Lecture Notes in Applied and Computational Mechanics 66

Thomas Apel
Olaf Steinbach *Editors*

Advanced Finite Element Methods and Applications

Springer

Lecture Notes in Applied and Computational Mechanics 66

Series Editors

For further volumes:
http://www.springer.com/series/4623

Advanced Finite Element Methods and Applications

Thomas Apel and Olaf Steinbach (Eds.)

Springer

Editors

Univ.-Prof. Dr. Thomas Apel
Institut für Mathematik und Bauinformatik
Universität der Bundeswehr München
Neubiberg
Germany

Univ.-Prof. Dr. Olaf Steinbach
Institut für Numerische Mathematik
TU Graz
Graz
Austria

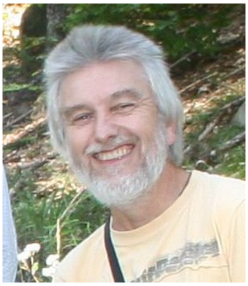
ISSN 1613-7736 e-ISSN 1860-0816
ISBN 978-3-642-30315-9 e-ISBN 978-3-642-30316-6
DOI 10.1007/978-3-642-30316-6
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012939401

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to Ulrich Langer and Arnd Meyer



Preface

This volume on some recent aspects of finite element methods and their applications is dedicated to Ulrich Langer and Arnd Meyer on the occasion of their 60th birthdays in 2012. After their studies of Mathematics at TU Chemnitz (TH Karl–Marx–Stadt) in the seventies, both working first on the numerical analysis of eigenvalue problems, they later became two of the leading scientists in the field of numerical discretization techniques for partial differential equations and the efficient solution of discrete finite element systems.

Ulrich Langer did his PhD with Vadim Korneev at St. Petersburg State University (Leningrad) on finite element schemes and their iterative solution in 1980. Together with Korneev, he published a Teubner monograph on the application of finite element methods in plasticity in 1984. After his habilitation in 1986 be became Associate Professor at TU Chemnitz. Since 1993 he holds a full professorship at the Johannes Kepler University in Linz, Austria. There, he initiated and actively participated in several research programs such as the collaborative research center (SFB) Numerical and Symbolic Scientific Computing and the Johann Radon Institute for Computational and Applied Mathematics (RICAM) of the Austrian Academy of Sciences. The scientific interests of Ulrich Langer are in the numerical analysis of parallel iterative solution strategies for finite and boundary element methods and their applications in solid mechanics and electromagnetism.

Arnd Meyer did his PhD with Wilgard Lang at TU Chemnitz in 1978 on the inclusion of the spectrum and the computation of matrix eigenvalues and eigenvectors. After his habilitation in 1986, again on the numerics of eigenvalue problems, he worked on the efficient implementation of finite element methods in fluid and solid mechanics. Since 1992 he holds a full professorship on Numerical Analysis at TU Chemnitz. Already before the political change in 1989 he started parallel computations on self–built transputer systems. Later this was continued with a research group and the collaborative research center (SFB) Numerical Simulation on Massively Parallel Computers with applications in physics and solid mechanics.

The work of Ulrich Langer and Arnd Meyer always combines the numerical analysis of finite element algorithms, their efficient implementation on state of the art hardware architectures, and the collaboration with engineers and practitioners. In this spirit, this volume contains contributions of former students and collaborators indicating the broad range of their interests in the theory and application of finite element methods.

Topics cover the analysis of domain decomposition and multilevel methods, including *hp* finite elements, hybrid discontinuous Galerkin methods, and the coupling of finite and boundary element methods; the efficient solution of eigenvalue problems related to partial differential equations with applications in electrical engineering and optics; and the solution of direct and inverse field problems in solid mechanics.

We would like to thank all authors for their contributions to this volume. Moreover, we also thank all anonymous referees for their work, their criticism, and their suggestions. These hints were very helpful to improve the contributions. Finally, we would like to thank Dr. T. Ditzinger of Springer Heidelberg for the continuing support and patience while preparing this volume.

Graz, München
March 2012

Thomas Apel
Olaf Steinbach

Contents

A Brief History of the Parallel Dawn in Karl-Marx-Stadt/Chemnitz

Gundolf Haase and Matthias Pester

Abstract. The paper recalls the period 1988-1993 when the research on parallel algorithms and their implementation started in Karl-Marx-Stadt (renamed to Chemnitz in 1990). We consider the research group formed at this time and the hardware available to this group. Parallel hardware as the transputer is considered and the ancient parallel computers from that time are depicted. The group has been formed by the series of workshops and seminars that took place; and the FEM-Symposium is still organized annually. We will focus on a few of these activities and present the developments in hardware, numerical methods, parallel algorithms and analysis that have been discussed between professors, research assistants and students. The paper contains also a brief view on parallel computers available to that group today and some examples document how the computing power has increased during a period of more than 20 years.

1 Introduction

Once upon a time, in the late 1980s of the last century, a handful mathematicians in Karl-Marx-Stadt, East Germany, were dreaming of large computers that are able to handle the mathematical methods they were investigating for solving discretized second order partial differential equations (PDEs) originating from applications. Instead of only dreaming they organized lectures and workshops to train the available

Gundolf Haase
Institut für Mathematik und Wissenschaftliches Rechnen, Karl–Franzens Universität Graz, Heinrichstrasse 36, 8010 Graz, Austria
e-mail: gundolf.haase@uni-graz.at

Matthias Pester
Fakultät für Mathematik, TU Chemnitz, Reichenhainer Strasse 41, 09107 Chemnitz, Germany
e-mail: m.pester@mathematik.tu-chemnitz.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 1–26.
springerlink.com

people with the knowledge on advanced numerical methods for discretization, iterative solution methods, preconditioners and parallelization techniques for the chosen problem class. Many research projects were realized with the sudden availability of parallel hardware and many of the team members went on to become full professors at other institutions.

This paper recalls the eventful period from 1988 to 1993 from the perspective of the authors. We will start in Section 2 with a review of those aspects of hardware and software development that influenced the group. Therefore, we will mainly focus on the transputer and computer systems by Parsytec. Consequently, some hardware of other vendors is not included. The scientific environment in Karl-Marx-Stadt/Chemnitz is illustrated in Section 3 through a review of several workshops that took place. The literature references are restricted to reflect the available information at a time with very limited access to journals, without internet and without the possibility to simply copy a paper or a talk. We finish the paper with some remarks on hardware in 2012 and which concepts survived 25 years of progress in hardware and solution methods.

2 Ancient Hardware and Its Development

In 1986 there was a BBC report [5] on parallel computing. The first part presented an introduction into parallel computing and a Non-Von (Neumann) parallel computer based on small processing units assembled with very small local memory. Eight of these units were planned to be assembled on one chip with a binary tree interconnection on this chip. Besides the binary tree that sounds like a description of a recent GPU. The project leader of this Non-Von supercomputer, David Shaw, made a very interesting statement in that BBC report: “There is one critical problem with that [supercomputers] and that is heat. . . . The latest supercomputers [Cray X-MP/48] really are very small, very highly efficient refrigerators.” Although the ratio of flops per Watt has increased by a factor of 200 000[1,2] during the last 25 years, we still face (and we will continue to face) the same problem in recent supercomputers. By the way, the Cray X-MP had already an SSD (Solid State Device) for very fast memory access.

2.1 Transputer

The whole BBC report [5] is worth watching for historical reasons. In the second part a new company from Bristol called *inmos* is introduced which presents its recent processor - a transputer: “A single piece of silicon with a powerful 32-bit processor, its own memory and the build-in ability to link directly to other transputers.” [5].

[1] Cray-2 in 1986: peak 1.9 Gflops, 200 kW, 0.0095 Mflops/Watt

[2] Blue Gene/Q in 2011: 1680 Mflops/Watt

An *inmos* employee shows a graphical demonstration on two screens with butterflies proceeding from one screen to the other. He briefly introduces the concept of parallelism and concurrency with that graphical demonstration and presents the board with a transputer and 2 MB memory which was a lot of memory in 1986 when a PC had between 64kB and 640 kB memory. The performance of one transputer was 10 MIPS - approximately 10 times faster than an IBM PC and one PC could accommodate 5 of those boards ("50 MIPS I can't believe it"). Such an upgraded PC reached 45 MIPS, i.e., 90 % of its peak performance. The programming languages used Fortran, C, and Occam[3] a concurrent programming language for communicating sequential processes named after William of Ockham[4]. The price for one board was £ 2900 (≈ DM 9200 ≈ € 4600) including one transputer for £ 420, OCCAM and the compilers cost extra. The board presented in that BBC report contained a T400 transputer with only integer arithmetic available. Its successor the T800 issued in 1987 had already a 64-bit floating point unit that achieved 3.6 Mflops and 25 MIPS peak performance.

The artificial word transputer originates from merging the words TRANSmitter and comPUTER. It had the unique feature of combining a processing unit with its own memory and additional communication channels. This was a new concept especially designed for parallel processing [47, 48]. The first transputers of the 16-bit T2 and 32-bit T4 series had only integer arithmetics available which had been extended with a 64-bit floating point unit in 1987 to the T8 series. Several companies combined these T805 (30 MHz, 4.4 Mflops) to parallel computers, e.g., Meiko and Parsytec, until the early 1990s.

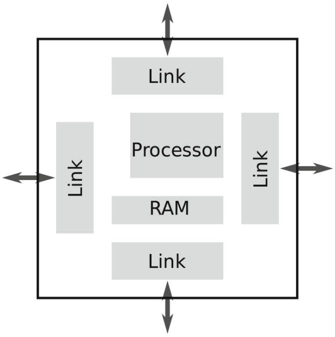


Fig. 1 Block draft of a transputer with its 4 links allowing a direct interconnect to 4 transputers of the same type.

The next transputer generation T9000, also preliminarily called H1, with an improved communication structure [46] and a peak performance of 20 Mflops [45] was released too late to be competitive with other CPUs. Parsytec switched to parallel systems with PowerPC$^{\text{TM}}$ 601 processors using transputers as communication chips.

[3] http://en.wikipedia.org/wiki/Occam_(programming_language)

[4] http://en.wikipedia.org/wiki/William_of_Ockham

2.2 Available Parallel Computers

2.2.1 Transputer Boards

We had first access to the transputer boards in 1989 that have been assembled by those colleagues from the university that later founded the Chemnitz branch of Parsytec. We refer the reader to their description of the transputer [8, 9] and the excellent book on parallelization [10]. There was one board with one transputer T805 (30 MHz, 4.4 Mflops) for code development (Fortran and Occam), and another board with four transputers T805 for real parallel computing, see Fig. 2.

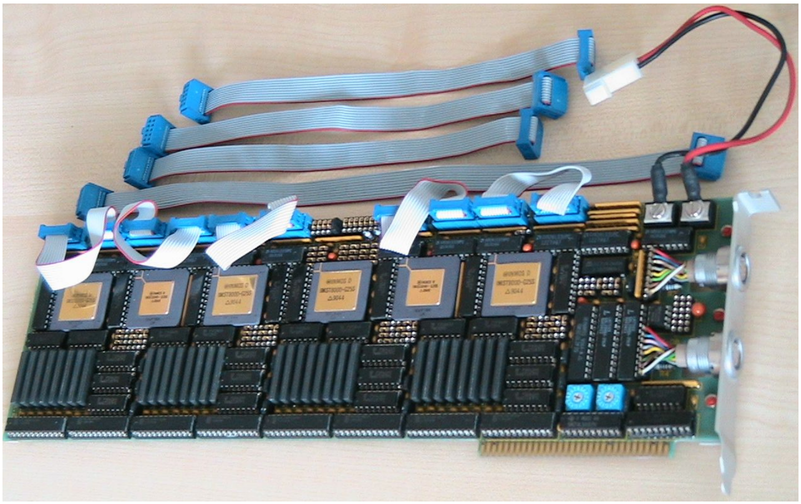


Fig. 2 Board with 4 transputers T800 and 2 link adapters (1989), hardwired links may connect to other boards. This was used by the mathematicians in Chemnitz in the early 90ies.

Both precious boards have been hidden at the Institute for Mechanics behind two locked doors, with special permission required to work with them. The happy people that had to share access to them were Arnd Meyer, Matthias Pester, Sergej Rjasanow, Gundolf Haase and later on Beate Junghans. The programming language was Fortran with an Occam harness supplying the primitive interprocessor communication. We learned a lot about parallel computer topologies, especially the hypercube, and how to set them up by ourselves with the available communication primitives from Occam. It is unimaginable nowadays but each of the four links of the transputer had to be addressed explicitly for communication to other transputers. A special configuration language had to be used for mapping software channels to hardware links.

2.2.2 MultiCluster-2

Thanks to the successful DFG project “Gebietszerlegungsmethoden für Finite Elemente und Randelemente” [Domain decomposition methods for finite and boundary elements] (1991-1993) the first commercial parallel computer in Chemnitz, the **MultiCluster-2** by Parsytec was delivered to the working group in 1991. During the same period (1991-1995) the Chemnitz group joined the DFG-Schwerpunktprogramm [DFG Priority Program] “Randelementmethoden” [Boundary element methods] with the subproject “Parallele Lösungsstrategien für Finite-Elemente- und Randintegralgleichungen” [Parallel solution strategies for finite and boundary element methods] [29].

The MultiCluster consisted of 16 boards with 2 T805 transputers running at 30 MHz placed on each of them and 8 MB memory per transputer. The interconnection network was fully reconfigurable by software and could be physically partitioned. This parallel computer had its own operating system Helios that started a micro-unix on each transputer to ensure basic services such as launching, spreading and terminating processes as well as communication between them. It needed an additional host, a SUN workstation to compile, link the code and launch the parallel application to the parallel computer. Also all IO operations had to be handled via the host. The MultiCluster-2 allowed us already to start more parallel processes than the physical processors available (not exceeding the 8 MB memory per processor).

In contrast to other parallel computers like MasPar with its SIMD programming model (Single Instruction Multiple Data) the MultiCluster and Helios supported the much more flexible MIMD model (Multiple Instructions Multiple Data) found in all parallel computers nowadays. As a consequence, Helios allowed us to start P different codes as one parallel application in which processes communicate with each other. Therefore, launching one code onto P processes required sending that code P times from the host to the parallel computers. That made testing (or even debugging) with 32 or even 256 processes on the MultiCluster-2 quite time consuming, emulators or MPI became available much later. An update to the parallel operating system Parix was already a relief to the user because it restricted the MIMD approach to an SPMD model (Single Program Multiple Data) and it transferred the code only once from host to parallel computer and launched the P parallel processes internally. Additionally much better compilers and pre-configured topologies like hypercube, tree and grid were available.

When Ulrich Langer became a full professor in Linz he was granted the permission of the DFG to take the two positions from the successful DFG project “Gebietszerlegungsmethoden für Finite Elemente und Randelemente” [Adaptive domain decomposition methods for finite and boundary elements] (1993-1995) together with the parallel computer (including the host *mephisto* and the two terminals *faust* and *gretchen*) with him to Austria in 1994.

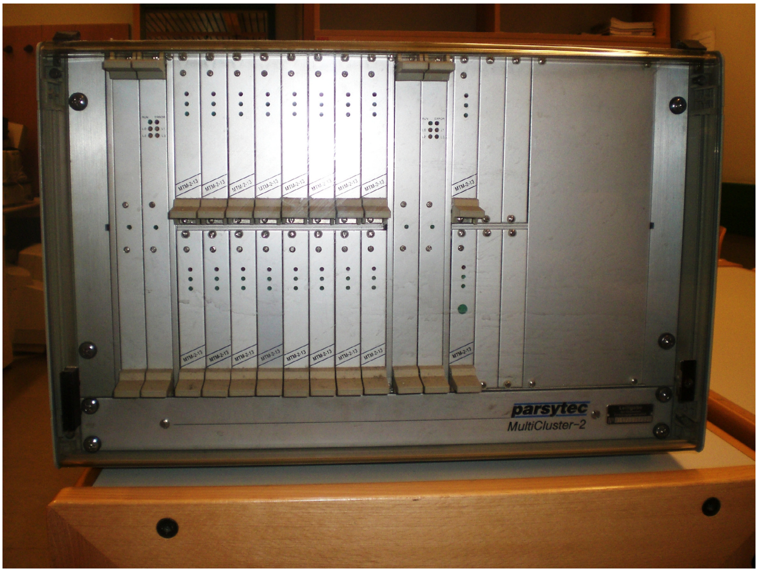


Fig. 3 The MultiCluster-2 that made it through the customs in 1994.

The very valuable parallel computer was transported by Ulrich Langer and Gundolf Haase using their private cars at a time when Austria still was outside the European Community. Besides a pile of paperwork for exporting the equipment from Germany and a careful drive, no further problems had been expected. Reaching the border at Suben the German customs officer told GH that the lane for goods had to be used and he explained how to turn to that lane. Because of the strong Bavarian dialect GH misinterpreted him and assumed that they should drive to the next freeway exit in Austria and return. They broke through the border to the astonishment of the officers. Returning from Austria a few minutes later they passed the border and turned at the next German freeway exit in order to be in the right lane at the border. Unfortunately, the Austrian customs required the Austrian VAT for the computers in cash from the two which were not willing to pay the 20% of the original price. They stayed overnight in Passau and made calls to Linz without the convenience of mobile phones. The colleagues in Linz found a passage in the law the next morning explaining which institutions don't have to pay that tax for certain import goods ["ausgenommen Gebietskörperschaften, dazu zählen . . . Universitäten"]. UL presented the updated version of the law to the Austrian authorities and, after convincing them that he is really an Austrian professor, the customs officer was very helpful. This allowed us to reach Linz at the coffee break and to celebrate the arrival with a sip of champagne.

2.2.3 PowerXplorer

The MultiCluster-2 transfered to Linz wasn't really fast enough to be competitive anymore. The new parallel computer was again delivered by Parsytec Chemnitz in 1994 and it contained the competitive PowerPC^{TM} processors which demoted the transputers in the system to pure communication chips. Each of the four cabinets contained 4 PowerPC^{TM} 601 64-bit processors with 80 MHz clock rate, 32 kByte internal cache and 32 MByte of memory for each processor. One processor reached 80 Mflops (double precision) such that the overall peak performance added to 1.28 Gflops on half a GByte memory. The PowerPC^{TM} has been derived from the RS/6000 chip-set and its great-grandsons (and -daughters) are still found in nowadays supercomputers by IBM.

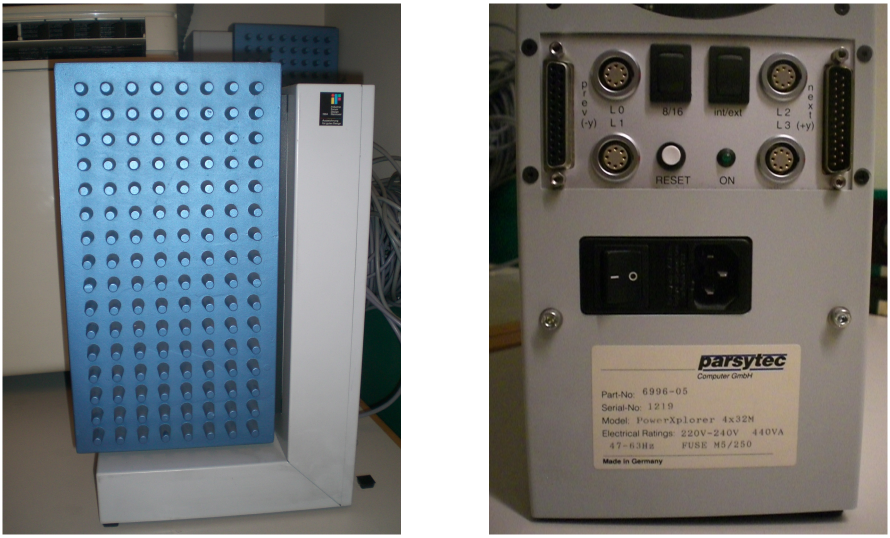


Fig. 4 One cabinet of the PowerXplorer [left] and its back side with plug-ins for the 4 links.

The cabinets had to be connected via special link cables plugged into links 0-3. This enabled the user to configure a hard wired hypercube but also other topologies as a mesh were possible. The PowerXplorer was derived from Parsytec's GC family with up to 1024 processors (up to 16.384 were planned with the GC-5) and the same operating system Parix was running on both of them. Therefore, code could be developed on smaller systems as in Graz and the production code ran on the larger parallel computer in Chemnitz. According to commercial material by Parsytec, the abbreviation GC means Grand Challenges as well as Giga Computer as well as GigaCluster.

When discussing new parallel computers for Graz one colleague from the computer science department remarked that the PowerPC$^{\text{TM}}$ processors were too fast for the interconnection network between the cabinets. Offering him the old MultiCluster-2 with a much better balance didn't satisfy him much. It turned out that this imbalance wasn't so important for the PDE solvers on the PowerXplorer, but nowadays we face that problem for clusters of GPUs.

The experience with parallel computers in Linz lead to the Special Research Program (SFB) on "Numerical and Symbolic Scientific Computing" (1998-2008) incorporating scientists from mathematics, computer science and engineering. Thanks to the funding agency FWF, the university and the local government in Upper Austria the university of Linz installed the largest Austrian computer in 1999. This parallel computer and all succeeding ones were delivered by SGI. The SFB is continued in terms of the funded graduate school on "Computational Mathematics: Numerical Analysis and Symbolic Computation" that runs since 2008. Additionally the Johann Radon Institute for Computational and Applied Mathematics (RICAM) were created in 2003. This research from Linz spread further when the first author left for Graz in 2004 and he participates in the SFB "Mathematical Optimization and Applications in Biomedical Sciences" in Graz since 2007.

2.2.4 Chemnitz as Center of Parallel Computing

Transputer boards and MultiCluster were only *small steps* followed by more or less *giant leaps* promoting the Chemnitz University to an internationally respected center of parallel computing and parallel computers. The MultiCluster-2 was an initial equipment for the research group "Scientific Parallel Computing" (SPC). The group consisted of mathematicians and mechanical engineers and was funded by the DFG between 1993 and 1995. Mainly initiated by Ulrich Langer and Arnd Meyer, this research group quickly found interested partners in other departments of the university, especially computer scientists, physicists, and mechanical engineers. Together they could convince the administration of the university (and the DFG) that the scientific potential in this field required and justified the acquisition of a parallel supercomputer.

There was also an interesting period of time before the new supercomputer was ordered. The most notable manufacturers that were able or believed to be able to offer the best parallel supercomputer worldwide found the way to Chemnitz, such as nCube, Kendall Square Research, MasPar, Thinking Machines, DSM Computer Systems, CRAY Research, Siemens Nixdorf, intel, Parsytec, GFTT, ibt computer, IBM, Hewlett Packard, Trust Computer, Silicon Graphics, ...

We heard many details about different hardware concepts and saw a lot of colorful slides. Casually, a nice business lunch should help to discuss more details. Almost each vendor granted the opportunity to access a test environment, either in Chemnitz or in another place where such a machine was available or just promoted in an exhibition, even once in the hotel *Occam* in Munich's *Occamstraße* – very matching.

All including this yielded to a lot of comparisons to be evaluated and analyzed. Finally, after the solicitation a decision had to be made due to the price-performance ratio that satisfied both the scientists and the administration.

In 1994, the first parallel supercomputer of the TU Chemnitz was installed, a GC PowerPlus with 128 processors. The name "GC" was mostly interpreted as "Grand Challenge" or "Giga-Cube".

The computing power of the GC PowerPlus (10 Gflops peak) was placed behind an attractive casing box, a tower of 2 by 4 modules (blue cubes, Fig. 5). Such a module contained 8 boards, each of them providing 2 PowerPC$^{\mathrm{TM}}$-601 processors, 32 MB of RAM to be shared between the 2 processors, and 4 transputers as helpers for communication.

The underlying topology for the communication was a 2D grid, but the operating system Parix supported any virtual topology. Thinking about topologies was important because other than in later systems with PVM or MPI, the communication in Parix was primarily controlled by links and not by destination node numbers.

A kind of prototype was available some months before in the computer architecture group of the Department of Computer Science. This GCel-192 (el = entry level) was a small tower of 3 modules, each of them housing 64 transputer nodes. The GCel was put at our research group's disposal and first tests with more than 16 or 32 processors were possible. The exterior view of the GCel was similar to that of the GCPP (Fig. 5), but the performance of its interior equipment exceeded that of a MultiCluster only by the quantity of 192 against 32 processors, whereas the performance of communication could be seen as a notable bottleneck. This bottleneck was a motivation to reconsider former simple strategies that were appropriate for up-to 16 processors [1] and improve them with respect to the communication effort.

By four years a renewal of the GCPP was needed in order to keep the hardware base for parallel computing up-to-date. Meanwhile, the research group was replaced by the *"Sonderforschungsbereich"* (SFB 393) on "Parallel numerical simulation for physics and continuum mechanics" that was funded by the DFG over 3 periods from 1996 to 2005 [36].

While the conceptual discussion for a new cluster was in its final stage, a cable fire destroyed one module of the GCPP irreparably († 1999). Fortunately, there was a couple of smaller systems (Power Xplorer, workstation cluster) available to continue the research at a certain level before the new acquisition had been finished in 2000.

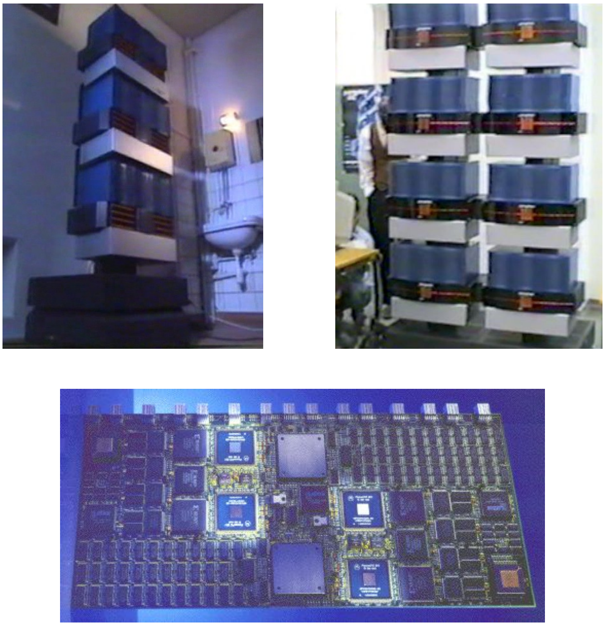


Fig. 5 The GCel-192 in an office of the computer architecture group, the GCPP-128, and one of the 64 single GCPP-boards. The upper left image illustrates that the product has found its (market) niche.

The next generation parallel supercomputer was the Chemnitz Linux Cluster (CLiC) which made headlines as Europe's fastest "self-made cluster computer" (and the 2nd of this category worldwide in the Top500 list). CLiC was a cluster of 528 merchantable PCs with Pentium III processors (800 MHz) and 512 MB RAM, see Fig. 6. The network was *only* based on fast ethernet (a question of price-performance ratio in combination with the real needs). With two network adapters per node the data flow was split into a management network for service access and an internal communication network. All nodes were connected via a high-performance switch (Extreme Network Black Diamond). The user access to the system was managed by an adapted Portable Batch System (OpenPBS).

The supplier for the slightly modified standard PCs was the *Megware* company domiciled in Chemnitz. Together with the technicians of the computer center of the university they established their first big cluster system. One can say, they gained so much experience with CLiC that this company has become a well-known provider of cluster systems afterwards. Thus, also on the Chemnitz High-Performance Linux Cluster (CHiC) the Megware logo appears next to the IBM logo (Fig. 6). Megware is still very successful in business in 2011 - besides the large installations in Vienna and Munich they nearly got the order for the new GPU server in Graz.

Fig. 6 The Linux clusters CLiC and CHiC, [source: TU Chemnitz].

With the High-Performance Linpack benchmark (HPL) for the Top500 evaluation this cluster attained a total performance of 221.6 Gflops.[5]

For the next 7 years CLiC had been intensively used by many groups and guest scientists of the TU Chemnitz but the scientists already thought about and discussed an extension or successor. The final shutoff for CLiC was in summer 2007 when again a next generation system, the CHiC, had already been in operation for a few months.

With 530 compute nodes containing 2 dual-core Opteron processors (at 2.6 GHz) and 4 GB RAM[6] CHiC made it to rank 117 of the Top500 in June 2007 gaining 8.2 Tflops sustained. This was only feasible, because the nodes are connected via a high-performance InfiniBand network, in addition to the Gigabit ethernet for management. The main characteristic values for the progress of the parallel computer systems in Chemnitz are summarized in Table 1.

Table 1 The main steps of high-performance parallel computers at TU Chemnitz in numbers.

				performance		best rank
year	system	#cores	total memory	peak	sustained	in Top500
1994	GCPP	128	2 GB	10 Gflops	5.25 Gflops	178
2000	CLiC	528+2	270 GB	424 Gflops	221.6 Gflops	126
2007	CHiC	530×4	>2 TB	11.19 Tflops	8.21 Tflops	117

Taking into account the vast amount of hard disk failures in the nodes of CLiC, the diskless nodes of CHiC were much less susceptible to faults in the long term. The cluster is provided with a 60 TB parallel storage file system (Lustre) whereon each node can access. A few extra nodes for special purposes complete the CHiC.

As on CLiC, the access to CHiC is controlled by a job queuing system similar to OpenPBS (TORQUE/Maui), where in both cases the *batch system* does not exclude

[5] By the way: using a 529th node (one of the service nodes) to have a 23×23 processor grid.

[6] Some of the nodes were upgraded to 8 or 16 GB later.

interactive jobs. A special feature for the users of CHiC is an easy way to test with different software environments having a modular concept to setup the appropriate environment, i.e. search paths and other environment variables. This way the user may choose among different compiler versions, MPI installations, and basic library versions for BLAS, BLACS, SCALAPACK, giving a simple ‘module’ command.

