Multigrid and Domain Decomposition Methods in Fault-Prone Environments

by

Christian Alexander Glusa

Ingénieur diplomé de l'École Polytechnique, 2011

Diplom-Mathematiker, Karlsruhe Institute of Technology, 2012

Sc. M., Brown University, 2013

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Division of Applied Mathematics at Brown University

Providence, Rhode Island

May 2017

This dissertation by Christian Alexander Glusa is accepted in its present form by the Division of Applied Mathematics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Mark Ainsworth, Director

Recommended to the Graduate Council

Date _____ _____

George Em Karniadakis, Reader

Date _____ _____

Johnny Guzmán, Reader

Approved by the Graduate Council

Date _____ _____

Andrew G. Campbell
Dean of the Graduate School

In memoriam Vytas.

Vita

Education

2012 – 2017 PhD, Brown University, Providence, RI, USA

Advisor:	Mark Ainsworth

2013 Masters of Science in Applied Mathematics, Brown University, Providence, RI, USA

2009 – 2011 Ingnieur polytechnicien, Ecole Polytechnique, Palaiseau, France

Double-degree program

Major:	Applied Mathematics
Minors:	Pure Mathematics, Physics, Mechanics, Computer Science

2006 – 2009 & 2011 – 2012, Diplom-Mathematiker, Karlsruhe Institute of Technology, Germany

Major:	Mathematics
Minors:	Theoretical Physics, Economics
Diploma thesis:	Sedimentation in Periodic Domains
Advisor:	Willy Dörfler

Talks & Presentations

- 3/2017, SIAM Conference on Computational Science and Engineering, Atlanta, *"The Resiliency of Multilevel Methods on Next Generation Computing Platforms: Probabilistic Model and Its Analysis"*
- 1/2017, CSRI, Sandia National Lab, Albuquerque, *"Resilience Analysis for Multigrid Methods"*
- 1/2017, Applied Math and Scientific Computing Seminar, Temple University, *"Resilience Analysis for Multigrid Methods"*
- 1/2017, Frontiers in Applied and Computational Mathematics, ICERM, Providence, *"Resilience Analysis for Multigrid Methods"* (Poster)
- 7/2016, SIAM Annual Meeting, Boston, *"Resilience Analysis for Multigrid Methods"* (Poster)
- 4/2016, CRUNCH seminar, Brown University, *"Resilience Analysis for Multigrid Methods"*
- 4/2016, Applied Math Days, Rensselaer Polytechnic Institute, *"Multigrid methods, random matrices and fault resilience"*
- 3/2016, 14^{th} Copper Mountain Conference on Iterative Methods, *"Resilience Analysis for Multigrid Methods"*
- 11/2015, Math Slam, Brown University, *"Multigrid methods, random matrices and fault resilience"*
- 8/2015, 2015 AARMS Workshop on Domain Decomposition Methods for PDEs, Dalhousie University, Halifax, Canada, *"Multigrid methods, random matrices and fault resilience"*

- 7/2015, Math Slam, Brown University, *"Multigrid methods, random matrices and fault resilience"*
- 3/2015, Graduate student seminar, Brown University, *"Multigrid methods, random matrices and fault resilience"*
- 10/2011, Workshop for Particulate Flows, Karlsruhe Institute of Technology, Germany, *"Sedimentation in Periodic Domains"*

Other Conferences & Workshops Attended

- 8/2016, Argonne Training Program on Extreme-Scale Computing, St. Charles, Illinois
- 5/2016, HPC Day 2016, CSCVR, University of Massachusetts, Dartmouth
- 10/2015, Finite Element Circus, University of Massachusetts, Dartmouth
- 9/2015, Topical Workshop on Numerical Methods for Large-Scale Nonlinear Problems and Their Applications, ICERM, Providence
- 5/2014, Topical Workshop on Robust Discretization and Fast Solvers for Computable Multi-Physics Models, ICERM, Providence
- 8/2013, Brown-Kobe Summer School on High-Performance Computing, Kobe University, Japan

Teaching

- 2015–, Math Teacher at English for Action, Providence, RI, Weekly classes for immigrants on High School level

- Summer 2014, Scientific Computing Summer School, Brown University, Organized a three week summer school for undergrad students from Universit Pierre-et-Marie-Curie (Paris VI)
- Spring 2014, Teaching Assistant, Methods of Applied Mathematics II, Brown University
- Fall 2013, Teaching Assistant, Methods of Applied Mathematics I, Brown University

Service & Outreach

- Fall 2016, Judge for Brown Mathematical Contest for Modelling
- 2016,AWM Grad-Undergrad Mentor, Brown University
- Fall 2015, Judge for Brown Mathematical Contest for Modelling
- 2015–2016, Treasurer for the Brown University SIAM Stundent Chapter
- 2010–2011, Treasurer for Aslive, Versailles, France, Charitable society for adults with developmental disabilities
- Summer 2009, Summer camp organizer, Foyer Les Sources, Pairis, France, Summer vacation camp for adults with developmental disabilities
- Summer 2008, Summer camp organizer, Foyer Les Sources, Pairis, France, Summer vacation camp for adults with developmental disabilities
- 2005 – 2006, Voluntary social year, Foyer Les Sources, Pairis, France, Care for adults with developmental disabilities

Honors

- 03/2014, Nominated to Sigma Xi Honor Society
- 08/2009–08/2011, Recipient of the ”Jacques Garaialde Bernard Oppetit” scholarship, Excellence Scholarship
- 09/2010, Recipient of the ”Prix du Projet Scientifique Collectif”, Scientific group project of 9 months,
  ”Compréhension de la Nage à faible nombre de Reynolds”.

Preface

Today, more than ever before, scientific exploration relies on large-scale computer simulation of complex phenomena. The installation of larger and faster supercomputers enables models of physical systems to include more features and make more accurate predictions. The advent of the new generation of exascale machines, as mandated by President Obama's executive order to establish the National Strategic Computing Initiative[1], poses several unique challenges. Historically, algorithms and their implementations were optimized with respect to the number of floating-point operations (flops) necessary. However, it has become apparent that the time to solution is more and more dependent on the amount of communication between the individual compute units of the system. A new constraint in the design of supercomputers is the power consumption of the installation. The energy necessary to send a message depends on distance of the communication, and simply up-scaling the design of current petascale systems, i.e. systems that can perform one petaflop, to the next generation of exascale machines is outright infeasible. The power consumption of such a system would easily surpass that of half a million homes. In order to observe the commonly accepted limit of 20MW, lower voltage thresholds and reduced cell capacitance are required. The inevitable consequence of these hardware level solutions are increased rates of faults that affect the correctness of the performed computations. This effect is amplified by the significant expansion

[1] Creating a National Strategic Computing Initiative, Executive Order No. 13702 of July 29th, 2015, Federal Register

of the number of components making up an exascale system. Therefore, fault detection and mitigation strategies will have to be considered in order to take full advantage of the potential of the next generation of supercomputers [27, 28, 29, 42].

Many scientific simulations rely on the solution of very large, but sparse, linear systems of equations. Multigrid and Domain Decomposition methods are widely used solvers due to their attractive properties. Their benefits include optimal computational complexity, scaling and easy of implementation on parallel machines. It is therefore of prime concern to ensure the transition of these methods and their preexisting implementations to the next generation of high-performance systems and study their application to the problems that become tractable on these new machines.

The contribution of this work is two-fold. In the first part, we study the impact of faults on the convergence of iterative solvers of linear systems. The first chapter serves as an introduction to linear iterative solvers, and briefly describes one of the theoretical frameworks used to prove convergence of multigrid methods in the ideal, fault-free case. In Chapter 2, we describe the different categories of faults and how they are modeled using random matrices. We derive the error propagator of a fault-prone multigrid algorithm, which will be instrumental to prove convergence estimates in the presence of faults. Theoretical results concerning products of random matrices and methods of determining their asymptotic rate of convergence are given in Chapter 3. As it turns out, the so-called Lyapunov spectral radius is difficult to compute, but can be estimated in some cases. In Chapters 4 and 5, we apply the theoretical framework first to a simple iteration to illustrate the approach, and then to a two grid method. The extension to the multigrid case is given in Chapter 6. It will transpire that multigrid is not a fault-resilient solver, since convergence can only be achieved for limited problem size. By protecting one of its steps, the prolongation operation from coarser to finer meshes, fault resilience

is obtained. All theoretical results are illustrated by means of numerical examples that stem from a variety of different applications. More practical implementation details concerning the detection and mitigation of faults are discussed in Chapter 7. We illustrate how the theoretical results can be leveraged to obtain a fault-tolerant multigrid implementation. Finally, we discuss fault mitigation for overlapping domain decomposition solvers in Chapter 8. We extend the framework to account for faults that appear in unknowns that are shared between compute nodes, and give convergence estimates for one- and two-level methods.

In the second part of this work, we turn our attention to the efficient solution of non-local and fractional models using iterative linear solvers. Non-local equations, as compared to classical local models, can be used to more accurately describe the underlying physical phenomena. The price to pay is an increased complexity in terms of discretization and solution of the arising system of equations. We introduce archetypes of elliptic and parabolic non-local equations in Chapter 10, and state results concerning the regularity of their solution. We give details concerning finite element approximations and explain how optimal rates of convergence can be achieved through adaptive mesh refinement based on the use of a posteriori error estimators. In Chapter 11, we turn to the question of efficient implementations of assembly, solution and computation of error indicators. The finite element assembly involves the computation of singular element contributions for which special quadrature rules are required. Since the considered problems are non-local, the straightforward approach results in dense system matrices. Similarly, straightforward computation of a posteriori error estimators scales quadratically in the number of unknowns. We circumvent these issues by exploiting the low rank of off-diagonal blocks and explain how all steps of the assemble-solve-estimate-refine cycle can be achieved in quasi-linear complexity. Finally, in Chapter 12, we illustrate optimality of computational complexity and rate of convergence for a number of one- and two-dimensional numerical examples, involving fractional Poisson and

heat equations and a fractional reaction-diffusion system.

Acknowledgements

First and foremost, I want to thank my advisor Mark Ainsworth. This work would not have been possible without his guidance and support. I am very fortunate to have benefited from his insight and eye for interesting problems. I will miss our weekly meetings and his sense of humor.

I would also like to thank Professor George Em Karniadakis and Professor Johnny Guzmán for serving on my thesis committee and generously providing feedback about this work.

Finally, I want to acknowledge all the support I received from my colleagues, friends and family. You kept me sane throughout grad school, and I am grateful.

Contents

⋆ Parts of this thesis have been previously published or have been submitted for publication:

- Mark Ainsworth and Christian Glusa. “Numerical Mathematics and Advanced

Applications - ENUMATH 2015: Proceedings of ENUMATH 2015". In: Cham: Springer International Publishing, 2015. Chap. Multigrid at Scale?, pp. 237–253. ISBN: 978-3-319-39929-4. DOI: 10.1007/978-3-319-39929-4_24

- Mark Ainsworth and Christian Glusa. "Is the Multigrid Method Fault Tolerant? The Two-Grid Case". In: *SIAM Journal on Scientific Computing* 39.2 (2017), pp. C116–C143. DOI: 10.1137/16M1100691

- Mark Ainsworth and Christian Glusa. "Is the multigrid method fault tolerant? The Multilevel Case." In: *SIAM Journal on Scientific Computing* (Submitted)

- Mark Ainsworth and Christian Glusa. "Towards an Efficient Finite Element Method for the Integral Fractional Laplacian on Polygonal Domains". In: *TBD* (2017)

- Mark Ainsworth and Christian Glusa. "Aspects of an adaptive finite element method for the fractional Laplacian: a priori and a posteriori error estimates, efficient implementation and multigrid solver". Submitted

List of Tables

List of Figures

Part I

Fault-Prone Iterative Methods

Chapter 1

Multigrid Algorithm

Solvers for linear systems of equations can be classified into two categories. *Direct methods* aim to obtain the exact solution of the linear system (up to rounding error). This is typically achieved computing a factorization of the system matrix. Solution of the linear systems involving the factors is significantly easier, since their special structure can be exploited. The most well-known example of a direct method is Gauss elimination in which the system matrix is factored into upper and lower triangular factors. The drawback of direct methods is that even if the system matrix is sparse, i.e. if only $\mathcal{O}(n)$ of its entries are non-zeros, where n is the number of unknowns, the arising factors will be dense. Moreover, the computational complexity of direct methods is generally of order $\mathcal{O}(n^3)$. These drawbacks in memory usage and complexity motivated the development of *iterative methods*. These methods aim to obtain approximations that converge to the exact solution. The iterates are obtained only by matrix-vector products with the system matrix (and sometimes its transpose), which has complexity proportional to the number of non-zero entries of the matrix. In practice, obtaining only an approximate solution is not an issue. It suffices that the error in the solution of the system is dominated by other inherent errors, such as the one stemming from discretization of a continuum model. In the present chapter, we give an introduction to linear iterative methods, and to the

multigrid algorithm in particular.

1.1 Classical Iterative Solvers

Let $\boldsymbol{A}$ be a symmetric, positive definite sparse matrix arising from the discretization of an elliptic partial differential equation in d spatial dimensions, and $b \in \mathbb{R}^n$. The equation

$$\boldsymbol{A}x = b$$

can be rewritten as the fixed point problem

$$x = x + \boldsymbol{N}\left(b - \boldsymbol{A}x\right),$$

where $\boldsymbol{N}$ is a suitably chosen invertible matrix. If we require that the application of the so-called *preconditioner* $\boldsymbol{N}$ is inexpensive, evaluation of the right-hand side will not be costly because of the sparsity of $\boldsymbol{A}$. This leads to the following linear iterative method

$$x_{j+1} = x_j + \boldsymbol{N}\left(b - \boldsymbol{A}x_j\right).$$

If the method is to be used as a solver, then we need that $x_j \to x$ as $j \to \infty$. The error of the method after the $j+1$-th step can be written as

$$\begin{aligned} e_{j+1} &= x_{j+1} - x \\ &= x_j - x + \boldsymbol{N}\left(b - \boldsymbol{A}x_j\right) \\ &= \left(\boldsymbol{I} - \boldsymbol{N}\boldsymbol{A}\right) e_j. \end{aligned}$$

We call $\boldsymbol{E} = \boldsymbol{I} - \boldsymbol{N}\boldsymbol{A}$ the *iteration matrix* or *error propagator*. The error converges to zero if and only if $\rho\left(\boldsymbol{E}\right) < 1$. The smaller the spectral radius of $\boldsymbol{E}$, the faster the convergence to the exact solution. The ideal preconditioner would be $\boldsymbol{A}^{-1}$, which would lead to convergence after just one iteration. Since finding $\boldsymbol{A}^{-1}$ is more

expensive than solving the original linear system $\boldsymbol{A}x = b$, $\boldsymbol{A}^{-1}$ is not a reasonable choice for the preconditioner. It demonstrates, however, that a good preconditioner should mimic the spectrum of $\boldsymbol{A}^{-1}$. On the other hand, taking $\boldsymbol{N} = \rho(\boldsymbol{A})\,\boldsymbol{I}$ leads to a very inexpensive preconditioner, which, in turn, has often poor convergence properties since almost no information of the original problem is taken into account. The construction of efficient preconditioners that are both inexpensive and a good approximation to $\boldsymbol{A}^{-1}$ is an extensive field of very active research.

A classical iterative scheme is the (damped) Jacobi method, where $\boldsymbol{N}$ is proportional to the inverse of the diagonal part $\boldsymbol{D}$ of $\boldsymbol{A}$. The convergence rate is given by the following theorem:

Theorem 1.
Let $\boldsymbol{E} = \boldsymbol{I} - \theta \boldsymbol{D}^{-1}\boldsymbol{A}$ be the iteration matrix of the damped Jacobi method.

1. *If $0 < \theta\boldsymbol{A} < 2\boldsymbol{D}$, then the damped Jacobi method converges.*
2. *If $0 < \lambda\boldsymbol{D} \leq \theta\boldsymbol{A} \leq \Lambda\boldsymbol{D}$, then the spectrum of $\boldsymbol{E}$ is contained in $[1-\Lambda, 1-\lambda]$ and the convergence rate is $\rho(\boldsymbol{E}) = \max\{\Lambda - 1, 1 - \lambda\}$.*

Example 1.
Let $\boldsymbol{A}_{1D} \in \mathbb{R}^{n\times n}$ be the finite difference approximation to the 1D-Laplacian with Dirichlet boundary conditions:

$$\boldsymbol{A}_{1D} = \operatorname{tridiag}_n(-1, 2, -1).$$

The eigenvalues of $\boldsymbol{A}_{1D}$ are

$$\lambda_k = 2 - 2\cos\frac{k\pi}{n+1} \qquad\qquad k = 1, \ldots, n$$

and the corresponding eigenvectors are the discrete sine waves

$$\vec{v}_k = \left\{\sin\frac{kj\pi}{n+1}\right\}_{j=1,\ldots,n} \qquad\qquad k = 1, \ldots, n$$

Because $\boldsymbol{D} = 2\boldsymbol{I}$*, we have* $0 < \lambda \boldsymbol{D} \leq \boldsymbol{A}_{1D} \leq \Lambda \boldsymbol{D}$ *with*

$$\lambda = 1 - \cos\left(\frac{\pi}{n+1}\right),$$
$$\Lambda = 1 + \cos\left(\frac{\pi}{n+1}\right).$$

Hence

$$\rho\left(\boldsymbol{E}\right) = \cos\left(\frac{\pi}{n+1}\right).$$

For problem size n*, we have*

$$\rho\left(\boldsymbol{E}\right) = 1 - \frac{\pi^2}{2(n+1)^2} + \mathcal{O}\left((n+1)^{-4}\right)$$

which makes this method quite slow. We observe though that the error components along the different eigenvectors $\vec{v}_k$ *get damped by*

$$\rho_k = 1 - \frac{\lambda_k}{2} = \cos\left(\frac{k\pi}{n+1}\right).$$

The components along $\vec{v}_k$*, where* k *is close to* $\frac{n+1}{2}$ *decrease indeed much faster.*

Example 1 shows that the Jacobi iteration reduces high-frequency error (with large λ_k) rapidly, but stalls on low frequency error. This suggests coupling the Jacobi iteration with a second type of iteration that specifically targets the low-frequency error.

1.2 Multigrid Algorithm

The fundamental observation that leads to the multigrid method is that low-frequency error varies slowly, whereas high frequency error is oscillatory. This means that low-frequency error can be transferred to a coarser discretization without loosing too much information.

In what follows, let $\Omega \subset \mathbb{R}^d$ be a convex polyhedral domain, $\Gamma_D \subseteq \partial\Omega$ be non-empty and $V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$. Let $\mathcal{T}_0$ be a partitioning of Ω into

the union of elements each of which is a $d+1$-simplex, such that he intersection of any pair of distinct elements is a single common vertex, edge, face or sub-simplex of both elements. For $\ell \in \mathbb{N}$, the partition $\mathcal{T}_\ell$ is obtained by uniformly subdividing each element in $\mathcal{T}_{\ell-1}$ into sub-simplices such that $\mathcal{T}_\ell$ satisfies the same conditions as $\mathcal{T}_0$ [73]. Let $\mathbb{P}_k$, $k \in \mathbb{N}$, consist of polynomials on $\mathbb{R}^d$ of total degree at most k. The subspace V_ℓ consists of continuous piecewise polynomials belonging to $\mathbb{P}_k$ on each element on the partition $\mathcal{T}_\ell$. In particular, this means that the spaces are nested: $V_0 \subset V_1 \subset \ldots \subset V_L \subset V$.

Let $a : V \times V \to \mathbb{R}$ be a continuous, V-elliptic, bilinear form and $F : V \to \mathbb{R}$ be a continuous linear form and consider the problem of obtaining $u_L \in V_L$ such that

$$a(u_L, v) = F(v) \quad \forall v \in V_L. \tag{1.1}$$

Let $\phi_\ell^{(i)}$, $i = 1, \ldots, n_\ell$, denote the usual nodal basis for V_ℓ, and ϕ_ℓ be the corresponding vector formed using the basis functions.

Problem eq. (1.1) is then equivalent to the matrix equation

$$\boldsymbol{A}_L x_L = F_L \tag{1.2}$$

where $\boldsymbol{A}_\ell = a(\phi_\ell, \phi_\ell)$ is the stiffness matrix, x_L is the coefficient vector of u_ℓ and $F_L = F(\phi_L)$ is the load vector. For future reference, we define the mass matrix by $\boldsymbol{M}_\ell = (\phi_\ell, \phi_\ell)$.

The multigrid method solves the linear system of equations

$$\boldsymbol{A} x = b$$

by introducing a hierarchy of coarser problems

$$\boldsymbol{A}_\ell x_\ell = b_\ell$$

of size n_ℓ, $\ell = 0, \ldots, L$ with $\boldsymbol{A}_L = \boldsymbol{A}$. Information gets transferred between levels through *restriction* and *prolongation* operators

$$\boldsymbol{R}_\ell^{\ell+1} : \mathbb{R}^{n_{\ell+1}} \to \mathbb{R}^{n_\ell}, \quad \boldsymbol{P}_{\ell+1}^{\ell} : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_{\ell+1}}.$$

We will assume that $\boldsymbol{P}_{\ell+1}^{\ell} = \left(\boldsymbol{R}_{\ell}^{\ell+1}\right)^T$ along with the *Galerkin relation*

$$\boldsymbol{A}_{\ell} = \boldsymbol{R}_{\ell}^{\ell+1} \boldsymbol{A}_{\ell+1} \boldsymbol{P}_{\ell+1}^{\ell}.$$

In the finite element context, the natural way to define these operators is to employ the nesting of the spaces $V_\ell \subset V_{\ell+1}$, which means that there exists a matrix $\boldsymbol{R}_{\ell+1}^{\ell}$ such that $\phi_\ell = \boldsymbol{R}_{\ell+1}^{\ell} \phi_{\ell+1}$. We will drop sub- and superscripts on restriction and prolongation operators in what follows.

Moreover, linear iterations called *smootheners* are defined on each level as

$$x_\ell \leftarrow x_\ell + \boldsymbol{N}_\ell \left(b_\ell - \boldsymbol{A}_\ell x_\ell\right),$$

where $\boldsymbol{N}_\ell$ is a suitable preconditioner, e.g. the damped Jacobi preconditioner $\boldsymbol{N}_\ell = \theta \boldsymbol{D}_\ell^{-1}$, with $\boldsymbol{D}_\ell$ the diagonal part of $\boldsymbol{A}_\ell$, as described above.

The multigrid method now consists in applying several steps of smoothening, i.e. application of an iterative methods that reduces high frequency error, followed by restriction of the *residual* $b_\ell - \boldsymbol{A}_\ell x_\ell$ to the next coarser grid. The coarser problem is solved by one or more recursive calls to the multigrid algorithm, or, in case of the coarsest mesh, by a direct solver. The multigrid method is given in Algorithm 1 and illustrated in Figure 1.1 and may be invoked to obtain an approximate solution of problem eq. (1.2) by a call to $MG_L(b, 0)$. The reason why the smoothening step appears twice is that it can be beneficial for the iteration to be symmetric, since the corresponding preconditioner will then be symmetric. This permits, for example, to use a multigrid iteration as a preconditioner in the Conjugate Gradient method. The number of recursive calls γ to the multigrid method on a coarse grid leads to different multigrid schemes. The most popular ones are called the *V-cycle* for which $\gamma = 1$, and the *W-cycle* with $\gamma = 2$.

Alternative choices for the smoothening iteration are possible. Popular options apart from the damped Jacobi iteration include Gauss-Seidel iterations, SOR smootheners, block-smootheners and polynomial smootheners.

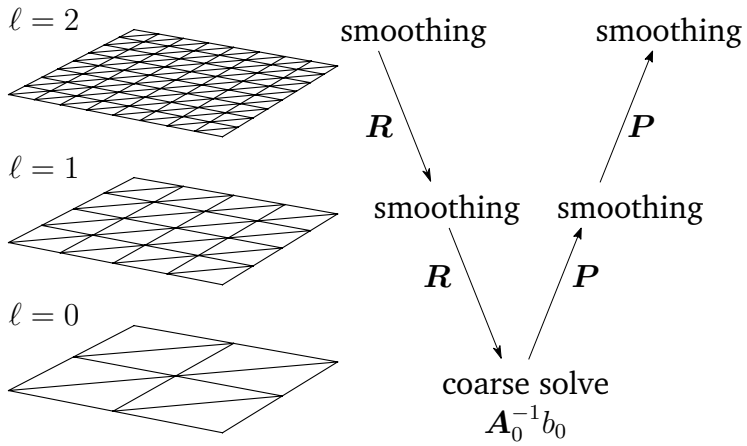


Figure 1.1: Multigrid hierarchy with three levels, and restrictions $\boldsymbol{R}$ and prolongations $\boldsymbol{P}$.

Algorithm 1 Multigrid method MG_ℓ

Input: Right hand side b_ℓ; Initial iterate x_ℓ
Output: $MG_\ell(b_\ell, x_\ell)$

if $\ell = 0$ **then return** $\boldsymbol{A}_0^{-1} b_0$ // Exact solve on coarsest grid
else
  for $i \leftarrow 1$ **to** ν_1 **do**
    $x_\ell \leftarrow S_\ell\,(b_\ell, x_\ell)$ // ν_1 pre-smoothening steps
  $d_{\ell-1} \leftarrow \boldsymbol{R}\,(b_\ell - \boldsymbol{A}_\ell x_\ell)$ // Restriction to coarser grid
  $e_{\ell-1} \leftarrow 0$
  for $j \leftarrow 1$ **to** γ **do**
    $e_{\ell-1} \leftarrow MG_{\ell-1}(d_{\ell-1}, e_{\ell-1})$ // γ coarse grid correction steps
  $x_\ell \leftarrow x_\ell + \boldsymbol{P} e_{\ell-1}$ // Prolongation to finer grid
  for $i \leftarrow 1$ **to** ν_2 **do**
    $x_\ell \leftarrow S_\ell\,(b_\ell, x_\ell)$ // ν_2 post-smoothening steps

What makes the multigrid method particularly attractive is its computational complexity. On every levels, only sparse operations with complexity proportional to n_ℓ are used. Provided that the coarsening ratio from level to level is large enough, a single iteration of the multigrid method can easily be shown to have complexity $\mathcal{O}(n_L)$ or $\mathcal{O}(n_L \log n_L)$ [62]. As we will discuss in the next section, the rate of convergence of the multigrid method does not depend on the problem size. This means that an essentially constant amount of iterations needs to be taken to converge to a given error tolerance, making the multigrid method a solver with linear complexity.

1.3 Convergence Theory for Multigrid

The analysis of the convergence behavior of multigrid algorithms is a rich topic. Two theoretical frameworks are available. The first one relies on splitting the problem into high and low frequencies for a two grid method. The results are then extended to the multigrid case by a perturbation argument and recursive inequalities [62, 63]. The second framework is inspired by the analysis of domain decomposition methods and relies on a nested decomposition of the solution space [21, 101, 102]. In what follows, we adhere to the first framework as it will be seen to extend to the analysis of fault-prone methods.

Let the iteration matrix of pre- and post-smoothener be given by $\boldsymbol{E}_\ell^S$. Then the iteration matrix of the multigrid method is given by

$$\begin{aligned} \boldsymbol{E}_0 &= \boldsymbol{0}, \\ \boldsymbol{E}_\ell &= \left(\boldsymbol{E}_\ell^S\right)^{\nu_2} \boldsymbol{E}_\ell^{CG} \left(\boldsymbol{E}_\ell^S\right)^{\nu_1}, \end{aligned} \tag{1.3}$$

where the coarse grid correction is given by

$$\boldsymbol{E}_\ell^{CG} = \boldsymbol{I} - \boldsymbol{P}\left(\boldsymbol{I} - \boldsymbol{E}_{\ell-1}^{\gamma}\right) \boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{R} \boldsymbol{A}_\ell.$$

In the case of a two grid method, the iteration matrix is given by

$$\boldsymbol{E}_\ell^{TG} = \left(\boldsymbol{E}_\ell^S\right)^{\nu_2} \left(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_{\ell-1}^{-1}\boldsymbol{R}\boldsymbol{A}_\ell\right) \left(\boldsymbol{E}_\ell^S\right)^{\nu_1}. \tag{1.4}$$

The standard assumptions for the convergence analysis of the multigrid method read as follows [20, 62, 63, 95]:

Assumption 1 (Smoothening property)*:* There exists $\eta : \mathbb{N} \to \mathbb{R}_{\geq 0}$ satisfying $\lim_{\nu\to\infty} \eta(\nu) = 0$ and such that

$$\left\| \boldsymbol{A}_\ell \left(\boldsymbol{E}_\ell^S \right)^\nu \right\|_2 \leq \eta(\nu) \left\| \boldsymbol{A}_\ell \right\|_2, \quad \nu \geq 0,\ \ell = 1, \ldots, L.$$

Assumption 2 (Approximation property)*:* There exists a constant C_A such that

$$\left\| \boldsymbol{E}_\ell^{CG} \boldsymbol{A}_\ell^{-1} \right\|_2 \leq \frac{C_A}{\left\| \boldsymbol{A}_\ell \right\|_2}, \quad \ell = 1, \ldots, L.$$

The smoothening and approximation property imply two grid convergence, with convergence rate given by

$$\begin{aligned} \rho\left(\boldsymbol{E}_\ell^{TG} \right) &\leq \left\| \boldsymbol{E}_\ell^{CG} \left(\boldsymbol{E}_\ell^S \right)^{\nu_1+\nu_2} \right\|_2 \\ &\leq \left\| \boldsymbol{E}_\ell^{CG} \boldsymbol{A}_\ell^{-1} \right\|_2 \left\| \boldsymbol{A}_\ell \left(\boldsymbol{E}_\ell^S \right)^{\nu_1+\nu_2} \right\|_2 \\ &\leq C_A \eta\left(\nu_1 + \nu_2 \right). \end{aligned} \tag{1.5}$$

The right-hand side is less than one provided $\nu_1 + \nu_2$ is large enough.

Assumption 3*:* The smoothener is non-expansive, i.e. $\rho\left(\boldsymbol{E}_\ell^S \right) = \left\| \boldsymbol{E}_\ell^S \right\|_A \leq 1$, and there exists $C_S > 0$ such that

$$\left\| \left(\boldsymbol{E}_\ell^S \right)^\nu \right\|_2 \leq C_S \quad \nu \geq 1,\ \ell = 1, \ldots, L.$$

Assumption 3 permits to show that the two grid method is also a contraction with respect to the spectral norm $\|\bullet\|_2$:

$$\left\| \boldsymbol{E}_\ell^{TG} \right\|_2 \leq \left\| \left(\boldsymbol{E}_\ell^S \right)^{\nu_2} \right\|_2 \left\| \boldsymbol{E}_\ell^{CG} \left(\boldsymbol{E}_\ell^S \right)^{\nu_1} \right\|_2 \leq C_S C_A \eta\left(\nu_1 \right).$$

While this result is weaker than eq. (1.5), it will be useful in the fault-prone case.

Assumption 4*:* There exist positive constants $\underline{C}_p$ and $\overline{C}_p$ such that

$$\underline{C}_p^{-1} \left\| x_\ell \right\|_2 \leq \left\| \boldsymbol{P} x_\ell \right\|_2 \leq \overline{C}_p \left\| x_\ell \right\|_2 \quad \forall x_\ell \in \mathbb{R}^{n_\ell},\ \ell = 0, \ldots, L-1.$$

Assumptions 3 and 4 allow to extend the convergence theory to the multilevel case with $\gamma \geq 2$. The most interesting case in practice is the W-cycle ($\gamma = 2$).

$$\begin{aligned}
\|\boldsymbol{E}_\ell\|_2 &= \left\| \left(\boldsymbol{E}_\ell^S\right)^{\nu_2} \left[\boldsymbol{I} - \boldsymbol{P}\left(\boldsymbol{I} - \boldsymbol{E}_{\ell-1}^\gamma\right) \boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{R}\boldsymbol{A}_\ell\right] \left(\boldsymbol{E}_\ell^S\right)^{\nu_1} \right\|_2 \\
&\leq \left\| \left(\boldsymbol{E}_\ell^S\right)^{\nu_2} \left[\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{R}\boldsymbol{A}_\ell\right] \left(\boldsymbol{E}_\ell^S\right)^{\nu_1} \right\|_2 \\
&\quad + \left\| \left(\boldsymbol{E}_\ell^S\right)^{\nu_2} \right\|_2 \|\boldsymbol{P}\|_2 \, \|\boldsymbol{E}_{\ell-1}\|_2^\gamma \left\| \boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{R}\boldsymbol{A}_\ell \left(\boldsymbol{E}_\ell^S\right)^{\nu_1} \right\|_2 \\
&\leq \left\| \boldsymbol{E}_\ell^{TG} \right\|_2 + \underline{C}_p \overline{C}_p C_S \, \|\boldsymbol{E}_{\ell-1}\|_2^\gamma \left\| \boldsymbol{P}\boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{R}\boldsymbol{A}_\ell \left(\boldsymbol{E}_\ell^S\right)^{\nu_1} \right\|_2 .
\end{aligned}$$

Since

$$\begin{aligned}
\left\| \boldsymbol{P}\boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{R}\boldsymbol{A}_\ell \left(\boldsymbol{E}_\ell^S\right)^{\nu_1} \right\|_2 &\leq \left\| \left(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{R}\boldsymbol{A}_\ell\right) \left(\boldsymbol{E}_\ell^S\right)^{\nu_1} \right\|_2 + \left\| \left(\boldsymbol{E}_\ell^S\right)^{\nu_1} \right\|_2 \\
&\leq C_S + C_A \eta\left(\nu_1\right),
\end{aligned}$$

we obtain the recursive inequality

$$\|\boldsymbol{E}_\ell\|_2 \leq \left\| \boldsymbol{E}_\ell^{TG} \right\|_2 + C_* \, \|\boldsymbol{E}_{\ell-1}\|_2^\gamma \tag{1.6}$$

with $C_* = C_S \underline{C}_p \overline{C}_p \left(C_S + C_A \eta\left(\nu_1\right)\right)$. A classical result [63] concerning this inequality is

Lemma 2.

Suppose the elements of the sequence $\{\eta_\ell\}_{\ell \geq 1}$ *satisfy* $0 \leq \eta_1 \leq \xi$ *and* $0 \leq \eta_\ell \leq \xi + C_* \eta_{\ell-1}^\gamma$, $\ell \geq 2$. *If* $\gamma \geq 2$, $C_* \gamma > 1$ *and* $\xi \leq \frac{\gamma-1}{\gamma} \frac{1}{(C_*\gamma)^{\frac{1}{\gamma-1}}}$, *then*

$$\eta_\ell \leq \left\{ \begin{array}{ll} \frac{\gamma}{\gamma-1}\xi, & \gamma \geq 2, \\ \frac{2\xi}{1+\sqrt{1-4C_*\xi}}, & \gamma = 2 \end{array} \right\} < 1.$$

The result show that two grid convergence along with a sufficient number of smoothening steps imply that the multilevel scheme is convergent.

Chapter 2

Fault-Prone Iterative Methods

The taxonomy of faults and failures by Avižienis et al. [11] distinguishes between *hard faults* and *soft faults*. Soft faults correspond to corruption of instructions as well as the data produced and used within the code but which allow the execution of the code to proceed albeit in a corrupted state. An example of such a fault would be flipping of individual bits in a floating point number. Some bit flips in the exponent, sign or in the significant components of the mantissa may result in a relatively large corruption that becomes detectable. Hard faults might correspond to a message from one or more compute nodes either being corrupted beyond all recognition (possibly resulting in an exception being thrown), or lost altogether in the event of a compute node failing completely. Hard faults result in the interruption of the execution of the actual code unless remedial action is taken. As such, hard faults constitute errors that are readily *detectable*. Although the likelihood ε of a hard fault occurring can be assumed small, it is not negligible. The fact that an algorithm cannot continue after a hard fault without some kind of remedial action means that such faults cannot simply be ignored in the hope that the algorithm will recover naturally.

Most, if not all, existing algorithms were derived under the assumption that such faults cannot occur and accordingly, no indication is generally offered as to

what might constitute an appropriate course of action should a fault be detected in a particular algorithm. Nevertheless, in the event of a hard fault occurring, a decision has to be made as to what remedial action should be taken in order to resume the execution of the algorithm. The action that is chosen can have a dramatic effect on the performance and characteristics of the algorithm and as such represents a vital part of the algorithm. Ideally, the resulting algorithm should be subjected to the same kind of mathematical analysis that was applied to the original, deterministic, variant of the algorithm. Historical considerations mean that such analysis is available for very few methods.

At the present time, the basic approach to dealing with faults consists of attempting to restore the algorithm to the state that would have existed had no fault occurred. Two common ways in which this might be accomplished include: *Checkpoint-restart* [56, 64, 89] whereby the state of the system is stored (or checkpointed) at pre-determined intervals so that in the event of a fault occurring, the computation can be restarted from the stored state; or, *Process Replication* [48, 53, 64, 89], whereby each critical component of the overall process is replicated one or more times so that in the event of a component being subject to a fault, the true state can be restored by making reference to an unaffected replica. Hybrid variants of these strategies can also be contemplated. The drawbacks of such approaches are self-evident.

Different techniques have been previously employed in order to achieve fault resilience for iterative methods in general and multigrid in particular. Replication was used in [30, 38], and checkpoint-restart in [3, 4, 26]. Stoyanov and Webster [90] proposed a method based on selective reliability for fixed point methods, a concept that has been introduced in [46, 47, 65]. Finally, different authors [59, 69, 70] have proposed forward recovery methods to mitigate the effect of faults. The efficient detection of faults has been addressed in the context of various applications[14, 26, 81].

2.1 Modeling Faults

In order to take account of the effect of detectable faults on the stability and convergence of a numerical algorithm, we develop a simple probabilistic model for this behavior and describe how it is incorporated into the formulation of iterative algorithms. As a matter of fact, it will transpire that our model for detectable faults will also apply to the case of silent faults that are relatively small.

Our preferred approach to mitigating the effects of faults in multigrid and related iterative algorithms is rather simple: *the lost or corrupted result expected from a node is replaced by a zero*. That is to say, if a value $x \in \mathbb{R}$ is prone to possible hard faults, then we propose to continue the global computation using $\tilde{x} \in \mathbb{R}$ in place of x, where

$$\tilde{x} = \begin{cases} 0 & \text{if a fault is detected,} \\ x & \text{otherwise.} \end{cases}$$

This minimalist, *laissez-faire*, approach has obvious attractions in terms of restoring the system to a valid state without the need to (i) halt the execution (other than to perhaps migrate data to a substitute node in the event of a node failure); (ii) take or read any data checkpoints; (iii) recompute any lost quantities; or, (iv) compromise on resources through having to replicate processes. However, the efficacy of *laissez-faire* depends on the extent to which the convergence characteristics of the resulting algorithm mirror those of the fault-free deterministic variant.

2.2 Probabilistic Model of Faults

In order to model the effects of the laissez-faire fault mitigation approach on an algorithm, we introduce a Bernoulli random variable given by

$$\chi = \begin{cases} 0 & \text{with probability } q, \\ 1 & \text{with probability } 1 - q. \end{cases} \tag{2.1}$$

If a scalar variable x is subject to faults, then the effect of the fault and the laissez-faire strategy is modeled by replacing the value of x by the new value $\tilde{x} = \chi x$. Evidently $\tilde{x}$ is a random variable with mean and variance given by $\mathbb{E}[\tilde{x}] = (1-q)x$ and $\mathbb{V}[\tilde{x}] = q(1-q)x^2$. By the same token, the effect of a fault on a vector-valued variable $x \in \mathbb{R}^n$ is modeled in a similar fashion by defining $\tilde{x} = \boldsymbol{\mathcal{X}} x$, where

$$\boldsymbol{\mathcal{X}} = \operatorname{diag}(\chi_1, \ldots, \chi_n)$$

and χ_i are identically distributed Bernoulli random variables. The variables χ_i can be independent, thus modeling componentwise faults, or block-dependent, as would be the case for a node failure.

More formally, for given $\varepsilon > 0$, let $\mathbb{S}_\varepsilon$ denote a set consisting of random matrices satisfying the following conditions:

Assumption 5*:*

1. Each $\boldsymbol{\mathcal{X}} \in \mathbb{S}_\varepsilon$ is a random diagonal matrix.
2. For every $\boldsymbol{\mathcal{X}} \in \mathbb{S}_\varepsilon$, there holds $\mathbb{E}[\boldsymbol{\mathcal{X}}] = e(\boldsymbol{\mathcal{X}})\boldsymbol{I}$, where $e(\boldsymbol{\mathcal{X}}) > 0$, and $|e(\boldsymbol{\mathcal{X}}) - 1| \leq C\varepsilon$ for some fixed $C > 0$.
3. For every $\boldsymbol{\mathcal{X}} \in \mathbb{S}_\varepsilon$ there holds $\|\mathbb{V}[\boldsymbol{\mathcal{X}}]\|_2 = \max_{i,j} |\operatorname{Cov}(\boldsymbol{\mathcal{X}}_{ii}, \boldsymbol{\mathcal{X}}_{jj})| \leq \varepsilon$.

The parameter ε is regarded as being small meaning that the random matrices $\boldsymbol{\mathcal{X}}$ behave, with high probability, like an identity matrix. While the value of ε could vary from operation to operation, for example to take into account that denser matrix-vector products take more time and are therefore more likely to be hit by faults, we neglect this aspect in the analysis for the sake of clarity.

Important examples of random faults covered by Assumption 5 are

1. *Componentwise detectable faults*

   A typical example of a detectable fault would be flipping of individual bits in a floating point number, resulting in a large enough upset to be detected,

or a pointer corruption that leads to an invalid memory address. Such cases can be treated using the laissez-faire strategy described above, and therefore modeled as componentwise faults with

$$\boldsymbol{\mathcal{X}} = \operatorname{diag}(\chi_1, \ldots, \chi_n), \tag{2.2}$$

where χ_i are independent and identically ε-distributed Bernoulli random variables, i. e.

$$\chi_i = \begin{cases} 0 & \text{with probability } \varepsilon, \\ 1 & \text{with probability } 1 - \varepsilon. \end{cases}$$

2. *Blockwise detectable faults*

   In the event of a node failure and application of the laissez-faire strategy, all the components of a vector that were residing on the node will be zeroed out. This can be modeled by a random matrix with independent diagonal *blocks*:

$$\boldsymbol{\mathcal{X}} = \operatorname{blockdiag}(\chi_1 \boldsymbol{I}_1, \ldots, \chi_N \boldsymbol{I}_N), \tag{2.3}$$

   where χ_i are independent and identically ε-distributed Bernoulli random variables and $\boldsymbol{I}_j$ are identity matrices.

3. *Silent faults*

   Some bit flips in the exponent, sign or in the significant components of the mantissa may result in a relatively large error that becomes detectable. Such cases could be treated using the laissez-faire strategy described earlier. However, in other cases such faults may give a relatively small error and, as a result, be difficult (or impossible) to identify. Such silent faults can be modeled by a random matrix

$$\boldsymbol{\mathcal{X}} = \boldsymbol{I} + \operatorname{diag}(\eta_1 \chi_1, \ldots, \eta_n \chi_n) \tag{2.4}$$

where η_i are independent and identically distributed random variables, and χ_i are independent and identically distributed Bernoulli random variables, such that $\boldsymbol{\mathcal{X}} \in \mathbb{S}_\varepsilon$. In particular, this in includes the cases of frequent faults with small impact and of rare but large upsets.

The matrices $\boldsymbol{\mathcal{X}}$ will appear at various points in our model for the Fault-Prone Multigrid Algorithm. It is easy to see that if $\boldsymbol{\mathcal{X}}_1$ and $\boldsymbol{\mathcal{X}}_2$ are independent diagonal matrices, then $\boldsymbol{\mathcal{X}}_1\boldsymbol{\mathcal{X}}_2$ is again a random diagonal matrix.

Sometimes, a random matrix will appear in two (or more) different places in a single equation and it will be important to clearly distinguish between the cases where (i) each of the two matrices is a *different* realization of the random matrix, or (ii) the matrices both correspond to the *same* realization of the random matrix. We shall adopt the convention whereby should a symbol appear twice, then (ii) holds: i.e. the two occurrences represent the *same realization* of the random matrix and the matrices are therefore identical. However, if the square or higher power of a random matrix appears then (i) holds: i.e. each of the matrices in the product is a *different* realization of the same random matrix.

2.3 Scope of Faults Covered by the Analysis

Our analysis will cover faults that can be represented by a random matrix belonging to $\mathbb{S}_\varepsilon$. This means that the analysis will cover each of the cases (i) when the fault is detectable and mitigated using the laissez-faire strategy, and (ii) when the fault is silent but is relatively small in the sense that it may be modeled using (2.4). However, our analysis will not cover the important case involving faults which result in entries in the matrices or right hand sides in the problem being corrupted. Equally well, our analysis will not cover the case of bit flips that result in a large relative error but which nevertheless remain undetected. Finally, this work is restricted to the solve phase of multigrid; we assume that the setup phase is protected from faults.

2.4 A Model for Fault-Prone Multigrid

The multigrid algorithm comprises a number of steps each of which may be affected by faults were the algorithm to be implemented on a fault-prone machine. In the absence of faults, a single iteration of multigrid replaces the current iterate x_ℓ by $MG_\ell(b_\ell, x_\ell)$ defined in Algorithm 1. However, in the presence of faults, a single iteration of multigrid means that x_ℓ is replaced by $\mathcal{MG}_\ell(b_\ell, x_\ell)$ where $\mathcal{MG}_\ell$ differs from MG_ℓ in general due to corruption of intermediate steps in the computations arising from faults. The object of this section is to develop a model for faults and define the corresponding operators $\mathcal{MG}_\ell$, $\ell \in \mathbb{N}$. For simplicity, we assume that the coarse grid is of moderate size meaning that the exact solve $\boldsymbol{A}_0^{-1}$ on the coarsest grid is not prone to faults, i.e. $\mathcal{MG}_0(d_0, \cdot) = \boldsymbol{A}_0^{-1} d_0$. This is a reasonable assumption when the size of the coarse grid problem is sufficiently small that either faults are not an issue or, if they are, then replication can be used to achieve fault resilience. The extension of the analysis to the case of fault-prone coarse grid correction does not pose any fundamental difficulties but is not pursued in the present work.

2.5 Faults in the Smoothener

A single application of the smoothener S_ℓ to an iterate x_ℓ takes the form

$$S_\ell(b_\ell, x_\ell) = x_\ell + \boldsymbol{N}_\ell(b_\ell - \boldsymbol{A}_\ell x_\ell) \tag{2.5}$$

with each of the interior sub-steps in eq. (2.5) being susceptible to faults.

The innermost step is the computation of the residual $\rho_\ell = b_\ell - \boldsymbol{A}_\ell x_\ell$. The action of the matrix $\boldsymbol{A}_\ell$ is applied repeatedly throughout the solution phase. We shall assume that neither the entries or structure of matrix $\boldsymbol{A}_\ell$ nor the right hand side b_ℓ are subject to corruption, only computation involving them. This would be the case were the information needed to compute the action of $\boldsymbol{A}_\ell$ placed in non-volatile random access memory (NVRAM) with relatively modest overhead. The net effect

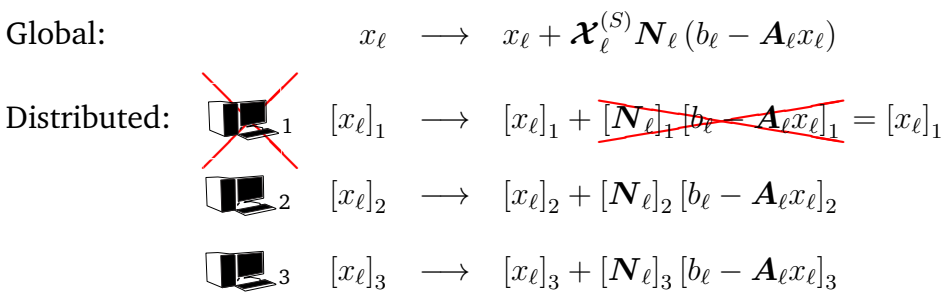


Figure 2.1: Schematic representation of a node failure during smoothening and the remedial action taken by the laissez-faire approach in the case of three compute nodes. $[\bullet]_i$ represents the part of the quantity $\bullet$ local to node i.

of faults in the computation of ρ_ℓ is that the preconditioner $\boldsymbol{N}_\ell$ does not act on the true residual but rather on a corrupted version modeled by $\boldsymbol{\mathcal{X}}_1 \rho_\ell$ where $\boldsymbol{\mathcal{X}}_1$ is a random diagonal matrix.

By the same token, the action of $\boldsymbol{N}_\ell$ is prone to faults, meaning that the true result $\boldsymbol{N}_\ell \boldsymbol{\mathcal{X}}_1 \rho_\ell$ may be corrupted and is therefore modeled by $\boldsymbol{\mathcal{X}}_2 \boldsymbol{N}_\ell \boldsymbol{\mathcal{X}}_1 \rho_\ell$, where $\boldsymbol{\mathcal{X}}_2$ is yet another random diagonal matrix. The matrix $\boldsymbol{N}_\ell$ corresponding to damped Jacobi is diagonal and hence $\boldsymbol{\mathcal{X}}_2 \boldsymbol{N}_\ell \boldsymbol{\mathcal{X}}_1 = \boldsymbol{\mathcal{X}}^{(S)} \boldsymbol{N}_\ell$, where $\boldsymbol{\mathcal{X}}^{(S)} = \boldsymbol{\mathcal{X}}_1 \boldsymbol{\mathcal{X}}_2$ is again a random diagonal matrix. Consequently, the combined effect of the two sources of error can be modeled by a single random diagonal matrix.

In summary, our model for the action of a smoothener prone to faults consists of replacing the true smoothener S used in the pre- and post-smoothening step in the multigrid algorithm by the *non-deterministic smoothener*

$$\mathcal{S}_\ell(b_\ell, x_\ell) = x_\ell + \boldsymbol{\mathcal{X}}_\ell^{(S)} \boldsymbol{N}_\ell (b_\ell - \boldsymbol{A}_\ell x_\ell) \tag{2.6}$$

in which $\boldsymbol{\mathcal{X}}_\ell^{(S)}$ is a random diagonal matrix which models the effect of the random faulty nature of the underlying hardware. The model eq. (2.6) tacitly assumes that the current iterate x_ℓ remains fault-free. We illustrate the remedial action to a node failure in Figure 2.1.

2.6 Faults in Restriction, Prolongation and Coarse Grid Correction

The restriction of the residual described by the step

$$d_{\ell-1} = \boldsymbol{R}(b_\ell - \boldsymbol{A}_\ell x_\ell) \tag{2.7}$$

is prone to faults. Firstly, as in the case of the smoothener, the true residual ρ_ℓ is prone to corruption and is modeled by $\boldsymbol{\mathcal{X}}_\ell^{(\rho)}\rho_\ell$. The resulting residual is then operated on by the restriction $\boldsymbol{R}$, which is itself prone to faults modeled using a random diagonal matrix $\boldsymbol{\mathcal{X}}_{\ell-1}^{(R)}$. We arrive at the following model for the effect of faults on eq. (2.7):

$$d_{\ell-1} = \boldsymbol{\mathcal{X}}_{\ell-1}^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}_\ell^{(\rho)}(b_\ell - \boldsymbol{A}_\ell x_\ell).$$

The coarse grid correction $e_{\ell-1}$ is obtained by performing γ iterations of $\mathcal{MG}_{\ell-1}$ with data $d_{\ell-1}$ and a zero initial iterate. The effect of faults when applying the prolongation to $e_{\ell-1}$ is also modeled by a random diagonal matrix $\boldsymbol{\mathcal{X}}_\ell^{(P)}$ leading to the following model for the effect of faults on the coarse grid correction and prolongation steps:

$$x_\ell \leftarrow x_\ell + \boldsymbol{\mathcal{X}}_\ell^{(P)}\boldsymbol{P}e_{\ell-1}.$$

2.7 Model for Multigrid Algorithm in Presence of Faults

Replacing each of the steps in the fault-free Multigrid Algorithm 1 with their non-deterministic equivalent yields the following model for the Fault-Prone Multigrid Algorithm 2 and defines the associated fault-prone multilevel operators $\mathcal{MG}_\ell$, $\ell \in \mathbb{N}$.

The classical approach to the analysis of iterative solution methods for linear systems uses the notion of an iteration matrix as given in 1.3. For example, in the case

Algorithm 2 Model for Fault-Prone Multigrid Algorithm $\mathcal{MG}_\ell$ where $\boldsymbol{\mathcal{X}}_\ell^{(\bullet)}$ are random diagonal matrices.

Input: Right hand side b_ℓ; Initial iterate x_ℓ
Output: $\mathcal{MG}_\ell(b_\ell, x_\ell)$

if $\ell = 0$ **then return** $\boldsymbol{A}_0^{-1} b_0$ // Exact solve on coarsest grid
else
  for $i \leftarrow 1$ **to** ν_1 **do**
    $x_\ell \leftarrow \mathcal{S}_\ell\left(b_\ell, x_\ell\right)$ // ν_1 pre-smoothening steps
  $d_{\ell-1} \leftarrow \boldsymbol{\mathcal{X}}_{\ell-1}^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}_\ell^{(\rho)} \left(b_\ell - \boldsymbol{A}_\ell x_\ell\right)$ // Restriction to coarser grid
  $e_{\ell-1} \leftarrow 0$
  for $j \leftarrow 1$ **to** γ **do**
    $e_{\ell-1} \leftarrow \mathcal{MG}_{\ell-1}(d_{\ell-1}, e_{\ell-1})$ // γ coarse grid correction steps
  $x_\ell \leftarrow x_\ell + \boldsymbol{\mathcal{X}}_\ell^{(P)} \boldsymbol{P} e_{\ell-1}$ // Prolongation to finer grid
  for $i \leftarrow 1$ **to** ν_2 **do**
    $x_\ell \leftarrow \mathcal{S}_\ell\left(b_\ell, x_\ell\right)$ // ν_2 post-smoothening steps

of the fault-free smoothening step $\boldsymbol{E}_\ell^S = \boldsymbol{I} - \boldsymbol{N}_\ell \boldsymbol{A}_\ell$ is the iteration matrix, with preconditioner $\boldsymbol{N}_\ell$. The corresponding result for the fault-prone smoothener eq. (2.6) is given by

$$x - \mathcal{S}_\ell(\boldsymbol{A}_\ell x, y) = \boldsymbol{\mathcal{E}}_\ell^S (x - y), \quad \forall x, y \in \mathbb{R}^{n_\ell} \tag{2.8}$$

where the iteration matrix $\boldsymbol{\mathcal{E}}_\ell^S = \boldsymbol{I} - \boldsymbol{\mathcal{X}}_\ell^{(S)} \boldsymbol{N}_\ell \boldsymbol{A}_\ell$ is now *random*. Similarly, we define the iteration matrix for the Fault-Prone Multigrid Algorithm 2 by the equation

$$x - \mathcal{MG}_\ell(\boldsymbol{A}_\ell x, y) = \boldsymbol{\mathcal{E}}_\ell (x - y), \quad \forall x, y \in \mathbb{R}^{n_\ell}. \tag{2.9}$$

In particular, $x - \mathcal{MG}_0(\boldsymbol{A}_0 x, \bullet) = 0$ and hence $\boldsymbol{\mathcal{E}}_0 = \boldsymbol{0}$, i.e. the zero matrix.

The ν_1 pre-smoothening steps in Algorithm 2 can be expressed in the form

$$x^{(0)} = x_\ell; \quad x^{(i)} = \mathcal{S}_\ell(b_\ell, x^{(i-1)}), \quad i = 1, \ldots, \nu_1.$$

Using eq. (2.8) gives

$$x - x_\ell^{(i)} = \boldsymbol{\mathcal{E}}_\ell^S (x - x_\ell^{(i-1)}), \quad i = 1, \ldots, \nu_1$$

and hence

$$x - x_\ell^{(\nu_1)} = \left(\boldsymbol{\mathcal{E}}_\ell^S\right)^{\nu_1} (x - x_\ell). \tag{2.10}$$

Using this notation means that the vector $d_{\ell-1}$ appearing in Algorithm 2 is given by

$$d_{\ell-1} = \boldsymbol{\mathcal{X}}_{\ell-1}^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}_{\ell}^{(\rho)} (b_\ell - \boldsymbol{A}_\ell x_\ell^{(\nu_1)}) = \boldsymbol{\mathcal{X}}_{\ell-1}^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}_{\ell}^{(\rho)} \boldsymbol{A}_\ell (x - x_\ell^{(\nu_1)}).$$

where $b_\ell = \boldsymbol{A}_\ell x$.

The coarse grid correction steps of Algorithm 2 can be written in the form

$$e_{\ell-1}^{(0)} = 0; \quad e_{\ell-1}^{(j)} = \mathcal{MG}_{\ell-1}(d_{\ell-1}, e_{\ell-1}^{(j-1)}), \quad j = 1, \ldots, \gamma.$$

Let $z = \boldsymbol{A}_{\ell-1}^{-1} d_{\ell-1}$ then

$$z - e_{\ell-1}^{(j)} = z - \mathcal{MG}_{\ell-1}(d_{\ell-1}, e_{\ell-1}^{(j-1)}) = \boldsymbol{\mathcal{E}}_{\ell-1}(z - e_{\ell-1}^{(j-1)}), \quad j = 1, \ldots, \gamma$$

thanks to eq. (2.9). Iterating this result and recalling that $e_{\ell-1}^{(0)} = 0$, gives $z - e_{\ell-1}^{(\gamma)} = \boldsymbol{\mathcal{E}}_{\ell-1}^{\gamma} z$, or, equally well,

$$e_{\ell-1}^{(\gamma)} = (\boldsymbol{I} - \boldsymbol{\mathcal{E}}_{\ell-1}^{\gamma}) \boldsymbol{A}_{\ell-1}^{-1} d_{\ell-1}.$$

Using these notations, the prolongation to the finer grid step in Algorithm 2 takes the form

$$x_\ell^{(P)} = x_\ell^{(\nu_1)} + \boldsymbol{\mathcal{X}}_{\ell}^{(P)} \boldsymbol{P} e_{\ell-1}^{(\gamma)}$$

and hence, by collecting results, we deduce that

$$\begin{aligned} x - x_\ell^{(P)} &= (x - x_\ell^{(\nu_1)}) - \boldsymbol{\mathcal{X}}_{\ell}^{(P)} \boldsymbol{P} e_{\ell-1}^{(\gamma)} \\ &= \left[\boldsymbol{I} - \boldsymbol{\mathcal{X}}_{\ell}^{(P)} \boldsymbol{P} (\boldsymbol{I} - \boldsymbol{\mathcal{E}}_{\ell-1}^{\gamma}) \boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{\mathcal{X}}_{\ell-1}^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}_{\ell}^{(\rho)} \boldsymbol{A}_\ell \right] (x - x_\ell^{(\nu_1)}). \end{aligned}$$

Arguments identical to those leading to eq. (2.10) give the following result

$$x - \mathcal{MG}_\ell(\boldsymbol{A}_\ell x, x_\ell) = \left(\boldsymbol{\mathcal{E}}_\ell^S\right)^{\nu_2} (x - x_\ell^{(P)}).$$

which, in view of identity eq. (2.9) and eq. (2.10), yields the following recursive formula for the iteration matrix of the Fault-Prone Multigrid Algorithm 2:

$$\boldsymbol{\mathcal{E}}_\ell = \left(\boldsymbol{\mathcal{E}}_\ell^{S,\text{post}}\right)^{\nu_2} \left[\boldsymbol{I} - \boldsymbol{\mathcal{X}}_{\ell}^{(P)} \boldsymbol{P} (\boldsymbol{I} - \boldsymbol{\mathcal{E}}_{\ell-1}^{\gamma}) \boldsymbol{A}_{\ell-1}^{-1} \boldsymbol{\mathcal{X}}_{\ell-1}^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}_{\ell}^{(\rho)} \boldsymbol{A}_\ell \right] \left(\boldsymbol{\mathcal{E}}_\ell^{S,\text{pre}}\right)^{\nu_1}, \qquad (2.11)$$

for $\ell = 1, \ldots, L$ with $\boldsymbol{\mathcal{E}}_0 = \mathbf{0}$. Here, we have used superscripts pre and post to reflect that the pre- and post-smootheners are independent realizations of the same random matrix.

By setting $\boldsymbol{\mathcal{E}}_{L-1} = \mathbf{0}$ and applying the recursion eq. (2.11) in the case $\ell = L$ yields a formula for the iteration matrix of the *Fault-Prone Two Grid Algorithm*:

$$\begin{aligned} \boldsymbol{\mathcal{E}}_L^{TG} &= \left(\boldsymbol{\mathcal{E}}_L^{S,\text{post}}\right)^{\nu_2} \left[\boldsymbol{I} - \boldsymbol{\mathcal{X}}_L^{(P)} \boldsymbol{P} \boldsymbol{A}_{L-1}^{-1} \boldsymbol{\mathcal{X}}_{L-1}^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}_L^{(\rho)} \boldsymbol{A}_L\right] \left(\boldsymbol{\mathcal{E}}_L^{S,\text{pre}}\right)^{\nu_1} \\ &= \left(\boldsymbol{\mathcal{E}}_L^{S,\text{post}}\right)^{\nu_2} \boldsymbol{\mathcal{E}}_L^{CG} \left(\boldsymbol{\mathcal{E}}_L^{S,\text{pre}}\right)^{\nu_1}, \end{aligned} \tag{2.12}$$

corresponding to using an exact solver on level $L-1$. Here $\boldsymbol{\mathcal{E}}_L^{CG}$ is the iteration matrix of the exact fault-prone coarse grid correction.

Algorithm 3 Model for Fault-Prone Two-Grid Algorithm $\mathcal{TG}_L$ where $\boldsymbol{\mathcal{X}}_\cdot^{(\bullet)}$ are random diagonal matrices.

Input: Right hand side b_L; Initial iterate x_L
Output: $\mathcal{TG}_L(b_L, x_L)$

for $i \leftarrow 1$ **to** ν_1 **do**
  $x_L \leftarrow \mathcal{S}_L(b_L, x_L)$ // ν_1 pre-smoothening steps
$d_{L-1} \leftarrow \boldsymbol{\mathcal{X}}_{L-1}^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}_L^{(\rho)} (b_L - \boldsymbol{A}_L x_L)$ // Restriction to coarser grid
$e_{L-1} \leftarrow \boldsymbol{A}_{L-1}^{-1} d_{L-1}$ // γ coarse grid correction
$x_L \leftarrow x_L + \boldsymbol{\mathcal{X}}_L^{(P)} \boldsymbol{P} e_{L-1}$ // Prolongation to finer grid
for $i \leftarrow 1$ **to** ν_2 **do**
  $x_L \leftarrow \mathcal{S}_L(b_L, x_L)$ // ν_2 post-smoothening steps

Having determined the iteration matrices for both Fault-Prone Two Grid and Multigrid Algorithm, the question of convergence of theses methods arises. In the fault-free setting, as presented in Section 1.3, the iteration matrix is constant and its asymptotic convergence rate is given by the spectral radius. In the present setting, however, the iteration matrix changes from iteration to iteration. Therefore, we discuss the theory of products of random matrices in the next chapter and generalizes the spectral radius to the random setting.

Chapter 3

Lyapunov Exponents and Their Computation

In the previous chapter, we determined the random iteration matrices corresponding to the Fault-Prone Two Grid and Multigrid Methods. To analyze their respective convergence, we introduce the theory of products of random matrices that concerns itself with the study of the asymptotic behavior of product sequences of random matrices. Most importantly, we will state a "law of large numbers" for random matrices, and discuss a multitude of approaches to the computation and estimation of the asymptotic convergence rate.

3.1 Lyapunov Exponents

Let $\boldsymbol{\mathcal{A}}$ be a random matrix of size $n \times n$ with distribution μ. We will assume that μ is a discrete probability distribution and picks values in a finite set $\mathcal{A} \subset GL_n$.

We define the *Lyapunov exponent* $\gamma(\boldsymbol{\mathcal{A}})$ as

$$\gamma(\boldsymbol{\mathcal{A}}) := \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\log \left\|\boldsymbol{\mathcal{A}}^N\right\|\right].$$

(Remember that according to our notation $\boldsymbol{\mathcal{A}}^N = \prod_{j=1}^N \boldsymbol{A_j}$, where $\boldsymbol{A_j}$ are all independent and of the same distribution as $\boldsymbol{\mathcal{A}}$.) This quantity does not depend on the choice of the norm $\|\cdot\|$ as all norms are equivalent in finite dimension. The definition is motivated by the equality

$$\log \rho(\boldsymbol{A}) = \lim_{N\to\infty} \frac{1}{N} \log \left\| \boldsymbol{A}^N \right\|$$

that holds for non-random matrices $\boldsymbol{A}$.

