subdomains in the plane. SIAM J. Numer. Anal. **46** (5), 2484–2504 (2008).
14. Klawonn, A. and Widlund, O.B.: Dual-Primal FETI methods for linear elasticity. Comm. Pure Appl. Math. **59** (11), 1523–1572 (2006).
15. Toselli, A. and Widlund, O.B.: Domain Decomposition Methods - Algorithms and Theory, *Springer Series in Computational Mathematics*, **34**. Springer-Verlag, Berlin, Heidelberg, New York (2005).