2.3 Some Remarks on Performance

Let us consider our old benchmark problem for testing the sequential performance of a processor. We solve the Poisson equation in the 2D domain $\Omega = (0,1)^2$,

$$-\Delta u(x) = 1 \quad \forall x \in \Omega, \quad u(x) = 0 \quad \forall x \in \Gamma = \partial\Omega. \tag{1}$$

The operator is discretized with an equidistant 5-point-stencil finite difference discretization. It takes our geometrical Multigrid V-cycle (2 Gauß-Seidel forward sweeps and 2 Gauß-Seidel backward sweeps as pre- and post-smoother, linear interpolation and linear prolongation) only 4 iterations to solve the problem above with initial guess $u = 0$ until a relative accuracy of $\varepsilon = 10^{-4}$ is reached. The number of iterations is independent of the number of grid cells, i.e., independent from the degrees of freedom.

Table 2 Solution time in seconds for solving the Poisson equation on one processor/core. Parix compiler ACE C or gcc with option -O3 have been used.

	multigrid levels	9 grids	14 grids	15 grids
	unknowns	$261,121$	$268 \cdot 10^6$	$1 \cdot 10^9$
processor	year	7.4 MB	14 GB	51 GB
T805 (30 MHz), 8 MB, Parix 1.0	1990	143.00		
T805 (30 MHz), 8 MB, Parix 1.2	1993	98.00		
i486DX (33 MHz), 8 MB, Linux	1993	41.00		
Xplorer-M601 (32 MB), 32 MB, Parix 1.2	1995	4.60		
Pentium (133 MHz), 16 MB, DOS 6.2	1996	12.80		
Pentium-II (350 MHz), 128 MB, Linux	1999	1.25		
Dual Xeon (2.4GHz), 4 GB, Linux	2002	0.23		
core i7-2600K (3.4 GHz), 16 GB, Ubuntu 10.04	2011	0.05	49	
Xeon X5660 (2.8 GHz), 96 GB, Ubuntu 11.04	2011	0.05	58	234

The first timing for the T805 in Tab. 2 has been presented already by G. Haase at Parsytec’s stand at the Systec 1990 in Munich. Comparing the solution times in 1990 and in 2011 we notice an acceleration factor of 3000 over 21 years, i.e., a factor of 1.5 per year, less than Moore’s law. The memory for a desktop PC increased by a factor of 2000. The performance boost by 2002 was a factor of 620, i.e., 1.7, per year which is slightly better than Moore's law. The permanent increase of the clock rate did not apply because the power consumption grows with the third power of the clock rate. Besides shrinking the structures on the chip (allowing lower voltage) the only way for further improvements consists in mounting multiple cores on one consumer chip (Intel: 8, AMD:12) and in constructing many-core chips. The many-core approach for scientific computing is applied by IBM's Blue Gene/Q, Fujitsu's SPARC64 VIIIfx and Nvidia's fermi architecture in 2011.

It is noteworthy that Intel presented the first 1 Tflops (single precision) single chip processor in 2007 which has been benchmarked with equation (1) using the same discretization but only Jacobi iterations for solving the resulting system of equations. Its eighty x86-cores had 4 links to their neighbors similar to the transputer in Fig. 1 and they have been arranged in an 8×10 grid. This concept is available in Intel's SCC (Single Chip Cloud-Computer). A further result of this experimental development is Intel's Knights Ferry delivered in 2010 as a co-processor card with 32 cores and 2GB GDDR5 RAM which achieves 515 Gflops in single precision performance. The cores are connected via a ring topology. Due to the apparent competition to Nvidia's high-performance GPUs, Intel attempts to establish the abbreviation MIC (many integrated core architecture) for that design.

3 The Scientific Environment

The working group for numerical analysis in Karl-Marx-Stadt has been very active with close connections to other numerical groups in East Germany (Magdeburg, Berlin, Greifswald, Ilmenau), the Soviet Union, Czechoslovakia and Bulgaria. Some scientific contacts were already established by Ulrich Langer even in countries west of the iron curtain. The research interests were already focused on analysis and numerics for second order partial differential equations starting with the discretization of the PDE by finite elements, approximation properties of the discrete solution and the efficient numerical solution of the resulting system of equations. The research was performed on the best computers available with state-of-the-art iterative methods, especially multigrid.

When the first author joined that group in 1988 most of the members had just recently received their PhD or were about to receive it. A series of very productive seminars and workshops was established. Besides the regular research seminars at the institute there have been Summer Schools in Breitenbrunn, the annual Multigrid Seminars and the annual FEM Symposium. We would like to mention a few of these activities in more detail. These talks and reports on talks had a significant impact on the young scientists attending these workshops. At this time only black board presentations were given and a lecturer usually presented several 90 minutes lectures without any hand-outs.

3.1 Summer School 1988

Probably few in East Germany have seen the BBC report on the transputer [5] mentioned on page 2. Nevertheless, the rumor about this fancy technology reached the scientists in Chemnitz. An article in the (single) East German computer journal [47] was the first official notice the first author received about the transputer and the following Breitenbrunn Summer School organized by the team in 1988 was partially dedicated to parallel algorithms, hardware and software. So, he had the pleasure to hear about the following topics during the first week of his affiliation at the TU Karl-Marx-Stadt:

- Ulrich Langer talked about classical super element techniques and recent domain decomposition methods on new computer architectures. That included Schur complement preconditioners by Bramble, Pasciak and Schatz [6, 7], multigrid preconditioning and the realization of these algorithms on MIMD computers. This presentation was the basis of several papers in the following years [23, 24, 20, 25].
- Alfred Tamme considered theoretically the challenges on projecting algorithms onto parallel computers starting with systolic arrays and via MIMD to supercomputers. He presented also the general construction of a parallel FEM code named PARFES on a MIMD architecture using OCCAM.
- Matthias Pester focused on transputers, for which constructs and data types were available in OCCAM and he presented the code for a dense matrix-vector multiplication on a grid of transputers.
- Ulrich Langer reported on an international conference in Sofia (Aug. 22-27, 1988) about talks by Raytcho Lazarov on superconvergence [15] and local grid refinement [14], a talk by Owe Axelsson on algebraic multilevel preconditioning methods [2, 3], a talk by Maksimilian Dryja on domain decomposition [13] and a talk by Wolfgang Hackbusch on the panel clustering method [32].

A few months later the second author and Sergej Rjasanow had the chance to do real computations on a parallel computer consisting of transputers at the Bulgarian Academy of Sciences. They gave presentations about their experiences afterwards and the first author still remembers the tuning for the best overlapping of communication and computation in the dense matrix-vector product by the second author. This experience also contributed to further results on basic algorithms and on the parallel boundary element methods [41, 42, 43, 44].

3.2 Summer School 1989

Arnd Meyer presented a parallel data decomposition following the paper by Law [39] integrated in a parallel cg implementation using domain decomposition preconditioners [40]. This data decomposition became the basis of all parallel iterative solution algorithms used in Karl-Marx-Stadt/Chemnitz as well as in Linz and Graz later on.

Data on P subdomains are represented in two different ways as *accumulated* and as *distributed* data. Accumulated data representation means that a process is storing the full numerical value of the nodes of its subdomain. Data stored in a distributed way has only the full value on nodes uniquely belonging to one process. On a shared node each process owning that node stores only a fraction of the full numerical value. This definition leads to the conclusion that accumulated and distributed data representation differ only on the shared nodes of local vectors.

A local accumulated vector $\underline{u}_s$ $(s = 1, \dots, P)$ stores a part of the global vector $\underline{u}$, without changing any numerical values, and both are connected through the mapping operation (or coincidence matrix) A_s with

$$\underline{u}_s = A_s \underline{u}, \tag{2}$$

$$A_s^{(i,j)} = \begin{cases} 1 & \text{iff global node } \underline{u}_j \text{ is stored locally at } \underline{u}_{s,i}, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

$$\dim A_s = \#\text{ local nodes} \times \#\text{ global nodes}.$$

The matrix A_s is not stored as a matrix in process s but as a vector $l2g$ (local to global node numbering) with $l2g_i := j$. A local distributed vector holds only a fraction of the data values in its shared nodes. To get the global data values, processes owning a certain node need to sum their values and next neighbor communication is needed. Therefore the global vector $\underline{u}$ can be obtained by applying the transposed mapping operation (A_s^T) to the local vectors $\underline{u}_s$ and adding the results over all processes. This requires communication via the interconnection network,

$$\underline{u} = \sum_{s=1}^{P} A_s^T \underline{u}_s. \tag{4}$$

Thanks to access to the 4-transputer system, this data distribution has been tested immediately with a cg algorithm accelerated by a domain decomposition preconditioner using multigrid as local solver, and the results have been presented first in January 1990 at the Sixth GAMM-Seminar in Kiel [19]. It finally led to a set of libraries for a series of parallel computers and environments like Parsytec machines under Occam and Parix, Kendall Square computers, nCube machines, IBM parallel machines, cluster computers using PVM and MPI, wherein the specialties of hardware and operating system were hidden by an abstract layer based on the hypercube topology and the basic communication routines needed therein [1, 28]. That allowed a very portable code development for the parallel computers available in the early 90-ies. Nowadays, we simply would use MPI (Message Passing Interface) but we have to remark that the first version of MPI was not issued before 1995. The PVM (Parallel Virtual Machine) software was earlier available but lost its importance over the years. Its good features like spawning processes during the calculation have been included in the MPI-2.0 standard and it is still used for special purposes.

The participants of this Summer School still remember discussions between Ulrich Langer and Arnd Meyer about the shape of these coincidence matrices and the shape of local stiffness matrices (K_{CF}): U.L.: "So the matrix is tall." - A.M.: "Nää, its wide!". During the regular Wednesday excursion the two stopped walking and drew matrices into the dust of the trail. Similar activities have also been reported during a train ride when the vaporized window was used as blackboard by them.

Such an engagement convinced us that this data distribution is the best one can obtain. As a consequence, this data decomposition has been used in many papers by scientists from the Chemnitz group during the last two decades [12].

3.3 Summer School 1990

This last summer school before the reunification of Germany contained some political discussions, but mainly mathematical topics were investigated.

- Ulrich Langer presented domain decomposition methods in the context of additive and multiplicative Schwarz methods [20] and introduced hierarchical preconditioners [52] which resulted also in publications afterwards [27, 21],
- Arnd Meyer pushed the topic of iterative solvers to eigenvalue solvers,
- Dieter Bahlmann investigated fast solvers for biharmonic equations [4],
- Ulrich Langer gave a lecture on the method of boundary potentials that finally ended also with a parallel implementation [11].

3.4 Multigrid Seminars

The Multigrid Seminar series was established in 1986 and it ran annually. In contrast to the summer schools, the Multigrid Seminars collected numerical mathematicians (working in PDEs) from East Germany for a one week workshop. Contributions from Karl-Marx-Stadt/Chemnitz in the years 1988-1990 reflect the main topics the group was working on.

- Multigrid preconditioners and their applications by Michael Jung, Ulrich Langer, Arnd Meyer, Werner Queck and Manfred Schneider [38].
- Gerhard Globisch and Ulrich Langer on the use of multigrid preconditioners in a multigrid software package [18] and together with Michael Jung on multigrid methods for interface problems [17].
- Presentations and papers by Bodo Heise on Multigrid-Newton methods and Full-Multigrid-Newton techniques to applications in electromagnetics [33, 34].
- Application of multigrid to mechanical and thermo-mechanical problems by Torsten Steidten [49].

- Domain decomposition methods with numerical tests on the 4-transputer board were presented by Gundolf Haase, Ulrich Langer and Arnd Meyer [22].

The first three contributions from the list above deal with the different aspects of solving discretized non-linear electromagnetic equations in non-trivial domains by means of multigrid, see Fig. 7.

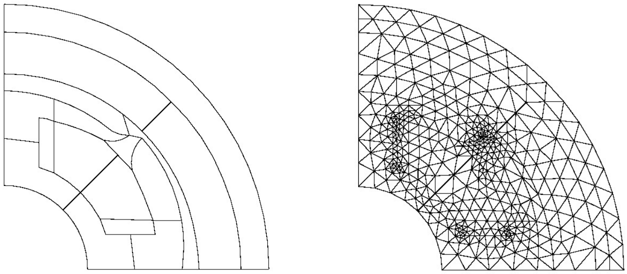


Fig. 7 Geometry with material boundaries [left] and coarse f.e. mesh [right] for one quarter of the ”Chemnitz motor”.

Later, this geometry was handled also in parallel by using domain decomposition methods [21]. The software package FEMGP written in the numerics group contained a mesh generator, a finite element package and a multigrid solver. All components were parallelized in the early 1990s [16, 37, 40].

Werner Queck and Gundolf Haase had the chance to perform numerical tests with one of the first i386 Intel processors (33 MHz) accessible in East Germany at the TH Cottbus in summer 1989. They ported the FEMGP package onto an IBM server with this new 32-bit processor. The AIX operating system was the first contact with a UNIX-like system and it was expensive at this time. So they did a three-day learning-by-doing crash course on UNIX while porting the code from DOS. Fortunately, Linus Thorvald started his LINUX project with the first available version in 1992 and thanks to Thomas Hommel the 32 floppy discs for a full LINUX project were also available in Chemnitz.

When the first author visited the group of Gabriel Wittum at the IWR Heidelberg for the first time in 1991 he was placed in front of a Mac computer that had many new features like a graphical user interface and a mouse with only one button. The time in Heidelberg became very productive after surviving this cultural shock. By the way, nearly all PhD students from that group in Heidelberg occupy full professorships now.

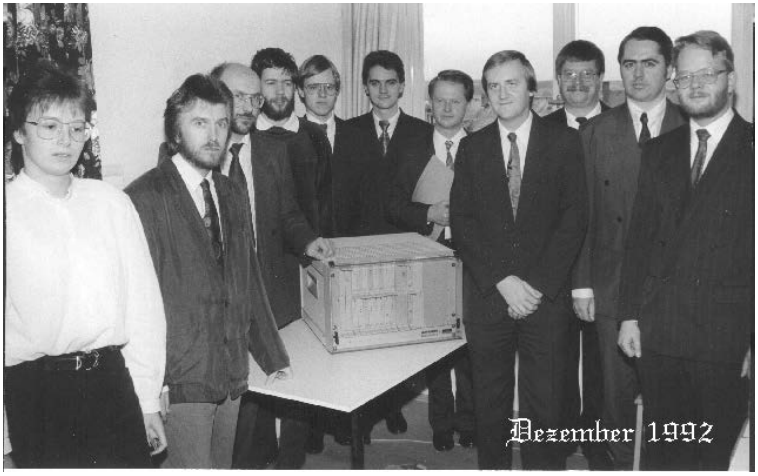


Fig. 8 Numerical analysis group in Chemnitz with the MultiCluster 2, from left to right: Beate Jung, Arnd Meyer, Stefan Meinel, Torsten Steidten, Thomas Hommel, Thomas Apel, Bernd Heinrich, Michael Jung, Ulrich Langer, Matthias Pester, Bodo Heise [Gundolf Haase visited the IWR Heidelberg when the photograph had been taken].

3.5 FEM Symposia

In 1992 the research group SPC ("Scientific Parallel Computing", long title "Algorithmische Grundlagen der Simulation von ausgewählten Problemen der Kontinuumsmechanik auf massiv parallelen Rechnern") was established in Chemnitz as a package of single projects in cooperation between mathematicians and mechanical engineers (Fig. 8). This time and the funds of the DFG were also used to organize annual workshops which are considered as a continuation of the FEM symposia that had taken place already 5 times in Karl-Marx-Stadt between 1978 and 1990. Then the field of parallel computing in FEM and BEM became one of its main topics.

Within a short time cooperation was extended to scientists from other departments with the main focus on parallel computing. For 10 years (1996–2005) this cooperation found its formal framework in the SFB 393 (Sonderforschungsbereich) which was funded by the DFG. Beside the financial support for staff and hardware equipment this was also a good base to make the FEM Symposia to what they are today, a popular tradition for participants from all over the world.

3.6 Historic Numerical Examples

Let us present two examples from the early 1990s published in [26] where parallel hierarchical preconditioners have been applied. Detailed numerical investigations of many additive and multiplicative Schwarz preconditioners can be found in [26].

We considered the Poisson equation in a domain Ω depicted in Fig. 9 with mixed Dirichlet and Neumann boundary conditions. The appropriate results on the MultiCluster-2 can be found in Tab. 3.

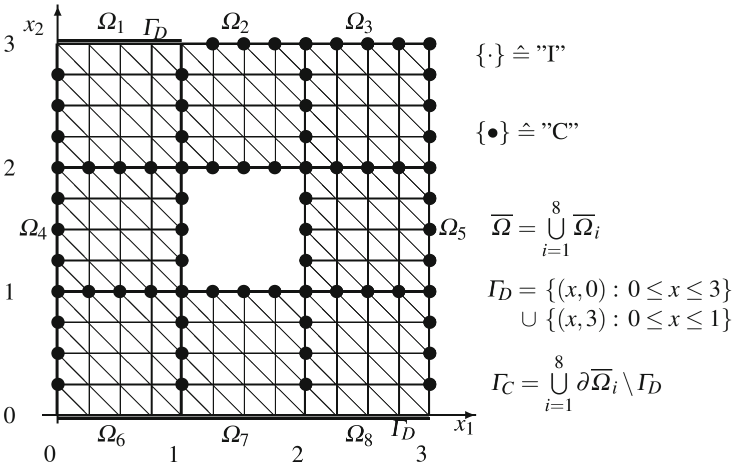

x_2
3
2
1
0
Ω_1
Γ_D
Ω_2
Ω_3
Ω_4
Ω_5
Ω_6
Ω_7
Ω_8
Γ_D
x_1
0
1
2
3
$\{\cdot\} \hat{=}$ "I"
$\{\bullet\} \hat{=}$ "C"
$\overline{\Omega} = \bigcup_{i=1}^{8} \overline{\Omega}_i$
$\Gamma_D = \{(x,0) : 0 \le x \le 3\}$
$\cup \{(x,3) : 0 \le x \le 1\}$
$\Gamma_C = \bigcup_{i=1}^{8} \partial\overline{\Omega}_i \setminus \Gamma_D$

Fig. 9 Domain decomposition and triangulation (level 2 in grid hierarchy) of domain Ω.

Table 3 Results on the MultiCluster-2 with p processors for Poisson equation $-\Delta u = -1$ in Ω (Fig. 9) with mixed boundary conditions.

Levels	number of unknowns	number of iterations	time [in sec.]			
l	N	$(\varepsilon = 10^{-4})$	$p = 2$	$p = 4$	$p = 8$	$p = 16$
2	160	23	3"	3"	3"	3"
3	576	28	4"	3"	3"	3"
4	2175	34	8"	5"	5"	3"
5	8448	39	28"	16"	10"	7"
6	33280	44	–	60"	33"	20"
7	132096	49	–	–	125"	70"

The second example uses the same domain Ω for the plain strain state equations from 2D linear elasticity. See Tab. 4 for the run times.

Table 4 Results on the Multicluster 2 with p processors for linear elasticity in Ω (Fig. 9) with mixed boundary conditions.

Level	number of unknowns	number of iterations	time [in sec.]		
l	N	$(\varepsilon = 10^{-4})$	$p = 8$	$p = 16$	$p = 32$
3	2178	40	6”	4”	4”
4	8450	42	13”	8”	7”
5	33282	45	41”	23”	17”
6	132096	47	–	84”	53”
7	526338	48	–	–	185”

Both examples show that the code ran in parallel and the speedup was good but not extraordinary. From the present-day point of view the number of unknowns was simply too small to achieve a very good speedup. But little else could have been expected from 8 MB memory per processor which is less than the L3-cache on recent server processors.

4 Parallel Computing in 2012

4.1 State of Hardware

Let us jump to the parallel computers that are locally available to the former members of the numerical mathematics group. We restrict ourselves to the locations Chemnitz, Linz and Graz.

The TU Chemnitz inaugurated their **C**hemnitz **Hi**gh Performance Linux **C**luster (CHiC) in 2007 containing 530 2× dual core AMD Opteron, 2.6 GHz and 2 TByte RAM resulting in 8.21 Tflops sustained performance. A 10 Gbit InfiniBand is used as interconnection network. The cluster was delivered from Megware, Chemnitz.

The CHiC project is conducted by a consortium of scientists from multiple departments of the university: the departments of Computer Science, Mathematics, Mechanical Engineering, Natural Sciences, Electrical Engineering and Information Technology, Humanities, and other institutes. Hence, the research fields using CHiC are accordingly manifold. They include high-performance simulation in engineering and natural science, optimization in nontechnical spheres (economic policy, financial management, sports science, psychology, artificial intelligence), the development of algorithms for high-performance computing and visualization techniques for virtual reality.

The management of hardware and software is mainly concentrated in the Department of Computer Science supported by the computer center. The mathematicians deal with adaptive methods for solving PDEs in cooperation with other institutes where the applications arise.

With respect to load balance, adaptive methods required new ideas for parallelization, differing from our *traditional* hypercube topology. The reorganization of distributed data after a few steps of local refinement would be a very hard job. This job is reserved for computer scientists. The more obvious solution is a kind of master–slave treatment where only the most time- and memory-consuming parts, i.e., computation of and on element matrices, are distributed to the slave nodes.

Some of those most recent projects are attached to the Cluster of Excellence eniPROD (Energy-efficient Product and Process Innovations in Production Engineering), e.g. the simulation of fiber reinforced polymers, large deformations for incompressible, elastic materials, or axisymmetric problems with transversely isotropic materials.

The University of Linz got a new parallel computer in October 2011 to which the universities from Linz, Innsbruck and Salzburg contributed financially. This shared memory system with 16 TB memory and 256 octocore Intel E7-8837 (2. 66 GHz, 24 GB L3 cache) has been installed by SGI.

The Institute for Mathematics and Scientific Computing at the University of Graz focuses on numerical algorithms for new hardware and so a GPU server with 5 compute nodes (and again, one host node) was bought in August 2011 in cooperation with the Medical University of Graz. Each compute node consists of 4 Tesla 2070 and two Xeon Westmere X5660 (2.8 GHz, 6 cores, 12 MB L3 Cache), 96 GB CPU memory. The Tesla GPU card is suited for scientific computing, possesses 448 cores (1.15 GHz), 8 GB GDDR5 ECC memory and achieves 515 Gflops double precision peak performance. So in sum the Graz Cluster named *mephisto* achieves nearly 11 Tflops double precision peak performance $(5 \times (2 \times 67.2 + 4 \times 515) = 10972)$, and a 40 GB/s QDR Infiniband is used as interconnect.

4.2 Performance Numbers

The group of Gundolf Haase in Graz is involved in several projects related to parallel computing in application areas. In the CARP project by Gernot Plank fast solvers for potential problems (as subproblem of the bidomain equations) with anisotropic varying coefficients in the domain of a heart are needed. The discretization is based on unstructured tetrahedrons. We use therein our own parallel algebraic multigrid (AMG) solver that solves the large systems of equations in parallel [30].

Let us present some recent numbers on the Cineca cluster (Italy) [IBM Power6, 4.7 GHz; Infiniband x4 DDR] on a problem size of 24,766,309 unknowns.

We would like to emphasize that a sparse linear system with 24 million unknowns and no underlying structure can be solved in 1 second on 256 cores nowadays, i.e., we are 12.000 times faster than the best result in Tab. 3 from 20 years ago.

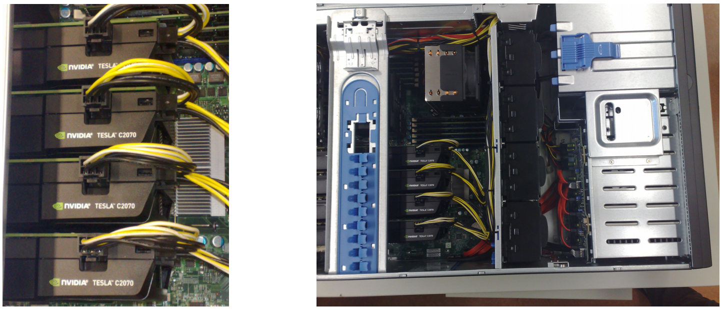


Fig. 10 One compute node of the GPU cluster in Graz with 4 Tesla 2070 and two Xeon Westmere X5660 (2.8 GHz, 6 cores, 12 MB L3 Cache), 96 GB CPU memory. Each Tesla card possesses 448 cores (1.15 GHz), 8 GB GDDR5 memory and achieves 515 Gflops peak performance.

Table 5 Time in seconds on Cineca for one PCCG-AMG (10^{-6}) solve with problem size 24,766,309.

CPU cores	4	8	16	32	64	128	256	512
AMG setup	30.3	15.2	8.0	4.1	2.5	1.5	**1.1**	1.3
solver	64.5	40.6	14.5	6.3	2.9	1.8	**1.2**	1.4

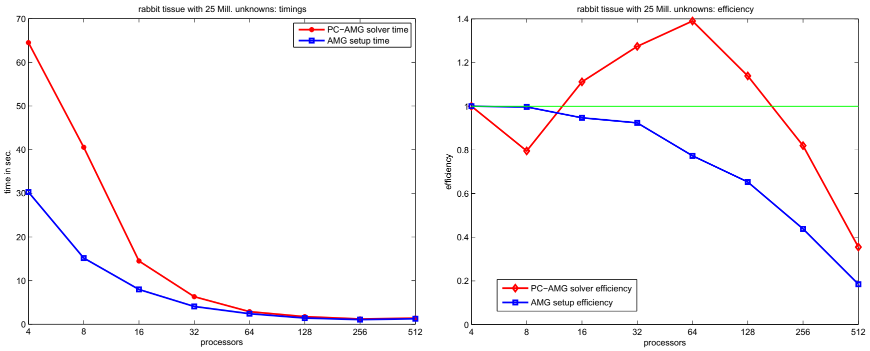


Fig. 11 Solution time and efficiency for AMG on Cineca [Computations performed by Aurel Neic and Manfred Liebmann].

On the other hand there is a high potential for using GPUs with several hundred computational cores. This allowed us to solve a sparse linear system with 2.1 Mill. unknowns (also originating from the CARP project) in 0.2 seconds on 6 GPUs GTX 285.

We observe even superscalar speedup in Fig. 11 thanks to cache effects until 128 cores but then the efficiency drops. This efficiency drop is observed much earlier in case of GPU clusters because the arithmetic performance is much higher in comparison to the available interconnection. The only solution consists in using methods with global multilevel solvers in combination with domain decomposition (DD) smoothers in order to reduce the communication costs in the smoothing sweeps. This closes the circle of numerical methods because now we are forced again to use those domain decomposition methods with which we started more than 20 years ago.

5 Conclusions

The performance of the computers increased by several orders of magnitude during the last 25 years but it still requires a detailed knowledge of numerical methods, domain decomposition methods and insight into specifications of advanced hardware to really achieve a fast and accurate code. Hardware concepts as many-core processing that seemed to be buried in 2000 have been resurrected on a new level as GPUs and again some older methods such as domain decomposition become popular again, because this is the only chance to perform PDE solvers on 10^6 and more processors.

With this background and in an ever changing scientific world, the authors join together with many former students, Ph.D. students and colleagues to thank Ulrich Langer and Arnd Meyer for the great opportunity to join their research group in the late 1980s. The seminars, discussions and supervisions from the past still influence us now.