Furstenberg and Kesten[55] showed that provided that $\mathbb{E}\left[\log \left(\boldsymbol{\mathcal{A}}\right)^+\right] < \infty$, $\gamma\left(\boldsymbol{\mathcal{A}}\right)$ exists and

$$\gamma\left(\boldsymbol{\mathcal{A}}\right) = \lim_{N\to\infty} \frac{1}{N} \log \left\| \boldsymbol{\mathcal{A}}^N x_0 \right\| \qquad \text{a.s.}$$

The *Lyapunov spectral radius*

$$\varrho\left(\boldsymbol{\mathcal{A}}\right) := e^{\gamma(\boldsymbol{\mathcal{A}})} = \lim_{N\to\infty} \left\| \boldsymbol{\mathcal{A}}^N x_0 \right\|^{\frac{1}{N}} \qquad \text{a.s.}$$

is the quantity corresponding to the spectral radius of a non-random matrix. $\varrho\left(\boldsymbol{\mathcal{A}}\right) < 1$ means that the matrix product is convergent, whereas $\varrho\left(\boldsymbol{\mathcal{A}}\right) > 1$ means that it is divergent.

We also define the *generalized Lyapunov exponents* $L(\boldsymbol{\mathcal{A}}, \alpha)$ for $\alpha \neq 0$ as

$$L(\boldsymbol{\mathcal{A}}, \alpha) := \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left[\left\| \boldsymbol{\mathcal{A}}^N \right\|^\alpha\right]$$

and the *generalized Lyapunov spectral radii* $\varrho_\alpha\left(\boldsymbol{\mathcal{A}}\right)$ as

$$\varrho_\alpha\left(\boldsymbol{\mathcal{A}}\right) := e^{\frac{1}{\alpha} L(\boldsymbol{\mathcal{A}}, \alpha)}.$$

It can be shown that $\frac{1}{N} \log \left\| \boldsymbol{\mathcal{A}}^N x_0 \right\|$ satisfies a large deviation principle with a rate function I which means that

$$P\left(\frac{1}{N} \log \left\| \boldsymbol{\mathcal{A}}^N x_0 \right\| = \lambda \right) \approx e^{-NI(\lambda)}.$$

From the above, we find that I satisfies

$$I\left(\gamma\left(\boldsymbol{\mathcal{A}}\right)\right) = 0,$$

$$I'\left(\gamma\left(\boldsymbol{\mathcal{A}}\right)\right)=0,$$
$$I''\left(\gamma\left(\boldsymbol{\mathcal{A}}\right)\right)>0.$$

We can approximately calculate

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\mathcal{A}}^{N}x_{0}\right\|^{\alpha}\right] &= \mathbb{E}\left[e^{N\alpha\frac{1}{N}\log\left\|\boldsymbol{\mathcal{A}}^{N}x_{0}\right\|}\right]\\ &\approx \int e^{N\alpha\lambda}e^{-NI(\lambda)}\,d\lambda\\ &\approx e^{N\sup_{\lambda}\{\alpha\lambda-I(\lambda)\}}\end{aligned}$$

and hence

$$L(\boldsymbol{\mathcal{A}},\alpha)=\sup_{\lambda}\alpha\lambda-I(\lambda).$$

This means that L is the Fenchel-Legendre transform of the rate function I. For each α^{*}, there is a characteristic growth rate λ^{*} that corresponds to it, given by

$$\frac{\partial L}{\partial\alpha}(\alpha^{*})=\lambda^{*}.$$

Therefore, the values of $L(\boldsymbol{\mathcal{A}},\alpha^{*})$ depend on the unlikely sequences $\left\{\boldsymbol{\mathcal{A}}^{N}x_{0}\right\}$ with growth rate λ^{*} different from $\gamma\left(\boldsymbol{\mathcal{A}}\right)$.

Finally, the Lyapunov spectral radius can also be estimated by neglecting the interactions between different values taken by the random matrix $\boldsymbol{\mathcal{A}}$. We define

$$\overline{\varrho}\left(\boldsymbol{\mathcal{A}}\right):=\exp\sup_{\|x\|=1}\mathbb{E}\left[\log\left\|\boldsymbol{\mathcal{A}}x\right\|\right],$$
$$\tilde{\varrho}\left(\boldsymbol{\mathcal{A}}\right):=\exp\mathbb{E}\left[\log\left\|\boldsymbol{\mathcal{A}}\right\|\right]$$

and

$$\tilde{\varrho}_{\alpha}\left(\boldsymbol{\mathcal{A}}\right):=\exp\frac{1}{\alpha}\log\mathbb{E}\left[\left\|\boldsymbol{\mathcal{A}}\right\|^{\alpha}\right]$$

for induced sub-multiplicative norms. These quantities obviously depend on the choice of the norms.

Hence, we have for $\alpha > 0$

$$\begin{array}{ccc} \varrho_{-\alpha}(\boldsymbol{\mathcal{A}}) \leq & \varrho(\boldsymbol{\mathcal{A}}) & \leq \varrho_{\alpha}(\boldsymbol{\mathcal{A}}) \\ & \leq & \\ \leq & \overline{\varrho}(\boldsymbol{\mathcal{A}}) & \leq \\ & \leq & \\ \tilde{\varrho}_{-\alpha}(\boldsymbol{\mathcal{A}}) \leq & \tilde{\varrho}(\boldsymbol{\mathcal{A}}) & \leq \tilde{\varrho}_{\alpha}(\boldsymbol{\mathcal{A}}) \end{array}$$

The horizontal inequalities follow from Jensen's inequality.

Moreover, if μ picks only one matrix $\boldsymbol{A}$, so that we are in fact in the non-random case, we find

$$\varrho(\boldsymbol{A}) = \varrho_{\alpha}(\boldsymbol{A}) = \rho(\boldsymbol{A}).$$

The quantities $\overline{\varrho}(\boldsymbol{A})$, $\tilde{\varrho}(\boldsymbol{A})$ and $\tilde{\varrho}_{\alpha}(\boldsymbol{A})$ can then be made arbitrarily close to $\rho(\boldsymbol{A})$ by choosing suitable vector and matrix norms.

The theory of Lyapunov exponents is thoroughly discussed in the book [19] by Bougerol and Lacroix.

The following example adapted from [37] shows that the interaction between the values taken by $\boldsymbol{\mathcal{A}}$ can be important:

Example 2.

Let $a \in \mathbb{R}$ and let the random matrix $\boldsymbol{\mathcal{A}}$ take values

$$\begin{pmatrix} a & 0 \\ 0 & \frac{1}{2a} \end{pmatrix} \quad \textit{with probability } (1-p) \quad \textit{and} \quad \boldsymbol{S} := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \textit{with probability } p.$$

We are interested in the growth and decay of

$$X_N := \boldsymbol{\mathcal{A}}^N x_0 = \prod_{j=1}^{N} \boldsymbol{A_j} x_0,$$

where

$$x_0 = \alpha \vec{e}_1 + \beta \vec{e}_2$$

and $\boldsymbol{A_j}$ *are independent random matrices that are distributed like* $\boldsymbol{\mathcal{A}}$*. Obviously, we have* $\varrho(\boldsymbol{\mathcal{A}}) = \max\left\{a, \frac{1}{2a}\right\}$ *for* $p = 0$ *and* $\varrho(\boldsymbol{\mathcal{A}}) = 1$ *for* $p = 1$*.*

We define the stopping times

$$\begin{aligned}
&\tau_0 = 0, \\
&\delta_k = \inf\left\{i \geq 1 : \boldsymbol{A}_{\boldsymbol{\tau_{k-1}+i}} = \boldsymbol{S}\right\}, && k \geq 1, \\
&\tau_k = \inf\left\{i \geq 1 : \boldsymbol{A}_{\boldsymbol{\delta_k+i}} = \boldsymbol{S}\right\}, && k \geq 1.
\end{aligned}$$

And set

$$\sigma_k = \sum_{n=1}^{k} \delta_n + \tau_n, \qquad k \geq 1.$$

Between σ_k *and* $\sigma_k + \delta_{k+1}$ *the first component is expanding by* a *and the second one is damped by* $\frac{1}{2a}$*. Between* $\sigma_k + \delta_{k+1}$ *and* σ_{k+1}*, the roles are exchanged. Moreover,* $\delta_k,\ \tau_k$ *are all independent and identically geometrically distributed with parameter* p*. We have*

$$\log|X_{j,1}| = \log|\alpha| + \log a\left[\sum_{n=1}^{k}(\delta_n - \tau_n) + N - \sigma_k\right] - \log 2\sum_{n=1}^{k}(\tau_n - 1)$$

if $\sigma_k \leq N < \sigma_k + \delta_{k+1}$ *or*

$$\begin{aligned}
\log|X_{j,1}| = \log|\beta| + \log a\left[\sum_{n=1}^{k}(\tau_n - \delta_n) + 2\delta_{k+1} - N + \sigma_k - 1\right] \\
- \log 2\left[\sum_{n=1}^{k+1}(\tau_n - 1) + N - \sigma_k - \delta_{k+1}\right]
\end{aligned}$$

if $\sigma_k + \delta_{k+1} \leq N < \sigma_{k+1}$*. Similar expressions hold for* $\log|X_{N,2}|$*. Now,*

$$\frac{\sigma_k}{k} \leq \frac{N}{k} \leq \frac{\sigma_{k+1}}{k}$$

and both upper and lower bound converges almost surely to $\mathbb{E}[\delta_1 + \tau_1] = \frac{2}{p}$*.*

$$\frac{1}{k}\sum_{n=1}^{k}\tau_n,\ \frac{1}{k}\sum_{n=1}^{k}\delta_n \xrightarrow{a.s.} \mathbb{E}[\tau_1] = \frac{1}{p},$$

$$\frac{\delta_{k+1}}{N} \xrightarrow{a.s.} 0,$$

so we find

$$\frac{1}{N}\log|X_{j,1}| \xrightarrow{a.s.} -\log(2)\frac{p}{2}\left(\frac{1}{p}-1\right)$$

and

$$|X_{N,1}|^{\frac{1}{N}} \xrightarrow{a.s.} 2^{\frac{p-1}{2}}.$$

The same is true for the second component and hence

$$\varrho(\boldsymbol{\mathcal{A}}) = \lim_{N\to\infty} \|X_N\|^{\frac{1}{N}} = \begin{cases} \max\left\{a, \frac{1}{2a}\right\} & \text{for } p=0, \\ 2^{\frac{p-1}{2}} & \text{for } 0<p<1, \\ 1 & \text{for } p=1. \end{cases}$$

On the other hand,

$$\log\tilde{\varrho}(\boldsymbol{\mathcal{A}}) = \mathbb{E}\left[\log\|\boldsymbol{\mathcal{A}}\|_2\right] = (1-p)\log\max\left\{a, \frac{1}{2a}\right\}$$

and

$$\log\overline{\varrho}(\boldsymbol{\mathcal{A}}) = \sup_{\|x\|_2=1} \mathbb{E}\left[\log\|\boldsymbol{\mathcal{A}}x\|_2\right] = (1-p)\log\max\left\{a, \frac{1}{2a}\right\}.$$

So for $0<p<1$ *and* $a \notin \left[\frac{1}{2}, 1\right]$*,*

$$\varrho(\boldsymbol{\mathcal{A}}) < 1 < \tilde{\varrho}(\boldsymbol{\mathcal{A}}) = \overline{\varrho}(\boldsymbol{\mathcal{A}}).$$

Moreover, $\tilde{\varrho}(\boldsymbol{\mathcal{A}}),\ \overline{\varrho}(\boldsymbol{\mathcal{A}}) \to \infty$ *as* $a\to\infty$ *or* $a\to 0$*, whereas* $\varrho(\boldsymbol{\mathcal{A}})$ *is independent of* a*.*

Remark: The previous example shows that the Lyapunov spectral radius can be discontinuous with respect to the weights and matrices in the support of $\boldsymbol{\mathcal{A}}$. In particular, this means that we cannot conclude from $\rho(\boldsymbol{E}_l) < 1$ that $\varrho(\boldsymbol{\mathcal{E}}_l) < 1$ for small perturbations.

The next example shows that measuring unlikely sequences can lead to serious over-estimation of the convergence rate:

Example 3.
Let $\boldsymbol{\mathcal{A}}$ be a random 1×1 matrix taking values 2^a with probability $\frac{1}{a}$ and 1 with probability $\frac{a-1}{a}$. Hence the growth rates of all possible sequences are between 1 and 2^a. Then

$$\varrho\left(\boldsymbol{\mathcal{A}}\right) = \overline{\varrho}\left(\boldsymbol{\mathcal{A}}\right) = \tilde{\varrho}\left(\boldsymbol{\mathcal{A}}\right) = 2,$$
$$\varrho_\alpha\left(\boldsymbol{\mathcal{A}}\right) = \left(\frac{1}{a}2^{a\alpha} + \frac{a-1}{a}\right)^{\frac{1}{\alpha}},$$

and for large a we find $\varrho\left(\boldsymbol{\mathcal{A}}\right) \ll \varrho_\alpha\left(\boldsymbol{\mathcal{A}}\right)$, $\alpha > 0$.

The following theorem is given as an exercise in the book [19].

Theorem 3.
Suppose that a random matrix A has block triangular form

$$\boldsymbol{\mathcal{A}} = \begin{pmatrix} \boldsymbol{\mathcal{B}} & \boldsymbol{\mathcal{C}} \\ 0 & \boldsymbol{\mathcal{D}} \end{pmatrix}$$

with square blocks $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\mathcal{D}}$. Then

$$\gamma(\boldsymbol{\mathcal{A}}) = \max\{\gamma(\boldsymbol{\mathcal{B}}),\ \gamma(\boldsymbol{\mathcal{D}})\}.$$

An immediate consequence is

Lemma 4.
Suppose that the support $\mathcal{A}$ of a random matrix $\boldsymbol{\mathcal{A}}$ can be simultaneously triangularized, i.e. there exists an invertible matrix $\boldsymbol{P}$ such that $\boldsymbol{P}^{-1}\boldsymbol{B}\boldsymbol{P}$ is upper triangular for every $\boldsymbol{B} \in \mathcal{A}$. Then

$$\gamma\left(\boldsymbol{\mathcal{A}}\right) = \max_k \mathbb{E}\left[\log\left(\boldsymbol{P}^{-1}\boldsymbol{\mathcal{A}}\boldsymbol{P}\right)_{kk}\right].$$

In general, unless the matrices in the support of $\boldsymbol{\mathcal{A}}$ can be simultaneously triangularized, it is impossible to calculate $\varrho\left(\boldsymbol{\mathcal{A}}\right)$ analytically. Several different methods to do so numerically have been proposed:

- Monte-Carlo simulation of trajectories of $\boldsymbol{\mathcal{A}}^N x_0$,
- Approximation of the invariant distribution of the Markov chain $\frac{\boldsymbol{\mathcal{A}}^N x_0}{\left\|\boldsymbol{\mathcal{A}}^N x_0\right\|}$,
- cycle expansions,
- weak disorder expansions.

A good review of these methods is given in [37]. We discuss and use the Monte Carlo approach as well as the calculation of the invariant distribution below. The cycle expansion can only be applied successfully if the cardinality of $\mathcal{A}$ is small, whereas the weak disorder expansion works only for small deviations around a deterministic matrix. As the invariant distribution method is limited to low dimensional matrices, the Monte-Carlo approach is the only generally available option to obtain the asymptotic growth rate $\varrho\left(\boldsymbol{\mathcal{A}}\right)$.

The situation is slightly better for the generalized Lyapunov spectral radius $\varrho_\alpha\left(\boldsymbol{\mathcal{A}}\right)$ which can be calculated using

1. Vanneste's [97] resampled Monte-Carlo algorithm,
2. the Replica Trick.

We shortly describe both methods below. The resampled Monte-Carlo algorithm can be seen as the equivalent of the Monte-Carlo approach for the calculation of $\varrho\left(\boldsymbol{\mathcal{A}}\right)$ in that it is generally available. The more specialized Replica Trick can only be used to calculate generalized Lyapunov exponents of positive even order.

3.2 Monte-Carlo Approximation

We can approximate the Lyapunov exponent $\gamma(\boldsymbol{\mathcal{A}})$ by Monte-Carlo simulations of a trajectory

$$X_N(\omega) = \prod_{j=1}^{N} \boldsymbol{A_j}(\omega) x_0$$

for normalized random initial vector x_0 and calculating

$$\gamma(\boldsymbol{\mathcal{A}}) \approx \frac{1}{N} \log \|X_N(\omega)\| .$$

This makes sense because of the almost sure convergence of X_N to $\gamma(\boldsymbol{\mathcal{A}})$.

In order to avoid over- and underflow of the components of the vector $X_N(\omega)$, we renormalize after every step, i.e. we set

$$\tilde{X}_j := \frac{\boldsymbol{A_j}(\omega)\tilde{X}_{j-1}}{\left\|\boldsymbol{A_j}(\omega)\tilde{X}_{j-1}\right\|}$$
$$\alpha_j := \left\|\boldsymbol{A_j}(\omega)\tilde{X}_{j-1}\right\|$$

and approximate

$$\gamma(\boldsymbol{\mathcal{A}}) \approx \frac{1}{N} \sum_{j=1}^{N} \log \alpha_j.$$

3.3 Approximation via Invariant Distribution

Another characterization of γ can be given in terms of the invariant probability distribution ν of the Markov chain $\tilde{X}_N = \frac{X_N}{\|X_N\|}$ on the projective space $P\mathbb{R}^{d-1}$:

$$\gamma = \int_{P\mathbb{R}^{d-1}} d\nu(x) \int_{\mathbb{R}^{d\times d}} d\mu(A) \log \|Ax\|$$

Here, ν satisfies the invariance equation

$$\int_{P\mathbb{R}^{d-1}} d\nu(x)\ f(x) = \int_{P\mathbb{R}^{d-1}} d\nu(x) \int_{\mathbb{R}^{d\times d}} d\mu(A)\ f\left(\frac{Ax}{\|Ax\|}\right)$$

for any Borel measurable bounded function f on $P\mathbb{R}^{d-1}$.

If d is small and only finitely many matrices are picked by μ, it is possible to approximate ν and then calculate γ. The following method has been proposed by Embree and Trefethen in [49] for the calculation of the Lyapunov spectral radius of the random Fibonacci sequence

$$x_{n+1} = x_n \pm x_{n-1},$$

where the signs $=$ or $-$ are chosen with equal probability. Rewritten as a one-step sequence, this corresponds to picking the matrices

$$\boldsymbol{\mathcal{A}} = \begin{pmatrix} 0 & 1 \\ \pm 1 & 1 \end{pmatrix}$$

with equal probability. This has become a test problem for different algorithms to determine the Lyapunov spectral radius, because Viswanath [98] was able to determine with high precision that $\varrho(\boldsymbol{\mathcal{A}}) = 1.1319882\ldots$

We restrict ourselves to 2×2 matrices. Vectors $\vec{x}$ of $P\mathbb{R}$ can be parametrized by their slope $m = \frac{x_2}{x_1} \in (-\infty, \infty) \cup \{\infty\}$ or by their angle $\theta = \arctan\left(\frac{x_2}{x_1}\right) \in (-\frac{\pi}{2}, \frac{\pi}{2}]$. The action of a matrix $\boldsymbol{B}$ on an angle θ or a slope m can be written as

$$\boldsymbol{B} \cdot \theta = \arctan \frac{y_2}{y_1}, \qquad \vec{y} = \boldsymbol{B} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$$

$$\boldsymbol{B} \cdot m = \frac{y_2}{y_1}, \qquad \vec{y} = \boldsymbol{B} \begin{pmatrix} 1 \\ m \end{pmatrix}$$

We call θ or m a point of discontinuity of $\boldsymbol{B}$, if $y_1 = 0$.

For numerical calculations, it is easier to work with a bounded interval, so we use the second option and partition the angular domain into K intervals:

$$(-\frac{\pi}{2}, \frac{\pi}{2}] = \bigcup_{k=1}^{K} I_k.$$

We then make the following ansatz:

$$\begin{aligned}\nu(J) &= \sum_{k=1}^{K} \nu\left(I_k \cap J\right) \\ &\approx \sum_{k=1}^{K} \nu_k \frac{l\left(I_k \cap J\right)}{l\left(I_k\right)}\end{aligned}$$

where $l(J)$ is the length of the interval J. We plug the ansatz into the invariance equation to obtain the set of equations

$$\begin{aligned}\nu_m = \nu(I_m) &= \sum_{k=1}^{K} \nu_k \int d\mu(\boldsymbol{A}) \frac{l\left(I_k \cap \boldsymbol{A} I_m\right)}{l\left(I_k\right)} \\ &=: \sum_{k=1}^{K} \nu_k B_{m,k}.\end{aligned}$$

Here, we write $\boldsymbol{A}I_k$ to designate the image of the angular interval I_k under the action of $\boldsymbol{A}$.

Supplemented with the condition

$$1 = \sum_{k=1}^{K} \nu_k$$

there exists a unique solution vector ν_k of the resulting system

$$\begin{pmatrix} \boldsymbol{B} - \boldsymbol{I} \\ \mathbf{1} \end{pmatrix} \nu = \begin{pmatrix} \vec{0} \\ 1 \end{pmatrix}.$$

It can be solved by either replacing one of the first K equations by the last one and then solving a square system or as a least squares problem.

Now, γ can be approximated as

$$\gamma \approx \sum_{j=1}^{K} \nu_k \int d\mu(\boldsymbol{A}) \log \left\| \boldsymbol{A} x_k \right\|.$$

We chose the evaluation point $x_k \in I_k$ to be the midpoint.

It remains to pick a suitable partition $\mathcal{I} = \{I_k\}$ of the angular domain. It is convenient to choose the points of discontinuity for all matrices in $\mathcal{A}$ as boundaries for

the intervals in order to simplify the calculation of $l\left(I_k \cap \boldsymbol{A} I_m\right)$. Embree and Trefethen [49] used uniformly spaced nodes for the Fibonacci problem. (This works, because the discontinuities are $\pm\frac{\pi}{4}$.) In order to get precision of 6 digits, they had to pick 2^{20} nodes, resulting in a matrix with 10^7 non-zero entries. Instead, we propose the following strategy of choosing a partition of $(-\frac{\pi}{2}, \frac{\pi}{2}]$:

Algorithm 4 Invariant distribution algorithm with refinement

Input: Number of refinement steps N, refinement threshold θ, number of subdivisions k

1: **for** $i \leftarrow 1$ **to** N **do**
2: $\quad \mathcal{I} \leftarrow$ refine$(\mathcal{I}, \nu; \theta, k)$
3: $\quad \nu \leftarrow$ FindInvariantDistribution$(\mathcal{I})$
4: $\quad \gamma \leftarrow$ CalculateLyapunovExponent$(\mathcal{I}, \nu)$

This strategy does worst, when the invariant distribution puts similar weight on the whole angular domain, and best if it has localized support. The method can obviously only be applied to low dimensional problems, as it requires the discretization of the hyper-cube $\left(-\frac{-\pi}{2}, \frac{-\pi}{2}\right)^{d-1}$.

3.4 Resampled Monte-Carlo Method

The generalized Lyapunov exponents $L(\alpha)$ could in principle be calculated just as the Lyapunov exponent γ using brute-force Monte Carlo simulations:

$$\begin{aligned} L(\alpha) &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left[\|\boldsymbol{X_N}\|^\alpha\right] \\ &\approx \frac{1}{N} \log\left[\frac{1}{K}\sum_{k=1}^{K} \left\|X_N^{(k)}\right\|^\alpha\right], \end{aligned}$$

where $X_N^{(k)}$ are independent realizations of $\boldsymbol{X_N} = \prod_{j=1}^{N} \boldsymbol{\mathcal{A}}_j x_0$. Vanneste[97] remarked that because the generalized Lyapunov exponents are associated with a class of sequences that display a growth behavior different from γ, it would be advantageous to give those rare sequences a higher weight. He proposed a resampling

strategy that selects exactly those realizations. At every step of the simulation, we calculate as before

$$\bar{X}_j^{(k)} := \frac{\boldsymbol{A}_j^{(k)} \tilde{X}_{j-1}^{(k)}}{\left\| \boldsymbol{A}_j^{(k)} \tilde{X}_{j-1}^{(k)} \right\|},$$
$$\alpha_j^{(k)} := \left\| \boldsymbol{A}_j^{(k)} \tilde{X}_{j-1}^{(k)} \right\|^{\alpha}.$$

But then we resample, i.e. we set

$$\beta_j = \sum_{k=1}^{K} \alpha_j^{(k)}$$

and pick J_k for $k = 1, \ldots, K$ in $\{1, \ldots, K\}$ with probabilities $\frac{\alpha_j^{(k)}}{\beta_j}$. Then we assign

$$\tilde{X}_j^{(k)} = \bar{X}_{J_k}^{(k)}.$$

Finally, we approximate

$$L(\alpha) \approx \frac{1}{N} \sum_{j=1}^{N} \beta_j - \log K.$$

Vanneste also shows that for an accurate estimate of L, we now only need $K \gg N$ and not, as in the brute-force case, $K \gg e^{cN}$ with $c > 0$.

3.5 Replica Trick

Just as the resampled Monte-Carlo method, the Replica Trick can be used to calculate the generalized Lyapunov exponent. It's attraction stems from the fact that the mean is taken directly of a random matrix, not of a nonlinear function. Hence the method is well suited for linear perturbations. The price to pay is that not all exponents $L(\boldsymbol{\mathcal{A}}, \alpha)$, $\alpha \in \mathbb{R}$ can be calculated, but only $\alpha \in 2\mathbb{N}_+$.

Definition 5.

Let $\boldsymbol{A} \in \mathbb{R}^{d \times e}$. The vectorization *of $\boldsymbol{A}$ by stacking its columns is defined as*

$$\operatorname{vec}(\boldsymbol{A})_i := \boldsymbol{A}_{(i-1 \operatorname{div} d)+1, (i-1 \bmod e)+1}$$

for $1 \leq i \leq de$.

Definition 6.

Let $\boldsymbol{Z},\ \boldsymbol{Y} \in \mathbb{R}^{d\times e}$, $\boldsymbol{W} \in \mathbb{R}^{f\times g}$. *Then the* Kronecker product $\boldsymbol{Z} \otimes \boldsymbol{W}$ *of* $\boldsymbol{Z}$ *and* $\boldsymbol{W}$ *is defined as*

$$(\boldsymbol{Z} \otimes \boldsymbol{W})_{i,j} := \boldsymbol{Z}_{(i-1 \text{ div } d)+1,(j-1 \text{ div } e)+1}\boldsymbol{W}_{(i-1 \bmod f)+1,(j-1 \bmod g)+1}$$

for $1 \leq i \leq df$ *and* $1 \leq j \leq eg$. *We also define the* k*-th* Kronecker power $\boldsymbol{Z}^{\otimes k}$ *of* $\boldsymbol{Z}$ *as*

$$\boldsymbol{Z}^{\otimes k} := \underbrace{\boldsymbol{Z} \otimes \boldsymbol{Z} \otimes \cdots \otimes \boldsymbol{Z}}_{k \text{ times}}.$$

Lemma 7.

Let $\boldsymbol{A},\ \boldsymbol{B},\ \boldsymbol{C} \in \mathbb{R}^{d\times d}$ *and* $\boldsymbol{I}$ *the identity in* $\mathbb{R}^d$. *We have the following identities:*

- $\operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B})$,
- $\operatorname{tr}(\boldsymbol{A}) = \operatorname{vec}(\boldsymbol{I}) \cdot \operatorname{vec}(\boldsymbol{A})$,
- $(\boldsymbol{A} \otimes \boldsymbol{C}) \operatorname{vec}(\boldsymbol{B}) = \operatorname{vec}\left(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}^T\right)$.

Lemma 8 (Replica trick)**.**

Let $\boldsymbol{\mathcal{A}}$ *be a random square matrix. Then*

$$L(\boldsymbol{\mathcal{A}}, 2k) = \log \rho\left(\mathbb{E}\left[\boldsymbol{\mathcal{A}}^{\otimes 2k}\right]\right)$$

for $k \in \mathbb{N}$.

Proof. Let $\boldsymbol{A}$ a nonrandom square matrix. We write using the above identities

$$\begin{aligned}
\|\boldsymbol{A}\|_F^{2k} &= \operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)^k \\
&= \operatorname{tr}\left((\boldsymbol{A}\boldsymbol{A}^T)^{\otimes k}\right) \\
&= \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \cdot \operatorname{vec}\left((\boldsymbol{A}\boldsymbol{A}^T)^{\otimes k}\right) \\
&= \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \cdot \operatorname{vec}\left(\boldsymbol{A}^{\otimes k}\boldsymbol{I}^{\otimes k}(\boldsymbol{A}^T)^{\otimes k}\right)
\end{aligned}$$

$$= \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \cdot \boldsymbol{A}^{\otimes 2k} \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right).$$

Hence, if we call $\lambda_{j,\otimes 2k}$ and $v_{j,\otimes 2k}$ the eigenvalues and eigenvectors of $\mathbb{E}\left[\boldsymbol{\mathcal{A}}^{\otimes 2k}\right]$, sorted in descending order with respect to their absolute value, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{\mathcal{A}}^{N}\right\|_{F}^{2k}\right] &= \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \cdot \mathbb{E}\left[\left(\boldsymbol{\mathcal{A}}^{\otimes 2k}\right)^{N}\right] \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \\
&= \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \cdot \mathbb{E}\left[\boldsymbol{\mathcal{A}}^{\otimes 2k}\right]^{N} \operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \\
&= \left[\operatorname{vec}\left(\boldsymbol{I}^{\otimes k}\right) \cdot v_{1,\otimes 2k}\right]^{2} \lambda_{1,\otimes 2k}^{N} + \mathcal{O}\left(\lambda_{2,\otimes 2k}^{N}\right).
\end{aligned}$$

Here, we assumed without loss of generality that $|\lambda_{1,\otimes 2k}| > |\lambda_{2,\otimes 2k}|$. Otherwise, we would only have a different constant multiplying the leading order term. Therefore, after taking the logarithm, dividing by N and letting $N \to \infty$, we obtain

$$L(\boldsymbol{\mathcal{A}}, 2k) = \log \lambda_{1,\otimes 2k}.$$

Moreover, we also see that $\left(\frac{\lambda_{2,\otimes 2k}}{\lambda_{1,\otimes 2k}}\right)^{N}$ can be used to estimate the finite N variations of the generalized Lyapunov exponents. □

The determination of the Lyapunov spectral radius is in general a hard problem. In what follows, we will limit ourselves to finding approximations numerically, as presented in Section 3.2, or estimates using the generalized Lyapunov exponent and the Replica Trick. While we have shown that the Lyapunov spectral radius can be bounded by

$$\varrho\left(\boldsymbol{\mathcal{A}}\right) \leq \varrho_2\left(\boldsymbol{\mathcal{A}}\right) \leq \sqrt{\rho\left(\mathbb{E}\left[\boldsymbol{\mathcal{A}}^{\otimes 2}\right]\right)},$$

the sharpness of the estimate is not always known and depends on the random matrix $\boldsymbol{\mathcal{A}}$, as shown in Example 3. Therefore, in what follows, we complement analytic convergence estimates for the Lyapunov spectral radius with Monte-Carlo simulations to determine if the predicted behavior matches the observed convergence rates of fault-prone methods.

Chapter 4

Application of the Replica Trick to a Simple Iteration

In this section, we will turn to the application of the Replica Trick to the analysis of a fault-prone linear iterative method. Before we discuss the more complicated cases of the Fault-Prone Two Grid and Multigrid Methods, we illustrate the approach by limiting ourselves to the case of a fault-prone smoothener.

The iteration matrix for the fault-prone smoothener eq. (2.6) is given by

$$\boldsymbol{\mathcal{E}}_\ell^S = \boldsymbol{I} - \boldsymbol{\mathcal{X}}_\ell^{(S)} \boldsymbol{N}_\ell \boldsymbol{A}_\ell,$$

Here, we take $\boldsymbol{\mathcal{X}}_\ell^{(S)} = \mathrm{diag}(\chi_1, \ldots, \chi_{n_\ell})$ to be a random diagonal matrix of componentwise faults with $\mathbb{E}(\chi_j) = 1 - \varepsilon$ and $\mathrm{Cov}(\chi_j, \chi_j) = \varepsilon(1-\varepsilon)\delta_{ij}$. The behavior of the fault-prone smoothener is governed by the Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}_\ell^S)$. Our goal is to determine under which conditions the smoothener will stay convergent, i.e. $\varrho(\boldsymbol{\mathcal{E}}_\ell^S) < 1$.

The following Lemma permits us to expand the second moment of a fault-prone iteration matrix in terms of expectations and variances of the fault matrices:

Lemma 9.

Let $\boldsymbol{\mathcal{B}}_i \in \mathbb{R}^{n_i \times n_{i+1}}$, $i = 1, \ldots, k$, be independent random matrices with $n_1 = n_{k+1}$. Then

$$\mathbb{E}\left[(\boldsymbol{I} - \boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k)^{\otimes 2}\right] = \mathbb{E}\left[I - \boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k\right]^{\otimes 2} + \sum_{\boldsymbol{E} \in S} \boldsymbol{E}_1 \cdots \boldsymbol{E}_k,$$

where

$$S = \left[\bigotimes_{j=1}^{k} \left\{\mathbb{E}\left[\boldsymbol{\mathcal{B}}_j\right]^{\otimes 2},\ \mathbb{V}\left[\boldsymbol{\mathcal{B}}_j\right]\right\}\right] \setminus \left\{\left(\mathbb{E}\left[\boldsymbol{\mathcal{B}}_1\right]^{\otimes 2}, \ldots, \mathbb{E}\left[\boldsymbol{\mathcal{B}}_k\right]^{\otimes 2}\right)\right\}.$$

Proof. We multiply out to find

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{I} - \boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k)^{\otimes 2}\right] &= \mathbb{E}\left[\boldsymbol{I}^{\otimes 2} - \boldsymbol{I} \otimes (\boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k) - (\boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k) \otimes \boldsymbol{I} + (\boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k)^{\otimes 2}\right] \\
&= \boldsymbol{I}^{\otimes 2} - \boldsymbol{I} \otimes \mathbb{E}\left[\boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k\right] - \mathbb{E}\left[\boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k\right] \otimes \boldsymbol{I} + \mathbb{E}\left[(\boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k)^{\otimes 2}\right] \\
&= \mathbb{E}\left[\boldsymbol{I} - \boldsymbol{\mathcal{B}}_1 \cdots \boldsymbol{\mathcal{B}}_k\right]^{\otimes 2} - \mathbb{E}\left[\boldsymbol{\mathcal{B}}_1\right]^{\otimes 2} \cdots \mathbb{E}\left[\boldsymbol{\mathcal{B}}_k\right]^{\otimes 2} + \mathbb{E}\left[\boldsymbol{\mathcal{B}}_1^{\otimes 2}\right] \cdots \mathbb{E}\left[\boldsymbol{\mathcal{B}}_k^{\otimes 2}\right].
\end{aligned}$$

In the last step, we used the independence of the random matrices. Now, because

$$\mathbb{E}\left[\boldsymbol{\mathcal{B}}_j^{\otimes 2}\right] = \mathbb{V}\left[\boldsymbol{\mathcal{B}}_j\right] + \mathbb{E}\left[\boldsymbol{\mathcal{B}}_j\right]^{\otimes 2},$$

we can combine the last two terms in the expression above to obtain the desired result. □

Theorem 10.

Let $\boldsymbol{E}_\ell^S = \boldsymbol{I} - \boldsymbol{N}_\ell \boldsymbol{A}_\ell$ be the iteration matrix for any convergent smoothener. Then, the corresponding fault-prone smoothener satisfies

$$\varrho(\boldsymbol{\mathcal{E}}_\ell^S) \leq \sqrt{(1-\varepsilon)\left\|\boldsymbol{E}_\ell^S\right\|_2^2 + \varepsilon\left(\left\|\boldsymbol{E}_\ell^S\right\|_2 + \left\|\boldsymbol{N}_\ell \boldsymbol{A}_\ell\right\|_2\right)^2},$$

for all $\varepsilon \in [0, 1]$.

Proof. Let $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{X}}^{(S)} \boldsymbol{N} \boldsymbol{A}$ and dispense with subscripts and superscripts for the duration of the proof. Then

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{\otimes 2}\right] &= \mathbb{E}\left[\boldsymbol{I}^{\otimes 2} - \boldsymbol{\mathcal{B}} \otimes \boldsymbol{I} - \boldsymbol{I} \otimes \boldsymbol{\mathcal{B}} + \boldsymbol{\mathcal{B}}^{\otimes 2}\right] \\
&= \boldsymbol{I}^{\otimes 2} - \mathbb{E}\left[\boldsymbol{\mathcal{B}}\right] \otimes \boldsymbol{I} - \boldsymbol{I} \otimes \mathbb{E}\left[\boldsymbol{\mathcal{B}}\right] + \mathbb{E}\left[\boldsymbol{\mathcal{B}}\right]^{\otimes 2}
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}\left[\boldsymbol{\mathcal{B}}\right]^{\otimes 2} + \mathbb{E}\left[\boldsymbol{\mathcal{B}}^{\otimes 2}\right] \\
&= \mathbb{E}\left[\boldsymbol{\mathcal{E}}\right]^{\otimes 2} + \mathbb{V}\left[\boldsymbol{\mathcal{B}}\right] \qquad (4.1)
\end{aligned}
$$

and

$$
\mathbb{V}\left[\boldsymbol{\mathcal{B}}\right] = \mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(S)}\right] (\boldsymbol{N}\boldsymbol{A})^{\otimes 2} = \varepsilon(1-\varepsilon)\boldsymbol{K}\,(\boldsymbol{N}\boldsymbol{A})^{\otimes 2}\,,
$$

where

$$
\boldsymbol{K} = \mathrm{blockdiag}\left(\vec{e}^{(j)} \otimes \left(\vec{e}^{(j)}\right)^T,\ j = 1, \ldots, n\right)
$$

and $\vec{e}^{(j)}$ the j-th canonical unit vector in $\mathbb{R}^n$. We will derive a more general form of eq. (4.1) in Lemma 9 in the appendix. Since $\left\|\boldsymbol{C}^{\otimes 2}\right\|_2 = \|\boldsymbol{C}\|_2^2$ and $\|\boldsymbol{K}_\ell \boldsymbol{C}\|_2 \leq \|\boldsymbol{C}\|_2$ for any compatible matrix $\boldsymbol{C}$, we obtain

$$
\left\|\mathbb{V}\left[\boldsymbol{\mathcal{B}}\right]\right\|_2 = \varepsilon(1-\varepsilon)\left\|\boldsymbol{K}\,(\boldsymbol{N}\boldsymbol{A})^{\otimes 2}\right\|_2 \leq \varepsilon(1-\varepsilon)\left\|\boldsymbol{N}\boldsymbol{A}\right\|_2^2
$$

and

$$
\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}\right]^{\otimes 2}\right\|_2 = \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}\right]\right\|_2^2 = \left\|\boldsymbol{I} - (1-\varepsilon)\boldsymbol{N}\boldsymbol{A}\right\|_2^2.
$$

Consequently,

$$
\begin{aligned}
\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{\otimes 2}\right]\right\|_2 &\leq \left\|\boldsymbol{I} - (1-\varepsilon)\boldsymbol{N}\boldsymbol{A}\right\|_2^2 + \varepsilon(1-\varepsilon)\left\|\boldsymbol{N}\boldsymbol{A}\right\|_2^2 \\
&= \left\|\boldsymbol{E} + \varepsilon\boldsymbol{N}\boldsymbol{A}\right\|_2^2 + \varepsilon(1-\varepsilon)\left\|\boldsymbol{N}\boldsymbol{A}\right\|_2^2 \\
&\leq \left(\left\|\boldsymbol{E}\right\|_2 + \varepsilon\left\|\boldsymbol{N}\boldsymbol{A}\right\|_2\right)^2 + \varepsilon(1-\varepsilon)\left\|\boldsymbol{N}\boldsymbol{A}\right\|_2^2 \\
&= (1-\varepsilon)\left\|\boldsymbol{E}\right\|_2^2 + \varepsilon(\left\|\boldsymbol{E}\right\|_2 + \left\|\boldsymbol{N}\boldsymbol{A}\right\|_2)^2
\end{aligned}
$$

and the result follows thanks to Lemma 8. □

In order to illustrate Theorem 10, we consider the situation where the smoothener is taken to be a convergent relaxation scheme so that $\left\|\boldsymbol{E}_\ell^S\right\|_2 = \Gamma \in [0, 1)$. The triangle inequality gives $\|\boldsymbol{N}_\ell \boldsymbol{A}_\ell\|_2 \leq 1 + \Gamma$ and hence, thanks to Theorem 10, we deduce that

$$
\varrho(\boldsymbol{\mathcal{E}}_\ell^S) \leq \sqrt{\Gamma^2 + \varepsilon(1 + 4\Gamma + 3\Gamma^2)} = \Gamma + \mathcal{O}\left(\varepsilon\right).
$$

As a consequence, we find that a sufficient condition for the fault-prone smoothener to remain convergent is that the fault rate is sufficiently small: $\varepsilon < (1-\Gamma)/(1+3\Gamma)$.

Chapter 5

Analysis of the Fault-Prone Two Grid Method

We now turn to the analysis of the Fault-Prone Two Grid Method as given in Algorithm 3. We use subscripts C for coarse-level quantities, and F for fine-level quantities.

The convergence analysis will be carried out with respect to the energy norm:

Definition 11 (Energy norms)**.**
For matrices $\boldsymbol{Z} \in \mathbb{R}^{n_\ell \times n_\ell}$, we define the usual matrix energy norm $\|\boldsymbol{Z}\|_A$ as well as the double energy norm $\|\boldsymbol{Z}\|_{A^2}$ to be $\|\boldsymbol{Z}\|_A = \left\|\boldsymbol{A}_\ell^{\frac{1}{2}} \boldsymbol{Z} \boldsymbol{A}_\ell^{-\frac{1}{2}}\right\|_2$, and $\|\boldsymbol{Z}\|_{A^2} = \left\|\boldsymbol{A}_\ell \boldsymbol{Z} \boldsymbol{A}_\ell^{-1}\right\|_2$. For matrices $\boldsymbol{W} \in \mathbb{R}^{n_\ell^2 \times n_\ell^2}$, we define the tensor energy norm $\|\boldsymbol{W}\|_A$ and the tensor double energy norm $\|\boldsymbol{W}\|_{A^2}$ to be $\|\boldsymbol{W}\|_A = \left\|\left(\boldsymbol{A}_\ell^{\frac{1}{2}}\right)^{\otimes 2} \boldsymbol{Z} \left(\boldsymbol{A}_\ell^{-\frac{1}{2}}\right)^{\otimes 2}\right\|_2$, and $\|\boldsymbol{W}\|_{A^2} = \left\|\boldsymbol{A}_\ell^{\otimes 2} \boldsymbol{Z} \left(\boldsymbol{A}_\ell^{-1}\right)^{\otimes 2}\right\|_2$. In all cases $\|\bullet\|_2$ is the spectral norm.

The first main result of the present work concerns the convergence of the Fault-Prone Two-Grid Algorithm for the special case of all $\boldsymbol{\mathcal{X}}_\bullet^{(\bullet)}$ having independent Bernoulli distributed diagonal entries with probability of fault ε:

Theorem 12.
Let $\boldsymbol{\mathcal{E}}^{TG} = \left(\boldsymbol{\mathcal{E}}^{S,post}\right)^{\nu_2} \boldsymbol{\mathcal{E}}^{CG} \left(\boldsymbol{\mathcal{E}}^{S,pre}\right)^{\nu_1}$ be the iteration matrix of the Fault-Prone Two Grid

Method (Algorithm 3) with Jacobi smoothener and component-wise faults of rate ε in prolongation, restriction, residual and in the smoothener, and let $\boldsymbol{E}^{TG}$ be its fault-free equivalent. Assume that $\Omega \in C^2$ or Ω a convex polyhedron, $\mathcal{T}_C$ quasiuniform and that Assumptions 1 to 4 hold. Then

$$\varrho\left(\boldsymbol{\mathcal{E}}^{TG}\right) \leq \left\|\boldsymbol{E}^{TG}\right\|_A + C \begin{cases} \varepsilon n_F^{(4-d)/2d} & d < 4, \\ \varepsilon\sqrt{\log n_F} & d = 4, \\ \varepsilon & d > 4, \end{cases} \tag{5.1}$$

where $\|\bullet\|_A$ is the matrix energy norm and the constant C is independent of ε and n_F.

Before we turn to the proof of Theorem 12, we record several useful definition and results.

Definition 13.
Let $\boldsymbol{Z},\ \boldsymbol{Y} \in \mathbb{R}^{n\times m}$, $n,\ m \in \mathbb{N}$. The element wise *or* Hadamard product $\boldsymbol{Z} \circ \boldsymbol{Y} \in \mathbb{R}^{n\times m}$ *of $\boldsymbol{Z}$ and $\boldsymbol{Y}$ is defined as*

$$(\boldsymbol{Z} \circ \boldsymbol{Y})_{ij} = \boldsymbol{Z}_{ij}\boldsymbol{Y}_{ij}$$

for $1 \leq i \leq n$, $1 \leq j \leq m$, and the k-th Hadamard power $\boldsymbol{Z}^{\circ k} \in \mathbb{R}^{n\times m}$ *of $\boldsymbol{Z}$ as*

$$\boldsymbol{Z}^{\circ k} := \underbrace{\boldsymbol{Z} \circ \boldsymbol{Z} \circ \cdots \circ \boldsymbol{Z}}_{k \text{ times}}.$$

Lemma 14 (Horn and Johnson [66])**.**
Let $\boldsymbol{H} = \left(\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{smallmatrix}\right) \in \mathbb{R}^{(p+q)\times(p+q)}$ be symmetric, $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\boldsymbol{B} \in \mathbb{R}^{p\times q}$, $\boldsymbol{C} \in \mathbb{R}^{q\times q}$. Then $\boldsymbol{H}$ is positive (semi-) definite if and only if $\boldsymbol{A}$ and $\boldsymbol{C}$ are positive (semi-) definite and there exists a contraction $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ (i.e. all singular values are smaller or equal than 1) such that $\boldsymbol{B} = \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{C}^{\frac{1}{2}}$.

Lemma 15 (Horn and Johnson [67])**.**
Let $\boldsymbol{Z}$, $\boldsymbol{Y}$ be symmetric positive (semi-) definite matrices. Then $\boldsymbol{Z} \circ \boldsymbol{Y}$ is symmetric positive (semi-) definite.

We now prove the following estimates for the Hadamard power of matrices:

Lemma 16.
Let $\boldsymbol{Z} \in \mathbb{R}^{n\times m}$. Then

$$\left\|\boldsymbol{Z}^{\circ 2}\right\|_2 \leq \max_{i=1,\dots,m} \left\|\boldsymbol{Z}\vec{e}_m^{(i)}\right\|_2 \max_{j=1,\dots,n} \left\|\boldsymbol{Z}^T\vec{e}_n^{(j)}\right\|_2 \leq \|\boldsymbol{Z}\|_2^2\,,$$

where $\left\{\vec{e}_m^{(i)}\right\}$ and $\left\{\vec{e}_n^{(j)}\right\}$ are the canonical unit basis vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively.

Proof. Both the matrix $\begin{pmatrix} \boldsymbol{Z} & \boldsymbol{I}_n \end{pmatrix}^T \begin{pmatrix} \boldsymbol{Z} & \boldsymbol{I}_n \end{pmatrix}$ and $\begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{Z}^T \end{pmatrix}^T \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{Z}^T \end{pmatrix}$ are symmetric positive semi-definite. Using Lemma 15, their Hadamard product

$$\begin{pmatrix} \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)\circ\boldsymbol{I}_m & \left(\boldsymbol{Z}^T\right)^{\circ 2} \\ \boldsymbol{Z}^{\circ 2} & \boldsymbol{I}_n \circ \left(\boldsymbol{Z}\boldsymbol{Z}^T\right) \end{pmatrix}$$

is also positive semi-definite. Hence, by Lemma 14,

$$\left\|\boldsymbol{Z}^{\circ 2}\right\|_2^2 \leq \left\|\left(\boldsymbol{Z}\boldsymbol{Z}^T\right)\circ\boldsymbol{I}_n\right\|_2 \left\|\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)\circ\boldsymbol{I}_m\right\|_2 = \max_i \left\|\boldsymbol{Z}\vec{e}_m^{(i)}\right\|_2^2 \max_i \left\|\boldsymbol{Z}^T\vec{e}_n^{(j)}\right\|_2^2.$$

The second inequality follows trivially. □

Proof of Theorem 12mg. We set

$$\mathbb{E}\left[\boldsymbol{\mathcal{X}}^{(\bullet)}\right] = e^{(\bullet)}\boldsymbol{I}, \qquad\qquad \mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\bullet)}\right] = \boldsymbol{V}^{(\bullet)}.$$

Using Lemma 9 the second moment of the fault-prone coarse grid correction and smoothener can be written as

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{CG}\right)^{\otimes 2}\right] &= \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right]^{\otimes 2} + \boldsymbol{C}^{(P)} + \boldsymbol{C}^{(R)} + \boldsymbol{C}^{(\rho)} \\ &\quad + \boldsymbol{C}^{(P,R)} + \boldsymbol{C}^{(P,\rho)} + \boldsymbol{C}^{(R,\rho)} + \boldsymbol{C}^{(P,R,\rho)}, \end{aligned} \tag{5.2}$$

$$\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{S}\right)^{\otimes 2}\right] = \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]^{\otimes 2} + \boldsymbol{C}^{(S)}, \tag{5.3}$$

with

$$\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right] = \boldsymbol{E}^{CG} + \left(1 - e^{(P)}e^{(R)}e^{(\rho)}\right)\left(\boldsymbol{I} - \boldsymbol{E}^{CG}\right),$$

$$
\begin{aligned}
\boldsymbol{C}^{(P)} &= \left(e^{(R)}e^{(\rho)}\right)^2 \boldsymbol{V}^{(P)} \left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}_F\right)^{\otimes 2},\\
\boldsymbol{C}^{(R)} &= \left(e^{(P)}e^{(\rho)}\right)^2 \left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\right)^{\otimes 2} \boldsymbol{V}^{(R)} \left(\boldsymbol{R}\boldsymbol{A}_F\right)^{\otimes 2},\\
\boldsymbol{C}^{(\rho)} &= \left(e^{(P)}e^{(R)}\right)^2 \left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\right)^{\otimes 2} \boldsymbol{V}^{(\rho)} \boldsymbol{A}_F^{\otimes 2},\\
\boldsymbol{C}^{(P,R)} &= \left(e^{(\rho)}\right)^2 \boldsymbol{V}^{(P)} \left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\right)^{\otimes 2} \boldsymbol{V}^{(R)} \left(\boldsymbol{R}\boldsymbol{A}_F^{\otimes 2}\right)^{\otimes 2},\\
\boldsymbol{C}^{(P,\rho)} &= \left(e^{(R)}\right)^2 \boldsymbol{V}^{(P)} \left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\right)^{\otimes 2} \boldsymbol{V}^{(\rho)} \boldsymbol{A}_F^{\otimes 2},\\
\boldsymbol{C}^{(R,\rho)} &= \left(e^{(P)}\right)^2 \left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\right)^{\otimes 2} \boldsymbol{V}^{(R)} \boldsymbol{R}^{\otimes 2} \boldsymbol{V}^{(\rho)} \boldsymbol{A}_F^{\otimes 2},\\
\boldsymbol{C}^{(P,R,\rho)} &= \boldsymbol{V}^{(P)} \left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\right)^{\otimes 2} \boldsymbol{V}^{(R)} \boldsymbol{R}^{\otimes 2} \boldsymbol{V}^{(\rho)} \boldsymbol{A}_F^{\otimes 2},\\
\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right] &= \boldsymbol{E}^S + e^{(S)}\left(\boldsymbol{I} - \boldsymbol{E}^S\right),\\
\boldsymbol{C}^{(S)} &= \boldsymbol{V}^{(S)} \left(\boldsymbol{N}\boldsymbol{A}_F\right)^{\otimes 2}.
\end{aligned}
$$

We have that

$$
e^{(\bullet)} = 1-\varepsilon, \qquad \boldsymbol{V}^{(P)} = \boldsymbol{V}^{(\rho)} = \boldsymbol{V}^{(S)} = \varepsilon(1-\varepsilon)\boldsymbol{K}_F, \qquad \boldsymbol{V}^{(R)} = \varepsilon(1-\varepsilon)\boldsymbol{K}_{\boldsymbol{C}},
$$

with

$$
\begin{aligned}
\boldsymbol{K} &= \text{blockdiag}\left(\vec{e}_F^{(i)} \otimes \left(\vec{e}_F^{(i)}\right)^T,\ i = 1, \ldots, n_F\right),\\
\boldsymbol{K}_C &= \text{blockdiag}\left(\vec{e}_C^{(i)} \otimes \left(\vec{e}_C^{(i)}\right)^T,\ i = 1, \ldots, n_C\right).
\end{aligned}
$$

Here, $\vec{e}_F^{(i)}$ and $\vec{e}_C^{(i)}$ are the canonical unit basis vectors of $\mathbb{R}^{n_F}$ and $\mathbb{R}^{n_C}$ respectively. Adding and subtracting $\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\right]^{\otimes 2}$ from $\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\right)^{\otimes 2}\right]$, we estimate using the energy norm on the tensor space

$$
\begin{aligned}
\rho\left(\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\right)^{\otimes 2}\right]\right) &\leq \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\right)^{\otimes 2}\right]\right\|_A\\
&\leq \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\right]\right\|_A^2 + \left\|\mathbb{E}\left[\left(\left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2} \boldsymbol{\mathcal{E}}^{CG} \left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1}\right)^{\otimes 2}\right]\right\|_A\\
&\quad - \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2} \boldsymbol{\mathcal{E}}^{CG} \left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1}\right]\right\|_A^2.
\end{aligned}
$$

We then use the subadditivity of $\|\bullet\|_A$ and equations eq. (5.2) and eq. (5.3) to write

$$
\begin{aligned}
\rho\left(\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\right)^{\otimes 2}\right]\right) \leq\ & \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\right]\right\|_A^2\\
&+ \Big(\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right]\right\|_A^2 + \left\|\boldsymbol{C}^{(P)}\right\|_A + \left\|\boldsymbol{C}^{(R)}\right\|_A + \left\|\boldsymbol{C}^{(\rho)}\right\|_A\\
&+ \left\|\boldsymbol{C}^{(P,R)}\right\|_A + \left\|\boldsymbol{C}^{(P,\rho)}\right\|_A + \left\|\boldsymbol{C}^{(R,\rho)}\right\|_A + \left\|\boldsymbol{C}^{(P,R,\rho)}\right\|_A \Big)
\end{aligned}
$$

$$\times \left(\left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right] \right\|_A^2 + \left\| \boldsymbol{C}^{(S)} \right\|_A \right)^{\nu_1+\nu_2} - \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right] \right\|_A^2 \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right] \right\|_A^{2(\nu_1+\nu_2)}. \tag{5.4}$$

First, we estimate the terms involving only first moments of the fault matrices. Using that $\boldsymbol{E}^{CG}$ is an A-orthogonal projection and that the damped Jacobi smoothener is convergent, we find

$$\begin{aligned} \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right] \right\|_A &\leq \left\| \boldsymbol{E}^{CG} \right\|_A + \left(1-(1-\varepsilon)^3\right) \left\| \boldsymbol{I} - \boldsymbol{E}^{CG} \right\|_A \leq 1 + C\varepsilon, \\ \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right] \right\|_A &\leq \left\| \boldsymbol{E}^{S} \right\|_A + \varepsilon \left\| \boldsymbol{I} - \boldsymbol{E}^{S} \right\|_A \leq 1 + C\varepsilon, \end{aligned}$$

and therefore

$$\begin{aligned} \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\right] \right\|_A &\leq \left\| \boldsymbol{E}^{TG} \right\|_A + \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right] \right\|_A^{\nu_1+\nu_2} \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right] \right\|_A - \left\| \boldsymbol{E}^{S} \right\|_A^{\nu_1+\nu_2} \left\| \boldsymbol{E}^{CG} \right\|_A \\ &= \left\| \boldsymbol{E}^{TG} \right\|_A + C\varepsilon. \end{aligned}$$

Next, we estimate all the terms $\boldsymbol{C}^{(\bullet)}$ involving variances of the fault matrices. We bound

$$\begin{aligned} \frac{1}{\varepsilon} \left\| \boldsymbol{C}^{(\rho)} \right\|_A &\leq \rho \left[\left(\boldsymbol{A}_F^{\frac{1}{2}} \boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{R} \right)^{\otimes 2} \boldsymbol{K}_F \boldsymbol{A}_F^{\otimes 2} \boldsymbol{K}_F \left(\boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{R} \boldsymbol{A}_F^{\frac{1}{2}} \right)^{\otimes 2} \right]^{\frac{1}{2}} \\ &= \rho \left[\left(\boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{R} \boldsymbol{A}_F \boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{R} \right)^{\otimes 2} \boldsymbol{K}_F \boldsymbol{A}_F^{\otimes 2} \boldsymbol{K}_F \right]^{\frac{1}{2}} \\ &= \rho \left[\left(\boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{R} \right)^{\circ 2} \boldsymbol{A}_F^{\circ 2} \right]^{\frac{1}{2}} \\ &\leq \left\| \left(\boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{R} \right)^{\circ 2} \right\|_2^{\frac{1}{2}} \left\| \boldsymbol{A}_F^{\circ 2} \right\|_2^{\frac{1}{2}}. \end{aligned}$$

Here, we used that $\boldsymbol{K}_F = \boldsymbol{K}_F^2$ and that for compatible $\boldsymbol{Z}$

$$\begin{aligned} \left(\boldsymbol{K}_F \boldsymbol{Z}^{\otimes 2} \boldsymbol{K}_F \right)_{in+p,jn+q} &= \left(\boldsymbol{K}_F \right)_{in+p,in+p} \boldsymbol{Z}_{ij} \boldsymbol{Z}_{pq} \left(\boldsymbol{K}_F \right)_{jn+q,jn+q} \\ &= \begin{cases} \left(\boldsymbol{Z}^{\circ 2} \right)_{ij} & \text{if } i = p,\ j = q, \\ 0 & \text{else} \end{cases} \end{aligned}$$

for $1 \leq i, j, p, q \leq n$. Similarly, we find

$$\frac{1}{\varepsilon} \left\| \boldsymbol{C}^{(P)} \right\|_A \leq \left\| \boldsymbol{A}_F^{\circ 2} \right\|_2^{\frac{1}{2}} \left\| \left(\boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{R} \right)^{\circ 2} \right\|_2^{\frac{1}{2}},$$

$$
\begin{aligned}
\frac{1}{\varepsilon}\left\|\boldsymbol{C}^{(R)}\right\|_A &\leq \left\|\boldsymbol{A}_C^{\circ 2}\right\|_2^{\frac{1}{2}}\left\|\left(\boldsymbol{A}_C^{-1}\right)^{\circ 2}\right\|_2^{\frac{1}{2}},\\
\frac{1}{\varepsilon^2}\left\|\boldsymbol{C}^{(P,R)}\right\|_A &\leq \left\|\boldsymbol{A}_F^{\circ 2}\right\|_2^{\frac{1}{2}}\left\|\boldsymbol{A}_C^{\circ 2}\right\|_2^{\frac{1}{2}}\left\|\left(\boldsymbol{A}_C^{-1}\boldsymbol{R}\right)^{\circ 2}\right\|_2,\\
\frac{1}{\varepsilon^2}\left\|\boldsymbol{C}^{(P,\rho)}\right\|_A &\leq \left\|\boldsymbol{A}_F^{\circ 2}\right\|_2\left\|\left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\right)^{\circ 2}\right\|_2,\\
\frac{1}{\varepsilon^2}\left\|\boldsymbol{C}^{(R,\rho)}\right\|_A &\leq \left\|\boldsymbol{P}^{\circ 2}\right\|_2\left\|\boldsymbol{R}^{\circ 2}\right\|_2\left\|\boldsymbol{A}_F^{\circ 2}\right\|_2^{\frac{1}{2}}\left\|\left(\boldsymbol{A}_C^{-1}\right)^{\circ 2}\right\|_2^{\frac{1}{2}},\\
\frac{1}{\varepsilon^3}\left\|\boldsymbol{C}^{(P,R,\rho)}\right\|_A &\leq \left\|\boldsymbol{P}^{\circ 2}\right\|_2^2\left\|\boldsymbol{A}_F^{\circ 2}\right\|_2\left\|\left(\boldsymbol{A}_C^{-1}\boldsymbol{R}\right)^{\circ 2}\right\|_2,\\
\frac{1}{\varepsilon}\left\|\boldsymbol{C}^{(S)}\right\|_A &\leq \left\|\left(\boldsymbol{N}\boldsymbol{A}_F\right)^{\circ 2}\right\|_2.
\end{aligned}
$$

In the last inequality, we used that $\boldsymbol{V}^{(S)}$ and $\boldsymbol{N}^{\otimes 2}$ commute. Using Lemma 16 and Assumptions 2 and 4 we find

$$
\begin{aligned}
\left\|\boldsymbol{P}^{\circ 2}\right\|_2 &\leq \|\boldsymbol{P}\|_2^2 \leq C,\\
\left\|\boldsymbol{A}_F^{\circ 2}\right\|_2 &\leq \|\boldsymbol{A}_F\|_2^2,\\
\left\|\boldsymbol{A_C}^{\circ 2}\right\|_2 &\leq \|\boldsymbol{A_C}\|_2^2 \leq \|\boldsymbol{A}_F\|_2^2,\\
\left\|\left(\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\right)^{\circ 2}\right\|_2 &\leq \max_i \left\|\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\vec{e}_F^{(i)}\right\|_2^2\\
&\leq \left\|\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}_F\right\|_2^2 \max_i \left\|\boldsymbol{A}_F^{-1}\vec{e}_F^{(i)}\right\|_2^2\\
&\leq C\max_i \left\|\boldsymbol{A}_F^{-1}\vec{e}_F^{(i)}\right\|_2^2,\\
\left\|\left(\boldsymbol{A}_C^{-1}\boldsymbol{R}\right)^{\circ 2}\right\|_2 &\leq \max_i \left\|\boldsymbol{P}\boldsymbol{A}_C^{-1}\vec{e}_C^{(i)}\right\|_2 \max_j \left\|\boldsymbol{A}_C^{-1}\boldsymbol{R}\vec{e}_F^{(j)}\right\|_2\\
&\leq C\max_i \left\|\boldsymbol{A}_C^{-1}\vec{e}_C^{(i)}\right\|_2 \left\|\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}_F\right\|_2 \max_j \left\|\boldsymbol{A}_F^{-1}\vec{e}_F^{(j)}\right\|_2\\
&\leq C\max_i \left\|\boldsymbol{A}_C^{-1}\vec{e}_C^{(i)}\right\|_2 \max_j \left\|\boldsymbol{A}_F^{-1}\vec{e}_F^{(j)}\right\|_2,\\
\left\|\left(\boldsymbol{N}\boldsymbol{A}_F\right)^{\circ 2}\right\|_2 &\leq \|\boldsymbol{N}\boldsymbol{A}_F\|_2^2 \leq C.
\end{aligned}
$$

Now, because $\boldsymbol{A}_F$ is the finite element discretization of a second order PDE over a quasi-uniform mesh with mesh size h, we obtain (see Theorem 9.11 in [50]),

$$\|\boldsymbol{A}_F\|_2 \leq Ch^{d-2}.$$

Since

$$\left\| \boldsymbol{A}_F^{-1} \vec{e}_F^{(i)} \right\|_2 \leq C h^{-\frac{d}{2}} \left\| u_F \right\|_{L^2},$$

where $u_F \in V_F$ is given by

$$a\left(u_F, v\right) = v\left(\vec{x}_i\right), \quad \forall v \in V_F,$$

we can apply Lemma 17 to obtain bounds for $\max_j \left\| \boldsymbol{A}_F^{-1} \vec{e}_F^{(j)} \right\|_2$ (and equally for $\max_i \left\| \boldsymbol{A}_C^{-1} \vec{e}_C^{(i)} \right\|_2$).

$$\frac{1}{\varepsilon}\left\|\boldsymbol{C}^{(P)}\right\|_A, \frac{1}{\varepsilon}\left\|\boldsymbol{C}^{(R)}\right\|_A, \frac{1}{\varepsilon}\left\|\boldsymbol{C}^{(\rho)}\right\|_A, \frac{1}{\varepsilon^2}\left\|\boldsymbol{C}^{(R,\rho)}\right\|_A \leq C \begin{cases} h^{\frac{d-4}{2}} & d<4, \\ \sqrt{1+|\log h|} & d=4, \\ 1 & d>4, \end{cases}$$

$$\frac{1}{\varepsilon^2}\left\|\boldsymbol{C}^{(P,R)}\right\|_A, \frac{1}{\varepsilon^2}\left\|\boldsymbol{C}^{(P,\rho)}\right\|_A, \frac{1}{\varepsilon^3}\left\|\boldsymbol{C}^{(P,R,\rho)}\right\|_A \leq C \begin{cases} h^{d-4} & d<4, \\ (1+|\log h|) & d=4, \\ 1 & d>4, \end{cases}$$

$$\frac{1}{\varepsilon}\left\|\boldsymbol{C}^{(S)}\right\|_A \leq C.$$

Therefore, we obtain from eq. (5.4)

$$\rho\left(\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\right)^{\otimes 2}\right]\right) \leq \left\|\boldsymbol{E}^{TG}\right\|_A^2 + C \begin{cases} \varepsilon^2 h^{d-4} & d<4, \\ \varepsilon^2\left(1+|\log h|\right) & d=4, \\ \varepsilon^2 & d>4, \end{cases}$$

and hence

$$\varrho\left(\boldsymbol{\mathcal{E}}^{TG}\right) \leq \left\|\boldsymbol{E}^{TG}\right\|_A + C \begin{cases} \varepsilon n_F^{\frac{4-d}{2d}} & d<4, \\ \varepsilon\sqrt{\log n_F} & d=4, \\ \varepsilon & d>4, \end{cases}$$

where we used that $n_F \approx h^{-d}$. □

In order to conclude, we need the following technical estimate:

Lemma 17.
Let $\Omega \in C^2$ or Ω a convex polyhedron and let $u_F \in V_F$ be the unique solution of

$$a\left(u_F, v\right) = v\left(\vec{x}_i\right), \quad \forall v \in V_F.$$

Then

$$\left\|u_F\right\|_{L^2} \leq C \begin{cases} 1 & d < 4, \\ \left(1 + |\log h|\right)^{\frac{1}{2}} & d = 4, \\ h^{2-\frac{d}{2}} & d > 4, \end{cases}$$

where C is a constant independent of h.

Proof. Let $f \in V_F$ be the unique function that corresponds to the load and write $f = \vec{F} \cdot \phi$ with $\vec{F}$ its coefficient vector and ϕ the vector of shape functions. Then

$$\vec{e}_F^{(i)} = \left(\phi_F, f\right)_{L^2} = \left(\phi_F, \phi_F\right)_{L^2} \cdot \vec{F} = \boldsymbol{M}_F \vec{F}.$$

Since by Theorem 9.8 in [50]

$$Ch^d \boldsymbol{I} \leq \boldsymbol{M}_F \leq Ch^d \boldsymbol{I}, \tag{5.5}$$

we have

$$\left\|f\right\|_{H^m} \leq Ch^{-d} \left\|\phi_F^{(i)}\right\|_{H^m}. \tag{5.6}$$

Now u_F is an approximation to $u \in V$ that solves

$$a(u, v) = \left(f, v\right)_{L^2}, \quad \forall v \in V.$$

Since $f \in V \subset L^2\left(\Omega\right)$, we find by the Aubin-Nitsche Lemma [50] that

$$\left\|u_F - u\right\|_{L^2} \leq Ch^2 \left\|f\right\|_{L^2}.$$

Consider the solution $u^* \in V$ to the dual problem

$$a(v, u^*) = (u, v)_{L^2}, \quad \forall v \in V.$$

By elliptic regularity [50], we have

$$\|u^*\|_{H^2} \leq C \|u\|_{L^2}.$$

Moreover,

$$(u, u)_{L^2} = a(u, u^*) = (f, u^*)_{L^2},$$

so

$$\|u\|_{L^2}^2 \leq \|f\|_{H^{-2}} \|u^*\|_{H^2} \leq C \|f\|_{H^{-2}} \|u\|_{L^2},$$

and hence $\|u\|_{L^2} \leq C \|f\|_{H^{-2}}$. Therefore, by triangle inequality and eq. (5.6), we find

$$\|u_F\|_{L^2} \leq Ch^2 \|f\|_{L^2} + C \|f\|_{H^{-2}} \leq Ch^{-d} \left(h^2 \left\| \phi_F^{(i)} \right\|_{L^2} + \left\| \phi_F^{(i)} \right\|_{H^{-2}} \right).$$

Applying the estimates for $\left\| \phi_F^{(i)} \right\|_{H^m}$ from Theorem 4.8 in [10], we obtain

$$\|u_F\|_{L^2} \leq C \begin{cases} 1 & d < 4, \\ (1 + |\log h|)^{\frac{1}{2}} & d = 4, \\ h^{2-\frac{d}{2}} & d > 4. \end{cases}$$

□

If the probability of a fault occurring is zero, i.e. $\varepsilon = 0$, then the bound eq. (5.1) reduces to the quantity $\left\| \boldsymbol{E}^{TG} \right\|_A$ which governs the convergence of the fault-free Two Grid Algorithm. Under the foregoing assumptions on the finite element discretization, it is known [20, 21, 63, 95] that $\left\| \boldsymbol{E}^{TG} \right\|_A \leq C_{TG} < 1$, where C_{TG} is a positive constant independent of n_F. If $\varepsilon \in (0, 1)$, then the bound eq. (5.1) depends

on the failure probability ε, the number of unknowns n_F and the spatial dimension d of the underlying discretization.

In the simplest case, in which the spatial dimension $d > 4$, the Fault-Prone Two Grid Method will converge uniformly in n_F provided that the failure probability ε is sufficiently small, i.e. $\varepsilon \in [0, (1 - C_{TG})/C) \subset [0, 1)$. The estimate in the cases of most practical interest, in which $d < 4$, is less encouraging with the bound depending on the size n_F of the problem being solved. In particular, the estimate means that no matter how small the probability ε of a failure may be, as the size of the problem being solved grows, the bound will exceed unity suggesting that the Two Grid Method will fail to converge *at all*.

5.1 Numerical Verification of Non-Resilience

In order to illustrate this behavior and to test the sharpness of the estimate eq. (5.1), we apply the Fault-Prone Two Grid Method to the system arising from a discretization of the Poisson problem on a square using piecewise linear functions on a uniform mesh of triangles (Figure 5.2).

We wish to investigate the behavior as the number of degrees of freedom n_F and the fault rate ε are varied. In order to do this, it will be necessary to simulate the faults that might be encountered on an actual system. To verify the scaling law eq. (5.1), we inject faults at a much higher rate than what one would expect to see on an exascale machine in practice. Equally well, the numbers of degrees of freedom n_F that we can simulate is chosen as large as feasible but falls short of what one may expect to achieve on an exascale machine. The laissez-faire mitigation strategy described in Section 2.1 is oblivious of the nature of the fault so long as they are reliably detected. Consequently, our fault simulation procedure does not need to take account of the type of fault. Hence, in order to simulate Algorithm 2, we generate independent realizations for the random matrices involved. In particular,

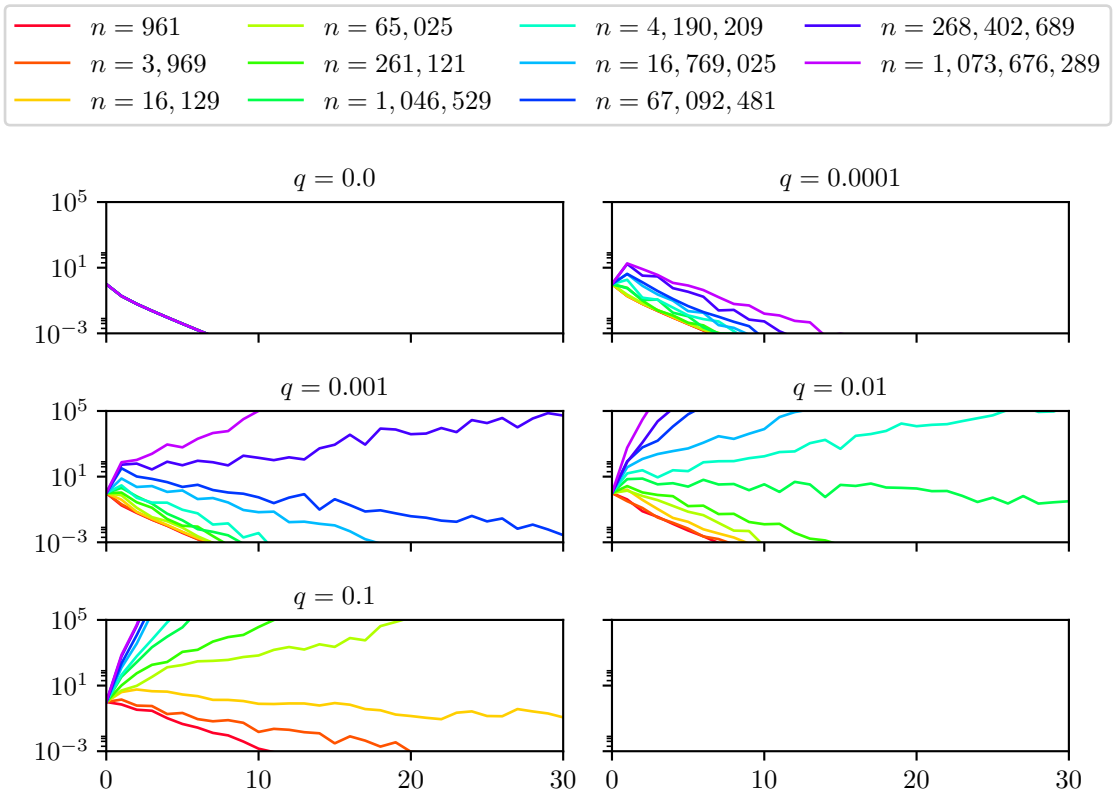


Figure 5.1: Plot of the norm of the residual against iteration number for the Fault-Prone Two Grid Method in the case of discretization of the Poisson problem on a square domain.

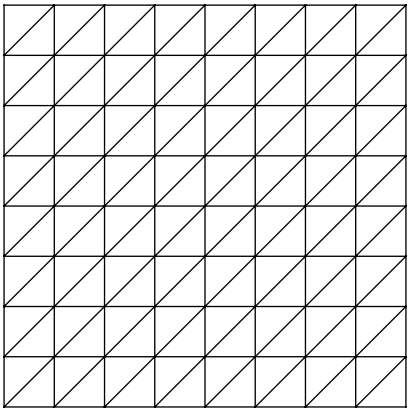


Figure 5.2: Uniform mesh for the square domain.

in order to realize line 5 of Algorithm 2, which contains two fault terms, we sample $n_C + n_F$ independent realizations from an ε-distributed Bernoulli variable (each resulting in a value of zero or one), and multiply residual and coarse-grid right-hand side element by element with the corresponding sampled values. The same approach is applied to simulate faults in lines 4, 9 and 11 of Algorithm 2, where the expression for the optimally damped Jacobi smoothener is given by eq. (2.6).

In Figure 5.1 we plot the norm of the residual after each iteration for a range of values of failure probability ε and system sizes n_F ranging from 1 thousand to 1 billion. These computations were performed on the Titan supercomputer located at Oak Ridge National Laboratory.

It is observed that when $\varepsilon = 0$, the method reaches residual norm 10^{-3} in about 7 iterations uniformly for all values of n_F up to around one billion. However, as ε grows, it is observed that the convergence deteriorates in all cases, with the deterioration being more and more severe as the number of unknowns is increased, until, eventually, the method fails to converge at all.

The fact that the Two Grid Method fails to converge even for such a simple classical problem means one cannot expect a more favorable behavior for realistic practical problems.

In order to verify the scaling law suggested by eq. (5.1) we examine the case

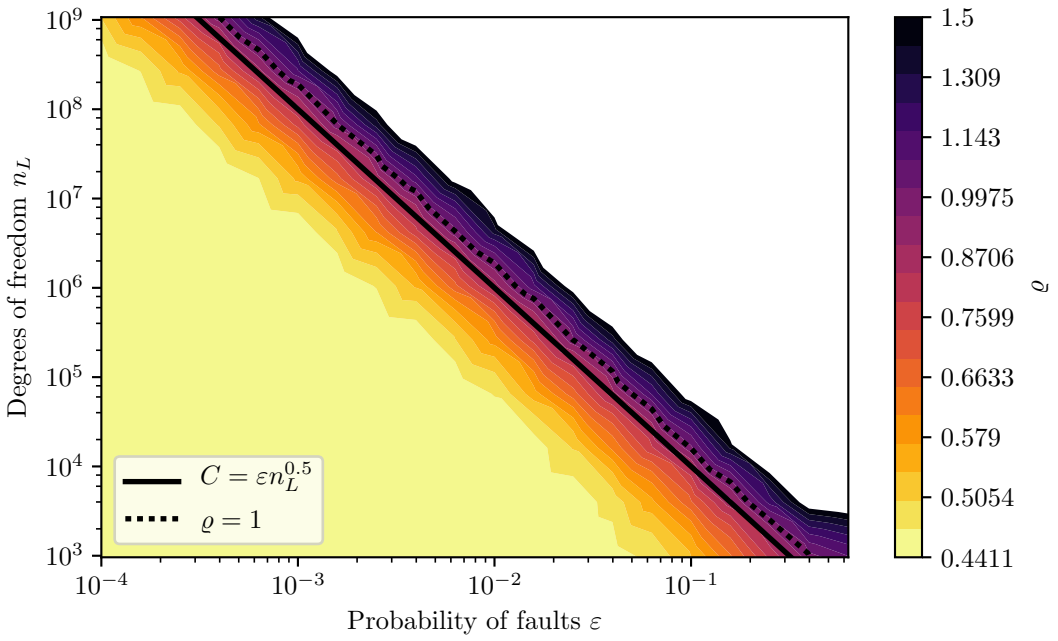


Figure 5.3: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}^{TG})$ for the iteration matrix for the Fault-Prone Two Grid Method in the case of discretization of the Poisson problem on a square domain.

$d = 2$ further. In particular, we expect convergence to fail when $\varepsilon\sqrt{n_F}$ exceeds some threshold. The Lyapunov spectral radius of the Fault-Prone Two Grid Method is approximated by repeatedly applying Algorithm 2. We perform 1000 iterations for each parameter combination of ε and n_F, and avoid under- and overflow of floating point numbers by renormalizing the iterates. Figure 5.3 shows a contour plot of the variation of the Lyapunov spectral radius with respect to the failure probability ε and the size n_F of the system. The plot indicates the boundary of the region in which the Lyapunov spectral radius exceeds unity, or equally, the region in which the fault-prone iteration no longer converges at all. It will be observed that the lines on which $\varepsilon\sqrt{n_F}$ remains constant (also indicated in the plot) coincide with the contours of the plot in the region in which failure to converge occurs. This supports the scaling law suggested in the estimate eq. (5.1).

5.2 Protecting the Prolongation Restores Resilience

Theorem 12 indicates that the Fault-Prone Two Grid Method will generally not be resilient to faults. What remedial action, in addition to laissez-faire, is needed to restore the uniform convergence of the Fault-Prone Multigrid Method to that of the fault-free scheme? Theorem 18 states that if the *prolongation* operation is protected, meaning that $\boldsymbol{\mathcal{X}}^{(P)} = \boldsymbol{I}$, then the rate of convergence of the Two Grid Method is independent of the number of unknowns and close to the rate of the fault-free method.

Theorem 18.
Let $\boldsymbol{\mathcal{E}}^{TG}(\nu_1,\nu_2) = \left(\boldsymbol{\mathcal{E}}^{S,post}\right)^{\nu_2} \boldsymbol{\mathcal{E}}^{CG} \left(\boldsymbol{\mathcal{E}}^{S,pre}\right)^{\nu_1}$ be the iteration matrix of the Fault-Prone Two Grid Method with faults in smoothener, residual and restriction, and protected prolongation. Provided Assumptions 1 to 5 hold, and that $\boldsymbol{N}$ and $\boldsymbol{\mathcal{X}}^{(S)}$ commute, we find that

$$\varrho\left(\boldsymbol{\mathcal{E}}^{TG}(\nu_1,\nu_2)\right) \leq \min_{\substack{\mu_1+\mu_2=\nu_1+\nu_2 \\ \mu_1,\mu_2\geq 0}} \left\|\boldsymbol{E}^{TG}(\mu_1,\mu_2)\right\|_2 + C\varepsilon,$$

where the constant C is independent of ε and n_F.

Proof. Adding and subtracting $\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}(\nu_1,\nu_2)\right]^{\otimes 2}$ from $\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}(\nu_1,\nu_2)\right)^{\otimes 2}\right]$, we estimate in $\|\bullet\|_{A^2}$

$$\begin{aligned}
&\rho\left(\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}(\nu_1,\nu_2)\right)^{\otimes 2}\right]\right) \leq \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}(\nu_1,\nu_2)\right)^{\otimes 2}\right]\right\|_{A^2} \\
&\leq \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}(\nu_1,\nu_2)\right]\right\|_{A^2}^2 + \left(\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right]\right\|_{A^2}^2 + \left\|\boldsymbol{C}^{(R)}\right\|_{A^2} + \left\|\boldsymbol{C}^{(\rho)}\right\|_{A^2} + \left\|\boldsymbol{C}^{(R,\rho)}\right\|_{A^2}\right) \\
&\qquad \times \left(\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]\right\|_{A^2}^2 + \left\|\boldsymbol{C}^{(S)}\right\|_{A^2}\right)^{\nu_1+\nu_2} \\
&\qquad\qquad - \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right]\right\|_{A^2}^2 \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]\right\|_{A^2}^{2(\nu_1+\nu_2)}.
\end{aligned}$$

We then get by Assumptions 2, 4 and 5 and $(e^{\bullet})^2 \leq 1 + 2C\varepsilon + C^2\varepsilon^2 \leq 1 + C\varepsilon$ that

$$\begin{aligned}
\left\|\boldsymbol{C}^{(R)}\right\|_{A^2} &\leq \varepsilon\left(e^{(R)}\right)^2 \left\|\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}_F\right\|_2^2 \|\boldsymbol{R}\|_2^2 \leq \varepsilon\left(e^{(R)}\right)^2 \underline{C}_p^2 \overline{C}_p^2 C_A^2 \leq C\varepsilon, \\
\left\|\boldsymbol{C}^{(\rho)}\right\|_{A^2} &\leq \varepsilon\left(e^{(\rho)}\right)^2 \left\|\boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}_F\right\|_2^2 \leq \varepsilon\left(e^{(\rho)}\right)^2 C_A^2 \leq C\varepsilon,
\end{aligned}$$

$$\left\|\boldsymbol{C}^{(R,\rho)}\right\|_{A^2} \leq \varepsilon^2 \left\|\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}_F\right\|_2^2 \|\boldsymbol{R}\|_2^2 \leq \varepsilon^2 \underline{C}_p^2 \overline{C}_p^2 C_A^2 \leq C\varepsilon^2.$$

We also estimate

$$\left\|\boldsymbol{C}^{(S)}\right\|_{A^2} = \left\|\boldsymbol{A}_F^{\otimes 2}\boldsymbol{V}^{(S)}\boldsymbol{N}^{\otimes 2}\right\|_2 = \left\|\left(\boldsymbol{A}_F\boldsymbol{N}\right)^{\otimes 2}\boldsymbol{V}^{(S)}\right\|_2 \leq \varepsilon \left\|\boldsymbol{A}_F\boldsymbol{N}\right\|_2^2 \leq C\varepsilon.$$

Here, we used Assumptions 3 and 5 and that $\boldsymbol{N}$ and $\boldsymbol{\mathcal{X}}^{(S)}$ commute. Moreover,

$$\begin{aligned}
\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right]\right\|_{A^2} &= \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right]^T\right\|_2 = \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{CG}\right]\right\|_2 \\
&\leq \left\|\boldsymbol{E}^{CG}\right\|_2 + \left|1 - e^{(R)}e^{(\rho)}\right| \left\|\boldsymbol{I} - \boldsymbol{E}^{CG}\right\|_2 \\
&\leq C_A\left(1 + \left|1 - e^{(R)}e^{(\rho)}\right|\right) \leq C\left(1+\varepsilon\right), \\
\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]\right\|_{A^2} &= \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]^T\right\|_2 = \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]\right\|_2 \\
&\leq \left\|\boldsymbol{E}^{S}\right\|_2 + \left|1 - e^{(S)}\right| \left\|\boldsymbol{I} - \boldsymbol{E}^{S}\right\|_2 \leq C(1+\varepsilon), \\
\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\left(\nu_1,\nu_2\right)\right]\right\|_{A^2} &= \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\left(\nu_1,\nu_2\right)\right]^T\right\|_2 = \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\left(\nu_2,\nu_1\right)\right]\right\|_2 \\
&= \left\|\boldsymbol{E}^{TG}\left(\nu_2,\nu_1\right)\right\|_2 + \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\left(\nu_2,\nu_1\right)\right]\right\|_2 - \left\|\boldsymbol{E}^{TG}\left(\nu_2,\nu_1\right)\right\|_2 \\
&\leq \left\|\boldsymbol{E}^{TG}\left(\nu_2,\nu_1\right)\right\|_2 + C\varepsilon.
\end{aligned}$$

Here, we used that by Assumption 5

$$\left|1 - e^{(\bullet)}\right| \leq C\varepsilon, \qquad \left(e^{(\bullet)}\right)^2 \leq 1 + C\varepsilon, \qquad \left|1 - e^{(R)}e^{(\rho)}\right| \leq C\varepsilon.$$

Collecting all the terms, we have

$$\rho\left(\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\left(\nu_1,\nu_2\right)\right)^{\otimes 2}\right]\right) \leq \left\|\boldsymbol{E}^{TG}\left(\nu_2,\nu_1\right)\right\|_2^2 + C\varepsilon.$$

so that we finally obtain

$$\varrho\left(\boldsymbol{\mathcal{E}}^{TG}\left(\nu_1,\nu_2\right)\right) \leq \left\|\boldsymbol{E}^{TG}\left(\nu_2,\nu_1\right)\right\|_2 + C\varepsilon$$

We conclude by observing that the Lyapunov spectral radius, just as the ordinary spectral radius, is invariant with respect to cyclic permutations. □

Theorem 18 makes no additional assumptions about the origin of the problem or the solver hierarchy, and allows for more general fault patterns and smootheners than Theorem 12 and as such, is more generally applicable.

We apply the Fault-Prone Two Grid Method incorporating protection of the prolongation to the same test problem as above. In Figure 5.5 we plot the evolution of the residual norm for varying system sizes and fault rates, and in Figure 5.4 we plot the Lyapunov spectral radius of the iteration matrix. Observe that the rate of convergence is independent of the number of unknowns and even stays below unity (indicating convergence) for fault probabilities as high as $\varepsilon \approx 0.6$. Moreover, Figure 5.5 seems to indicate that while the asymptotic behavior is independent of the number of unknowns, the first iteration is not.

In practice, we are faced with the problem of protecting the prolongation. Some developers have proposed architectures on which certain components are less susceptible to faults, and the prolongation operator would be a candidate for this treatment. We shall present a practical approach to the protection of the prolongation in Chapter 7.

5.3 Protecting the Restriction Fails to Give Resilience

The previous result raises the question as to whether protection of any other component of the algorithm can give similar (or better) results. From a theoretical aspect, this is does not seem to be the case, as careful examination of the proof of Theorem 18 reveals. A provably resilient method can only be obtained if either the prolongation or both restriction and residual in the coarse-grid correction are protected. To confirm whether or not this proves to be the case in practice, we repeat the previous simulations, using instead a protected restriction operation. The measured rates of convergence as shown in Figure 5.6 resemble the behavior in the case without protection. Another possible alternative might be to protect the calculation

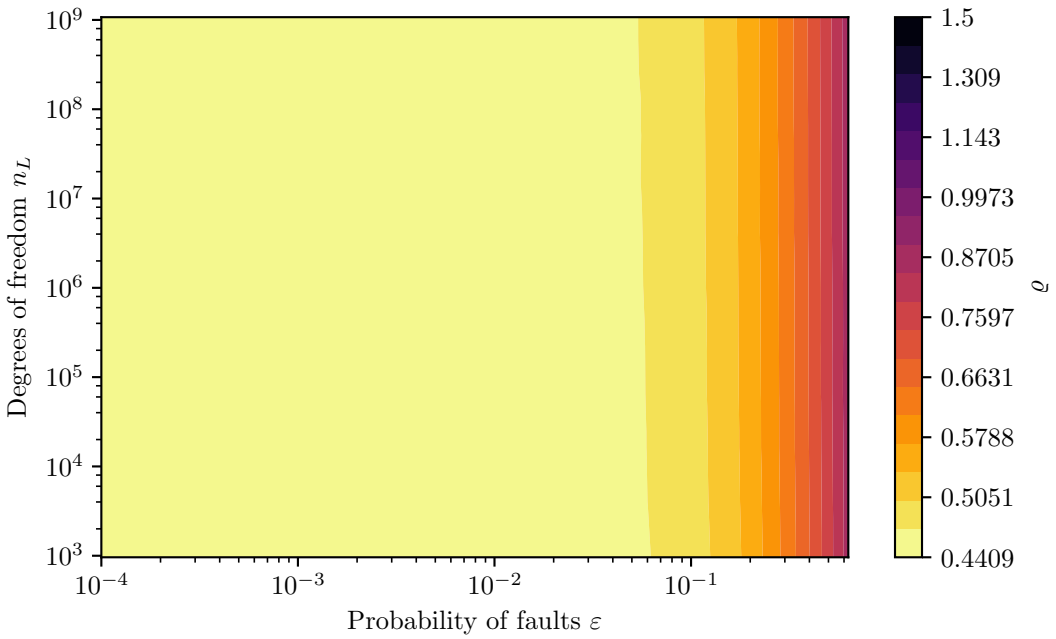


Figure 5.4: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}^{TG})$ for the iteration matrix for the Fault-Prone Two Grid Method with protected prolongation in the case of discretization of Poisson problem on a square domain.

of the residual. However, this is computationally unattractive, since the residual constitutes typically the most expensive part of the method and will therefore have a larger overhead than protection of the prolongation or restriction.

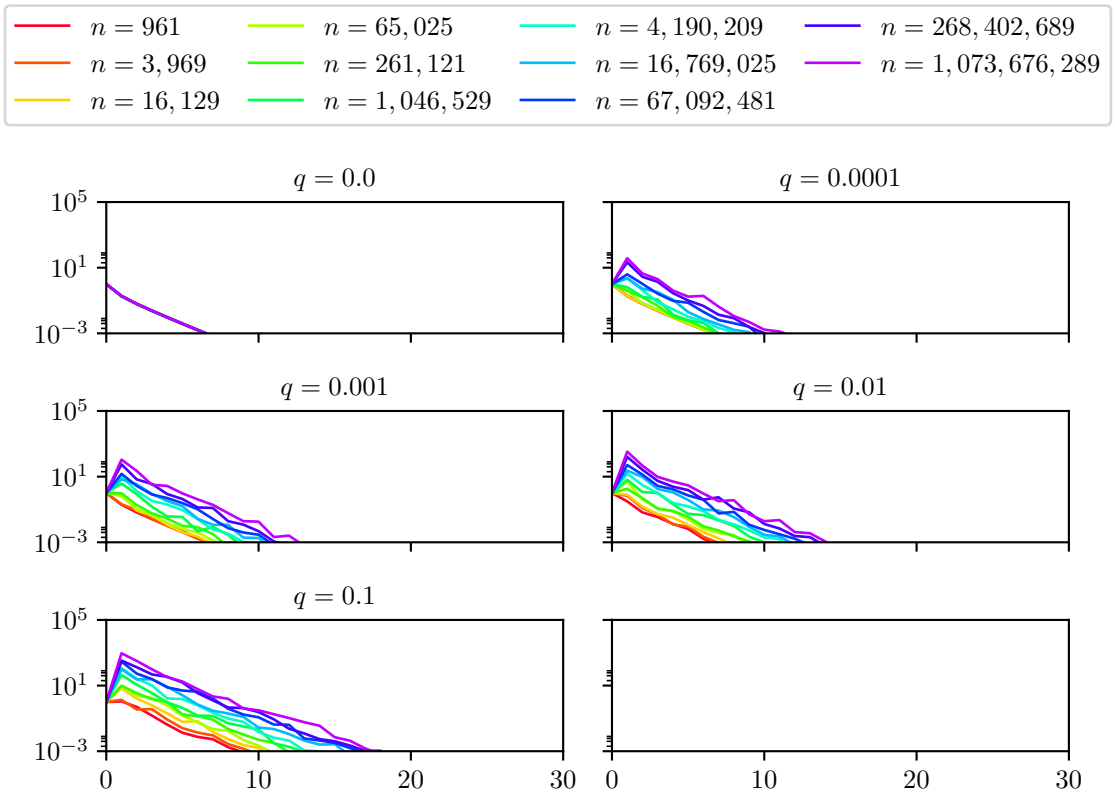


Figure 5.5: Plot of the norm of the residual against iteration number for the Fault-Prone Two Grid Method with protected prolongation in the case of discretization of Poisson problem on square domain.

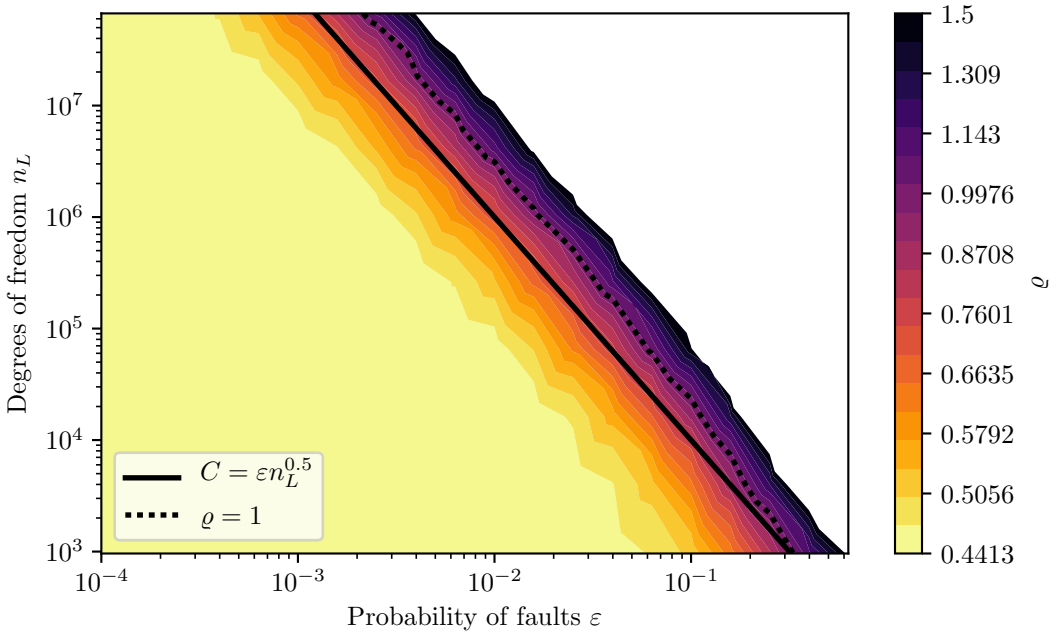


Figure 5.6: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}^{TG})$ for the iteration matrix for the Fault-Prone Two Grid Method with protected restriction in the case of discretization of Poisson problem on a square domain.

Chapter 6

Analysis of the Fault-Prone Multigrid Method

While the two grid method is, as a solver, mostly of academic interest, the behavior of the multigrid method under the impact of faults is of great practical importance. We extend the result of Theorem 18 to the W-cycle. This theorem mirrors the classical implication of W-cycle convergence by two grid convergence outlined in Section 1.3, but applies to the Lyapunov spectral radius needed for the analysis of the Fault-Prone Multigrid Method. No additional assumptions are required beyond these needed for the classical multigrid analysis.

Theorem 19.
Provided Assumptions 1 to 5 hold, that the prolongation is protected, that the number of smoothing steps is sufficient, and that ε is sufficiently small, the fault-prone multigrid method converges with a rate bounded by

$$\varrho\left(\boldsymbol{\mathcal{E}}_\ell\left(\nu_1,\nu_2,\gamma\right)\right) \leq \begin{cases} \frac{\gamma}{\gamma-1}C_{TG} + C\varepsilon, & \gamma \geq 2, \\ \frac{2}{1+\sqrt{1-4C_*C_{TG}}}C_{TG} + C\varepsilon, & \gamma = 2, \end{cases}$$

where C is independent of ε and ℓ and

$$C_{TG} = \max_\ell \left\| \boldsymbol{E}_\ell^{TG}\left(\nu_2,\nu_1\right)\right\|_2 \leq C_S C_A \eta(\nu_2),$$

$$C_* = C_S \underline{C}_p \overline{C}_p \left(C_S + C_A \eta(\nu_2) \right).$$

Setting the fault rate $\varepsilon = 0$ in the above bounds recovers the classical estimates from Section 1.3 for the fault-free multigrid method, since $\varrho\left(\boldsymbol{\mathcal{E}}_\ell\right)$ reduces to $\rho\left(\boldsymbol{E}_\ell\right)$ for $\varepsilon = 0$. Just as the Two Grid result, this Theorem makes no assumptions about the origin of the solver hierarchy, and does not rely on a particular choice of smoothener. The proof of Theorem 19 requires several steps that are recorded in the results below.

Throughout this section, C will be a generic constant whose value can change from line to line, but which is independent of ℓ and ε.

We set

$$\mathbb{E}\left[\boldsymbol{\mathcal{X}}_\ell^{(\bullet)}\right] = e_\ell^{(\bullet)} \boldsymbol{I}, \qquad \mathbb{V}\left[\boldsymbol{\mathcal{X}}_\ell^{(\bullet)}\right] = \boldsymbol{V}_\ell^{(\bullet)}.$$

and expand

$$\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}_\ell^S\right)^{\otimes 2}\right] = \mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell^S\right]^{\otimes 2} + \boldsymbol{C}_\ell^{(S)},$$
$$\boldsymbol{C}_\ell^{(S)} = \boldsymbol{V}_\ell^{(S)} \left(\boldsymbol{N}_\ell \boldsymbol{A}_\ell\right)^{\otimes 2},$$

using Lemma 9. Similar to the fault-free case, we derive recursive inequalities for the iteration matrix of the multigrid method.

Theorem 20.

If Assumptions 1 to 5 hold, then the iteration matrix of the Fault-Prone Multigrid Method satisfies the following recursive inequalities:

$$\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell\left(\nu_1, \nu_2, \gamma\right)\right]\right\|_{A^2} \leq \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell^{TG}\left(\nu_1, \nu_2\right)\right]\right\|_{A^2} + C\varepsilon + C_* \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_{l-1}\left(\nu_1, \nu_2, \gamma\right)\right]\right\|_{A^2}^{\gamma}, \tag{6.1}$$

and

$$\begin{aligned}\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell\left(\nu_1, \nu_2, \gamma\right)^{\otimes 2}\right]\right\|_{A^2} \leq{}& \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell^{TG}\left(\nu_1, \nu_2\right)^{\otimes 2}\right]\right\|_{A^2} + C\varepsilon \\ &+ 2C_* \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell^{TG}\left(\nu_1, \nu_2\right)\right]\right\|_{A^2} \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_{l-1}\left(\nu_1, \nu_2, \gamma\right)\right]\right\|_{A^2}^{\gamma}\end{aligned} \tag{6.2}$$

$$+ C_*^2 \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}_{l-1}\left(\nu_1,\nu_2,\gamma\right)^{\otimes 2} \right] \right\|_{A^2}^{\gamma},$$

with

$$C_* = C_S \underline{C}_p \overline{C}_p \left(C_S + C_A \eta(\nu_2) \right).$$

The constants C and C_* are independent of the level ℓ and ε.

If $\varepsilon = 0$, the second inequality matches exactly the recursive inequality that we would use for the fault-free case, while the first inequality matches the tensor square power of the same inequality.

Proof. Since the proof on level ℓ only involves the levels ℓ and $\ell - 1$, we will drop the first subscript and replace the second one with a subscript C.

We can write the iteration matrix as

$$\begin{aligned}
\boldsymbol{\mathcal{E}}\left(\nu_1,\nu_2,\gamma\right) &= \left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2} \left(\boldsymbol{I} - \boldsymbol{P}\left(\boldsymbol{I} - \left(\boldsymbol{\mathcal{E}}_C\right)^{\gamma}\right) \boldsymbol{A}_C^{-1} \boldsymbol{\mathcal{X}}_C^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}^{(\rho)} \boldsymbol{A} \right) \left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1} \\
&= \underbrace{\left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2} \left(\boldsymbol{I} - \boldsymbol{P} \boldsymbol{A}_C^{-1} \boldsymbol{\mathcal{X}}_C^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}^{(\rho)} \boldsymbol{A} \right) \left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1}}_{\boldsymbol{\mathcal{E}}^{TG}\left(\nu_1,\nu_2\right)} \\
&\quad + \underbrace{\left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2} \boldsymbol{P} \left(\boldsymbol{\mathcal{E}}_C\right)^{\gamma} \boldsymbol{A}_C^{-1} \boldsymbol{\mathcal{X}}_C^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}^{(\rho)} \boldsymbol{A} \left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1}}_{=:\boldsymbol{\mathcal{C}}}.
\end{aligned}$$

Since

$$\begin{aligned}
\left\| \mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{S}\right)^{\nu_2} \boldsymbol{P} \right] \right\|_{A^2} &= \left\| \boldsymbol{A} \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S} \right]^{\nu_2} \boldsymbol{P} \boldsymbol{A}_C^{-1} \right\|_2 \\
&= \left\| \boldsymbol{A}_C^{-1} \boldsymbol{R} \boldsymbol{A} \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S} \right]^{\nu_2} \right\|_2 \\
&\leq \underline{C}_p \left(\left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S} \right]^{\nu_2} \right\|_2 + \left\| \boldsymbol{E}^{CG} \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S} \right]^{\nu_2} \right\|_2 \right) \\
&\leq \underline{C}_p \left(C_S + C_A \eta(\nu_2) \right) + C\varepsilon, \\
\left\| \mathbb{E}\left[\boldsymbol{A}_C^{-1} \boldsymbol{\mathcal{X}}_C^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}^{(\rho)} \boldsymbol{A} \right] \right\|_{A^2} &\leq \left| e_C^{(R)} e^{(\rho)} \right| \left\| \boldsymbol{R} \right\|_2 \leq \overline{C}_p + C\varepsilon, \\
\left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S} \right]^{\nu_1} \right\|_{A^2} &\leq C_S + C\varepsilon,
\end{aligned}$$

we have by Assumptions 4 and 5 that

$$\left\| \mathbb{E}\left[\boldsymbol{\mathcal{C}} \right] \right\|_{A^2} \leq \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S} \right]^{\nu_2} \boldsymbol{P} \right\|_{A^2} \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}_C \right] \right\|_{A^2}^{\gamma} \left\| \mathbb{E}\left[\boldsymbol{A}_C^{-1} \boldsymbol{\mathcal{X}}_C^{(R)} \boldsymbol{R} \boldsymbol{\mathcal{X}}^{(\rho)} \boldsymbol{A} \right] \right\|_{A^2} \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S} \right]^{\nu_1} \right\|_{A^2}$$

$$\leq C_* \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C\right]\right\|_{A^2}^{\gamma} + C\varepsilon. \tag{6.3}$$

Therefore the first inequality of the theorem follows from

$$\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}\right]\right\|_{A^2} \leq \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\right]\right\|_{A^2} + \left\|\mathbb{E}\left[\boldsymbol{\mathcal{C}}\right]\right\|_{A^2}.$$

The second inequality is more work, but follows the same path:

$$\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{\otimes 2}\right]\right\|_{A^2} \leq \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\right)^{\otimes 2}\right]\right\|_{A^2} + \left\|\mathbb{E}\left[\boldsymbol{\mathcal{C}}^{\otimes 2}\right]\right\|_{A^2} + 2\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG} \otimes \boldsymbol{\mathcal{C}}\right]\right\|_{A^2}. \tag{6.4}$$

We start by bounding the second term on the right-hand side.

$$\begin{aligned}\left\|\mathbb{E}\left[\boldsymbol{\mathcal{C}}^{\otimes 2}\right]\right\|_{A^2} \leq & \left\|\mathbb{E}\left[\left(\left(\boldsymbol{\mathcal{E}}^{S}\right)^{\nu_2} \boldsymbol{P}\right)^{\otimes 2}\right]\right\|_{A^2} \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C^{\otimes 2}\right]\right\|_{A^2}^{\gamma} \\ & \times \left\|\mathbb{E}\left[\left(\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right)^{\otimes 2}\right]\right\|_{A^2} \left\|\mathbb{E}\left[\left(\left(\boldsymbol{\mathcal{E}}^{S}\right)^{\nu_1}\right)^{\otimes 2}\right]\right\|_{A^2}\end{aligned} \tag{6.5}$$

We have by

$$\begin{aligned}\left\|\mathbb{E}\left[\left(\left(\boldsymbol{\mathcal{E}}^{S}\right)^{\nu_2} \boldsymbol{P}\right)^{\otimes 2}\right]\right\|_{A^2} &\leq \left\|\left(\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]^{\nu_2} \boldsymbol{P}\right)^{\otimes 2}\right\|_{A^2} + C\left\|\boldsymbol{C}^{(S)}\right\|_{A^2} \\ &= \left\|\boldsymbol{A}\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]^{\nu_2} \boldsymbol{P}\boldsymbol{A}_C^{-1}\right\|_2^2 + C\varepsilon \\ &= \left\|\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]^{\nu_2}\right\|_2^2 + C\varepsilon \\ &\leq \underline{C}_p^2\left(\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]^{\nu_2}\right\|_2 + \left\|\boldsymbol{E}^{CG}\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{S}\right]^{\nu_2}\right\|_2\right)^2 + C\varepsilon \\ &\leq \underline{C}_p^2\left(C_S + C_A\eta(\nu_2)\right)^2 + C\varepsilon.\end{aligned} \tag{6.6}$$

Since by Assumptions 4 and 5

$$\begin{aligned}\left\|\mathbb{E}\left[\left(\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right)^{\otimes 2}\right]\right\|_{A^2} &= \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\right)^{\otimes 2}\right]\right\|_2 \\ &\leq \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{X}}_C^{(R)}\right)^{\otimes 2}\right]\right\|_2 \|\boldsymbol{R}\|_2^2 \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{X}}^{(\rho)}\right)^{\otimes 2}\right]\right\|_2 \\ &\leq \left(\left\|\mathbb{V}\left[\boldsymbol{\mathcal{X}}_C^{(R)}\right]\right\|_2 + \left\|\mathbb{E}\left[\boldsymbol{\mathcal{X}}_C^{(R)}\right]\right\|_2^2\right)\|\boldsymbol{R}\|_2^2 \\ &\quad \times \left(\left\|\mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right]\right\|_2 + \left\|\mathbb{E}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right]\right\|_2^2\right) \\ &\leq \left(\varepsilon + \left(e^{(R)}\right)^2\right)\left(\varepsilon + \left(e^{(\rho)}\right)^2\right)\|\boldsymbol{R}\|_2^2\end{aligned} \tag{6.7}$$

$$\leq \overline{C}_p^2 + C\varepsilon.$$

and

$$\begin{aligned}\left\|\mathbb{E}\left[\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_1}\right)^{\otimes 2}\right]\right\|_{A^2} &= \left\|\left(\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\otimes 2} + \boldsymbol{C}^{(S)}\right)^{\nu_1}\right\|_{A^2} \\ &\leq \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right\|_{A^2}^2 + C\varepsilon \leq C_S^2 + C\varepsilon,\end{aligned} \tag{6.8}$$

Inserting eqs. (6.8), (6.6) and (6.7) into eq. (6.5), we have

$$\left\|\mathbb{E}\left[\boldsymbol{\mathcal{C}}^{\otimes 2}\right]\right\|_{A^2} \leq C_*^2 \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C^{\otimes 2}\right]\right\|_{A^2}^{\gamma} + C\varepsilon. \tag{6.9}$$

Now we turn to the third term of eq. (6.4). We split

$$\begin{aligned}\boldsymbol{\mathcal{E}}^{TG} &= \left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2}\left(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right)\left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1} \\ &= \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\boldsymbol{E}^{CG}\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1} \\ &\quad + \left(\left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\right)\boldsymbol{E}^{CG}\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1} \\ &\quad + \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\boldsymbol{E}^{CG}\left(\left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right) \\ &\quad + \left(\left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\right)\boldsymbol{E}^{CG}\left(\left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right) \\ &\quad + \left(\boldsymbol{\mathcal{E}}^{S,\text{post}}\right)^{\nu_2}\boldsymbol{P}\left(\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A} - \boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right)\left(\boldsymbol{\mathcal{E}}^{S,\text{pre}}\right)^{\nu_1}\end{aligned}$$

so that

$$\begin{aligned}&\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\otimes\boldsymbol{\mathcal{C}}\right]\right\|_{A^2} \\ &\leq \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\boldsymbol{E}^{CG}\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right\|_{A^2}\left\|\mathbb{E}\left[\boldsymbol{\mathcal{C}}\right]\right\|_{A^2} \\ &\quad + \left\|\boldsymbol{E}^{CG}\right\|_{A^2}\left\|\boldsymbol{P}\right\|_{A^2}\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C\right]\right\|_{A^2}^{\gamma}\left\|\mathbb{E}\left[\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right]\right\|_{A^2} \\ &\quad \left\{\left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_2}\otimes\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_2} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\right)\right]\right\|_{A^2}\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right\|_{A^2}^2\right. \\ &\quad + \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\right\|_{A^2}^2\left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_1}\otimes\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_1} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right)\right]\right\|_{A^2} \\ &\quad + \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_2}\otimes\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_2} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\right)\right]\right\|_{A^2} \\ &\quad \left.\times\left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_1}\otimes\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_1} - \mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right)\right]\right\|_{A^2}\right\} \\ &\quad + \left\|\mathbb{E}\left[\left(\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}\right)\otimes\left(\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right) - \left(\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right)^{\otimes 2}\right]\right\|_{A^2}\end{aligned} \tag{6.10}$$

$$\times \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C\right]\right\|_{A^2}^{\gamma} \left\|\mathbb{E}\left[\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu}\boldsymbol{P}\right)^{\otimes 2}\right]\right\|_{A^2} \left\|\mathbb{E}\left[\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu_1}\right)^{\otimes 2}\right]\right\|_{A^2}$$

Now, we have by Assumptions 2 and 4 that

$$\left\|\boldsymbol{E}^{CG}\right\|_{A^2} = \left\|\boldsymbol{E}^{CG}\right\|_2 \leq C_A, \tag{6.11}$$

$$\left\|\boldsymbol{P}\right\|_{A^2} = \left\|\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}_C^{-1}\right\|_2 \leq \underline{C}_p, \tag{6.12}$$

and by Assumptions 4 and 5 that

$$\begin{aligned}
&\left\|\mathbb{E}\left[\left(\boldsymbol{A}_C^{-1}\boldsymbol{R}\boldsymbol{A}\right)\otimes\left(\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right)-\left(\boldsymbol{A}_C^{-1}\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\boldsymbol{A}\right)^{\otimes 2}\right]\right\|_{A^2} \qquad (6.13)\\
=&\left\|\mathbb{E}\left[\boldsymbol{R}\otimes\left(\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\right)-\left(\boldsymbol{\mathcal{X}}_C^{(R)}\boldsymbol{R}\boldsymbol{\mathcal{X}}^{(\rho)}\right)^{\otimes 2}\right]\right\|_2\\
\leq&\left\|\boldsymbol{R}\right\|_2^2\left\{\left|e_C^{(R)}e^{(\rho)}\right|\left|1-e_C^{(R)}e^{(\rho)}\right|+\left(e^{(\rho)}\right)^2\left\|\mathbb{V}\left[\boldsymbol{\mathcal{X}}_C^{(R)}\right]\right\|_2\right.\\
&\left.+\left(e_C^{(R)}\right)^2\left\|\mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right]\right\|_2+\left\|\mathbb{V}\left[\boldsymbol{\mathcal{X}}_C^{(R)}\right]\right\|_2\left\|\mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right]\right\|_2\right\}\\
\leq&C\varepsilon
\end{aligned}$$

and that

$$\begin{aligned}
&\left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu}\otimes\left(\left(\boldsymbol{\mathcal{E}}^S\right)^{\nu}-\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu}\right)\right]\right\|_{A^2} \qquad (6.14)\\
=&\left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^S\right)^{\otimes 2}\right]^{\nu}-\left(\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\otimes 2}\right)^{\nu}\right\|_{A^2}\\
=&\left\|\left(\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\otimes 2}+\boldsymbol{C}^{(S)}\right)^{\nu}-\left(\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\otimes 2}\right)^{\nu}\right\|_{A^2}\\
\leq&C\left\|\boldsymbol{C}^{(S)}\right\|_{A^2}\leq C\varepsilon.
\end{aligned}$$

Also,

$$\begin{aligned}
&\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\boldsymbol{E}^{CG}\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\right\|_{A^2} \qquad (6.15)\\
\leq&\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\left(\nu_1,\nu_2\right)\right]\right\|_{A^2}+\left|1-e_C^{(R)}\right|\left|1-e^{(\rho)}\right|\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_1}\left(\boldsymbol{I}-\boldsymbol{E}^{CG}\right)\mathbb{E}\left[\boldsymbol{\mathcal{E}}^S\right]^{\nu_2}\right\|_2\\
=&\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\left(\nu_1,\nu_2\right)\right]\right\|_{A^2}+C\varepsilon.
\end{aligned}$$

Hence we find by inserting eqs. (6.3), (6.11), (6.12) and (6.13) to (6.14) into eq. (6.10)

$$\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\otimes\boldsymbol{\mathcal{C}}\right]\right\|_{A^2}\leq C_*\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\right]\right\|_{A^2}\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C\right]\right\|_{A^2}^{\gamma}+C\varepsilon. \tag{6.16}$$

Therefore, the second recursive inequality

$$\begin{aligned}\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{\otimes 2}\right]\right\|_{A^2} \leq & \left\|\mathbb{E}\left[\left(\boldsymbol{\mathcal{E}}^{TG}\right)^{\otimes 2}\right]\right\|_{A^2} + C\varepsilon \\ & + 2C_* \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}^{TG}\right]\right\|_{A^2} \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C\right]\right\|_{A^2}^{\gamma} \\ & + \left(C_*\right)^2 \left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_C^{\otimes 2}\right]\right\|_{A^2}^{\gamma}\end{aligned}$$

follows from eqs. (6.4), (6.9) and (6.16). □

We adapt Lemma 2 to the tensor space setting:

Lemma 21.

Let the elements of the sequences $\{\eta_\ell\}_{\ell\geq 1}$ and $\{\eta_{\ell,\otimes 2}\}_{\ell\geq 1}$ be bounded as follows

$$\begin{aligned}\eta_1 &\leq \xi, & \eta_\ell &\leq \xi + C_*\eta_{\ell-1}^{\gamma}, \quad \ell \geq 2, \\ \eta_{1,\otimes 2}^2 &\leq \xi_{\otimes 2}^2, & \eta_{\ell,\otimes 2}^2 &\leq \xi_{\otimes 2}^2 + 2\xi C_*\eta_{\ell-1}^{\gamma} + C_*^2\eta_{\ell-1,\otimes 2}^{2\gamma}, \quad \ell \geq 2.\end{aligned}$$

If $\gamma \geq 2$, $C_*\gamma > 1$ and $\max\{\xi, \xi_{\otimes 2}\} \leq \frac{\gamma-1}{\gamma}\frac{1}{(C_*\gamma)^{\frac{1}{\gamma-1}}}$, then

$$\max\{\eta_\ell, \eta_{\ell,\otimes 2}\} \leq \frac{\gamma}{\gamma-1}\max\{\xi, \xi_{\otimes 2}\} < 1, \qquad \ell \geq 1.$$

In the case of $\gamma = 2$, we have

$$\max\{\eta_\ell, \eta_{\ell,\otimes 2}\} \leq \frac{2\max\{\xi, \xi_{\otimes 2}\}}{1 + \sqrt{1 - 4C_*\max\{\xi, \xi_{\otimes 2}\}}} < 1, \qquad \ell \geq 1.$$

Proof. Set $\alpha_\ell := \max\{\eta_\ell, \eta_{\ell,\otimes 2}\}$ and $\beta := \max\{\xi, \xi_{\otimes 2}\}$. Then

$$\begin{aligned}\eta_1 &\leq \beta, & \eta_\ell &\leq \beta + C_*\alpha_{\ell-1}^{\gamma}, \quad \ell \geq 2, \\ \eta_{1,\otimes 2}^2 &\leq \beta^2, & \eta_{\ell,\otimes 2}^2 &\leq \beta^2 + 2\beta C_*\alpha_{\ell-1}^{\gamma} + C_*^2\alpha_{\ell-1}^{2\gamma}, \quad \ell \geq 2,\end{aligned}$$

and hence

$$\alpha_1 \leq \beta, \qquad \alpha_\ell \leq \beta + C_*\alpha_{\ell-1}^{\gamma}, \quad \ell \geq 2.$$

Application of lemma 2 gives for $\gamma \geq 2$

$$\alpha_\ell \leq \frac{\gamma}{\gamma-1}\beta < 1$$

and for $\gamma = 2$

$$\alpha_\ell \leq \frac{2\beta}{1+\sqrt{1-4C_*\beta}} < 1.$$

□

This finally permits us to give prove the main result of this chapter:

Proof of Theorem 19. Set

$$\begin{aligned}
\eta_\ell &:= \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell\left(\nu_1,\nu_2,\gamma\right)\right]\right\|_{A^2}, \\
\eta_{\ell,\otimes 2} &:= \left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell\left(\nu_1,\nu_2,\gamma\right)^{\otimes 2}\right]\right\|_{A^2}^{\frac{1}{2}}, \\
\xi &:= \max_\ell \left\| \boldsymbol{E}_\ell^{TG}\left(\nu_2,\nu_1\right)\right\|_2 + C\varepsilon, \\
\xi_{\otimes 2} &:= \max_\ell \left\| \boldsymbol{E}_\ell^{TG}\left(\nu_2,\nu_1\right)\right\|_2 + C\varepsilon.
\end{aligned}$$

It follows from

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell^S\right] &= \boldsymbol{E}_\ell^S + \left(1-e_\ell^{(S)}\right)\left(\boldsymbol{I}-\boldsymbol{E}_\ell^S\right), \\
\mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell^{CG}\right] &= \boldsymbol{E}_\ell^{CG} + \left(1-e_{\ell-1}^{(R)}e_\ell^{(\rho)}\right)\left(\boldsymbol{I}-\boldsymbol{E}_\ell^{CG}\right),
\end{aligned}$$

and Assumptions 2, 3 and 5 that

$$\left\| \mathbb{E}\left[\boldsymbol{\mathcal{E}}_\ell^{TG}\left(\nu_1,\nu_2\right)\right]\right\|_{A^2} \leq \left\| \boldsymbol{E}_\ell^{TG}\left(\nu_2,\nu_1\right)\right\|_2 + C\varepsilon,$$

and therefore with eq. (6.1) that $\eta_\ell \leq \xi + C_*\eta_{\ell-1}^\gamma$. By Theorem 18 and eq. (6.2), we find $\eta_{\ell,\otimes 2}^2 \leq \xi_{\otimes 2}^2 + 2\xi C_*\eta_{\ell-1}^\gamma + C_*^2\eta_{\ell-1,\otimes 2}^2$. Since C_*, ξ and $\xi_{\otimes 2}$ decrease as the number of smoothening steps increases, Lemma 21 can be applied for a sufficient number of smoothing steps. □

6.1 The Effect of Protection of the Prolongation

Theorem 19 assumes that the prolongation operator is protected. We investigate whether this assumption can be relaxed by considering a numerical example where

we do not protect the prolongation. Specifically, we consider the Poisson equation in two dimensions

$$\begin{cases} -\Delta u &= f \quad \text{in } \Omega = [0,1]^2, \\ u &= 0 \quad \text{on } \partial\Omega, \end{cases} \tag{6.17}$$

and in order to rule out extraneous effects, we use a uniform mesh, a discretization with piecewise linear finite elements, and optimally damped Jacobi pre- and post-smootheners. In particular, it is well established that the fault-free multigrid method converges in this setting. Assumptions 1 to 4 are satisfied, as for example shown in [62]. On every level, residual, restriction, prolongation and smootheners are subject to componentwise faults, as given in eq. (2.2). We consider problems of size between 1 million and 1 billion degrees of freedom, and fault probabilities between 10^{-4} and 0.6. To minimize floating point contamination in the approximation of the Lyapunov spectral radius, the right-hand side f is taken to be zero, a non-zero random initial iterate is chosen, and after each iteration the current iterate is renormalized. In Figure 6.1, we plot the evolution of the residual norm with respect to the iteration number for the case of laissez-faire mitigation in prolongation, restriction, residual and smoothener. We can see that the number of degrees of freedom n_L adversely affects the rate of convergence, even leading to divergence for large number of unknowns. Estimates of the Lyapunov spectral radius are obtained as the geometric average of 1000 iterations of Fault-Prone Multigrid, and are displayed in Section 6.1.

In the Two Grid setting which was discussed in Chapter 5, we found that the asymptotic rate for this problem scales like $\sqrt{n_L \varepsilon^2}$. Similarly, as seen in Section 6.1, we find

$$\varrho\left(\boldsymbol{\mathcal{E}}_L\right) = C_0 + \mathcal{O}\left(\sqrt{n_L \varepsilon^2}\right) \tag{6.18}$$

for the multilevel case, with C_0 being a constant related to the fault-free method. This means that the method is not fault resilient, since for any given rate of faults

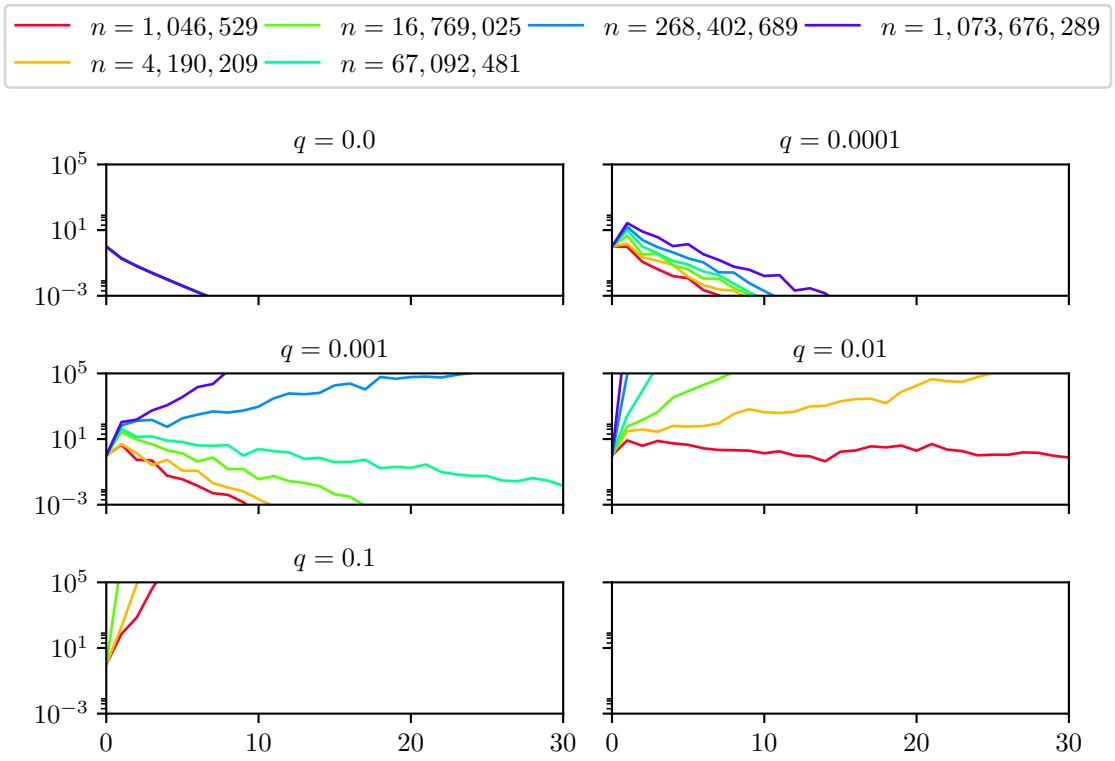


Figure 6.1: Plots of the norm of the residual against iteration number for the Fault-Prone W-Cycle Multigrid Method in the case of discretization of Poisson problem on square domain.

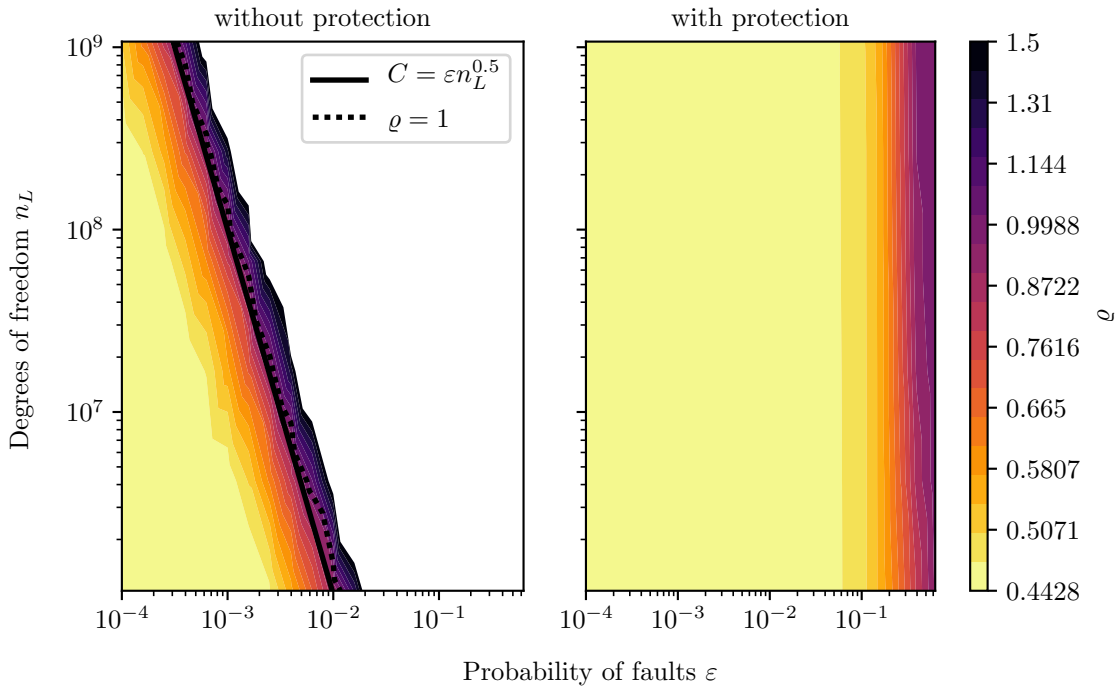


Figure 6.2: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}_L)$ for the iteration matrix for the Fault-Prone W-Cycle Multigrid method in the case of discretization of Poisson problem on a square domain. Without protected prolongation (*left*), and with protected prolongation (*right*).

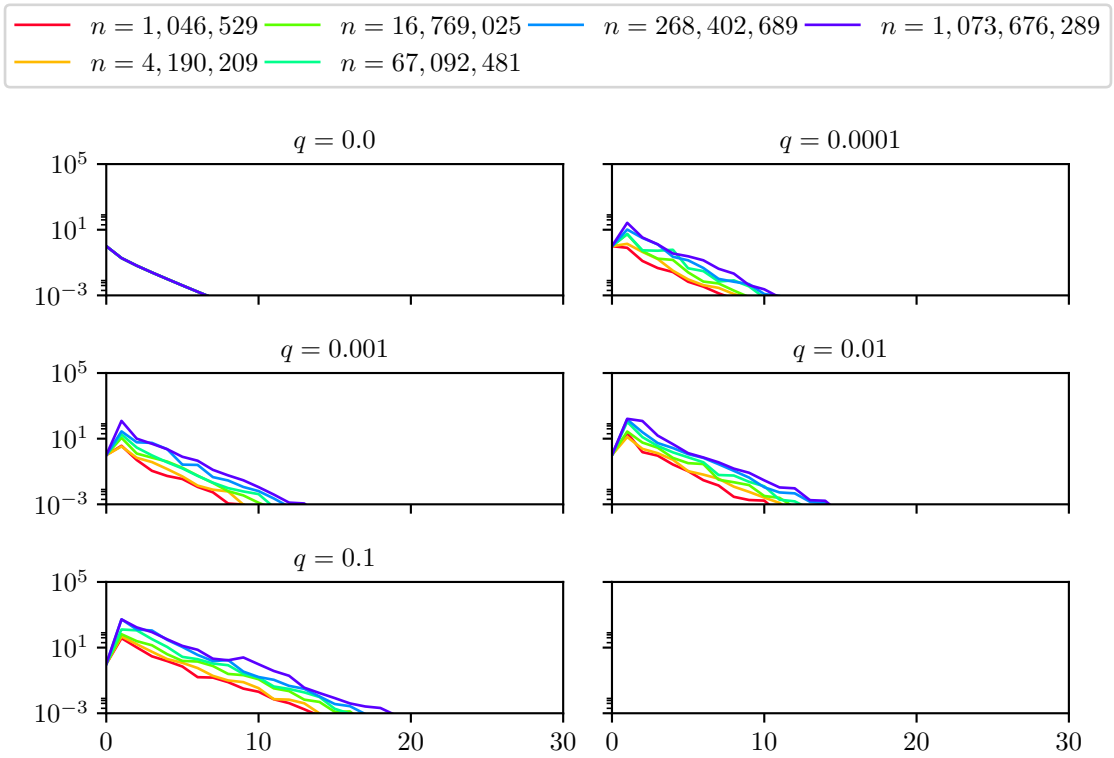


Figure 6.3: Plot of the norm of the residual against iteration number for the Fault-Prone W-Cycle Multigrid Method with protected prolongation in the case of discretization of Poisson problem on square domain.