References

[1] Apel, T., Haase, G., Meyer, A., Pester, M.: Parallel solution of finite element equation systems: efficient inter-processor communication. Preprint-Reihe der Chemnitzer DFG-Forschergruppe “Scientific Parallel Computing” SPC 95-5, TU Chemnitz-Zwickau (1995)

[2] Axelsson, O., Vassilevski, P.: Algebraic multilevel preconditioning methods I. Numer. Math. 56, 157–177 (1989)

[3] Axelsson, O., Vassilevski, P.: Algebraic multilevel preconditioning methods II. SIAM J. Numer. Anal. 27, 1569–1590 (1990)

[4] Bahlmann, D., Langer, U.: A fast solver for the first biharmonic boundary value problem. Numer. Math. 63(1), 297–313 (1992)

[5] BBC: MICRO LIVE television (1986),
`http://www.youtube.com/watch?v=Bn35IbEcEMM`

[6] Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring, I. Math. Comp. 47(175), 103–134 (1986)

[7] Bramble, J.H., Pasciak, J.E., Schatz, A.H.: An iterative method for elliptic problems on regions partitioned into substructures. Math. Comp. 46(173), 361–369 (1986)

[8] Burkhardt, S., Fritzsche, M., Nowak, O.: Der Transputer (Teil 1). Radio Fernsehen Elektronik 38(11), 687–690 (1989)

[9] Burkhardt, S., Fritzsche, M., Nowak, O.: Der Transputer (Teil 2). Radio Fernsehen Elektronik 38(12), 760–763, 808 (1989)

[10] Burkhardt, S., Drey, K.D., Friedrich, V., Nowak, O.: Parallele Rechnertsysteme: Programmierung und Anwendung. Verlag Technik, Berlin/München (1993)

[11] Carstensen, C., Kuhn, M., Langer, U.: Fast parallel solvers for symmetric boundary element domain decomposition equations. Numer. Math. 79(3), 321–347 (1998)

[12] Douglas, C., Haase, G., Langer, U.: A Tutorial on Elliptic PDE Solvers and Their Parallelization. Software, Environments, and Tools. SIAM, Philadelphia (2003)

[13] Dryja, M., Widlund, O.B.: Some domain decomposition algorithms for elliptic problems. In: Hayes, L., Kincaid, D. (eds.) Proceeding of the Conference on Iterative Methods for Large Linear Systems held in Austin, Iterative Methods for Large Linear Systems, October 19-21, pp. 273–291. Academic Press, San Diego (1988)

[14] Ewing, R., Lazarov, R.: Local refinement techniques in the finite element and finite difference methods. In: Numerical Methods and Applications, pp. 148–159. Publishing House of the Bulgarian Acad. Sci. (1989)

[15] Ewing, R., Lazarov, R., Wang, J.: Superconvergence of the velocity along the Gauss lines in the mixed finite element methods. SIAM J. Numer. Anal. 4(28), 1015–1029 (1991)

[16] Globisch, G.: PARMESH - a parallel mesh generator. Parallel Computing 21(3), 509–524 (1995)

[17] Globisch, G., Jung, M.: Mehrgitterverfahren für Interfaceprobleme. In: [35], pp. 60–84 (1990)

[18] Globisch, G., Langer, U.: On the use of multigrid preconditioners in a multigrid software package. In: [50], pp. 105–134 (1990)

[19] Haase, G., Langer, U.: On the use of multigrid preconditioners in the domain decomposition method. In: [31], pp. 101–110 (1991)

[20] Haase, G., Langer, U.: The non-overlapping domain decomposition multiplicative Schwarz method. Inter. J. Computer. Math. 44, 223–242 (1992)

[21] Haase, G., Nepomnyaschikh, S.V.: Explicit extension operators on hierarchical grids. East-West J. Numer. Math. 5(4), 231–248 (1997)

[22] Haase, G., Langer, U., Meyer, A.: A new approach to the Dirichlet domain decomposition method. In: [35], pp. 1–59 (1990)

[23] Haase, G., Langer, U., Meyer, A.: The approximate Dirichlet decomposition method. Part I: An algebraic approach. Computing 47, 137–151 (1991)

[24] Haase, G., Langer, U., Meyer, A.: The approximate Dirichlet decomposition method. Part II: Application to 2nd-order elliptic BVPs. Computing 47, 153–167 (1991)

[25] Haase, G., Langer, U., Meyer, A.: Domain decomposition preconditioners with inexact subdomain solvers. J. Numer. Lin. Alg. Appl. 1, 27–42 (1992)

[26] Haase, G., Langer, U., Meyer, A.: Parallelisierung und Vorkonditionierung des CG-Verfahrens durch Gebietszerlegung. In: Parallele Algorithmen auf Transputersystemen. Tagungsbericht der GAMM-Tagung, Heidelberg, May 31-June 1, 1991, vol. III, Teubner-Scripten zur Numerik, Teubner, Stuttgart (1992)

[27] Haase, G., Langer, U., Meyer, A., Nepomnyaschikh, S.V.: Hierarchical extension operators and local multigrid methods in domain decomposition preconditioners. East-West J. Numer. Math. 2(3), 173–193 (1994)

[28] Haase, G., Hommel, T., Meyer, A., Pester, M.: Bibliotheken zur Entwicklung paralleler Algorithmen. Preprint-Reihe der Chemnitzer DFG-Forschergruppe ”Scientific Parallel Computing” SPC 95-20, TU Chemnitz-Zwickau (1995)

[29] Haase, G., Heise, B., Kuhn, M., Langer, U.: Adaptive domain decomposition methods for finite and boundary element equations. In: Wendland, W.L. (ed.) Boundary Element Topics. Reports from the Final Conference of the Priority Research Programme of the German Research Foundation (DFG), Stuttgart, October 2-4, 1995, pp. 121–148. Springer, Berlin (1997)

[30] Haase, G., Liebmann, M., Douglas, C.C., Plank, G.: A Parallel Algebraic Multigrid Solver on Graphics Processing Units. In: Zhang, W., Chen, Z., Douglas, C.C., Tong, W. (eds.) HPCA 2009. LNCS, vol. 5938, pp. 38–47. Springer, Heidelberg (2010)

[31] Hackbusch, W. (ed.): Parallel Algorithms for Partial Differential Equations. Proceedings of the Sixth GAMM-Seminar, Kiel, January 19-21,1990. Vieweg, Braunschweig (1991)

[32] Hackbusch, W., Nowak, Z.P.: On the fast matrix multiplication in the boundary element method by panel clustering. Numer. Math. 54, 463–491 (1989)

[33] Heise, B.: Multigrid-Newton-methods for the calculation of electromagnetic fields. In: [50], pp. 53–73 (1989)

[34] Heise, B.: Nichtlineare Berechnung stationärer Magnetfelder einer Gleichstrommaschine mittels Full-Multigrid-Newton-Techniken. In: [51], pp. 135–146 (1990)

[35] Hengst, S. (ed.): Proceedings of the ”5-th Multigrid Seminar” held at Eberswalde, May 14-18. R-MATH-09/90. Academy of Sciences, Berlin (1990)

[36] Hoffmann, K.H., Meyer, A. (eds.): Parallel Algorithms and Cluster Computing. Springer, Heidelberg (2006)

[37] Jung, M.: On the parallelization of multi-grid methods using a non-overlapping domain decomposition data structure. Appl. Numer. Math. 23(1), 119–137 (1997)

[38] Jung, M., Langer, U., Meyer, A., Queck, W., Schneider, M.: Multigrid preconditioners and their applications. In: [51], pp. 11–52 (1990)

[39] Law, K.H.: A parallel finite element solution method. Computer and Structures 23(6), 845–858 (1989)

[40] Meyer, A.: A parallel preconditioned conjugate gradient method using domain decomposition and inexact solvers on each subdomain. Computing 45, 217–234 (1990)

[41] Pester, M.: Implementation und Test paralleler Basisalgorithmen der linearen Algebra. In: Grebe, R., Ziemann, C. (eds.) Parallele Datenverarbeitung mit dem Transputer. 2. Transputer–Anwender–Treffen TAT 1990. Proceedings X. Informatik–Fachberichte, vol. 272, pp. 111–118. Springer (1991)

[42] Pester, M.: Parallele Implementation der iterativen Auflösung von Randelementgleichungen. In: 3. Transputer–Anwender–Treffen TAT 1991. Proceedings, RWTH Aachen and Parsytec GmbH Aachen (1991)

[43] Pester, M., Rjasanow, S.: A parallel version of the preconditioned conjugate gradient method for boundary element equations. J. Numer. Lin. Alg. Appl. 2(1), 1–16 (1995)

[44] Pester, M., Rjasanow, S.: A parallel preconditioned iterative realization of the panel method in 3D. J. Numer. Lin. Alg. Appl. 3(1), 65–80 (1996)

[45] Pichlik, H.: Transputer-(R)evolution H1. c't 1, 62–69 (1991)

[46] Pountain, D.: H1 - neue Transputergeneration. c't 7, 34–43 (1990)

[47] Sattelkau, M.: Transputer. Mikroprozessortechnik 2(5), 131–132 (1988)

[48] Schlechter, J.: Innovative Computerarchitektur - Transputer. Mikroprozessortechnik 3(1), 11–12 (1989)

[49] Steidten, T.: Application of multigrid methods to mechanical and thermo-mechanical problems. In: [35], pp. 85–96 (1990)

[50] Telschow, G. (ed.): Proceedings of the ”3-rd Multigrid Seminar” held at Biesenthal, May 2-6, 1988, R-MATH-03/89, Academy of Sciences, Berlin (1989)

[51] Telschow, G. (ed.): Proceedings of the ”4-th Multigrid Seminar” held at Unterwirrbach, May 2-6, 1989, R-MATH-03/90. Academy of Sciences, Berlin (1990)

[52] Yserentant, H.: Hierarchical bases give conjugate gradient type methods a multigrid speed of convergence. Appl. Math. Comp. 19, 347–358 (1986)

Domain Decomposition Preconditioning for High Order Hybrid Discontinuous Galerkin Methods on Tetrahedral Meshes

Joachim Schöberl and Christoph Lehrenfeld

Abstract. Hybrid discontinuous Galerkin methods are popular discretization methods in applications from fluid dynamics and many others. Often large scale linear systems arising from elliptic operators have to be solved. We show that standard *p*-version domain decomposition techniques can be applied, but we have to develop new technical tools to prove poly-logarithmic condition number estimates, in particular on tetrahedral meshes.

1 Introduction

In this paper we are concerned with discontinuous Galerkin (DG) finite element methods for elliptic problems [4, 12, 24]. The motivation might be to have dominant convection, or one wants to build exactly divergence free finite element spaces for incompressible flows [11, 31], or other. We think of operator splitting methods, where one has to solve a large scale symmetric matrix equation in each time-step.

In recent years hybridization methods appeared, which allow to reduce the discrete system to the element interfaces [10]. This paper is concerned with the construction and analysis of domain decomposition methods for the Hybrid Discontinuous Galerkin (HDG) method. We consider one element as sub-domain, and the coarse grid problem consists of mean values on element interfaces. We prove

Joachim Schöberl
Institut für Analysis und Scientific Computing, TU Wien,
Wiedner Hauptstrasse 8–10, 1040 Wien, Austria
e-mail: joachim.schoeberl@tuwien.ac.at

Christoph Lehrenfeld
Institut für Geometrie und Praktische Mathematik, RWTH Aachen,
Templergraben 55, 52056 Aachen, Germany
e-mail: lehrenfeld@igpm.rwth-aachen.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 27–56.
springerlink.com

robustness with respect to the mesh-size, and a poly-logarithmic growth of the condition number with the polynomial order p.

There is now an established literature on high order finite element methods, from the more theoretical point of view as well as from an applied one [13, 26, 41, 43].

We consider two strategies for domain decomposition algorithms [45], non-overlapping Schwarz type methods [15, 20, 21] and balancing domain decomposition with constraints (BDDC) [14, 33]. There is a big literature, in particular high order methods and three dimensional problems are treated in [2, 5, 7, 8, 9, 19, 23, 27, 28, 30, 30, 36, 37, 38, 40, 42]. There is a classical paper on multi-level analysis for h-version DG methods by Gopalakrishnan and Kanschat [18], and a recent one studying higher order methods by Antonietti and Houston [3] showing a polynomial growth of the condition number in p. We will see that the conditioning is significantly improved by hybridization, namely to a poly-logarithmic growth. We are not aware of particular analysis for preconditioners for high order HDG methods, even not in 2D.

The main result of the present paper is Theorem 3 proving that optimal extension from faces to elements with Dirichlet constraints is nearly as good as extension without constraints. With this result condition number estimates follow with the usual techniques.

The main difficulty is to build optimal extension operators from an edge to a tetrahedron. This problem was solved for hexahedral elements by multiplying with fast decaying functions by Pavarino and Widlund [38]. Polynomial extension operators for simplicial elements are usually based on smoothing operators [5, 34]. Heuer and Leydecker have analyzed such operators also for boundary elements, i.e, for three dimensional edge to face extension.

We cannot use the existing simplicial extension operators to prove quasi-optimality of HDG methods since they do not decay fast enough in the jump-norm. We give a new construction of discrete edge-to-tetrahedron extension operators which are motivated by the multiplication with low-energy functions of Pavarino and Widlund, but are contained in the polynomial space on tetrahedra.

We declare some notation. With $a \preceq b$ we mean the existence of a generic constant c such that $a \leq cb$, where c is independent of parameters h and p. Otherwise, we denote the dependence as $c(p)$. The space of univariate polynomials of order p is P^p, and $P^p(T)$ is the space of multivariate polynomials of total order p on a simplex T. To simplify notation we redefine $\log p := 1$ for $p \in \{0,1\}$.

In Sect. 2 we give the hybrid DG formulation, in Sect. 3 we prove the main result, Theorem 3, and show how to apply it to analyze domain decomposition algorithms for HDG. Technical lemmas are shifted to Sect. 4, 5, and 6. In Sect. 4 we collect properties of orthogonal polynomials, and prove one dimensional trace estimates and construct one-dimensional extension operators with respect to different norms. The short Sect. 5 gives the proofs for extension from vertices, the technical proofs for the extension from edges are in Sect. 6.

2 HDG Discretization

Let $\Omega \subset \mathbb{R}^3$ be a polyhedral domain. Let $\mathcal{T} = \{T\}$ be a conforming triangulation of Ω consisting of shape regular tetrahedral elements. With $\mathcal{F} = \{F\}$ we denote the set of all faces, and $\mathcal{F}_T$ are the faces of the element T. As usual $h_T = \operatorname{diam} T$ is the local mesh-size.

We consider the Dirichlet problem of the Poisson equation problem, namely

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

with the source $f \in L_2(\Omega)$. We define the p^{th} order hybrid discontinuous Galerkin finite element space

$$V_N := P^p(\mathcal{T}) \times P^p(\mathcal{F}) := \prod_{T\in\mathcal{T}} P^p(T) \times \prod_{F\in\mathcal{F}} P^p(F),$$

its subspace $V_{N,0} = \{(u,\lambda) \in V_N : \lambda = 0 \text{ on } \partial\Omega\}$, and the hybrid discontinuous Galerkin (HDG) method as: find $(u_N, \lambda_N) \in V_{N,0}$:

$$A(u_N, \lambda_N; v, \mu) = (f, v)_{L_2(\Omega)} \qquad \forall (v,\mu) \in V_{N,0}.$$

The HDG bilinear-form is

$$A(u,\lambda;v,\mu) = \sum_{T\in\mathcal{T}} A_T(u,\lambda;v,\mu)$$

with the element contributions

$$A_T(u,\lambda;v,\mu) := \int_T \nabla u \nabla v + \int_{\partial_T} \frac{\partial u}{\partial n}(\mu - v) + \int_{\partial T} \frac{\partial v}{\partial n}(\lambda - u) + \alpha(u-\lambda, v-\mu)_{j,\partial T}$$

with a fixed $\alpha > 4 = |\mathcal{F}_T|$. We choose the stabilization similar to the stabilized Bassi-Rebay method [4, 6, 25] as

$$(u-\lambda, v-\mu)_{j,\partial T} = \sum_{F\in\mathcal{F}_T} (r_F(u-\lambda), r_F(v-\mu))_{L_2(T)}.$$

The discrete lifting operator $r_F : P^p(F) \to [P^p(T)]^3$ is defined by

$$(r_F(\mu), v)_{L_2(T)} = (\mu, v\cdot n)_{L_2(F)} \qquad \forall v \in [P^p(T)]^3.$$

The norm

$$\|u-\lambda\|_{j,F} = \|r_F(u-\lambda)\|_{L_2(T)}$$

is realized by

$$\|u-\lambda\|_{j,F} = \sup_{\sigma\in[P^p(T)]^3} \frac{(u-\lambda, \sigma\cdot n)_{L_2(F)}}{\|\sigma\|_{L_2(T)}} = \sup_{\sigma\in P^p(T)} \frac{(u-\lambda, \sigma)_{L_2(F)}}{\|\sigma\|_{L_2(T)}}. \qquad (1)$$

The last equality holds since the normal vector n is constant on F.

We define the norm

$$\|(u,\lambda)\|_{1,HDG}^2 := \sum_{T\in\mathcal{T}} \left\{ \|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,\partial T}^2 \right\}$$

with $\|u-\lambda\|_{j,\partial T}^2 = \sum_{F\in\mathcal{F}_T} \|u-\lambda\|_{j,F}^2$. We note that more general elliptic equations, with mixed boundary conditions, variable coefficients as well as variable polynomial orders can be treated the same way.

Theorem 1. *The bilinear-form* $A(.,.)$ *is continuous and coercive on* $(V_{N,0}, \|\cdot\|_{1,HDG})$.

Proof. Continuity and coercivity are proven element-wise, i.e.,

$$\|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,\partial T}^2 \preceq A_T(u,\lambda;u,\lambda) \preceq \|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,\partial T}^2$$

is shown for all $u \in P^p(T), \lambda \in P^p(\mathcal{F}_T)$, and for all $T \in \mathcal{T}$. For $F \in \mathcal{F}_T$ we use Young's inequality $ab \le \frac{1}{2\gamma}a^2 + \frac{\gamma}{2}b^2$ with $4 < \gamma < \alpha$ to obtain

$$\begin{aligned}\int_F \frac{\partial u}{\partial n}(u-\lambda) &\le \|\nabla u\|_{L_2(T)} \sup_{\sigma\in[P^p]^3} \frac{\int \sigma_n (u-\lambda)}{\|\sigma\|_{L_2(T)}} \\ &\le \frac{1}{2\gamma}\|\nabla u\|_{L_2(T)}^2 + \frac{\gamma}{2}\|u-\lambda\|_{j,F}^2.\end{aligned}$$

Summing over the 4 faces of T we obtain

$$\begin{aligned}A_T(u,\lambda;u,\lambda) &= \|\nabla u\|_{L_2(T)}^2 + 2\sum_{F\in\mathcal{F}_T} \int_F \frac{\partial u}{\partial n}(u-\lambda) + \alpha(u-\lambda,u-\lambda)_{j,T} \\ &\ge \|\nabla u\|_{L_2(T)}^2 - \frac{4}{\gamma}\|\nabla u\|_{L_2(T)}^2 - \gamma\sum_F \|u-\lambda\|_{j,F}^2 + \alpha\|u-\lambda\|_{j,\partial T}^2 \\ &\succeq \|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,\partial T}^2,\end{aligned}$$

continuity is verified similar. □

Theorem 1 allows to reduce the analysis of preconditioners for $A(.,.)$ to the form generated by the norm $\|(u,\lambda)\|_{1,HDG}$. Theorem 1 is also the basis for a-priori error estimates, for example the h-version estimate

$$\|(u-u_N, u-\lambda_N)\|_{1,HDG} \preceq h^s \|u\|_{H^{1+s}(\Omega)}$$

for $1 \le s \le p$, see [31].

Theorem 2. *For* $F \in \mathcal{F}_T$ *let* $\mathbf{P}^k$ *denote the* $L_2(F)$*-orthogonal projector onto* $P^k(F)$, *with* $\mathbf{P}^{-1} = 0$. *For* $\lambda \in P^p(F)$ *there holds*

$$\|\lambda\|_{j,F}^2 \simeq h_T^{-1} \sum_{k=0}^{p} p(p-k+1)\,\|(\mathbf{P}^k - \mathbf{P}^{k-1})\lambda\|_{L_2(F)}^2.$$

Proof. By an affine-linear transformation to the reference tetrahedron and reference face

$$T = \{(x,y,z) : y \geq 0, z \geq 0, |x| + y + z \leq 1\}, \tag{2}$$
$$F = \{(x,y,0) : y \geq 0, |x| + y \leq 1\} \tag{3}$$

one obtains the scaling in the mesh-size. By means of Jacobi and Legendre polynomials (see Sect. 4), the Dubiner basis polynomials [16, 26]

$$\varphi_{ij}(x,y) = P_i\Big(\frac{x}{1-y}\Big)(1-y)^i P_j^{(0,2i+1)}(1-2y) \qquad \text{for } i+j \leq p$$

form an $L_2(F)$-orthogonal basis for $P^p(F)$. Expand

$$\lambda(x,y) = \sum_{i+j \leq p} \lambda_{ij}\varphi_{ij}(x,y)$$
$$\sigma(x,y,z) = \sum_{i+j \leq p} \varphi_{ij}\Big(\frac{x}{1-z}, \frac{y}{1-z}\Big)(1-z)^{i+j}\sigma_{ij}(z)$$

with $\sigma_{ij} \in P^{p-i-j}$. By the change of variables

$$g : F \times [0,1] \to T : (\xi, \eta, z) \mapsto (x,y,z) := ((1-z)\xi, (1-z)\eta, z)$$

with $\det g' = (1-z)^2$ we express

$$\|\sigma\|^2_{L_2(T)} = \int_F \int_0^1 (1-z)^2 \sigma((1-z)\xi, (1-z)\eta, z)^2 \, dz d(\xi, \eta).$$

Due to orthogonality there holds

$$\|\sigma\|^2_{L_2(T)} = \sum_{i+j \leq p} \|\varphi_{ij}\|^2_{L_2(F)} \int_0^1 (1-z)^{2i+2j+2} \sigma^2_{ij}(z) \, dz$$

and

$$(\lambda, \sigma)_{L_2(F)} = \sum_{i+j \leq p} \|\varphi_{ij}\|^2_{L_2(F)} \lambda_{ij}\sigma_{ij}(0).$$

There holds

$$\sup_{\sigma \in P^p(T)} \frac{(\lambda, \sigma)^2_{L_2(F)}}{\|\sigma\|^2_{L_2(T)}} = \|\sigma^*\|^2_{L_2(T)},$$

where $\sigma^* \in P^p(T)$ solves

$$(\sigma^*, \tau)_{L_2(T)} = (\lambda, \tau)_{L_2(F)} \qquad \forall \tau \in P^p(T).$$

The components $\sigma_{ij}^* \in P^{p-i-j}$ of the $L_2(T)$-orthogonal decomposition

$$\sigma^*(x,y,z) = \sum_{i+j\leq p} \varphi_{ij}\Big(\frac{x}{1-z}, \frac{y}{1-z}\Big)(1-z)^{i+j}\sigma_{ij}^*(z)$$

solve

$$\int_0^1 (1-z)^{2i+2j+2}\sigma_{ij}^*(z)\,\tau(z)\,dz = \lambda_{ij}\tau(0) \qquad \forall\,\tau \in P^{p-i-j},$$

and there holds

$$\int_0^1 (1-z)^{2i+2j+2}\sigma_{ij}^*(z)^2\,dz = \sup_{\sigma_{ij}\in P^{p-i-j}} \frac{(\lambda_{ij}\sigma_{ij}(0))^2}{\int_0^1 (1-z)^{2i+2j+2}\sigma_{ij}^2(z)\,dz}.$$

From Lemma 1 below we get

$$|\sigma_{ij}(0)|^2 \preceq p(p-i-j+1)\int_0^1 (1-z)^{2i+2j+2}\sigma_{ij}^2(z)\,dz$$

is sharp, and thus

$$\int_0^1 (1-z)^{2i+2j+2}\sigma_{ij}^*(z)^2\,dz \simeq p(p-i-j+1)\lambda_{ij}^2.$$

Thus there holds

$$\begin{aligned}\sup_{\sigma\in P^p(T)} \frac{(\lambda,\sigma)_{L_2(F)}^2}{\|\sigma\|_{L_2(T)}^2} &\simeq \sum_{i+j\leq p} p(p-i-j+1)\,\lambda_{ij}^2\|\varphi_{ij}\|_{L_2(F)}^2 \\ &= \sum_{k=0}^{p} p(p-k+1)\sum_{i+j=k}\lambda_{ij}^2\|\varphi_{ij}\|^2 \\ &= \sum_{k=0}^{p} p(p-k+1)\|(\mathbf{P}^k-\mathbf{P}^{k-1})\lambda\|_{L_2(F)}^2. \qquad \square\end{aligned}$$

We observe that

$$\frac{p}{h}\|u-\lambda\|_{L_2(F)}^2 \preceq \|u-\lambda\|_{j,F}^2 \preceq \frac{p^2}{h}\|u-\lambda\|_{L_2(F)}^2. \tag{4}$$

Often $\frac{\alpha p^2}{h}\|u-\lambda\|_{L_2(F)}^2$ with a sufficiently large parameter α is chosen as penalty term. Usually α is chosen on the safe side. We will see in the numerical examples that the condition number does increase with α. In this paper we prove quasi-optimal condition numbers for the presented stabilization, it does not carry over to the weighted L_2-stabilization.

The benefit is two-fold, on one side the method is guaranteed to be stable, for any $\alpha > |\mathbf{F}_T|$, on the other side the condition number is proven to have only poly-logarithmic growth.

3 Domain Decomposition Preconditioning

The analysis of non-overlapping DD preconditioners is based on stable decompositions of finite element functions. For that, quasi-optimal extension procedures are essential. The main result of our work is to construct an extension operator, and bound its norm.

For $F \in \mathcal{F}$ and a fixed $T \in \mathcal{T}$ such that $F \subset T$ we define the trace semi-norm

$$\|\lambda\|_F^2 = \inf_{u \in P^p(T)} \left\{ \|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,F}^2 \right\}$$

and the trace norm

$$\|\lambda\|_{F,0}^2 = \inf_{u \in P^p(T)} \left\{ \|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,F}^2 + \sum_{\substack{F' \in \mathcal{F}_T \\ F' \neq F}} \|u-0\|_{j,F'}^2 \right\}.$$

The semi-norm $\|\lambda\|_F$ mimics the $H^{1/2}(F)$ semi-norm, i.e. the trace semi-norm corresponding to arbitrary H^1-optimal extension onto the element T, while the norm $\|\lambda\|_{F,0}$ mimics the $H_{00}^{1/2}$-norm, i.e., the trace norm corresponding to H^1-optimal extension under Dirichlet constraints on $\partial T \setminus F$. Note that for continuous finite element spaces $\|\lambda\|_{H_{00}^{1/2}}$ is defined only for $\lambda = 0$ on ∂F. For hybrid DG, both norms $\|\lambda\|_F$ and $\|\lambda\|_{F,0}$ are defined for the same space $P^p(F)$.

Theorem 3. *Let $\lambda_F \in P^p(F)$ with $\int_F \lambda = 0$. Then here holds*

$$\|\lambda\|_{F,0}^2 \preceq (\log p)^\gamma \, \|\lambda\|_F^2$$

with $\gamma = 3$.

Proof. It is enough to consider the reference element T. Let u be the minimizer corresponding to $\|\lambda\|_F$. Thanks to a Poincare-type inequality and Theorem 2 there holds

$$\begin{aligned} \|u\|_{H^1(T)}^2 &\preceq \|\nabla u\|_{L_2(T)}^2 + \Big(\int_F u\Big)^2 \\ &\preceq \|\nabla u\|_{L_2(T)}^2 + \Big(\int_F u - \lambda\Big)^2 + \Big(\int_F \lambda\Big)^2 \\ &\preceq \|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,F}^2. \end{aligned}$$

We modify the function u by subtracting vertex and edge contributions:

$$u_2 = u - \sum_{V \subset F} \mathscr{E}_{V \to T}\, u(V),$$
$$u_3 = u_2 - \sum_{E \subset F} \mathscr{E}_{E \to T}\, u_2|_E.$$

In Theorem 6 and Theorem 7 below we prove that the function u_3 is in P^p, vanishes on ∂F, and satisfies

$$\|\nabla u_3\|^2_{L_2(T)} + \|u - u_3\|^2_{j,F} \preceq \log p\, \|u\|^2_{H^1(T)}.$$

There holds [8, Lemma 4.7]

$$\|u_3\|^2_{H^{1/2}_{00}(F)} \preceq (\log p)^2\, \|u_3\|^2_{H^{1/2}(F)}.$$

Now take

$$\tilde{u} := \mathscr{E}^{MS}_{F \to T}\, u_3|_F$$

as the Muñoz-Sola extension [34]. Finally we get

$$\|\nabla \tilde{u}\|^2_{L_2(T)} + \|\tilde{u} - \lambda\|^2_{j,F} \preceq (\log p)^\gamma \|\lambda\|^2_F,$$

and together with $\tilde{u} = 0$ on $\partial T \setminus F$ we have proven the result. □

We note that in [8, 37, 38] and others estimates with $(\log p)^2$ have been obtained for continuous finite elements. It might be that our result can also be improved to $(\log p)^2$. One approach would be to directly estimate the $\int_F \frac{1}{\operatorname{dist}(x,\partial F)} u(x)^2\, dx$ term of the $H^{1/2}_{00}(F)$-norm. If one succeeds with that estimate, then that improved γ can be used immediately in the following condition number estimates.

3.1 Schwarz Type Domain Decomposition

To analyze Scharz-type domain decomposition methods one has to prove stable decompositions into sub-spaces [32].

For $\lambda \in P^p(\mathcal{F})$ we define the Schur-complement norm

$$\|\lambda\|^2_S = \inf_{u \in P^p(\mathcal{T})} \|(u, \lambda)\|^2_{1,HDG}.$$

Theorem 4. *Let $\lambda \in P^p(\mathcal{F})$. Define the coarse grid component as*

$$\lambda_H \in P^0(\mathcal{F}) \text{ such that } \int_F \lambda_H = \int_F \lambda,$$

and for $F \in \mathcal{F}$ define the local components λ_F as

$$\lambda_F = \begin{cases} \lambda_{|F} - \lambda_{H|F} & \text{on } F, \\ 0 & \text{for } F' \neq F. \end{cases}$$

Then there holds

$$\|\lambda_H\|_S^2 + \sum_{F \in \mathcal{F}} \|\lambda_F\|_S^2 \preceq (\log p)^\gamma \|\lambda\|_S^2.$$

Thus, the additive Schwarz preconditioner C_{ASM} applied to the facet Schur-complement S_A of A leads to a condition number estimate

$$\kappa(C_{ASM}^{-1} S_A) \preceq (\log p)^\gamma.$$

Proof. From the definitions of the norms there follows

$$\sum_{F \in \mathcal{F}} \|\mu|_F\|_F^2 \preceq \|\mu\|_S^2 \preceq \sum_{F \in \mathcal{F}} \|\mu|_F\|_{F,0}^2 \qquad \forall \mu \in P^p(\mathcal{F}).$$

Since $\int_F \lambda_F = 0$ we have

$$\begin{aligned} \sum_{F \in \mathcal{F}} \|\lambda_F\|_S^2 \preceq \sum_{F \in \mathcal{F}} \|\lambda_F|_F\|_{F,0}^2 &\preceq (\log p)^\gamma \sum_{F \in \mathcal{F}} \|\lambda_F|_F\|_F^2 \\ &= (\log p)^\gamma \sum_{F \in \mathcal{F}} \|\lambda|_F\|_F^2 \preceq (\log p)^\gamma \|\lambda\|_S^2, \end{aligned}$$

and

$$\begin{aligned} \|\lambda_H\|_S^2 = \|\lambda - \sum_{F \in \mathcal{F}} \lambda_F\|_S^2 &\preceq \|\lambda\|_S^2 + \|\sum_{F \in \mathcal{F}} \lambda_F\|_S^2 \\ &\preceq \|\lambda\|_S^2 + \sum_{F \in \mathcal{F}} \|\lambda_{F|F}\|_{F,0}^2 \preceq (\log p)^\gamma \|\lambda\|_S^2. \end{aligned}$$

Due to finite overlap of the sub-spaces, the largest eigenvalue of $C_{ASM}^{-1} S$ is bounded by a constant, and thus the condition number is bounded by $(\log p)^\gamma$. □

3.2 BDDC Preconditioners

To define a BDDC preconditioner one sub-divides degrees of freedom into primal and dual, see [14, 33]. The dual ones are treated discontinuous, and thus can be eliminated on the element-level. In our case we choose the mean value on the face as primal, all others are dual degrees of freedom. Thus, the remaining global system involves only one degree of freedom per face.

Theorem 5. *The BDDC preconditioner with mean values on faces leads to a condition number*

$$\kappa(C_{BDDC}^{-1} S_A) \preceq (\log p)^\gamma.$$

Proof. Let λ be double-valued on faces with consistent mean-values, this means

$$\lambda = (\lambda_T)_{T\in\mathcal{T}} \in \prod_{T\in\mathcal{T}} P^p(\mathcal{F}_T),$$

such that

$$\int\limits_F \lambda_T = \int\limits_F \lambda_{T'} \qquad \text{for } F = T \cap T'.$$

Define the average $\tilde{\lambda} \in P^p(\mathcal{F})$ as

$$\tilde{\lambda} = \frac{\sum_{T:F\subset T} \lambda_{T|F}}{\sum_{T:F\subset T} 1}.$$

We have to prove continuity of the averaging operator, i.e.

$$\|\tilde{\lambda}\|_S^2 \leq c(p) \sum_{T\in\mathcal{T}} \|\lambda_T\|_{S,T}^2,$$

where $\|\lambda_T\|_{S,T}^2 := \inf_{u\in P^p(T)} \left\{ \|\nabla u\|_{L_2(T)}^2 + \|u-\lambda\|_{j,\partial T}^2 \right\}$.

We use $\int_F \tilde{\lambda} = \int_F \lambda_T$ to apply Theorem 3 for estimating

$$\begin{aligned}
\|\tilde{\lambda}\|_S^2 = \sum_{T\in\mathcal{T}} \|\tilde{\lambda}_{|\partial T}\|_{S,T}^2 &\preceq \sum_{T\in\mathcal{T}} \left\{ \|\lambda_T\|_{S,T}^2 + \|\tilde{\lambda}_{|\partial T} - \lambda_T\|_{S,T}^2 \right\} \\
&\leq \sum_{T\in\mathcal{T}} \left\{ \|\lambda_T\|_{S,T}^2 + \sum_{F\in\mathcal{F}_T} \|\tilde{\lambda}_{\partial T} - \lambda_T\|_{F,0}^2 \right\} \\
&\preceq (\log p)^\gamma \sum_{T\in\mathcal{T}} \left\{ \|\lambda_T\|_{S,T}^2 + \sum_{F\in\mathcal{F}_T} \|\tilde{\lambda}_{\partial T} - \lambda_T\|_F^2 \right\} \\
&\preceq (\log p)^\gamma \sum_{T\in\mathcal{T}} \left\{ \|\lambda_T\|_{S,T}^2 + \sum_{F\in\mathcal{F}_T} \|\lambda_T\|_F^2 \right\} \\
&\preceq (\log p)^\gamma \sum_{T\in\mathcal{T}} \|\lambda_T\|_{S,T}^2.
\end{aligned}$$

The condition number $\kappa(C_{BDDC}^{-1}A)$ is given by the continuity bound $c(p) = (\log p)^\gamma$. □

4 Traces and Polynomial Extensions on the Interval

In this section we collect some properties of Jacobi polynomials which can be found in [44, Chapter 4], or [1], then we prove trace and extension estimates on the interval. Let $w = (1-x)^\alpha(1+x)^\beta$ be the weight function, for us $\alpha, \beta \in \mathbb{N}_0$ is sufficient. The

n^{th}-order Jacobi polynomial $P_n^{(\alpha,\beta)}$ is defined by Rodrigues' formula as

$$P_n^{(\alpha,\beta)}(x) := \frac{1}{(-2)^n n!\, w(x)} \frac{d^n}{dx^n} \Big(w(x)(1-x^2)^n \Big).$$

There holds the orthogonality relation

$$\int_{-1}^{1} w P_n^{(\alpha,\beta)} P_m^{(\alpha,\beta)}\, dx = \delta_{n,m} \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{(n+\alpha)!\,(n+\beta)!}{n!\,(n+\alpha+\beta)!},$$

and boundary values are

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}.$$

The Legendre polynomials are $P_n := P_n^{(0,0)}$, and the integrated Legendre polynomials are defined as

$$L_n(x) = \int_{-1}^{x} P_{n-1}(s)\, ds.$$

We often use

$$\|P_n\|^2_{L_2([-1,1])} = \frac{2}{2n+1},$$

and we need

$$(2n+1)L_{n+1} = P_{n+1} - P_{n-1}.$$

Parameters can be shifted by

$$(2n+\alpha+\beta)P_n^{(\alpha-1,\beta)} = (n+\alpha+\beta)P_n^{(\alpha,\beta)} - (n+\beta)P_{n-1}^{(\alpha,\beta)},$$

and by telescoping one obtains for the particular choice $\alpha = 1$

$$(m+\beta+1)P_m^{(1,\beta)} = \sum_{n=0}^{m} (2n+\beta+1)P_n^{(0,\beta)}. \tag{5}$$

Differentiating Jacobi polynomials gives

$$\frac{d}{dx} P_n^{(\alpha,\beta)} = \frac{1}{2}(n+\alpha+\beta+1)P_{n-1}^{(\alpha+1,\beta+1)}. \tag{6}$$

Lemma 1 (Trace inequality 1D). *For $v \in P^n$ there holds*

$$v(0)^2 \leq S(\alpha,\beta,n) \int_0^1 y^\alpha (1-y)^\beta v(y)^2\, dy, \tag{7}$$

and for every n there exists an $l_n^{(\alpha,\beta)} \in P^n$ such that $l_n^{(\alpha,\beta)}(0) = 1$ and

$$\int_0^1 y^\alpha (1-y)^\beta l_n^{(\alpha,\beta)}(y)^2 \, dy \le S(n,\alpha,\beta)^{-1},$$

with

$$S(n,\alpha,\beta) = \frac{(n+\alpha+1)!\,(n+\alpha+\beta+1)!}{\alpha!\,(\alpha+1)!\,n!\,(n+\beta)!}.$$

For a fixed α there holds

$$S(n,\alpha,\beta) \simeq (n+1)^{\alpha+1}(n+1+\beta)^{\alpha+1}.$$

Proof. Estimate (7) is sharp for the solution of the constrained minimization problem

$$\min_{v(0)=1} \int_0^1 y^\alpha (1-y)^\beta v(y)^2 \, dy.$$

By choosing the representation

$$v(y) = \sum_{k=0}^{n} c_k P_k^{(\alpha,\beta)}(1-2y),$$

the minimization problem can be rephrased as

$$\min_{\substack{c\in\mathbb{R}^{n+1}\\ b^\top c=1}} c^\top D c$$

with $b \in \mathbb{R}^{n+1}$ and $D \in \mathbb{R}^{(n+1)\times(n+1)}$ diagonal with components

$$\begin{aligned}
b_k &= P_k^{(\alpha,\beta)}(1) = \frac{(k+\alpha)!}{\alpha!\,k!},\\
D_{k,k} &= \int_0^1 y^\alpha (1-y)^\beta P_k^{(\alpha,\beta)}(1-2y)^2 \, dy = \int_{-1}^1 \left(\frac{1-z}{2}\right)^\alpha \left(\frac{1+z}{2}\right)^\beta P_k^{(\alpha,\beta)}(z)^2 \, \frac{1}{2} dz\\
&= \frac{1}{2k+\alpha+\beta+1} \frac{(k+\alpha)!\,(k+\beta)!}{k!\,(k+\alpha+\beta)!}.
\end{aligned}$$

Using the method of Lagrange multipliers we obtain

$$\begin{pmatrix} D & b \\ b^\top & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

With the Schur complement

$$S = b^\top D^{-1} b = \sum_{k=0}^{n} \frac{b_k^2}{D_{k,k}} = \sum_{k=0}^{n} (2k+\alpha+\beta+1) \frac{(k+\alpha)!\,(k+\alpha+\beta)!}{(\alpha!)^2\, k!\,(k+\beta)!},$$

the solution is given by $\lambda = \frac{-1}{S}$ and $c = \frac{1}{S} D^{-1} b$. The value of the minimum is S^{-1}.

By means of the Paule/Schorn implementation [35] of Gosper's algorithm, V. Pillwein computed

$$S = \frac{(n+\alpha+1)!\,(n+\alpha+\beta+1)!}{\alpha!\,(\alpha+1)!\,n!\,(n+\beta)!}.$$

More on computer algebra techniques in finite element methods is found in [39].

We continue with a hand-proof for the asymptotic behavior:

$$\begin{aligned}
S &\simeq c(\alpha)\sum_{k=0}^{n}(k+1)^{\alpha}(k+\beta+1)^{\alpha+1}\\
&= c(\alpha)\sum_{k=0}^{n}(k+1)^{\alpha}\sum_{j=0}^{\alpha+1}\binom{\alpha+1}{j}(k+1)^{j}\beta^{\alpha+1-j}\\
&\simeq c(\alpha)\sum_{j=0}^{\alpha+1}\binom{\alpha+1}{j}(n+1)^{j+\alpha+1}\beta^{\alpha+1-j}\\
&= c(\alpha)(n+1)^{\alpha+1}(n+\beta+1)^{\alpha+1}.
\end{aligned}$$ □

Lemma 2. *For $v \in P^n$ there holds*

$$(v(0)-v(1))^2 \preceq \log n \int_0^1 y(1-y)v'(y)^2\,dy, \tag{8}$$

$$v(0)^2 \preceq \log n \int_0^1 y(1-y)(v'(y)^2+v(y)^2)\,dy. \tag{9}$$

Proof. To verify (8) we follow the lines of Lemma 1. Now we expand

$$v(y) = \sum_{k=0}^{n} c_k P_k^{(0,0)}(1-2y),$$

from (6) there follows

$$v'(y) = -\sum_{k=1}^{n} c_k(k+1)P_{k-1}^{(1,1)}(1-2y),$$

and now

$$\begin{aligned}
b_k &= P_k^{(0,0)}(1) - P_k^{(0,0)}(-1) = 1+(-1)^k,\\
D_{k,k} &= (k+1)^2\int_0^1 y(1-y)P_{k-1}^{(1,1)}(1-2y)^2\,dy = (k+1)^2\frac{k}{(2k+1)(k+1)},
\end{aligned}$$

and thus

$$S = \sum_{k=0}^{p} \frac{b_k^2}{D_{k,k}} = \sum_{\substack{k=0\\k \text{ even}}}^{n} \frac{4\,(2k+1)}{(k+1)k} \simeq \sum_{k=0}^{n} \frac{1}{k+1} \simeq \log n.$$

Estimate (9) follows from (8) as follows: for $\tilde{v}(y) := (1-y)v(y)$ we apply (8) to obtain

$$\begin{aligned} v(0)^2 = \tilde{v}(0)^2 &\preceq \log n \int_0^1 y(1-y)\tilde{v}'(y)^2\,dy \\ &= \log n \int_0^1 y(1-y)[-v(y) + (1-y)v'(y)]^2\,dy \\ &\preceq \log n \int_0^1 y(1-y)\big(v(y)^2 + v'(y)^2\big)\,dy. \end{aligned}$$ □

Next we prove that the minimal energy extension in certain norms is also quasi-optimal in related norms:

Lemma 3. *We define for* $n, \beta \in \mathbb{N}_0$

$$l_n^\beta := \operatorname*{argmin}_{v \in P^n, v(0)=1} \int_0^1 y(1-y)^\beta v(y)^2\,dy.$$

Then there holds

$$\int_0^1 y(1-y)^\beta l_n^\beta(y)^2\,dy \preceq \frac{1}{(n+1)^2(n+\beta+1)^2}, \tag{10}$$

$$\int_0^1 (1-y)^\beta l_n^\beta(y)^2\,dy \preceq \frac{1}{(n+1)(n+\beta+1)}, \tag{11}$$

$$\int_0^1 y(1-y)^{\beta+1}\big((l_n^\beta)'(y)\big)^2\,dy \preceq 1. \tag{12}$$

Proof. The optimizer l_n^β was calculated in the proof of Lemma 1 with $\alpha = 1$, namely

$$l_n^\beta(y) = \sum_{k=0}^{n} c_k P_k^{(1,\beta)}(1-2y),$$

with $b_k = k+1$ and $D_{k,k} = \dfrac{1}{2k+\beta+2}\dfrac{k+1}{k+\beta+1}$. We get

$$c_k = \frac{b_k}{D_{k,k}S} = \frac{(2k+\beta+2)(k+\beta+1)}{S},$$

and $S \simeq (n+1)^2(n+\beta+1)^2$. Inequality (10) was proven in Lemma 1. To verify (11) we utilize (5) to re-expand l_n^β in terms of Jacobi-polynomials $P_n^{(0,\beta)}$:

$$\begin{aligned} l_n^\beta(y) &= \sum_{k=0}^{n} c_k \sum_{j=0}^{k} \frac{2j+\beta+1}{k+\beta+1} P_j^{(0,\beta)}(1-2y) \\ &= \sum_{j=0}^{n}\sum_{k=j}^{n} \frac{2j+\beta+1}{k+\beta+1} c_k P_j^{(0,\beta)}(1-2y) \\ &= \frac{1}{S}\sum_{j=0}^{n}\sum_{k=j}^{n}(2k+\beta+2)(2j+\beta+1)P_j^{(0,\beta)}(1-2y) \\ &= \frac{1}{S}\sum_{j=0}^{n}(n+j+\beta+2)(n-j+1)(2j+\beta+1)P_j^{(0,\beta)}(1-2y). \end{aligned} \tag{13}$$

Thus there holds

$$\begin{aligned} &\int_0^1 (1-y)^\beta l_n^\beta(y)^2 dy = \\ &= \sum_{j=0}^{n} \int_0^1 (1-y)^\beta (P_j^{(0,\beta)}(1-2y))^2 dy \, \frac{(n+j+\beta+1)^2(n-j+1)^2(2j+\beta+1)^2}{S^2} \\ &\simeq \sum_{j=0}^{n} \frac{1}{2j+\beta+1} \frac{(n+j+\beta+1)^2(n-j+1)^2(2j+\beta+1)^2}{(n+1)^4(n+\beta+1)^4} \\ &\preceq \frac{1}{(n+1)(n+\beta+1)}. \end{aligned}$$

From (13) and (6) we get

$$(l_n^\beta)'(y) = \frac{1}{S}\sum_{j=1}^{n} \frac{j+\beta+1}{2}(n+j+\beta+2)(n-j+1)(2j+\beta+1)P_{j-1}^{(1,\beta+1)},$$

and thus with similar arguments as above

$$\int_0^1 y(1-y)^{\beta+1}(l_n^\beta)'(y)^2\, dy \preceq 1. \qquad \square$$

We construct a family of minimal extensions $\{e_i^p : 0 \le i \le p\}$ similar to Lemma 3, such that the differences between two consecutive functions is small. We obtain this by weighted averaging of the previously defined l_n^β.

Lemma 4. *For i such that $p/2 \le i \le p$ we define the weighted average*

$$e_i^p(y) = \frac{1}{\sum_{k=i}^p w_k} \sum_{k=i}^p w_k (1-y)^{k-i} l_{p-k}^{2k-1}(y) \qquad \text{with } w_k = (p-k+1)$$

and for $i < p/2$ we set

$$e_i^p(y) := (1-y)^{\lceil p/2 \rceil - i} e_{\lceil p/2 \rceil}^p(y).$$

There holds $e_i^p \in P^{p-i}$, it satisfies the boundary condition $e_i^p(0) = 1$ and the estimates

$$\int_0^1 y(1-y)^{2i-1} e_i^p(y)^2 \, dy \preceq \frac{1}{p^2(p-i+1)^2}, \tag{14}$$

$$\int_0^1 (1-y)^{2i-1} e_i^p(y)^2 \, dy \preceq \frac{1}{p(p-i+1)}, \tag{15}$$

$$\int_0^1 y(1-y)\Big(\frac{d}{dy}\big((1-y)^i e_i^p(y)\big)\Big)^2 dy, \preceq 1. \tag{16}$$

We define differences of consecutive functions as

$$d_i^p(y) := e_i^p(y) - (1-y) e_{i+1}^p(y),$$

they satisfy $d_i^p \in P^{p-i}$, $d_i^p = 0$ for $i < p/2$, and

$$\int_0^1 (1-y)^{2i-1} d_i^p(y)^2 \, dy \preceq \frac{i^2}{p^3(p-i+1)^3}, \tag{17}$$

$$\int_0^1 y(1-y)\Big(\frac{d}{dy}\big((1-y)^i d_i^p(y)\big)\Big)^2 dy \preceq \frac{i^2}{p^2(p-i+1)^2}. \tag{18}$$

Proof. We apply the triangle inequality, use (10), and $\sum\limits_{k=i}^{p} w_k \simeq (p-i+1)^2$ to prove (14):

$$\Big(\int_0^1 y(1-y)^{2i-1} e_i^p(y)^2 \, dy\Big)^{1/2}$$

$$\le \frac{1}{\sum_{k=i}^p w_k} \sum_{k=i}^p w_k \Big(\int_0^1 y(1-y)^{2i-1} \big((1-y)^{k-i} l_{p-k}^{2k-1}(y)\big)^2 dy\Big)^{1/2}$$

$$\preceq \frac{1}{\sum w_k} \sum_{k=i}^{p} (p-k+1) \frac{1}{(p-k+1)(p-k+1+2k-1)} \preceq \frac{1}{p(p-i+1)}.$$

The estimates (15) and (16) follow similarly.

The differences d_i^p vanish for $i < p/2$, and (17) is trivially fulfilled for $i < p/2$. Thus we assume $i \geq p/2$. We realize that

$$\begin{aligned} d_i^p(y) &= e_i^p(y) - (1-y) e_{i+1}^p(y) \\ &= \frac{1}{\sum_{k=i}^p w_k} \sum_{k=i}^{p} w_k (1-y)^{k-i} l_{p-k}^{2k-1}(y) - \frac{1}{\sum_{k=i+1}^p w_k} \sum_{k=i+1}^{p} w_k (1-y)^{k-i} l_{p-k}^{2k-1}(y) \\ &= \frac{w_i}{\sum_{k=i}^p w_k} \Big(l_{p-i}^{2i-1}(y) - (1-y) e_{i+1}^p(y) \Big). \end{aligned}$$

Since

$$\frac{w_i}{\sum_{k=i}^p w_k} \simeq \frac{1}{p-i+1},$$

we get

$$\begin{aligned} &\int_0^1 (1-y)^{2i-1} d_i^p(y)^2 \, dy \\ &\preceq \frac{1}{(p-i+1)^2} \Big(\int_0^1 (1-y)^{2i-1} l_{p-i}^{2i-1}(y)^2 \, dy + \int_0^1 (1-y)^{2i-1} e_{i+1}^p(y)^2 \, dy \Big) \\ &\preceq \frac{1}{(p-i+1)^3 p}. \end{aligned}$$

The additional factor $\frac{i^2}{p^2}$ follows trivially since $p/2 \leq i \leq p$. Estimate (18) follows similarly. □

5 Extension from a Vertex

In this section we define and analyze minimal extensions from a vertex of the reference tetrahedron T.

Lemma 5. *We define*

$$\tilde{e}_V = \underset{v \in P^{p-1}, v(0)=1}{\operatorname{argmin}} \int_0^1 y^2 v(y)^2 \, dy$$

and

$$e_V(y) = (1-y)\tilde{e}_V(y).$$

Then $e_V \in P^p$ with $e_V(0) = 1$ and $e_V(1) = 0$, and there holds

$$\int_0^1 y^2 e_V(y)^2\,dy \preceq p^{-6}, \qquad \int_0^1 y e_V(y)^2\,dy \preceq p^{-4}, \qquad \int_0^1 y^2 e_V'(y)^2\,dy \preceq p^{-2}.$$

Proof. With Lemma 1 there follows

$$\int_0^1 y^2 e_V(y)^2\,dy \le \int_0^1 y^2 \tilde{e}_V(y)^2\,dy \preceq p^{-6}.$$

With, see [41, Theorem 3.96],

$$\int_0^1 v(y)^2 \preceq p^2 \int_0^1 y(1-y)v(y)^2\,dy,$$

we get

$$\int_0^1 y e_V(y)^2\,dy = \int_0^1 y(1-y)^2 \tilde{e}_V(y)^2\,dy \preceq p^2 \int_0^1 y^2(1-y)^3 \tilde{e}_V(y)^2\,dy \preceq p^{-4},$$

and with

$$\int_0^1 y(1-y)v'(y)^2\,dy \preceq p^2 \int_0^1 v^2(y)\,dy,$$

which is [41, Theorem 3.95], we get

$$\begin{aligned}\int_0^1 y^2 e_V'(y)^2\,dy &= \int_0^1 y^2((1-y)\tilde{e}_V'(y) - \tilde{e}_V(y))^2 dy\\ &\preceq \int_0^1 y^2(1-y)^2 \tilde{e}_V'(y)^2 + \int_0^1 y^2 \tilde{e}_V(y)^2\,dy\\ &\preceq p^2 \int_0^1 y(1-y)\tilde{e}_V(y)^2\,dy + p^{-6} \preceq p^{-2}.\end{aligned}$$ □

Theorem 6 (Extension from a vertex). *Let V be a vertex of the reference tetrahedron T, and λ_V the corresponding barycentric coordinate. Define the vertex-to-element extension $\mathscr{E}_{V\to T} : \mathbb{R} \to P^p(T)$ as*

$$\mathscr{E}_{V\to T} v := e_V(1-\lambda_V)\,v.$$

Then $u := \mathscr{E}_{V\to T}\, v(V)$ vanishes on the face opposite to V and satisfies

$$\|\nabla u\|^2_{L_2(T)} + \sum_{F\in\mathcal{F}_T} \|u\|^2_{j,F} \preceq \|v\|^2_{H^1(T)}.$$

Proof. We recall the inverse estimate $|v(V)| \preceq p^2\|v\|^2_{H^1(T)}$. There holds

$$\|\nabla e_V(1-\lambda_V)\|^2_{L_2(T)} \simeq \int_0^1 (1-\lambda_V)^2 e_V'(1-\lambda_V)^2\, d\lambda_V \preceq p^{-2},$$

and for a face F containing V we have

$$\|e_V\|^2_{j,F} \preceq p^2\|e_V\|^2_{L_2(F)} \simeq p^2 \int_0^1 (1-\lambda_V) e_V(1-\lambda_V)^2\, d\lambda_V \preceq p^{-2},$$

and thus the powers of p cancel out. □

6 Extension from an Edge

In this section we analyze trace operator on edges, and define edge-to-element extension operators. We consider the edge $E = \{(x,0,0) : |x| \le 1\}$ of the reference tetrahedron T defined by (2). We split the construction into two pieces, one is face-to-element extension, the other one is edge-to-face extension.

Lemma 6. *Define the face-to-element extension $\mathscr{E}_{F\to T}$ and the element-to-face restriction operator $\mathscr{R}_{T\to F}$ between the reference tetrahedron T and the reference face F from (2), (3) as*

$$(\mathscr{E}_{F\to T} u)\,(x,y,z) = u(x, y+z), \tag{19}$$

$$(\mathscr{R}_{T\to F} w)\,(x,y) = \int_0^1 w(x, sy, (1-s)z)\, ds. \tag{20}$$

These operators are mappings between $P^p(T)$ and $P^p(F)$, preserve the function on the edge $y = z = 0$, and are continuous with respect to the norms

$$\|\mathscr{R}_{T\to F} w\|_{L_2(F),y} \preceq \|w\|_{L_2(T)}, \qquad \|\mathscr{R}_{T\to F} w\|_{H^1(F),y} \preceq \|w\|_{H^1(T)},$$

$$\|\mathscr{E}_{F\to T} u\|_{L_2(T)} \preceq \|u\|_{L_2(F),y}, \qquad \|\mathscr{E}_{F\to T} u\|_{H^1(T)} \preceq \|u\|_{H^1(F),y}$$

with the norms

$$\|u\|^2_{L_2(F),y} := \int_F y u(x,y)^2\, d(x,y), \qquad \|u\|^2_{H^1(F),y} := \|u\|^2_{L_2(F),y} + \|\nabla u\|^2_{L_2(F),y}.$$

Proof. The proof follows from change of variables via

$$g : [0,1] \times F \to T : (s,x,y) \mapsto (x, sy, (1-s)y)$$

with $|\det g'| = y$, and thus

$$\int_T u(\xi,\eta,\zeta)^2 \, d(\xi,\eta,\zeta) = \int_0^1 \int_F y\, u(x,sy,(1-s)y)^2 \, d(x,y)\, ds. \qquad \square$$

Next we study trace and extension operators between the face F and the edge E. Continuity is proven with respect to the weighted norm $\|\cdot\|_{H^1(F),y}$, and a proper norm on the edge $\|\cdot\|_E$.

Expand $u \in P^p(F)$ as

$$u(x,y) = \sum_{i=2}^{p} L_i\Big(\frac{x}{1-y}\Big)(1-y)^i u_i(y) + x u_1(y) - u_0(y), \tag{21}$$

where $u_i \in P^{p-i}$. Utilize $L_i = \frac{1}{2i-1}(P_i - P_{i-2})$, define $u_i = 0$ for $i > p$, and shift indices

$$\begin{aligned} u(x,y) &= \sum_{i=2}^{p} \frac{1}{2i-1} P_i(\cdot)(1-y)^i u_i(y) - \sum_{i=2}^{p} \frac{1}{2i-1} P_{i-2}(\cdot)(1-y)^i u_i(y) \\ &\quad + x u_1(y) - u_0(y) \\ &= \sum_{i=2}^{p} P_i(\cdot)(1-y)^i \Big(\frac{u_i}{2i-1} - \frac{u_{i+2}(1-y)^2}{2i+3}\Big) \\ &\quad -\frac{1}{3}(1-y)^2 u_2 - \frac{1}{5} x (1-y)^2 u_3 + x u_1(y) - u_0(y) \\ &= \sum_{i=0}^{p} P_i\Big(\frac{x}{1-y}\Big)(1-y)^i \Big(\frac{u_i(y)}{2i-1} - \frac{(1-y)^2 u_{i+2}(y)}{2i+3}\Big). \end{aligned}$$

Thus, u can be re-expanded as

$$u(x,y) = \sum_{i=0}^{p} P_i\Big(\frac{x}{1-y}\Big)(1-y)^i v_i(y), \tag{22}$$

where the $v_i \in P^{p-i}$ are given as

$$v_i(y) = \frac{u_i(y)}{2i-1} - \frac{(1-y)^2 u_{i+2}(y)}{2i+3}. \tag{23}$$

Lemma 7. *Let $u \in P^p(F)$, and $u_i, v_i \in P^{p-i}$ be the expansion coefficients in (21) and (22). Then there holds*

$$\|u\|^2_{L_2(F),y} \simeq \sum_{i=0}^{p} \frac{1}{i+1} \int_0^1 y(1-y)^{2i+1} v_i(y)^2\, dy$$

and

$$\begin{aligned}\|\nabla u\|^2_{L_2(F),y} \simeq {} & \sum_{i=1}^{p} \frac{1}{i} \int_0^1 y(1-y)^{2i-1} u_i^2(y)\, dy \\ & + \sum_{i=0}^{p} \frac{1}{i+1} \int_0^1 y(1-y) \Big(\frac{d}{dy} \big((1-y)^i v_i(y) \big) \Big)^2 dy.\end{aligned}$$

Proof. Use the Duffy transform

$$g : [-1,1] \times [0,1] \to F : (\xi, y) \mapsto (x,y) = (\xi(1-y), y)$$

with $\det g' = (1-y)$ to transform the norm

$$\begin{aligned}\|u\|^2_{L_2(F),y} &= \int_F y u(x,y)^2\, d(x,y) = \int_0^1 \int_{-1}^1 y \det g'\, u(\xi,y)^2\, d\xi\, dy \\ &= \int_0^1 \int_{-1}^1 y(1-y) \Big(\sum_{i=0}^{p} P_i(\xi)(1-y)^i v_i(y) \Big)^2 d\xi dy \\ &= \sum_{i=0}^{p} \int_{-1}^1 P_i(\xi)^2\, d\xi \int_0^1 y(1-y)^{2i+1} v_i(y)^2\, dy.\end{aligned}$$

To transform the gradient-norm we calculate

$$(g')^{-\top} = \begin{pmatrix} \frac{1}{1-y} & 0 \\ \frac{\xi}{1-y} & 1 \end{pmatrix}$$

and note that

$$|(g')^{-\top} v|^2 \simeq (1-y)^{-2} v_1^2 + v_2^2 \quad \forall v \in \mathbb{R}^2.$$

Then we get

$$\|\nabla u\|^2_{L_2(F),y} = \int_0^1 \int_{-1+y}^{1-y} y\, |\nabla_{(x,y)} u|^2\, dx dy = \int_0^1 \int_{-1}^1 y \det g'\, |(g')^{-T} \nabla_{(\xi,y)} u|^2\, d\xi dy$$

$$\simeq \int_0^1 \int_{-1}^1 y(1-y)^{-1} \left| \frac{\partial u}{\partial \xi}(\xi,y) \right|^2 d\xi dy + \int_0^1 \int_{-1}^1 y(1-y) \left| \frac{\partial u}{\partial y}(\xi,y) \right|^2 d\xi dy.$$

For the first term we use representation (21):

$$\begin{aligned}
&\int_0^1 \int_{-1}^1 y(1-y)^{-1} \Big(\sum_{i=2}^p L_i'(\xi)(1-y)^i u_i(y) + u_1(y) \Big)^2 d\xi\, dy \\
&= \int_0^1 \int_{-1}^1 y(1-y)^{-1} \Big(\sum_{i=1}^p P_{i-1}(\xi)(1-y)^i u_i(y) \Big)^2 d\xi\, dy \\
&= \sum_{i=1}^p \int_{-1}^1 P_{i-1}(\xi)^2\, d\xi \int_0^1 y(1-y)^{2i-1} u_i(y)^2\, dy.
\end{aligned}$$

For the second term we use representation (22):

$$\begin{aligned}
&\int_0^1 \int_{-1}^1 y(1-y) \Big(\frac{\partial}{\partial y} \sum_{i=0}^p P_i(\xi)(1-y)^i v_i(y) \Big)^2 d\xi\, dy \\
&= \sum_{i=0}^p \int_{-1}^1 P_i(\xi)^2\, d\xi \int_0^1 y(1-y) \Big(\frac{d}{dy} \big((1-y)^i v_i(y)\big) \Big)^2 dy. \qquad \square
\end{aligned}$$

For

$$u(x) = \sum_{i=2}^p u_i L_i(x) + u_1 x - u_0 = \sum_{i=0}^p v_i P_i(x) \in P^p(E)$$

we define the norm

$$\|u\|_E^2 := \sum_{i=1}^p \frac{u_i^2}{ip^2(p-i+1)^2} + \sum_{i=0}^p \frac{v_i^2}{i+1}. \tag{24}$$

We note that

$$\sum_{i=0}^p \frac{v_i^2}{i+1} \simeq \|u\|_{L_2(E)}^2.$$

Numerical tests indicate that the first sum in (24) is bounded by $\log p\, \|u\|_{L_2(E)}^2$, and we decided to keep it in the definition of the norm $\|\cdot\|_E$ instead of loosing another log-factor.