ε, there exists a maximum problem size above which multigrid diverges. Therefore additional protection is mandatory in the multilevel case as well.

We repeat the same experiments, this time with a protected prolongation, i.e. $\boldsymbol{\mathcal{X}}_{\ell}^{(P)} = \boldsymbol{I}$. This produces the desired independence of the problem size, as predicted by Theorem 19 and as shown by the evolution of the residual in Figure 6.3 and by the estimated rate of convergence in Section 6.1. We further illustrate the results with several test problems that pose more of a challenge to the multigrid solver.

6.2 Adaptively Refined Meshes

A second numerical example illustrates the results of Theorem 19 for the case of an adaptively refined mesh. We consider the 2D magnetostatics problem for a three phase 6/4 switched reluctance motor as depicted in Figure 6.4. Gauss's law and Ampère's law are given by

$$div\vec{B} = 0, \qquad\qquad curl\vec{H} = J,$$

where $\vec{B}$ is the magnetic flux density, $\vec{H}$ the magnetic field intensity and J the current density. $\vec{B}$ and $\vec{H}$ are linked through the constitutive relation $\vec{B} = \mu\vec{H}$ with magnetic permeability μ. Using the magnetic vector potential $\vec{B} = c\vec{u}rl\ u$, the system can be rewritten as

$$-\frac{1}{\mu}\Delta u = J.$$

This gives rise to a variational problem with

$$a(u,v) = \int_\Omega \frac{1}{\mu} gradu \cdot gradv, \qquad\qquad F(v) = \int_\Omega Jv.$$

The permeability μ is 5200 in the rotor and the stator, and 1 everywhere else. The current density J is unity in the coils and 0 everywhere else. Using continuous piecewise linear finite elements and classical residual-based local a posteriori error indicators [50], we adaptively refine an initial mesh shown in Figure 6.4. We use Red-Green refinement [12] coupled with a Dörfler marking strategy [43] with refinement parameter 0.5. The solver hierarchy is generated without injection of faults; then the problem is solved for componentwise faults and varying fault rates.

Section 6.2 depicts the approximate rate of convergence of the W-Cycle with one step of Jacobi pre- and post-smoothening each and without protection of the prolongation. We can see that although neither the W-Cycle nor the adaptively refined mesh are covered by the two grid result from Chapter 5, the convergence estimate is qualitatively correct. Added protection of the prolongation operation

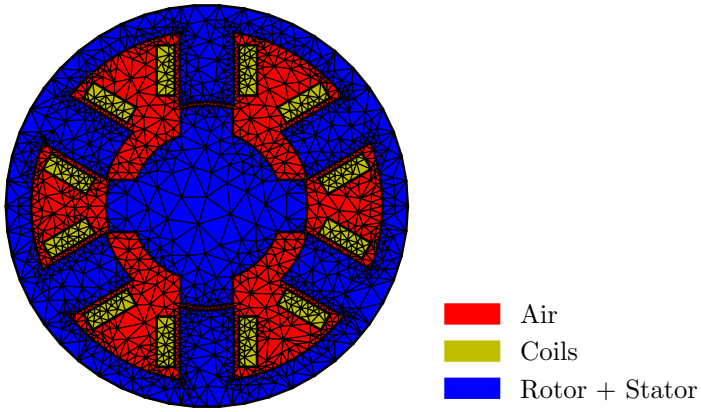


Figure 6.4: Initial mesh for the motor problem.

leads to a fault resilient method, as shown in Section 6.2. Small variations with respect to the number of degrees of freedom are due to varying coarsening ratios of the levels. The results match Theorem 19, even though Assumption 2 is not satisfied.

6.3 Higher Spatial Dimension and Higher Order Finite Elements

We now demonstrate that fault resilience of the W-cycle is retained for a 3D partial differential equation and higher order (quadratic) continuous finite elements, and solve the Poisson equation on the Fichera cube.

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega = [0,2]^3 \setminus [1,2]^3, \\ u &= u_0 && \text{on } \partial\Omega \end{aligned}$$

The geometry and its uniform meshing are shown in Figure 6.6. The system is partitioned using METIS [71] into blocks of size 2^{14} and distributed over the compute nodes. We inject nodewise faults as given by eq. (2.3). The problem sizes considered range from 1.7 million to about 940 million degrees of freedom. In Section 6.3, we show the approximate rate of convergence as well as a level set of the

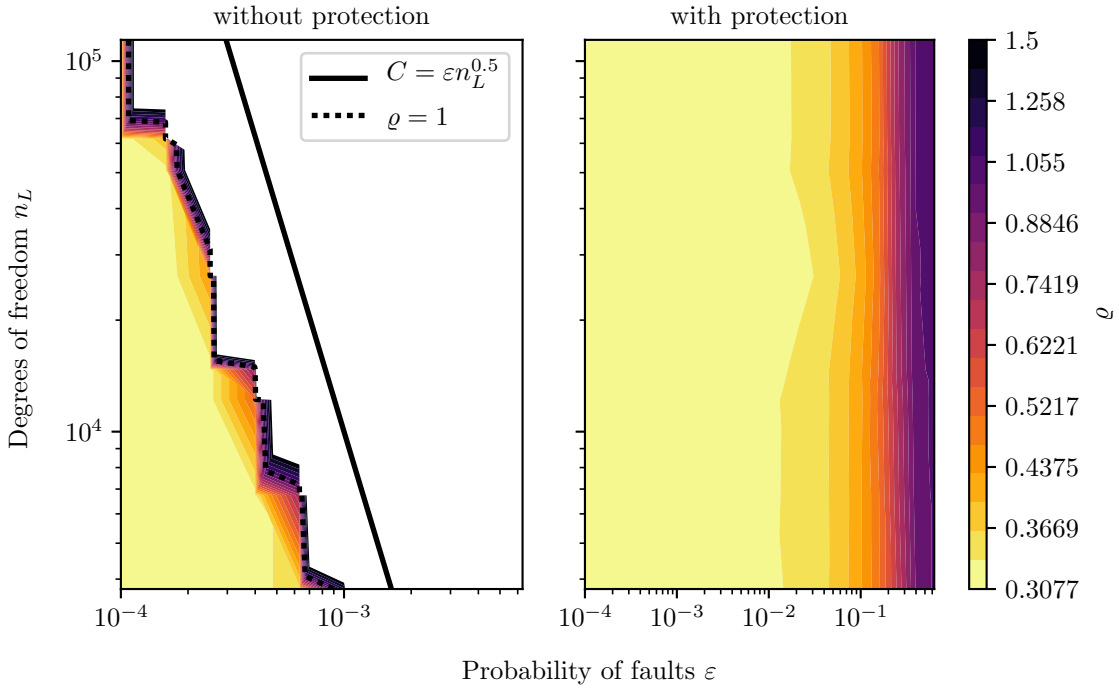


Figure 6.5: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}_L)$ for the iteration matrix for the Fault-Prone W-Cycle Multigrid Method in the case of discretization of the motor problem. Without protected prolongation (*left*), and with protected prolongation (*right*).

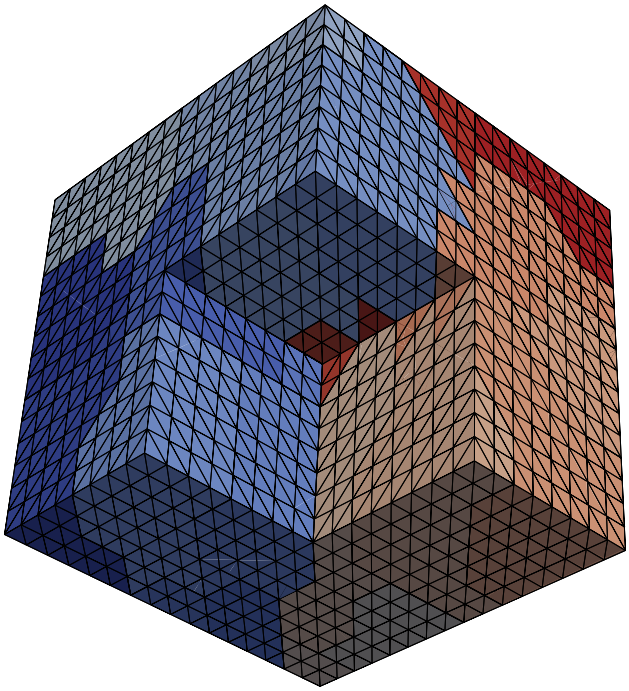


Figure 6.6: Mesh for the Fichera cube, indicating partitioning from METIS.

error bound obtained for the Fault-Prone Two Grid Method in Chapter 5. It can be seen that the rate of convergence in the multilevel case behaves like

$$\varrho\left(\boldsymbol{\mathcal{E}}_L\right) = C_0 + \varepsilon \sqrt[6]{n_L},$$

where C_0 is a constant that is related to the fault-free method. We notice that the onset of divergent behavior indeed happens at a slower rate as compared to the 2D setting. Protection of the prolongation again makes the method fault resilient, as shown in Section 6.3. The choice of higher order elements plays no significant role in the convergence behavior of the method.

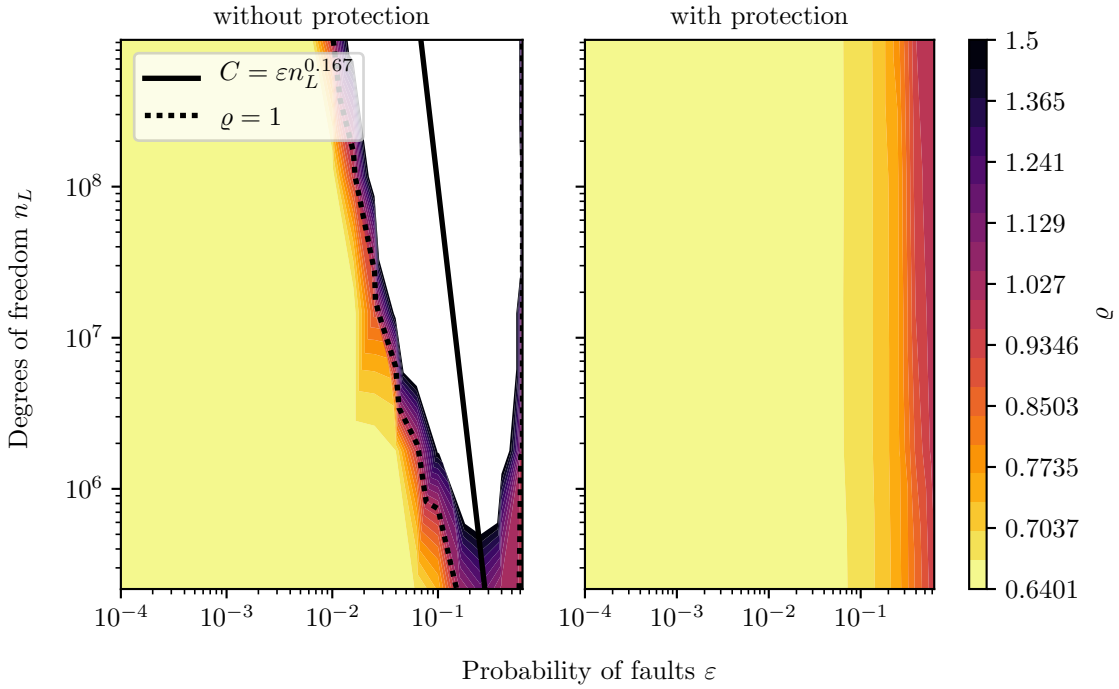


Figure 6.7: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}_L)$ for the iteration matrix for the Fault-Prone W-Cycle Multigrid Method in the case of discretization of the Fichera cube. Without protected prolongation (*left*), and with protected prolongation (*right*).

Chapter 7

Implementation Issues

In the previous chapters, we assumed that faults can be perfectly detected. In practice, faults cannot be perfectly detected, nor is is it possible to perfectly protect the prolongation. The question therefore arises of whether Theorem 19 has any practical relevance. In Section 7.1, we present one possible simple approach for fault detection and show that the behavior of Fault-Prone Multigrid found in eq. (6.18) is recovered. In Section 7.2 we turn to the issue of protecting the prolongation operator.

7.1 Detection of Soft Faults

The laissez-faire fault mitigation strategy requires fault detection. While this is straight-forward in the case of hard faults, soft faults are more problematic. Various techniques have been suggested [89]. Here, we present a simple approach based on replication [64] in which a fault-prone component is repeated K times ($K \geq 2$) and the results are compared for consistency. Algorithm 5 shows how the strategy is used in conjunction with laissez-faire in the detection of faults in the computation of a generic matrix-vector product $y = \boldsymbol{M}x$. Instead of the action of $\boldsymbol{M}$, only its fault-prone equivalent $\boldsymbol{\mathcal{M}}$ is available in practice, where $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{n\times n}$ is a random

matrix.

Algorithm 5 Detection of faults using K replicas and laissez-faire mitigation for the fault-prone matrix-vector product $y = \boldsymbol{\mathcal{M}}x$.

```
1: for i ← 1 to n do
2:     for j ← 1 to K do
3:         w_j ← (M x)_i                    // j-th replica
4:     if w_1 = w_2 = ··· = w_K and |w_1| < 10^16 then
5:         y_i ← w_1                        // value accepted
6:     else
7:         y_i ← 0                          // Laissez-faire
```

The basic idea behind Algorithm 5 is to declare an operation as being fault-free if all replicas are in agreement, otherwise a fault is deemed to have occurred and laissez-faire is triggered. This means that there are three possible outcomes for the i-th component of the output y from Algorithm 5:

- (C) the correct value of $y_i = (\boldsymbol{M}x)_i$ is returned,
- (M) an error is detected, the laissez-faire mitigation is triggered, and $y_i = 0$,
- (U) a fault occurs but remains undetected, and y_i is a corrupted value.

Our analysis in Chapter 6 caters for the cases (C) and (M), but does not take account of case (U). Suppose that the probability of a replica being corrupted is $\varepsilon \ll 1$. Then the probability of *all* K replicas being corrupted in *exactly* the same way is $\mathcal{O}(\varepsilon^K)$, showing that the likelihood of (U) occurring is exponentially small in K. Hence, the probabilities of the outcomes of Algorithm 5 are

$$\begin{aligned}\mathbb{P}\,(C) &= 1 - \mathcal{O}(\varepsilon),\\ \mathbb{P}\,(M) &= \mathcal{O}(\varepsilon),\\ \mathbb{P}\,(U) &= \mathcal{O}(\varepsilon^K).\end{aligned}$$

We illustrate the effect of using our ad-hoc fault detection strategy given in Algorithm 5 in Figure 7.1 for the Poisson problem eq. (6.17). In particular, the faults

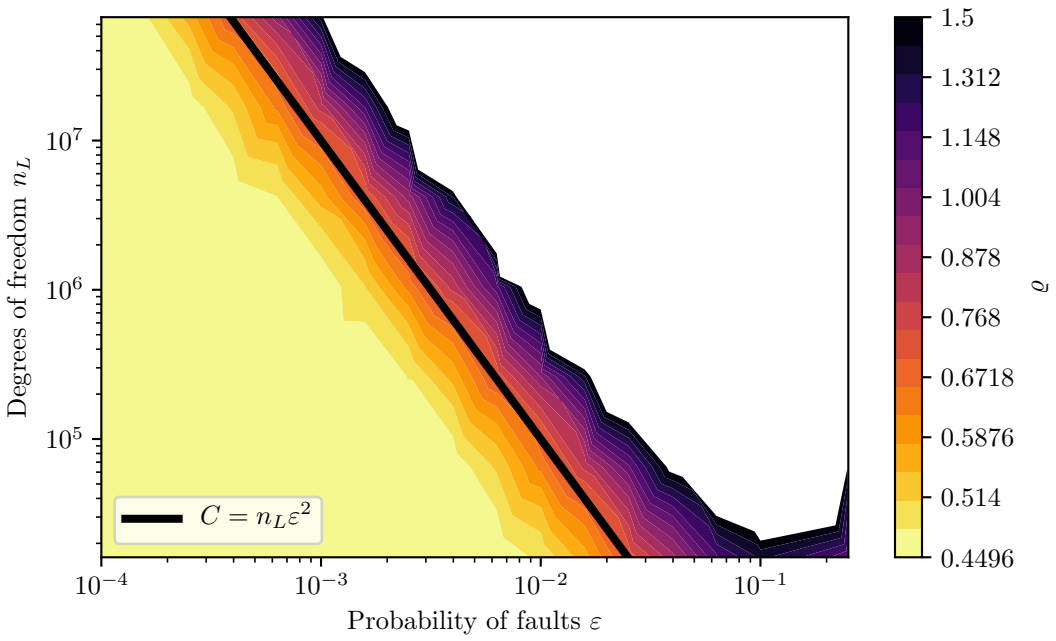


Figure 7.1: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}_L)$ for the iteration matrix for the Fault-Prone W-Cycle Multigrid Method in the case of discretization of the 2D Poisson problem with fault detection using $K = 2$ replicas in every operation.

are modeled using bit flipping in which any bit in the floating point representation is flipped with a small but non-zero probability such that the overall probability of a floating point number being corrupted is ε. The results are given in Figure 7.1 in the simplest case where $K = 2$ replicas are used in Algorithm 5 and all operations. It is observed that, even in this simplest case, the convergence behavior mirrors that which would be obtained with perfect detection.

7.2 Protection of the Prolongation

Theorem 19 requires that the prolongation operations

$$x_\ell \leftarrow x_\ell + \boldsymbol{P} e_{\ell-1}$$

are computed exactly, i.e. without *any* faults. This is clearly not possible in practice. We have seen that in the absence of full protection of the prolongation, the Fault-Prone Multigrid converges at a rate given by eq. (6.18), where ε is the underlying failure rate of the machine. Moreover, Theorem 19 suggests that the $\mathcal{O}(\sqrt{n_L\varepsilon^2})$ term is entirely due to faults in the prolongation (with the faults coming purely from other components of multigrid contributing with higher order terms). At any rate, it is clear that if we can enhance the reliability of the prolongation *sufficiently* (without being necessarily exact), then one can expect to ameliorate the factor $\mathcal{O}\left(\sqrt{n_L\varepsilon^2}\right)$ in the bound of the Lyapunov spectral radius, e.g. by obtaining a growth of $\mathcal{O}\left(n_L\varepsilon^{\alpha}\right)$ for some $\alpha > 2$.

Is it possible to improve the likelihood of detecting and mitigating faults in the prolongation beyond $\mathbb{P}\left(M\right) = \mathcal{O}\left(\varepsilon\right)$? Algorithm 5 may be regarded as being overly conservative. For instance, if all but one of the replicas are in agreement, then intuitively it seems likely that the majority are correct. Algorithm 6 implements the approach suggested by this argument by looking for agreement amongst a subset of k_P of the K_P replicas for the prolongation. This added freedom alters the likelihood

Algorithm 6 Protection of the fault-prone prolongation $x_\ell \leftarrow x_\ell + \boldsymbol{\mathcal{P}} e_{\ell-1}$ with up to K_P replicas, acceptance threshold k_P and laissez-faire mitigation.

```
1: for i ← 1 to n_ℓ do
2:     for j ← 1 to K_P do
3:         w_j ← (P e_{ℓ−1})_i                    // j-th replica
4:         if k_P replicas have matching value w and |w| < 10^16 then
5:             (x_ℓ)_i ← (x_ℓ)_i + w              // value accepted
6:             break
```

at which the three events occur in the prolongation:

$$\mathbb{P}\left(C_P\right) = 1 - \mathcal{O}(\varepsilon^{K_P - k_P + 1}),$$
$$\mathbb{P}\left(M_P\right) = \mathcal{O}(\varepsilon^{K_P - k_P + 1}),$$
$$\mathbb{P}\left(U_P\right) = \mathcal{O}(\varepsilon^{k_P}).$$

In particular, when Algorithm 6 is applied to the computation of the prolongation

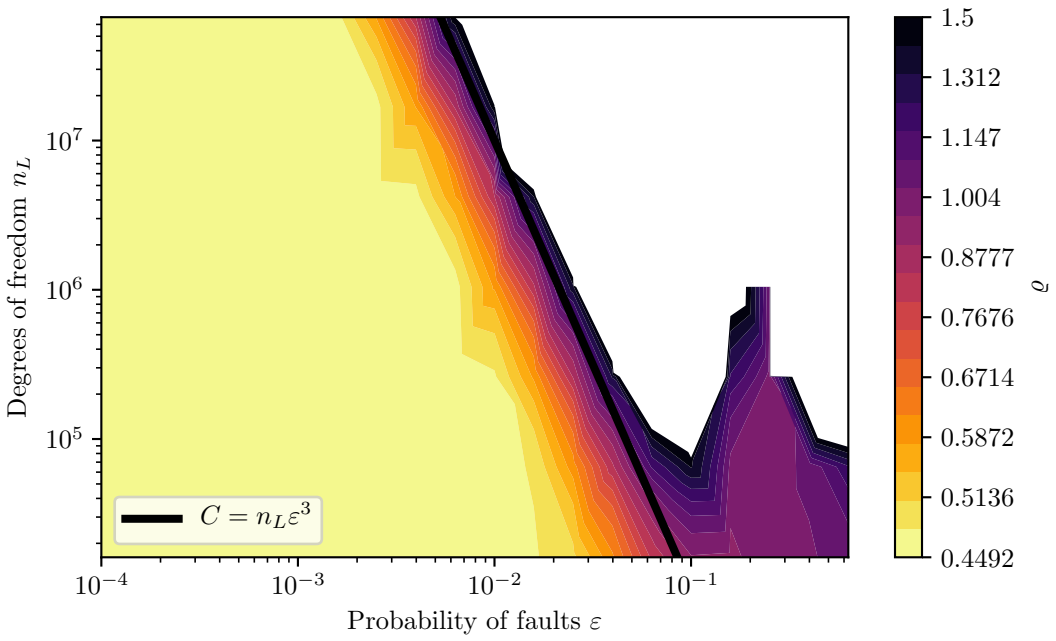


Figure 7.2: Lyapunov spectral radius $\varrho(\boldsymbol{\mathcal{E}}_L)$ for the iteration matrix for the Fault-Prone W-Cycle Multigrid Method with $K_P = 4$ replicas and acceptance threshold $k_P = 3$ in the prolongation and $K = 3$ replicas in all other operations.

with parameters $k_P = 3$ and $K_P = 4$, we obtain an overall rate of $\mathbb{P}(M_P) = \mathcal{O}(\varepsilon^2)$ at which mitigation occurs. Figure 7.2 shows the results obtained by applying Algorithm 6 with $k_P = 3$ and $K_P = 4$ to the prolongation, and Algorithm 5 with $K = 3$ to all other operations. It is seen that the strategy has resulted in the factor $\mathcal{O}\left(\sqrt{n_L\varepsilon^2}\right)$ in the Lyapunov spectral radius being replaced by $\mathcal{O}\left(n_L\varepsilon^\alpha\right)$ with $\alpha = 3$.

Would a cheaper detection and protection strategy (e.g. with $K_P = 3$, $k_P = 2$ and $K = 2$) suffice? Figure 7.3 shows the results obtained employing a variety of different choices for K, K_P and k_P. It is seen that the cheapest strategy that results in the growth $\mathcal{O}\left(\sqrt{n_L\varepsilon^2}\right)$ being replaced by $\mathcal{O}\left(n_L\varepsilon^\alpha\right)$, $\alpha > 2$, is indeed $K_P = 4$, $k_P = 3$ and $K = 3$.

Moreover, if one wishes to obtain a given rate $\alpha \geq 2$, then one can choose $K = k_P = \alpha$ and $K_P = 2\alpha - 2$. As problems in three spatial dimensions are more resilient to the effect of laissez-faire mitigation, as reflected by Theorem 12 and the

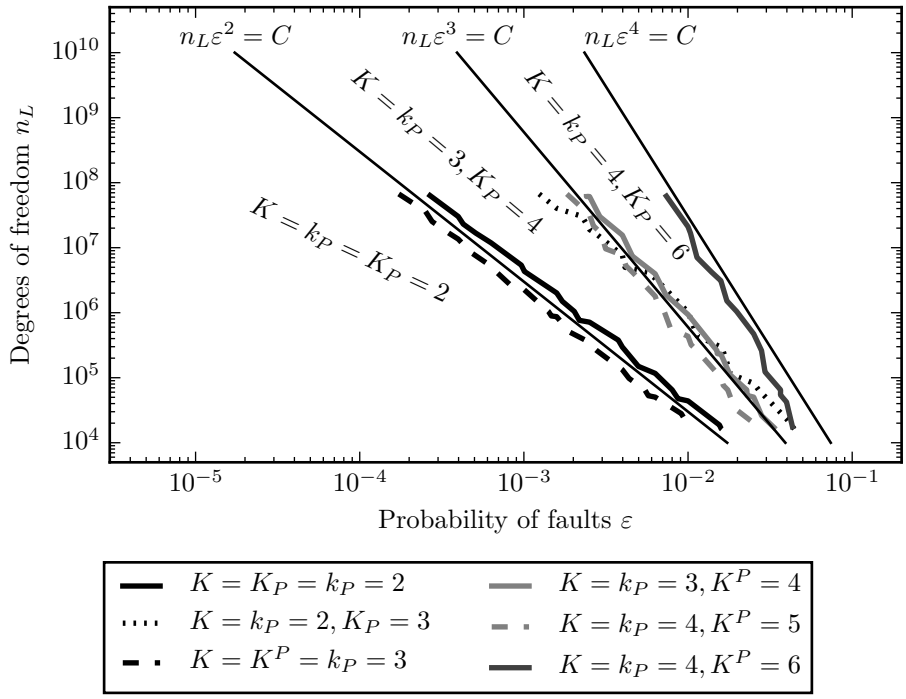


Figure 7.3: Level sets $\varrho(\boldsymbol{\mathcal{E}}_L) = 0.57$ (50% more iterations than the ideal fault-free case) for different levels of detection and protection in the case of the 2D Poisson problem. Optimal parameter combinations are marked in thick solid lines. The obtained optimal detection and protection parameters for each region in the ε-n_L-plane are marked in the plot.

numerical evidence in Chapter 6, we expect the strategy $K = K_P = k_P = \alpha$ to be sufficient for $2 \leq \alpha \leq 6$.

Chapter 8

Fault-Prone Domain Decomposition Methods

A second class of widely used iterative linear solvers are so-called domain decomposition methods. The main idea behind domain decomposition methods is to divide up the problem domain into several potentially overlapping subdomains, and solve the system restricted to every subdomain concurrently. Therefore, a domain decomposition method is similar in spirit to a block Jacobi smoothener. In typical applications, the subdomain size is chosen so that each compute node in the parallel computation can fit one subdomain problem. In order to obtain a scalable method, a coarse grid solve needs to be incorporated for global exchange of information. While multigrid methods achieve optimal complexity by coarsening the problem only mildly and using inexpensive solvers on each level, domain decomposition methods coarsen aggressively, so that typically only two levels are used. In turn, the considered problem on the different levels are significantly more expensive to solve [41, 88, 94].

Domain decomposition methods can be divided up into two categories: solvers without and with overlap of the subdomains. In this chapter, we introduce the later kind, and discuss an extension of the framework for the analysis of laissez-faire fault

mitigation to blockwise faults in one and two level domain decomposition solvers. In particular, we explain the effect on faults in unknowns that are shared by several subdomains.

8.1 Overlapping Domain Decomposition Methods

Let the polygonal domain $\Omega \subset \mathbb{R}^d$ be divided up into a shape regular, quasi-uniform partition $\mathcal{T}_h$ of simplices of size h. Partition $\mathcal{T}_h$ into N parts $\mathcal{T}_h^{(i,0)}$, $i = 1, \ldots, N$, corresponding to subdomains $\Omega^{(i,0)}$. The extended triangulations $T_h^{(i,\delta)}$, $\delta = 1, 2, \ldots$ are obtained by adding δ exterior layers of cells to $T_h^{(i,0)}$. We denote the corresponding overlapping subdomains by $\Omega^{(i,\delta)}$. Let $V := H_0^1(\Omega)$ and V_h be the space of piecewise linears on $\mathcal{T}_h$, and $V_h^{(i,\delta)}$ the space of piecewise linears on $\mathcal{T}_h^{(i,\delta)}$, with respective dimensions $n = \dim V_h$ and $n^{(i,\delta)} = \dim V_h^{(i,\delta)}$. Let ϕ_j the nodal basis functions of V_h, and let $\vec{\phi}$ and $\vec{\phi}^{(i,\delta)}$ be the vectors of basis functions in V_h and $V_h^{(i,\delta)}$ respectively, with index sets $\mathcal{I}$ and $\mathcal{I}^{(i,\delta)} \subset \mathcal{I}$. Moreover, let $z_j \in \Omega$ be the node corresponding to the basis function ϕ_j, $j \in \mathcal{I}$. The canonical injection

$$\mathcal{P}^{(i,\delta)} : V_h^{(i,\delta)} \hookrightarrow V_h$$

gives rise to the prolongation matrix $\boldsymbol{P}^{(i,\delta)}$ defined by the action of $\mathcal{P}^{(i,\delta)}$ on $u = \vec{x} \cdot \phi^{(i,\delta)} \in V_h^{(i,\delta)}$ as

$$\mathcal{P}^{(i,\delta)} u = \vec{x} \cdot \mathcal{P}^{(i,\delta)} \vec{\phi}^{(i,\delta)} = \left(\boldsymbol{P}^{(i,\delta)} \vec{x} \right) \cdot \vec{\phi}.$$

Moreover, define the restriction matrices $\boldsymbol{R}^{(i,\delta)} := \left(\boldsymbol{P}^{(i,\delta)} \right)^T$, and the corresponding operator $\mathcal{R}^{(i,\delta)}$.

Let a be a symmetric, continuous, bilinear form on $V \times V$ that is V-elliptic with constant η. For $u = \vec{x} \cdot \vec{\phi}$, $v = \vec{y} \cdot \vec{\phi} \in V_h$,

$$a(u, v) = \vec{x} \cdot a\left(\vec{\phi}, \vec{\phi}\right) \cdot \vec{y} =: \vec{x} \cdot \boldsymbol{A} \vec{y},$$

where $\boldsymbol{A}$ is the stiffness matrix. Similarly, we define the subdomain stiffness matrices $\boldsymbol{A}^{(i,\delta)} := a(\vec{\phi}^{(i,\delta)}, \vec{\phi}^{(i,\delta)})$. Hence, for $u = \vec{x} \cdot \vec{\phi}^{(i,\delta)}$, $v = \vec{y} \cdot \vec{\phi}^{(i,\delta)} \in V_h^{(i,\delta)}$

$$\begin{aligned}
\vec{x} \cdot \boldsymbol{A}^{(i,\delta)} \cdot \vec{y} &= a\left(\vec{x} \cdot \vec{\phi}^{(i,\delta)}, \vec{y} \cdot \vec{\phi}^{(i,\delta)}\right) \\
&= a\left(\left(\boldsymbol{P}^{(i,\delta)}\vec{x}\right) \cdot \vec{\phi}, \left(\boldsymbol{P}^{(i,\delta)}\vec{y}\right) \cdot \vec{\phi}\right) \\
&= \left(\boldsymbol{P}^{(i,\delta)}\vec{x}\right) \cdot a\left(\vec{\phi}, \vec{\phi}\right) \cdot \left(\boldsymbol{P}^{(i,\delta)}\vec{y}\right) \\
&= \vec{x} \cdot \boldsymbol{R}^{(i,\delta)}\boldsymbol{A}\boldsymbol{P}^{(i,\delta)} \cdot \vec{y}.
\end{aligned}$$

Therefore, $\boldsymbol{A}^{(i,\delta)} = \boldsymbol{R}^{(i,\delta)}\boldsymbol{A}\boldsymbol{P}^{(i,\delta)}$. We also set $\boldsymbol{A}^{(i,\delta,\delta+1)} := \boldsymbol{R}^{(i,\delta)}\boldsymbol{A}\boldsymbol{P}^{(i,\delta+1)}$ so that $\boldsymbol{R}^{(i,\delta)}\boldsymbol{A} = \boldsymbol{A}^{(i,\delta,\delta+1)}\boldsymbol{R}^{(i,\delta+1)}$.

Moreover, let a partition of unity be given by $\theta^{(i,\delta)} \in C(\Omega, [0,1])$ such that $\sum_i \theta^{(i,\delta)} = 1$ and $\operatorname{supp}\theta^{(i,\delta)} \subset \Omega^{(i,\delta)}$. The operator

$$\mathcal{D}^{(i,\delta)}(u) := \mathcal{I}_h\left(\theta^{(i,\delta)}u\right), \quad u \in V_h^{(i,\delta)}$$

has the matrix representation

$$\boldsymbol{D}^{(i,\delta)} = \operatorname{diag}\left(\theta^{(i,\delta)}(z_j)\right).$$

Then the discrete partition of unity property reads as

$$\sum_i \boldsymbol{P}^{(i,\delta)}\boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)} = \boldsymbol{I}_n.$$

A particular choice of such a partition function is $\theta^{(i,-1)}$ such that $\theta^{(i,-1)}(z_j) \in \{0,1\}$, which leads to an algebraically non-overlapping partition of the unknowns. We write the restriction and prolongation matrices and operators using the superscript $(i,-1)$.

For every subdomain i, we define its set of neighbors by

$$\mathcal{O}_i := \left\{j \neq i \mid \mathcal{I}^{(i,\delta)} \cap \mathcal{I}^{(j,\delta)} \neq \emptyset\right\}.$$

Lemma 22.

For $i = 1, \ldots, N$

$$\boldsymbol{P}^{(i,\delta)}\boldsymbol{A}^{(i,\delta)}\boldsymbol{D}^{(i,\delta)} = \boldsymbol{A}\boldsymbol{P}^{(i,\delta)}\boldsymbol{D}^{(i,\delta)}, \qquad \boldsymbol{D}^{(i,\delta)}\boldsymbol{A}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)} = \boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)}\boldsymbol{A}.$$

Proof.

$$\begin{aligned}
\boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)}\boldsymbol{A} - \boldsymbol{D}^{(i,\delta)}\boldsymbol{A}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)} &= \boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)} a\left(\vec{\phi},\phi\right) - \boldsymbol{D}^{(i,\delta)} a\left(\phi^{(i)},\phi^{(i)}\right)\boldsymbol{R}^{(i,\delta)} \\
&= a\left(\boldsymbol{D}^{(i,\delta)}\phi^{(i)},\phi\right) - a\left(\boldsymbol{D}^{(i,\delta)}\phi^{(i)},\left(\boldsymbol{P}^{(i,\delta)}\phi^{(i)}\right)\right) \\
&= a\left(\boldsymbol{D}^{(i,\delta)}\phi^{(i)},\left(\phi - \boldsymbol{P}^{(i,\delta)}\phi^{(i)}\right)\right) \\
&= \left\{a\left(\theta(z_j)\phi_j, \delta_{\left\{k\notin\mathcal{I}^{(i)}\right\}}\phi_k\right)\right\}_{j\in\mathcal{I}^{(i)},k\in\mathcal{I}} \\
&= \left\{\theta(z_j)\delta_{\left\{k\notin\mathcal{I}^{(i)}\right\}} a\left(\phi_j,\phi_k\right)\right\}_{j\in\mathcal{I}^{(i)},k\in\mathcal{I}}
\end{aligned}$$

Now, $a(\phi_j,\phi_k) \neq 0$ only if i and j belong to the same element. For $k \notin \mathcal{I}^{(i)}$, $a(\phi_j,\phi_k) \neq 0$ therefore implies that z_j must be on the boundary of subdomain $\Omega^{(i,\delta)}$. Since θ_i is a continuous function that is zero outside of $\Omega^{(i,\delta)}$, we have $\theta(z_j) = 0$ and therefore that $\boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)}\boldsymbol{A} = \boldsymbol{D}^{(i,\delta)}\boldsymbol{A}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)}$. The second equality is the transpose of the first. □

An immediate consequence of the lemma is that

$$\sum_i \boldsymbol{P}^{(i,\delta)}\boldsymbol{D}^{(i,\delta)}\boldsymbol{A}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)} = \sum_i \boldsymbol{P}^{(i,\delta)}\boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)}\boldsymbol{A} = \boldsymbol{A}.$$

We define the additive Schwarz preconditioner

$$\boldsymbol{N}_{AS} = \sum_i \boldsymbol{P}^{(i,\delta)}\left(\boldsymbol{A}^{(i,\delta)}\right)^{-1}\boldsymbol{R}^{(i,\delta)},$$

the additive Schwarz preconditioner with harmonic overlap [23]

$$\boldsymbol{N}_{ASH} = \sum_i \boldsymbol{P}^{(i,\delta)}\left(\boldsymbol{A}^{(i,\delta)}\right)^{-1}\boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)},$$

and the restricted additive Schwarz preconditioner [24]

$$\boldsymbol{N}_{RAS} = \sum_i \boldsymbol{P}^{(i,\delta)}\boldsymbol{D}^{(i,\delta)}\left(\boldsymbol{A}^{(i,\delta)}\right)^{-1}\boldsymbol{R}^{(i,\delta)}.$$

The corresponding iterative schemes $\mathbb{S}_\bullet$ are given by

$$\vec{x}^{k+1} = \mathbb{S}_{AS}\left(\vec{b},\vec{x}^k\right) := \vec{x}^k + \theta\boldsymbol{N}_{AS}\left(\vec{b} - \boldsymbol{A}\vec{x}^k\right),$$

$$\vec{x}^{k+1} = \mathbb{S}_{ASH}\left(\vec{b}, \vec{x}^k\right) := \vec{x}^k + \boldsymbol{N}_{ASH}\left(\vec{b} - \boldsymbol{A}\vec{x}^k\right),$$
$$\vec{x}^{k+1} = \mathbb{S}_{RAS}\left(\vec{b}, \vec{x}^k\right) := \vec{x}^k + \boldsymbol{N}_{RAS}\left(\vec{b} - \boldsymbol{A}\vec{x}^k\right)$$

in global form. The additive Schwarz iteration is only convergent if the damping parameter θ is small enough.

It is typically observed that the number of iterations of the above methods deteriorates with a growing number of subdomains. In order to obtain a scalable solver, the inclusion of a coarse grid is therefore often mandatory. Let $V_H \subset V_h$, and designate the canonical injection by

$$\mathcal{P}^0 : V_H \hookrightarrow V_h.$$

The associated matrix form is called $\boldsymbol{P}^{(0)}$, and the corresponding restriction is $\boldsymbol{R}^{(0)} = \left(\boldsymbol{P}^{(0)}\right)^T$. This permits to define a coarse grid system matrix $\boldsymbol{A}^{(0)}$ through the Galerkin relationship

$$\boldsymbol{A}^{(0)} = \boldsymbol{R}^{(0)}\boldsymbol{A}\boldsymbol{P}^{(0)},$$

and a coarse grid preconditioner and correction scheme as

$$\boldsymbol{N}_{CG} = \boldsymbol{P}^{(0)}\left(\boldsymbol{A}^{(0)}\right)^{-1}\boldsymbol{R}^{(0)},$$
$$\vec{x}^{k+1} = \mathbb{S}_{CG}\left(\vec{b}, \vec{x}^k\right) := \vec{x}^k + \boldsymbol{N}_{CG}\left(\vec{b} - \boldsymbol{A}\vec{x}^k\right).$$

Two grid methods can then be obtained by combining a domain decomposition iteration with the coarse grid correction, for example as

$$\vec{x}^{k+1} = \left[\mathbb{S}_{CG}\left(\vec{b}, \cdot\right) \circ \mathbb{S}_{ASH}\left(\vec{b}, \cdot\right)\right]\left(\vec{x}^k\right).$$

We will now turn to a detailed description of how faults will affect the convergence properties of one and two level domain decomposition solvers.

8.2 Modeling Blockwise Faults in Iterative Schemes

For the purposes of implementation, and to study the impact of faults, the local form of the above iterations is more helpful. We notice that faults of subdomain solves can not be modeled as before, since, due to the overlap, one unknown can be owned by several subdomains. This means that we cannot model faults by a random diagonal matrix with blockwise dependent diagonal entries.

The additive Schwarz iteration can be written as

$$\begin{aligned}\vec{x}_i^{k+1} &= \vec{x}_i^k + \boldsymbol{R}^{(i,\delta+1)}\boldsymbol{N}_{AS}\left(\vec{b} - \boldsymbol{A}\vec{x}^k\right) \\ &= \vec{x}_i^k + \sum_{j\in\mathcal{O}_i} \boldsymbol{R}^{(i,\delta+1)}\boldsymbol{P}^{(j,\delta)}\left(\boldsymbol{A}^{(j,\delta)}\right)^{-1}\left(\vec{b}_j - \boldsymbol{A}^{(i,\delta,\delta+1)}\vec{x}_i^k\right),\end{aligned}$$

where $\vec{x}_i^k = \boldsymbol{R}^{(i,\delta+1)}\vec{x}^k$ and $\vec{b}_i = \boldsymbol{R}^{(i,\delta)}\vec{b}$ are the node local part of the solution and the right-hand side respectively.

In practice, one subdomain will be assigned to one compute node. We can model the effect of failing compute nodes by Bernoulli distributed random variables

$$\chi^{(i)} = \begin{cases} 1 & \text{with probability } 1-\varepsilon, \\ 0 & \text{with probability } \varepsilon. \end{cases}$$

Here, the probability of a node failure ε is taken to be small. We rewrite the additive Schwarz scheme introduced above by introducing faults:

$$\vec{x}_i^{k+1} = \vec{x}_i^k + \sum_{j\in\mathcal{O}_i} \boldsymbol{R}^{(i,\delta+1)}\chi^{(j)}\boldsymbol{P}^{(j,\delta)}\left(\boldsymbol{A}^{(j,\delta)}\right)^{-1}\left(\vec{b}_j - \boldsymbol{A}^{(j,\delta,\delta+1)}\vec{x}_j^k\right).$$

Since all operations up to the communication step are node local, we only need to include one random variable per node to account for any failure on that node.

In order to analyze the convergence behavior, we derive the error propagator $\boldsymbol{\mathcal{E}}_{AS}$ associated with the error equation. The error $\vec{e}^{k+1}$ after $k+1$ iterations in terms of the error after k iterations by

$$\vec{e}^{k+1} = \vec{x}^{k+1} - \vec{x} = = \vec{e}^k + \sum_j \chi^{(j)}\boldsymbol{P}^{(j,\delta)}\left(\boldsymbol{A}^{(i,\delta)}\right)^{-1}\boldsymbol{R}^{(j,\delta)}\boldsymbol{A}\left(\vec{x} - \vec{x}^k\right)$$

$$= \left(\boldsymbol{I} - \sum_j \chi^{(j)} \boldsymbol{P}^{(j,\delta)} \left(\boldsymbol{A}^{(i,\delta)} \right)^{-1} \boldsymbol{R}^{(j,\delta)} \boldsymbol{A} \right) \vec{e}^{k}$$
$$=: \boldsymbol{\mathcal{E}}_{AS} \vec{e}^{k}.$$

The local version of the additive Schwarz method with harmonic overlap reads as

$$\begin{aligned} \vec{x}_i^{k+1} &= \vec{x}_i^k + \boldsymbol{R}^{(i,\delta)} \boldsymbol{N}_{ASH} \left(\vec{b} - \boldsymbol{A} \vec{x}^k \right) \\ &= \vec{x}_i^k + \sum_{j \in \mathcal{O}_i} \boldsymbol{R}^{(i,\delta)} \boldsymbol{P}^{(j,\delta)} \left(\boldsymbol{A}^{(i,\delta)} \right)^{-1} \boldsymbol{D}^{(j,\delta)} \left(\boldsymbol{R}^{(j,\delta)} \vec{b} - \boldsymbol{R}^{(j,\delta)} \boldsymbol{A} \vec{x}^k \right) \\ &= \vec{x}_i^k + \sum_{j \in \mathcal{O}_i} \boldsymbol{R}^{(i,\delta)} \boldsymbol{P}^{(j,\delta)} \left(\boldsymbol{A}^{(i,\delta)} \right)^{-1} \boldsymbol{D}^{(j,\delta)} \left(\vec{b}_j - \boldsymbol{A}^{(j,\delta)} \vec{x}_j^k \right) \end{aligned}$$

with $\vec{x}_i^k = \boldsymbol{R}^{(i,\delta)} \vec{x}^k$, where we used Lemma 22 in the last equality. This shows that again, only one random variable per node is necessary to capture all failures that that node might suffer in one iteration. Similar to the additive Schwarz method, we obtain the error propagator

$$\boldsymbol{\mathcal{E}}_{ASH} = \boldsymbol{I} - \sum_j \chi^{(j)} \boldsymbol{P}^{(j,\delta)} \left(\boldsymbol{A}^{(i,\delta)} \right)^{-1} \boldsymbol{D}^{(j,\delta)} \boldsymbol{R}^{(j,\delta)} \boldsymbol{A}.$$

The reduced additive Schwarz method

$$\begin{aligned} \vec{x}_i^{k+1} &= \vec{x}_i^k + \boldsymbol{R}^{(i,\delta)} \boldsymbol{N}_{RAS} \left(\vec{b} - \boldsymbol{A} \vec{x}^k \right) \\ &= \vec{x}_i^k + \sum_{j \in \mathcal{O}_i} \boldsymbol{R}^{(i,\delta)} \boldsymbol{P}^{(j,\delta)} \boldsymbol{D}^{(j,\delta)} \left(\boldsymbol{A}^{(i,\delta)} \right)^{-1} \left(\vec{b}_j - \sum_{l \in \mathcal{O}_j} \boldsymbol{R}^{(j,\delta)} \boldsymbol{P}^{(l,\delta)} \boldsymbol{D}^{(l,\delta)} \boldsymbol{A}^{(l,\delta)} \vec{x}_l^k \right), \end{aligned}$$

can be rewritten as

$$\begin{aligned} \vec{y}_i^0 &= x_i^0, \\ \vec{y}_i^{k+1} &= \vec{y}_i^k + \left(\boldsymbol{A}^{(i,\delta)} \right)^{-1} \left(\vec{b}_i - \sum_{j \in \mathcal{O}_i} \boldsymbol{R}^{(i,\delta)} \boldsymbol{P}^{(j,\delta)} \boldsymbol{A}^{(j,\delta)} \boldsymbol{D}^{(j,\delta)} \vec{y}_j^k \right), \\ \vec{x}_i^k &= \sum_{j \in \mathcal{O}_i} \boldsymbol{R}^{(i,\delta)} \boldsymbol{P}^{(j,\delta)} \boldsymbol{D}^{(j,\delta)} \vec{y}_j^k. \end{aligned}$$

which leads to

$$\boldsymbol{\mathcal{E}}_{RAS} = \boldsymbol{I} - \sum_j \chi^{(j)} \boldsymbol{P}^{(j,\delta)} \boldsymbol{D}^{(j,\delta)} \left(\boldsymbol{A}^{(i,\delta)}\right)^{-1} \boldsymbol{R}^{(j,\delta)} \boldsymbol{A}.$$

Finally, the coarse grid correction can be given as

$$\begin{aligned}
\vec{x}_i^{k+1} &= \vec{x}_i^k + \boldsymbol{R}^{(i,\delta)} \boldsymbol{N}_{CG} \left(\vec{b} - \boldsymbol{A}\vec{x}^k\right) \\
&= \vec{x}_i^k + \boldsymbol{R}^{(i,\delta)} \boldsymbol{P}^{(0)} \left(\boldsymbol{A}^{(0)}\right)^{-1} \sum_j \boldsymbol{R}^{(0)} \boldsymbol{P}^{(j,-1)} \left(\boldsymbol{R}^{(j,-1)} \vec{b} - \boldsymbol{R}^{(j,-1)} \boldsymbol{A} \vec{x}^k\right) \\
&= \vec{x}_i^k + \boldsymbol{R}^{(i,\delta)} \boldsymbol{P}^{(0)} \left(\boldsymbol{A}^{(0)}\right)^{-1} \sum_j \boldsymbol{R}^{(0)} \boldsymbol{P}^{(j,-1)} \left(\vec{b}_j - \boldsymbol{A}^{(j,-1)} \boldsymbol{R}^{(j,-1)} \boldsymbol{P}^{(j,\delta)} \vec{x}_j^k\right)
\end{aligned}$$

with $\vec{b}_j = \boldsymbol{R}^{(j,-1)}$. The error propagator of the coarse grid is given by

$$\boldsymbol{\mathcal{E}}_{CG} = \boldsymbol{I} - \boldsymbol{\mathcal{X}}^{(N)} \boldsymbol{N}_{CG} \sum_j \chi^{(j)} \boldsymbol{P}^{(j,-1)} \boldsymbol{A}^{(j,-1)} \boldsymbol{R}^{(j,-1)}.$$

Here, two sources of faults are possible: faults can occur during the computation of the residual, or during the coarse grid solve and update. The random variables $\chi^{(j)}$ model the former, whereas the diagonal fault matrix $\boldsymbol{\mathcal{X}}^{(N)}$ encodes the fault structure of the coarse grid solve. Since we chose an algebraically non-overlapping partition, we can write

$$\boldsymbol{\mathcal{E}}_{CG} = \boldsymbol{I} - \boldsymbol{\mathcal{X}}^{(N)} \boldsymbol{N}_{CG} \boldsymbol{\mathcal{X}}^{(\rho)} \boldsymbol{A},$$

where the diagonal faults matrix $\boldsymbol{\mathcal{X}}^{(\rho)}$ has block structure:

$$\begin{aligned}
\boldsymbol{\mathcal{X}}_{ii}^{(\rho)} &= \chi^{(j)} && \text{if } \theta^{(j,-1)}(z_i) = 1, \\
\boldsymbol{\mathcal{X}}_{ik}^{(\rho)} &= 0 && \text{if } i \neq k.
\end{aligned}$$

The corresponding error propagators for the classical fault free methods are obtained by replacing the random variables with identity, and are designated by $\boldsymbol{E}_{AS}$, $\boldsymbol{E}_{ASH}$, $\boldsymbol{E}_{RAS}$ and $\boldsymbol{E}_{CG}$ respectively.

8.3 Convergence Analysis for Domain Decomposition Methods

Our first result concerns the convergence behavior of the single level domain decomposition methods introduced above.

A standard assumption needed in the convergence analysis of fault-free domain decomposition methods concerns the interaction of the local solution spaces $V^{(i,\delta)}$:

Assumption 6 (Strengthened Cauchy-Schwarz inequality)*:* Define $0 \leq \epsilon_{ij} \leq 1$ to be the smallest constants such that

$$|a(u,v)| \leq \epsilon_{ij} a(u,u)^{1/2} a(v,v)^{1/2} \quad \forall u \in V^{(i,\delta)},\ v \in V^{(j,\delta)}.$$

If N_c denotes the smallest number of colors necessary to color the subdomains, then

$$\rho(\boldsymbol{\epsilon}) \leq N_c,$$

and N_c is independent of h and N.

Theorem 23.
Let $\boldsymbol{\mathcal{E}}_{DD} \in \{\boldsymbol{\mathcal{E}}_{AS}, \boldsymbol{\mathcal{E}}_{ASH}, \boldsymbol{\mathcal{E}}_{RAS}\}$, and let $\boldsymbol{E}_{DD}$ the iteration matrix of its fault-free equivalent. Provided Assumption 6 holds, the convergence rate of the fault-prone iterations is bounded as

$$\varrho(\boldsymbol{\mathcal{E}}_{DD}) \leq \|\boldsymbol{E}_{DD}\|_A + C\varepsilon,$$

where $\|\bullet\|_A$ is the energy norm and C is independent of ε, h and N.

Proof. For easy of notation, we define the extended matrices

$$\overline{\boldsymbol{A}} := \text{blockdiag}\left(\boldsymbol{A}^{(i,\delta)}\right), \qquad \overline{\boldsymbol{D}} := \text{blockdiag}\left(\boldsymbol{D}^{(i,\delta)}\right),$$
$$\overline{\boldsymbol{R}} := \text{vec}\left(\boldsymbol{R}^{(i,\delta)}\right), \qquad \overline{\boldsymbol{P}} := \text{vec}\left(\boldsymbol{P}^{(i,\delta)}\right),$$
$$\overline{\boldsymbol{\mathcal{X}}} := \text{blockdiag}\left(\chi^{(i)} \boldsymbol{I}_{n^{(i,\delta)}}\right),$$

and write

$$\begin{aligned}
\boldsymbol{E}_{AS} &= \boldsymbol{I} - \overline{\boldsymbol{PA}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}, & \boldsymbol{\mathcal{E}}_{AS} &= \boldsymbol{I} - \overline{\boldsymbol{P\mathcal{X}A}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}, \\
\boldsymbol{E}_{ASH} &= \boldsymbol{I} - \overline{\boldsymbol{PA}}^{-1}\overline{\boldsymbol{DR}}\boldsymbol{A}, & \boldsymbol{\mathcal{E}}_{ASH} &= \boldsymbol{I} - \overline{\boldsymbol{P\mathcal{X}A}}^{-1}\overline{\boldsymbol{DR}}\boldsymbol{A} \\
\boldsymbol{E}_{RAS} &= \boldsymbol{I} - \overline{\boldsymbol{PDA}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}, & \boldsymbol{\mathcal{E}}_{RAS} &= \boldsymbol{I} - \overline{\boldsymbol{P\mathcal{X}DA}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}.
\end{aligned}$$

Using the fact that $\overline{\boldsymbol{A}}^{1/2}$ and $\overline{\boldsymbol{\mathcal{X}}}$ commute, we find

$$\begin{aligned}
\left\|\mathbb{V}\left[\boldsymbol{\mathcal{E}}_{AS}\right]\right\|_A &= \left\|\overline{\boldsymbol{P}}^{\otimes 2}\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\left(\overline{\boldsymbol{A}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}\right)^{\otimes 2}\right\|_A \\
&= \left\|\left(\overline{\boldsymbol{A}}^{1/2}\boldsymbol{P}\right)^{\otimes 2}\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\left(\overline{\boldsymbol{A}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}^{1/2}\right)^{\otimes 2}\right\|_2 \\
&= \left\|\left(\overline{\boldsymbol{A}}^{1/2}\boldsymbol{P}\overline{\boldsymbol{A}}^{-1/2}\right)^{\otimes 2}\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\left(\overline{\boldsymbol{A}}^{-1/2}\overline{\boldsymbol{R}}\boldsymbol{A}^{1/2}\right)^{\otimes 2}\right\|_2 \\
&\leq \left\|\overline{\boldsymbol{A}}^{1/2}\boldsymbol{P}\overline{\boldsymbol{A}}^{-1/2}\right\|_2^2 \left\|\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\right\|_2 \left\|\overline{\boldsymbol{A}}^{-1/2}\overline{\boldsymbol{R}}\boldsymbol{A}^{1/2}\right\|_2^2 \\
&\leq C\varepsilon\rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right)^2.
\end{aligned}$$

Similarly, we find

$$\begin{aligned}
\left\|\mathbb{V}\left[\boldsymbol{\mathcal{E}}_{ASH}\right]\right\|_A &= \left\|\overline{\boldsymbol{P}}^{\otimes 2}\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\left(\overline{\boldsymbol{A}}^{-1}\overline{\boldsymbol{DR}}\boldsymbol{A}\right)^{\otimes 2}\right\|_A \\
&= \left\|\left(\overline{\boldsymbol{PA}}^{-1/2}\right)^{\otimes 2}\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\left(\overline{\boldsymbol{A}}^{-1/2}\overline{\boldsymbol{DR}}\boldsymbol{A}\right)^{\otimes 2}\right\|_A \\
&= \left\|\left(\boldsymbol{A}^{1/2}\overline{\boldsymbol{PA}}^{-1/2}\right)^{\otimes 2}\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\left(\overline{\boldsymbol{A}}^{-1/2}\overline{\boldsymbol{DR}}\boldsymbol{A}^{1/2}\right)^{\otimes 2}\right\|_2 \\
&\leq \left\|\boldsymbol{A}^{1/2}\overline{\boldsymbol{PA}}^{-1/2}\right\|_2^2 \left\|\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\right\|_2 \left\|\overline{\boldsymbol{A}}^{-1/2}\overline{\boldsymbol{DR}}\boldsymbol{A}^{1/2}\right\|_2^2 \\
&= \rho\left(\overline{\boldsymbol{PA}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}\right)\left\|\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\right\|_2 \rho\left(\boldsymbol{A}^{1/2}\overline{\boldsymbol{PDA}}^{-1}\overline{\boldsymbol{DR}}\boldsymbol{A}^{1/2}\right) \\
&= \rho\left(\overline{\boldsymbol{PA}}^{-1}\overline{\boldsymbol{R}}\boldsymbol{A}\right)\left\|\mathbb{V}\left[\overline{\boldsymbol{\mathcal{X}}}\right]\right\|_2 \rho\left(\overline{\boldsymbol{PDA}}^{-1}\overline{\boldsymbol{DR}}\boldsymbol{A}\right) \\
&\leq C\varepsilon\rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right)\rho\left(\boldsymbol{N}_{RASH}\boldsymbol{A}\right)
\end{aligned}$$

where we have set

$$\boldsymbol{N}_{RASH} = \overline{\boldsymbol{PDA}}^{-1}\overline{\boldsymbol{DR}} = \sum_i \boldsymbol{P}^{(i,\delta)}\boldsymbol{D}^{(i,\delta)}\left(\boldsymbol{A}^{(i,\delta)}\right)^{-1}\boldsymbol{D}^{(i,\delta)}\boldsymbol{R}^{(i,\delta)}.$$

It can be shown that the same holds for the reduced additive Schwarz method. Therefore we have shown that

$$\begin{aligned}
\varrho\left(\boldsymbol{\mathcal{E}}_{AS}\right) &\leq \sqrt{\left(\left\|\boldsymbol{E}_{AS}\right\|_A + \varepsilon\left\|\boldsymbol{N}_{AS}\boldsymbol{A}\right\|_A\right)^2 + C\varepsilon\rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right)^2},\\
\varrho\left(\boldsymbol{\mathcal{E}}_{ASH}\right) &\leq \sqrt{\left(\left\|\boldsymbol{E}_{ASH}\right\|_A + \varepsilon\left\|\boldsymbol{N}_{ASH}\boldsymbol{A}\right\|_A\right)^2 + C\varepsilon\rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right)\rho\left(\boldsymbol{N}_{RASH}\boldsymbol{A}\right)},\\
\varrho\left(\boldsymbol{\mathcal{E}}_{RAS}\right) &\leq \sqrt{\left(\left\|\boldsymbol{E}_{RAS}\right\|_A + \varepsilon\left\|\boldsymbol{N}_{RAS}\boldsymbol{A}\right\|_A\right)^2 + C\varepsilon\rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right)\rho\left(\boldsymbol{N}_{RASH}\boldsymbol{A}\right)}.
\end{aligned}$$

We conclude using Lemmas 24 and 25 below. □

Lemma 24.

The partition of unity $\theta^{(i,\delta)}$ can be chosen such that $\left\|\nabla\theta^{(i,\delta)}\right\|_{L^\infty} \leq C/(\delta h)$. Then, for all $u \in V_h^{(i,\delta)}$,

$$a\left(\mathcal{D}^{(i,\delta)}u, \mathcal{D}^{(i,\delta)}u\right) \leq C_D a(u,u)$$

for some constant C_D.

Proof. Set $B^{(i,\delta)} := \Omega^{(i,\delta)} \cap \left\{\theta^{(i,\delta)} < 1\right\}$. For $u \in V_h^{(i,\delta)}$, we have that

$$\begin{aligned}
a\left(\mathcal{D}^{(i,\delta)}u, \mathcal{D}^{(i,\delta)}u\right) &\leq C\left|\mathcal{D}^{(i,\delta)}u\right|^2_{H^1\left(\Omega^{(i,\delta)}\right)}\\
&= C\left(|u|^2_{H^1\left(\Omega^{(i,\delta)}\setminus B^{(i,\delta)}\right)} + \left|\mathcal{I}_h\left[\theta^{(i,\delta)}u\right]\right|^2_{H^1\left(B^{(i,\delta)}\right)}\right).
\end{aligned}$$

The second term is estimated as

$$\begin{aligned}
\left|\mathcal{I}_h\left[\theta^{(i,\delta)}u\right]\right|_{H^1\left(B^{(i,\delta)}\right)} &\leq \left|\theta^{(i,\delta)}u\right|_{H^1\left(B^{(i,\delta)}\right)} + \left|\theta^{(i,\delta)}u - \mathcal{I}_h\left[\theta^{(i,\delta)}u\right]\right|_{H^1\left(B^{(i,\delta)}\right)}\\
&\leq C\left|\theta^{(i,\delta)}u\right|_{H^1\left(B^{(i,\delta)}\right)}.
\end{aligned}$$

Now

$$\begin{aligned}
\left|\theta^{(i,\delta)}u\right|^2_{H^1\left(B^{(i,\delta)}\right)} &\leq 2\int_{B^{(i,\delta)}}\left(\theta^{(i,\delta)}\right)^2|\nabla u|^2 + 2\int_{B^{(i,\delta)}}u^2\left|\nabla\theta^{(i,\delta)}\right|^2\\
&\leq 2|u|^2_{H^1\left(B^{(i,\delta)}\right)} + 2\left\|\nabla\theta^{(i,\delta)}\right\|^2_{L^\infty}\|u\|^2_{L^2\left(B^{(i,\delta)}\right)}. \qquad (8.1)
\end{aligned}$$

Application of a Poincare inequality then leads to

$$\left|\theta^{(i,\delta)}u\right|^2_{H^1\left(B^{(i,\delta)}\right)} \leq 2|u|^2_{H^1\left(B^{(i,\delta)}\right)} + 2C\left\|\nabla\theta^{(i,\delta)}\right\|_{L^\infty}(\delta h)^2|u|^2_{H^1\left(B^{(i,\delta)}\right)}$$

$$\leq 2\left|u\right|^2_{H^1\left(B^{(i,\delta)}\right)} + 2C\left|u\right|^2_{H^1\left(B^{(i,\delta)}\right)}.$$

Therefore, we have shown

$$\begin{aligned} a\left(\mathcal{D}^{(i,\delta)}u, \mathcal{D}^{(i,\delta)}u\right) &\leq C\left|u\right|^2_{H^1\left(\Omega^{(i,\delta)}\right)} \\ &\leq Ca(u,u). \end{aligned}$$

□

Lemma 25.

$$\begin{aligned} \rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right) &= \left\|\boldsymbol{N}_{AS}\boldsymbol{A}\right\|_A \leq N_c, \\ \rho\left(\boldsymbol{N}_{ASH}\boldsymbol{A}\right) &\leq \left\|\boldsymbol{N}_{ASH}\boldsymbol{A}\right\|_A \leq N_c C_D^{1/2}, \\ \rho\left(\boldsymbol{N}_{RAS}\boldsymbol{A}\right) &\leq \left\|\boldsymbol{N}_{RAS}\boldsymbol{A}\right\|_A \leq N_c C_D^{1/2}, \\ \rho\left(\boldsymbol{N}_{RASH}\boldsymbol{A}\right) &\leq \left\|\boldsymbol{N}_{RASH}\boldsymbol{A}\right\|_A \leq N_c C_D. \end{aligned}$$

Proof. The proof of the first inequality is classical and makes use of the operator form $\mathcal{T}_{AS}$ of $\boldsymbol{N}_{AS}\boldsymbol{A}$. This operator as well as the ones for $\boldsymbol{N}_{ASH}\boldsymbol{A}$, $\boldsymbol{N}_{RAS}\boldsymbol{A}$ and $\boldsymbol{N}_{RASH}\boldsymbol{A}$ are given by

$$\begin{aligned} \mathcal{T}_{AS} &= \sum_i \mathcal{T}^{(i,\delta)}, \\ \mathcal{T}_{ASH} &= \sum_i \tilde{\mathcal{T}}^{(i,\delta)}, \\ \mathcal{T}_{RAS} &= \sum_i \mathcal{D}^{(i,\delta)}\mathcal{T}^{(i,\delta)}, \\ \mathcal{T}_{RASH} &= \sum_i \mathcal{D}^{(i,\delta)}\tilde{\mathcal{T}}^{(i,\delta)}, \end{aligned}$$

where the subdomain solution operators $\mathcal{T}^{(i,\delta)}$ and $\tilde{\mathcal{T}}^{(i,\delta)}$ are defined by

$$\begin{aligned} a\left(\mathcal{T}^{(i,\delta)}u, v\right) &= a\left(u, v\right) && \forall u \in V_h,\ v \in V_h^{(i,\delta)}, \\ a\left(\tilde{\mathcal{T}}^{(i,\delta)}u, v\right) &= a\left(u, \mathcal{D}^{(i,\delta)}v\right) && \forall u \in V_h,\ v \in V_h^{(i,\delta)}. \end{aligned}$$

We then have

$$\begin{aligned}\rho\left(\boldsymbol{N}_\bullet \boldsymbol{A}\right)^2 &\leq \|\boldsymbol{N}_\bullet \boldsymbol{A}\|_A^2 \\ &= \sup_{\vec{x}} \frac{(\boldsymbol{N}_\bullet \boldsymbol{A}\vec{x}) \cdot (\boldsymbol{A}\boldsymbol{N}_\bullet \boldsymbol{A}\vec{x})}{\vec{x} \cdot \boldsymbol{A}\vec{x}} \\ &= \sup_{u \in V_h} \frac{a(\mathcal{T}_\bullet u, \mathcal{T}_\bullet u)}{a(u,u)}.\end{aligned}$$

For $\mathcal{T}_\bullet = \sum_i \mathcal{T}_\bullet^{(i)}$, we have by the strengthened Cauchy-Schwarz inequality that

$$\begin{aligned}a\left(\mathcal{T}_\bullet u, \mathcal{T}_\bullet u\right) &= \sum_{i,j} a\left(\mathcal{T}_\bullet^{(i)} u, \mathcal{T}_\bullet^{(j)} u\right) \\ &\leq \sum_{i,j} \epsilon_{ij} a\left(\mathcal{T}_\bullet^{(i)} u, \mathcal{T}_\bullet^{(i)} u\right)^{1/2} a\left(\mathcal{T}_\bullet^{(j)} u, \mathcal{T}_\bullet^{(j)} u\right)^{1/2} \\ &\leq \rho(\boldsymbol{\epsilon}) \sum_i a\left(\mathcal{T}_\bullet^{(i)} u, \mathcal{T}_\bullet^{(i)} u\right).\end{aligned}$$

For $\mathcal{T}_\bullet = \mathcal{T}_{AS}$ and $\mathcal{T}_\bullet^{(i)} = \mathcal{T}^{(i,\delta)}$, we can use that $\mathcal{T}^{(i,\delta)}$ are idempotent and self-adjoint with respect to $a(\cdot,\cdot)$ and obtain

$$\begin{aligned}a\left(\mathcal{T}_{AS} u, \mathcal{T}_{AS} u\right) &\leq \rho(\boldsymbol{\epsilon}) \sum_i a\left(u, \mathcal{T}^{(i,\delta)} u\right) \\ &= \rho(\boldsymbol{\epsilon}) a\left(u, \mathcal{T}_{AS} u\right) \\ &\leq \rho(\boldsymbol{\epsilon}) a\left(u,u\right)^{1/2} a\left(\mathcal{T}_{AS} u, \mathcal{T}_{AS} u\right)^{1/2}.\end{aligned}$$

Therefore $\rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right) \leq \rho(\boldsymbol{\epsilon})$.