Lemma 8 (Trace theorem on edges). *Let $u \in P^p(F)$. Then there holds*

$$\|u|_E\|_E^2 \preceq \log p\, \|u\|_{H^1(F),y}^2.$$

Proof. Follows immediately from the definition of $\|\cdot\|_E$, trace inequalities, Lemma 1 and Lemma 2, and the representation Lemma 7:

$$\begin{aligned}
\|u\|_E^2 &= \sum_{i=1}^{p} \frac{u_i^2(0)}{ip^2(p-i+1)^2} + \sum_{i=0}^{p} \frac{v_i^2(0)}{i+1} \\
&\preceq \sum_{i=1}^{p} \frac{1}{ip^2(p-i+1)^2} p^2(p-i+1)^2 \int_0^1 y(1-y)^{2i-1} u_i(y)^2\, dy \\
&\quad + \sum_{i=0}^{p} \frac{1}{1+i} \log p \int_0^1 y(1-y) \Big[\Big(\frac{d}{dy}\big((1-y)^i v_i(y)\big)\Big)^2 + \big((1-y)^i v_i(y)\big)^2\Big]\, dy \\
&\preceq \log p\, \|u\|_{H^1(F),y}^2.
\end{aligned}$$

□

Lemma 9 (Extension from edges). *For $u(x) = \sum_{i=2}^{p} u_i L_i(x) + u_1 x - u_0 \in P^p(E)$ and the functions e_i^p from Lemma 4 we define the extension operator as*

$$(\mathscr{E}_{E\to F} u)(x,y) := \sum_{i=2}^{p} u_i L_i\Big(\frac{x}{1-y}\Big)(1-y)^i e_i^p(y) + u_1 x e_1^p(y) - u_0 e_0^p(y).$$

Then here holds

$$\|\mathscr{E}_{E\to F} u\|_{H^1(F),y} \preceq \|u\|_E.$$

Proof. We convert the extended function into the Legendre basis as

$$\mathscr{E}_{E\to F} u = \sum_{i=0}^{p} P_i\Big(\frac{x}{1-y}\Big)(1-y)^i v_i(y),$$

where $v_i \in P^{p-i}$ are

$$v_i(y) = \frac{u_i e_i^p(y)}{2i-1} - \frac{u_{i+2}(1-y)^2 e_{i+2}^p(y)}{2i+3}.$$

We rewrite

$$\begin{aligned}
v_i(y) &= \Big(\frac{u_i}{2i-1} - \frac{u_{i+2}}{2i+3}\Big) e_i^p(y) + \big(e_i^p(y) - (1-y)^2 e_{i+2}^p(y)\big)\frac{u_{i+2}}{2i+3} \\
&= v_i(0) e_i^p(y) + \frac{u_{i+2}}{2i+3}\big(d_i^p(y) + (1-y) d_{i+1}^p(y)\big).
\end{aligned} \tag{25}$$

Note that there holds $u(x) = \sum_{i=0}^{p} v_i(0) P_i(x)$.

From Lemma 4 and Lemma 7 there follows

$$\begin{aligned}
&\|\mathscr{E}_{E\to F} u\|_{L_2(F),y}^2 \\
&\simeq \sum_{i=0}^{p} \frac{1}{i+1} \int_0^1 y(1-y)^{2i+1} \Big(e_i^p(y) v_i(0) + \big(d_i^p(y) + (1-y) d_{i+1}^p(y)\big)\frac{u_{i+2}}{2i+3}\Big)^2 dy
\end{aligned}$$

$$\preceq \sum_{i=0}^{p} \frac{v_i(0)^2}{i+1} \frac{1}{p^2(p-i+1)^2} + \sum_{i=0}^{p} \frac{u_{i+2}^2}{(i+1)^3} \frac{i^2}{p^3(p-i+1)^3} \preceq \|u\|_E^2$$

and

$$\|\nabla \mathscr{E}_{E\to F} u\|_{L_2(F),y}^2 = \sum_{i=1}^{p} \frac{u_i^2}{i} \int y(1-y)^{2i-1} e_i^p(y)^2\, dy$$

$$+ \sum_{i=0}^{p} \frac{1}{i+1} \int_0^1 y(1-y) \Big(\frac{d}{dy} \Big((1-y)^i \big(e_i^p(y) v_0$$

$$+ \big(d_i^p(y) + (1-y) d_{i+1}^p(y)\big) \frac{u_{i+2}}{2i+3} \big) \Big) \Big)^2 dy$$

$$\preceq \sum_{i=1}^{p} \frac{u_i^2}{i} \int y(1-y)^{2i-1} e_i^p(y)^2\, dy + \sum_{i=0}^{p} \frac{v_i^2}{i+1} \int y(1-y) \Big(\frac{d}{dy} \big((1-y)^i e_i^p(y) \big) \Big)^2 dy$$

$$+ \sum_{i=0}^{p} \frac{u_{i+2}^2}{(i+1)^3} \int y(1-y) \Big(\frac{d}{dy} \big((1-y)^i (d_i^p(y) + (1-y) d_{i+1}^p(y) \big) \Big)^2 dy.$$

Now we apply Lemma 4 to estimate

$$\|\nabla u\|_{L_2(F),y}^2 \preceq \sum_{i=1}^{p} \frac{u_i^2}{i} \frac{1}{p^2(p-i+1)^2} + \sum_{i=0}^{p} \frac{v_i^2(y)}{i+1} \simeq \|u\|_E^2. \quad \square$$

Next we estimate the contributions from the jump - norms. For this, we prove a face-to-edge trace lemma in weighted L_2-norms:

Lemma 10. *Let* $D = \{(y,z) : y \ge 0, z \ge 0, y+z \le 1\}$. *For* $v \in P^n(D)$ *there holds*

$$\int_0^1 y^\alpha (1-y)^\beta v(y,0)^2\, dy \preceq (n+1)(n+\alpha+\beta+1) \int_D y^\alpha (1-y-z)^\beta v(y,z)^2\, d(y,z).$$

Proof. We expand

$$v(y,z) = \sum_{j=0}^{n} P_j^{(\alpha,\beta)} \Big(2\frac{y}{1-z} - 1 \Big) (1-z)^j v_j(z)$$

with $v_j \in P^{n-j}$, and calculate

$$\int_0^1 y^\alpha (1-y)^\beta v(y,0)^2\, dy = \sum_{j=0}^{n} \int_0^1 y^\alpha (1-y)^\beta P_j^{(\alpha,\beta)} (2y-1)^2\, dy\, v_j(0)^2,$$

and with the change of variables $(y,z) = (\eta(1-z), z)$

$$\begin{aligned}
&\int_D y^\alpha (1-y-z)^\beta v(y,z)^2 \, d(y,z) \\
&= \int_0^1 \int_0^1 \eta^\alpha (1-\eta)^\beta (1-z)^{\alpha+\beta+1} v(\eta(1-z),z)^2 \, d\eta dz \\
&= \sum_{j=0}^p \int_0^1 \eta^\alpha (1-\eta)^\alpha P_j^{(\alpha,\beta)}(2\eta-1)^2 \, d\eta \int_0^1 (1-z)^{\alpha+\beta+1+2j} v_j(z)^2 \, dz.
\end{aligned}$$

The estimate follows with Lemma 1, i.e.

$$v_j(0)^2 \preceq (n-j+1)(n-j+\alpha+\beta+1+2j) \int_0^1 (1-z)^{\alpha+\beta+1+2j} v_j(z)^2 \, dz,$$

for $0 \leq j \leq n$. □

Lemma 11. *For $u \in P^p(E)$ there holds*

$$\|\mathscr{E}_{E\to F} u\|_{j,F} \preceq \|u\|_E.$$

Proof. By characterization (1) we have to prove the estimate

$$(\mathscr{E}_{E\to F} u, \sigma)_{L_2(F)} \preceq \|u\|_E \, \|\sigma\|_{L_2(T)} \qquad \forall u \in P^p(E), \forall \sigma \in P^p(T).$$

We recall

$$\mathscr{E}_{E\to F} u = \sum_{i=0}^p v_i(y) P_i\Big(\frac{x}{1-y}\Big)(1-y)^i$$

with (25)

$$v_i(y) = v_i(0) e_i^p(y) + \frac{u_{i+2}}{2i+3}(d_i^p(y) + (1-y) d_{i+1}^p).$$

We expand σ as

$$\sigma = \sum_{i=0}^p P_i\Big(\frac{x}{1-y-z}\Big)(1-y-z)^i \sigma_i(y,z)$$

with $\sigma_i \in P^{p-i}(D)$. By the change of variables $(x,y,z) = (\xi(1-y-z), y, z)$ we have

$$\begin{aligned}
\|\sigma\|_{L_2(T)}^2 &= \int_D \int_{-1+y+z}^{1-y-z} \sigma(x,y,z)^2 \, dx d(y,z) \\
&= \int_D \int_{-1}^1 (1-y-z) \sigma(\xi(1-y-z),y,z)^2 d\xi \, d(y,z) \\
&= \sum_i \|P_i\|_0^2 \int_D (1-y-z)^{2i+1} \sigma_i(y,z)^2 \, d(y,z).
\end{aligned}$$

We expand the inner product, use Lemma 4 and Lemma 10 to estimate

$$
\begin{aligned}
&(\mathscr{E}_{E\to F}u,\sigma)_{L_2(F)}\\
&=\int_0^1\int_{-1+y}^{1-y}\Big(\sum_{i=0}^{p}P_i\Big(\frac{x}{1-y}\Big)(1-y)^i v_i(y)\Big)\Big(\sum_{j=0}^{p}P_i\Big(\frac{x}{1-y}\Big)(1-y)^i\sigma_i(y,0)\Big)\,dxdy\\
&=\sum_{i=0}^{p}\|P_i\|^2\int_0^1(1-y)^{2i+1}v_i(y)\sigma_i(y,0)\,dy\\
&\le\sum_{i=0}^{p}\|P_i\|^2\Big(\int_0^1(1-y)^{2i+1}v_i(y)^2\,dy\Big)^{1/2}\Big(\int_0^1(1-y)^{2i+1}\sigma_i(y,0)^2\,dy\Big)^{1/2}\\
&\preceq\sum_{i=0}^{p}\|P_i\|^2\Big(\frac{v_i(0)^2}{p(p-i+1)}+\frac{u_i^2}{p^3(p-i+1)^3}\Big)^{1/2}\\
&\qquad\qquad\Big((p-i+1)p\int_D(1-y-z)^{2i+1}\sigma_i(y,z)^2\,d(y,z)\Big)^{1/2}\\
&\le\Big(\sum_{i=0}^{p}\|P_i\|^2\Big(v_i(0)^2+\frac{u_i^2}{p^2(p-i+1)^2}\Big)\Big)^{1/2}\\
&\qquad\qquad\Big(\sum_{i=0}^{p}\|P_i\|^2\int_D(1-y-z)^{2i+1}\sigma_i(y,z)^2\,d(y,z)\Big)^2\\
&\simeq\|u\|_E\,\|\sigma\|_{L_2(T)}.
\end{aligned}
$$

□

Finally we define the edge to element extension $\mathscr{E}_{E\to T}:P^p(E)\to P^p(T)$ as

$$\mathscr{E}_{E\to T}:=\mathscr{E}_{F\to T}\mathscr{E}_{E\to F}.$$

Theorem 7. *For $v\in P^p(T)$ define*

$$u:=\mathscr{E}_{E\to T}\,v|_E.$$

Then $u|_E=v|_E$ and there holds

$$\|u\|^2_{H^1(T)}+\sum_{F:E\subset F}\|u\|^2_{j,F}\preceq\log p\,\|v\|^2_{H^1(T)}.$$

If in addition v vanishes at the end-points of the edge E, then u vanishes on faces not containing E, and there holds

$$\|u\|^2_{H^1(T)}+\|u\|^2_{j,\partial T}\preceq\log p\,\|v\|^2_{H^1(T)}.$$

Proof. Follows from the construction of $\mathscr{E}_{F\to T}$ and $\mathscr{E}_{E\to F}$, and Lemmas 6, 8, 9, and 11. □

7 Numerical Results

In this section we give some computational results for different versions of stabilization terms. The first one is the facet-wise Bassi-Rebay stabilization as we have analyzed. The second one is an element-wise Bassi-Rebay stabilization where

$$\|u-\lambda\|_{j,\partial T} := \sup_{\sigma \in [P^p(T)]^3} \frac{\int_{\partial T}(u-\lambda)\sigma_n\,ds}{\|\sigma\|_{L_2(T)}}.$$

Here it is enough to choose the stabilization factor $\alpha > 1$. The norm is equivalent to the analyzed one (the proof is at some point tricky, and not given here). The developed theory carrys over. The third one is weighted L_2-stabilization with

$$\|u-\lambda\|_j^2 := \alpha\frac{p^2}{h}\|u-\lambda\|_{L_2(\partial T)}^2$$

Here, the choice of a sufficiently large α is not trivial.

We have chosen $\Omega = (0,1)^3$, and used Netgen to generated an unstructured mesh consisting of 184 tetrahedral elements. The condition numbers using a BDDC preconditioner are given in Table 1. Choosing $\alpha < 3$ for the method with L_2-stabilization does not lead to a coercive discrete problem.

It is clearly seen that the condition number depends on the stabilization term, and it is an advantage of having a method for which small stabilization factors are guaranteed to be stable. As we have proven, the condition numbers show a polylogarithmic growth for the BR-facet method. It is left the reader to interpret the numbers for L_2-stabilization, from our analysis there follows only $\kappa \preceq p(\log p)^\gamma$ due to norm equivalence (4).

Table 1 Condition numbers of the BDDC preconditioned system depending on p and the stabilization method.

pol deg	BR - facet	BR - element	L_2-stab	L_2-stab	L_2-stab	L_2-stab
	$\alpha = 5$	$\alpha = 1.5$	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 40$
2	24.91	10.62	12.91	23.74	45.50	88.96
4	41.41	18.64	23.62	41.19	75.63	144.65
8	59.44	33.16	42.27	67.20	116.47	214.49
16	80.70	54.78	65.97	94.73	152.47	268.62

References

[1] Abramowitz, M., Stegun, I.A.: Handbook of mathematical functions. John Wiley & Sons (1993)

[2] Ainsworth, M.: A preconditioner based on domain decomposition for *h-p* finite element approximation on quasi-uniform meshes. SIAM J. Numer. Anal. 33, 1358–1376 (1996)

[3] Antonietti, P.-F., Houston, P.: A class of domain decomposition preconditioners for hp-discontinuous Galerkin finite element methods. J. Sci. Comput. 46, 124–149 (2011)

[4] Arnold, D.N., Brezzi, F., Cockburn, B., Marini, D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. SIAM J. Numer. Anal. 39(5), 1749–1779 (2002)

[5] Babuška, I., Craig, A.W., Mandel, J., Pitkäranta, J.: Efficient preconditioning for the p version of the finite element method in $\mathbb{R}^2$. SIAM J. Numer. Anal. 28, 624–661 (1991)

[6] Bassi, F., Rebay, S.: High-order accurate discontinuous finite element solution of the 2D Euler equations. J. Comp. Phys. 138, 251–285 (1997)

[7] Beuchler, S., Schneider, R., Schwab, C.: Multiresolution weighted norm equivalences and applications. Numer. Math. 98, 67–97 (2004)

[8] Bică, I.: Iterative substructuring algorithms for the *p*-version finite element method for elliptic problems. PhD thesis. Courant Institute of Mathematical Sciences, New York University (1997)

[9] Casarin, M.: Quasi-optimal Schwarz methods for the conforming spectral element discretization. SIAM J. Numer. Anal. 34, 2482–2502 (1997)

[10] Cockburn, B., Gopalakrishnan, J., Lazarov, R.: Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. SIAM J. Numer. Anal. 47, 1319–1365 (2009)

[11] Cockburn, B., Kanschat, G., Schötzau, D.: A note on discontinuous Galerkin divergence-free solutions of the Navier-Stokes equations. J. Sci. Comput. 31, 61–73 (2007)

[12] Cockburn, B., Karniadakis, G.E., Shu, C.W.: Discontinuous Galerkin Methods: Theory, Computation and Applications. Springer (2000)

[13] Demkowicz, L.: Computing with hp-adaptive finite elements. One and two dimensional elliptic and Maxwell problems, vol. 1. Chapman & Hall/CRC (2007)

[14] Dohrmann, C.R.: A preconditioner for substructuring based on constrained energy minimization. SIAM J. Sci. Comput. 25(1), 246–258 (2003)

[15] Dryja, M., Widlund, O.B.: Towards a unified theory of domain decomposition algorithms for elliptic problems. In: Chan, T.F., Glowinski, R., Périaux, J., Widlund, O.B. (eds.) Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, pp. 3–21. SIAM, Philadelphia (1990)

[16] Dubiner, M.: Spectral methods on triangles and other domains. J. Sci. Comput. 6(4), 345–390 (1991)

[17] Georgoulis, E.H., Süli, E.: Optimal error estimates for the *hp*-version interior penalty discontinuous Galerkin finite element method. IMA J. Numer. Anal. 25, 205–220 (2005)

[18] Gopalakrishnan, J., Kanschat, G.: A multilevel discontinuous Galerkin method. Numer. Math. 95(3), 527–550 (2003)

[19] Guo, B., Cao, W.: An additive Schwarz method for the *h-p* version of the finite element method in three dimensions. SIAM J. Numer. Anal. 35, 632–654 (1998)

[20] Griebel, M., Oswald, P.: On the abstract theory of additive and multiplicative Schwarz algorithms. Numer. Math. 70, 163–180 (1995)

[21] Haase, G., Langer, U., Meyer, A.: The approximate Dirichlet domain decomposition method. Part I: An algebraic approach. Part II: Applications to 2nd-order elliptic boundary value problems. Computing 47, 137–151, 153–167 (1991)

[22] Heuer, N., Leydecker, F.: An extension theorem for polynomials on triangles. Calcolo 45(2), 69–85 (2008)

[23] Heuer, N., Leydecker, F., Stephan, E.P.: An iterative substructuring method for the hp-version of the BEM on quasi-uniform triangular meshes. Numer. Methods Partial Differential Eq. 23(4), 879–903 (2007)

[24] Hesthaven, J.S., Warburton, T.: Nodal Discontinuous Galerkin Methods—Algorithms, Analysis and Applications. Text in Applied Mathematics. Springer (2007)

[25] Juntunen, M., Stenberg, R.: On a mixed discontinuous Galerkin method. ETNA 32, 17–32 (2008)

[26] Karniadakis, G.E., Sherwin, S.J.: Spectral/hp Element Methods for Computational Fluid Dynamics. Oxford Science Publications (2005)

[27] Klawonn, A., Widlund, O.B., Dryja, M.: Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients. SIAM J. Numer. Anal. 40(1), 159–179 (2002)

[28] Korneev, V.G., Jensen, S.: Domain decomposition preconditioning in the hierarchical *p*-version of the finite element method. Appl. Numer. Math. 29, 479–518 (1999)

[29] Korneev, V.G., Langer, U.: Domain Decomposition and Preconditioning. In: Stein, E., de Borst, R., Hughes, T.J.R. (eds.) Encyclopedia of Computational Mechanics, Part I, ch. 19, 44 p. John Wiley & Sons (2004)

[30] Korneev, V.G., Langer, U., Xanthis, L.: On fast domain decomposition solving procedures for hp-discretizations of 3D elliptic problems. Comput. Meth. Appl. Math. 3(4), 536–559 (2003)

[31] Lehrenfeld, C.: Hybrid discontinuous Galerkin methods for solving incompressible flow problems. Master thesis, RWTH Aachen (2010)

[32] Lions, P.–L.: On the Schwarz alternating method I. In: First International Symposium on Domain Decomposition Methods for Partial Differential Equations, Paris, 1987, pp. 1–42. SIAM, Philadelphia (1988)

[33] Li, J., Widlund, O.B.: FETI-DP, BDDC, and block Cholesky methods. Int. J. Numer. Meth. Engrg. 66(2), 250–271 (2006)

[34] Muñoz-Sola, R.: Polynomial liftings on a tetrahedron and applications to the *hp*-version of the finite element method in three dimensions. SIAM J. Numer. Anal. 34(1), 282–314 (1997)

[35] Paule, P., Schorn, M.: A Mathematica version of Zeilberger's algorithm for proving binomial coefficient identities. J. Symbolic Comput. 20, 673–698 (1995)

[36] Pavarino, L.: Additive Schwarz methods for the *p*-version finite element method. Numer. Math. 66, 493–515 (1994)

[37] Pavarino, L.: BDDC and FETI-DP preconditioners for spectral element discretizations. Comp. Meth. Appl. Mech. Engrg. 196, 1380–1388 (2007)

[38] Pavarino, L.F., Widlund, O.B.: A polylogarithmic bound for an iterative substructuring method for spectral elements in three dimensions. SIAM J. Numer. Anal. 33(4), 1303–1335 (1996)

[39] Pillwein, V.: Computer Algebra Tools for Special Functions in High Order Finite Element Methods. PhD thesis, Johannes Kepler University Linz (2008)

[40] Schöberl, J., Melenk, J.M., Pechstein, C., Zaglmayr, S.: Additive Schwarz preconditioning for p-version triangular and tetrahedral finite elements. IMA J. Numer. Anal. 28, 1–24 (2008)

[41] Schwab, C.: *p*- and *hp*-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics. Oxford Science Publications (1998)

[42] Sherwin, S.J., Casarin, M.: Low energy basis preconditioning for elliptic substructured solvers based on unstructured spectral/hp element discretisations. J. Comput. Phys. 171, 394–417 (2001)

[43] Szabó, B., Babuška, I.: Finite Element Analysis. Wiley (1991)

[44] Szegö, G.: Orthogonal Polynomials, vol. 23. AMS Colloquium Publications (1939) (reprinted 2003)

[45] Toselli, A., Widlund, O.B.: Domain Decomposition Methods - Algorithms and Theory. Springer Series in Computational Mathematics, vol. 34 (2005)

Fast Domain Decomposition Algorithms for Elliptic Problems with Piecewise Variable Orthotropism

Vadim G. Korneev

Abstract. Second order elliptic equations are considered in the unit square, which is decomposed into subdomains by an arbitrary nonuniform orthogonal grid. For the elliptic operator we assume that the energy integral contains only squares of first order derivatives with coefficients, which are arbitrary positive finite numbers but different for each subdomain. The orthogonal finite element mesh has to satisfy only one condition: it is uniform on each subdomain. No other conditions on the coefficients of the elliptic equation and on the step sizes of the discretization and decomposition are imposed. For the resulting discrete finite element problem, we suggest domain decomposition algorithms of linear total arithmetical complexity, not depending on any of the three factors contributing to the orthotropism of the discretization on subdomains. The main problem of designing such an algorithm is the preconditioning of the inter-subdomain Schur complement, which is related in part to obtaining boundary norms for discrete harmonic functions on the shape irregular domains.

1 Introduction

The aim of this paper is to present fast DD (domain decomposition) algorithms for the discretization of partial differential equations with piecewise variable orthotropism, which is modelled by the discrete problem described below. Suppose, the domain $\Omega = (0,1) \times (0,1)$ is decomposed into subdomains

$$\Omega_j = (z_{1,j_1-1}, z_{1,j_1}) \times (z_{2,j_2-1}, z_{2,j_2}), \quad j = (j_1, j_2), \tag{1}$$

Vadim G. Korneev
St. Petersburg State University, St. Petersburg State Polytechnical University, Russia
e-mail: VadimKorneev@yahoo.com

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 57–89.
springerlink.com

by the rectangular *decomposition grid*

$$x_k = z_{k,j_k}, \quad j_k = 0,1,..,J_k, \quad z_{k,j_k} - z_{k,j_k-1} = H_{k,j_k} > 0, \quad z_{k,0} = 0, \quad z_{k,J_k} = 1\,. \quad (2)$$

The decomposition grid is imbedded in the nonuniform rectangular finer *source grid*

$$x_k = x_{k,i_k}, \quad i_k = 0,1,..,N_k, \quad x_{k,0} = 0, \quad x_{k,N_k} = 1, \qquad (3)$$

i.e. $x_{k,\gamma_k} = z_{k,j_k}$ for some numbers $\gamma_k = \varkappa_k(j_k)$, $k = 1,2$. Assume for simplicity, that this grid is uniform on each subdomain and has sizes $h_{k,j_k} = H_{k,j_k}/n_{k,j_k}$, where n_{k,j_k} is the number of the source grid intervals on the decomposition grid interval (z_{k,j_k-1}, z_{k,j_k}).

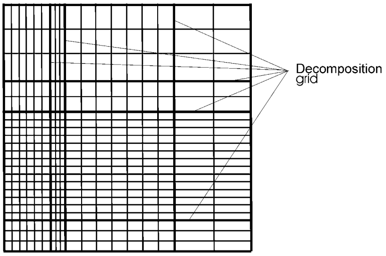


Fig. 1 Decomposition grid and subdomain wise uniform rectangular source grid.

Let $\mathring{\mathscr{V}}(\Omega)$ be the FE (finite element) space of globally continuous functions which are bilinear on each nest of the source grid and which vanish on $\partial\Omega$. We consider the problem

$$\alpha_\Omega(u,v) = \langle f, v\rangle, \quad \forall\, v \in \mathring{H}^1(\Omega), \text{ where } \alpha_\Omega(u,v) = \int_\Omega \nabla u(x)\cdot \wp(x)\,\nabla v(x)\,dx, \quad (4)$$

$\wp(x)$ is a 2×2 matrix satisfying the inequalities

$$\mu_1\rho(x) \le \wp(x) \le \mu_2\rho(x), \quad 0 \le \mu_1, \mu_2 = \text{const}, \qquad (5)$$

and $\rho = \operatorname{diag}[\rho_1, \rho_2]$ is a diagonal matrix with piecewise constant positive functions $\rho_k(x)$. In other words, $\rho_k(x) = \rho_{k,j} = \text{const} > 0$ for $x \in \Omega_j$, and $\rho_{k,j}$ are arbitrary positive numbers. The integral identity (4) on the space $\mathring{\mathscr{V}}(\Omega)$ is reduced to the system of linear algebraic equations

$$\mathbf{K}\mathbf{u} = \mathbf{f}\,. \qquad (6)$$

The answer, we are looking for, is whether there exists a DD algorithm, which is robust and fast uniformly for arbitrary positive H_{k,j_k}, $\rho_{k,j}$ and h_{k,j_k}. Indeed, we suggest DD preconditioners $\mathscr{K}_{\mathrm{DD}}$ of the Dirichlet-Dirichlet type such that

$$\operatorname{ops}\left[\mathscr{K}_{\mathrm{DD}}^{-1}\mathbf{f}\right] \times \left(\kappa\left[\mathscr{K}_{\mathrm{DD}}^{-1}\mathbf{K}\right]\right)^{1/2} \le cN_{\Omega}, \quad c = \mathrm{const}, \tag{7}$$

where N_{Ω} is the number of unknowns in (6), $\kappa[\mathbb{A}]$ is the spectral condition number of the matrix $\mathbb{A}$, ops$[\ldots]$ is the number of arithmetical operations for performance of the operation inside the brackets, and c is an absolute constant. Therefore, the bound (7) approves that the DD preconditioner provides a solution procedure for the system (6) of a linear total arithmetical complexity. The bound retains, if the number of subdomains $J = J_1 J_2$ grows along with the number of FE unknowns, but not too fast, for instance, when $J_k \le N_k^{1/2} / \log^{3/2} \overline{N}$, $\overline{N} = \max N_k$. It is worth to stress that the right part of (7) does not depend on any of the three factors contributing to the orthotropism of the discretization on subdomains.

The results are retained, if Ω is the union of any number of nests of the decomposition mesh, they are also retained for the respective FE discretizations by linear triangular finite elements with vertices at the nodes of the orthogonal discretization mesh. Moreover, for both types of discretizations, the preconditioners will be often defined by means of triangular elements, which provide simpler explicit representations.

In the DD algorithm, Dirichlet problems in rectangular subdomains Ω_j can be efficiently solved by numerous direct and iterative methods, including FDFT (Fast Discrete Fourier Transform). Solvers for orthotropic and some more complex types of discretizations on rectangular domains have been intensively studied. We can find respective results in Schieweck [47], Wittum [50], Dahmen [17], Griebel & Oswald [22], Grauschopf et al. [21], Cohen et al. [14], Schneider [48], Oswald [44], Pflaum [45], and many other papers. These works allow also to obtain for each subdomain optimal low energy prolongation operators and spectrally equivalent Schur complement preconditioners which are almost optimal for inversion. Therefore, at designing DD algorithm subordinate to the bound (7), the key problem is obtaining an interface preconditioner, which would not compromise this bound.

An analysis of DD interface preconditioners for isotropic elliptic equations in domains, composed of thin rectangles, can be found, e.g., in Chen et al. [13] and Nepomnyaschikh [40, 42]. In relation with some deteriorating elliptic equations, DD algorithms for discretizations with a subdomain-wise more general variation of orthotropism were studied analytically in Korneev [31] and numerically in Rytov [46] and Anufriev & Korneev [4] for 2-d and 3-d problems, respectively. Other techniques, e.g., boundary element methods, $\mathscr{H}$-matrices, and tensor-train decompositions were also attracted for obtaining efficient interface preconditioners for elliptic problems with orthotropism with a subdomain-wise variation which is subjected to some restrictions. Papers of Hsiao et al. [26], Hackbusch et al. [25], Dolgov et al. [16] are only a few representatives of this vast area of research.

The case of an assemblage of rectangular subdomains, having different arbitrary aspect ratios, accompanied by an orthotropism of the differential equation which is chaotically strongly changing from subdomain to subdomain, causes specific difficulties. They are strengthened by a jumping orthotropism of a rectangular subdomain-wise uniform, but otherwise arbitrary, finer mesh for the discretization. The results for solvers of uniformly anisotropic problems on arbitrary rectangle or corresponding Schur complement solvers are not directly applicable. The reason is that multilevel decompositions, which are efficient for each subdomain separately, are not compatible, and, therefore, in general can't be assembled to obtain an efficient interface preconditioner.

Discrete problems close to (1)-(6) with $\wp(x) = \rho(x)$ were independently addressed in Khoromskij & Wittum [27, 28], in Kwak et al. [38], Nepomnyaschikh [43], and in Korneev et al. [33, 34]. In particular [27, 28] concentrated on the interface solvers, whereas the others considered DD Dirichlet-Dirichlet solvers. There are other differences, pertaining specific components of algorithms and techniques of their analysis, which resulted in different bounds for the relative condition number of the DD preconditioner and its total computational complexity. The bounds of [27, 28, 38] depend on $\nu_{\rho,j}$, $\nu_{\rho,j}^2 = \max_{k=1,2}(\rho_{k,j}/\rho_{3-k,j})$, whereas the bounds of [33, 34] do not. In this paper, we improve the preconditioner of [33, 34] and come to estimates of linear complexity.

Compatibility of subdomain Schur complement preconditioners usually included a splitting of the vertex degrees of freedom from the rest. However, for thin rectangles this damages the relative condition number more severely, than for shape regular rectangles. One way to circumvent these obstacles is the use of an iterative preconditioner $\mathscr{S}_{1,\mathrm{it}}$, resulting from an inexact solver defined by means of two Schur complement preconditioners $\mathscr{S}_k$, $k = 1,2$. One is aimed to provide a good relative condition number and cheap matrix-vector multiplications. The other may have a not so good relative condition number, but is cheap for inversion. Seemingly, the idea of such a preconditioning was introduced by Nepomnyaschikh [40]; in Kwak et al. [38] it was implemented in the DD solver for a model problem, close to (1)-(6), but with a different, than in this paper, choice of $\mathscr{S}_k$, $k = 1,2$.

The preconditioner $\mathscr{S}_1^j$, used in this paper for one subdomain $\Omega = \Omega_j$, stems from the shape dependent boundary seminorm, which is equivalent to the $H^1(\Omega)$-seminorm for discrete harmonic functions in shape irregular rectangles. The corresponding norm was introduced in Korneev [31], see also Korneev et al. [33, 34], following the technique of Maz'ya & Poborchi [39] used for harmonic functions. The validity of this norm for discrete harmonic functions is established with the use of a Scott & Zhang [49] result on a special interpolation operator for functions from $H^1(\Omega)$. Then the shape dependent norm is simplified by means of finite-difference norms, equivalent to the $H^{1/2}$-norms on some 1-d sets. It is, by the way, worth noting that it well reflects the fact, known in applications, e.g., of boundary FE methods and referred as *absorption of singularities*. Suppose, we have a spectrally equivalent preconditioner for the boundary Schur complement for a FE discretization on a quasiuniform mesh. Then, this preconditioner retains the spectral equivalence to the Schur complement generated on any shape regular mesh, which is imbedded in the quasiuniform mesh and coincides with it on the boundary. Even more general meshes can be considered. Basic facts related to the preconditioner are presented in Subsect. 2.1 and 2.2.

There are obvious reasons to use Schur complement preconditioners, in which, apart from splitting vertices, each edge is split at least from a part of others. In the DD algorithm of this paper, $\mathscr{S}_2$ can be taken as one of the slightly different and compatible preconditioners used in Khoromskij & Wittum [27, 28], Korneev [31], and Korneev et al. [33, 34]. The proof of the bounds of its relative spectrum, independent of $\nu_{\rho,j}$, is made in a way of comparison with the sample preconditioner, which has a structure similar to $\mathscr{S}_2$, but which is more suitable for an analysis. The purpose of Subsect. 2.3 and Sect. 3 is to describe the sample preconditioner and to prepare subsidiary results for the analysis of the basic preconditioner $\mathscr{S}_2$. The derivation of the sample preconditioner for one subdomain is accomplished by means of a secondary DD technique with a secondary shape regular nonoverlapping domain decomposition. The preconditioner $\mathscr{S}_2$ is presented in Sect. 4, Sect. 5 summarizes results for separate subdomains, obtained in preceding sections, into a bound of computational cost of a DD method for the problem (1)-(6).

We do not study solvers for the subproblem of the DD algorithm, which is related to the vertices of the subdomains. Often its dimension is much smaller than N_Ω, and at varying J_k, we assume that there is a solver, which does not compromise (7). Obviously, if $J_k \le N_k^{1/3}$, then even a direct elimination procedure will satisfy this assumption.

Let us list some notations as used in this paper. For matrices we primarily use capital letters of the styles $\mathbf{A}$, $\mathbb{A}$, $\mathscr{A}$, $\mathbf{I}$ stands for identity matrices, small boldface letters – for vectors. $(\cdot,\cdot)_\Omega$, and $\|\cdot\|_{0,\Omega}$ are the scalar product and the norm in $L^2(\Omega)$, whereas $|\cdot|_{k,\Omega}$, $\|\cdot\|_{k,\Omega}$ are the semi-norm and the norm in the Sobolev space $H^k(\Omega)$, i.e.,

$$|v|_{k,\Omega}^2 = \sum_{|q|=k} \int_\Omega (D_x^q v)^2 dx, \qquad \|v\|_{k,\Omega}^2 = \|v\|_{0,\Omega}^2 + \sum_{l=1}^{k} |v|_{l,\Omega}^2,$$

with

$$D_x^q v := \partial^{|q|} v/\partial x_1^{q_1} \partial x_2^{q_2}, \quad q = (q_1, q_2), \quad q_1, q_2 \ge 0, \quad |q| = q_1 + q_2.$$

$\mathring{H}^1(\Omega)$ is the subspace of $H^1(\Omega)$ of functions having zero traces on $\partial\Omega$. For $I = (a,b)$, $\|\cdot\|_{1/2,I}$ and ${}_{00}\|\cdot\|_{1/2,I}$ are the norms in the space $H^{1/2}(I)$ and the subspace ${}_{00}H^{1/2}(I) \subset H^{1/2}(I)$ of functions having zero values at $x = a, b$, see, e.g., [1]. Expressions for these norms are

$$\|v\|_{1/2,I}^2 = \|v\|_{0,I}^2 + |v|_{1/2,I}^2, \quad |v|_{1/2,I}^2 = \int_a^b \int_a^b \left(\frac{v(x) - v(y)}{x - y} \right)^2 dx dy,$$

$${}_{00}\|v\|_{1/2,I}^2 = \|v\|_{1/2,I}^2 + \int_a^b \frac{v^2(x)}{x-a} dx + \int_a^b \frac{v^2(x)}{b-x} dx.$$

The norm $\|\cdot\|_{1/2,\gamma_i}$, when γ_i is an edge of $\sqcap = (0,1) \times (0,1)$ is defined for the traces on this edge analogously with $\|\cdot\|_{1/2,I}$. For instance, for the edge γ_i, which is on the line $x_1 = c$, $c = 0, 1$, we have

$$\|v\|^2_{1/2,\gamma_i}=\|v\|^2_{0,\gamma_i}+|v|^2_{1/2,\gamma_i},\quad |v|^2_{1/2,\gamma_i}=\int\limits_0^1\int\limits_0^1\left(\frac{v(c,t)-v(c,\tau)}{t-\tau}\right)^2 dtd\tau.$$

For a sufficiently smooth and shape regular domain Ω, the norm $\|v\|_{1/2,\partial\Omega}$ is given by the formulas

$$\|v\|^2_{1/2,\partial\Omega}=\|v\|^2_{0,\partial\Omega}+|v|^2_{1/2,\partial\Omega},\quad |v|^2_{1/2,\partial\Omega}=\int\limits_{\partial\Omega}\int\limits_{\partial\Omega}\left(\frac{v(x)-v(y)}{x-y}\right)^2 ds_x ds_y,$$

in which ds_x, ds_y are the length elements at the points $x,y\in\partial\Omega$. In the case, e.g., $\Omega=\sqcap:=(0,1)\times(0,1)$ this norm is equivalent to the norm

$$\|v\|^2_{1/2,\partial\sqcap}=\|v\|^2_{0,\partial\sqcap}+|v|^2_{1/2,\partial\sqcap} \tag{8}$$

with

$$|v|^2_{1/2,\partial\sqcap}=\sum_{i=1}^4|v|^2_{1/2,\gamma_i}+\sum_{i=1}^4\int\limits_0^1\frac{\left(v_{j(i)}(t)-v_{l(i)}(t)\right)^2}{|t|}dt,$$

where $u_{j(i)}$ denotes the restriction of u onto the edge $\gamma_{j(i)}$, and t is the distance to the vertex V_i of $\sqcap$, which is common for $\gamma_{j(i)}$ and $\gamma_{l(i)}$. To each V_i, we associate a preceding edge $\gamma_{j(i)}$ and a succeeding edge $\gamma_{l(i)}$, e.g., according to a counter-clockwise orientation of the boundary. The norm and semi-norm defined in this way for the space $H^{1/2}(\partial\sqcap)$ are equivalent to $\|v\|_{1/2,\partial\sqcap}:=\inf\|w\|_{1,\sqcap}$ and $|v|_{1/2,\partial\sqcap}:=\inf|w|_{1,\sqcap}$ with infima taken over $w\in H^1(\sqcap)$ for which $w=v$ on $\partial\sqcap$. We refer to Grisvard [23], Ben Belgacem [7] and [39] for additional details on the introduced boundary norms.

$\mathbf{A}^+$ is the pseudo-inverse to a matrix $\mathbf{A}$, $\|\mathbf{v}\|_{\mathbf{A}}=(\mathbf{v}^\top\mathbf{A}\mathbf{v})^{1/2}$ is the norm or the seminorm, induced by a nonnegative symmetric matrix $\mathbf{A}$. If $\mathbf{A}$ is a nonnegative symmetric matrix, the notation $\mathbf{A}^{1/2}$ stands for the nonnegative symmetric matrix $\mathbf{B}$ satisfying $\mathbf{A}=\mathbf{B}\mathbf{B}$, $\ker[\mathbf{A}]=\ker[\mathbf{B}]$. The spectral condition number of a matrix $\mathbb{A}$ is denoted $\kappa[\mathbb{A}]$, ops$[\cdot]$ is the number of arithmetic operations needed for the procedure in the square brackets. Symbols $\prec$, $\succ$ denote one-sided and $\asymp$ – two-sided inequalities, which hold for some, mostly absolute, constants omitted, whereas $\mathbf{A}\prec\mathbf{B}$ with nonnegative matrices $\mathbf{A}$, $\mathbf{B}$ implies $\mathbf{v}^\top\mathbf{A}\mathbf{v}\prec\mathbf{v}^\top\mathbf{B}\mathbf{v}$ for any vector $\mathbf{v}$, and similarly for signs $\succ$, $\asymp$. We write $\mathbf{v}\Leftrightarrow v$, if the vector $\mathbf{v}$ represents the FE function v in a chosen basis. Whenever we write "inversion of matrix $\mathbb{A}$" or $\mathbb{A}^{-1}\mathbf{y}$, we imply solving of the system $\mathbb{A}\mathbf{x}=\mathbf{y}$.

We avoid the use of special notations for perturbed matrices and matrices expanded by zero entries. Accordingly, sums of matrices are typically understood as topological sums, and a $n\times n$ matrix $\mathbf{A}$, initially defined for some n, is considered, when necessary, as expanded by zero entries up to a $m\times m$, $m>n$, matrix without special explanations.

2 Single Thin Rectangle

2.1 *Discrete Analogues of Boundary Norms for Harmonic Functions in Thin Rectangles*

Let $\Omega = (0,1) \times (0,\varepsilon)$ and ε, δ satisfy $0 < \varepsilon,\ \delta \leq 1$. For the traces of functions $v \in H^1(\Omega)$ on $\partial\Omega$, we consider two norms and two seminorms. Norm and seminorm of one pair, denoted by $\|\cdot\|$ and $|\cdot|$, minimize the H^1-norm and H^1-seminorm, respectively, among all functions $\phi \in H^1(\Omega)$ coinciding with a given function on the boundary:

$$\|v\|^2_{\partial\Omega} = \inf_{\phi_{|\partial\Omega}=v} \left((\delta\varepsilon^{-1})^2\|\phi\|^2_{0,\Omega} + \|\nabla\phi\|^2_{0,\Omega}\right), \quad |v|^2_{\partial\Omega} = \inf_{\phi_{|\partial\Omega}=v} \|\nabla\phi\|^2_{0,\Omega}\,. \qquad (9)$$

For another norm and seminorm we use notations $]|\cdot|[_{\partial\Omega}$ and $]\cdot[_{\partial\Omega}$ and introduce them by the expressions

$$]|v|[^2_{\partial\Omega} = \delta^2\varepsilon^{-1}\|v\|^2_{0,\partial\Omega} +]v[^2_{\partial\Omega} \qquad (10)$$

with

$$\begin{aligned}]v[^2_{\partial\Omega} = {} & \varepsilon^{-1}\int_0^1 (v(x_1,\varepsilon) - v(x_1,0))^2 dx_1 + \int_0^1 \int_{|x_1-y_1|\leq\varepsilon} \frac{(v(x_1,0) - v(y_1,0))^2}{(x_1-y_1)^2} dx_1 dy_1 \\ & + \int_0^1 \int_{|x_1-y_1|\leq\varepsilon} \frac{(v(x_1,\varepsilon) - v(y_1,\varepsilon))^2}{(x_1-y_1)^2} dx_1 dy_1 + \int_{\Gamma_0}\int_{\Gamma_0} \frac{(v(s) - v(\bar{s}))^2}{(s-\bar{s})^2} ds d\bar{s} \\ & + \int_{\Gamma_1}\int_{\Gamma_1} \frac{(v(s) - v(\bar{s}))^2}{(s-\bar{s})^2} ds d\bar{s}. \end{aligned}$$

Here $\Gamma_0 = \{x \in \partial\Omega : x_1 < \varepsilon\}$, $\Gamma_1 = \{x \in \partial\Omega_\varepsilon : x_1 > 1-\varepsilon\}$ and ds, $d\bar{s}$ are the length elements of $\partial\Omega$. The set Γ_1 is symmetric to Γ_0 with respect to the line $x_1 \equiv 1/2$, see Fig.2.

Theorem 1. *For the traces of functions from $H^1(\Omega)$, the norms (9), (10) are equivalent uniformly in $\varepsilon, \delta \in (0,1]$.*

Proof. The proof can be found in Korneev et al. [33, 34]. □

We will call $]|\cdot|[_{\partial\Omega}$ and $]\cdot[_{\partial\Omega}$ the *shape dependent norm and seminorm* for boundary functions.

In this paper, discretizations on rectangular meshes are considered. Accordingly, we use the FE space $\mathscr{V}(\Omega)$ of piecewise bilinear functions on the rectangular mesh $x_k \equiv x_{k,l}$ with the steps $h_{k,l} = x_{k,l} - x_{k,l-1}$, $l = 1,2,..,n_k$, satisfying the quasiuniformity conditions

$$\underline{c}h \le h_{k,l} \le \overline{c}h, \quad 0 < \underline{c}, \overline{c} = \mathrm{const}, \tag{11}$$

and $x_{k,0} = 0$, $x_{1,n_1} = 1$, $x_{2,n_2} = \varepsilon$.

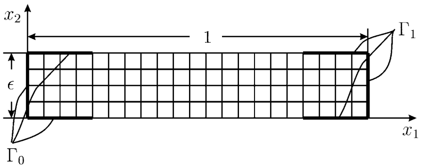


Fig. 2 High aspect ratio rectangular domain triangulated by a square mesh.

By $\mathscr{V}_{\mathrm{tr}}(\partial\Omega)$ we denote the space of traces of functions from $\mathscr{V}(\Omega)$ on $\partial\Omega$. However, most of the results hold for much more general discretizations. The mesh as described above may represent a *skeleton mesh*, while calculations are performed on a mesh, called the *source or fine mesh*, which

α) *is finer only in the interior of the domain,*
β) *has the same trace with the skeleton mesh on the boundary and*
γ) *covers the skeleton mesh, whereas the skeleton mesh itself may be a general quasiuniform quadrangular unstructured mesh with the mesh parameter h.*

For simplicity, it is convenient also to assume that there are mesh nodes at the ends of the sets Γ_k, $k = 0, 1$, and that the number of nodes on the opposite edges of Ω are equal, as in the case of an orthogonal mesh. Clearly, the skeleton and the source meshes can be as well triangular meshes, from which the former is quasiuniform.

For the traces of FE functions $v \in \mathscr{V}_{\mathrm{tr}}(\partial\Omega)$, the simpler norm than (9)

$$\begin{aligned} \|v\|^2_{h,\partial\Omega} &= \inf_{\phi \in \mathscr{V}(\Omega):\phi_{|\partial\Omega} = v} \left((\delta\varepsilon^{-1})^2 \|\phi\|^2_{0,\Omega} + \|\nabla\phi\|^2_{0,\Omega}\right), \\ |v|^2_{h,\partial\Omega} &= \inf_{\phi \in \mathscr{V}(\Omega):\phi_{|\partial\Omega} = v} \|\nabla\phi\|^2_{0,\Omega}, \end{aligned} \tag{12}$$

can be justified, in which inf is taken only over the subspace of FE functions.

Theorem 2. *Let the FE space $\mathscr{V}(\Omega)$ be induced by the quasiuniform triangulation with the mesh parameter h or by its refinement satisfying α)-γ). Then for any $h > 0$ and $v \in \mathscr{V}_{tr}(\partial\Omega)$, the norms and seminorms (12), (10), respectively, are equivalent uniformly in $\varepsilon, \delta \in (0,1]$.*

Proof. The proof is based, first, on Theorem 1, and, second, on the quasi-interpolation result of Lemma 2, given after the proof of the theorem.

Since $\mathscr{V}(\Omega) \subset H^1(\Omega)$, one has the inequalities

$$]|v|[_{\partial\Omega} \prec \|v\|_{\partial\Omega} \le \|v\|_{h,\partial\Omega} \quad \forall v \in \mathscr{V}_{\mathrm{tr}}(\partial\Omega), \tag{13}$$

with the first one following from Theorem 1 and the definitions of the norms $\|\cdot\|_{\partial\Omega}$ and $\|\cdot\|_{h,\partial\Omega}$. For the proof of the opposite bound

$$\|v\|_{h,\partial\Omega} \prec]|v|[_{\partial\Omega}, \quad \forall v \in \mathscr{V}_{\mathrm{tr}}(\partial\Omega) \tag{14}$$

it is sufficient to use in addition Lemma 2.

Indeed, let $\mathscr{H}(\Omega)$ be the subspace of $\mathscr{V}(\Omega)$, induced by the skeleton quasiuniform triangulation, and $v \in \mathscr{V}_{\mathrm{tr}}(\partial\Omega)$. Suppose also that $v_{\mathrm{inf}} \in H^1(\Omega)$ and $v_{d/\mathrm{inf}} \in \mathscr{V}(\Omega)$ are the functions on which the inf's in the first relationships of (9) and (12), respectively, are reached. Let also $\widetilde{v}$ be the interpolation of v_{inf} from the space $\mathscr{H}(\Omega)$, satisfying i) and ii) of Lemma 2. First of all, we note that according to Lemma 2

$$(\delta\varepsilon^{-1})^2\|\widetilde{v}\|^2_{0,\Omega} + \|\nabla\widetilde{v}\|^2_{0,\Omega} \prec (\delta\varepsilon^{-1})^2\|v_{\mathrm{inf}}\|^2_{0,\Omega} + \|\nabla v_{\mathrm{inf}}\|^2_{0,\Omega}. \tag{15}$$

Therefore, we can write

$$\begin{aligned} \|v\|^2_{h,\partial\Omega} &:= (\delta\varepsilon^{-1})^2\|v_{d/\mathrm{inf}}\|^2_{0,\Omega} + \|\nabla v_{d/\mathrm{inf}}\|^2_{0,\Omega} \\ &\prec (\delta\varepsilon^{-1})^2\|\widetilde{v}\|^2_{0,\Omega} + \|\nabla\widetilde{v}\|^2_{0,\Omega} \\ &\prec (\delta\varepsilon^{-1})^2\|v_{\mathrm{inf}}\|^2_{0,\Omega} + \|\nabla v_{\mathrm{inf}}\|^2_{0,\Omega} \prec]|v|[^2_{\partial\Omega}, \end{aligned} \tag{16}$$

where the first inequality follows by the definition of $\|v\|^2_{h,\partial\Omega}$, the second inequality – by the definition of the same norm and the inclusion $\mathscr{H}(\Omega) \subset \mathscr{V}(\Omega)$, the third inequality is simply (15), and the last one is a consequence of Theorem 1.

For the seminorms the proof is similar. □

Lemma 2, used above, is practically a corollary of a result of Scott & Zhang [49] on a special quasi-interpolation operator, which we present first.

Let $\Omega \subset \mathbb{R}^n$ be a n-dimensional domain with an arbitrary quasiuniform triangulation $\mathscr{S}_h$ with nodal points $x^{(i)}$, $i = 1,2,\dots,I$, and maximal edge size h. To each node $x^{(i)}$, we relate the $(n-1)$-dimensional simplex τ_i, which is the face of one of the n-dimensional simplices of the triangulation $\mathscr{S}_h$ having the vertex $x^{(i)}$. For n vertices of the simplex τ_i, we also use the notations $z_l^{(i)}$, $l = 1,2,\dots,n$, assuming for definiteness that $z_1^{(i)} = x^{(i)}$. The choice of τ_i is not unique, but for $x^{(i)} \in \partial\Omega$ we always take $\tau_i \subset \partial\Omega$. By $\mathscr{V}_\Delta(\Omega)$ and $\mathscr{V}_{\mathrm{tr}}(\partial\Omega)$ we denote the space of functions, which are continuous on $\overline{\Omega}$ and linear on each simplex of the triangulation, and the space of their traces on $\partial\Omega$, respectively. Let $\theta_i \in \mathscr{P}(\tau_i)$ be the function satisfying

$$\int\limits_{\tau_i} \theta_i \lambda_l^{(i)}\, dx = \delta_{1,l}, \quad l = 1,2,..,n,$$

where $\lambda_l^{(i)}$ are the barycentric coordinates in τ_i related to its vertices $z_l^{(i)}$, and $\delta_{i,l}$ is the Kronecker symbol. If $\phi_i \in \mathscr{V}_\Delta(\Omega)$ are Galerkin FE basis functions such that $\phi_i(x_j) = \delta_{i,j}$, $i,j = 1,2,\dots,I$, then for each $v \in H^1(\Omega)$ the quasi-interpolation $\mathscr{I}_h v \in \mathscr{V}_\Delta(\Omega)$ is defined as

$$\mathscr{I}_h v = \sum_{i=1}^{I} \left(\int_{\tau_i} \theta_i v \, dx \right) \phi_i(x).$$

A triangulation $\mathscr{T}_h$ by simplices is called quasiuniform with respect to some mesh parameter $h > 0$ in the usual sense, see Ciarlet [15] and Korneev [29]. In two dimensions, quasiuniformity of quadrangular meshes is controlled by the conditions that lengths of edges and angles at vertices of quadrangles belong to intervals $(\alpha^{(1)}h, h)$ and $(\theta, \pi - \theta)$, respectively, with $0 < \alpha^{(1)}$, $\theta = \text{const}$.

Lemma 1. *The quasi-interpolation operator $\mathscr{I}_h$ satisfies*

a) $\mathscr{I}_h v : H^1(\Omega) \mapsto \mathscr{V}_\Delta(\Omega)$*, and, if* $v \in \mathscr{V}_\Delta(\Omega)$*, then* $\mathscr{I}_h v = v$*,*
b) $(v - \mathscr{I}_h v) \in \mathring{H}^1(\Omega)$*, if* $v_{|\partial\Omega} \in \mathscr{V}_{\text{tr}}(\partial\Omega)$*,*
c) $\|v - \mathscr{I}_h v\|_{t,\Omega} \le c_{\text{int}} h^{s-t} \|v\|_{s,\Omega}$ *for* $t = 0, 1$*, and* $s = 1, 2$*,*
d) $|\mathscr{I}_h v|_{1,\Omega} \le c_{\text{int}} |v|_{1,\Omega}$ *and* $\|\mathscr{I}_h v\|_{1,\Omega} \le c_{\text{int}} \|v\|_{1,\Omega}$ *for all* $v \in H^1(\Omega)$*, where* c_{int} *is a constant, depending only on* $\alpha^{(1)}$*,* θ *from the quasiuniformity conditions.*

Proof. The proof was given by Scott & Zhang [49] and can be found also in Xu & Zou [51]. □

The operator $\mathscr{I}_h$, obviously, is a projection onto the space $v \in \mathscr{V}_\Delta(\Omega)$.

Now we will formulate for $n = 2$ the interpolation result used in the proof. Suppose, $\mathscr{V}(\Omega)$ is the FE space induced by first order quadrangular finite elements. We define a triangulation of Ω by subdividing each quadrangular nest of the mesh in two triangles by one of the diagonals of the nest and then denote by $\mathscr{V}_\Delta(\Omega)$ the space of continuous piecewise linear functions, induced by the obtained triangulation. For each $v \in H^1(\Omega)$, we define $\Pi_h v$ by the equality $(\Pi_h v)(x^{(i)}) = (\mathscr{I}_h v)(x^{(i)})$ for all vertices $x^{(i)}$ of the triangulation. It is easy to note that Π_h is not a projection operator, but it retains some other useful properties of the operator $\mathscr{I}_h$.

Lemma 2. *For any* $v \in H^1(\Omega)$ *the interpolation* $\Pi_h v \in \mathscr{V}(\Omega)$ *is such that*

i) if $v_{|\partial\Omega} \in \mathscr{V}_{\text{tr}}(\partial\Omega)$*, then the traces of* $\Pi_h v$ *and* v *on the boundary* $\partial\Omega$ *coincide,*
ii) the interpolation satisfies the stability estimates

$$|\Pi_h v|_{1,\Omega} \prec |v|_{1,\Omega}, \quad \|\Pi_h v\|_{1,\Omega} \prec \|v\|_{1,\Omega}, \tag{17}$$

iii) and approximation estimates

$$\|v - \Pi_h v\|_{t,\Omega} \prec h^{s-t} \|v\|_{s,\Omega}, \quad t = 0, 1, \; s = 1, 2. \tag{18}$$

Proof. We omit the proof, which in part can be found in Korneev et al. [33, 34]. □

Obviously, the norm (10) defines a matrix $\mathbb{B}_{\text{KPS}}$ such that $\|\mathbf{v}\|_{\mathbb{B}_{\text{KPS}}} =]v[_{\partial\Omega}, \forall \mathbf{v} \Leftrightarrow v \in \mathscr{V}_{\text{tr}}(\partial\Omega)$. Let the FE matrix $\mathbf{A}$ be induced by the Dirichlet integral over Ω and the FE space $\mathscr{V}(\Omega)$ with the nodal basis functions. Representing it in the block form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_I & \mathbf{A}_{I,B} \\ \mathbf{A}_{B,I} & \mathbf{A}_B \end{pmatrix}, \tag{19}$$

we denote by **B** the Schur complement

$$\mathbf{B} = \mathbf{A}_B - \mathbf{A}_{B,I}\mathbf{A}_I^{-1}\mathbf{A}_{I,B}, \tag{20}$$

where the lower indices I and B are related to the degrees of freedom at the nodes, living in the interior of the domain Ω and on its boundary, respectively.

Corollary 1. $\mathbb{B}_{\text{KPS}} \prec \mathbf{B} \prec \mathbb{B}_{\text{KPS}}$ *uniformly in h.*

Thus, $\mathbb{B}_{\text{KPS}}$ can be used as a spectrally equivalent preconditioner for the Schur complement **B**. In the next subsection, we introduce a simpler boundary seminorm for discrete harmonic functions, which is more convenient for matrix vector multiplications.

2.2 Finite-Difference Shape Dependent Boundary Norm for Finite Element Functions

Let us turn to a simpler case of the skeleton mesh, whose trace on $\partial\Omega$ coincides with the trace of an auxiliary orthogonal quasiuniform grid satisfying (11). We denote by γ_k for $k = 0,1$ the left and the right vertical edges of Ω, and by γ_k, $k = 2,3$, the horizontal lower and upper edges, respectively.

The auxiliary *coarse (quasiuniform) grid*, by its definition, is the coarsest rectangular imbedded quasiuniform grid. It has rectangular nests as much as possible close to the square $\varepsilon \times \varepsilon$ and is obtained by subdividing the domain Ω by vertical lines $x_1 = t_{1,i}$ in such a way that $t_{1,0} = 0$, $t_{1,n_\varepsilon} = 1$ for some integer $n_\varepsilon \geq 1$ and sizes $\eta_{1,i} := t_{1,i} - t_{1,i-1}$, $i = 1,2,\ldots,n_\varepsilon$, satisfy the inequalities

$$\varepsilon \leq \eta_{1,i} \leq \overline{c}_\circ \varepsilon, \tag{21}$$

with $\overline{c}_\circ = \text{const} \leq 2$. The notation $\mathcal{V}_{\text{c}}(\Omega)$ will stand for the space of functions which are continuous on Ω and bilinear on each nest of the coarse quasiuniform mesh. This space will be called the *coarse finite element space*.