Similarly, for $\mathcal{T}_\bullet = \mathcal{T}_{RAS}$ and $\mathcal{T}_\bullet^{(i)} = \mathcal{D}^{(i,\delta)}\mathcal{T}^{(i,\delta)}$, we have by Lemma 24

$$\begin{aligned}a\left(\mathcal{T}_{RAS} u, \mathcal{T}_{RAS} u\right) &\leq \rho(\boldsymbol{\epsilon}) \sum_i a\left(\mathcal{D}^{(i,\delta)}\mathcal{T}^{(i,\delta)} u, \mathcal{D}^{(i,\delta)}\mathcal{T}^{(i,\delta)} u\right) \\ &\leq \rho(\boldsymbol{\epsilon}) C_D \sum_i a\left(\mathcal{T}^{(i,\delta)} u, \mathcal{T}^{(i,\delta)} u\right) \\ &\leq \rho(\boldsymbol{\epsilon}) C_D a\left(u,u\right)^{1/2} a\left(\mathcal{T}_{AS} u, \mathcal{T}_{AS} u\right)^{1/2} \\ &\leq \rho(\boldsymbol{\epsilon})^2 C_D a\left(u,u\right)\end{aligned}$$

which permits to conclude that $\rho\left(\boldsymbol{N}_{RAS}\boldsymbol{A}\right) \leq \rho(\boldsymbol{\epsilon}) C_D^{1/2}$.

For $\mathcal{T}_\bullet = \mathcal{T}_{RASH}$ and $\mathcal{T}_\bullet^{(i)} = \mathcal{D}^{(i,\delta)}\tilde{\mathcal{T}}^{(i,\delta)}$, we find

$$\begin{aligned} a\left(\mathcal{T}_{RASH}u, \mathcal{T}_{RASH}u\right) &\leq \rho(\boldsymbol{\epsilon}) \sum_i a\left(\mathcal{D}^{(i,\delta)}\tilde{\mathcal{T}}^{(i,\delta)}u, \mathcal{D}^{(i,\delta)}\tilde{\mathcal{T}}^{(i,\delta)}u\right) \\ &\leq \rho(\boldsymbol{\epsilon}) C_D \sum_i a\left(\tilde{\mathcal{T}}^{(i,\delta)}u, \tilde{\mathcal{T}}^{(i,\delta)}u\right) \\ &= \rho(\boldsymbol{\epsilon}) C_D \sum_i a\left(u, \mathcal{D}^{(i,\delta)}\tilde{\mathcal{T}}^{(i,\delta)}u\right) \\ &= \rho(\boldsymbol{\epsilon}) C_D a\left(u, \mathcal{T}_{RASH}u\right) \\ &\leq \rho(\boldsymbol{\epsilon}) C_D a\left(u,u\right)^{1/2} a\left(\mathcal{T}_{RASH}u, \mathcal{T}_{RASH}u\right)^{1/2} \end{aligned}$$

and hence $\rho\left(\boldsymbol{N}_{RASH}\boldsymbol{A}\right) \leq \rho\left(\boldsymbol{\epsilon}\right) C_D$.

Finally, for $\mathcal{T}_\bullet = \mathcal{T}_{ASH}$ and $\mathcal{T}_\bullet^{(i)} = \tilde{\mathcal{T}}^{(i,\delta)}$,

$$\begin{aligned} a\left(\mathcal{T}_{ASH}u, \mathcal{T}_{ASH}u\right) &\leq \rho(\boldsymbol{\epsilon}) \sum_i a\left(\tilde{\mathcal{T}}^{(i,\delta)}u, \tilde{\mathcal{T}}^{(i,\delta)}u\right) \\ &\leq \rho(\boldsymbol{\epsilon}) \sum_i a\left(u, \mathcal{D}^{(i,\delta)}\tilde{\mathcal{T}}^{(i,\delta)}u\right) \\ &= \rho(\boldsymbol{\epsilon}) a\left(u, \mathcal{T}_{RASH}u\right) \\ &\leq \rho(\boldsymbol{\epsilon}) a\left(u,u\right)^{1/2} a\left(\mathcal{T}_{TASH}u, \mathcal{T}_{TASH}u\right)^{1/2} \\ &\leq \rho(\boldsymbol{\epsilon})^2 C_D a\left(u,u\right), \end{aligned}$$

where the previous inequality was used. Therefore $\rho\left(\boldsymbol{N}_{RASH}\boldsymbol{A}\right) \leq \rho\left(\boldsymbol{\epsilon}\right) C_D$. □

In order to verify the above result, we consider the following test case:

$$\text{Find } u \in V_h : \int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v \quad \forall v \in V_h,$$

where $\Omega = [0,1]^2$, V_h the space of piecewise linears on a uniform triangular mesh and $f \in L^2\left(\Omega\right)$. The resulting linear system is solved using the additive Schwarz method with harmonic overlap on $N = 16$ subdomains for varying mesh sizes h and fault rates ε. We iterate for 1000 iterations, as to get a good estimate of the asymptotic convergence rate, and renormalize the residual between iterations. The resulting approximation of $\varrho\left(\boldsymbol{\mathcal{E}}_{ASH}\right)$ is shown in Figure 8.1. It can be observed that

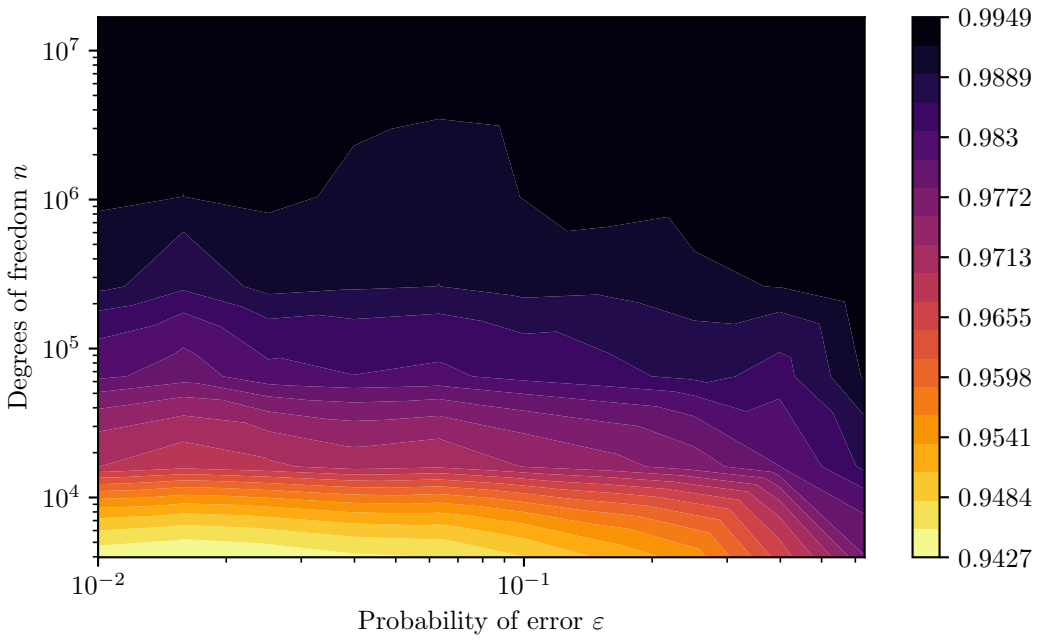


Figure 8.1: Asymptotic convergence rate $\varrho(\boldsymbol{\mathcal{E}}_{ASH})$ for $N = 16$ subdomains.

the convergence rate varies mainly in the dependence on the number of degrees of freedom n, but is largely unaffected by the fault rate ε.

If the number of subdomains N gets scaled proportionally to the number of degrees of freedom n, similar behavior is observed. (See Figure 8.2.) In both cases, the method stays convergent even for considerably large fault rates ε, and essentially mirror their fault free equivalent. This shows that the estimate in Theorem 23 is sharp.

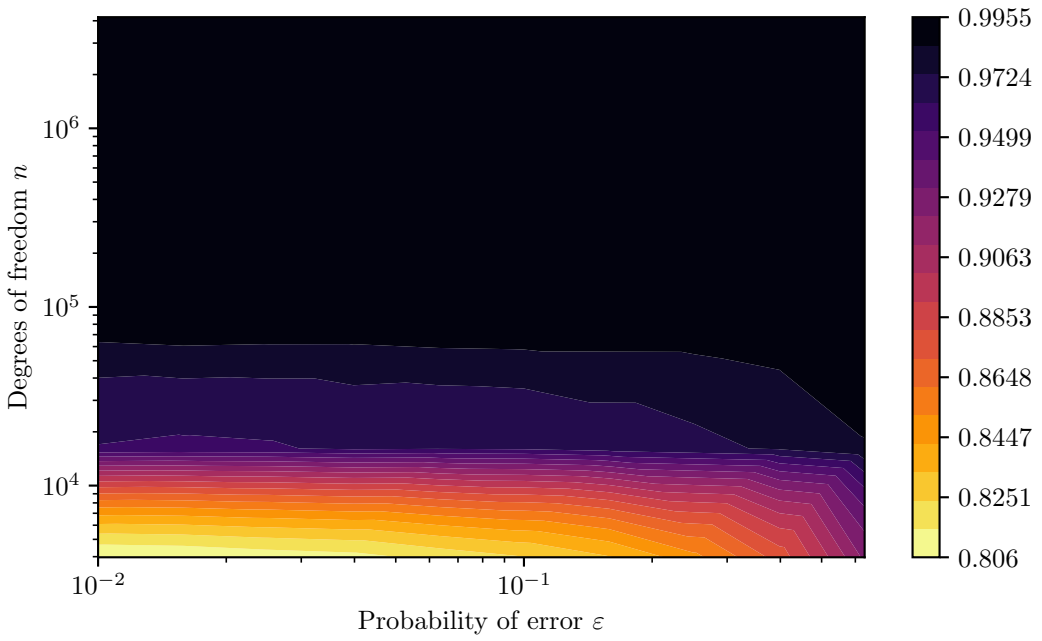


Figure 8.2: Asymptotic convergence rate $\varrho(\boldsymbol{\mathcal{E}}_{ASH})$. The number of subdomains N is proportional to the number of degrees of freedom n.

8.4 Two-Level Methods

We now turn to the case when a coarse grid is incorporated into the solver. It has already been established in Chapters 5 and 6 for multigrid solvers that in order to obtain a fault-resilient method, protection of the prolongation from the coarse grid to the fine grid is mandatory. For the same test problem as above, we numerically verify that this also the case for two level domain decomposition solvers. (See Figures 8.3 and 8.4.) Obviously, the single node case does not require protection of the prolongation, since a fault in the coarse solve simply leads to a skipped coarse grid correction. If the number of subdomains is fixed to $N = 16$, divergence can be observed that for large values of ε. The dependence on the number of degrees of freedom seems to be very mild. For the more realistic case of $N \sim n$ seen in Figure 8.4, we observe however that the problem size for which the two level method is convergent scales with $\varepsilon n^{1/2}$. This behavior mirrors what was derived in

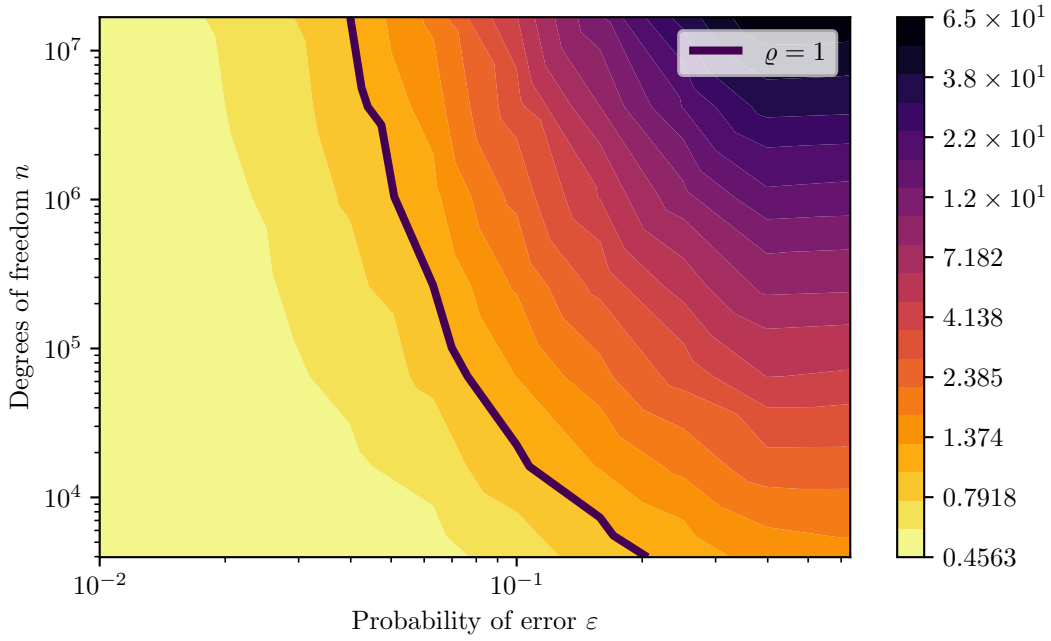


Figure 8.3: Asymptotic convergence rate $\varrho\left(\boldsymbol{\mathcal{E}}_{ASH}\boldsymbol{\mathcal{E}}_{CG}\right)$ without protected prolongation for $N = 16$ subdomains.

Chapter 6 for the multigrid case.

The next result concerns the variance of the coarse grid solver if the prolongation is protected.

Lemma 26.

Take η_{ij} to be the smallest constants such that

$$\left|\vec{x} \cdot \boldsymbol{A}^{-1}\vec{y}\right| \leq \eta_{ij} \left|\vec{x} \cdot \boldsymbol{A}^{-1}\vec{x}\right|^{1/2} \left|\vec{y} \cdot \boldsymbol{A}^{-1}\vec{y}\right|^{1/2}$$

for all $\vec{x} \in \operatorname{range} \boldsymbol{P}^{(j,-1)}$, $\vec{y} \in \operatorname{range} \boldsymbol{P}^{(i,-1)}$. Then, if prolongation is protected, we can bound

$$\left\|\mathbb{V}\left[\boldsymbol{\mathcal{E}}_{CG}\right]\right\|_A \leq C\varepsilon h^{-1} \sqrt{\max_i \sum_j \eta_{ij}^2} \leq C\varepsilon h^{-1}\sqrt{N}, \tag{8.2}$$

where the tensor energy norm is given by $\|\bullet\|_A := \left\| \left(\boldsymbol{A}^{1/2}\right)^{\otimes 2} \bullet \left(\boldsymbol{A}^{-1/2}\right)^{\otimes 2} \right\|_2$.

The proof of Lemma 26 requires several definitions and estimates that concern the block structure of the variance of the fault matrix.

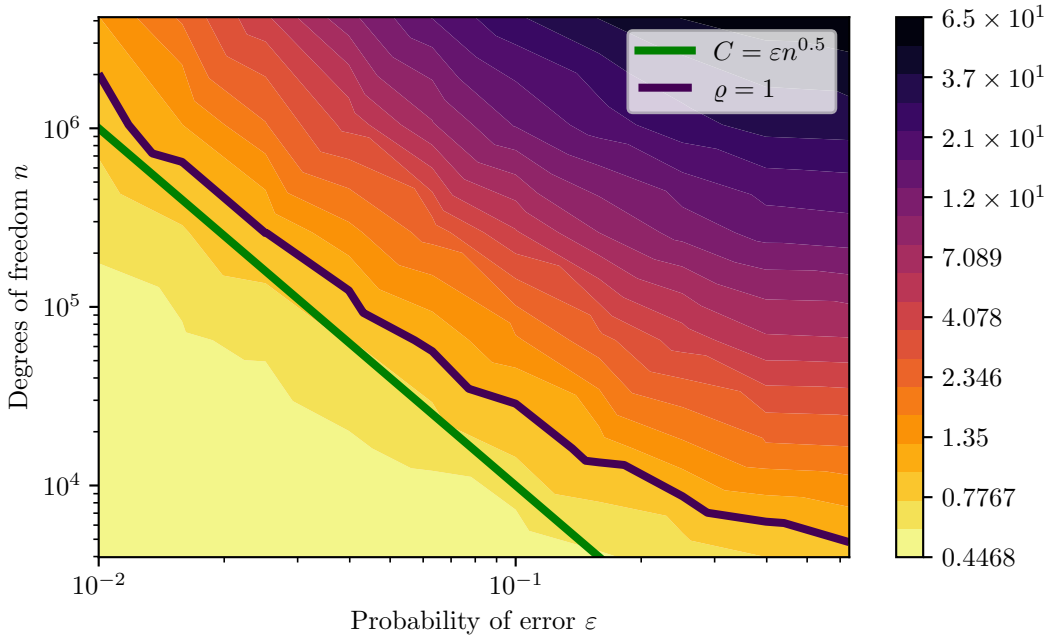


Figure 8.4: Asymptotic convergence rate $\varrho\left(\boldsymbol{\mathcal{E}}_{ASH}\boldsymbol{\mathcal{E}}_{CG}\right)$ without protected prolongation for $N \sim n$.

The variance of $\boldsymbol{\mathcal{X}}^{(\rho)}$ is a diagonal matrix as well, and is given by

$$\mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right]_{in+j,in+j} = \begin{cases} 1 & \text{if } \theta^{(i,-1)}(x_j) = 1 \\ 0 & \text{else} \end{cases}.$$

After reordering, it is a block diagonal matrix with identity in the blocks corresponding to interactions within a node, and zero blocks elsewhere. Before we introduce the notation for the treatment of block matrices, we show the

Example 4.

Consider the case of two subdomains, where the degrees of freedom of the first subdomain are ordered first, and then the unknowns of the second partition. We then have

for the variance of $\boldsymbol{\mathcal{X}}^{(\rho)}$ *the form*

$$\mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right] = \varepsilon(1-\varepsilon)\begin{pmatrix} \boldsymbol{I} & & & \\ & \mathbf{0} & & \\ & & \mathbf{0} & \\ & & & \boldsymbol{I} \end{pmatrix}.$$

If a matrix $\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_{11} & \boldsymbol{Z}_{12} \\ \boldsymbol{Z}_{21} & \boldsymbol{Z}_{22} \end{pmatrix} \in \mathbb{R}^{n\times n}$ *is block partitioned with respect to the sub-domains, we have*

$$\begin{aligned}\rho\left(\mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right]\boldsymbol{Z}^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}^{(\rho)}\right]\right) &= \varepsilon^2(1-\varepsilon)^2\rho\left[\begin{pmatrix} \boldsymbol{Z}_{11}^{\otimes 2} & & & \boldsymbol{Z}_{12}^{\otimes 2} \\ & \mathbf{0} & & \\ & & \mathbf{0} & \\ \boldsymbol{Z}_{21}^{\otimes 2} & & & \boldsymbol{Z}_{22}^{\otimes 2} \end{pmatrix}\right] \\ &= \varepsilon^2(1-\varepsilon)^2\rho\left[\begin{pmatrix} \boldsymbol{Z}_{11}^{\otimes 2} & \boldsymbol{Z}_{12}^{\otimes 2} \\ \boldsymbol{Z}_{21}^{\otimes 2} & \boldsymbol{Z}_{22}^{\otimes 2} \end{pmatrix}\right].\end{aligned}$$

The last block matrix is called the Khatri-Rao square power of $\boldsymbol{Z}$*, and we are interested in estimating its spectral radius.*

Definition 27 (Blocks of matrices)**.**
Given ordered vectors of indices $1 \le \alpha_1 < \alpha_2 < \cdots < \alpha_{\ell(\alpha)} \le m$*,* $1 \le \beta_1 < \beta_2 < \cdots < \beta_{\ell(\beta)} \le n$*, and a matrix* $\boldsymbol{A} \in \mathbb{R}^{m\times n}$*, we designate the sub-matrix corresponding to* α *and* β *by*

$$\boldsymbol{A}[\alpha,\beta] = \left\{\boldsymbol{A}_{\alpha_i,\beta_j}\right\}_{i,j}.$$

If $M = N$ *and* $\alpha = \beta$*, we write*

$$\boldsymbol{A}[\alpha] := \boldsymbol{A}[\alpha,\alpha].$$

We write the length of the index vector α *as* $\ell(\alpha)$*, so that* $\boldsymbol{A}[\alpha,\beta]$ *has size* $\ell(\alpha)\times\ell(\beta)$*.*

Definition 28 (Partitioned matrix)**.**
Let $\mathcal{A}$ and $\mathcal{B}$ be sets of index vectors as above, with $N_\mathcal{A} = |\mathcal{A}|$ and $N_\mathcal{B} = |\mathcal{B}|$ the respective number of blocks, such that $\bigcup_{\alpha\in\mathcal{A}} \alpha = \{1,\dots,m\}$ with $\alpha \cap \gamma = \emptyset$ for $\alpha \neq \gamma$, and $\bigcup_{\beta\in\mathcal{B}} \beta = \{1,\dots,n\}$ with $\beta \cap \eta = \emptyset$ for $\beta \neq \eta$. We call $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ partitioned by $\mathcal{A}\times\mathcal{B}$ into the blocks $\boldsymbol{A}[\alpha,\beta]$, $\alpha \in \mathcal{A},\ \beta \in \mathcal{B}$. If we prescribe an ordering of $\mathcal{A} = \{\alpha_i\}_{i=1}^{N_\mathcal{A}}$ and of $\mathcal{B} = \{\beta_j\}_{j=1}^{N_\mathcal{B}}$, we can define the block-ordered permutation $\overline{\boldsymbol{A}}$ of $\boldsymbol{A}$:

$$\overline{\boldsymbol{A}}_{ij} = \boldsymbol{A}[\alpha_i,\beta_j].$$

Definition 29 (Khatri-Rao product)**.**
Let $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{m\times n}$ be compatibly partitioned by $\mathcal{A}\times\mathcal{B}$. The Khatri-Rao product $\boldsymbol{A} * \boldsymbol{B}$ *is defined block-wise as*

$$(\boldsymbol{A} * \boldsymbol{B})_{ij} = \boldsymbol{A}[\alpha_i,\beta_j] \otimes \boldsymbol{B}[\alpha_i,\beta_j], \quad 1 \le i \le N_\mathcal{A},\ 1 \le j \le N_\mathcal{B}.$$

Then $\boldsymbol{A} * \boldsymbol{B} \in \mathbb{R}^{\sum_{i=1}^{N_\mathcal{A}} \ell(\alpha_i)^2 \times \sum_{j=1}^{N_\mathcal{B}} \ell(\beta_j)^2}$. We define the k-th Khatri-Rao power *of $\boldsymbol{A}$ by*

$$\boldsymbol{A}^{*k} := \underbrace{\boldsymbol{A} * \boldsymbol{A} * \cdots * \boldsymbol{A}}_{k \text{ times}}.$$

Lemma 30 ([74])**.**
Let $\boldsymbol{A}$, $\boldsymbol{B}$ be compatibly partitioned positive (semi-)definite matrices. Then $\boldsymbol{A} * \boldsymbol{B}$ is positive (semi-) definite.

Lemma 31.
Let $\boldsymbol{A},\boldsymbol{B} \in \mathbb{R}^{n\times m}$ be compatibly partitioned matrices. Then

$$\|\boldsymbol{A} * \boldsymbol{B}\|_2 \le \left\|(\boldsymbol{A}^T\boldsymbol{A}) * \boldsymbol{I}_m\right\|_2^{\frac{1}{2}} \left\|\boldsymbol{I}_n * (\boldsymbol{B}\boldsymbol{B}^T)\right\|_2^{\frac{1}{2}}.$$

Proof. The matrices $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{I}_n \end{pmatrix}^T \begin{pmatrix} \boldsymbol{A} & \boldsymbol{I}_n \end{pmatrix}$ and $\begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{B}^T \end{pmatrix}^T \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{B}^T \end{pmatrix}$ are trivially positive semi-definite. Using lemma 30 their Khatri-Rao product

$$\begin{pmatrix} (\boldsymbol{A}^T\boldsymbol{A}) * \boldsymbol{I}_m & \boldsymbol{A}^T * \boldsymbol{B}^T \\ \boldsymbol{A} * \boldsymbol{B} & \boldsymbol{I}_n * (\boldsymbol{B}\boldsymbol{B}^T) \end{pmatrix}$$

is positive semi-definite as well. Hence, by Lemma 14,

$$\|\boldsymbol{A} * \boldsymbol{B}\|_2^2 \leq \left\|(\boldsymbol{A}^T\boldsymbol{A}) * \boldsymbol{I}_m\right\|_2 \left\|\boldsymbol{I}_n * (\boldsymbol{B}\boldsymbol{B}^T)\right\|_2 .$$

□

Therefore, if we write the partition of a matrix $\boldsymbol{Z}$ in terms of the non-overlapping partition given by the restriction and prolongation matrices $\boldsymbol{R}^{(i,-1)}$ and $\boldsymbol{P}^{(i,-1)}$ as

$$\boldsymbol{Z} = \sum_{i,j} \boldsymbol{P}^{(i,-1)} \left(\boldsymbol{R}^{(i,-1)}\boldsymbol{Z}\boldsymbol{P}^{(j,-1)}\right) \boldsymbol{R}^{(j,-1)},$$

we obtain

$$\begin{aligned}\rho\left(\boldsymbol{Z}^{*2}\right) &\leq \max_i \left\|\boldsymbol{R}^{(i,-1)}\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{P}^{(i,-1)}\right\|_2^{1/2} \max_i \left\|\boldsymbol{R}^{(i,-1)}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{P}^{(i,-1)}\right\|_2^{1/2} \\ &= \max_i \left\|\boldsymbol{Z}\boldsymbol{P}^{(i,-1)}\right\|_2 \max_i \left\|\boldsymbol{Z}^T\boldsymbol{P}^{(i,-1)}\right\|_2 . \end{aligned} \tag{8.3}$$

Lemma 32.
For all $u \in V_h$,

$$a\left(\mathcal{R}^{(i,-1)}u, \mathcal{R}^{(i,-1)}u\right) \leq Ch^{-1}a(u,u)$$

for some constant C that is independent of h and N.

The proof of Lemma 32 follows immediately from (8.1) in Lemma 24.

Proof of Lemma 26. Set $\boldsymbol{D} := \sum_i \boldsymbol{P}^{(i,-1)}\boldsymbol{A}^{(i,-1)}\boldsymbol{R}^{(i,-1)}$. We then find that

$$\begin{aligned}\|\mathbb{V}\left[\boldsymbol{\mathcal{E}}_{CG}\right]\|_A &= \left\|\boldsymbol{N}_{CG}^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]\boldsymbol{A}^{\otimes 2}\right\|_A \\ &= \left\|\left(\boldsymbol{A}^{1/2}\boldsymbol{N}_{CG}\right)^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]\left(\boldsymbol{A}^{1/2}\right)^{\otimes 2}\right\|_2 \\ &= \left\|\left(\boldsymbol{A}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{D}^{1/2}\right)^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]\left(\boldsymbol{D}^{-1/2}\boldsymbol{A}^{1/2}\right)^{\otimes 2}\right\|_2 \\ &= \rho\left(\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]^{1/2}\left(\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{A}\boldsymbol{N}_{CG}\boldsymbol{D}^{1/2}\right)^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]\right. \\ &\qquad \left.\left(\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\right)^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]^{1/2}\right)^{1/2}\end{aligned}$$

$$\leq \rho\left(\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]^{1/2}\left(\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{D}^{1/2}\right)^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]^{1/2}\right)^{1/2}$$
$$\rho\left(\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]^{1/2}\left(\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\right)^{\otimes 2}\mathbb{V}\left[\boldsymbol{\mathcal{X}}\right]^{1/2}\right)^{1/2}$$
$$= \varepsilon\rho\left(\left(\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{D}^{1/2}\right)^{*2}\right)^{1/2}\rho\left(\left(\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\right)^{*2}\right)^{1/2}.$$

Using the result of eq. (8.3) and the discrete partition of unity property

$$\boldsymbol{I} = \sum_j \boldsymbol{P}^{(j,-1)}\boldsymbol{R}^{(j,-1)},$$

one obtains that

$$\begin{aligned}
\rho\left(\left(\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\right)^{*2}\right) &\leq \max_i \left\|\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{P}^{(i,-1)}\right\|_2^2 \\
&\leq \max_i \sum_j \left\|\boldsymbol{R}^{(j,-1)}\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{P}^{(i,-1)}\right\|_2^2 \\
&= \max_i \sum_j \left\|\left(\boldsymbol{A}^{(j,-1)}\right)^{-1/2}\boldsymbol{R}^{(j,-1)}\boldsymbol{A}\boldsymbol{P}^{(i,-1)}\left(\boldsymbol{A}^{(i,-1)}\right)^{-1/2}\right\|_2^2 \\
&\leq \max_i \sum_j \epsilon_{ij}^2 \\
&\leq N_c.
\end{aligned}$$

By the same token,

$$\rho\left(\left(\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{D}^{1/2}\right)^{*2}\right) \leq \max_i \left\|\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{P}^{(i,-1)}\left(\boldsymbol{A}^{(i,-1)}\right)^{1/2}\right\|_2^2.$$

Since

$$\begin{aligned}
&\left\|\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{P}^{(i,-1)}\left(\boldsymbol{A}^{(i,-1)}\right)^{1/2}\right\|_2 \\
\leq &\left\|\boldsymbol{D}^{1/2}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\left(\boldsymbol{A}^{(i,-1)}\right)^{1/2}\right\|_2 + \left\|\boldsymbol{D}^{1/2}\left(\boldsymbol{A}^{-1}-\boldsymbol{N}_{CG}\right)\boldsymbol{P}^{(i,-1)}\left(\boldsymbol{A}^{(i,-1)}\right)^{1/2}\right\|_2 \\
\leq &\left\|\boldsymbol{D}^{1/2}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\left(\boldsymbol{A}^{(i,-1)}\right)^{1/2}\right\|_2 + \left\|\boldsymbol{D}\right\|_2\left\|\boldsymbol{A}^{-1}-\boldsymbol{N}_{CG}\right\|_2 \\
\leq &\left\|\boldsymbol{D}^{1/2}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\left(\boldsymbol{A}^{(i,-1)}\right)^{1/2}\right\|_2 + C,
\end{aligned}$$

we have

$$\rho\left(\left(\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{D}^{1/2}\right)^{*2}\right) \leq C + \max_i \sum_j \left\| \left(\boldsymbol{A}^{(j,-1)}\right)^{1/2} \boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)} \left(\boldsymbol{A}^{(i,-1)}\right)^{1/2} \right\|_2^2.$$

We expand the summand into

$$\begin{aligned}
&\left\| \left(\boldsymbol{A}^{(j,-1)}\right)^{1/2} \boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)} \left(\boldsymbol{A}^{(i,-1)}\right)^{1/2} \right\|_2 \\
\leq &\left\| \left(\boldsymbol{A}^{(j,-1)}\right)^{1/2} \left(\boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(j,-1)}\right)^{-1/2} \left(\boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(j,-1)}\right)^{1/2} \boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)} \right. \\
&\qquad \left. \left(\boldsymbol{R}^{(i,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\right)^{1/2} \left(\boldsymbol{R}^{(i,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\right)^{-1/2} \left(\boldsymbol{A}^{(i,-1)}\right)^{1/2} \right\|_2 \\
\leq &\rho\left(\boldsymbol{A}^{(j,-1)} \left(\boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(j,-1)}\right)^{-1}\right)^{1/2} \rho\left(\boldsymbol{A}^{(i,-1)} \left(\boldsymbol{R}^{(i,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\right)^{-1}\right)^{1/2} \\
&\qquad \left\| \left(\boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(j,-1)}\right)^{1/2} \boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)} \left(\boldsymbol{R}^{(i,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\right)^{1/2} \right\|_2
\end{aligned}$$

Using that

$$\rho\left(\boldsymbol{A}^{(i,-1)} \left(\boldsymbol{R}^{(i,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\right)^{-1}\right) = \sup_{u\in V_h} \frac{a\left(\mathcal{R}^{(i,-1)}u, \mathcal{R}^{(i,-1)}u\right)}{a(u,u)}$$

and that

$$\left\| \left(\boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(j,-1)}\right)^{1/2} \boldsymbol{R}^{(j,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)} \left(\boldsymbol{R}^{(i,-1)}\boldsymbol{A}^{-1}\boldsymbol{P}^{(i,-1)}\right)^{1/2} \right\|_2 \leq \eta_{ij}$$

we therefore obtain that

$$\rho\left(\left(\boldsymbol{D}^{1/2}\boldsymbol{N}_{CG}\boldsymbol{D}^{1/2}\right)^{*2}\right) \leq C + \left\{ \max_i \sup_{u\in V_h} \frac{a\left(\mathcal{R}^{(i,-1)}u, \mathcal{R}^{(i,-1)}u\right)}{a(u,u)} \right\}^2 \max_i \sum_j \eta_{ij}^2.$$

We conclude using Lemma 32. □

While the last inequality in (8.2) is not sharp in general, we verify numerically that it describes the correct behavior for one dimensional problems. We calculate $\|\mathbb{V}\left[\boldsymbol{\mathcal{E}}_{CG}\right]\|_A$ directly for varying mesh spacing h and number of partitions N. The resulting scaling curves are shown in Figure 8.5. This shows that the bound in Lemma 26 is sharp in 1D, and that we cannot conclude the convergence of the two level methods using the energy norm.

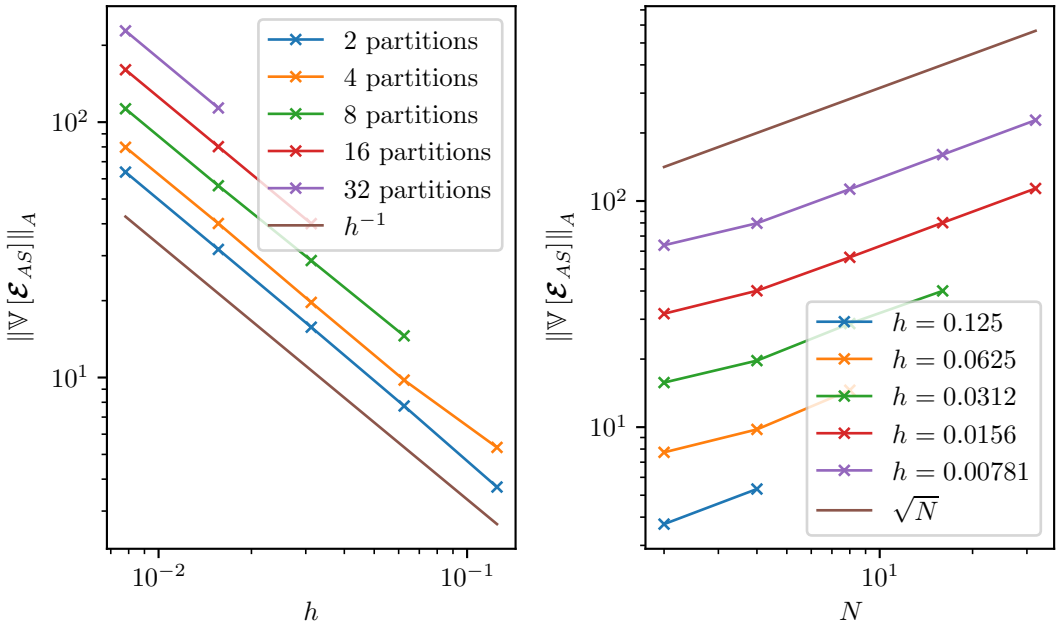


Figure 8.5: Energy norm of the variance of the fault-prone coarse-grid correction. Scaling with respect to mesh size h (*left*), and number of subdomains N (*right*).

On the other hand, if we consider the coarse grid correction on its own, we find that its Lyapunov spectral radius does not display *any* dependence on h or N:

Lemma 33.

If the prolongation in the coarse grid correction is protected, then it holds for the double energy norm $\|\bullet\|_{A^2} := \left\|\boldsymbol{A} \bullet \boldsymbol{A}^{-1}\right\|_2$, $\|\bullet\|_{A^2} := \left\|\boldsymbol{A}^{\otimes 2} \bullet \left(\boldsymbol{A}^{-1}\right)^{\otimes 2}\right\|_2$ *that*

$$\|\mathbb{V}[\boldsymbol{\mathcal{E}}_{CG}]\|_{A^2} \leq C\varepsilon$$

with C *independent of* ε, h *and* N, *so that*

$$\varrho\left(\boldsymbol{\mathcal{E}}_{CG}\right) \leq \|\boldsymbol{E}_{CG}\|_2 + C\varepsilon.$$

Proof.

$$\begin{aligned}\|\mathbb{V}[\boldsymbol{\mathcal{E}}_{CG}]\|_{A^2} &\leq \left\|\boldsymbol{N}_{CG}^{\otimes 2}\mathbb{V}[\boldsymbol{\mathcal{X}}]\,\boldsymbol{A}^{\otimes 2}\right\|_{A^2} \\ &\leq \|\boldsymbol{A}\boldsymbol{N}_{CG}\|_2^2\,\|\mathbb{V}[\boldsymbol{\mathcal{X}}]\|_2\end{aligned}$$

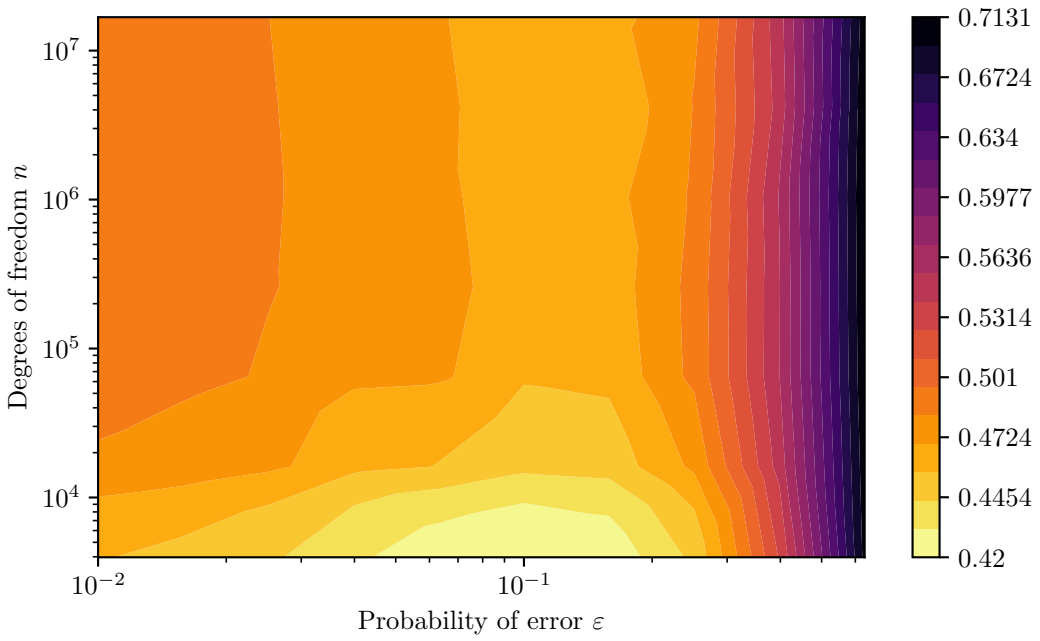


Figure 8.6: Asymptotic convergence rate $\varrho\left(\boldsymbol{\mathcal{E}}_{ASH}\boldsymbol{\mathcal{E}}_{CG}\right)$ with protected prolongation and $N = 16$ subdomains.

$$\leq \varepsilon \left\|\boldsymbol{N}_{CG}\boldsymbol{A}\right\|_2^2 .$$

Now, since $\left\|\boldsymbol{E}_{CG}\right\|_2 \leq C$ (see for example [62]), we have that $\left\|\mathbb{V}\left[\boldsymbol{\mathcal{E}}_{CG}\right]\right\|_{A^2} \leq C\varepsilon$ and also that

$$\begin{aligned}\varrho\left(\boldsymbol{\mathcal{E}}_{CG}\right) &\leq \sqrt{\left\|\mathbb{E}\left[\boldsymbol{\mathcal{E}}_{CG}\right]\right\|_{A^2}^2 + \left\|\mathbb{V}\left[\boldsymbol{\mathcal{E}}_{CG}\right]\right\|_{A^2}} \\ &\leq \left\|\boldsymbol{E}_{CG}\right\|_2 + C\varepsilon.\end{aligned}$$

□

While the coarse grid correction on itself is not a convergent iteration, this result shows that we can expect its fault-prone version to perform similarly to its fault-free variant. This is indeed what can be observed through numerical simulations, as seen in Figures 8.6 and 8.7. (The convergence is the fastest for $\varepsilon \approx 0.1$, not $\varepsilon = 0$. This indicates that some amount of damping might be beneficial for ASH.)

In order to obtain an analytic result for the full two level method, it would seem logical to analyze the domain decomposition solver in turn in the double energy

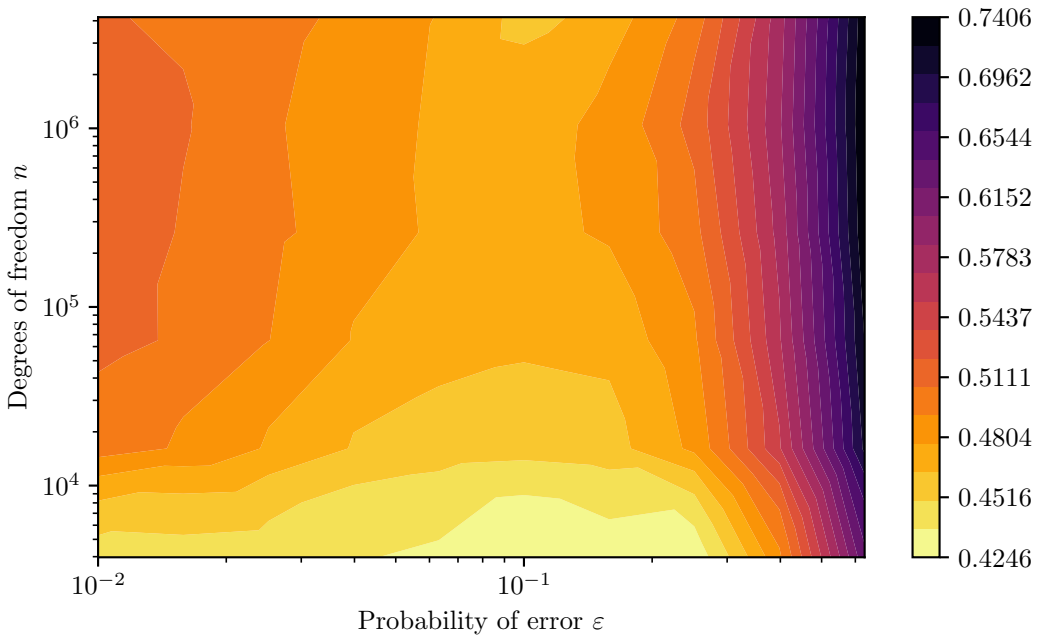


Figure 8.7: Asymptotic convergence rate $\varrho\left(\boldsymbol{\mathcal{E}}_{ASH}\boldsymbol{\mathcal{E}}_{CG}\right)$ with protected prolongation and $N \sim n$ subdomains.

or L^2 norm. The results in [25] suggest however that $\|\boldsymbol{N}_{AS}\boldsymbol{A}\|_{A^2} = \|\boldsymbol{N}_{AS}\boldsymbol{A}\|_2$ diverges as h goes to zero. It can be easily verified numerically (see Figure 8.8) that for a 1D test problem $\|\boldsymbol{N}_{AS}\boldsymbol{A}\|_2 \sim h^{-1/2}$, while $\rho\left(\boldsymbol{N}_{AS}\boldsymbol{A}\right) = \|\boldsymbol{N}_{AS}\boldsymbol{A}\|_A \leq C$.

8.5 Discussion

In this chapter, we generalized the framework to the analysis of fault-prone iterative schemes with overlap. We modeled blockwise faults, where special care had to be taken to account for the fact that affect unknowns can be owned by several subdomains at once. We gave convergence results for three commonly used one-level methods and observed in numerical examples that one-level solvers are resilient to faults, mirroring the behavior observed for the smoothening iterations of multigrid.

Inclusion of a coarse grid, as is mandatory for many applications, only results in a fault-resilient method, if the prolongation operation is protected. While the

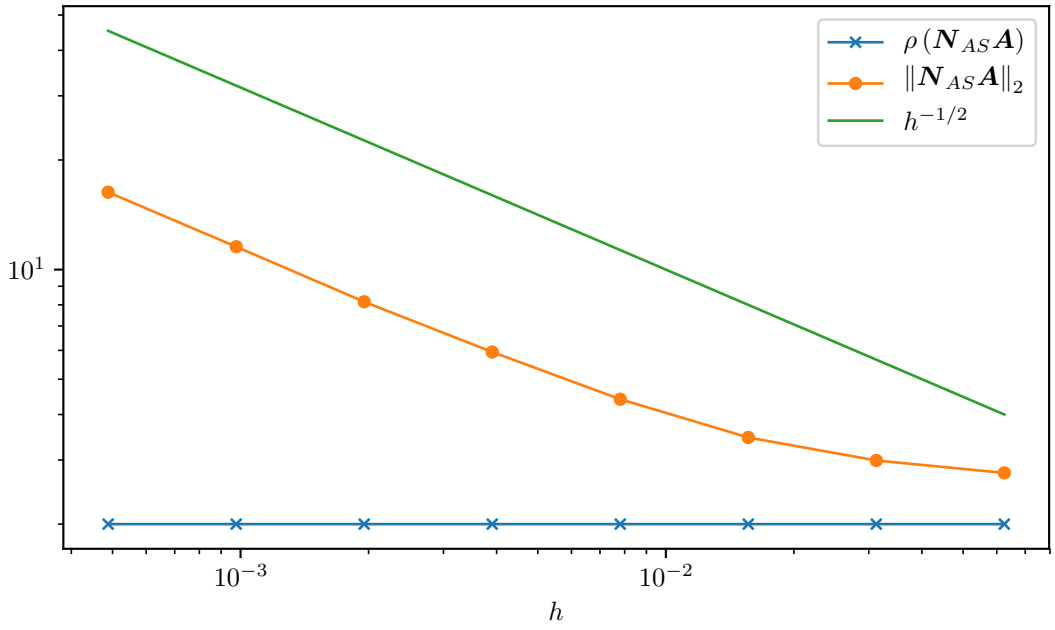


Figure 8.8: $\|\boldsymbol{N}_{AS}\boldsymbol{A}\|_2$ in 1D for several mesh sizes h and $N = 4$ subdomains.

numerical examples closely resemble the ones obtained for the multigrid method, the analysis of two-level methods in unsatisfactory. The issue that arises is that convergence analysis of the domain decomposition iteration can not be performed in spectral norm, contrary to multigrid. The isolated fault-prone coarse grid correction is non-divergent, as is the isolated domain decomposition iteration, but the analysis of their combination is inconclusive.

Chapter 9

Future Work on Fault Resilience and Related Topics

9.1 Fault-Resilient Domain Decomposition Methods

While the numerical experiments in Section 8.4 suggest that the same techniques as in the multigrid case can be used to obtain a fault-resilient two-level domain decomposition method, the application of the theoretical framework to this case is more intricate. The random faults matrices are best analyzed in L^2-type norms, whereas domain decomposition methods are analyzed in energy norm. I believe that by a more tailored approach it is nevertheless possible to obtain convergence estimates for this important class of solvers. Equally well, it would be of interest to assert that domain decomposition methods without overlap (also called substructuring methods) can be made fault-resilient in the same way.

In Chapter 7, we discussed a simple approach to realize fault detection in all operations and protection of the prolongation. While the replication approach was chosen to illustrate the practical consequences of the theoretical results of the previous chapters, its implementation suffers from a large overhead. It might be prudent

to explore the use of node-level checksums for fault detection [68]. The computation of checksums would only introduce a negligible overhead, but come at the price of a coarser fault detection. Instead of applying the laissez-faire approach to a single faulty component, a whole block of values would be zeroed out. Moreover, protection of the prolongation might be achieved by using values from coarser levels of the hierarchy as implicit checkpoints.

Finally, there is interest to combine our work for linear iterative methods with the results obtained by other researchers for fault-prone Krylov methods and study the complete stack of linear systems solvers in high-performance computing.

9.2 Asynchronous Iterations

More generally, random matrices can be used not to model faults, but the delayed arrival of updates in an asynchronous iterative method. This touches on the challenge of communication costs in evermore parallel computation. Classically, iterative methods alternate phases of high communication volume between compute nodes with phases of high computational intensity. As supercomputers are built from more and more compute units, communication becomes a bottleneck for performance. Therefore, algorithms that are as asynchronous as possible will be required for the next generation of supercomputers. The study of such asynchronous iterative methods by itself is not a new topic [16, 54, 76, 91], but generally seems to be lacking sharp convergence estimates. In order to employ our novel approach based on random matrices, we would need to extend the theoretical framework to allow for memory terms, as to model the staggered arrival of updates.

Part II

Fast Solvers for Non-Local Equations

Chapter 10

The Finite Element Method for the Fractional Laplacian

In recent years, there has been a burgeoning of interest in the use of non-local and fractional models. To some extent, this move reflects the fact that with present day computational resources coupled with state of the art numerical algorithms, attention is now shifting back to the fidelity of the underlying mathematical models as opposed to their approximation. Fractional equations have been used to describe phenomena in anomalous diffusion [15], material science [13, 85], image processing [57, 75], finance [93] and electromagnetic fluids [77] [100]. Fractional order equations arise naturally as the limit of discrete diffusion governed by stochastic processes [79].

Whilst the development of fractional derivatives dates back to essentially the same time as their integer counterparts, the computational methods available for their numerical resolution drastically lags behind the vast array of numerical techniques from which one can choose to treat integer order partial differential equations. The recent literature abounds with work on numerical methods for fractional partial differential equations in one spatial dimension and fractional order temporal derivatives. However, with most applications of interest being posed on domains

in two or more spatial dimensions, the solution of fractional equations posed on complex domains is a problem of considerable practical interest.

The archetypal elliptic partial differential equation is the Poisson problem involving the standard Laplacian. By analogy, one can consider a fractional Poisson problem involving the fractional Laplacian. The first problem one encounters is that of how to define a fractional Laplacian, particularly in the case where the domain is compact, and a number of alternatives have been suggested. The *integral fractional Laplacian* is obtained by restriction of the Fourier definition to functions that have prescribed value outside of the domain of interest, whereas the spectral fractional Laplacian is based on the spectral decomposition of the regular Laplace operator. In general, the two operators are different [84], and only coincide when the domain of interest is the full space.

The approximation of the integral fractional Laplacian using finite elements was considered by D'Elia and Gunzburger [39]. The important work of Acosta and Borthagaray [2] and Grubb [61] gave regularity results for the analytic solution of the fractional Poisson problem and obtained convergence rates for the finite element approximation supported by numerical examples computed using techniques described in [1].

The numerical treatment of fractional partial differential equations is rather different from the integer order case owing to the fact that the fractional derivative is a non-local operator. This creates a number of issues including the fact that the resulting stiffness matrix is dense and, moreover, the entries in the matrix are given in terms of singular integrals. In turn, these features create issues in the numerical computation of the entries and the need to store the entries of a dense matrix, not to mention the fact that a solution of the resulting matrix equation has to be computed. The seasoned reader will readily appreciate that many of these issues are shared by boundary integral equations arising from classical integer order differential operators [86, 87, 103]. This similarity is not altogether surprising given

that the boundary integral operators are pseudo-differential operators of fractional order.

A different, integer order operator based approach, was taken by Nochetto, Otárola and Salgado [31, 80] for the case of the spectral Laplacian. Caffarelli and Silvestre [22] showed that the operator can be realized as a Dirichlet-to-Neumann operator of an extended problem in the half space in $d+1$ dimensions.

In this part of the work, we explore the connection with boundary integral operators to develop techniques that enable one to efficiently tackle the integral fractional Laplacian. In particular, we develop techniques for the treatment of the stiffness matrix including the computation of the entries, the efficient storage of the resulting dense matrix and the efficient solution of the resulting equations. Since solutions of the fractional Poisson problem are generally non-smooth in a neighborhood of the boundary, we introduce a posteriori error estimators that are used for adaptive mesh refinement. The main ideas consist of generalizing proven techniques for the treatment of boundary integral equations to general fractional orders. Importantly, the approximation does not make any strong assumptions on the shape of the underlying domain and does not rely on any special structure of the matrix that could exploited by fast transforms. We demonstrate the flexibility and performance of this approach in a couple of one- and two-dimensional numerical examples. It will be observed that optimal rates of convergence are achieved in all numerical examples.

10.1 Fractional Laplacians

The fractional Laplacian in $\mathbb{R}^d$ of order s, for $0 < s < 1$ and $d \in \mathbb{N}$, of a function u can be defined by the Fourier transform $\mathcal{F}$ as

$$(-\Delta)^s u = \mathcal{F}^{-1}\left[|\xi|^{2s}\,\mathcal{F}u\right].$$

Alternatively, this expression can be rewritten [96] in integral form as

$$(-\Delta)^s u(\vec{x}) = C(d,s)\,\mathrm{p.\,v.} \int_{\mathbb{R}^d} d\vec{y}\, \frac{u(\vec{x}) - u(\vec{y})}{|\vec{x} - \vec{y}|^{d+2s}}$$

where

$$C(d,s) = \frac{2^{2s} s \Gamma\left(s + \frac{d}{2}\right)}{\pi^{d/2} \Gamma(1-s)}$$

is a normalization constant and $\mathrm{p.\,v.}$ denotes the Cauchy principal value of the integral [78, Chapter 5]. In the case where $s = 1$ this operator coincides with the usual Laplacian. If $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain, we define the *integral fractional Laplacian* $(-\Delta)^s$ to be the restriction of the full-space operator to functions with compact support in Ω. This generalizes the homogeneous Dirichlet condition applied in the case $s = 1$ to the case $s \in (0,1)$.

The *fractional Poisson problem* is then given by

$$\begin{aligned} (-\Delta)^s u &= f \text{ in } \Omega, \\ u &= 0 \text{ in } \Omega^c \end{aligned} \tag{10.1}$$

This generalizes the homogeneous integer order Poisson problem from the case $s = 1$ to the case $s \in (0,1)$.

A related operator on Ω is the *regional fractional Laplacian*

$$(-\Delta)^s_R u(\vec{x}) = C(d,s)\,\mathrm{p.\,v.} \int_{\Omega} d\vec{y}\, \frac{u(\vec{x}) - u(\vec{y})}{|\vec{x} - \vec{y}|^{d+2s}},$$

which generalizes the Laplacian with homogeneous Neumann boundary condition to the fractional order case. It will be seen that the techniques developed below for the integral fractional Laplacian also apply for the regional fractional Laplacian.

In practice, we are usually interested in dynamic behavior rather than steady state problems. The archetype of a time-dependent problem is the *fractional heat equation*

$$\partial_t u + (-\Delta)^s u = f \text{ in } [0,T] \times \Omega, \tag{10.2}$$

$$u = u_0 \text{ in } \{0\} \times \Omega.$$

Equally well, we might consider the fractional heat equation with the regional fractional Laplacian instead of the integral one, if a Neumann boundary condition is more appropriate.

10.2 Weak formulation

Define the usual fractional Sobolev space $H^s\left(\mathbb{R}^d\right)$ via the Fourier transform. If Ω is a sub-domain as above, then we define the Sobolev space $H^s(\Omega)$ to be

$$H^s(\Omega) := \left\{ u \in L^2(\Omega) \mid \|u\|_{H^s(\Omega)} < \infty \right\},$$

equipped with the norm

$$\|u\|^2_{H^s(\Omega)} = \|u\|^2_{L^2(\Omega)} + \int_\Omega d\vec{x} \int_\Omega d\vec{y} \frac{(u(\vec{x}) - u(\vec{y}))^2}{|\vec{x} - \vec{y}|^{d+2s}}.$$

The space

$$\widetilde{H}^s(\Omega) := \left\{ u \in H^s\left(\mathbb{R}^d\right) \mid u = 0 \text{ in } \Omega^c \right\}$$

can be equipped with the energy norm

$$\|u\|_{\widetilde{H}^s(\Omega)} := a(u,u)^{1/2} = \sqrt{\frac{C(d,s)}{2}} |u|_{H^s\left(\mathbb{R}^d\right)},$$

where the non-standard factor $\sqrt{C(d,s)/2}$ is included for convenience. For $s > 1/2$, $\widetilde{H}^s(\Omega)$ coincides with the space $H^s_0(\Omega)$ which is the closure of $C^\infty_0(\Omega)$ with respect to the $H^s(\Omega)$-norm. For $s < 1/2$, $\widetilde{H}^s(\Omega)$ is identical to $H^s(\Omega)$. In the critical case $s = 1/2$, $\widetilde{H}^s(\Omega) \subset H^s_0(\Omega)$, and the inclusion is strict. (See for example [78, Chapter 3].)

The usual approach to dealing with elliptic PDEs consists of obtaining a weak form of the operator by multiplying the equation by a test function and applying

integration by parts [50]. In contrast, for equations involving the fractional Laplacian $(-\Delta)^s u$, we again multiply by as test function $v \in \widetilde{H}^s(\Omega)$ and integrate over $\mathbb{R}^d$, and then, instead of integration by parts, we use the identity

$$\int_{\mathbb{R}^d} d\vec{x} \int_{\mathbb{R}^d} d\vec{y} \frac{(u(\vec{x}) - u(\vec{y})\, v(\vec{x}))}{|\vec{x}-\vec{y}|^{d+2s}} = -\int_{\mathbb{R}^d} d\vec{x} \int_{\mathbb{R}^d} d\vec{y} \frac{(u(\vec{x}) - u(\vec{y})\, v(\vec{y}))}{|\vec{x}-\vec{y}|^{d+2s}}.$$

Following this approach, since both u and v vanish outside of Ω, we arrive at the bilinear form

$$\begin{aligned} a(u,v) = \frac{C(d,s)}{2} \int_{\Omega} d\vec{x} \int_{\Omega} d\vec{y} \frac{(u(\vec{x}) - u(\vec{y}))\,(v(\vec{x}) - v(\vec{y}))}{|\vec{x}-\vec{y}|^{d+2s}} \\ + C(d,s) \int_{\Omega} d\vec{x} \int_{\Omega^c} d\vec{y} \frac{u(\vec{x})\, v(\vec{x})}{|\vec{x}-\vec{y}|^{d+2s}}, \end{aligned}$$

corresponding to $(-\Delta)^s$ on $\widetilde{H}^s(\Omega) \times \widetilde{H}^s(\Omega)$. The bilinear form $a(\cdot,\cdot)$ is trivially seen to be $\widetilde{H}^s(\Omega)$-coercive and continuous and, as such, is amenable to treatment using the Lax-Milgram Lemma.

In this chapter we shall concern ourselves with the computational details needed to implement the finite element approximation of problems involving the fractional Laplacian. To this end, the presence of the unbounded domain Ω^c in the bilinear form $a(\cdot,\cdot)$ is somewhat undesirable. Fortunately, we can dispense with Ω^c using the following argument. The identity

$$\frac{1}{|\vec{x}-\vec{y}|^{d+2s}} = \frac{1}{2s} \nabla_{\vec{y}} \cdot \frac{\vec{x}-\vec{y}}{|\vec{x}-\vec{y}|^{d+2s}},$$

enables the second integral to be rewritten using the Gauss theorem as

$$\frac{C(d,s)}{2s} \int_{\Omega} d\vec{x} \int_{\partial\Omega} d\vec{y} \frac{u(\vec{x})\, v(\vec{x})\ \ \vec{n}_{\vec{y}} \cdot (\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}},$$

where $\vec{n}_y$ is the *inward* normal to $\partial\Omega$ at $\vec{y}$, so that the bilinear form can be expressed equivalently as

$$\begin{aligned} a(u,v) = \frac{C(d,s)}{2} \int_{\Omega} d\vec{x} \int_{\Omega} d\vec{y} \frac{(u(\vec{x}) - u(\vec{y}))\,(v(\vec{x}) - v(\vec{y}))}{|\vec{x}-\vec{y}|^{d+2s}} \\ + \frac{C(d,s)}{2s} \int_{\Omega} d\vec{x} \int_{\partial\Omega} d\vec{y} \frac{u(\vec{x})\, v(\vec{x})\ \ \vec{n}_{\vec{y}} \cdot (\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}}. \end{aligned}$$

As an aside, we note that dropping the boundary term in $a(\cdot,\cdot)$ leads to the bilinear form

$$b(u,v) = \frac{C(d,s)}{2}\int_\Omega d\vec{x}\int_\Omega d\vec{y}\frac{(u(\vec{x})-u(\vec{y}))(v(\vec{x})-v(\vec{y}))}{|\vec{x}-\vec{y}|^{d+2s}}$$

which represents the regional fractional Laplacian [17, 32]. The regional fractional Laplacian can be interpreted as a generalization of the usual Laplacian with homogeneous Neumann boundary condition for $s = 1$ to the case of fractional orders $s \in (0,1)$. It will transpire from our work that most of the presented techniques carry over to the regional fractional Laplacian by simply omitting the boundary integral terms.