Simultaneously, we have defined overlapping intervals

$$\tau_0 = (0, t_{1,1}), \quad \tau_i = (t_{1,i-1}, t_{1,i+1}), \quad i = 1,2,\ldots,n_\varepsilon - 1, \quad \tau_{n_\varepsilon} = (t_{1,n_\varepsilon - 1}, 1),$$

and intersections $\gamma_{k,i}$ of τ_i, $i = 1,2,\ldots,n_\varepsilon - 1$, with the edges γ_k, $k = 2,3$. We use the notations $\gamma_0 = \Gamma_0$, $\gamma_{n_\varepsilon} = \Gamma_1$ and for simplicity we assume that $\overline{\tau}_0 \cap \partial\Omega = \overline{\Gamma}_0$ and $\overline{\tau}_{n_\varepsilon} \cap \partial\Omega = \overline{\Gamma}_1$ and that the numbers of nodes on these sets are the same and equal to ν.

Let

$$\Delta_{1/2,k} = \left(\begin{array}{cccccc} 1 & -1 & & & & \\ -1 & 2 & -1 & & \mathbb{O} & \\ & & \cdot\ \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot \\ & \mathbb{O} & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{array} \right)^{1/2}, \quad k = 0,1, \tag{22}$$

be $\nu \times \nu$ matrices, acting on vectors of degrees of freedom corresponding to the nodes on $\overline{\gamma}_k$. If ν_i is the number of the nodes on $\overline{\tau}_i$, then $\Delta_{1/2,k,i}$ is the, similar to (22), $\nu_i \times \nu_i$ matrix related to the segment $\overline{\gamma}_{k,i}$, $k = 2,3$. In the $(n_1+1)\times(n_1+1)$ matrix

$$\nabla = \frac{h}{\varepsilon} \begin{pmatrix} \mathbf{I} & -\mathbf{I} \\ -\mathbf{I} & \mathbf{I} \end{pmatrix},$$

the unity matrices $\mathbf{I}$ on the diagonal correspond to the nodes on the edges $\overline{\gamma}_2, \overline{\gamma}_3$, respectively.

Let us remind that, as it was stated in the introduction, matrices $\Delta_{1/2,}$, $\Delta_{1/2,k,i}$ and ∇ are considered as defined on degrees of freedom at the nodes of the sets $\overline{\Gamma}_0$, $\overline{\Gamma}_1$, $\overline{\gamma}_{2,i}$, $\overline{\gamma}_{3,i}$ and $\overline{\gamma}_2$, $\overline{\gamma}_3$, respectively, and continued by zeroes on the degrees of freedom of the remaining nodes of the boundary, when necessary. This implies that sums like (23) below should be understood as *topological* sums.

Lemma 3. *The matrix*

$$\mathbf{C} = \nabla + \Delta_{1/2,0} + \Delta_{1/2,1} + \sum_{k=2,3} \sum_{i=1}^{n_\varepsilon - 1} \Delta_{1/2,k,i} \tag{23}$$

is spectrally equivalent to the matrix $\mathbb{B}_{\mathrm{KPS}}$ *uniformly in* h *and* $\varepsilon \in (0,1]$. *Besides, for any vector* $\mathbf{v}_B$ *the arithmetical costs of the matrix-vector multiplication* $\mathbf{C}\mathbf{v}_B$ *are* $ops\,[\mathbf{C}\mathbf{v}_B] = \mathscr{O}((n_1+n_2)(1+\log n_2))$.

Proof. We outline the proof, omitting details, which is completed in two steps. First, we introduce the seminorm $\rfloor \cdot \lfloor_{\partial\Omega}$ by the expression

$$\begin{aligned} \rfloor v \lfloor^2_{\partial\Omega} = {} & \varepsilon^{-1} \int_0^1 (v(x_1,\varepsilon) - v(x_1,0))^2 dx_1 \\ & + \sum_{i=1}^{n_\varepsilon - 1} \int_{\tau_i} \int_{\tau_i} \left[\frac{(v(x_1,0) - v(y_1,0))^2}{(x_1-y_1)^2} + \frac{(v(x_1,\varepsilon) - v(y_1,\varepsilon))^2}{(x_1-y_1)^2}) \right] dx_1 dy_1 \\ & + \int_{\Gamma_0} \int_{\Gamma_0} \frac{(v(s) - v(\overline{s}))^2}{(s-\overline{s})^2} ds d\overline{s} + \int_{\Gamma_1} \int_{\Gamma_1} \frac{(v(s) - v(\overline{s}))^2}{(s-\overline{s})^2} ds d\overline{s} \end{aligned} \tag{24}$$

and show that it is equivalent to the seminorm $]v[_{\partial\Omega}$, i.e.,

$$\underline{\gamma}_0 \rfloor v \lfloor_{\partial\Omega} \leq]v[_{\partial\Omega} \leq 3 \rfloor v \lfloor_{\partial\Omega}, \quad 0 < \underline{\gamma}_0 = \underline{\gamma}_0(\overline{c}) = \mathrm{const}, \quad \forall v \in H^1(\Omega). \tag{25}$$

Result of the second step are the inequalities

$$\underline{\gamma}_C \mathbf{v}^\top \mathbf{C}\mathbf{v} \leq \rfloor v \lfloor^2_{\partial\Omega} \leq \overline{\gamma}_C \mathbf{v}^\top \mathbf{C}\mathbf{v}, \quad \forall \mathbf{v} \Leftrightarrow v \in \mathscr{V}(\Omega), \tag{26}$$

which hold with some constants $\underline{\gamma}_C$, $\overline{\gamma}_C > 0$. These bounds follow from the equivalences

$$\varepsilon^{-1} \int\limits_0^1 (v(x_1,\varepsilon) - v(x_1,0))^2 dx_1 \asymp \mathbf{v}^\top \nabla \mathbf{v} \quad \forall \mathbf{v} \Leftrightarrow v \in \mathscr{V}_{\mathrm{tr}}(\gamma_2 \cup \gamma_3),$$

$$\int\limits_{\tau_i} \int\limits_{\tau_i} \frac{(v(x_1,0) - v(y_1,0))^2}{(x_1 - y_1)^2} dx_1 dy_1 \asymp \mathbf{v}^\top \Delta_{1/2,2,i} \mathbf{v}, \quad \forall \mathbf{v} \Leftrightarrow v \in \mathscr{V}_{\mathrm{tr}}(\gamma_2),$$

$$\int\limits_{\tau_i} \int\limits_{\tau_i} \frac{(v(x_1,\varepsilon) - v(y_1,\varepsilon))^2}{(x_1 - y_1)^2} dx_1 dy_1 \asymp \mathbf{v}^\top \Delta_{1/2,3,i} \mathbf{v}, \quad \forall \mathbf{v} \Leftrightarrow v \in \mathscr{V}_{\mathrm{tr}}(\gamma_3),$$

$$\int\limits_{\Gamma_k} \int\limits_{\Gamma_k} \frac{(v(s) - v(\overline{s}))^2}{(s - \overline{s})^2} ds d\overline{s} \asymp \mathbf{v}^\top \Delta_{1/2,k} \mathbf{v}, \quad \forall \mathbf{v} \Leftrightarrow v \in \mathscr{V}_{\mathrm{tr}}(\Gamma_k),$$

where $\mathscr{V}_{\mathrm{tr}}(\gamma_2 \cup \gamma_3)$, $\mathscr{V}_{\mathrm{tr}}(\gamma_2)$, $\mathscr{V}_{\mathrm{tr}}(\gamma_3)$ and $\mathscr{V}_{\mathrm{tr}}(\Gamma_k)$, $k = 0,1$, are the trace spaces of the FE space on the corresponding subsets of the boundary. Bounds of the first line follow by the spectral equivalence of the FE 1-d mass matrix to its diagonal. Lines 1-3 express another known fact. Suppose, some interval τ is subdivided by a quasiuniform grid with ν intervals and $\mathscr{H}(\tau)$ is the corresponding space of continuous piecewise linear functions. Then the matrix of the quadratic form $|v|^2_{1/2,\tau}$ on the space $\mathscr{H}(\tau)$ is spectrally equivalent to the matrix $\Delta_{1/2}$ of the form (22). We found an early proof of this fact in Andreev [2, 3].

Combining (25) and (26), one comes to

$$\underline{\beta}_C \mathbf{C} \prec \mathbb{B}_{\mathrm{KPS}} \prec \overline{\beta}_C \mathbf{C} \tag{27}$$

with positive constants $\underline{\beta}_C$, $\overline{\beta}_C > 0$ depending only on the constants from the quasiuniformity conditions (11).

The matrix-vector multiplication $\nabla \mathbf{v}$ requires $6n_2 + 1$ arithmetical operations. The matrix-vector multiplications by each matrix $\Delta_{1/2,k}$ or $\Delta_{1/2,k,i}$ can be completed by FDFT and requires $\mathscr{O}(n_2 \log n_2)$ arithmetical operations. Hence, the vector-matrix multiplication by the topological sum of these matrices requires in total $\mathscr{O}((n_1 + n_2)(1 + \log n_2))$ arithmetical operations. This approves the estimate of the arithmetic work for the multiplication $\mathbf{C}\mathbf{v}$ given in the Lemma. □

By Theorem 2, the definition of the matrix $\mathbb{B}_{\partial\Omega}$, and Lemma 3, it follows that:

Corollary 2. *For the matrix* $\mathbf{C}$ *and the Schur complement* $\mathbf{B}$ *of (20), it holds* $\mathbf{B} \asymp \mathbf{C}$.

Similar to (23) we can define preconditioners in the case of orthotropic discretizations, e.g., by the mesh, which is quasiuniform in each direction and has characteristic sizes h_1, h_2. We cover such a discretization mesh by a finer mesh, called *condensed mesh*, which has nests as close as possible in the shape to the square $h \times h$, $h = \min(h_1, h_2)$ and define the matrix (23) for this mesh, which we denote $\mathbf{C}_{\text{cond}}$. After that, we restrict this matrix to the set of nodes corresponding to the space $\mathscr{V}_{\text{tr}}(\partial\Omega)$ denoting the new matrix $\mathbf{C}_{\text{ff}}$. At that time, we represent it in the form

$$\mathbf{C}_{\text{ff}} = \nabla + \Delta^{(0)}_{1/2,\text{ff}} + \Delta^{(1)}_{1/2,\text{ff}} + \sum_{k=2,3} \sum_{i=1}^{n_\varepsilon - 1} \Delta_{1/2,k,i}, \tag{28}$$

where the matrices ∇, $\Delta_{1/2,k,i}$ are defined as above on the discretization mesh, whereas $\Delta^{(k)}_{1/2,\text{ff}}$, $k = 1,2$, are defined as the respective matrices $\Delta_{1/2,k,i}$, but for the condensed mesh with understanding that added nodes are treated as hanging nodes.

Corollary 3. *Let the discretization mesh be quasiuniform in each direction with the characteristic sizes h_1, h_2, $\mathbf{B}$ be the Schur complement, generated by the corresponding FE space $\mathscr{V}(\Omega)$. Then $\mathbf{B} \asymp \mathbf{C}_{\text{ff}}$ and*

$$ops\,[\mathbf{C}_{\text{ff}}\mathbf{v}] \prec (n_1 + n_2)\log\max(n_2, n_1/n_\varepsilon) \prec \overline{n}\log\overline{n}, \quad \overline{n} = \max(n_1, n_2), \quad \forall \mathbf{v},$$

uniformly in $h_1 \in (0,1)$, $h_2 \in (0,\varepsilon)$.

Note that for matrix vector multiplications by $\mathbf{C}_{\text{ff}}$, the FDFT can be used and that for $h_1 \ll h_2$ or $h_2 \ll h_1$ we have $\dim\Delta_{1/2,k,i} \leq \dim\Delta^{(l)}_{1/2,\text{ff}} \approx 3\max(n_2, n_1/n_\varepsilon)$, $l = 0,1$.

2.3 Preconditioning by Nonoverlapping Domain Decomposition

In this paragraph we consider a thin rectangle and derive a preconditioner for the boundary Schur complement by an implementation of the DD procedure with nonoverlapping subdomains. It allows to split degrees of freedom of each vertical edge and degrees of freedom of the pair of longest edges from other degrees of freedom. Then we additionally split the degrees of freedom at the vertices of the rectangle from all other ones.

The coarse grid, introduced in the preceding subsection, defines a nonoverlapping domain decomposition of the rectangle Ω into subdomains

$$\Omega^i = (t_{1,i-1}, t_{1,i}) \times (0,\varepsilon), \quad i = 1,2,\ldots,n_\varepsilon.$$

For their edges, we use the notations Γ^i_k, $k = 0,1,2,3$, and order them counterclockwise starting from the lower edge of Ω^i. The FE space can be represented by the direct sum

$$\mathscr{V}(\Omega) = \mathscr{V}_{\text{c}}(\Omega) \oplus \mathscr{W}_{\text{r}}(\Omega), \tag{29}$$

where $\mathscr{V}_{\text{c}}(\Omega)$ is the space of continuous functions which are bilinear on each subdomain Ω^i. The second subspace in (29) is supplied by the index ”r”, because in what follows it will play the role of the subspace, induced by the rarefied mesh. Notations $\mathscr{V}_{\text{c}}(\Omega^i)$ and $\mathscr{W}_{\text{r}}(\Omega^i)$ will stand for restrictions to Ω^i of the spaces $\mathscr{V}_{\text{c}}(\Omega)$ and $\mathscr{W}_{\text{r}}(\Omega)$.

We consider any $v = (v_{\mathrm{c}} + v_W) \in \mathscr{V}(\Omega^i)$, such that $v_0 \in \mathscr{V}_{\mathrm{c}}(\Omega^i)$ and $v_W \in \mathscr{W}_{\mathrm{r}}(\Omega^i)$ is discrete harmonic on Ω^i. Subdomains Ω^i are shape regular, and according to a result of Bramble et al. [9]–[12]

$$\frac{1}{(1+\log n_2)^2}\left(|v_{\mathrm{c}}|^2_{1,\Omega^i} + \sum_{k=0}^{3} {}_{00}|v_W|^2_{1/2,\Gamma^i_k}\right) \prec |v|^2_{1,\Omega^i} \prec |v_{\mathrm{c}}|^2_{1,\Omega^i} + \sum_{k=0}^{3} {}_{00}|v_W|^2_{1/2,\Gamma^i_k}. \tag{30}$$

From here, for a FE function $v \in \mathscr{V}(\Omega)$ which is discrete harmonic in Ω, we directly come to the inequalities

$$\begin{aligned}&\frac{1}{(1+\log n_2)^2}\left(|v_{\mathrm{c}}|^2_{1,\Omega} + {}_{00}|v_W|^2_{1/2,\Gamma^1_3} + {}_{00}|v_W|^2_{1/2,\Gamma^{n\varepsilon}_1} + \sum_{i=1}^{n_\varepsilon}\sum_{k=0,2} {}_{00}|v_W|^2_{1/2,\Gamma^i_k}\right)\\ &\prec |v|^2_{1,\Omega} \prec |v_{\mathrm{c}}|^2_{1,\Omega} + {}_{00}|v_W|^2_{1/2,\Gamma^1_3} + {}_{00}|v_W|^2_{1/2,\Gamma^{n\varepsilon}_1} + \sum_{i=1}^{n_\varepsilon}\sum_{k=0,2} {}_{00}|v_W|^2_{1/2,\Gamma^i_k}.\end{aligned} \tag{31}$$

Let $\mathscr{B}_W$ be the matrix which is spectrally equivalent to the matrix of the quadratic form (31) on the subspace $\mathscr{W}(\Omega)$, i.e.,

$$\mathbf{v}_W^\top \mathscr{B}_W \mathbf{v}_W \prec {}_{00}|v_W|^2_{1/2,\Gamma^1_3} + {}_{00}|v_W|^2_{1/2,\Gamma^{n\varepsilon}_k} + \sum_{i=1}^{n_\varepsilon}\sum_{k=0,2} {}_{00}|v_W|^2_{1/2,\Gamma^i_k} \prec \mathbf{v}_W^\top \mathscr{B}_W \mathbf{v}_W. \tag{32}$$

Then (31) is equivalent to the inequalities

$$\frac{1}{(1+\log n_2)^2}\mathbf{v}^\top \mathbb{C}_{\mathrm{hi}} \mathbf{v} \prec \mathbf{v}^\top \mathbf{B}_{\mathrm{hi}} \mathbf{v} \prec \mathbf{v}^\top \mathbb{C}_{\mathrm{hi}} \mathbf{v}, \tag{33}$$

where

$$\mathbb{C}_{\mathrm{hi}} = \begin{pmatrix} \mathscr{B}_W & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{\mathrm{c}} \end{pmatrix}$$

and the notation $\mathbf{B}_{\mathrm{hi}}$ for the Schur complement reflects that it is written in the two level basis, corresponding to the representation $\mathscr{V}(\Omega) = \mathscr{V}_{\mathrm{c}}(\Omega) \oplus \mathscr{W}_{\mathrm{r}}(\Omega)$ of the FE space. The matrix $\mathbf{B}_{\mathrm{c}}$ is the block, corresponding to the subspace $\mathscr{V}_{\mathrm{c}}(\Omega)$.

Let the notation $\mathbf{B}_{\mathrm{Hi}}$ stand for the Schur complement corresponding to the three level representation of the FE space

$$\mathscr{V}(\Omega) = \mathscr{W}_{\mathrm{r}}(\Omega) \oplus \mathscr{W}_{\mathrm{c}}(\Omega) \oplus \mathscr{V}_0(\Omega),$$

and let $\mathbf{B}^{(\mathrm{w})}$ and $\mathbf{B}^{(\mathrm{v})}$ be the notations for the blocks on the diagonal of $\mathbf{B}_{\mathrm{Hi}}$, corresponding to the subspaces $\mathscr{W}_{\mathrm{c}}(\Omega)$ and $\mathscr{V}_0(\Omega)$, respectively, and $\underline{n} = \min(n_1, n_2)$, $\overline{n} = \max(n_1, n_2)$. With a slightly changing reasoning, one also can get

$$\min\left(\frac{1}{n_\varepsilon(1+\log \underline{n})}, \frac{1}{(1+\log \underline{n})^2}\right)\mathbf{v}^\top \mathscr{C}_{\mathrm{Hi}} \mathbf{v} \prec \mathbf{v}^\top \mathbf{B}_{\mathrm{Hi}} \mathbf{v} \prec \mathbf{v}^\top \mathscr{C}_{\mathrm{Hi}} \mathbf{v}, \tag{34}$$

where

$$\mathscr{C}_{\mathrm{Hi}} = \begin{pmatrix} \mathscr{B}_{\mathrm{W}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{(\mathrm{w})} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}^{(\mathrm{v})} \end{pmatrix}. \tag{35}$$

We reorder the sets $\Gamma_k^i \subset \partial\Omega$ consecutively counter clockwise, starting from Γ_0^1, introduce for them the notations $\mathscr{T}^i$, $i = 1,2,\dots,2n_\varepsilon + 2$, and by ν_i the number of the intervals of the source mesh on $\mathscr{T}^i$. Estimates (32) and therefore (33), (34) hold for

$$\mathscr{B}_{\mathrm{W}} = \mathrm{diag}[\Delta_{1/2,i}]_{i=1}^{2(n_\varepsilon+1)} \tag{36}$$

with $\Delta_{1/2,i}^{1/2} = \mathrm{tridiag}\,[-1,2,-1]_1^{\nu_i-1}$. Thus, we have proved:

Lemma 4. *With $\mathscr{B}_{\mathrm{W}}$ defined in (36), the estimates (34) hold for all $\varepsilon \in (0,1]$ and $h \le (0,\varepsilon]$.*

The preconditioner $\mathscr{C}_{\mathrm{hi}}$ is sufficiently simple. In particular, it is given explicitly, is easily invertible, and, therefore, can be used for assembling the Schur complement preconditioner for a domain which is decomposed into rectangular subdomains. The system of algebraic equations with the preconditioner $\mathscr{C}_{\mathrm{hi}}$ of Lemma 4 for the matrix is solved in $\mathscr{O}(n_\varepsilon \underline{n} \log \underline{n}) = \mathscr{O}(\overline{n} \log \underline{n})$ arithmetical operations, where for the case under consideration $\overline{n} = n_1$, $\underline{n} = n_2$. However, the subspaces $\mathscr{V}_0$ and $\mathscr{W}_{\mathrm{r}}$ depend on the aspect ratio of the rectangle Ω, and, therefore, even the assembling procedure of the interface Schur complement preconditioner can be not simple, not to speak about its inversion.

3 Orthotropic Discretization with Arbitrary Aspect Ratio on Thin Rectangles

3.1 Finite Element Space Decomposition

A more complicated situation arises, when we consider a heat conduction problem in a slim domain with different heat conduction coefficients along different axes, and a uniform rectangular mesh is used for discretization. No restrictions are imposed on the aspect ratios of conductivity coefficients and sizes of the mesh, except that they are finite. Therefore, the model problem, we turn here to, is

$$\alpha_\Omega(u,v) = \langle f,v\rangle, \quad \alpha_\Omega(u,v) = \int\limits_\Omega \nabla u(x)\cdot\rho(x)\nabla v(x)\,dx, \quad \forall v \in H^1(\Omega), \tag{37}$$

in a slim rectangle $\Omega = (0,1)\times(0,\varepsilon)$. Now, $\rho = \mathrm{diag}\,[\rho_1,\rho_2]$ with an arbitrary constant $\rho_k > 0$. For simplicity, we restrict ourselves to a uniform rectangular mesh of arbitrary sizes $h_1, h_2 > 0$.

For the ease of future references, we use, different from the previously used notations, $\mathbf{Q}$, $\mathbf{Y}$ for the FE stiffness matrix and its boundary Schur complement

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_I & \mathbf{Q}_{I,B} \\ \mathbf{Q}_{B,I} & \mathbf{Q}_B \end{pmatrix}, \quad \mathbf{Y} = \mathbf{Q}_B - \mathbf{Q}_{B,I}\mathbf{Q}_I^{-1}\mathbf{Q}_{I,B}. \tag{38}$$

The derivation of a good preconditioner for the Schur complement will be completed in three steps. At step 1), we change variables and reduce the problem (37) to a transformed isotropic problem on some domain Ω_ξ. At step 2), we introduce the rarefied transformed mesh, which is the finest and, if possible, quasiuniform mesh imbedded in the transformed source mesh. It is obtained by rarefication of the transformed source mesh in one direction, corresponding to the smallest size of the latter mesh. Then the block diagonal preconditioner for the FE matrix $\mathbf{Q}$ is introduced, containing two independent blocks on the diagonal, one of which is the FE stiffness matrix, induced by the rarefied transformed mesh. In turn, the preconditioner for the FE matrix $\mathbf{Q}$ allows us to obtain the block diagonal preconditioner for the Schur complement $\mathbf{Y}$ with two independent blocks. At step 3), a further decoupling is accomplished. The block of the Schur complement preconditioner, corresponding to the transformed rarefied mesh, obviously, can be handled as in the preceding section. Another block does not require an additional treatment, because it itself is a block diagonal matrix with simple explicitly written down blocks, specified on the unknowns, subjected to rarefication.

The domain Ω represents one subdomain $\Omega = \Omega_j$ of the decomposition. The described process is based on the sequence of the FE spaces, which is related to the sequence of the image spaces

$$\mathbb{V} = \mathbb{V}_{\mathrm{r}} \oplus \mathbb{W}, \quad \mathbb{V}_{\mathrm{r}} = \mathbb{V}_{\mathrm{c}} \oplus \mathbb{W}_{\mathrm{r}}, \quad \mathbb{V}_{\mathrm{c}} = \mathbb{V}_0 \oplus \mathbb{W}_{\mathrm{c}}, \tag{39}$$

with

$$\mathbb{V} = \mathbb{W} \oplus \mathbb{W}_{\mathrm{r}} \oplus \mathbb{W}_{\mathrm{c}} \oplus \mathbb{V}_0, \quad \mathbb{V}_0 \subseteq \mathbb{V}_{\mathrm{c}} \subseteq \mathbb{V}_{\mathrm{r}} \subseteq \mathbb{V},$$

defined for the transformed problem and its discretization on the transformed subdomain $\Omega_\xi = \Omega_{j,\xi}$. In other words, the spaces $\mathbb{V}(\Omega_\xi)$, $\mathbb{V}_{\mathrm{r}}(\Omega_\xi)$, $\mathbb{V}_{\mathrm{c}}(\Omega_\xi)$, $\mathbb{V}_0(\Omega_\xi)$ are induced by the transformed source FE mesh, rarefied and coarse meshes, imbedded in the transformed source mesh, and by the space of bilinear functions on Ω_ξ, respectively. The spaces $\mathbb{V}_{\mathrm{r}}(\Omega_\xi),\dots,\mathbb{W}_{\mathrm{c}}(\Omega_\xi)$ define preimage spaces denoted by $\mathsf{V}_{\mathrm{r}}(\Omega),\dots,\mathsf{W}_{\mathrm{c}}(\Omega)$. In the preceding sections, the space $\mathscr{V}(\Omega)$ played the role of $\mathbb{V}_{\mathrm{r}}(\Omega_\xi)$.

3.2 Reducing to Isotropic Discretization

The change of variables $\xi_1 = x_1$, $\xi_2 = \sqrt{\rho_1/\rho_2}\,x_2$ transforms the bilinear form $\alpha_\Omega(\cdot,\cdot)$ into

$$\alpha_\Omega(u,v)=\sqrt{\rho_1\rho_2}\,\widetilde{\alpha}(u,v),\quad \widetilde{\alpha}(u,v)=\int\limits_{\Omega_\xi}\nabla_\xi u\cdot\nabla_\xi v\,d\xi, \tag{40}$$

with the notations ∇_ξ for the gradient in the variables ξ, $\Omega_\xi=(0,1)\times(0,\widetilde{\varepsilon})$ for the new domain and $\widetilde{\varepsilon}=\varepsilon\sqrt{\rho_1/\rho_2}$. With this the FE space $\mathscr{V}(\Omega)$ is transformed into the space $\mathbb{V}(\Omega_\xi)$ of piecewise bilinear functions on the rectangular transformed source mesh of the sizes $\hbar_1=h_1$, $\hbar_2=\sqrt{\rho_1/\rho_2}\,h_2$ with the mesh lines $\xi_k\equiv\xi_{k,l}=l\hbar_k$. We have $\mathbf{Q}=\sqrt{\rho_1\rho_2}\,\mathbb{Q}$, $\mathbf{Y}=\sqrt{\rho_1\rho_2}\,\mathbb{Y}$,

$$\mathbb{Y}=\mathbb{Q}_B-\mathbb{Q}_{B,I}\mathbb{Q}_I^{-1}\mathbb{Q}_{I,B} \tag{41}$$

where $\mathbb{Q}_I$, $\mathbb{Q}_{I,B}$, $\mathbb{Q}_B$ are blocks of the stiffness matrix $\mathbb{Q}$ generated by the bilinear form $\widetilde{\alpha}(u,v)$ on the space $\mathbb{V}(\Omega_\xi)$. Therefore, the preconditioning of $\mathbf{Y}$ is reduced to the preconditioning of the Schur complement $\mathbb{Y}$.

We can restrict ourselves to the consideration of the case $\widetilde{\varepsilon}<1$, since the case $\widetilde{\varepsilon}>1$ is reduced to the former one by the interchange of variables. Under the condition $\widetilde{\varepsilon}\le 1$, three cases can be distinguished:

$$i)\quad \hbar_2\le\hbar_1\le\widetilde{\varepsilon},\qquad ii)\quad \hbar_2\le\widetilde{\varepsilon}\le\hbar_1,\qquad iii)\quad \hbar_1\le\hbar_2\le\widetilde{\varepsilon}, \tag{42}$$

and we start with $i)$. Under the stated conditions, the embedded rarefied quasiuniform rectangular grid

$$\xi_k\equiv\widetilde{\xi}_{k,i},\quad k=1,2,\quad \text{with the steps}\quad \eta_{k,i}=\widetilde{\xi}_{k,i}-\widetilde{\xi}_{k,i-1},$$

is introduced by coarsening only in one direction ξ_2. In other words, it is the same uniform grid in the direction ξ_1 with $\eta_{1,i}\equiv\hbar_1\equiv h_1$ and nonuniform in the direction ξ_2 with the sizes $\eta_{2,j}$ as much close as possible to $\hbar_1$. The mesh lines $\xi_2\equiv\widetilde{\xi}_{2,j}$ can be defined as follows. We find $m_2=\text{integer}[\widetilde{\varepsilon}/h_1]$, then define the uniform mesh $\zeta_{2,j}=j\widetilde{\varepsilon}/m_2$, $j=0,1,\dots,m_2$, and then shift the lines of this uniform nonembedded coarse mesh $\xi_2=\zeta_{2,j}$ to the nearest lines $\xi_2\equiv\xi_{2,l}=l\hbar_2$ of the transformed source mesh of the size $\hbar_2$ in the direction ξ_2. We retain the notation m_2 for the number of the rarefied mesh intervals in the direction ξ_2 whereas the number of the rarefied mesh intervals in the direction ξ_1 is $m_1=n_1$. Obviously, the sizes of this mesh satisfy inequalities

$$\underline{c}h_1\le\eta_{k,i}\le\overline{c}h_1,\quad \underline{c}>0,\quad k=1,2, \tag{43}$$

with positive constants, for which we retain the notations as in (11).