10.3 Finite Element Approximation of the Fractional Poisson Equation

The fractional Poisson problem

$$\begin{aligned} (-\Delta)^s u &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \Omega^c \end{aligned}$$

takes the variational form

$$\text{Find } u \in \widetilde{H}^s(\Omega): \quad a(u,v) = \langle f, v\rangle \quad \forall v \in \widetilde{H}^s(\Omega). \tag{10.3}$$

The existence of a unique solution to the fractional Poisson problem eq. (10.1) (and its subsequent finite element approximation) follows from the Lax-Milgram Lemma.

As a special case of a result by Grubb [61], the regularity of the solution was recorded in [1]:

Theorem 34.
Let $\partial\Omega \in C^\infty$, $f \in H^r(\Omega)$ for $r \geq -s$ and $u \in \widetilde{H}^s(\Omega)$ be the solution of the fractional

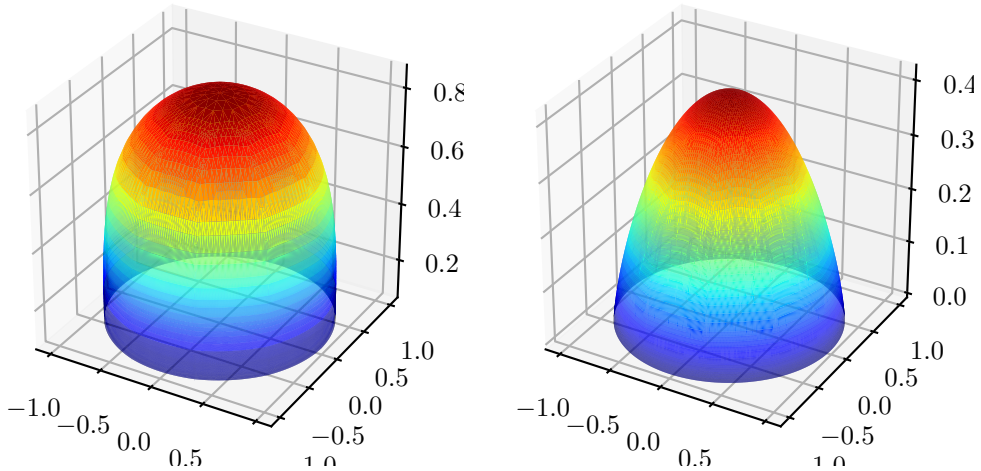


Figure 10.1: Solutions u corresponding to the constant right-hand side for $s = 0.25$ and for $s = 0.75$.

Poisson problem (10.3). *Then the following regularity estimate holds:*

$$u \in \begin{cases} H^{2s+r}(\Omega) & \text{if } 0 < s + r < 1/2, \\ H^{s+1/2-\varepsilon}(\Omega) \ \forall \varepsilon > 0 & \text{if } 1/2 \leq s + r. \end{cases}$$

The result shows that increasing the regularity of the right-hand side only results in an increase of the regularity of the solution, as long as $r < 1/2 - s$. We contrast the result with the standard "lifting property" of the integer order Laplacian [50, Theorem 3.10], whereby a gain of two derivatives takes place on smooth domains. Examining a particular solution reveals why the lifting property does not extend to the fractional case. On $\Omega = B(0,1) \subset \mathbb{R}^2$, the constant right-hand side $f = 2^{2s}\Gamma(1+s)^2$ has corresponding solution $u(\vec{x}) = \left(1 - |\vec{x}|^2\right)_+^s$, where $y_+ = \max(y, 0)$, and u is singular at the boundary [45, 58]. (See Figure 10.1.) It generally holds that the solutions of the fractional Poisson problem contain a power $\delta(\vec{x})^s \in H^{s+1/2-\varepsilon}$, where $\delta(\vec{x})$ is the distance of $\vec{x}$ to the boundary $\partial\Omega$.

For polygonal domains, the infinite lifting property does not hold in the integer order case either [50, Theorem 3.11].

Henceforth, let Ω be a polygon, and let $\mathcal{P}_h$ be a family of shape-regular and

locally quasi-uniform triangulations of Ω [50], and

$$\partial\mathcal{P}_h = \{\text{edge } e \mid \exists K \in \mathcal{P}_h : e \subset \partial K \cap \partial\Omega\}$$

the trace of the interior mesh. Let $\mathcal{N}_h$ be the set of vertices of $\mathcal{P}_h$ and h_K be the diameter of the element $K \in \mathcal{P}_h$, and h_e the diameter of $e \in \partial\mathcal{P}_h$. Moreover, let

$$h := \max_{K\in\mathcal{P}_h} h_K,$$

$$h_{\min} := \min_{K\in\mathcal{P}_h} h_K,$$

$$h_\partial := \max_{e\in\partial\mathcal{P}_h} h_e.$$

Let ϕ_i be the usual piecewise linear basis function associated with a node $\vec{z}_i \in \mathcal{N}_h$, satisfying $\phi_i(\vec{z}_j) = \delta_{ij}$ for $\vec{z}_j \in \mathcal{N}_h$, and let $X_h := \operatorname{span}\{\phi_i \mid \vec{z}_i \in \mathcal{N}_h\}$. The finite element subspace $V_h \subset \widetilde{H}^s(\Omega)$ is given by $V_h = X_h$ when $s < 1/2$ and by

$$V_h = \{v_h \in X_h \mid v_h = 0 \text{ on } \partial\Omega\} = \operatorname{span}\{\phi_i \mid \vec{z}_i \notin \partial\Omega\}$$

when $s \geq 1/2$. The corresponding set of degrees of freedom $\mathcal{I}_h$ for V_h is given by $\mathcal{I}_h = \mathcal{N}_h$ when $s < 1/2$ and otherwise consists of nodes in the interior of Ω. In both cases we denote the cardinality of $\mathcal{I}_h$ by n. The set of degrees of freedom on an element $K \in \mathcal{P}_h$ is denoted by $\mathcal{I}_K$.

The stiffness matrix associated with the fractional Laplacian is defined to be $\boldsymbol{A}^s = \{a(\phi_i, \phi_j)\}_{i,j}$, where

$$\begin{aligned} a(\phi_i, \phi_j) = &\frac{C(d,s)}{2}\int_\Omega d\vec{x}\int_\Omega d\vec{y}\frac{(\phi_i(\vec{x}) - \phi_i(\vec{y}))(\phi_j(\vec{x}) - \phi_j(\vec{y}))}{|\vec{x}-\vec{y}|^{d+2s}} \\ &+ \frac{C(d,s)}{2s}\int_\Omega d\vec{x}\int_{\partial\Omega} d\vec{y}\frac{\phi_i(\vec{x})\,\phi_j(\vec{x})\;\vec{n}_{\vec{y}}\cdot(\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}}. \end{aligned}$$

Let $u_h \in V_h$ be the finite element solution. By Céa's Lemma and the inclusion $\widetilde{H}^s(\Omega) \subseteq H^s(\Omega)$,

$$\|u - u_h\|_{\widetilde{H}^s(\Omega)} \leq C \inf_{v_h\in V_h} \|u - v_h\|_{\widetilde{H}^s(\Omega)} \leq C \inf_{v_h\in V_h} \|u - v_h\|_{H^s(\Omega)}. \tag{10.4}$$

The proof of the rate of convergence is based on approximation properties of the Scott-Zhang interpolation operator with respect to the $\|\cdot\|_{H^s(\Omega)}$-norm. For $u \in H^\ell(\Omega)$, $\ell > 1/2$, let

$$\Pi_h u = \sum_i \left(\int_{K_i} \psi_i(\vec{y})\, u(\vec{y})\ d\vec{y} \right) \phi_i$$

denote the Scott-Zhang interpolation operator [83]. Here, K_i is either a simplex K of the triangulation such that $i \in \mathcal{I}_K$, or a sub-simplex e such that i is associated with e, and $\{\psi_i\}_i$ is an $L^2(K_i)$-dual basis chosen such that $\int_{K_i} \psi_i(\vec{y})\, \phi_j(\vec{y})\ d\vec{y} = \delta_{ij}$.

Acosta and Borthagaray [2] showed that if $u \in H^\ell(S_K)$, where $S_K = \bigcup_{\tilde{K} \text{s.t. } \overline{K} \cap \overline{\tilde{K}} \neq \emptyset} \tilde{K}$ is the patch of an element $K \in \mathcal{P}_h$, the local approximation property

$$\int_K d\vec{x} \int_{S_K} d\vec{y} \frac{|(u - \Pi_h u)(\vec{x}) - (u - \Pi_h u)(\vec{y})|^2}{|\vec{x} - \vec{y}|^{d+2s}} \leq C h_K^{2\ell - 2s} |u|^2_{H^\ell(S_K)}$$

holds for $0 < s < \ell < 1$ and for $1/2 < s < 1$ and $1 < \ell < 2$. Below, we extend the result to include the case of $0 < s < 1/2$ and $1 \leq \ell \leq 2$ as well. This result will be needed to take advantage of higher regularity of the solution in the interior of the domain.

Lemma 35.

If $u \in H^\ell(S_K)$, for $\ell \in (1/2, 2]$ and $0 < s \leq \ell$, then

$$\int_K d\vec{x} \int_{S_K} d\vec{y} \frac{|\Pi_h u(\vec{x}) - \Pi_h u(\vec{y})|^2}{|\vec{x} - \vec{y}|^{d+2s}} \leq C \left[h_K^{-2s} \|u\|^2_{L^2(S_K)} + h_K^{2t-2s} |u|^2_{H^t(S_K)} \right], \tag{10.5}$$

$$\int_K d\vec{x} \int_{S_K} d\vec{y} \frac{|(u - \Pi_h u)(\vec{x}) - (u - \Pi_h u)(\vec{y})|^2}{|\vec{x} - \vec{y}|^{d+2s}} \leq C h_K^{2\ell - 2s} |u|^2_{H^\ell(S_K)}. \tag{10.6}$$

Proof. Similar to the proof in [36], for $t \in (1/2, \min\{1, \ell\}]$, we obtain

$$\int_K d\vec{x} \int_{S_K} d\vec{y} \frac{|\Pi_h u(\vec{x}) - \Pi_h u(\vec{y})|^2}{|\vec{x} - \vec{y}|^{d+2s}} \leq C \sum_{i \in \mathcal{I}_K} \|\psi_i\|^2_{L^\infty(K_i)} \|u\|^2_{L^1(K_i)} |\phi_i|^2_{H^s(S_K)}.$$

From [10, Theorem 4.8], we obtain

$$|\phi_i|^2_{H^s(S_K)} \leq C h_K^{d-2s},$$

and from [83, Lemma 3.1] that

$$\|\psi_i\|^2_{L^\infty(K_i)} \leq Ch_K^{-2\dim K_i}.$$

If K_i is a simplex K, then

$$\|u\|^2_{L^1(K_i)} \leq h_K^d \|u\|^2_{L^2(S_K)}.$$

On the other hand, if K_i is a sub-simplex e, then by [36, 83]

$$\|u\|^2_{L^1(K_i)} \leq C\left(h_K^{d-2}\|u\|^2_{L^2(S_K)} + h_K^{d-2+2t}|u|^2_{H^t(S_K)}\right).$$

Combining these estimates gives the following stability estimate

$$\int_K d\vec{x}\int_{S_K} d\vec{y}\frac{\left|\Pi_h u(\vec{x}) - \Pi_h u(\vec{y})\right|^2}{|\vec{x}-\vec{y}|^{d+2s}} \leq C\left[h_K^{-2s}\|u\|^2_{L^2(S_K)} + h_K^{2t-2s}|u|^2_{H^t(S_K)}\right].$$

Now, let p be a polynomial of degree 1 on S_K. Then, by $\Pi_h p = p$ and the above stability result

$$\begin{aligned}
&\int_K d\vec{x}\int_{S_K} d\vec{y}\frac{\left|(u-\Pi_h u)(\vec{x}) - (u-\Pi_h u)(\vec{y})\right|^2}{|\vec{x}-\vec{y}|^{d+2s}}\\
\leq& C\left[\int_K d\vec{x}\int_{S_K} d\vec{y}\frac{\left|(u-p)(\vec{x}) - (u-p)(\vec{y})\right|^2}{|\vec{x}-\vec{y}|^{d+2s}}\right.\\
&\left.+\int_K d\vec{x}\int_{S_K} d\vec{y}\frac{\left|(\Pi_h p-\Pi_h u)(\vec{x}) - (\Pi_h p-\Pi_h u)(\vec{y})\right|^2}{|\vec{x}-\vec{y}|^{d+2s}}\right]\\
\leq& C\left[|u-p|^2_{H^s(S_K)} + h_K^{-2s}\|u-p\|^2_{L^2(S_K)} + h_K^{2t-2s}|u-p|^2_{H^t(S_K)}\right].
\end{aligned}$$

We distinguish two cases. In the first case, assume that $\ell < 1$ and take $t = \ell$. By the Bramble-Hilbert Lemma, it holds that there exists a polynomial p of order zero such that $\|u-p\|_{L^2(S_K)} \leq Ch_K^\ell |u|_{H^\ell(S_K)}$. Moreover $|u-p|_{H^\ell(S_K)} = |u|_{H^\ell(S_K)}$ and by interpolation of spaces $|u-p|_{H^s(S_K)} \leq Ch_K^{\ell-s}|u|_{H^\ell(S_K)}$. Hence

$$\int_K d\vec{x}\int_{S_K} d\vec{y}\frac{\left|(u-\Pi_h u)(\vec{x}) - (u-\Pi_h u)(\vec{y})\right|^2}{|\vec{x}-\vec{y}|^{d+2s}} \leq Ch_K^{2\ell-2s}|u|^2_{H^\ell(S_K)}. \tag{10.7}$$

In the second case, assume that $\ell \in [1,2]$, and take $t = 1$. Again, by the Bramble-Hilbert Lemma, a polynomial p of degree one can be found such that

$\|u-p\|_{L^2(S_K)} \leq Ch_K^{\ell} |u|_{H^{\ell}(S_K)}$ and $|u-p|_{H^1(S_K)} \leq Ch_K^{\ell-1} |u|_{H^1(S_K)}$. By interpolation of spaces, its is obtained that $|u-p|_{H^s(S_K)} \leq Ch_K^{\ell-s} |u|_{H^s(S_K)}$, and therefore the local approximability result (10.7) holds in this case as well. □

The proof of the rate of convergence of the finite element solution is an immediate consequence of Lemma 35.

Lemma 36.
If $u \in H^t(\Omega) \cap H^{\ell}_{loc}(\Omega)$, for $t, \ell \in (1/2, 2]$ and $0 < s \leq t \leq \ell$, i.e. if u has Sobolev regularity t and interior regularity ℓ, then

$$\|u-u_h\|_{\widetilde{H}^s(\Omega)} \leq C\left(h^{\ell-s} |u|_{H^{\ell}_{loc}(\Omega)} + h_{\partial}^{t-s} |u|_{H^t(\Omega)}\right), \tag{10.8}$$

where h_{∂} is the maximum size of all elements K whose patch S_K touches the boundary. In particular, if the family of triangulations $\mathcal{P}_h$ is globally quasi-uniform, and $u \in H^t(\Omega)$, for $t \in (1/2, 2]$ and $0 < s \leq t$, then

$$\|u-u_h\|_{\widetilde{H}^s(\Omega)} \leq Ch^{t-s} |u|_{H^t(\Omega)}. \tag{10.9}$$

Proof. Assume that $u \in H^t(\Omega) \cap H^{\ell}_{\text{loc}}(\Omega)$, for $t, \ell \in (1/2, 2]$ and $0 < s \leq t \leq \ell$. By a localization result of Faermann [51, 52], we can estimate

$$\begin{aligned}
&\|u-\Pi_h u\|^2_{H^s(\Omega)} \\
\leq& C\sum_K \left[\int_K d\vec{x} \int_{S_K} d\vec{y} \frac{|(u-\Pi_h u)(\vec{x}) - (u-\Pi_h u)(\vec{y})|^2}{|\vec{x}-\vec{y}|^{d+2s}} + h_K^{-2s} \|u-\Pi_h u\|^2_{L^2(K)}\right].
\end{aligned} \tag{10.10}$$

The local approximability of the Scott-Zhang Interpolation (10.6) then gives

$$\begin{aligned}
\|u-\Pi_h u\|^2_{H^s(\Omega)} &\leq C\left\{\sum_{\substack{K \\ S_K \cap \partial\Omega = \emptyset}} h_K^{2\ell-2s} |u|^2_{H^{\ell}(S_K)} + \sum_{\substack{K \\ S_K \cap \partial\Omega \neq \emptyset}} h_K^{2t-2s} |u|^2_{H^t(S_K)}\right\} \\
&\leq C\left(h^{2\ell-2s} |u|^2_{H^{\ell}_{\text{loc}}(\Omega)} + h_{\partial}^{2t-2s} |u|^2_{H^t(\Omega)}\right),
\end{aligned}$$

so that we can conclude using eq. (10.4). □

Moreover, using a standard Aubin-Nitsche argument [50, Lemma 2.31] gives estimates in $L^2(\Omega)$:

Theorem 37 ([18])**.**
If the family of triangulations $\mathcal{P}_h$ is shape regular and globally quasi-uniform, and, for $\epsilon > 0$, $u \in H^{s+1/2-\epsilon}(\Omega)$, then

$$\|u - u_h\|_{L^2} \leq \begin{cases} C(s,\epsilon) h^{1/2+s-\epsilon} |u|_{H^{s+1/2-\epsilon}(\Omega)} & \text{if } 0 < s < 1/2, \\ C(s,\epsilon) h^{1-2\epsilon} |u|_{H^{s+1/2-\epsilon}(\Omega)} & \text{if } 1/2 \leq s < 1. \end{cases}$$

Estimate (10.9) implies that if $u \in H^2(\Omega)$, then the expected rate of convergence on a globally quasi-uniform mesh is $h^{2-s} \sim n^{(s-2)/d}$. However, the solutions of (10.1) generally have limited regularity as described in Theorem 34, owing to the lack of regularity in the neighborhood of the boundary. This suggests one should use a more finely graded mesh near the boundary so that $h_\partial \ll h$. The estimate (10.8) distinguishes between elements in the interior of the domain Ω and elements touching the boundary $\partial\Omega$. Generally, one would hope to restore the optimal rate of convergence $\mathcal{O}\left(n^{(s-2)/d}\right)$ observed for smooth solutions by using appropriately graded meshes.

To this end, we choose $h_\partial \sim h^{(\ell-s)/(1-2\varepsilon)}$, where h denotes the maximum element size in the interior. Applying estimate (10.8) with $t = s + 1/2 - \varepsilon$ we obtain

$$\|u - u_h\|_{\widetilde{H}^s(\Omega)} \leq C h^{\ell-s} \left(|u|_{H^\ell_{\text{loc}}(\Omega)} + |u|_{H^{1/2+s-\varepsilon}(\Omega)} \right).$$

In the one-dimensional case, $d = 1$, it is a simple matter to construct meshes for which $h_K = h$ for all elements K such that $S_K \cap \partial\Omega = \emptyset$, and $h_K = h^{(\ell-s)/(1-2\varepsilon)}$ for the four elements whose patches touch the boundary. Then the number of unknowns $n \sim h^{-1}$ and therefore

$$\|u - u_h\|_{\widetilde{H}^s(\Omega)} \leq C n^{s-\ell} \left(|u|_{H^\ell_{\text{loc}}(\Omega)} + |u|_{H^{1/2+s-\varepsilon}(\Omega)} \right). \tag{10.11}$$

This means that, provided $u \in H^2_{\text{loc}}(\Omega)$, the optimal rate of convergence is indeed achieved.

However, in the two-dimensional case one is constrained in the construction of the meshes if, as we do here, the use of hanging nodes is prohibited. This raises the possibility of the optimal rate of convergence of $\mathcal{O}\left(n^{(s-2)/d}\right)$ not being achievable in practice. Acosta and Borthagaray [2] proposed a graded mesh for the unit disc domain for approximating the solution obtained with a constant right-hand side, in which a grading parameter $\mu \geq 1$ is selected and elements are distributed in M concentric layers of radii $r_i = 1 - \left(1 - \frac{i}{M}\right)^{\mu}$, $i = 1, \ldots, M$, so that the element sizes are $h_i = r_i - r_{i-1}$. The dimension of the associated finite element space is $n \sim M^2$ if $\mu \leq 2$, and $n \sim M^{\mu}$ if $\mu > 2$. By following the arguments in [2] it is shown that

$$\|u - u_h\|_{\widetilde{H}^s(\Omega)} \leq C n^{(\max\{s-\ell,-1+\varepsilon\})/2} \left(|u|_{H^{\ell}_{\text{loc}}(\Omega)} + |u|_{H^{1/2+s-\varepsilon}(\Omega)} \right). \tag{10.12}$$

This suggests that the optimal rate in $d = 2$ dimensions is $n^{-1/2+\varepsilon}$, which can be achieved if u has interior regularity of at least order $1 + s - \varepsilon$. We note that contrary to the result stated in [2], this holds for all $s \in (0, 1)$, provided that the solution u enjoys sufficient interior regularity.

A priori mesh grading of the type discussed above can give considerable improvements but does rely on the availability of sharp estimates for the regularity which, unfortunately, is not always the case for practical problems on complicated domains. Therefore, we explore a posteriori error indicators for the fractional Laplacian in Section 10.4.

10.4 Adaptive Mesh Refinement

The fractional Laplacian $(-\Delta)^s : \widetilde{H}^s(\Omega) \to H^{-s}(\Omega)$ is continuous with continuous inverse. Hence, the discretization error

$$e = u - u_h$$

can be bounded from above and below in terms of the residual $r = f - (-\Delta)^s u_h$ as

$$c \|r\|_{H^{-s}(\Omega)} \leq \|e\|_{\widetilde{H}^s(\Omega)} \leq C \|r\|_{H^{-s}(\Omega)},$$

where c and C are positive constants depending on the norm of $(-\Delta)^s$ and its inverse. It was shown in [51, 52] that the Babuška-Rheinboldt type error indicators

$$\eta_i := \sup_{\substack{v \in \widetilde{H}^s(\Omega) \\ v\phi_i \neq 0}} \frac{|a(e, v\phi_i)|}{\|v\phi_i\|_{\widetilde{H}^s(\Omega)}}$$

are reliable and efficient for non-negative order differential operators. Unfortunately, the estimators are not computable owing to the presence of the supremum. However, η_i can be estimated up to a generic (unknown) factor, independent of the solution and the mesh-size, by

$$\hat{\eta}_i := \sqrt{\sum_{K \in S_i} h_K^{2s} \|r\|^2_{L^2(K)}}, \tag{10.13}$$

where S_i is the patch formed by the elements which contain the node $\vec{z}_i$. Therefore, the following a posteriori estimate holds:

$$\|e\|^2_{\widetilde{H}^s(\Omega)} \leq \hat{\eta} := \sum_i \hat{\eta}_i^2.$$

The estimator (10.13) involves only the local L^2-norms of the residual and can easily be evaluated using quadrature, but then one requires pointwise values for $(-\Delta)^s u_h$. A direct evaluation using quadrature rules is discouraged due to the presence of the singularity of the kernel $|\vec{x} - \vec{y}|^{-d-2s}$ as well as the fact that the domain of integration is unbounded. The next result gives a method whereby this issue can be circumvented.

Lemma 38.

For $u \in V_h$ and $\vec{x} \in K_0$ for an element $K_0 \in \mathcal{P}_h$, the fractional Laplacian $(-\Delta)^s u$ can be evaluated as

$$\begin{aligned}\frac{(-\Delta)^s u(\vec{x})}{C(d,s)} = &\frac{1}{d+2s-2} \int_{\partial K_0} \frac{\nabla u\big|_{K_0} \cdot \vec{n}_{\vec{y}}}{|\vec{x} - \vec{y}|^{d+2s-2}} \, d\vec{y} - \frac{u(\vec{x})}{2s} \int_{\partial K_0} \frac{\vec{n}_{\vec{y}} \cdot (\vec{x} - \vec{y})}{|\vec{x} - \vec{y}|^{d+2s}} \, d\vec{y} \\ &+ \sum_{K \neq K_0} \left\{ \frac{1}{2s(d+2s-2)} \int_{\partial K} \nabla u\big|_K \cdot \vec{n}_{\vec{y}} \frac{1}{|\vec{x} - \vec{y}|^{d+2s-2}} \, d\vec{y} \right.\end{aligned} \tag{10.14}$$

$$-\frac{1}{2s}\int_{\partial K} u\left(\vec{y}\right)\frac{\vec{n}_{\vec{y}}\cdot\left(\vec{x}-\vec{y}\right)}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y}\Bigg\} \tag{10.15}$$

if $s \neq 1-d/2$*, and as*

$$\begin{aligned}\frac{\left(-\Delta\right)^{s} u\left(\vec{x}\right)}{C\left(d,s\right)} &= \int_{\partial K_0} \nabla u\big|_{K_0}\cdot\vec{n}_{\vec{y}}\log\frac{1}{\left|\vec{x}-\vec{y}\right|}\,d\vec{y} - \frac{u\left(\vec{x}\right)}{2s}\int_{\partial K_0}\frac{\vec{n}_{\vec{y}}\cdot\left(\vec{x}-\vec{y}\right)}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y}\\ &\quad+\sum_{K\neq K_0}\left\{\frac{1}{2s}\int_{\partial K}\nabla u\big|_{K}\cdot\vec{n}_{\vec{y}}\log\frac{1}{\left|\vec{x}-\vec{y}\right|}\,d\vec{y} - \frac{1}{2s}\int_{\partial K} u\left(\vec{y}\right)\frac{\vec{n}_{\vec{y}}\cdot\left(\vec{x}-\vec{y}\right)}{\left|\vec{x}-\vec{y}\right|^{2}}\,d\vec{y}\right\}.\end{aligned} \tag{10.16}$$

when $s = 1-d/2$*.*

Proof. Let $u \in V_h$ and $\vec{x} \in K_0$, for some $K_0 \in \mathcal{P}_h$. Then

$$\begin{aligned}\frac{\left(-\Delta\right)^{s} u\left(\vec{x}\right)}{C\left(d,s\right)} &= \text{p. v.}\int_{\mathbb{R}^d}\frac{u\left(\vec{x}\right)-u\left(\vec{y}\right)}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y}\\ &= \text{p. v.}\int_{K_0}\frac{u\left(\vec{x}\right)-u\left(\vec{y}\right)}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y} + u\left(\vec{x}\right)\int_{\mathbb{R}^d\setminus K_0}\frac{1}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y}\\ &\quad-\sum_{K\neq K_0}\int_{K}\frac{u\left(\vec{y}\right)}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y}.\end{aligned} \tag{10.17}$$

We have the following helpful identities:

$$\frac{\vec{x}-\vec{y}}{\left|\vec{x}-\vec{y}\right|^{d+2s}} = \begin{cases}\frac{1}{d+2s-2}\nabla_{\vec{y}}\frac{1}{\left|\vec{x}-\vec{y}\right|^{d+2s-2}}, & s\neq 1-d/2,\\ \nabla_{\vec{y}}\log\frac{1}{\left|\vec{x}-\vec{y}\right|}, & s = 1-d/2,\end{cases} \tag{10.18}$$

$$\frac{1}{\left|\vec{x}-\vec{y}\right|^{d+2s}} = \frac{1}{2s}\nabla_{\vec{y}}\cdot\frac{\vec{x}-\vec{y}}{\left|\vec{x}-\vec{y}\right|^{d+2s}}, \tag{10.19}$$

$$\frac{1}{\left|\vec{x}-\vec{y}\right|^{d+2s}} = \begin{cases}\frac{1}{2s(d+2s-2)}\Delta_{\vec{y}}\frac{1}{\left|\vec{x}-\vec{y}\right|^{d+2s-2}}, & s\neq 1-d/s,\\ \frac{1}{d-2}\Delta_{\vec{y}}\log\frac{1}{\left|\vec{x}-\vec{y}\right|}, & s = 1-d/2.\end{cases} \tag{10.20}$$

For $\vec{x},\vec{y}\in K_0$,

$$u\left(\vec{x}\right)-u\left(\vec{y}\right) = \nabla u\big|_{K_0}\cdot\left(\vec{x}-\vec{y}\right)$$

since $u\big|_{K_0}$ is linear. Hence, the first term of (10.17) can be rewritten using (10.18) as

$$\text{p. v.}\int_{K_0}\frac{u\left(\vec{x}\right)-u\left(\vec{y}\right)}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y} = \lim_{\varepsilon\to 0}\nabla u\big|_{K_0}\cdot\int_{K_0\setminus B(\vec{x},\varepsilon)}\frac{\vec{x}-\vec{y}}{\left|\vec{x}-\vec{y}\right|^{d+2s}}\,d\vec{y}$$

$$= \lim_{\varepsilon \to 0} \left[\nabla u\big|_{K_0} \cdot \frac{1}{d+2s-2} \int_{\partial K_0} \frac{\vec{n}_{\vec{y}}}{|\vec{x}-\vec{y}|^{d+2s-2}} \, d\vec{y} \right.$$

$$\left. + \nabla u\big|_{K_0} \cdot \frac{1}{d+2s-2} \underbrace{\int_{\partial B(\vec{x},\varepsilon)} \frac{\vec{n}_{\vec{y}}}{|\vec{x}-\vec{y}|^{d+2s-2}} \, d\vec{y}}_{=0} \right]$$

$$= \frac{1}{d+2s-2} \int_{\partial K_0} \frac{\nabla u\big|_{K_0} \cdot \vec{n}_{\vec{y}}}{|\vec{x}-\vec{y}|^{d+2s-2}} \, d\vec{y}.$$

Here, $\vec{n}_{\vec{y}}$ is the outward normal at $\vec{y}$. When $s = 1 - d/2$, we obtain instead that

$$\text{p. v.} \int_{K_0} \frac{u(\vec{x}) - u(\vec{y})}{|\vec{x}-\vec{y}|^2} \, d\vec{y} = \int_{\partial K_0} \nabla u\big|_{K_0} \cdot \vec{n}_{\vec{y}} \log \frac{1}{\left|\vec{x-y}\right|} \, d\vec{y}.$$

The second term of (10.17) can be rewritten using (10.19) as

$$\int_{\mathbb{R}^d \setminus K_0} \frac{1}{|\vec{x}-\vec{y}|^{d+2s}} \, d\vec{y} = \frac{1}{2s} \int_{\mathbb{R}^d \setminus K_0} \nabla_{\vec{y}} \frac{\vec{x}-\vec{y}}{|\vec{x}-\vec{y}|^{d+2s}} \, d\vec{y}$$
$$= -\frac{1}{2s} \int_{\partial K_0} \frac{\vec{n}_{\vec{y}} \cdot (\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}} \, d\vec{y}.$$

Finally, for every $K \in \mathcal{P}_h$, $K \neq K_0$, it follows from (10.20), Green's second identity and (10.18) that

$$\int_K \frac{u(\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}} \, d\vec{y} = \frac{1}{2s(d+2s-2)} \int_K u(\vec{y}) \, \Delta_{\vec{y}} \frac{1}{|\vec{x}-\vec{y}|^{d+2s-2}} \, d\vec{y}$$
$$= \frac{1}{2s(d+2s-2)} \int_K \underbrace{(\Delta_{\vec{y}} u(\vec{y}))}_{=0} \frac{1}{|\vec{x}-\vec{y}|^{d+2s-2}} \, d\vec{y}$$
$$+ \frac{1}{2s(d+2s-2)} \int_{\partial K} \left\{ u(\vec{y}) \, (d+2s-2) \frac{\vec{n}_{\vec{y}} \cdot (\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}} \right.$$
$$\left. -\nabla u\big|_K \cdot \vec{n}_{\vec{y}} \frac{1}{|\vec{x}-\vec{y}|^{d+2s-2}} \right\} \, d\vec{y}$$
$$= \frac{1}{2s} \int_{\partial K} u(\vec{y}) \, \frac{\vec{n}_{\vec{y}} \cdot (\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}} \, d\vec{y}$$
$$- \frac{1}{2s(d+2s-2)} \int_{\partial K} \nabla u\big|_K \cdot \vec{n}_{\vec{y}} \frac{1}{|\vec{x}-\vec{y}|^{d+2s-2}} \, d\vec{y}.$$

When $s = 1 - d/2$, we obtain instead that

$$\int_K \frac{u(\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}}\, d\vec{y} = \frac{1}{2s}\int_{\partial K} u(\vec{y})\,\frac{\vec{n}_{\vec{y}}\cdot(\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^2}\, d\vec{y} - \frac{1}{2s}\int_{\partial K} \nabla u\big|_K \cdot \vec{n}_{\vec{y}} \log\frac{1}{|\vec{x}-\vec{y}|}\, d\vec{y}.$$

In conclusion, we find that

$$\begin{aligned}\frac{(-\Delta)^s u(\vec{x})}{C(d,s)} = {} & \frac{1}{d+2s-2}\int_{\partial K_0} \frac{\nabla u\big|_{K_0}\cdot\vec{n}_{\vec{y}}}{|\vec{x}-\vec{y}|^{d+2s-2}}\, d\vec{y} - \frac{u(\vec{x})}{2s}\int_{\partial K_0}\frac{\vec{n}_{\vec{y}}\cdot(\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}}\, d\vec{y} \\ & + \sum_{K\neq K_0}\left\{\frac{1}{2s(d+2s-2)}\int_{\partial K}\nabla u\big|_K\cdot\vec{n}_{\vec{y}}\frac{1}{|\vec{x}-\vec{y}|^{d+2s-2}}\, d\vec{y}\right. \\ & \left. -\frac{1}{2s}\int_{\partial K} u(\vec{y})\,\frac{\vec{n}_{\vec{y}}\cdot(\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}}\, d\vec{y}\right\}\end{aligned}$$

and when $s = 1 - d/2$ then

$$\begin{aligned}\frac{(-\Delta)^s u(\vec{x})}{C(d,s)} = {} & \int_{\partial K_0}\nabla u\big|_{K_0}\cdot\vec{n}_{\vec{y}}\log\frac{1}{\left|\vec{x-y}\right|}\, d\vec{y} - \frac{u(\vec{x})}{2s}\int_{\partial K_0}\frac{\vec{n}_{\vec{y}}\cdot(\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{d+2s}}\, d\vec{y} \\ & + \sum_{K\neq K_0}\left\{\frac{1}{2s}\int_{\partial K}\nabla u\big|_K\cdot\vec{n}_{\vec{y}}\log\frac{1}{|\vec{x}-\vec{y}|}\, d\vec{y} - \frac{1}{2s}\int_{\partial K} u(\vec{y})\,\frac{\vec{n}_{\vec{y}}\cdot(\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^2}\, d\vec{y}\right\}.\end{aligned}$$

□

The integrals over the element surface in eqs. (10.15) and (10.16) can be evaluated using standard quadrature rules since the integrals in eqs. (10.15) and (10.16) are regular. Although we are primarily concerned here with the case of piecewise linear basis functions, the above expression can easily be generalized to higher orders finite elements.

Chapter 11

Overcoming Computational Challenges of the Fractional Laplacian

11.1 Computation of Entries of the Stiffness Matrix

The computation of entries of the stiffness matrix $\boldsymbol{A}^s$ in the case of the usual Laplacian ($s = 1$) is straightforward. However, for $s \in (0, 1)$, the bilinear form contains factors $|\vec{x} - \vec{y}|^{-d-2s}$ which means that simple closed forms for the entries are no longer available and suitable quadrature rules therefore must be identified. Moreover, the presence of a repeated integral over Ω (as opposed to an integral over just Ω in the case $s = 1$) means that the matrix needs to be assembled in a double loop over the elements of the mesh so that the computational cost is potentially much larger than in the integer $s = 1$ case. Additionally, every degree of freedom is coupled to all other degrees of freedom and the stiffness matrix is therefore dense.

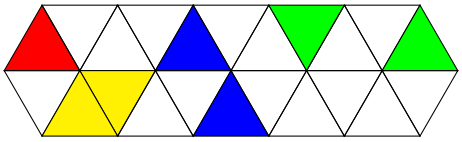


Figure 11.1: Element pairs that are treated separately. We distinguish element pairs of identical elements (*red*), element pairs with common edge (*yellow*), with common vertex (*blue*) and separated elements (*green*).

11.1.1 Reduction to Smooth Integrals

In order to compute the entries of $\boldsymbol{A}^s = \{a(\phi_i, \phi_j)\}_{ij}$ we decompose the expression for the entries into contributions from elements $K, \tilde{K} \in \mathcal{P}_h$ and external edges $e \in \mathcal{P}_{h,\partial}$:

$$a(\phi_i, \phi_i) = \sum_K \sum_{\tilde{K}} a^{K\times\tilde{K}}(\phi_i, \phi_j) + \sum_K \sum_e a^{K\times e}(\phi_i, \phi_j),$$

where the contributions $a^{K\times\tilde{K}}$ and $a^{K\times e}$ are given by:

$$a^{K\times\tilde{K}}(\phi_i, \phi_j) = \frac{C(d,s)}{2} \int_K d\vec{x} \int_{\tilde{K}} d\vec{y} \frac{(\phi_i(\vec{x}) - \phi_i(\vec{y}))\,(\phi_j(\vec{x}) - \phi_j(\vec{y}))}{|\vec{x} - \vec{y}|^{d+2s}}, \tag{11.1}$$

$$a^{K\times e}(\phi_i, \phi_j) = \frac{C(d,s)}{2s} \int_K d\vec{x} \int_e d\vec{y} \frac{\phi_i(\vec{x})\,\phi_j(\vec{x})\;\vec{n}_e \cdot (\vec{x} - \vec{y})}{|\vec{x} - \vec{y}|^{d+2s}}. \tag{11.2}$$

Although the following approach holds for arbitrary spatial dimension d, we restrict ourselves to $d = 2$ dimensions. In evaluating the contributions $a^{K\times\tilde{K}}$ over element pairs $K \times \tilde{K}$, several cases need to be distinguished:

1. K and $\tilde{K}$ have empty intersection,
2. K and $\tilde{K}$ are identical,
3. K and $\tilde{K}$ share an edge,
4. K and $\tilde{K}$ share a vertex.

These cases are illustrated in Figure 11.1. In case 1, where the elements do not touch, the Stroud conical quadrature rule [92] (or any other suitable Gauss rule on

simplices) of *sufficiently high order* can be used to approximate the integrals. More details as to what constitutes a sufficiently high order are given in Section 11.1.2.

Special care has to be taken in the remaining cases 2-4, in which the elements are touching, owing to the presence of a singularity in the integrand. The contributions $a^{K\times\tilde{K}}$ and $a^{K\times e}$ as given in eqs. (11.1) and (11.2) for touching elements K and $\tilde{K}$ contain singularities. Fortunately, the singularity is removable and can, as pointed out in [2], be treated using standard techniques from the boundary element literature [82]. More specifically, we write the integral as a sum of integrals over sub-simplices. Each sub-simplex is then mapped onto the hyper-cube $[0,1]^4$ using the Duffy transformation [44]. The advantage of pursuing this approach is that the singularity arising from the degenerate nature of the Duffy transformation offsets the singularity present in the integrals. Removing the singularity means that the integrals in eqs. (11.1) and (11.2) will become amenable to approximation using standard Gaussian quadrature rules of *sufficiently high order* as discussed in Section 11.1.2. The same idea is applicable in any number of space dimensions.

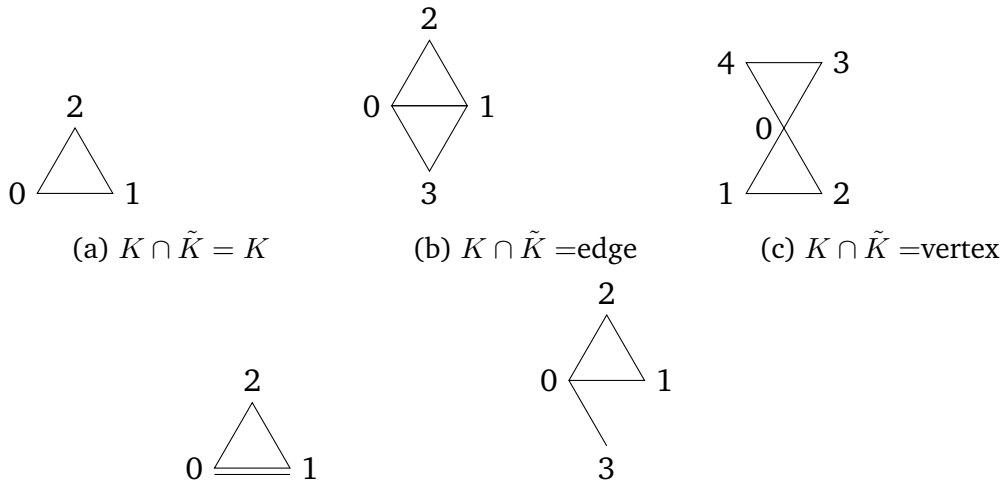
The derivations of the terms involved can be found in [1, 82] and is outlined below for $d = 2$ dimensions.

The expression for $a^{K\times\tilde{K}}$ can be transformed into integrals over the reference element $\hat{K}$:

$$
\begin{aligned}
&a^{K\times\tilde{K}}(\phi_i,\phi_j)\\
=&\frac{C(2,s)}{2}\int_K d\vec{x}\int_{\tilde{K}} d\vec{y}\frac{\left(\phi_i(\vec{x})-\phi_i(\vec{y})\right)\left(\phi_j(\vec{x})-\phi_j(\vec{y})\right)}{|\vec{x}-\vec{y}|^{2+2s}}\\
=&\frac{C(2,s)}{2}\frac{|K|}{\left|\hat{K}\right|}\frac{\left|\tilde{K}\right|}{\left|\hat{K}\right|}\int_{\hat{K}} d\hat{\vec{x}}\int_{\hat{K}} d\hat{\vec{y}}\frac{\left(\phi_i(\vec{x}\left(\hat{\vec{x}}\right))-\phi_i(\vec{y}\left(\hat{\vec{y}}\right))\right)\left(\phi_j(\vec{x}\left(\hat{\vec{x}}\right))-\phi_j(\vec{y}\left(\hat{\vec{y}}\right))\right)}{\left|\vec{x}\left(\hat{\vec{x}}\right)-\vec{y}\left(\hat{\vec{y}}\right)\right|^{2+2s}}.
\end{aligned}
$$

Similarly, by introducing the reference edge $\hat{e}$, we obtain

$$
a^{K\times e}(\phi_i,\phi_j)=\frac{C(2,s)}{2s}\int_K d\vec{x}\int_e d\vec{y}\frac{\phi_i\left(\vec{x}\right)\phi_j\left(\vec{x}\right)\ \vec{n}_e\cdot(\vec{x}-\vec{y})}{|\vec{x}-\vec{y}|^{2+2s}}
$$



(a) $K \cap \tilde{K} = K$ (b) $K \cap \tilde{K} =$edge (c) $K \cap \tilde{K} =$vertex

(d) $K \cap e = e$ (e) $K \cap e =$vertex

Figure 11.2: Numbering of local nodes for touching triangular elements K and $\tilde{K}$ or element K and edge e.

$$= \frac{C(2,s)}{2s} \frac{|K|}{\left|\hat{K}\right|} \frac{|e|}{|\hat{e}|} \int_{\hat{K}} d\hat{\vec{x}} \int_{\hat{e}} d\hat{\vec{y}} \frac{\phi_i\left(\vec{x}\left(\hat{\vec{x}}\right)\right) \phi_j\left(\vec{x}\left(\hat{\vec{x}}\right)\right) \ \vec{n}_e \cdot \left(\vec{x}\left(\hat{\vec{x}}\right) - \vec{y}\left(\hat{\vec{y}}\right)\right)}{\left|\vec{x}\left(\hat{\vec{x}}\right) - \vec{y}\left(\hat{\vec{y}}\right)\right|^{2+2s}}$$

for touching elements K and edges e. If K and $\tilde{K}$ or e have $c \geq 1$ common vertices, and if we designate by λ_k, $k = 0, \ldots, 6-c$ the barycentric coordinates of $K \cup \tilde{K}$ or $K \cup e$ respectively (cf. Figure 11.2), we have

$$\lambda_{k(i)}(\hat{\vec{x}}) = \phi_i\left(\vec{x}\left(\hat{\vec{x}}\right)\right),$$

where $k(i)$ is the local index on $K \cup \tilde{K}$ or $K \cup e$ of the global degree of freedom i. Moreover, we have that

$$\begin{aligned} \vec{x}\left(\hat{\vec{x}}\right) - \vec{y}\left(\hat{\vec{y}}\right) &= \sum_{k=0}^{6-c} \lambda_k\left(\hat{\vec{x}}\right) \vec{x}_k - \sum_{k=0}^{6-c} \lambda_k\left(\hat{\vec{y}}\right) \vec{y}_k \\ &= \sum_{k=0}^{6-c} \left[\lambda_k\left(\hat{\vec{x}}\right) - \lambda_k\left(\hat{\vec{y}}\right)\right] \vec{x}_k. \end{aligned}$$

Here, $\vec{x}_k$, $k = 0, \ldots, 6-c$ are the vertices that span $K \cup \tilde{K}$ or $K \cup e$ respectively.

By setting

$$\psi_k\left(\hat{\vec{x}}, \hat{\vec{y}}\right) := \lambda_k\left(\hat{\vec{x}}\right) - \lambda_k\left(\hat{\vec{y}}\right),$$

we can therefore write

$$a^{K\times\tilde{K}}(\phi_i,\phi_j)=\frac{C(2,s)}{2}\frac{|K|}{\left|\hat{K}\right|}\frac{\left|\tilde{K}\right|}{\left|\hat{K}\right|}\int_{\hat{K}}d\hat{\vec{x}}\int_{\hat{K}}d\hat{\vec{y}}\frac{\psi_{k(i)}\left(\hat{\vec{x}},\hat{\vec{y}}\right)\psi_{k(j)}\left(\hat{\vec{x}},\hat{\vec{y}}\right)}{\left|\sum_{k=0}^{6-c}\psi_k\left(\hat{\vec{x}},\hat{\vec{y}}\right)\vec{x}_k\right|^{2+2s}}.$$

By carefully splitting the integration domain $\hat{K}\times\hat{K}$ into L_c parts and applying a Duffy transformation to each part, the contributions can be rewritten into integrals over a unit hyper-cube, where the singularities are lifted.

$$a^{K\times\tilde{K}}(\phi_i,\phi_j)=\frac{C(2,s)}{2}\frac{|K|}{\left|\hat{K}\right|}\frac{\left|\tilde{K}\right|}{\left|\hat{K}\right|}\sum_{\ell=1}^{L_c}\int_{[0,1]^4}d\vec{\eta}\;\bar{J}^{(\ell,c)}\frac{\bar{\psi}_{k(i)}^{(\ell,c)}\left(\vec{\eta}\right)\bar{\psi}_{k(j)}^{(\ell,c)}\left(\vec{\eta}\right)}{\left|\sum_{k=0}^{2d-c}\bar{\psi}_k^{(\ell,c)}\left(\vec{\eta}\right)\vec{x}_k\right|^{2+2s}}. \quad (11.3)$$

The details of this approach can be found in Chapter 5 of [82] for the interactions between K and $\tilde{K}$. We record the obtained expressions in this case.

- K and $\tilde{K}$ are identical, i.e. $c=3$

$$L_3=3,\qquad \bar{J}^{(1,3)}=\bar{J}^{(2,3)}=\bar{J}^{(3,3)}=\eta_0^{3-2s}\eta_1^{2-2s}\eta_2^{1-2s},$$

$$\bar{\psi}_k^{(1,3)}=\begin{cases}-\eta_3\\ \eta_3-1\\ 1\end{cases}\qquad \bar{\psi}_k^{(2,3)}=\begin{cases}-1\\ 1-\eta_3\\ \eta_3\end{cases}\qquad \bar{\psi}_k^{(3,3)}=\begin{cases}\eta_3\\ -1\\ 1-\eta_3\end{cases}$$

- K and $\tilde{K}$ share an edge, i.e. $c=2$

$$L_2=5,\qquad \bar{J}^{(1,2)}=\eta_0^{3-2s}\eta_1^{2-2s},$$
$$\bar{J}^{(2,2)}=\bar{J}^{(3,2)}=\bar{J}^{(4,2)}=\bar{J}^{(5,2)}=\eta_0^{3-2s}\eta_1^{2-2s}\eta_2$$

$$\bar{\psi}_k^{(1,2)}=\begin{cases}-\eta_2\\ 1-\eta_3\\ \eta_3\\ \eta_2-1\end{cases}\qquad \bar{\psi}_k^{(2,2)}=\begin{cases}-\eta_2\eta_3\\ \eta_2-1\\ 1\\ \eta_2\eta_3-\eta_2\end{cases}\qquad \bar{\psi}_k^{(3,2)}=\begin{cases}\eta_2\\ \eta_2\eta_3-1\\ 1-\eta_2\\ -\eta_2\eta_3\end{cases}$$

$$\bar{\psi}_k^{(4,2)} = \begin{cases} \eta_2\eta_3 \\ 1-\eta_2 \\ \eta_2 - \eta_2\eta_3 \\ -1 \end{cases} \qquad \bar{\psi}_k^{(5,2)} = \begin{cases} \eta_2\eta_3 \\ \eta_2 - 1 \\ 1-\eta_2\eta_3 \\ -\eta_2 \end{cases}$$

- K and $\tilde{K}$ share a vertex, i.e. $c = 1$

$$L_1 = 2, \qquad \bar{J}^{(1,1)} = \bar{J}^{(2,1)} = \eta_0^{3-2s}\eta_2$$

$$\bar{\psi}_k^{(1,1)} = \begin{cases} \eta_2 - 1 \\ 1-\eta_1 \\ \eta_1 \\ \eta_2\eta_3 - \eta_2 \\ -\eta_2\eta_3 \end{cases} \qquad \bar{\psi}_k^{(2,1)} = \begin{cases} 1-\eta_2 \\ \eta_2 - \eta_2\eta_3 \\ \eta_2\eta_3 \\ \eta_1 - 1 \\ -\eta_1 \end{cases}$$

We notice that the contributions for identical elements only depend on η_3, so that in fact only one-dimensional integrals needs to be computed. Similarly, the cases of common edges or common vertices only require two and three dimensional integration.

In a similar fashion, the integration domain of $a^{K\times e}$ can be split into several parts, so that the singularity can be lifted:

$$a^{K\times e}(\phi_i,\phi_j) = \frac{C(2,s)}{2s}\frac{|K|}{\left|\hat{K}\right|}\frac{|e|}{|\hat{e}|}\int_{[0,1]^3} d\vec{\eta}\frac{\phi_{k(i)}^{(\ell,c)}(\vec{\eta})\,\phi_{k(j)}^{(\ell,c)}(\vec{\eta})\;\sum_{k=0}^{5-c}\bar{\psi}_k^{(\ell,c)}(\vec{\eta})\,\vec{n}_e\cdot\vec{x}_k}{\left|\sum_{k=0}^{5-c}\bar{\psi}_k^{(\ell,c)}(\vec{\eta})\,\vec{x}_k\right|^{2+2s}}. \tag{11.4}$$

Here, $\phi_k^{(\ell,c)}$ are the expressions for the local shape functions under the Duffy transformations. The obtained expressions are

- e is an edge of K, i.e. $c = 2$

$$L_2 = 3, \qquad \bar{J}^{(1,2)} = \bar{J}^{(2,2)} = \bar{J}^{(3,2)} = \eta_0^{-2s}\left(1-\eta_0\right),$$

$$\phi_k^{(1,2)} = \begin{cases} 1-\eta_0-\eta_2+\eta_0\eta_2 \\ \eta_0+\eta_2-\eta_0\eta_1-\eta_0\eta_1 \\ \eta_0\eta_1 \end{cases} \qquad \phi_k^{(2,2)} = \begin{cases} 1-\eta_0-\eta_2+\eta_0\eta_2 \\ \eta_2-\eta_0\eta_2 \\ \eta_0 \end{cases}$$

$$\phi_k^{(3,2)} = \begin{cases} 1-\eta_2+\eta_0\eta_2-\eta_0\eta_1 \\ \eta_2-\eta_0\eta_2 \\ \eta_0\eta_1 \end{cases}$$

$$\bar{\psi}_k^{(1,2)} = \begin{cases} -1 \\ 1-\eta_1 \\ \eta_1 \end{cases} \qquad \bar{\psi}_k^{(2,2)} = \begin{cases} -\eta_1 \\ \eta_1-1 \\ 1 \end{cases} \qquad \bar{\psi}_k^{(3,2)} = \begin{cases} 1-\eta_1 \\ -1 \\ \eta_1 \end{cases}$$

We notice that for $s \geq 1/2$, the integrand still contains a singularity. In this case, the finite element space V_h does not include the degrees of freedom on the boundary. For the interaction of the single degree of freedom that is not on the boundary ($k = 2$), we obtain

$$\bar{J}^{(1,2)} = \bar{J}^{(2,2)} = \bar{J}^{(3,2)} = \eta_0^{2-2s}\left(1-\eta_0\right),$$

$$\phi_2^{(1,2)} = \eta_1 \qquad \phi_2^{(2,2)} = 1 \qquad \phi_2^{(3,2)} = \eta_1$$

and $\bar{\psi}_2^{\ell,c}$ as above.

- K and e share a vertex, i.e. $c = 1$

$$L_1 = 2, \qquad \bar{J}^{(1,1)} = \eta_0^{1-2s}, \bar{J}^{(2,1)} = \eta_0^{1-2s}\eta_1$$

$$\bar{\psi}_k^{(1,1)} = \begin{cases} \eta_2 - 1 \\ 1 - \eta_1 \\ \eta_1 \\ -\eta_2 \end{cases} \qquad \bar{\psi}_k^{(2,1)} = \begin{cases} 1 - \eta_1 \\ \eta_1 - \eta_1\eta_2 \\ \eta_1\eta_2 \\ -1 \end{cases}$$

11.1.2 Determining the Order of the Quadrature Rules

The foregoing considerations show that the evaluation of the entries of the stiffness matrix boils down to the evaluation of integrals with smooth integrands, i.e. expressions eqs. (11.1) and (11.2) for case 1 and expressions eqs. (11.3) and (11.4) for case 2-4. As mentioned earlier, it is necessary to use a *sufficiently high order* quadrature rule to approximate these integrals. We now turn to the question of how high is sufficient.

The arguments used to prove the ensuing estimates follow a pattern similar to the proofs of Theorems 5.3.29, 5.3.23 and 5.3.24 in [82]. The main difference from [82] is the presence of the boundary integral term.

Theorem 39 ([82], Theorems 5.3.23 and 5.3.24)**.**
If K and $\tilde{K}$ (K and e) are touching elements, then

$$\left|E^{i,j}_{K\times\tilde{K}}\right| \leq C h_K^{2-2s} \rho_1^{-2k_T},$$
$$\left|E^{i,j}_{K\times e}\right| \leq C h_K^{2-2s} \rho_3^{-2k_{T,\partial}},$$

where $\rho_1, \rho_3 > 1$ and k_T, $k_{T,\partial}$ are the quadrature orders in every dimension of eqs. (11.1) *and* (11.2) *(after regularization).*

If K and $\tilde{K}$ (K and e) are not touching, then

$$\left|E^{i,j}_{K\times\tilde{K}}\right| \leq C \left(h_K h_{\tilde{K}}\right)^2 d_{K,\tilde{K}}^{-2-2s} \left\{ \left(\frac{d_{K,\tilde{K}}}{h_K}\right)^2 \tilde{\rho}_2\left(K,\tilde{K}\right)^{-2k_{NT}} + \left(\frac{d_{K,\tilde{K}}}{h_{\tilde{K}}}\right)^2 \tilde{\rho}_2\left(\tilde{K},K\right)^{-2k_{NT}} \right\},$$

$$\left|E_{K\times e}^{i,j}\right| \leq C\left(h_K h_e\right)^2 d_{K,e}^{-2-2s}\left\{\left(\frac{d_{K,e}}{h_K}\right)^2 \tilde{\rho}_4\left(K,e\right)^{-2k_{NT}} + \left(\frac{d_{K,e}}{h_e}\right)^2 \tilde{\rho}_4\left(e,K\right)^{-2k_{NT}}\right\},$$

where $d_{K,\tilde{K}} := dist(K,\tilde{K})$, $d_{K,e} := dist(K,e)$, $\tilde{\rho}_2(K,\tilde{K}) := \rho_2 \max\left\{\frac{d_{K,\tilde{K}}}{h_K}, 1\right\}$, $\tilde{\rho}_4(K,e) := \rho_4 \max\left\{\frac{d_{K,e}}{h_K}, 1\right\}$ *and* $\rho_2, \rho_4 > 1$, *and* k_{NT}, $k_{NT,\partial}$ *are the quadrature order in every dimension of eqs.* (11.1) *and* (11.2).

Theorem 40.

For $d = 2$, *let* $\mathcal{I}_K$ *index the degrees of freedom on* $K \in \mathcal{P}_h$, *and define* $\mathcal{I}_{K\times\tilde{K}} := \mathcal{I}_K \cup \mathcal{I}_{\tilde{K}}$. *Let* $k_T\left(K,\tilde{K}\right)$ *(respectively* $k_{T,\partial}\left(K,e\right)$*) be the quadrature order used for touching pairs* $K\times\tilde{K}$ *(respectively* $K\times e$*), and let* $k_{NT}\left(K,\tilde{K}\right)$ *(respectively* $k_{NT,\partial}\left(K,e\right)$*) be the quadrature order used for pairs that have empty intersection. Denote the resulting approximation to the bilinear form* $a\left(\cdot,\cdot\right)$ *by* $a^Q\left(\cdot,\cdot\right)$. *Then the consistency error due to quadrature is bounded by*

$$\left|a(u,v) - a_Q(u,v)\right| \leq C\left(E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial}\right)\left\|u\right\|_{L^2(\Omega)}\left\|v\right\|_{L^2(\Omega)} \quad \forall u,v \in V_h,$$

where the errors are given by

$$\begin{aligned}
E_T &= n \max_{K,\tilde{K}\in\mathcal{P}_h, \overline{K}\cap\overline{\tilde{K}}\neq\emptyset} h_K^{-2s}\rho_1^{-2k_T\left(K,\tilde{K}\right)},\\
E_{NT} &= n \max_{K,\tilde{K}\in\mathcal{P}_h, \overline{K}\cap\overline{\tilde{K}}=\emptyset}\left[h_{\tilde{K}}^2\tilde{\rho}_2\left(K,\tilde{K}\right)^{-2k_{NT}\left(K,\tilde{K}\right)} + h_K^2\tilde{\rho}_2\left(\tilde{K},K\right)^{-2k_{NT}\left(K,\tilde{K}\right)}\right]\\
&\qquad d_{K,\tilde{K}}^{-2s}\max\left\{h_K^{-2}, h_{\tilde{K}}^{-2}\right\},\\
E_{T,\partial} &= n^{1/2}\max_{K\in\mathcal{P}_h, e\in\partial\mathcal{P}_h, \overline{K}\cap\overline{e}\neq\emptyset} h_K^{-2s}\rho_3^{-2k_{T,\partial}(K,e)},\\
E_{NT,\partial} &= n^{1/2}\max_{K\in\mathcal{P}_h, e\in\partial\mathcal{P}_h, \overline{K}\cap\overline{e}=\emptyset}\left[h_e^2\tilde{\rho}_4\left(K,e\right)^{-2k_{NT}(K,e)} + h_K^2\tilde{\rho}_4\left(e,K\right)^{-2k_{NT}(K,e)}\right] d_{K,e}^{-2s}h_K^{-2},
\end{aligned}$$

and where $d_{K,\tilde{K}} := \inf_{\vec{x}\in K,\vec{y}\in\tilde{K}}\left|\vec{x}-\vec{y}\right|$, $d_{K,e} := \inf_{\vec{x}\in K,\vec{y}\in e}\left|\vec{x}-\vec{y}\right|$, *and* $\rho_1, \tilde{\rho}_2, \rho_3, \tilde{\rho}_4$ *are as in Theorem 39.*

Proof of Theorem 40. Let the quadrature rules for the pairs $K\times\tilde{K}$ and $K\times e$ be denoted by $a_Q^{K\times\tilde{K}}\left(\cdot,\cdot\right)$ and $a_Q^{K\times e}\left(\cdot,\cdot\right)$. Set

$$E_{K\times\tilde{K}}^{i,j} = a^{K\times\tilde{K}}\left(\phi_i,\phi_j\right) - a_Q^{K\times\tilde{K}}\left(\phi_i,\phi_j\right),$$

$$E^{i,j}_{K\times e} = a^{K\times e}(\phi_i,\phi_j) - a_Q^{K\times e}(\phi_i,\phi_j).$$

For $u, v \in V_h$, we set

$$E_{K\times\tilde{K}}(u,v) = \sum_{i\in\mathcal{I}_{K\times\tilde{K}}}\sum_{j\in\mathcal{I}_{K\times\tilde{K}}} u_i v_j E^{i,j}_{K\times\tilde{K}},$$
$$E_{K\times e}(u,v) = \sum_{i\in\mathcal{I}_K}\sum_{j\in\mathcal{I}_K} u_i v_j E^{i,j}_{K\times e}$$

so that

$$\begin{aligned}
\left|E_{K\times\tilde{K}}(u,v)\right| &\le \left(\max_{i,j}\left|E^{i,j}_{K\times\tilde{K}}\right|\right)\sum_{i\in\mathcal{I}_{K\times\tilde{K}}}|u_i|\sum_{j\in\mathcal{I}_{K\times\tilde{K}}}|v_j| \\
&\le \left(\max_{i,j}\left|E^{i,j}_{K\times\tilde{K}}\right|\right)\left|\mathcal{I}_{K\times\tilde{K}}\right|\sqrt{\sum_{i\in\mathcal{I}_{K\times\tilde{K}}}|u_i|^2}\sqrt{\sum_{j\in\mathcal{I}_{K\times\tilde{K}}}|v_j|^2}, \\
|E_{K\times e}(u,v)| &\le \left(\max_{i,j}\left|E^{i,j}_{K,e}\right|\right)\sum_{i\in\mathcal{I}_K}|u_i|\sum_{j\in\mathcal{I}_K}|v_j| \\
&\le \left(\max_{i,j}\left|E^{i,j}_{K,e}\right|\right)|\mathcal{I}_K|\sqrt{\sum_{i\in\mathcal{I}_K}|u_i|^2}\sqrt{\sum_{j\in\mathcal{I}_K}|v_j|^2}
\end{aligned}$$

Since

$$\sum_{i\in\mathcal{I}_{K\times\tilde{K}}}|u_i|^2 \le C\left[h_K^{-d}\int_K u^2 + h_{\tilde{K}}^{-d}\int_{\tilde{K}} u^2\right],$$
$$\sum_{i\in\mathcal{I}_K}|u_i|^2 \le Ch_K^{-d}\int_K u^2,$$

we find

$$\begin{aligned}
|a(u,v)-a_Q(u,v)| &\le \sum_K\sum_{\tilde{K}}\left|E_{K\times\tilde{K}}(u,v)\right| + \sum_K\sum_e |E_{K\times e}(u,v)| \\
&\le C\sum_K\sum_{\tilde{K}}\left(\max_{i,j}\left|E^{i,j}_{K\times\tilde{K}}\right|\right)\max\left\{h_K^{-d},h_{\tilde{K}}^{-d}\right\}\left[\|u\|^2_{L^2(K)}+\|u\|^2_{L^2(\tilde{K})}\right]^{1/2} \\
&\qquad \left[\|v\|^2_{L^2(K)}+\|v\|^2_{L^2(\tilde{K})}\right]^{1/2} \\
&\quad + C\sum_K\sum_e\left(\max_{i,j}\left|E^{i,j}_{K\times e}\right|\right)h_K^{-d}\|u\|_{L^2(K)}\|v\|_{L^2(K)}
\end{aligned}$$

$$\leq C \left(\max_{K,\tilde{K}} \max_{i,j} \left| E^{i,j}_{K\times\tilde{K}} \right| \max\left\{ h_K^{-d}, h_{\tilde{K}}^{-d} \right\} \right) \sum_K \sum_{\tilde{K}} \|u\|_{L^2(K\cup\tilde{K})} \|v\|_{L^2(K\cup\tilde{K})}$$
$$+ C \left(\max_{K,e} \max_{i,j} \left| E^{i,j}_{K\times e} \right| h_K^{-d} \right) \sum_K \sum_e \|u\|_{L^2(K)} \|v\|_{L^2(K)} .$$

Because

$$\sum_K \sum_{\tilde{K}} \|u\|_{L^2(K\cup\tilde{K})} \|v\|_{L^2(K\cup\tilde{K})} \leq \sqrt{\sum_K \sum_{\tilde{K}} \|u\|^2_{L^2(K\cup\tilde{K})}} \sqrt{\sum_K \sum_{\tilde{K}} \|v\|^2_{L^2(K\cup\tilde{K})}}$$
$$\leq 2 \left|\mathcal{P}_h\right| \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}$$
$$\leq Cn \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}$$

and

$$\sum_K \sum_e \|u\|_{L^2(K)} \|v\|_{L^2(K)} \leq \left|\partial\mathcal{P}_h\right| \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}$$
$$\leq Cn^{(d-1)/d} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} ,$$

we obtain

$$|a(u,v) - a_Q(u,v)| \leq C \left[n \left(\max_{K,\tilde{K}} \max_{i,j} \left| E^{i,j}_{K\times\tilde{K}} \right| \max\left\{ h_K^{-d}, h_{\tilde{K}}^{-d} \right\} \right) \right.$$
$$\left. + n^{(d-1)/d} \left(\max_{K,e} \max_{i,j} \left| E^{i,j}_{K\times e} \right| h_K^{-d} \right) \right] \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} .$$

For $d = 2$, using Theorem 39 stated below permits to conclude. □

The impact of the use of quadrature rules on the accuracy of the resulting finite element approximation can be quantified using Strang's first lemma [50, Lemma 2.27]:

$$\|u - u_h\|_{\widetilde{H}^s(\Omega)} \leq C \inf_{v_h\in V_h} \left[\|u - v_h\|_{\widetilde{H}^s(\Omega)} + \sup_{w_h\in V_h} \frac{|a(v_h,w_h) - a_Q(v_h,w_h)|}{\|w_h\|_{\widetilde{H}^s(\Omega)}} \right]$$
$$\leq C \inf_{v_h\in V_h} \left[\|u - v_h\|_{\widetilde{H}^s(\Omega)} \right.$$
$$\left. + \left(E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial}\right) \|v_h\|_{L^2(\Omega)} \sup_{w_h\in V_h} \frac{\|w_h\|_{L^2(\Omega)}}{\|w_h\|_{\widetilde{H}^s(\Omega)}} \right]$$

$$\leq C \inf_{v_h \in V_h} \left[\|u - v_h\|_{\widetilde{H}^s(\Omega)} + \left(E_T + E_{NT} + E_{T,\partial} + E_{NT,\partial}\right) \|v_h\|_{L^2(\Omega)} \right],$$

where we used the Poincare inequality $\|w_h\|_{L^2(\Omega)} \leq C \|w_h\|_{\widetilde{H}^s(\Omega)}$ in the last step. Based on the error estimates of Theorem 40, the optimal quadrature orders in order to allow convergence of up to order $n^{-\alpha}$ should be chosen as

$$\begin{aligned}
&k_T\left(K, \tilde{K}\right) \\
\geq &\frac{(\alpha/2 + 1/2)\log n + s\left|\log h_K\right|}{\log(\rho_1)} - C, \\
&k_{NT}\left(K, \tilde{K}\right) \\
\geq &\max\left\{ \frac{(\alpha/2 + 1/2)\log n + (s-1)\left|\log h_{\tilde{K}}\right| + \max\left\{\left|\log h_K\right|, \left|\log h_{\tilde{K}}\right|\right\} - s\log\frac{d_{K,\tilde{K}}}{h_{\tilde{K}}} - C}{\log_+\frac{d_{K,\tilde{K}}}{h_K} + \log(\rho_2)}, \right. \\
&\qquad \left. \frac{(\alpha/2 + 1/2)\log n + (s-1)\left|\log h_K\right| + \max\left\{\left|\log h_K\right|, \left|\log h_{\tilde{K}}\right|\right\} - s\log\frac{d_{K,\tilde{K}}}{h_K} - C}{\log_+\frac{d_{K,\tilde{K}}}{h_{\tilde{K}}} + \log(\rho_2)} \right\} \\
&k_{T,\partial}\left(K, e\right) \\
\geq &\frac{(\alpha/2 + 1/4)\log n + s\left|\log h_K\right|}{\log(\rho_3)} - C, \\
&k_{NT,\partial}\left(K, e\right) \\
\geq &\max\left\{ \frac{(\alpha/2 + 1/4)\log n + (s-1)\left|\log h_e\right| + \max\left\{\left|\log h_K\right|, \left|\log h_e\right|\right\} - s\log\frac{d_{K,e}}{h_e} - C}{\log_+\frac{d_{K,e}}{h_K} + \log(\rho_4)}, \right. \\
&\qquad \left. \frac{(\alpha/2 + 1/4)\log n + (s-1)\left|\log h_K\right| + \max\left\{\left|\log h_K\right|, \left|\log h_e\right|\right\} - s\log\frac{d_{K,e}}{h_K} - C}{\log_+\frac{d_{K,e}}{h_e} + \log(\rho_4)} \right\}.
\end{aligned}$$

It transpires from the expressions derived above and the fact that $n \sim h^{-2}$ for a quasi-uniform mesh that the complexity to calculate the contributions by a single pair of elements K and $\tilde{K}$ scales like

- $\log n$ if the elements coincide,
- $(\log n)^2$ if the elements share only an edge,
- $(\log n)^3$ if the elements share only a vertex,

- $(\log n)^4$ if the elements have empty intersection, but are "near neighbors", and
- C if the elements are well separated.

Since $n \sim |\mathcal{P}_h|$, we cannot expect a straightforward assembly of the stiffness matrix to scale better than $\mathcal{O}(n^2)$. Similarly, its memory requirement is n^2, and a single matrix-vector product has complexity $\mathcal{O}(n^2)$, which severely limits the size of problems that can be considered.

11.2 Solving the Linear Systems

The fractional Poisson equation (10.1) leads to a dense linear algebraic system

$$\boldsymbol{A}^s \vec{u} = \vec{b}. \tag{11.5}$$

The solution of the dense matrix equation (11.5) also poses its own problems. For instance, using matrix factorization to solve the resulting system has complexity $\mathcal{O}(n^3)$, where n is the number of degrees of freedom.

Alternatively, if conjugate gradient iteration is used, the number of iterations necessary to converge to a given error tolerance is known to scale as the square root of the condition number of the matrix. The condition numbers of the fractional Laplacian, as well as the its diagonally preconditioned version grow with the number of unknowns n, as shown by Theorem 41.

Theorem 41 ([10])**.**
For $s < d/2$, and a family of shape regular triangulations $\mathcal{P}_h$ with minimal and maximal element size $h_{\min}$ and h, the spectrum of the stiffness matrix satisfies

$$cn^{-2s/d} h_{\min}^{d-2s} \boldsymbol{I} \leq \boldsymbol{A}^s \leq Ch^{d-2s} \boldsymbol{I},$$
$$cn^{-2s/d} \boldsymbol{I} \leq (\boldsymbol{D}^s)^{-1/2} \boldsymbol{A}^s (\boldsymbol{D}^s)^{-1/2} \leq C\boldsymbol{I},$$

where $\boldsymbol{D}^s$ is the diagonal part of $\boldsymbol{A}^s$. The condition number of the stiffness matrix satisfies

$$\kappa\left(\boldsymbol{A}^s\right) = C\left(\frac{h}{h_{\min}}\right)^{d-2s} n^{2s/d}, \tag{11.6}$$

$$\kappa\left(\left(\boldsymbol{D}^s\right)^{-1}\boldsymbol{A}^s\right) = Cn^{2s/d}. \tag{11.7}$$

The exponent of the growth of the condition number depends on the fractional order s. For small s, the matrix is better conditioned. For large s, the growth of the condition number approaches that of the usual integer order Laplacian. Moreover, (11.6) shows that the ill-conditioning becomes increasingly severe on non-uniform meshes. Fortunately, estimate (11.7) shows that using diagonal scaling removes the effects of the non-uniformity of the triangulation. The conjugate gradient method applied to eq. (11.5) is therefore expected to converge in $\sqrt{\kappa\left(\left(\boldsymbol{D}^s\right)^{-1}\boldsymbol{A}^s\right)} \sim n^{s/d}$ iterations. Alternatively, a geometric multigrid solver can be used. In the integer order case, multigrid iterations are used to good effect for solving systems involving both the mass matrix and the stiffness matrix which arises from the discretisation of the regular Laplacian. It is known that multigrid converges with a rate that is independent of the problem size for pseudo-differential operators of positive order in general, which in particular includes the fractional Laplacian [62, 63, 82]. Although a single multigrid iteration is more expensive than a single iteration of conjugate gradient, a multigrid solver is more efficient, especially for larger fractional order s that are troublesome for the conjugate gradient method.

As mentioned earlier, the temporal problem (10.2) is often of more interest than the steady state problem (10.1). We first observe that using an explicit schemes for the time discretization will lead to a CFL conditions on the time-step size Δt of type $\Delta t \leq Ch^{2s}$. The discretization using an implicit integration scheme in the time variable lead to systems of the form

$$\left(\boldsymbol{M} + \Delta t\boldsymbol{A}^s\right)\vec{u} = \vec{b}, \tag{11.8}$$

where $\boldsymbol{M}$ is the mass matrix. In typical examples, the time-step will be chosen so that the error from the approximation in time matches the spatial order, and hence Δt depends on the spatial discretisation. The following Theorem shows that if the time-step size is $\mathcal{O}\left(h^{2s}\right)$, we can expect the conjugate gradient method to converge rapidly, since the condition number of the system is bounded by a constant.

Lemma 42.

For a shape regular and globally quasi-uniform family of triangulations $\mathcal{P}_h$ and time-step $\Delta t \leq 1$,

$$\kappa\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right) \leq C\left(1 + \frac{\Delta t}{h^{2s}}\right).$$

More generally, for a family of triangulations that is only locally quasi-uniform and $\Delta t \leq h_{\min}^{2s} n^{2s/d}$, it holds that

$$\kappa\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right) \leq C\left(\frac{h}{h_{\min}}\right)^d \left(1 + \frac{\Delta t}{h^{2s}}\right). \tag{11.9}$$

If $\boldsymbol{D}^0$ is taken to be the diagonal part of the mass matrix and $\Delta t \leq h^{2s} n^{2s/d}$, then

$$\kappa\left(\left(\boldsymbol{D}^0\right)^{-1}\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right)\right) \leq C\left(1 + \frac{\Delta t}{h_{\min}^{2s}}\right) \tag{11.10}$$

Proof. Since $c h_{\min}^d \boldsymbol{I} \leq \boldsymbol{M} \leq C h^d \boldsymbol{I}$, this also permits us to deduce that

$$c\left(h_{\min}^d + \Delta t\ n^{-2s/d} h_{\min}^{d-2s}\right)\boldsymbol{I} \leq \boldsymbol{M} + \Delta t \boldsymbol{A}^s \leq C\left(h^d + \Delta t\ h^{d-2s}\right)\boldsymbol{I}$$

and so

$$\begin{aligned}
\kappa\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right) &\leq C \frac{h^d + \Delta t\ h^{d-2s}}{h_{\min}^d + \Delta t\ n^{-2s/d} h_{\min}^{d-2s}} \\
&= C\left(\frac{h}{h_{\min}}\right)^d \frac{1 + \Delta t\ h^{-2s}}{1 + \Delta t\ n^{-2s/d} h_{\min}^{-2s}} \\
&\leq C\left(\frac{h}{h_{\min}}\right)^d \left(1 + \frac{\Delta t}{h^{2s}}\right).
\end{aligned}$$

For globally quasi-uniform meshes, $\frac{h}{h_{\min}} \sim C$ and $h_{\min}^{2s} n^{2s/d} \sim 1$.

Since $\|\phi_i\|_{L^2}^2 \sim h_i^d$ and $\|\phi_i\|_{\widetilde{H}^s}^2 \sim h_i^{d-2s}$ [10] and $\left(\boldsymbol{D}^0\right)^{-1/2} \boldsymbol{M} \left(\boldsymbol{D}^0\right)^{-1/2} \sim \boldsymbol{I}$, we find

$$c\left(1+\Delta t\, n^{-2s/d} h^{-2s}\right) \boldsymbol{I} \leq \left(\boldsymbol{D}^0\right)^{-1/2} \left(\boldsymbol{M}+\Delta t \boldsymbol{A}^s\right) \left(\boldsymbol{D}^0\right)^{-1/2} \leq C\left(1+\Delta t\, h_{\min}^{-2s}\right) \boldsymbol{I},$$

and hence

$$\kappa\left(\left(\boldsymbol{D}^0\right)^{-1}\left(\boldsymbol{M}+\Delta t \boldsymbol{A}^s\right)\right) \leq C \frac{1+\Delta t\, h_{\min}^{-2s}}{1+\Delta t\, n^{-2s/d} h^{-2s}} \leq C\left(1+\frac{\Delta t}{h_{\min}^{2s}}\right).$$

□

Lemma 42 shows that if the time step Δt is chosen to be $\Delta t \sim h_{\min}^{2s}$, then using the conjugate gradient method to solve the diagonally preconditioned system (11.8) gives a uniform bound on the number of iterations. However, if larger time steps are chosen, e.g. independent of h_{min}, then (11.10) shows that the simple diagonal scaling becomes inefficient resulting in a number of iterations growing as $\mathcal{O}\left(\frac{\Delta t}{h_{\min}^{2s}}\right)$. Nevertheless, by applying a multigrid solver (as in the steady state case), one can restore a uniform bound on the number of iterations [62, 63, 82].

In this section we have concerned ourselves with the effect the mesh and the fractional order has on the rate of convergence of iterative solvers. This, of course, ignores the cost of carrying out the iteration. The complexity of both multigrid and conjugate gradient iterations depends on how efficiently the matrix-vector product $\boldsymbol{A}^s \vec{x}$ can be computed. (The mass matrix in eq. (11.8) has $\mathcal{O}(n)$ entries, so its matrix-vector product scales linearly in the number of unknowns.) Since all the basis functions ϕ_i interact with each other, the matrix $\boldsymbol{A}^s$ is dense and its matrix-vector product has complexity n^2. In the following section, we discuss a sparse approximation that will preserve the order of the approximation error of the fractional Laplacian, but display significantly better scaling in memory and operations for both assembly and matrix-vector product.

11.3 Sparse Approximation of the Stiffness Matrix and of the Strong Form

The presence of a factor $|\vec{x} - \vec{y}|^{-d-2s}$ in the integrand in the expression for the entries of the stiffness matrix means that the contribution of pairs of elements that are well separated is significantly smaller than the contribution arising from pairs of elements that are close to one another. The evaluation of the strong form of the fractional Laplacian as given in eqs. (10.15) and (10.16) involves integration over all edges of the mesh weighted by the same kernel. This suggests the use of the *cluster method* from the boundary element literature, whereby such far field contributions are replaced by less expensive low-rank blocks rather than computing and storing the all individual entries from the original matrix. Conversely, the near-field contributions are more significant but involve only local couplings and hence the cost of storing the individual entries is a practical proposition. A full discussion of the clustering method is beyond the scope of the present work but can be found in [82, Chapter 7]. Here, we confine ourselves to stating only the necessary definitions and steps needed to describe our approach.

Definition 43 ([82])**.**
A cluster *is a union of one or more indices from the set of degrees of freedom $\mathcal{I}$. The nodes of a hierarchical cluster tree $\mathcal{T}$ are clusters. The set of all nodes is denoted by T and satisfies*

1. *$\mathcal{I}$ is a node of $\mathcal{T}$.*

2. *The set of leaves Leaves$(\mathcal{T}) \subset T$ corresponds to the degrees of freedom $i \in \mathcal{I}$ and is given by*

$$\textit{Leaves}(\mathcal{T}) := \{\{i\} : i \in \mathcal{I}\}.$$

3. *For every $\sigma \in T \setminus$ Leaves $(\mathcal{T})$ there exists a minimal set $\Sigma(\sigma)$ of nodes in $T \setminus \{\sigma\}$*

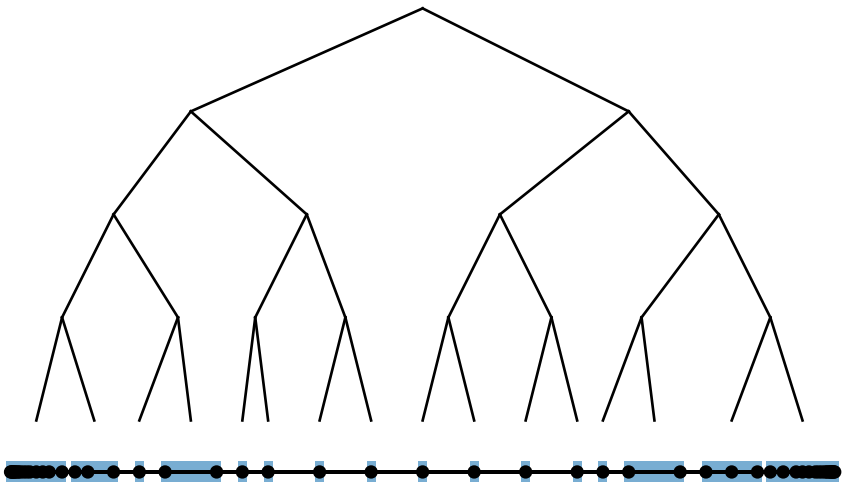


Figure 11.3: Cluster tree for a one-dimensional problem. The mesh with its nodal degrees of freedom is plotted at the bottom. The leaf clusters are colored in blue.

that satisfies

$$\sigma = \bigcup_{\tau \in \Sigma(\sigma)} \tau.$$

The set $\Sigma(\sigma)$ is called the sons of σ. The edges of the cluster tree $\mathcal{T}$ are the pairs of nodes $(\sigma, \tau) \in T \times T$ such that $\tau \in \Sigma(\sigma)$.

An example of a cluster tree for a one-dimensional problem is given in Figure 11.3.

Definition 44 ([82])**.**
The cluster box Q_σ *of a cluster $\sigma \in T$ is the minimal hyper-cube which contains $\bigcup_{i \in \sigma} \operatorname{supp} \phi_i$. The* diameter *of a cluster is the diameter of its cluster box* $\operatorname{diam}(\sigma) := \sup_{\vec{x}, \vec{y} \in Q_\sigma} |\vec{x} - \vec{y}|$. *The* distance *of two clusters σ and τ is* $\operatorname{dist}(\sigma, \tau) := \inf_{\vec{x} \in Q_\sigma, \vec{y} \in Q_\tau} |\vec{x} - \vec{y}|$. *The subspace V_σ of V_h is defined as* $V_\sigma := \operatorname{span}\{\phi_i \mid i \in \sigma\}$.

For given $\eta > 0$, a pair of clusters (σ, τ) is called *admissible*, if

$$\eta \operatorname{dist}(\sigma, \tau) \geq \max\{\operatorname{diam}(\sigma), \operatorname{diam}(\tau)\}.$$

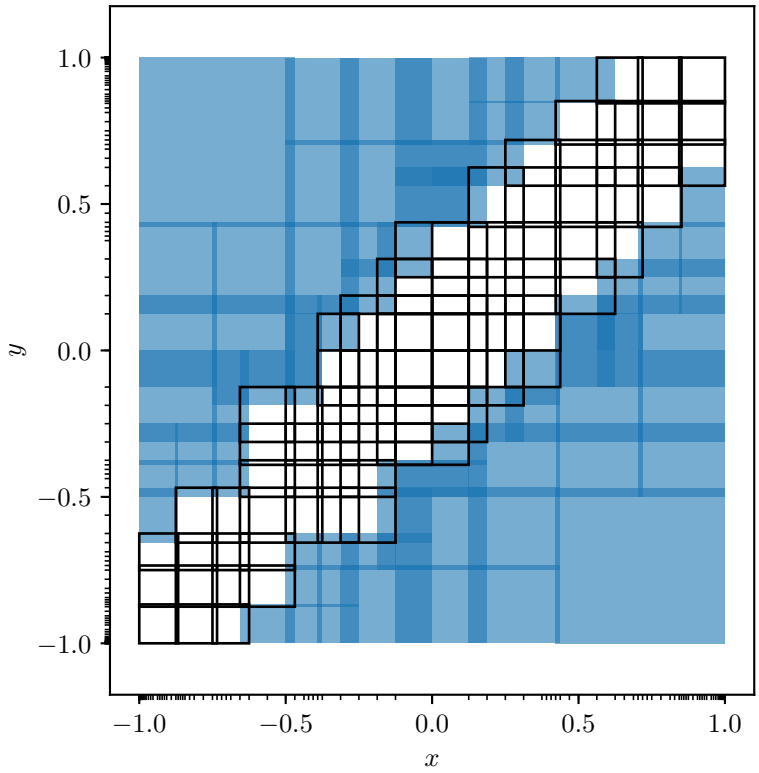


Figure 11.4: Cluster pairs for a one-dimensional problem. The cluster boxes of the admissible cluster pairs are colored in light blue, and their overlap in darker blue. The diagonal cluster pairs are not admissible and are not approximated, but assembled in full. The nodes of the nodal degrees of freedom are plotted on x- and y-axis.

The admissible cluster pairs can be determined recursively. Cluster pairs that are not admissible and have no admissible sons are part of the near field and are assembled into a sparse matrix. The admissible cluster pairs for a one dimensional problem are shown in Figure 11.4.

For admissible pairs of clusters σ and τ and any functions ϕ, ψ with $\operatorname{supp}\phi \subset Q_\sigma$ and $\operatorname{supp}\psi \subset Q_\tau$, the bilinear form a evaluates to

$$a(\phi,\psi) = -C(d,s)\int_\Omega\int_\Omega k(\vec{x},\vec{y})\,\phi(\vec{x})\,\psi(\vec{y})\,,$$

since $Q_\sigma \cap Q_\tau = \emptyset$. Here, $k(\vec{x},\vec{y}) = |\vec{x}-\vec{y}|^{-(d+2s)}$ is the interaction kernel, which can be approximated on $Q_\sigma \times Q_\tau$ using Chebyshev interpolation of order m in every

spatial dimension by

$$k_m(\vec{x},\vec{y}) = \sum_{\alpha,\beta=1}^{m^d} k\left(\vec{\xi}_\alpha^\sigma, \vec{\xi}_\beta^\tau\right) L_\alpha^\sigma(\vec{x})\, L_\beta^\tau(\vec{y}).$$

L_α^σ are the tensor Chebyshev polynomials of order m on the cluster box Q_σ, and $\vec{\xi}_\alpha^\sigma$ are the tensor Chebyshev points on Q_σ with $L_\alpha^\sigma\left(\vec{\xi}_\beta^\sigma\right) = \delta_{\alpha\beta}$. This leads to the following approximation:

$$a(\phi,\psi) \approx -C(d,s) \sum_{\alpha,\beta=1}^{m^2} k\left(\vec{\xi}_\alpha^\sigma, \vec{\xi}_\beta^\tau\right) \int_{\operatorname{supp}\phi} \phi(\vec{x})\, L_\alpha^\sigma(\vec{x})\; d\vec{x} \int_{\operatorname{supp}\psi} \psi(\vec{y})\, L_\beta^\tau(\vec{y})\; d\vec{y}.$$

Expressions of the type $\int_{\operatorname{supp}\phi} \phi(\vec{x})\, L_\alpha^\sigma(\vec{x})\; d\vec{x}$ can be computed recursively starting from the finest level of the cluster tree, since for $\tau \in \Sigma(\sigma)$ and $\vec{x} \in Q_\tau$

$$L_\alpha^\sigma(\vec{x}) = \sum_\beta L_\alpha^\sigma\left(\vec{\xi}_\beta^\tau\right) L_\beta^\tau(\vec{x}).$$

This means that for all leaves $\sigma = \{i\}$, and all $1 \le \alpha \le m^d$, the *basis far-field coefficients*

$$\int_{\operatorname{supp}\phi \cap Q_\sigma} \phi(\vec{x})\, L_\alpha^\sigma(\vec{x})\; d\vec{x}$$

need to be evaluated (e.g. by $m+1$-th order Gaussian quadrature). Moreover, the *shift coefficients*

$$L_\alpha^\sigma\left(\vec{\xi}_\beta^\tau\right)$$

for $\tau \in \Sigma(\sigma)$ must be computed, as well as the kernel approximations

$$k\left(\vec{\xi}_\alpha^\sigma, \vec{\xi}_\beta^\tau\right)$$

for every admissible pair of clusters (σ,τ). We refer the reader to [82] for further details.

In order for the consistency error of the cluster method to be dominated by the discretization error of the method, the interpolation order m essentially needs to grow with $|\log h_{\min}|$. For further details, we refer to [9, 82].

By following the arguments in [82], it can be shown that the number of near field entries scales linearly in n. The same conclusion holds for the number of far field cluster pairs. Since the four dimensional integral contributions $a^{K\times\tilde{K}}$ are evaluated using Gaussian quadrature rules with at most $k \sim \log n$ quadrature nodes per dimension, the assembly of the near field contributions scales with $n \log^{2d} n$. The far field kernel approximations and the shift coefficients have size $m^{2d} \sim \log^{2d} n$, and are also calculated in $\log^{2d} n$ complexity. This means that all the kernel approximations and shift coefficients are obtained in $n \log^{2d} n$ time. Finally, the nm^d basis far-field coefficients require the evaluation of integrals using $m+1$-th order Gaussian quadrature, leading to a complexity of $n \log^{2d}$ as well. The overall complexity of the cluster method is therefore $\mathcal{O}\left(n \log^{2d} n\right)$, and the sparse approximation requires $\mathcal{O}\left(n \log^{2d} n\right)$ memory. In practice, this means that the assembly of the near-field matrix dominates the other steps but involves only local computations.

The computation of the matrix-vector product involving upward and downward recursion in the cluster tree and multiplication by the kernel approximations can also be shown to scale with $\mathcal{O}\left(n \log^{2d} n\right)$.

The same principle can be used to evaluate the strong form in all the quadrature nodes necessary for the computation of the BR error indicators. According to eqs. (10.15) and (10.16), a naive implementation scales as $\mathcal{O}\left(n^2\right)$. Many of the components that are used for the cluster method in the matrix-vector product can be reused. Let $u = \sum_{i\in\mathcal{I}_h} u_i\phi_i \in V_h$, and let $\vec{x}_0 \in K_0$ be a quadrature node in the evaluation of the local L^2-norms in (10.13). Moreover, let $i_0 \in \mathcal{I}_h$ be the closest degree of freedom to $\vec{x}_0$ on K_0.

Let (σ, τ) be an admissible cluster pair such that $i_0 \in \sigma$, and set $u_\tau = \sum_{j\in\tau} u_j\phi_j$, so that $\operatorname{supp} u_\tau \subset Q_\tau$. The contribution to the fractional Laplacian is approximated by

$$-\int_{Q_\tau} k\left(\vec{x},\vec{y}\right) u_\tau\left(\vec{y}\right)\ d\vec{y} = -\int_{Q_\sigma}\int_{Q_\tau} k\left(\vec{x},\vec{y}\right)\delta\left(\vec{x}-\vec{x}_0\right) u_\tau\left(\vec{y}\right)\ d\vec{y}\, d\vec{x}$$

$$\approx \sum_{\alpha,\beta=1}^{m^d} k\left(\vec{\xi}_\alpha^\sigma, \vec{\xi}_\beta^\tau\right) \int_{Q_\sigma} L_\alpha^\sigma(\vec{x})\, \delta(\vec{x}-\vec{x}_0)\ d\vec{x} \int_{Q_\tau} L_\beta^\tau(\vec{y})\, u_\tau(\vec{y})\ d\vec{y}$$

$$= \sum_{\alpha,\beta=1}^{m^d} k\left(\vec{\xi}_\alpha^\sigma, \vec{\xi}_\beta^\tau\right) L_\alpha^\sigma(\vec{x}_0) \int_{Q_\tau} L_\beta^\tau(\vec{y})\, u_\tau(\vec{y})\ d\vec{y}.$$

The integrals $\int_{Q_\tau} L_\beta^\tau(\vec{y})\, u_\tau(\vec{y})\ d\vec{y}$ already appeared in the computation of the matrix-vector product, and can be recursively computed. The same holds for the factors $L_\alpha^\sigma(\vec{x}_0)$.

On the other hand, if the cluster pair (σ, τ) is not admissible and therefore in the near-field, then formulas eqs. (10.15) and (10.16) with Ω set to Q_σ can be used to evaluate the fractional Laplacian.

This shows that the evaluation of the strong form of the fractional Laplacian on all the quadrature nodes $\{\vec{x}_j\}_j$ can be seen as an application of the cluster method with respect to the sets of functions $\{\delta(\cdot - \vec{x}_j)\}_j$ and $\{\phi_i\}_i$.

We illustrate the above results by assembling both the full matrix as well as its sparse approximation on the unit disk for fractional orders $s = 0.25$ and $s = 0.75$. The memory usage of the matrices are compared in Figure 11.5. For low number of degrees of freedom, none of the cluster pairs are admissible, so the full matrix and its approximation have the same size. Starting with roughly 2000 degrees of freedom, the memory footprint of the sparse approximate starts to follow the $n \log^4 n$ curve and therefore outperforms the full assembly. The same behavior can be observed for the assembly times, as seen in Figure 11.6.

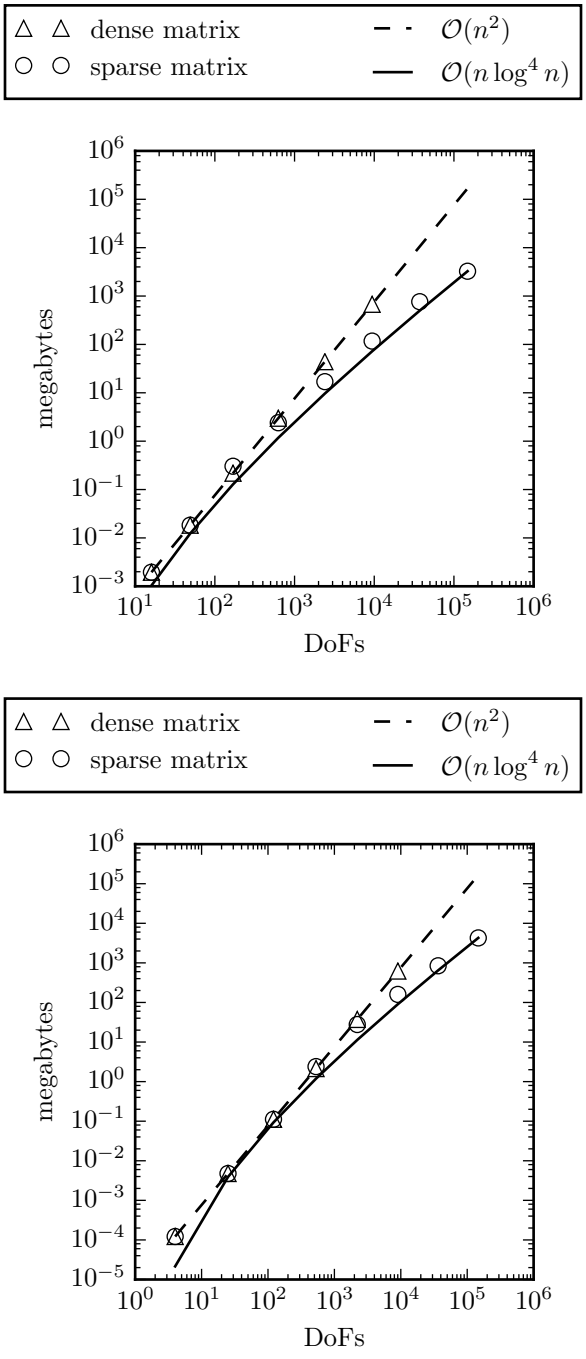


Figure 11.5: Memory usage of the dense matrix and its sparse approximation. $s = 0.25$ (*top*), $s = 0.75$ (*bottom*). While the dense matrix uses n^2 floating-point numbers, the sparse approximation can be seen to require only $\mathcal{O}\left(n \log^4 n\right)$ memory. At roughly 2000 unknowns, the memory footprint of the sparse approximation separates from the $\mathcal{O}\left(n^2\right)$ curve.

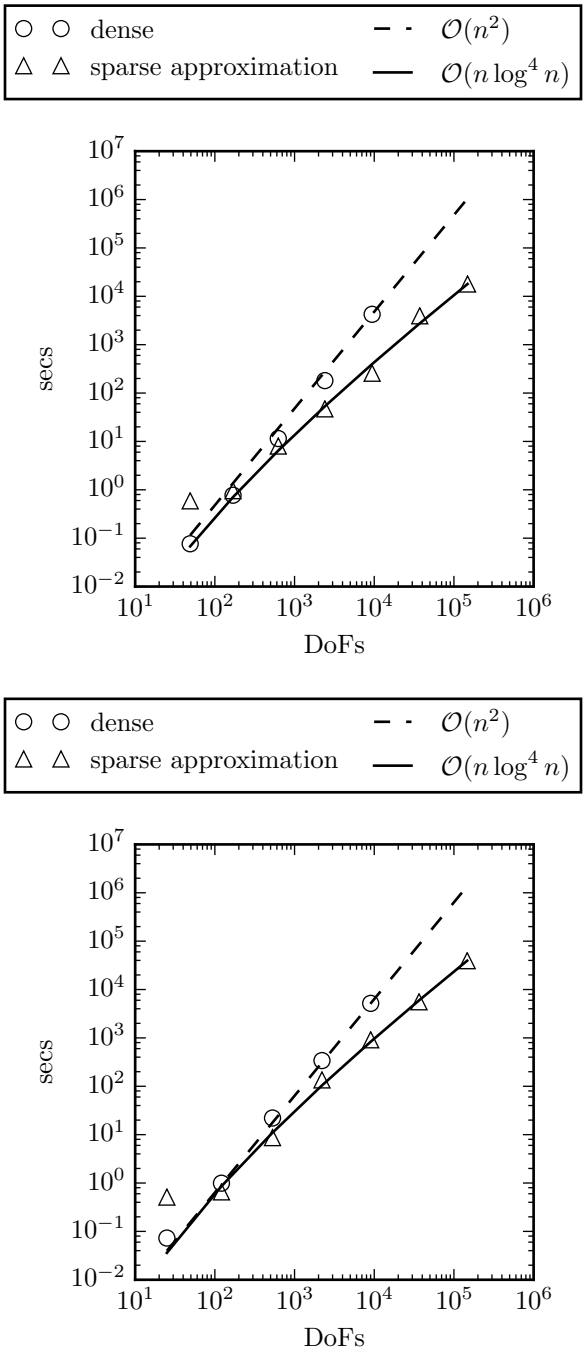


Figure 11.6: Assembly time of the dense matrix and its sparse approximation. $s = 0.25$ (*top*), $s = 0.75$ (*bottom*). The time to assemble the full matrix grows quadratically in the number of unknowns, whereas the sparse approximation starts to follow the $n \log^4 n$ curve at about 2000 degrees of freedom.

Chapter 12

Numerical Examples

12.1 Fractional Poisson Equation

For $s \in (0,1)$, we consider the fractional Poisson problem

$$(-\Delta)^s u = f \text{ in } \Omega = B(0,1),$$
$$u = 0 \text{ in } \Omega^c$$

in $d \in \{1,2\}$ dimensions. The reason for choosing a unit disk domain is that analytic solutions to the problem are known. In $d = 1$ dimensions, the right-hand sides

$$f_{n,0}^{1D} = 2^{2s}\Gamma(1+s)^2 \binom{s+n-1/2}{s}\binom{s+n}{s} P_n^{(s,-1/2)}\left(2x^2-1\right), \qquad n \geq 0$$

have corresponding solutions

$$u_{n,0}^{1D} = P_n^{(s,-1/2)}\left(2x^2-1\right)\left(1-x^2\right)_+^s$$

and the right-hand sides

$$f_{n,1}^{1D} = 2^{2s}\Gamma(1+s)^2 \binom{s+n+1/2}{s}\binom{s+n}{s} x P_n^{(s,1/2)}\left(2x^2-1\right), \qquad n \geq 0$$

have solutions

$$u_{n,1}^{1D} = x P_n^{(s,1/2)}\left(2x^2-1\right)\left(1-x^2\right)_+^s$$

as shown in [45]. Here, $\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}$ are generalized binomial coefficients, $P_n^{(\alpha,\beta)}$ are the Jacobi polynomials, and $x_+ = \max\{0, x\}$. In $d = 2$ dimensions, the right-hand sides

$$f_{n,\ell}^{2D} = 2^{2s}\Gamma(1+s)^2 \binom{s+n+\ell}{s}\binom{s+n}{s} r^\ell \cos(\ell\theta) P_n^{(s,\ell)}\left(2r^2-1\right), \qquad \ell, n \geq 0$$

have known analytical solutions given by

$$u_{n,\ell}^{2D} = r^\ell \cos(\ell\theta) P_n^{(s,\ell)}\left(2r^2-1\right)\left(1-r^2\right)_+^s.$$

For more details on how these analytical solutions are determined, see [45]. We note that both the one-dimensional and the two-dimensional solutions contain a term $(1-x^2)_+^s$ (respectively $(1-r^2)_+^s$) which behaves like $\delta(\vec{x})^s$, where δ is the distance to the boundary.

12.1.1 One-Dimensional Examples

We first solve the fractional Poisson problem with constant right-hand side $f = f_{0,0}^{1D}$ and solution $u = u_{0,0}^{1D}$ for $s = 0.25$ and $s = 0.75$. According to the regularity results in Theorem 34, u is in $H^{s+1/2-\varepsilon}(\Omega)$, so that we expect essentially $h^{1/2} \sim n^{-1/2}$ convergence if a globally quasi-uniform mesh is used. Using graded meshes, the expected rate in $d = 1$ dimensions is n^{s-2}, as shown in eq. (10.11).

We adaptively refine the mesh based on the BR error indicators and a maximum marking strategy with threshold 0.8. The cluster method is used both for the stiffness matrix and for the evaluation of the strong form of the fractional Laplacian. The error plots in Figure 12.1 clearly display n^{s-2} convergence in $\widetilde{H}^s$-norm, both for $s = 0.25$ and for $s = 0.75$, as compared to the $n^{-1/2}$ convergence obtained by uniform mesh refinement. This means that the optimal rate of convergence for $d = 1$ is achieved. The measured L^2-convergence is of order $n^{-1/2-s}$ for uniform refinement and n^{-2} for adaptive refinement for both values of s.

In Figure 12.2, we plot the number of degrees of freedom n against timing results for the assembly of the stiffness matrix, the solution of the linear system,

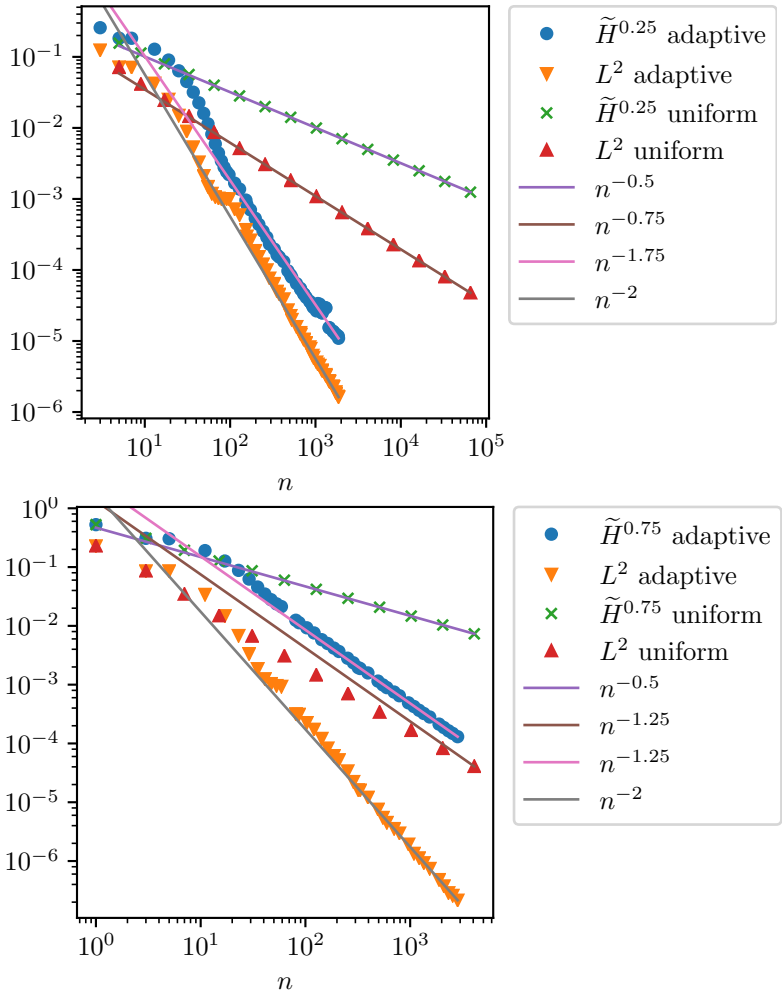


Figure 12.1: $\widetilde{H}^s$- and L^s-error for the one-dimensional problem with constant right-hand side $f = f^{1D}_{0,0}$ and for uniform and adaptive refinement. $s = 0.25$ at the top, $s = 0.75$ at the bottom. It can be seen that while uniform refinement results in convergence in $\widetilde{H}^s$-norm with rate $n^{-1/2}$, adaptive refinement achieves the optimal rate of n^{s-2}. The convergence in L^2-norm is of order $n^{-1/2-s}$ for uniform refinement, and of order n^{-2} for adaptive refinement.

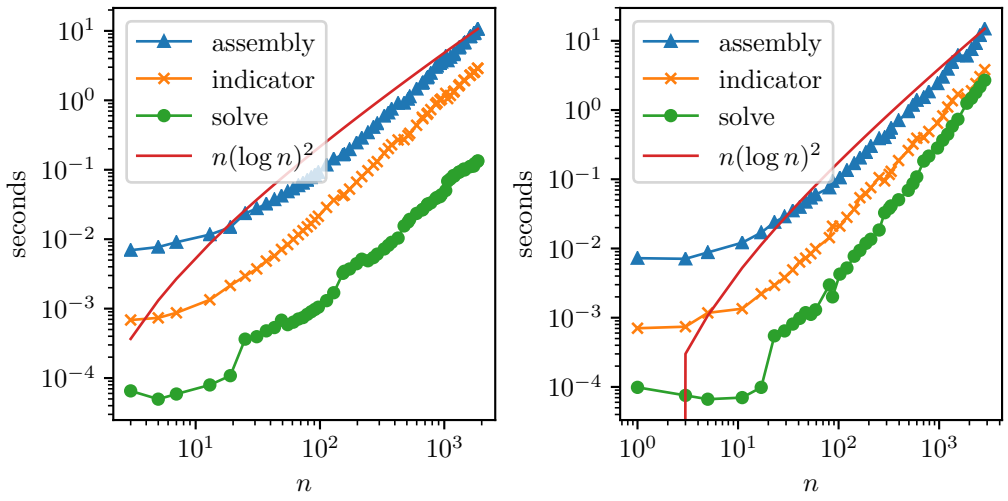


Figure 12.2: Timings for assembly of the stiffness matrix, solution of linear system and computation of the error indicators for the one-dimensional problem with constant right-hand side $f = f^{1D}_{0,0}$. $s = 0.25$ on the left, $s = 0.75$ on the right. All steps can be seen to scale quasi-linearly with the number of unknowns n.

and the computation of the error indicators (10.13). It can be observed that all three scale with $n\,(\log n)^2$.

Next, we consider the solution of the problem with right-hand side $f(x) = \operatorname{sign} x \in H^{1/2-\varepsilon}(\Omega)$. Based on decomposition of f with respect to $f^{1D}_{n,\ell}$, the analytic solution u is found to be

$$u = 2^{-2s} \sum_{n\geq 0} (-1)^n \frac{2n+s+3/2}{n+1/2} \frac{\binom{n+s+1/2}{-1/2}}{\binom{n+s+1/2}{s}\binom{n+s}{s}} x P_n\left(2x^2-1\right)\left(1-x^2\right)_+^s$$

so that the discretization error can be calculated as

$$\|u-u_h\|^2_{\widetilde{H}^s(\Omega)} = a\left(u-u_h, u-u_h\right) = (f,u) - (f,u_h)\,, \tag{12.1}$$

with

$$(f,u) = \frac{2^{1-2s}}{(2s+1)\Gamma\left(1+s\right)^2}$$

Solutions for $s = i/10$, $i = 1,\ldots,9$ are plotted in Figure 12.3. For smaller values of s, the transition at the boundary and at $x = 0$ is more pronounced, whereas for

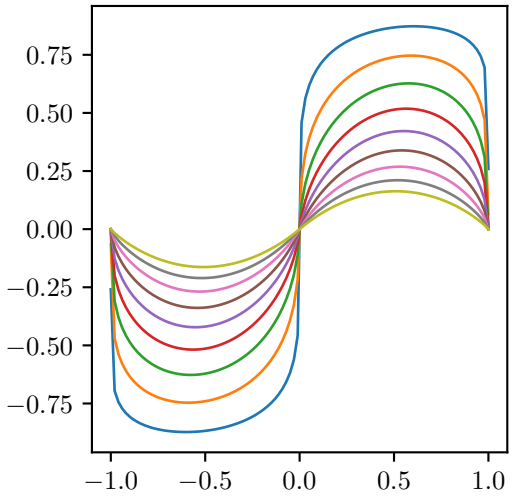


Figure 12.3: Solutions to the one-dimensional fractional Poisson problem with right-hand side $f(x) = \operatorname{sign} x$ for $s = i/10$, $i = 1, \dots, 9$. For smaller values of s, the transition at the boundary and at $x = 0$ is more pronounced, whereas for larger values of s the solutions resemble the solution of the standard integer-order Poisson problem.

larger values of s the solutions resemble the solution of the standard integer-order Poisson problem.

The error plots in Figure 12.4 show that n^{s-2}-convergence in $\widetilde{H}^s$-norm is also achieved for the discontinuous right-hand side $f(x) = \operatorname{sign} x$.

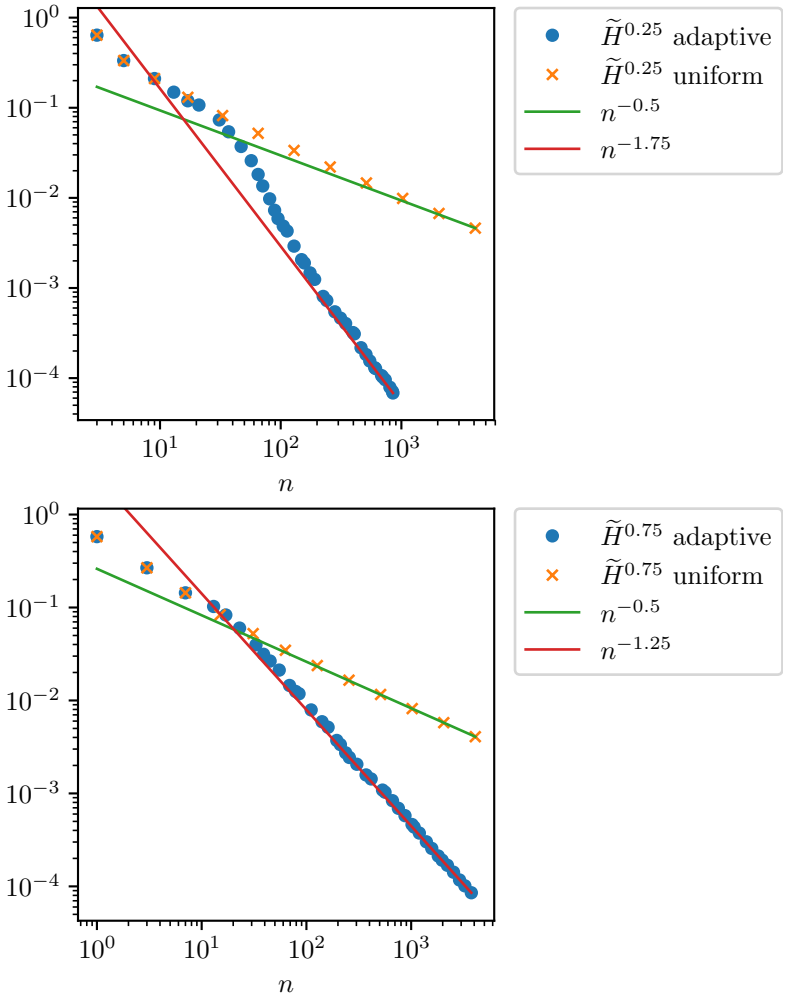


Figure 12.4: $\widetilde{H}^s$-error for the one-dimensional problem with discontinuous right-hand side $f(x) = \operatorname{sign} x$ and for uniform and adaptive refinement. $s = 0.25$ at the top, $s = 0.75$ at the bottom. It can be seen that while uniform refinement results in convergence in $\widetilde{H}^s$-norm with rate $n^{-1/2}$, adaptive refinement achieves the optimal rate of n^{s-2}.

12.1.2 Two-Dimensional Examples

We consider the two-dimensional fractional Poisson problem on a disk with constant right-hand side $f = f_{0,0}^{2D}$. The solutions u for $s = 0.25$ and $s = 0.75$ are shown in Figure 10.1.

Using a globally quasi-uniform mesh, we would obtain $h^{1/2} \sim n^{-1/4}$ convergence. Using a graded mesh, $n^{-1/2}$ convergence was predicted, owing to the fact that the mesh is required to be locally quasi-uniform. Again, we adaptively refine the mesh, based on the error indicators (10.13). The obtained meshes (Figure 12.5) are highly refined in regions close to the boundary, where the solution is singular.

We plot the $\widetilde{H}^s(\Omega)$-error in Figure 12.6. It can be observed that $n^{-1/2}$ convergence is obtained, thus matching the optimal rate obtained using a graded mesh for a constant right-hand side in (10.12). As in the one-dimensional case, the rate in L^2-norm is s orders higher than the rate in $\widetilde{H}^s$-norm. By considering the timing results shown in Figure 12.7, we see that all phases of the adaptive refinement loop scale with $n(\log n)^4$. The computation of the error indicators is cheaper than the assembly of the system matrix, and the solution of the linear system is significantly cheaper than both. Relative to the matrix assembly, the computation of the error indicators is more expensive than for the one dimensional problem, since it involves numerical quadrature over element edges instead of pointwise evaluations.

We now consider the solution of the problem with right-hand side $f(r,\theta) = 1_{|\theta|<\pi/2} \in H^{1/2-\varepsilon}(\Omega)$. Based on decomposition of f with respect to $f_{n,\ell}^{2D}$, the analytic solution u is found to be

$$u = 2^{-2s}\frac{(1-r^2)_+^s}{\Gamma(1+s)^2}\left\{\frac{1}{2} + \sum_{n\geq 0,\ \ell\geq 1 \text{ and odd}} \frac{(-1)^{(\ell+1)/2+n}(2n+s+l+1)}{\pi(n+l/2)(s+1)}\frac{\cos(\ell\theta)\, r^\ell P_n^{(s,\ell)}(2r^2-1)}{\binom{n+s+l/2+1}{n+l/2}\binom{s+n}{n}}\right\},$$

so that the discretization error can be calculated as

$$\|u-u_h\|_{\widetilde{H}^s(\Omega)}^2 = a(u-u_h, u-u_h) = (f,u) - (f,u_h),$$

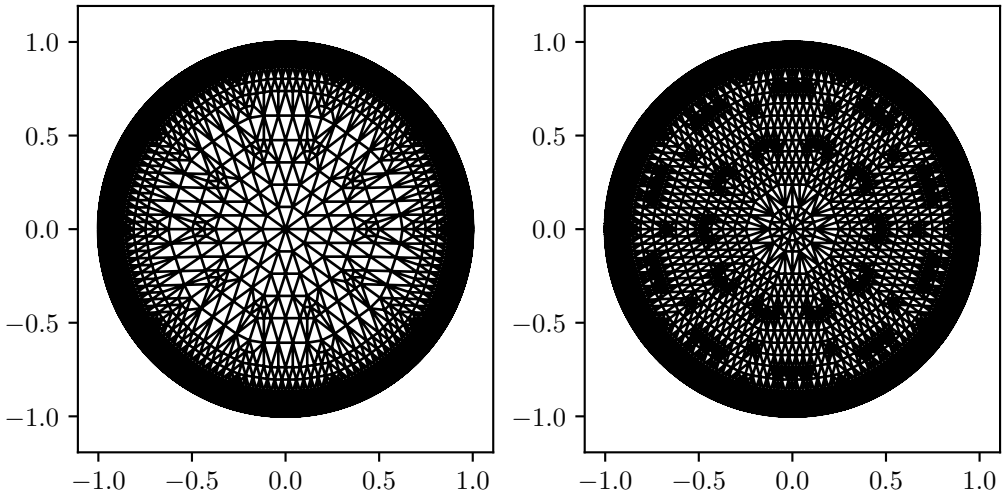


Figure 12.5: Adaptively refines meshes for the two-dimensional problem with constant right-hand side. $s = 0.25$ on the left, $s = 0.75$ on the right.

with

$$(f,u) = 2^{-2s}\left\{\frac{\pi}{4(s+1)\Gamma\left(s+1\right)^2} - \frac{1}{\pi}G^{3,2}_{4,4}\left(\begin{array}{cccc} 1, & 1+s/2, & 5/2+s, & 5/2+s \\ 2, & 1/2, & 1/2, & 2+s/2 \end{array}\middle| -1\right)\right\},$$

where G is the Meijer G-function [99].

The solutions u for $s = 0.25$ and $s = 0.75$ are shown in Figure 12.8. The obtained meshes (Figure 12.9) are highly refined in regions close to the boundary and along the discontinuity $\theta = \pm\pi/2$. We plot the $\widetilde{H}^s$-error in Figure 12.10. In the case $s = 0.25$, we recover the optimal rate of convergence of $n^{-1/2}$, whereas for $s = 0.75$, we only find $n^{-1/4}$, which coincides with the rate that would be obtained by uniform refinement.

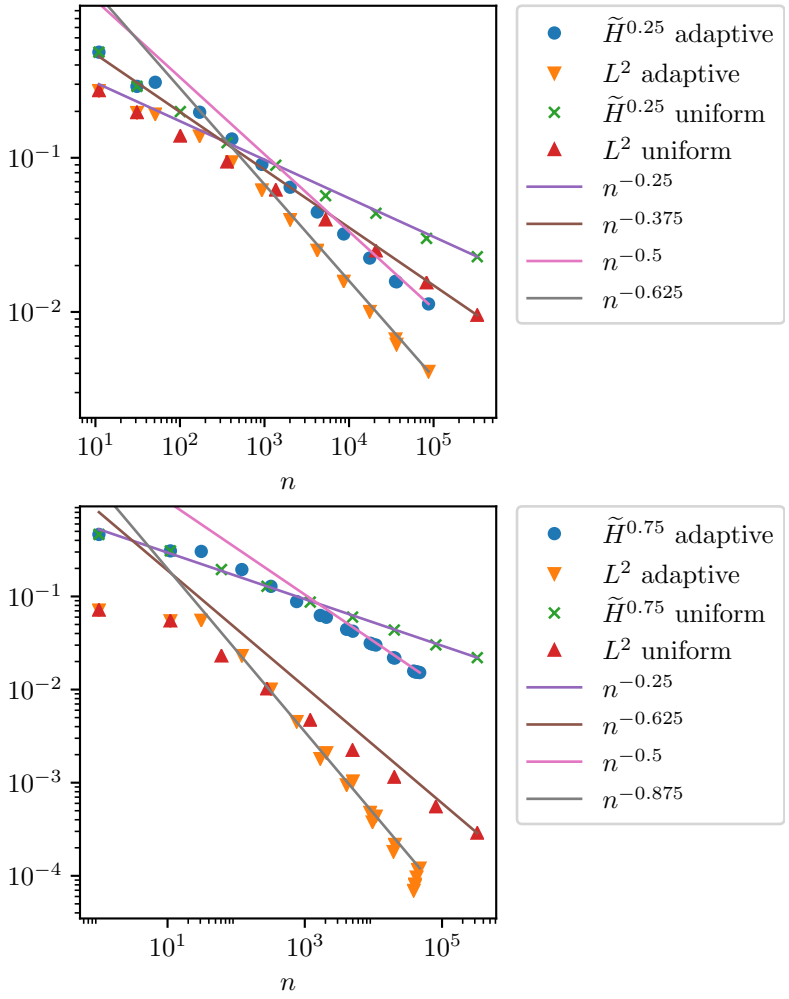


Figure 12.6: $\widetilde{H}^s$- and L^s-error for the two-dimensional problem with constant right-hand side $f = f_{0,0}^{2D}$ and for uniform and adaptive refinement. $s = 0.25$ at the top, $s = 0.75$ at the bottom. It can be seen that while uniform refinement results in convergence in $\widetilde{H}^s$-norm with rate $n^{-1/4}$, adaptive refinement achieves the optimal rate of $n^{-1/2}$. The convergence in L^2-norm is of order $n^{-1/4-s/2}$ for uniform refinement, and of order $n^{-1/2-s/2}$ for adaptive refinement.

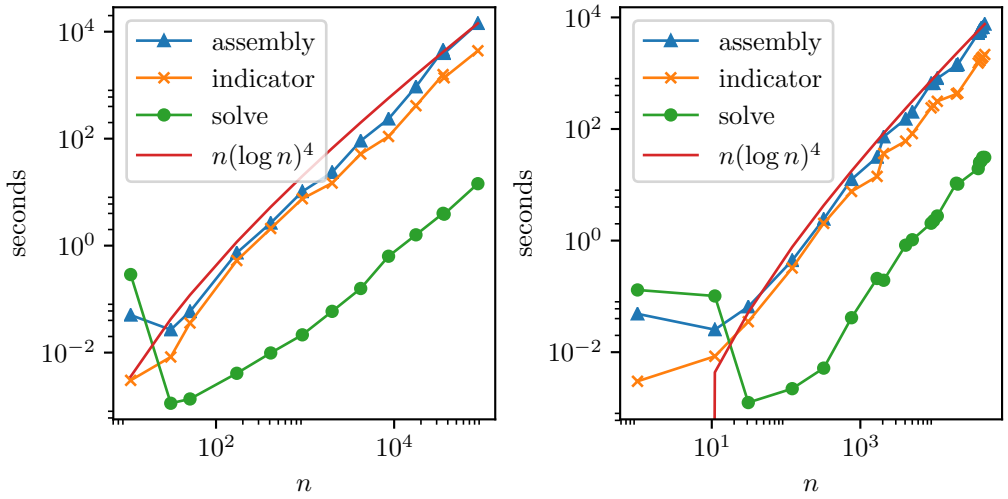


Figure 12.7: Timings for assembly of the stiffness matrix, solution of linear system and computation of the error indicators for the two-dimensional problem with constant right-hand side $f = f^{2D}_{0,0}$. $s = 0.25$ on the left, $s = 0.75$ on the right. All steps can be seen to scale quasi-linearly with the number of unknowns n.

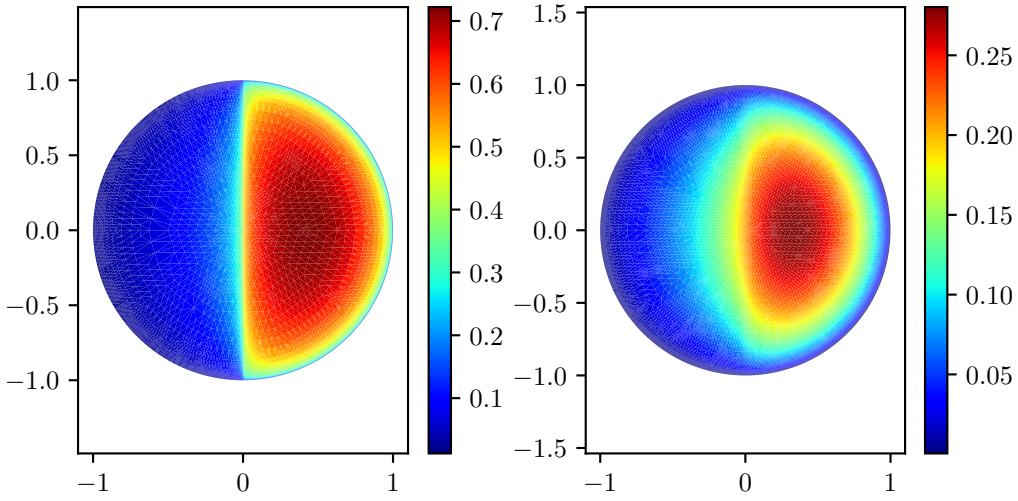


Figure 12.8: Solutions u corresponding to the discontinuous right-hand side $f(r, \theta) = 1_{|\theta|<\pi/2}$ for $s = 0.25$ and for $s = 0.75$.

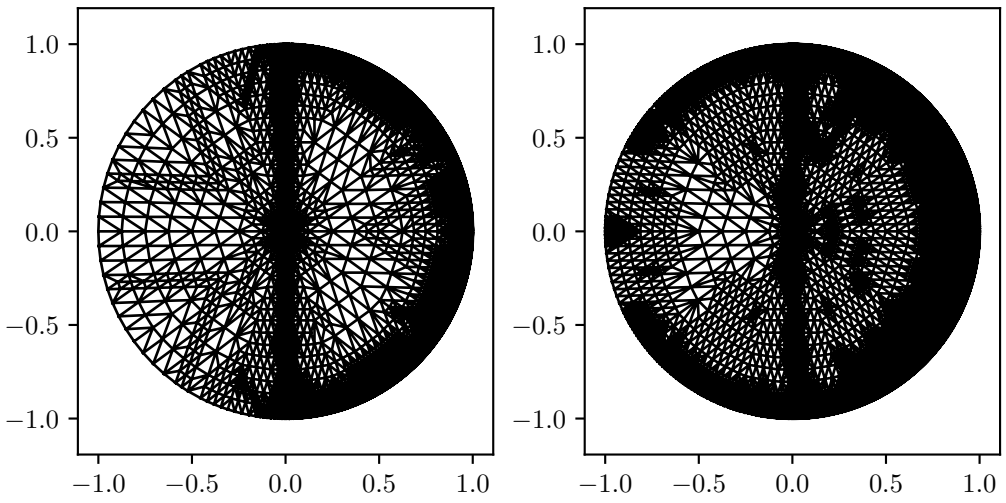


Figure 12.9: Adaptively refined meshes for the discontinuous right-hand side. $s = 0.25$ on the left, $s = 0.75$ on the right.

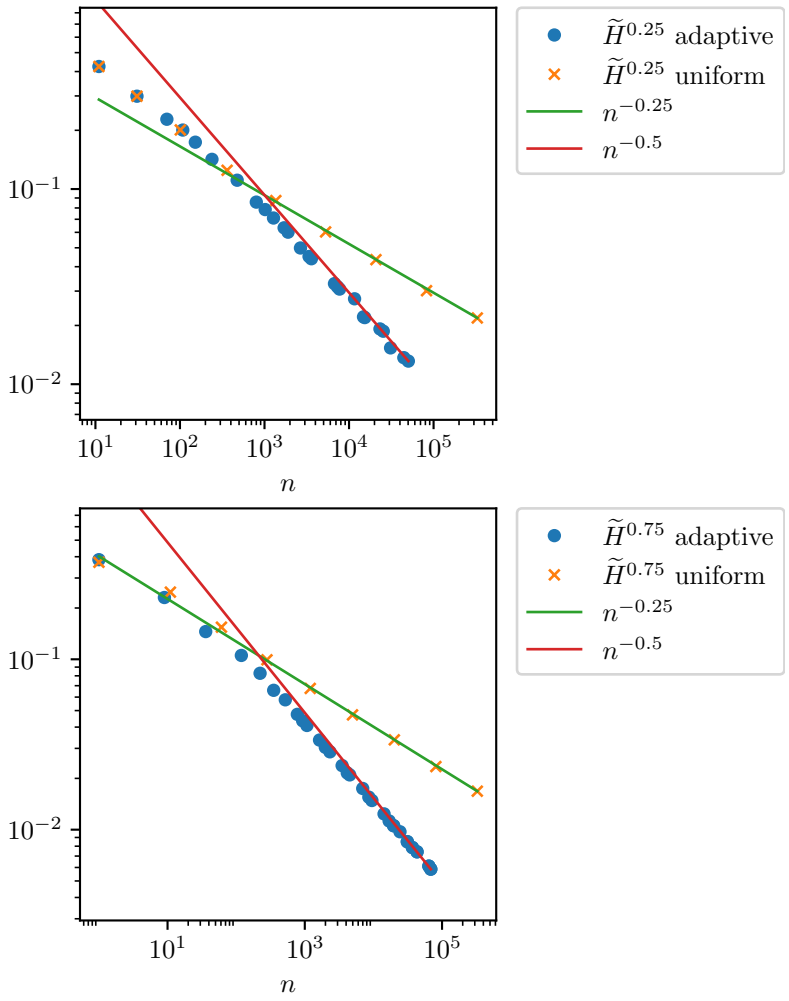


Figure 12.10: $\widetilde{H}^s$-error for the two-dimensional problem with discontinuous right-hand side $f(r,\theta) = 1_{|\theta|<\pi/2}$ and for uniform and adaptive refinement. $s = 0.25$ at the top, $s = 0.75$ at the bottom. For $s = 0.25$, uniform refinement results in convergence in $\widetilde{H}^s$-norm with rate $n^{-1/2}$, adaptive refinement achieves the optimal rate of n^{s-2}. For $s = 0.75$, both uniform and adaptive refinement results in convergence in $\widetilde{H}^s$-norm with rate $n^{-1/2}$.

12.1.3 L-shaped Domain

We propose to solve the fractional Poisson problem on the L-shaped domain $\Omega = [0,2]^2 \setminus [1,2]^2$ with constant right-hand side. Since no analytic solution is available, we solve the problem on a highly refined mesh and use this reference solution $u_{\underline{h}}$ to approximate the error. In Figure 12.13, we show $\widetilde{H}^s$- and L^2 error, as well as the a posteriori error estimate $\hat{\eta} = (\sum_i \hat{\eta}_i^2)^{1/2}$, which is based on the local a posteriori error estimators $\hat{\eta}_i$ given in (10.13). It can be observed that $\|u_h - u_{\underline{h}}\|_{\widetilde{H}^s}$ and the a posteriori error estimate η converge with optimal rate $n^{-1/2}$, and that $\|u_h - u_{\underline{h}}\|_{L^2}$ converges with rate $n^{-1/2-s/2}$. The apparent speed-up of convergence in $\widetilde{H}^s$- and L^2-norm for larger number of unknowns is due to the fact that the reference solution $u_{\underline{h}}$ is used in their computation instead of the true solution u. The a posteriori error estimate is obviously not suffer this deficiency, because it does not involve the reference solution $u_{\underline{h}}$. This example demonstrates that more complicated domains can easily be considered in the presented framework, and that convergence of optimal order can equally well be achieved.

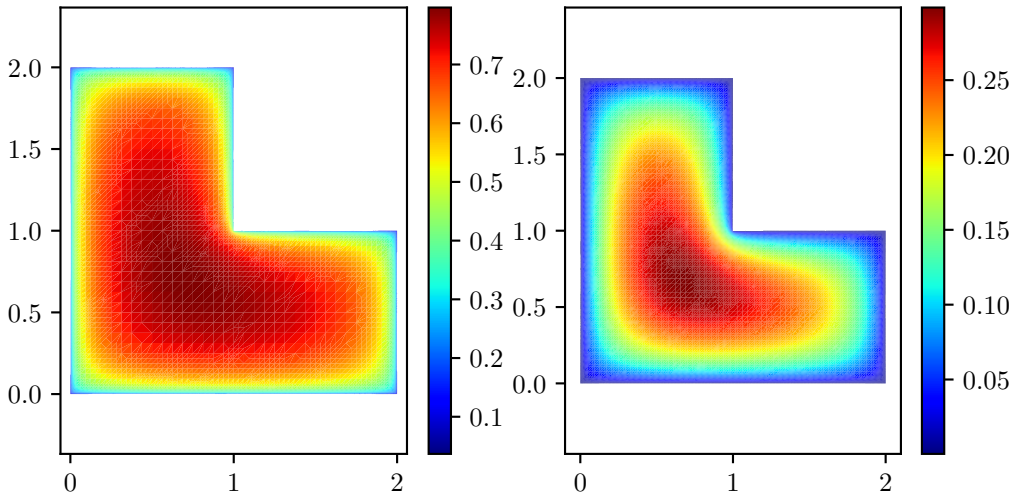


Figure 12.11: Solutions u on the L-shaped domain corresponding to a constant right-hand side for $s = 0.25$ and for $s = 0.75$.

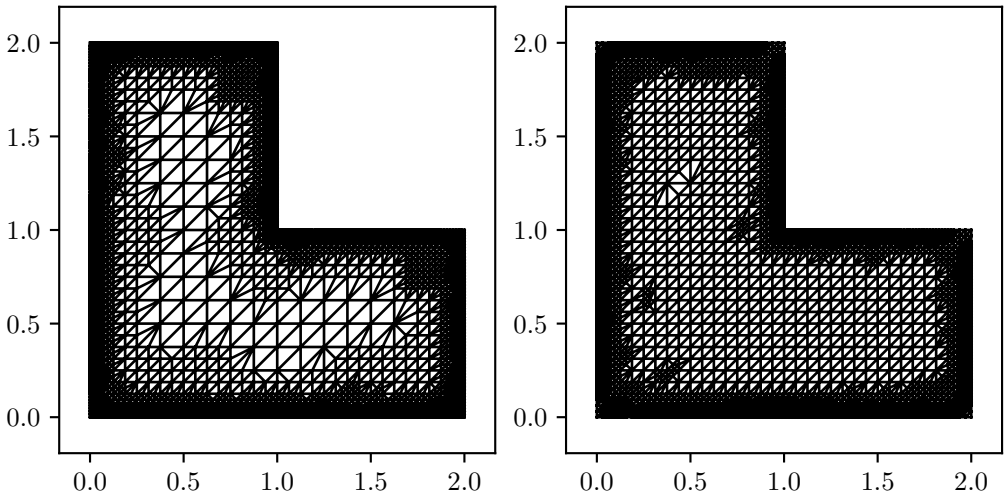


Figure 12.12: Adaptively refined meshes for the L-shaped domain. $s = 0.25$ on the left, $s = 0.75$ on the right.

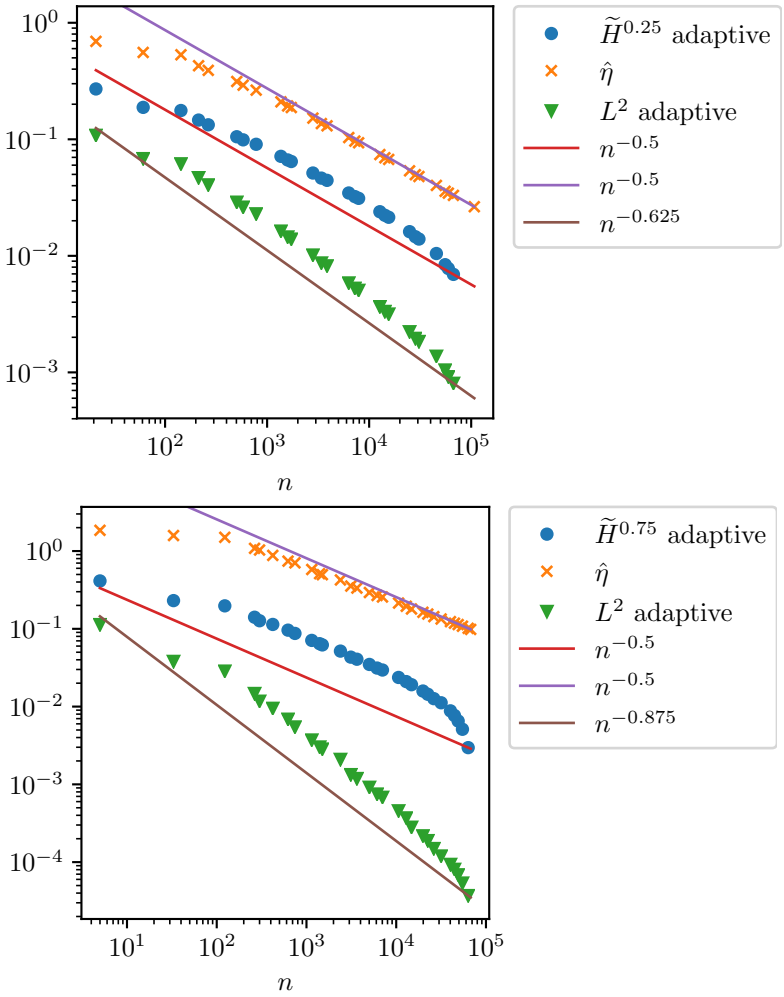


Figure 12.13: $\widetilde{H}^s$-, L^2-error and estimated error η for the L-shaped domain. $s = 0.25$ at the top, $s = 0.75$ at the bottom. Adaptive refinement results in convergence in $\widetilde{H}^s$-norm with optimal rate of $n^{-1/2}$, and the same holds for the a posteriori error $\hat{\eta}$. In L^2-norm, the rate $n^{-1/2-s/2}$ is observed.

12.1.4 Exploiting Symmetry

The example of the previous sections displayed symmetries and periodicity in the shape of the domain and the right-hand side and therefore in the solution. In particular, the solution shown in Figure 12.8 suggests that one should be able to exploit symmetry to halve the number of degrees of freedom required.

Let $\boldsymbol{B}$ be an orthogonal transformation such that for some K, $\boldsymbol{B}^K = \mathrm{Id}$, $\boldsymbol{B}\Omega = \Omega$, and let $\omega \subset \tilde{\omega} \subset \mathbb{R}^d$ be representative domains such that $\mathbb{R}^d = \bigcup_{i=0}^{K-1} \boldsymbol{B}^i \tilde{\omega}$ and $\Omega = \bigcup_{i=0}^{K-1} \boldsymbol{B}^i \omega$. Moreover, assume that $u\left(\boldsymbol{B}^i \vec{x}\right) = u\left(\vec{x}\right)$ and $f\left(\boldsymbol{B}^i \vec{x}\right) = f\left(\vec{x}\right)$. Examples of such transformations for $\Omega = B(0,1)$ include the reflection $\boldsymbol{B} = -\mathrm{Id}$ with $\tilde{\omega}$ and ω being the half-space and the half disc, or the rotation by angle $\theta = \frac{2\pi}{K}$ and the domains $\tilde{\omega} = \{(r\cos\phi, r\sin\phi) \in \mathbb{R}^2 \mid r \geq 0, \phi \in [0, 2\pi/K]\}$ and $\omega = \tilde{\omega} \cap \{r \leq 1\}$.

The variational formulation of the fractional Poisson problem eq. (10.3) then reduces to

$$\begin{aligned}
a\left(u,v\right) &= \frac{C(d,s)}{2} \int_\omega \int_\omega \sum_{i,j=0}^{K} \frac{\left(u\left(\boldsymbol{B}^i\vec{x}\right) - u\left(\boldsymbol{B}^j\vec{y}\right)\right)\left(v\left(\boldsymbol{B}^i\vec{x}\right) - v\left(\boldsymbol{B}^j\vec{y}\right)\right)}{\left|\boldsymbol{B}^i\vec{x} - \boldsymbol{B}^j\vec{y}\right|^{d+2s}} \\
&\quad + C(d,s) \int_\omega \int_{\tilde{\omega}\setminus\omega} \sum_{i,j=0}^{K} \frac{u\left(\boldsymbol{B}^i\vec{x}\right) v\left(\boldsymbol{B}^i\vec{x}\right)}{\left|\boldsymbol{B}^i\vec{x} - \boldsymbol{B}^j\vec{y}\right|^{d+2s}} \\
&= K\frac{C(d,s)}{2} \int_\omega \int_\omega \left(u\left(\vec{x}\right) - u\left(\vec{y}\right)\right)\left(v\left(\vec{x}\right) - v\left(\vec{y}\right)\right) k_{\text{eff}}\left(\vec{x},\vec{y}\right) \\
&\quad + KC(d,s) \int_\omega \int_{\tilde{\omega}\setminus\omega} u\left(\vec{x}\right) v\left(\vec{x}\right) k_{\text{eff}}\left(\vec{x},\vec{y}\right)
\end{aligned}$$

and

$$\begin{aligned}
\langle f, v\rangle &= \sum_{i=0}^{K-1} \int_\omega f\left(\boldsymbol{B}^i\vec{x}\right) v\left(\boldsymbol{B}^i\vec{x}\right) \\
&= K\int_\omega fv
\end{aligned}$$

with the effective kernel

$$k_{\text{eff}}\left(\vec{x},\vec{y}\right) = \sum_{i=0}^{K-1} \left|\vec{x} - \boldsymbol{B}^i\vec{y}\right|^{-d-2s}.$$

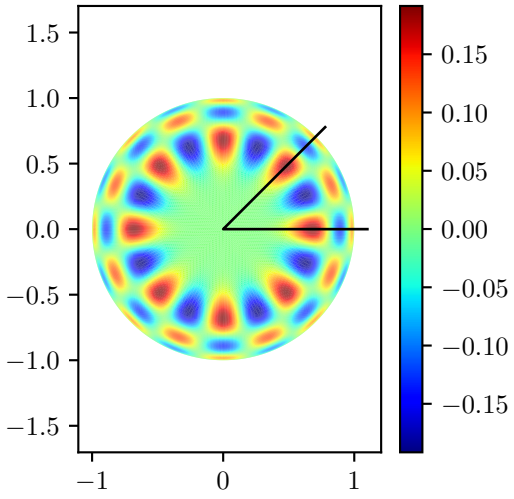


Figure 12.14: Periodic solution $u = u^{2D}_{2,8}$. Only the circular sector on the right is discretized, with periodic images taken into account through the effective kernel k_{eff}.

that accounts for the interaction with mirror images. Effectively, the computational effort for the assembly and solution of systems involving the dense stiffness matrix is divided by a factor of K, since only the sub-domain ω needs to be discretized. The complexity of the cluster method is expected to decrease from $\mathcal{O}\left(n\left(\log n\right)^{2d}\right)$ to $\mathcal{O}\left(n/K\left(\log n\right)^{2d}\right)$ since the assembly of the near-field contributions constitutes the bulk of the work.

To illustrate the approach, we solve the fractional Poisson problem with right-hand side $f_{2,8}$ on the unit disk. By leveraging the periodicity in the angular direction (and neglecting the mirror symmetry), we reduce to a problem on the circular sector $\tilde{\omega} = \{(r\cos\phi, r\sin\phi) \in \mathbb{R}^2 \mid r \geq 0, \phi \in [0, \pi/4]\}$ and $\omega = \tilde{\omega} \cap \{r \leq 1\}$. For simplicity, we employ a uniform mesh and assemble the dense system matrix. The solution is shown in Figure 12.14.

12.2 Fractional Heat Equation

The fractional heat equation is given by

$$\begin{aligned} u_t + (-\Delta)^s u &= f && \text{in } \Omega, \\ u &= 0 && \text{in } \Omega^c. \end{aligned}$$

We propose to approximate the problem using an implicit method in time. The simplest such scheme is the backward Euler method

$$(\boldsymbol{M} + \Delta t\ \boldsymbol{A}^s)\, \vec{u}^{k+1} = \boldsymbol{M}\vec{u}^k + \vec{f}^{k+1},$$

where $u(\cdot, k\Delta t) \approx \sum_i u_i^k \phi_i$ and $f_i^k = (f(\cdot, k\Delta t), \phi_i)$.

More generally, let us assume that a scheme of order α is used in time, and that a globally quasi-uniform mesh with mesh size h is used in space. In order to obtain optimal convergence in L^2-norm, in view of Theorem 37, we shall choose $\Delta t^\alpha \sim h^{1/2+\min(1/2,s)}$, i.e.

$$\Delta t_{L^2} \sim h^{\min(2,1+2s)/(2\alpha)}.$$

On the other hand, if optimal $\widetilde{H}^s(\Omega)$-convergence is desired, we need $\Delta t_{\widetilde{H}^s(\Omega)} \sim h^{1/(2\alpha)}$, see Lemma 36. Consequently, if an order α scheme is used for time stepping, with optimal time step Δt_{L^2} or $\Delta t_{\widetilde{H}^s(\Omega)}$, we find by Lemma 42 that the condition numbers of the iteration matrix satisfy

$$\begin{aligned} \kappa\left(\boldsymbol{M} + \Delta t_{L^2}\ \boldsymbol{A}^s\right) &\le C\left(1 + h^{\min(2,1+2s)/(2\alpha)-2s}\right), \\ \kappa\left(\boldsymbol{M} + \Delta t_{\widetilde{H}^s(\Omega)}\ \boldsymbol{A}^s\right) &\le C\left(1 + h^{1/(2\alpha)-2s}\right). \end{aligned}$$

In particular, this shows that the condition number will not grow as the mesh size decreases if $s \le 1/(4\alpha - 2)$ in the L^2 case, or if $s \le 1/(4\alpha)$ in the $\widetilde{H}^s(\Omega)$ case.

We illustrate the consequences of the above result in the case of a second order accurate time stepping scheme ($\alpha = 2$), and for $s = 0.25$ and $s = 0.75$. In the case of $s = 0.25$, $\Delta t_{L^2} \sim h^{3/8}$ and $\kappa\left(\boldsymbol{M} + \Delta t_{L^2}\ \boldsymbol{A}^s\right) \sim 1 + h^{-1/8}$. This suggests

Method	$\Delta t = \Delta t_{L^2}$	$\Delta t = \Delta t_{\widetilde{H}^s(\Omega)}$
Conjugate Gradient	$n^{1+2s/d-\min(2,1+2s)/(2\alpha d)} \left(\log n\right)^{2d}$	$n^{1+2s/d-1/(2\alpha d)} \left(\log n\right)^{2d}$
Multigrid	$n \left(\log n\right)^{2d}$	$n \left(\log n\right)^{2d}$

Table 12.1: Complexity of different solvers for $\left(\boldsymbol{M} + \Delta t \boldsymbol{A}^s\right) \vec{u} = \vec{b}$ for $\Delta t = \Delta t_{L^2}$ and $\Delta t = \Delta t_{\widetilde{H}^s(\Omega)}$ for an α-order time stepping scheme.

that the conjugate gradient method will uniformly in h deliver good results. The convergence of the multigrid method does not depend on the condition number and is essentially independent of h. This is indeed what is observed in the top part of Figure 12.15. In Figure 12.16, the number of iterations is shown. It can be observed that for $s = 0.25$ both the multigrid and the conjugate gradient solver require an essentially constant number of iterations for varying values of Δt.

On the other hand, for $s = 0.75$, $\Delta t_{L^2} \sim h^{1/2}$ and $\kappa\left(\boldsymbol{M} + \Delta t_{L^2}\, \boldsymbol{A}^s\right) \sim 1 + h^{-1}$. Therefore, the condition number increases as h goes to zero, and we expect that multigrid asymptotically outperforms the CG solver. This is indeed what is observed in Figures 12.15 and 12.16.

The complexities of the different solvers for different choices of time step size are summarized in Table 12.1.

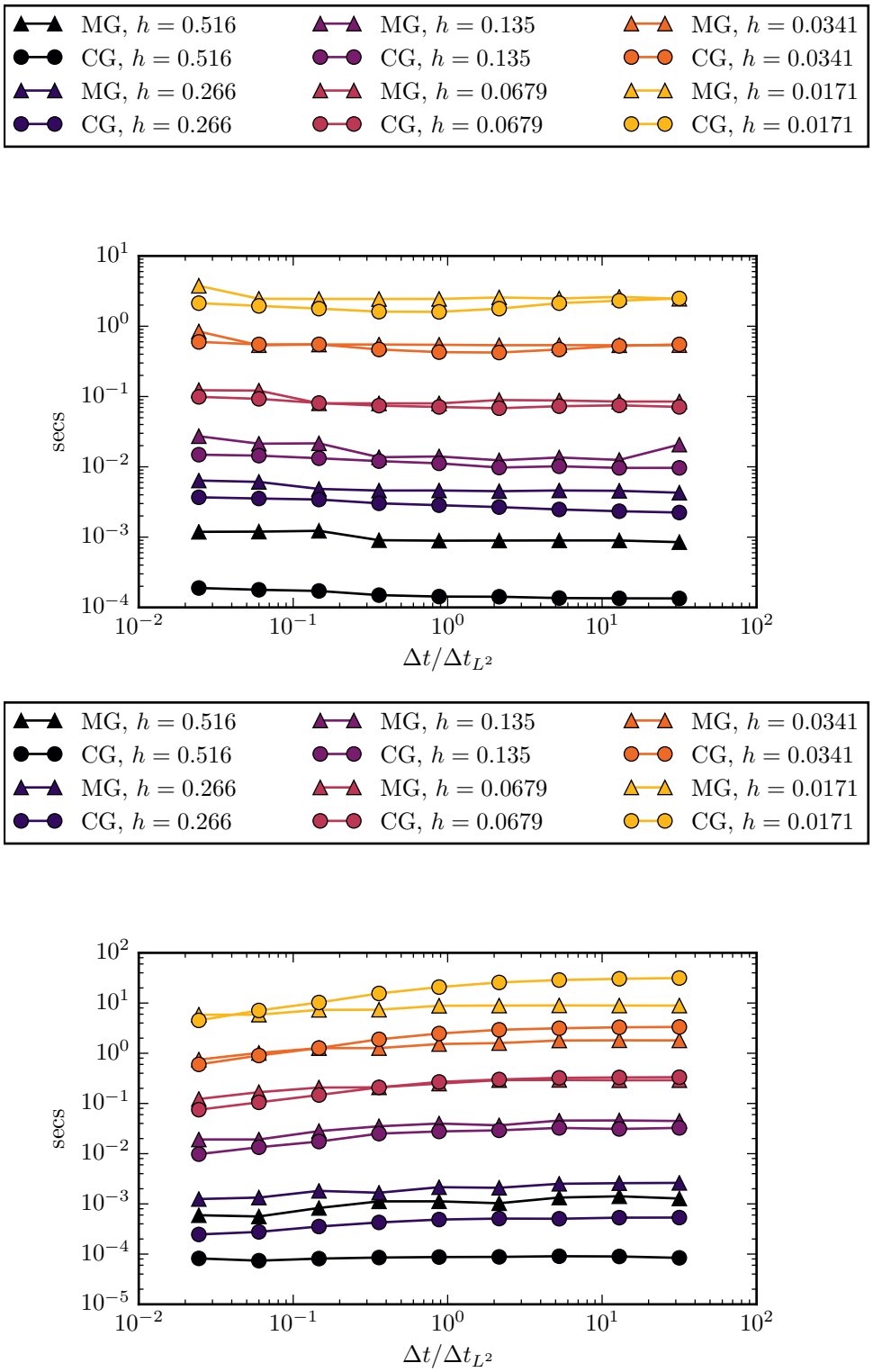


Figure 12.15: Timings in seconds for CG and MG depending on Δt for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*). It can be observed that, for $s = 0.25$, the conjugate gradient method is essentially on par with the multigrid solver. For $s = 0.75$, the multigrid solver asymptotically outperforms the conjugate gradient method, since the condition number $\kappa\left(\boldsymbol{M} + \Delta t_{L^2} \boldsymbol{A}^s\right)$ grows as h^{-1}.

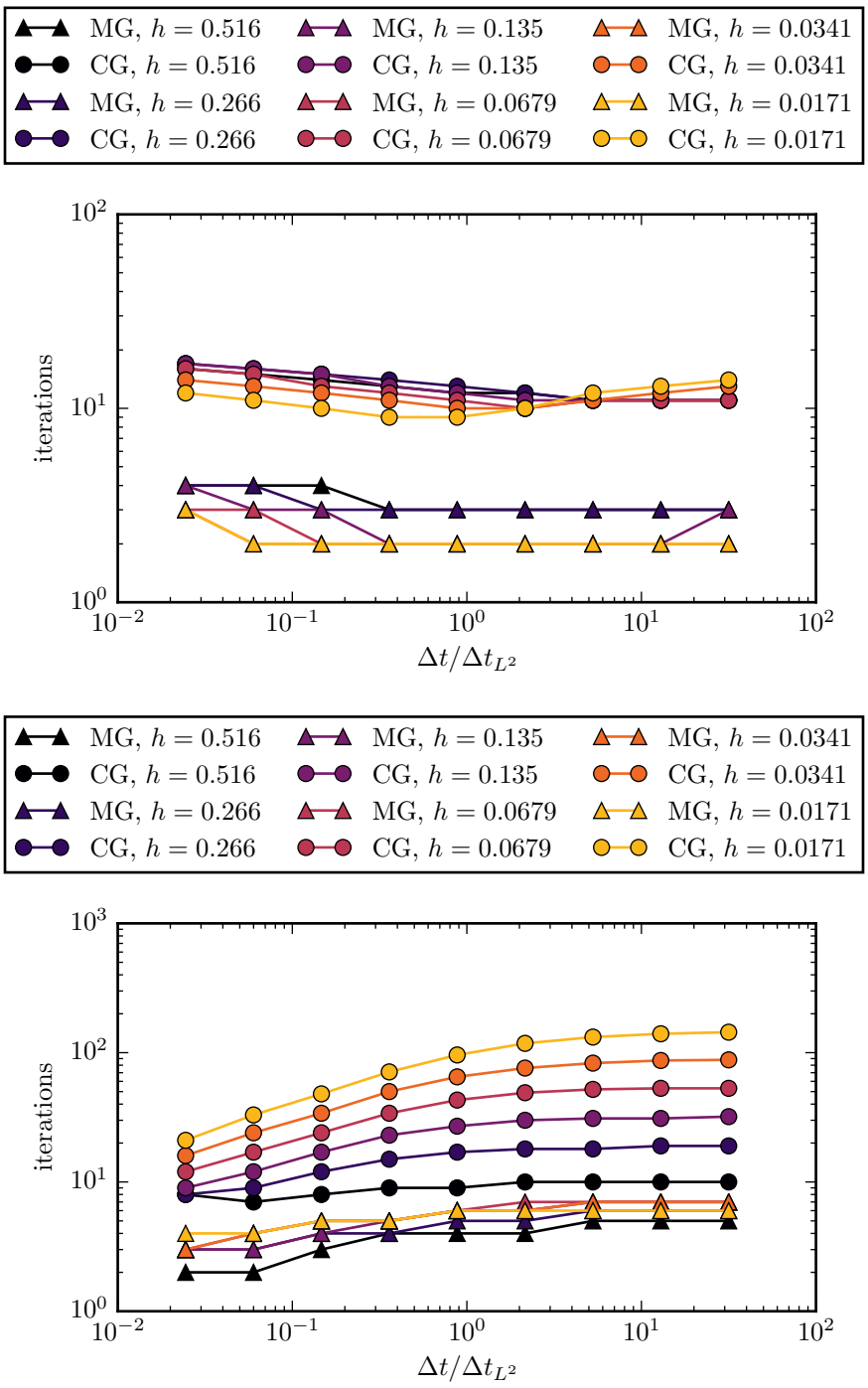


Figure 12.16: Number of iterations for CG and MG depending on Δt for $s = 0.25$ (*top*) and $s = 0.75$ (*bottom*). For $s = 0.25$, the number of iterations is essentially independent of Δt. For $s = 0.75$, the number of iterations of the multigrid solver is independent of Δt, but the iterations count for conjugate gradient grows with $h^{-1/2}$.

12.3 Fractional Reaction-Diffusion Systems

In [60], a space-fractional Brusselator model was analyzed and compared to the classical integer-order case. The coupled system of equations is given by

$$\frac{\partial X}{\partial t} = -D_X \left(-\Delta\right)^{\alpha} X + A - (B+1)X + X^2 Y,$$
$$\frac{\partial Y}{\partial t} = -D_Y \left(-\Delta\right)^{\beta} Y + BX - X^2 Y.$$

Here, D_X and D_Y are diffusion coefficients, A and B are reaction parameters, and α and β determine the type of diffusion. By rewriting the solutions as deviations from the stationary solution $X = A$, $Y = B/A$ and rescaling, one obtains

$$\frac{\partial u}{\partial t} = -\left(-\Delta\right)^{\alpha} u + (B-1)u + Q^2 v + \frac{B}{Q}u^2 + 2Quv + u^2 v, \tag{12.2}$$
$$\eta^2 \frac{\partial v}{\partial t} = -\left(-\Delta\right)^{\beta} v - Bu - Q^2 v - \frac{B}{Q}u^2 - 2Quv - u^2 v, \tag{12.3}$$

with $\eta = \sqrt{D_Y / D_X^{\beta/\alpha}}$ and $Q = A\eta$.

In [60] the equations were augmented with periodic boundary conditions and approximated using a pseudospectral method for various different parameter combinations. Here, thanks to the foregoing developments, we have the flexibility to handle more general domains and, in particular, we consider the case where Ω corresponds to a Petri-dish. We solve the above set of equations using a second order accurate IMEX scheme proposed by Koto [72], whose Butcher tableaux are given by Table 12.2. The diffusive parts are treated implicitly and therefore require the solution of several systems all of which are of the type $\boldsymbol{M} + c\Delta t \boldsymbol{A}^s$ with appropriate values of c.

0	0			
1	0	1		
1/2	0	-1/2	1	
1	0	-1	1	1
	0	-1	1	1

0				
1	1			
1/2	0	0		
1	0	0	1	
	0	0	1	0

Table 12.2: IMEX scheme by Koto. Implicit scheme on the left, explicit on the right.

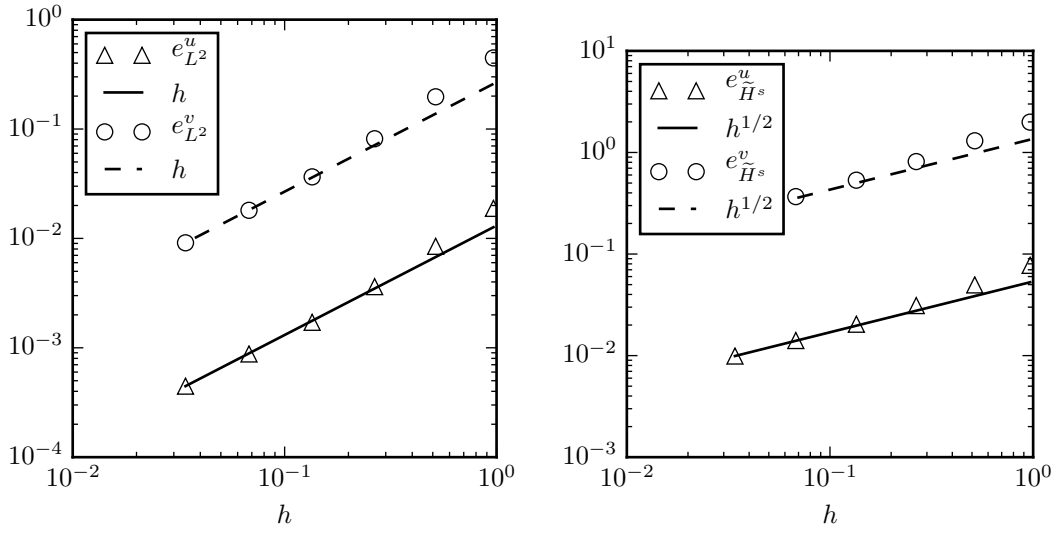


Figure 12.17: Error in L^2-norm (*top*) and $\widetilde{H}^s(\Omega)$-norm (*bottom*) in the Brussela-tor model. Optimal orders of convergence are achieved. (Compare Theorem 37 and lemma 36.)

In order to verify the correct convergence behavior, we add forcing functions f and g to the system, chosen such that the analytic solution is given by

$$u = \eta \sin(t) u^s(\vec{x}),$$
$$v = \eta^{-1} \cos(2t) u^s(\vec{x}),$$

for suitable initial conditions, where u^s is the solution of the fractional Poisson problem with constant right-hand side. We take $\alpha = \beta = 0.75$, and choose $\Delta t \sim h^{1/2}$, since we already saw that the rate of the spatial approximation in L^2-norm is of order h. We measure the error as

$$e^u_{L^2} = \max_{0 \le t_i \le 10} \left\| u(t_i, \cdot) - u^i_h \right\|_{L^2}, \qquad e^v_{L^2} = \max_{0 \le t_i \le 10} \left\| v(t_i, \cdot) - v^i_h \right\|_{L^2},$$
$$e^u_{\widetilde{H}^s(\Omega)} = \max_{0 \le t_i \le 10} \left\| u(t_i, \cdot) - u^i_h \right\|_{\widetilde{H}^s(\Omega)}, \qquad e^v_{\widetilde{H}^s(\Omega)} = \max_{0 \le t_i \le 10} \left\| v(t_i, \cdot) - v^i_h \right\|_{\widetilde{H}^s(\Omega)}.$$

From the error plots in Figure 12.17, it can be observed that $e_{L^2} \sim h$ and $e_V \sim h^{1/2}$, as expected.

Having verified the accuracy of the method, we turn to the solution of the system eqs. (12.2) and (12.3) augmented with exterior Neumann conditions as described

in Section 10.2. Golovin, Matkowsky and Volpert [60] observed that for $\eta = 0.2$, $B = 1.22$ and $Q = 0.1$, a single localized perturbation would first form a ring and then break up into spots. The radius of the ring and the number of resulting spots increases as the fractional orders are decreased. In Figure 12.18, simulation results for $\alpha = \beta = 0.625$ and $\alpha = \beta = 0.75$ are shown. We observe that in both cases, an initially circular perturbation develops into a ring. Lower diffusion coefficients do lead to a larger ring, which breaks up later and into more spots. In the last row, we can see that the resulting spots start to replicate and spread out over the whole domain.

Another choice of parameters leads to stripes in the solution. For $\alpha = \beta = 0.75$, $\eta = 0.2$, $B = 6.26$ and $Q = 2.5$, and a random initial condition, stripes without directionality form in the whole domain. This is in alignment with the theoretical considerations of Golovin, Matkowsky and Volpert [60].

12.4 Discussion

In this work, we have presented the numerical approximation of the fractional Laplacian operator for the case of globally quasi-uniform meshes as well as locally refined meshes. In order to overcome slow convergence due to the inherent singularity of the solution close to the boundary, we introduced a posteriori error indicators and adaptive mesh refinement. The non-locality of the problem leads to dense matrices and solution complexity of $\mathcal{O}(n^3)$ using direct solvers and at best $\mathcal{O}(n^2)$ using an optimal iterative method. We showed that by employing a cluster method, assembly, matrix-vector product and computation of the error estimators scale quasi-linear in the number of unknowns. This means that the solution of the fractional Poisson problem (10.1) and the fractional Heat equation (10.2) can be performed in the same complexity and with equal (or even better) rate of convergence as of the standard integer order equivalents. Through several one- and

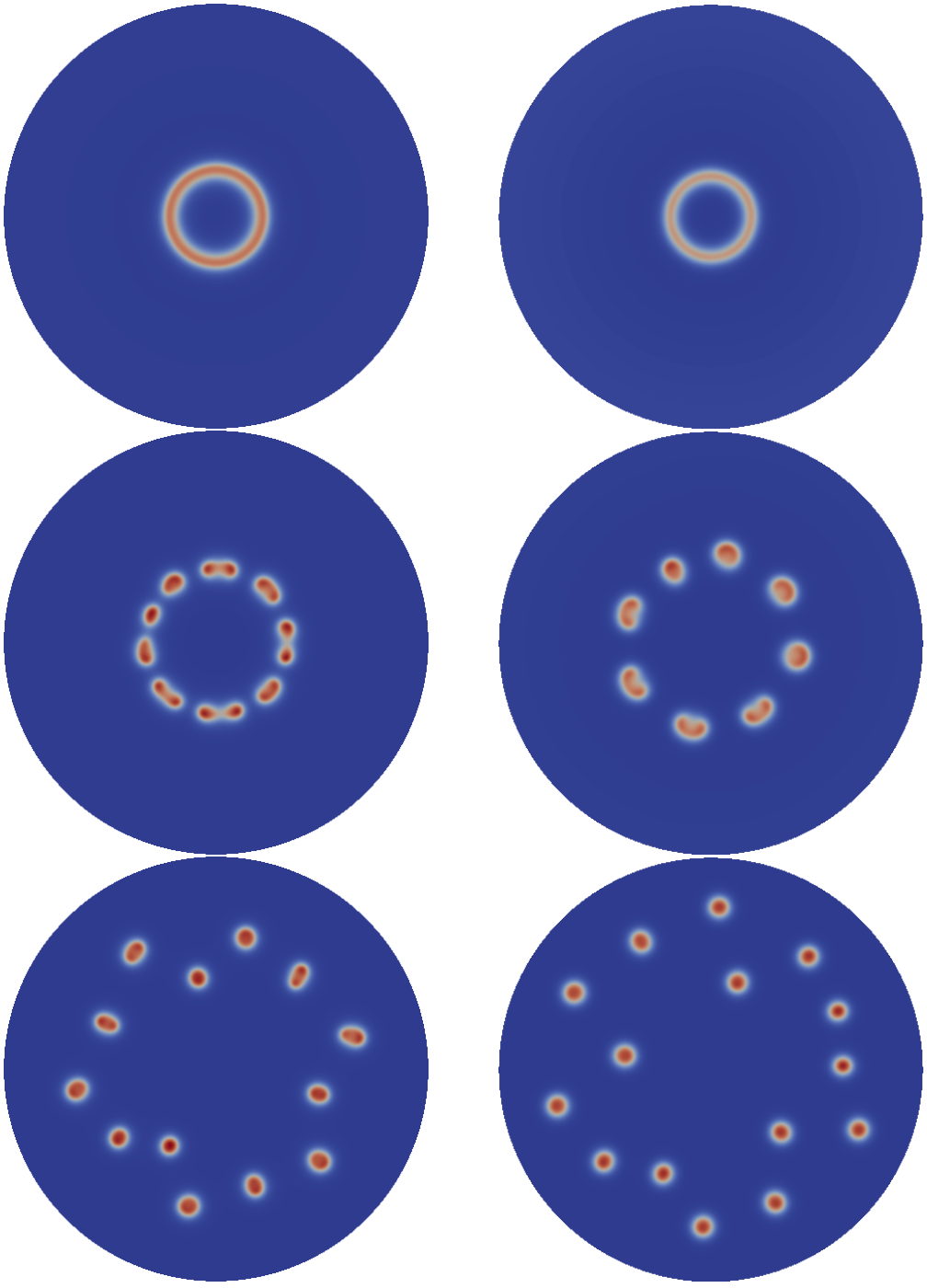


Figure 12.18: Localized spot solutions of the Brusselator system with $\alpha = \beta = 0.625$ (*left*) and $\alpha = \beta = 0.75$ (*right*). u is shown in both cases, and time progresses from top to bottom. The initial perturbation was identical in both cases. The initial perturbation in the centre of the domain forms a ring, whose radius is bigger if the fractional orders of diffusion α, β are smaller. The ring breaks up into several spots, which start to replicate and spread out over the whole domain. $n \approx 50,000$ unknowns were used in the finite element approximation.

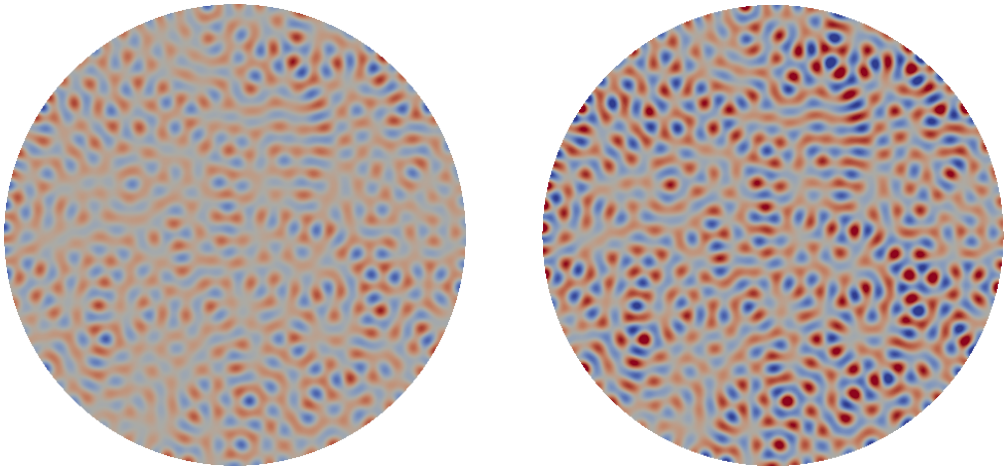


Figure 12.19: Stripe solutions of the Brusselator system with $\alpha = \beta = 0.75$. u is shown on the left, and v on the right. The random initial condition leads to the formation of stripes throughout the domain. $n \approx 50,000$ unknowns were used in the finite element approximation.

two-dimensional examples, we illustrated that the predicted optimal rates of convergence are obtained. We demonstrated how symmetries and periodicity in the problem can be leveraged to reduce the problem size even further.

Chapter 13

Future Work on Fractional and Non-Local Equations

In the present work, we focused on the case of fractional equations posed in $d = 1$ and $d = 2$ dimensions. The generalization to higher dimensions does not pose any fundamental difficulties. Expressions for the contributions of touching pairs of elements that contain a singularity would have to be rederived. Alternatively, Chernov, von Petersdorff, and Schwab proposed a procedure to transform quadrature rules with suitable weights from the hypercube $[0, 1]^{2d}$ to pairs of simplices [33, 34, 35]. The advantage of this method is that simplices in any dimension d can be handled. A preliminary implementation seems to suggest, however, that comparatively more quadrature points are necessary to obtain the same quadrature error.

Moreover, the use of higher order elements might be worth exploring. Preliminary results using piecewise quadratic instead of piecewise linear finite elements suggests that faster convergence can indeed be achieved, but a more thorough comparison of computational costs would be in order, since, for example, the computation of the error indicators introduced in Section 10.4 becomes more expensive.

At present, we also limited ourselves to homogeneous boundary conditions of Dirichlet and Neumann type. Non-homogeneous Dirichlet conditions can be taken

into account in a relatively straightforward way, provided that the contribution to the right-hand side of the exterior condition can be evaluated efficiently using numerical quadrature. The case of the Neumann condition is more subtle, as several generalizations of the integer order non-homogeneous Neumann condition have been proposed in the literature [40].

The approach described to obtain a sparse approximation to the dense system matrix for the fractional Laplacian does not rely strongly on the form of the interaction kernel $k(\vec{x}, \vec{y}) = |\vec{x} - \vec{y}|^{-(d+2s)}$, and generalizations to different kernels such as the one used in peridynamics [85] or kernels that can model anisotropic behavior are therefore possible.

Finally, it would be of interest to compare solutions to the fractional Poisson problem involving the integral definition of the fractional Laplacian with solutions obtained using the spectral fractional Laplacian. Solutions to the later can be obtained by solving an extension problem in $d+1$ dimensions [22]. Nochetto, Otárola, and Salgado leveraged this and proposed a method that converges essentially with rate $\mathcal{O}\left(n^{-1/(d+1)}\right)$, n being the number of unknows, provided that the solution is sufficiently regular [31, 80]. Since the two fractional operators, albeit different [84], seem to enjoy very similar properties, the question arises whether a method can be devised to solve equations involving the spectral fractional Laplacian with error decaying at the same rate as its integral counterpart, namely $\mathcal{O}\left(n^{-1/d}\right)$.

Bibliography

[1] Gabriel Acosta, Francisco M. Bersetche, and Juan-Pablo Borthagaray. “A short FE implementation for a 2d homogeneous Dirichlet problem of a Fractional Laplacian”. In: *ArXiv e-prints* (Oct. 2016).

[2] Gabriel Acosta and Juan Pablo Borthagaray. “A fractional Laplace equation: regularity of solutions and Finite Element approximations”. In: *ArXiv e-prints* (July 2015).

[3] Emmanuel Agullo, Luc Giraud, Pablo Salas, and Mawussi Zounon. “Interpolation-Restart Strategies for Resilient Eigensolvers”. In: *SIAM Journal on Scientific Computing* 38.5 (2016), pp. C560–C583. DOI: 10.1137/15M1042115.

[4] Emmanuel Agullo, Luc Giraud, Pablo Salas, and Mawussi Zounon. “Recover-Restart Strategies for Resilient Parallel Numerical Linear Algebra Solvers”. In: *International Workshop on Parallel Matrix Algorithms and Applications (PMAA 2014)*. Lugano, Switzerland, July 2014.

[5] Mark Ainsworth and Christian Glusa. “Aspects of an adaptive finite element method for the fractional Laplacian: a priori and a posteriori error estimates, efficient implementation and multigrid solver”.

[6] Mark Ainsworth and Christian Glusa. “Is the multigrid method fault tolerant? The Multilevel Case.” In: *SIAM Journal on Scientific Computing* (Submitted).

[7] Mark Ainsworth and Christian Glusa. “Is the Multigrid Method Fault Tolerant? The Two-Grid Case”. In: *SIAM Journal on Scientific Computing* 39.2 (2017), pp. C116–C143. DOI: 10.1137/16M1100691.

[8] Mark Ainsworth and Christian Glusa. “Numerical Mathematics and Advanced Applications - ENUMATH 2015: Proceedings of ENUMATH 2015”. In: Cham: Springer International Publishing, 2015. Chap. Multigrid at Scale?, pp. 237–253. ISBN: 978-3-319-39929-4. DOI: 10.1007/978-3-319-39929-4_24.

[9] Mark Ainsworth and Christian Glusa. “Towards an Efficient Finite Element Method for the Integral Fractional Laplacian on Polygonal Domains”. In: *TBD* (2017).

[10] Mark Ainsworth, William McLean, and Thanh Tran. “The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling”. In: *SIAM Journal on Numerical Analysis* 36.6 (1999), pp. 1901–1932.

[11] Algirdas Avižienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr. “Basic concepts and taxonomy of dependable and secure computing”. In: *IEEE Transactions on Dependable and Secure Computing* 1.1 (2004), pp. 11–33.

[12] Randolph E. Bank, Andrew H. Sherman, and Alan Weiser. “Some refinement algorithms and data structures for regular local mesh refinement”. In: *Scientific Computing, Applications of Mathematics and Computing to the Physical Sciences* 1 (1983), pp. 3–17.

[13] Peter W. Bates. “On some nonlocal evolution equations arising in materials science”. In: *Nonlinear dynamics and evolution equations* 48 (2006), pp. 13–52.

[14] Austin R Benson, Sven Schmit, and Robert Schreiber. “Silent Error Detection in Numerical Time-stepping Schemes”. In: *Int. J. High Perform. Comput. Appl.* 29.4 (Nov. 2015), pp. 403–421. ISSN: 1094-3420. DOI: 10.1177/1094342014532297.

[15] David A. Benson, Stephen W. Wheatcraft, and Mark M. Meerschaert. “Application of a fractional advection-dispersion equation”. In: *Water Resources Research* 36.6 (2000), pp. 1403–1412.

[16] Dimitri P Bertsekas. “Distributed asynchronous computation of fixed points”. In: *Mathematical Programming* 27.1 (1983), pp. 107–120.

[17] Krzysztof Bogdan, Krzysztof Burdzy, and Zhen-Qing Chen. “Censored stable processes”. In: *Probability theory and related fields* 127.1 (2003), pp. 89–152.

[18] Juan Pablo Borthagaray, Leandro M. Del Pezzo, and Sandra Martinez. “Finite element approximation for the fractional eigenvalue problem”. In: *ArXiv e-prints* (Mar. 2016).

[19] Philippe Bougerol and Jean Lacroix. *Products of random matrices with applications to Schrödinger operators*. Vol. 8. Progress in Probability and Statistics. Boston, MA: Birkhäuser Boston Inc., 1985, pp. xii+283. ISBN: 0-8176-3324-3. DOI: 10.1007/978-1-4684-9172-2.

[20] Dietrich Braess. *Finite elements. Theory, fast solvers and applications in solid mechanics. Translated from German by Larry L. Schumaker. 3rd ed.* English. Cambridge: Cambridge University Press, 2007. DOI: 10.1017/CBO9780511618635.

[21] James H. Bramble. *Multigrid methods*. Vol. 294. CRC Press, 1993.

[22] Luis Caffarelli and Luis Silvestre. “An extension problem related to the fractional Laplacian”. In: *Communications in Partial Differential Equations* 32.8 (2007), pp. 1245–1260.

[23] Xiao-Chuan Cai, Maksymilian Dryja, and Marcus Sarkis. “Restricted additive Schwarz preconditioners with harmonic overlap for symmetric positive definite linear systems”. In: *SIAM Journal on Numerical Analysis* 41.4 (2003), pp. 1209–1231.

[24] Xiao-Chuan Cai and Marcus Sarkis. “A restricted additive Schwarz preconditioner for general sparse linear systems”. In: *SIAM Journal on Scientific Computing* 21.2 (1999), pp. 792–797.

[25] Xiao-Chuan Cai and Jun Zou. “Some observations on the l2 convergence of the additive Schwarz preconditioned GMRES method”. In: *Numerical linear algebra with applications* 9.5 (2002), pp. 379–397.

[26] Jon Calhoun, Luke Olson, Marc Snir, and William D. Gropp. “Towards a More Fault Resilient Multigrid Solver”. In: *Proceedings of the Symposium on High Performance Computing*. HPC ’15. Alexandria, Virginia: Society for Computer Simulation International, 2015, pp. 1–8. ISBN: 978-1-5108-0101-1.

[27] Franck Cappello. “Fault tolerance in petascale/exascale systems: Current knowledge, challenges and research opportunities”. In: *International Journal of High Performance Computing Applications* 23.3 (2009), pp. 212–226.

[28] Franck Cappello, Al Geist, Bill Gropp, Laxmikant Kale, Bill Kramer, and Marc Snir. “Toward exascale resilience”. In: *International Journal of High Performance Computing Applications* 23.4 (2009), pp. 374–388.

[29] Franck Cappello, Al Geist, William Gropp, Sanjay Kale, Bill Kramer, and Marc Snir. “Toward exascale resilience: 2014 update”. In: *Supercomputing frontiers and innovations* 1.1 (2014), pp. 5–28.

[30] Marc Casas, Bronis R. de Supinski, Greg Bronevetsky, and Martin Schulz. “Fault Resilience of the Algebraic Multi-grid Solver”. In: *Proceedings of the 26th ACM International Conference on Supercomputing*. ICS ’12. San Servolo

Island, Venice, Italy: ACM, 2012, pp. 91–100. ISBN: 978-1-4503-1316-2. DOI: 10.1145/2304576.2304590.

[31] Long Chen, Ricardo Nochetto, Enrique Otárola, and Abner Salgado. "Multilevel methods for nonuniformly elliptic operators and fractional diffusion". In: *Mathematics of Computation* (2016).

[32] Zhen-Qing Chen and Panki Kim. "Green function estimate for censored stable processes". In: *Probability Theory and Related Fields* 124.4 (2002), pp. 595–610.

[33] Alexey Chernov and Christoph Schwab. "Exponential convergence of Gauss–Jacobi quadratures for singular integrals over simplices in arbitrary dimension". In: *SIAM Journal on Numerical Analysis* 50.3 (2012), pp. 1433–1455.

[34] Alexey Chernov, Tobias von Petersdorff, and Christoph Schwab. "Exponential convergence of hp quadrature for integral operators with Gevrey kernels". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 45.3 (2011), pp. 387–422.

[35] Alexey Chernov, Tobias von Petersdorff, and Christoph Schwab. "Quadrature algorithms for high dimensional singular integrands on simplices". In: *Numerical Algorithms* 70.4 (2015), pp. 847–874.

[36] Patrick Ciarlet. "Analysis of the Scott–Zhang interpolation in the fractional order Sobolev spaces". In: *Journal of Numerical Mathematics* 21.3 (2013), pp. 173–180.

[37] Andrea Crisanti, Giovanni Paladin, and Angelo Vulpiani. *Products of random matrices*. Springer, 1993.

[38] Tao Cui, Jinchao Xu, and Chen-Song Zhang. "An error-resilient redundant subspace correction method". In: *Computing and Visualization in Science* 18.2 (2017), pp. 65–77. ISSN: 1433-0369. DOI: 10.1007/s00791-016-0270-6.

[39] Marta D'Elia and Max Gunzburger. "The fractional Laplacian operator on bounded domains as a special case of the nonlocal diffusion operator". In: *Computers & Mathematics with Applications* 66.7 (2013), pp. 1245–1260. ISSN: 0898-1221. DOI: http://dx.doi.org/10.1016/j.camwa.2013.07.022.

[40] Serena Dipierro, Xavier Ros-Oton, and Enrico Valdinoci. "Nonlocal problems with Neumann boundary conditions". In: *ArXiv e-prints* (July 2014).

[41] Victorita Dolean, Pierre Jolivet, and Frédéric Nataf. "An Introduction to Domain Decomposition Methods: algorithms, theory and parallel implementation". France, Jan. 2015.

[42] Jack Dongarra, Jeffrey Hittinger, John Bell, Luis Chacon, Robert Falgout, Michael Heroux, Paul Hovland, Esmond Ng, Clayton Webster, and Stefan Wild. *Applied Mathematics Research for Exascale Computing*. Tech. rep. Lawrence Livermore National Laboratory (LLNL), Livermore, CA, Feb. 2014. DOI: 10.2172/1149042.

[43] Willy Dörfler. "A convergent adaptive algorithm for Poisson's equation". In: *SIAM Journal on Numerical Analysis* 33.3 (1996), pp. 1106–1124.

[44] Michael G Duffy. "Quadrature over a pyramid or cube of integrands with a singularity at a vertex". In: *SIAM journal on Numerical Analysis* 19.6 (1982), pp. 1260–1262.

[45] Bartłomiej Dyda, Alexey Kuznetsov, and Mateusz Kwaśnicki. "Fractional Laplace Operator and Meijer G-function". In: *Constructive Approximation* (2016), pp. 1–22. ISSN: 1432-0940. DOI: 10.1007/s00365-016-9336-4.

[46] James Elliott, Mark Hoemmen, and Frank Mueller. "Evaluating the impact of SDC on the GMRES iterative solver". In: *Parallel and Distributed Processing Symposium, 2014 IEEE 28th International*. IEEE. May 2014, pp. 1193–1202. DOI: 10.1109/IPDPS.2014.123.

[47] James Elliott, Mark Hoemmen, and Frank Mueller. “Exploiting data representation for fault tolerance”. In: *Journal of Computational Science* (2016).

[48] James Elliott, Kishor Kharbas, David Fiala, Frank Mueller, Kurt Ferreira, and Christian Engelmann. “Combining Partial Redundancy and Checkpointing for HPC”. In: *Proceedings of the 2012 IEEE 32Nd International Conference on Distributed Computing Systems*. ICDCS ’12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 615–626. ISBN: 978-0-7695-4685-8. DOI: 10.1109/ICDCS.2012.56.

[49] Mark Embree and Lloyd N. Trefethen. “Growth and Decay of Random Fibonacci Sequences”. English. In: *Proceedings: Mathematical, Physical and Engineering Sciences* 455.1987 (1999), pp. 2471–2485. ISSN: 13645021.

[50] Alexandre Ern and Jean-Luc Guermond. *Theory and Practice of Finite Elements*. Applied Mathematical Sciences 159. New York, NY: Springer, 2004. DOI: 10.1007/978-1-4757-4355-5.

[51] Birgit Faermann. “Localization of the Aronszajn-Slobodeckij norm and application to adaptive boundary element methods Part II. The three-dimensional case”. In: *Numerische Mathematik* 92.3 (2002), pp. 467–499.

[52] Birgit Faermann. “Localization of the Aronszajn-Slobodeckij norm and application to adaptive boundary elements methods. Part I. The two-dimensional case”. In: *IMA Journal of Numerical Analysis* 20.2 (2000), pp. 203–234.

[53] Kurt Ferreira, Jon Stearley, James H. Laros III, Ron Oldfield, Kevin Pedretti, Ron Brightwell, Rolf Riesen, Patrick G. Bridges, and Dorian Arnold. “Evaluating the Viability of Process Replication Reliability for Exascale Systems”. In: *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. SC ’11. Seattle, Washington: ACM, 2011, 44:1–44:12. ISBN: 978-1-4503-0771-0. DOI: 10.1145/2063384.2063443.

[54] Andreas Frommer and Daniel B Szyld. “On asynchronous iterations”. In: *Journal of Computational and Applied Mathematics* 123.1 (2000), pp. 201–216.

[55] H. Furstenberg and H. Kesten. “Products of Random Matrices”. In: *The Annals of Mathematical Statistics* 31.2 (June 1960), pp. 457–469. DOI: 10.1214/aoms/1177705909.

[56] Marc Gamell, Daniel S. Katz, Hemanth Kolla, Jacqueline Chen, Scott Klasky, and Manish Parashar. “Exploring Automatic, Online Failure Recovery for Scientific Applications at Extreme Scales”. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC ’14. New Orleans, Louisana: IEEE Press, 2014, pp. 895–906. ISBN: 978-1-4799-5500-8. DOI: 10.1109/SC.2014.78.

[57] Paolo Gatto and Jan S Hesthaven. “Numerical approximation of the fractional laplacian via hp-finite elements, with an application to image denoising”. In: *Journal of Scientific Computing* 65.1 (2015), pp. 249–270.

[58] R. K. Getoor. “First passage times for symmetric stable processes in space”. In: *Transactions of the American Mathematical Society* 101.1 (1961), pp. 75–90.

[59] Dominik Göddeke, Mirco Altenbernd, and Dirk Ribbrock. “Fault-tolerant finite-element multigrid algorithms with hierarchically compressed asynchronous checkpointing”. In: *Parallel Computing* 49 (2015), pp. 117–135.

[60] Alexander A. Golovin, Bernard J. Matkowsky, and Vladimir A. Volpert. “Turing pattern formation in the Brusselator model with superdiffusion”. In: *SIAM Journal on Applied Mathematics* 69.1 (2008), pp. 251–272.

[61] Gerd Grubb. “Fractional Laplacians on domains, a development of Hörmander’s theory of μ-transmission pseudodifferential operators”. In: *Advances in Mathematics* 268 (2015), pp. 478–528.

[62] Wolfgang Hackbusch. *Iterative solution of large sparse systems of equations*. Vol. 95. Applied Mathematical Sciences. New York: Springer-Verlag, 1994, pp. xxii+429. ISBN: 0-387-94064-2. DOI: 10.1007/978-1-4612-4288-8.

[63] Wolfgang Hackbusch. *Multi-grid methods and applications*. Vol. 4. Springer-Verlag Berlin, 1985. DOI: 10.1007/978-3-662-02427-0.

[64] Thomas Herault and Yves Robert. *Fault-Tolerance Techniques for High-Performance Computing*. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-20943-2.

[65] Mark Hoemmen and Michael A Heroux. "Fault-tolerant iterative methods via selective reliability". In: *Proceedings of the 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC). IEEE Computer Society*. Vol. 3. 2011, p. 9.

[66] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University press, 2012.

[67] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis*. Cambridge University press, 1991. DOI: 10.1017/cbo9780511840371.

[68] Kuang-Hua Huang and Jacob Abraham. "Algorithm-based fault tolerance for matrix operations". In: *Computers, IEEE Transactions on* 100.6 (1984), pp. 518–528.

[69] Markus Huber, Björn Gmeiner, Ulrich Rüde, and Barbara Wohlmuth. "Resilience for Multigrid Software at the Extreme Scale". In: *arXiv preprint arXiv:1506.06185* (2015).

[70] Luc Jaulmes, Marc Casas, Miquel Moretó, Eduard Ayguadé, Jesús Labarta, and Mateo Valero. "Exploiting Asynchrony from Exact Forward Recovery for DUE in Iterative Solvers". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC '15.

Austin, Texas: ACM, 2015, 53:1–53:12. ISBN: 978-1-4503-3723-6. DOI: 10.1145/2807591.2807599.

[71] George Karypis and Vipin Kumar. "A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs". In: *SIAM Journal on Scientific Computing* 20.1 (1998), pp. 359–392. DOI: 10.1137/S1064827595287997.

[72] Toshiyuki Koto. "IMEX Runge–Kutta schemes for reaction–diffusion equations". In: *Journal of Computational and Applied Mathematics* 215.1 (2008), pp. 182–195.

[73] Tim Kröger and Tobias Preusser. "Stability of the 8-tetrahedra shortest-interior-edge partitioning method". In: *Numerische Mathematik* 109.3 (2008), pp. 435–457. ISSN: 0945-3245. DOI: 10.1007/s00211-008-0148-8.

[74] Shuangzhe Liu and Gõtz Trenkler. "Hadamard, Khatri-Rao, Kronecker and other matrix products". In: *Int. J. Inform. Syst. Sci* 4.1 (2008), pp. 160–177.

[75] Yifei Lou, Xiaoqun Zhang, Stanley Osher, and Andrea Bertozzi. "Image recovery via nonlocal operators". In: *Journal of Scientific Computing* 42.2 (2010), pp. 185–197.

[76] Frédéric Magoulès, Daniel B. Szyld, and Cédric Venet. "Asynchronous optimized Schwarz methods with and without overlap". In: *Numerische Mathematik* (2017), pp. 1–29. ISSN: 0945-3245. DOI: 10.1007/s00211-017-0872-z.

[77] B. M. McCay and M. N. L. Narasimhan. "Theory of nonlocal electromagnetic fluids". In: *Archives of Mechanics* 33.3 (1981), pp. 365–384.

[78] William Charles Hector McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2000.

[79] Mark M. Meerschaert and Alla Sikorskii. *Stochastic models for fractional calculus*. Vol. 43. de Gruyter Studies in Mathematics. Walter de Gruyter & Co., Berlin, 2012, pp. x+291. ISBN: 978-3-11-025869-1.

[80] Ricardo H Nochetto, Enrique Otárola, and Abner J Salgado. “A PDE approach to fractional diffusion in general domains: a priori error analysis”. In: *Foundations of Computational Mathematics* 15.3 (2015), pp. 733–791.

[81] Piyush Sao and Richard Vuduc. “Self-stabilizing Iterative Solvers”. In: *Proceedings of the Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems*. ScalA ’13. Denver, Colorado: ACM, 2013, 4:1–4:8. ISBN: 978-1-4503-2508-0. DOI: 10.1145/2530268.2530272.

[82] Stefan A. Sauter and Christoph Schwab. “Boundary element methods”. In: *Boundary Element Methods*. Springer, 2010, pp. 183–287.

[83] L Ridgway Scott and Shangyou Zhang. “Finite element interpolation of non-smooth functions satisfying boundary conditions”. In: *Mathematics of Computation* 54.190 (1990), pp. 483–493.

[84] Raffaella Servadei and Enrico Valdinoci. “On the spectrum of two different fractional operators”. In: *Proceedings of the Royal Society of Edinburgh: Section A Mathematics* 144.04 (2014), pp. 831–855.

[85] Stewart A. Silling. “Reformulation of elasticity theory for discontinuities and long-range forces”. In: *Journal of the Mechanics and Physics of Solids* 48.1 (2000), pp. 175–209.

[86] Ian H. Sloan. “Error analysis of boundary integral methods”. In: *Acta Numerica* 1 (1992), pp. 287–339.

[87] Ian H. Sloan and Alastair Spence. “The Galerkin method for integral equations of the first kind with logarithmic kernel: Theory”. In: *IMA Journal of Numerical Analysis* 8.1 (1988), pp. 105–122.

[88] Barry Smith, Petter Bjorstad, and William Gropp. *Domain decomposition: parallel multilevel methods for elliptic partial differential equations*. Cambridge university press, 2004.

[89] Marc Snir, Robert W Wisniewski, Jacob A Abraham, Sarita V Adve, Saurabh Bagchi, Pavan Balaji, Jim Belak, Pradip Bose, Franck Cappello, Bill Carlson, et al. "Addressing failures in exascale computing". In: *International Journal of High Performance Computing Applications* 28.2 (2014), pp. 129–173.

[90] Miroslav Stoyanov and Clayton Webster. "Numerical Analysis of Fixed Point Algorithms in the Presence of Hardware Faults". In: *SIAM Journal on Scientific Computing* 37.5 (2015), pp. C532–C553. DOI: 10.1137/140991406.

[91] John C. Strikwerda. "A probabilistic analysis of asynchronous iteration". In: *Linear Algebra Appl.* 349 (2002), pp. 125–154. ISSN: 0024-3795. DOI: 10.1016/S0024-3795(02)00258-6.

[92] Arthur H. Stroud. *Approximate calculation of multiple integrals*. Prentice-Hall, 1971.

[93] Peter Tankov. *Financial modelling with jump processes*. Vol. 2. CRC press, 2003.

[94] Andrea Toselli and Olof Widlund. *Domain decomposition methods: algorithms and theory*. Vol. 3. Springer, 2005.

[95] Ulrich Trottenberg, Cornelius W. Oosterlee, and Anton Schüller. *Multigrid*. San Diego, CA: Academic Press Inc., 2001, pp. xvi+631. ISBN: 0-12-701070-X.

[96] Enrico Valdinoci. "From the long jump random walk to the fractional Laplacian". In: *SeMA Journal: Boletín de la Sociedad Española de Matemática Aplicada* 49 (2009), pp. 33–44.

[97] Jacques Vanneste. "Estimating generalized Lyapunov exponents for products of random matrices". In: *Physical Review E* 81.3 (2010), p. 036701.

[98] Divakar Viswanath. "Random Fibonacci sequences and the number $1.13198824\ldots$". In: *Math. Comp.* 69.231 (2000), pp. 1131–1155. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-99-01145-X.

[99] Eric W. Weisstein. *Meijer G-Function. From MathWorld—A Wolfram Web Resource*. 2017.

[100] Bruce J. West. *Fractional Calculus View of Complexity: Tomorrow's Science*. CRC Press, 2016.

[101] Jinchao Xu. *An introduction to multilevel methods*. 1997.

[102] Jinchao Xu. "Iterative methods by space decomposition and subspace correction". In: *SIAM Rev.* 34.4 (1992), pp. 581–613. ISSN: 0036-1445. DOI: 10.1137/1034116.

[103] Yeli Yan, Ian H. Sloan, et al. *On integral equations of the first kind with logarithmic kernels*. University of NSW, 1988.