The space $\mathbb{V}(\Omega_\xi)$ may be represented by the direct sum

$$\mathbb{V}(\Omega_\xi)=\mathbb{V}_{\mathrm{r}}(\Omega_\xi)\oplus\mathbb{W}(\Omega_\xi), \tag{44}$$

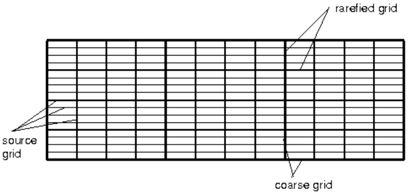


Fig. 3 Transformed rectangular domain and the source, rarefied and coarse grids.

where $\mathbb{V}_{\mathrm{r}}(\Omega_\xi)$ is the space of FE functions which are continuous on $\overline{\Omega}_\xi$ and bilinear on each nest of the rarefied grid. Obviously, the space $\mathbb{W}(\Omega_\xi)$ contains FE functions, which vanish on the lines $\xi_2 \equiv \widetilde{\xi}_{2,j}$ of the rarefied grid. Let $\mathbb{V}_{\mathrm{tr}}(\partial\Omega_\xi)$ be the space of traces of functions from $\mathbb{V}(\Omega_\xi)$ on $\partial\Omega_\xi$. For $u \in \mathbb{V}(\Omega_\xi)$ and $v \in \mathbb{V}_{\mathrm{tr}}(\partial\Omega_\xi)$, respectively, we introduce the norms

$$|u|_{1,\Omega_\xi} = \|u\|_{\Omega_\xi} = (\widetilde{\alpha}(u,u))^{1/2}, \quad \|v\|_{h,\partial\Omega_\xi} = \inf_{\phi\in\mathbb{V}(\Omega_\xi):\phi_{|\partial\Omega}=v} \|\phi\|_{\Omega_\xi}. \tag{45}$$

The matrix $\mathbb{Q}$ can be represented in the block form

$$\mathbb{Q} = \begin{pmatrix} \mathbb{Q}_{\mathrm{s}} & \mathbb{Q}_{\mathrm{sr}} \\ \mathbb{Q}_{\mathrm{rs}} & \mathbb{Q}_{\mathrm{r}} \end{pmatrix}, \tag{46}$$

with blocks $\mathbb{Q}_{\mathrm{s}}$ and $\mathbb{Q}_{\mathrm{r}}$ corresponding to subspaces $\mathbb{W}(\Omega_\xi)$ and $\mathbb{V}_{\mathrm{r}}(\Omega_\xi)$, respectively. Let $\nu_{2,j}$ denote the number of the fine mesh intervals on the rarefied mesh interval $(\widetilde{\xi}_{2,j-1}, \widetilde{\xi}_{2,j})$ and

$$\Delta_{2,j} = \mathrm{tridiag}\,[-1,2,-1]_1^{\nu_{2,j}-1}. \tag{47}$$

An intermediate preconditioner $\mathscr{Q}^*$ for $\mathbb{Q}$ may be defined in the following block form:

$$\mathscr{Q}^* = \mathrm{diag}\,[\mathscr{Q}_{\mathrm{s}}, \mathbb{Q}_{\mathrm{r}}], \quad \mathscr{Q}_{\mathrm{s}}^2 = \mathrm{diag}\,[\underbrace{\Delta_{2,j}, \Delta_{2,j}, \ldots, \Delta_{2,j}}_{(n_1+1)\text{ times}}]_{j=1}^{m_2}. \tag{48}$$

Note that $\ker\mathscr{Q}^* = \ker\mathbb{Q}_{\mathrm{r}}$. For a given j, the i-th block $\Delta_{2,j}$ in the square brackets is related to the nodes on the interval $(\widetilde{\xi}_{2,j-1}, \widetilde{\xi}_{2,j})$ of the mesh line $\xi_1 \equiv \widetilde{\xi}_{1,i}$.

In the case *ii*), the rarefied quasiuniform mesh of the characteristic size $\widetilde{\varepsilon}$ does not exist, and we introduce the *transformed rarefied uniform rectangular mesh* of the characteristic sizes h_1, $\widetilde{\varepsilon}$. Therefore, we have only one layer of n_1 cells $h_1 \times \widetilde{\varepsilon}$, meaning $m_2 = 1$, and similarly to (48) we can set

$$\mathscr{Q}^* = \mathrm{diag}\,[\mathscr{Q}_{\mathrm{s}}, \mathbb{Q}_{\mathrm{r}}], \quad \mathscr{Q}^2_{\mathrm{s}} = \mathrm{diag}\,[\underbrace{\Delta_2, \Delta_2, \dots, \Delta_2}_{(n_1+1)\text{ times}}],$$

with the $(n_2-1)\times(n_2-1)$ blocks Δ_2. The matrix $\mathbb{Q}_{\mathrm{r}}$ is defined on the uniform rectangular coarse transformed grid, which coincides with the rarefied transformed grid and has all nodes on $\partial\Omega_\xi$. This matrix is block-tridiagonal with the blocks 2×2 and does not require preconditioning.

If we have *iii*), the mesh parameter for the quasiuniform rectangular coarse grid is $\hbar_2$ and it satisfies

$$\underline{c}\hbar_2 \le \eta_{k,i} \le \overline{c}\hbar_2, \quad \underline{c} > 0, \quad k = 1,2. \tag{49}$$

At a proper ordering of unknowns, we have again $\mathscr{Q}^* = \mathrm{diag}\,[\mathscr{Q}_{\mathrm{s}}, \mathbb{Q}_{\mathrm{r}}]$, but

$$\mathscr{Q}^2_{\mathrm{s}} = \mathrm{diag}\,[\underbrace{\Delta_{1,i}, \Delta_{1,i}, \dots, \Delta_{1,i}}_{(n_2+1)\text{ times}}]_{j=1}^{m_1}, \quad \Delta_{1,i} = \mathrm{tridiag}\,[-1,2,-1]_1^{\nu_{1,i}-1}, \tag{50}$$

where $\nu_{1,i}$ is the number of the fine mesh intervals on the coarse mesh interval $(\widetilde{\xi}_{1,i-1}, \widetilde{\xi}_{1,i})$. For a given i, the j-th block $\Delta_{1,i}$ in the square brackets is related to the nodes on the interval $(\widetilde{\xi}_{1,i-1}, \widetilde{\xi}_{1,i})$ of the mesh line $\xi_2 \equiv \widetilde{\xi}_{2,j}$.

Lemma 5. *For any positive h_1, h_2, ρ_1, ρ_2 and ε,*

$$\frac{1}{1+\log\delta}\,\mathscr{Q}^* \prec \mathbb{Q} \prec \mathscr{Q}^*, \quad \delta = \max_k \min\left(n_k, \frac{h_{3-k}\sqrt{\rho_k}}{h_k\sqrt{\rho_{3-k}}}\right). \tag{51}$$

Proof. In the case *i*), we consider the transformed discretization mesh and add mesh lines subdividing each interval $(\widetilde{\xi}_{1,i-1}, \widetilde{\xi}_{1,i})$ in $\nu_2 = \mathrm{integer}[h_1/\hbar_2]$ parts and come to a shape regular orthogonal mesh. To FE spaces on this mesh, we can apply results of Bramble *et. al* [9], which allow us to write

$$\frac{1}{1+\log\nu_2}\,\mathscr{Q}^* \prec \mathbb{Q} \prec c_2\,\mathscr{Q}^*. \tag{52}$$

For *ii)*, we have

$$\nu_2 = \min(n_2, \overline{c}h_1/\hbar_2) = \min\left(n_2, \frac{h_1\sqrt{\rho_2}}{h_2\sqrt{\rho_1}}\right), \tag{53}$$

which in the general case should be replaced by δ. □

We will now use Lemma 5 for defining some preconditioner for the Schur complement $\mathbb{Y}$, see (41), restricting ourselves for simplicity to the case *i)*. Taking into account, similar to (19) and (20), representations for the matrices $\mathbb{Q}_{\mathrm{r}}$ and $\mathscr{Q}_{\mathrm{s}}$

$$\mathbb{Q}_{\mathrm{r}} = \begin{pmatrix} \mathbb{Q}_I^{\mathrm{r}} & \mathbb{Q}_{I,B}^{\mathrm{r}} \\ \mathbb{Q}_{B,I}^{\mathrm{r}} & \mathbb{Q}_B^{\mathrm{r}} \end{pmatrix}, \qquad \mathscr{Q}_{\mathrm{s}} = \begin{pmatrix} \mathscr{Q}_I^{\mathrm{s}} & \mathscr{Q}_{I,B}^{\mathrm{s}} \\ \mathscr{Q}_{B,I}^{\mathrm{s}} & \mathscr{Q}_B^{\mathrm{s}} \end{pmatrix},$$
$$\mathbb{Y}_{\mathrm{r}} = \mathbb{Q}_B^{\mathrm{r}} - \mathbb{Q}_{B,I}^{\mathrm{r}}(\mathbb{Q}_I^{\mathrm{r}})^{-1}\mathbb{Q}_{I,B}^{\mathrm{r}}, \quad \mathscr{G}_{\mathrm{s}} = \mathscr{Q}_B^{\mathrm{s}} - \mathscr{Q}_{B,I}^{\mathrm{s}}(\mathscr{Q}_I^{\mathrm{s}})^{-1}\mathscr{Q}_{I,B}^{\mathrm{s}}, \tag{54}$$

we conclude that the Schur complement $\mathscr{G}^*$ for $\mathscr{Q}^*$ has the form

$$\mathscr{G}^* = \mathrm{diag}\,[\mathscr{G}_{\mathrm{s}}, \mathbb{Y}_{\mathrm{r}}]. \tag{55}$$

According to (48), in the matrix $\mathscr{Q}_{\mathrm{s}}$ internal degrees of freedom are not coupled with the boundary degrees of freedom, i.e., $\mathscr{Q}_{I,B}^{\mathrm{s}} = (\mathscr{Q}_{B,I}^{\mathrm{s}})^\top = \mathbf{0}$. Therefore,

$$\mathscr{G}_{\mathrm{s}} = \mathscr{Q}_B^{\mathrm{s}} = \frac{h_1}{\hbar_2}\,\mathrm{diag}\,[\Delta_{2,j}, \Delta_{2,j}]_{j=1}^{m_2}, \tag{56}$$

where the two matrices in the square brackets correspond to nodes in the interval $(\widetilde{\xi}_{2,j-1}, \widetilde{\xi}_{2,j})$ on the left and right vertical edges of $\partial\Omega_\xi$, respectively.

Corollary 4. *Let $\mathscr{G}^* = diag\,[\mathscr{Q}_{\mathrm{B}}^{\mathrm{s}}, \mathbb{Y}_{\mathrm{r}}]$. Then for all positive h_1, h_2, ρ_1, ρ_2, and ε we have*

$$\frac{1}{1+\log\delta}\,\mathscr{G}^* \prec \mathbb{Y} \prec \mathscr{G}^*. \tag{57}$$

For the reason of a complete analogy between the Schur complements $\mathbf{B}_{\mathrm{Hi}}$ and $\mathbb{Y}_{\mathrm{r}}$, Corollary 4 reduces the Schur complement preconditioning to the case, considered in the preceding section. There, the FE space $\mathscr{V}(\Omega) = \mathscr{W}_{\mathrm{r}}(\Omega) \oplus \mathscr{W}_{\mathrm{c}}(\Omega) \oplus \mathscr{V}_0(\Omega)$ plays here the role of the rarefied transformed space $\mathbb{V}_{\mathrm{r}}(\Omega_\xi) = \mathbb{W}_{\mathrm{r}}(\Omega_\xi) \oplus \mathbb{W}_{\mathrm{c}}(\Omega_\xi) \oplus \mathbb{V}_0(\Omega_\xi)$. Hence, we can introduce the preconditioner for $\mathbb{Y}_{\mathrm{r}}$

$$\mathscr{C}_{\mathrm{r}} = \begin{pmatrix} \mathscr{B}_{W_{\mathrm{r}}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{W_{\mathrm{c}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_0 \end{pmatrix} \tag{58}$$

in a completely similar way to the preconditioner $\mathscr{C}_{\mathrm{Hi}}$ of (35) for $\mathbf{B}_{\mathrm{Hi}}$ and the preconditioner for the Schur complement $\mathbb{Y}$

$$\mathscr{G} = \mathrm{diag}\,[\mathscr{Q}_B^{\mathrm{s}}, \mathscr{C}_{\mathrm{r}}] = \mathrm{diag}\,[\mathscr{Q}_B^{\mathrm{s}}, \mathscr{B}_{W_{\mathrm{r}}}, \mathbf{B}_{W_{\mathrm{c}}}, \mathbf{B}_0]. \tag{59}$$

For the cases (42), the proof of the bounds

$$\frac{1}{(1+\log\delta)(1+\log\underline{m})}\min\left(\frac{1}{m_\varepsilon}, \frac{1}{(1+\log\underline{m})}\right)\mathscr{G} \prec \mathbb{Y} \prec \mathscr{G}, \tag{60}$$

where $\underline{m} = \min(m_1, m_2)$ and $m_\varepsilon = 1/\widetilde{\varepsilon}$, follows by combining the bounds of Lemma 5 and Lemma 4.

For a general rectangle $\Omega = H_1 \times H_2$, in the same way we introduce the rarefied source and the coarsest meshes, while the role of ε is played by $\min_k (H_k/H_{3-k})$. The above form of the preconditioner is retained, if after the transformation to the isotropic problem the shortest edge is directed along the axis x_2. In general, with the notation

$$\theta = \max_k \min\left(m_k, \frac{H_k\sqrt{\rho_{3-k}}}{H_{3-k}\sqrt{\rho_k}}\right),$$

the counterpart of (60) for all positive $H_k, h_k \le H_k, \rho_k$ is

$$\underline{\mu}\mathscr{G} \prec \mathbb{Y} \prec \mathscr{G} \tag{61}$$

with

$$\underline{\mu} = \frac{1}{(1+\log\delta)(1+\log\underline{m})} \min\left(\frac{1}{\theta}, \frac{1}{(1+\log\underline{m})}\right). \tag{62}$$

For slim rectangular domains Ω, a Schur complement preconditioner with a less degree of decoupling can be introduced. It is based on the FE space decomposition

$$\mathbb{V}(\Omega_\xi) = \mathbb{W}_{\mathrm{d}}(\Omega_\xi) \oplus \mathbb{V}_{\mathrm{c}}(\Omega_\xi), \quad \mathbb{V}_{\mathrm{c}}(\Omega_\xi) = \mathbb{W}_{\mathrm{c}}(\Omega_\xi) \oplus \mathbb{V}_0(\Omega_\xi),$$

where $\mathbb{W}_d(\Omega_\xi)$ is the space of functions $v \in \mathbb{V}(\Omega_\xi)$ with zero values at the nodes of the coarse mesh. The preconditioner, for which we retain the notation $\mathscr{G}$, gets the form

$$\mathscr{G} = \mathrm{diag}\,[\mathscr{B}_{\mathrm{W_d}}, \mathbf{B}_{W_{\mathrm{c}}}, \mathbf{B}_0] \tag{63}$$

with the same $\mathbf{B}_{W_{\mathrm{c}}}$, $\mathbf{B}_0$ as in (59). The block $\mathscr{B}_{\mathrm{W_d}}$ looks like $\mathscr{B}_W$ in (36), i.e.,

$$\mathscr{B}_{\mathrm{W_d}} = \mathrm{diag}\,[\Delta_{1/2,i}]_{i=1}^{2(n_\varepsilon+1)}, \quad \Delta_{1/2,i}^{1/2} = \mathrm{tridiag}\,[-1,2,-1]_1^{\nu_i-1},$$

but now ν_i denotes the number of the discretization mesh on the corresponding coarse mesh interval belonging to $\partial\Omega$. In the proof of the relative spectrum bounds results of Bramble *et. al* [9] are applied to the domain decomposition mesh by the coarse mesh. For this, FE functions are considered as elements of the space $\mathscr{V}_{\mathrm{ff}}(\Omega)$, induced by the condensed transformed discretization mesh, which is the coarsest shape regular orthogonal mesh, covering the transformed discretization mesh. In the resulting inequalities (61)

$$\underline{\mu} = \frac{1}{(1+\log n_{\mathrm{ff}})} \min\left(\frac{1}{\theta}, \frac{1}{(1+\log n_{\mathrm{ff}})}\right), \quad n_{\mathrm{ff}} = \max_{k=1,2} \frac{n_k}{\max[1, \frac{H_k\sqrt{\rho_{3-k}}}{H_{3-k}\sqrt{\rho_k}}]}. \tag{64}$$

Thus, we have proved:

Theorem 3. *For all positive H_k, $h_k \le H_k$, ρ_k, the Schur complement preconditioners $\mathscr{G}$ of (59) and (63) satisfy (61) with $\underline{\mu}$ from (62), (64), respectively.*

Solving systems $\mathscr{G}\mathbf{v}_B = \mathbf{f}_B$ requires not more than $\mathscr{O}(\overline{n}\log\underline{n})$ arithmetical operations, but the relative condition depends on $\theta \le m_k$. At the same time this Schur complement preconditioner, as others previously considered, has an obvious drawback, if we turn to the problem (1)-(5). Let $\mathscr{G} = \mathscr{G}_j$ be such a preconditioner for the subdomain Ω_j and $\mathscr{S}$ be the preconditioner for the interface Schur complement $\mathbf{S} = \mathbf{K}_B - \mathbf{K}_{B,I}\mathbf{K}_I^{-1}\mathbf{K}_{I,B}$ of the matrix $\mathbf{K}$, assembled from preconditioners $\sqrt{\rho_1\rho_2}\mathscr{G}_j$. For different subdomains Ω_j, decompositions $\mathscr{V}(\Omega_j) = \mathsf{W}(\Omega_j) \oplus \mathsf{W}_\mathrm{r}(\Omega_j) \oplus \mathsf{W}_\mathrm{c}(\Omega_j) \oplus \mathsf{V}_0(\Omega_j)$ are not compatible and, therefore, a fast solving procedure for systems with such a matrix $\mathscr{S}$ requires a new consideration.

4 Compatible Schur Complement Preconditioner

Khoromskij & Wittum [27, 28], Korneev [31], Korneev et al. [33, 34] and Rytov [46] used in relation with the problem (37) slightly different compatible subdomain edge Schur complement preconditioners. They were obtained from the Schur complement $\mathbf{Y}$ by decoupling some of its blocks on the diagonal. In [31], the preconditioner has on the diagonal the independent block of vertex degrees of freedom and two independent blocks each related to a pair of parallel edges of the rectangle Ω_ξ, whereas in [27, 28] two short edges were additionally decoupled. In this section, we present bounds of the relative condition numbers of these preconditioners, which follow from (61).

We call by the source triangulation the one obtained by subdivision of each rectangular nest of the source mesh in two triangles by the diagonal of the same direction and denote $\mathscr{U}_\Delta(\Omega)$ the space of continuous functions which are linear on each triangle. It is represented by the direct sums $\mathscr{U}(\Omega) = \mathscr{U}_I(\Omega) \oplus \mathscr{U}^\mathrm{B}(\Omega)$, $\mathscr{U}^\mathrm{B}(\Omega) = \mathscr{U}^\mathrm{E}(\Omega) \oplus \mathscr{U}^\mathrm{V}(\Omega)$, spanned over internal, boundary, edge and vertex FE functions, which are a nodal basis for the source triangulation. Respectively, the stiffness matrix induced by the space $\mathscr{U}(\Omega)$, which is denoted $\mathbf{L}$, is represented in the block forms

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_I & \mathbf{L}_{IB} \\ \mathbf{L}_{BI} & \mathbf{L}_B \end{pmatrix} = \begin{pmatrix} \mathbf{L}_I & \mathbf{L}_{IE} & \mathbf{L}_{IV} \\ \mathbf{L}_{EI} & \mathbf{L}_E & \mathbf{L}_{EV} \\ \mathbf{L}_{VI} & \mathbf{L}_{VE} & \mathbf{L}_V \end{pmatrix}, \tag{65}$$

and $\mathbf{L}_{IV} = \mathbf{L}_{VI}^\top = \mathbf{0}$. Therefore, the Schur complement $\mathscr{L} = \mathbf{L}_B - \mathbf{L}_{BI}\mathbf{L}_I^{-1}\mathbf{L}_{IB}$ has the form

$$\mathscr{L} = \begin{pmatrix} \mathscr{L}_E & \mathbf{L}_{EV} \\ \mathbf{L}_{VE} & \mathbf{L}_V \end{pmatrix}, \quad \mathscr{L}_E = \mathbf{L}_E - \mathbf{L}_{EI}\mathbf{L}_I^{-1}\mathbf{L}_{IE},$$

and, due to the spectral equivalence $\mathbf{L} \prec \mathbf{Q} \prec \mathbf{L}$, we have

$$\mathscr{L} \prec \mathbf{Y} \prec \mathscr{L}\,.$$

Let us represent $\mathscr{L}_E$ in the 4×4 block form $\mathscr{L}_E = \left\{\mathscr{L}_{k,l}^\mathrm{E}\right\}_{k,l=0}^3$ with the blocks corresponding to the edges γ_k and note that all 16 blocks are nonzero, see, e.g., Korneev [31] and Rytov [46]. If θ is not big, say $\theta \le 2$, it is possible to use the preconditioner

$$\mathscr{Y}^\mathrm{E} = \mathrm{diag}\,[\mathscr{L}_{0,0}^\mathrm{E}, \mathscr{L}_{1,1}^\mathrm{E}, \mathscr{L}_{2,2}^\mathrm{E}, \mathscr{L}_{3,3}^\mathrm{E}]\,.$$

For $\theta > 2$ any of the two preconditioners

$$\mathscr{Y}^{\mathrm{E}} = \begin{pmatrix} \mathscr{L}^{\mathrm{E}}_{0,0} & \mathscr{L}^{\mathrm{E}}_{0,1} & \mathbf{0} & \mathbf{0} \\ \mathscr{L}^{\mathrm{E}}_{1,0} & \mathscr{L}^{\mathrm{E}}_{1,1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathscr{L}^{\mathrm{E}}_{2,2} & \mathscr{L}^{\mathrm{E}}_{2,3} \\ \mathbf{0} & \mathbf{0} & \mathscr{L}^{\mathrm{E}}_{3,2} & \mathscr{L}^{\mathrm{E}}_{3,3} \end{pmatrix}, \quad \mathscr{Y}^{\mathrm{E}} = \begin{pmatrix} \mathscr{L}^{\mathrm{E}}_{0,0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathscr{L}^{\mathrm{E}}_{1,1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathscr{L}^{\mathrm{E}}_{2,2} & \mathscr{L}^{\mathrm{E}}_{2,3} \\ \mathbf{0} & \mathbf{0} & \mathscr{L}^{\mathrm{E}}_{3,2} & \mathscr{L}^{\mathrm{E}}_{3,3} \end{pmatrix} \tag{66}$$

can be used. The first one is obtained by decoupling adjacent edges, and the second one additionally assumes a decoupling the pair of parallel edges, which became shortest after mapping to Ω_ξ and are the edges γ_0, γ_1 in the above expression. Edge Schur complement preconditioners can be defined similarly by means of the matrix $\mathbf{Q}$ or $\widetilde{\mathbf{Q}}$ with the latter obtained from $\mathbf{Q}$ by setting $\mathbf{Q}_{IV} = \mathbf{Q}_{VI}^{\top} = \mathbf{0}$.

The Schur complement preconditioner

$$\mathscr{Y} = \mathrm{diag}\,[\mathscr{Y}^{\mathrm{E}}, \mathscr{Y}_0], \quad \mathscr{Y}_0 = \sqrt{\rho_1 \rho_2}\,\mathbf{B}_0 \tag{67}$$

corresponds to the two-level decomposition $\mathscr{U}^{\mathrm{B}}(\partial\Omega) = \mathscr{U}^{\mathrm{E}}(\partial\Omega) \oplus \mathscr{U}_0(\partial\Omega)$, of the boundary FE space, written for the traces of the spaces entering the decomposition $\mathscr{U}^{\mathrm{B}}(\Omega) = \mathscr{U}^{\mathrm{E}}(\Omega) \oplus \mathscr{U}_0(\Omega)$. Here $\mathbf{B}_0$, is the matrix generated by the space $\mathscr{U}_0(\Omega)$ of continuous functions which are linear on each of the two triangles, having vertices at the vertices of Ω. This matrix may be also generated by the subspace $\mathscr{V}_0(\Omega)$ of bilinear polynomials on Ω.

Theorem 4. *For all positive ρ_k, H_k, and $h_k \le H_k$ the preconditioners $\mathscr{Y}^{\mathrm{E}}$, $\mathscr{Y}$ satisfy the inequalities*

$$\underline{\beta}_{\mathrm{E}}\, \mathscr{Y}^{\mathrm{E}} \prec \mathbf{Y}^{\mathrm{E}} \prec \mathscr{Y}^{\mathrm{E}}, \tag{68}$$

$$\underline{\mu}\mathscr{Y} \prec \mathbf{Y} \prec \mathscr{Y}, \tag{69}$$

with $\underline{\mu}$, defined by the maximum of the values (62), (64) and

$$\underline{\beta}_{\mathrm{E}} = \max\left[\frac{1}{(1+\log\delta)(1+\log\underline{m})^2}, \frac{1}{(1+\log n_{\mathrm{ff}})^2}\right].$$

Proof. We will consider the case of $\theta > 2$ and the preconditioner $\mathscr{Y}^{\mathrm{E}}$ defined by the second expression in (66). Suppose, that for some positive $\underline{\mu}$, $\overline{\mu}$ the preconditioner $\Upsilon = \mathrm{diag}\,[\Upsilon_E, \Upsilon_0]$ satisfies

$$\underline{\mu}\,\Upsilon \le \mathbf{Y} \le \overline{\mu}\,\Upsilon, \tag{70}$$

where Υ_E has the structure, similar to $\mathscr{Y}^{\mathrm{E}}$, i.e.,

$$\Upsilon_E = \begin{pmatrix} \Upsilon_{E,0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Upsilon_{E,1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Upsilon_{E,2} & \Upsilon_{E,23} \\ \mathbf{0} & \mathbf{0} & \Upsilon_{E,32} & \Upsilon_{E,3} \end{pmatrix}. \tag{71}$$

Then

$$(\underline{\mu}/\overline{\mu})\mathscr{Y} \prec \mathbf{Y} \prec 4\overline{\mu}\mathscr{Y}. \tag{72}$$

Indeed, let V_0, V_{E_0}, V_{E_1}, V_{E_2,E_3} be the vector spaces of degrees of freedom corresponding to the independent blocks on the diagonal of the matrix $\mathscr{Y}$. From (70), considered on these subspaces, and (66), (71) it follows that

$$\underline{\mu}\,\varUpsilon_E \prec \mathscr{Y}^{\rm E} \prec \overline{\mu}\,\varUpsilon_E, \quad \underline{\mu}\,\varUpsilon_0 \prec \mathscr{Y}_0 \prec \overline{\mu}\,\varUpsilon_0, \tag{73}$$

and combining (70) and (73) we get

$$\mathbf{Y} \geq \underline{\mu}\,\mathrm{diag}\,[\varUpsilon_E, \varUpsilon_0] \geq (\underline{\mu}/\overline{\mu})\,\mathrm{diag}\,[\mathscr{Y}^{\rm E}, \mathscr{Y}_0] = (\underline{\mu}/\overline{\mu})\,\mathscr{Y}.$$

This proves the left inequality (72). The right inequality (72) follows from the inequality of Cauchy and the last one in (73).

Let us turn now to the block $\mathscr{G}_E = \mathrm{diag}\,[\mathscr{Q}^{\rm s}_B, \mathscr{B}_{W_{\rm r}}, \mathbf{B}_{W_{\rm c}}]$, of the preconditioner $\mathscr{G}$ in Theorem 3. It has the same structure as $\mathscr{Y}^{\rm E}$. If we repeat the derivation of (60), however omitting steps related to the splitting of vertices, we come to the bounds

$$\underline{\beta}_E\,\mathscr{G}_E \prec \mathbf{Y}_E \prec \mathscr{G}_E, \tag{74}$$

where without change of the notation it is implied that $\mathscr{G}$ is transformed to the basis common with $\mathbf{Y}$. Therefore, one can take $\sqrt{\rho_1\rho_2}\,\mathscr{G}_E$ for $\varUpsilon_E$ and obtain (68).

Similarly, on the basis (72) and Theorem 3, inequalities (69) are proved, including the case of the use of the preconditioner given by first expression (66). □

Several facts are important for numerical implementations of the preconditioners (66). For each subdomain, we have introduced the discretization, transformed rarefied and transformed coarse imbedded meshes, and in general all these meshes can be nonuniform. However, in the DD preconditioner they all can be replaced by uniform orthogonal non-imbedded meshes, what is assumed in what follows. This does not influence the asymptotic computational cost. FDFT, applied edge-wise to the first of the preconditioners (66), makes the preconditioner a block diagonal matrix with 2 × 2 blocks. Obviously, the matrix of FDFT, which is designated $\mathscr{F}_E$, is the block diagonal matrix with the identical blocks for the opposite among edges γ_k, $k = 0, 1, 2, 3$:

$$\mathscr{F}_E = [\mathscr{F}_0, \mathscr{F}_0, \mathscr{F}_2, \mathscr{F}_2]. \tag{75}$$

For the problem (37) the block diagonal matrix $\varLambda := \mathscr{F}_E^\top \mathscr{Y}^{\rm E} \mathscr{F}_E$ has the form

$$\varLambda = \mathrm{diag}\left[\mathrm{diag}\,[\varLambda_i^{(0)}]_{i=1}^{n_2-1}, \mathrm{diag}\,[\varLambda_k^{(2)}]_{k=1}^{n_1-1}\right],$$

$$\varLambda = \mathrm{diag}\left[\mathrm{diag}\,[\mathrm{diag}\,[\varLambda_{0,i}, \varLambda_{0,i}]]_{i=1}^{n_2-1}, \mathrm{diag}\,[\varLambda_k^{(2)}]_{k=1}^{n_1-1}\right],$$

respectively to the first and second expressions (66). Each block $\varLambda_i^{(0)}$ couples a pair of opposite nodes $(0, x_{2,i})$ and $(1, x_{2,i})$ on the vertical edges and each block $\varLambda_k^{(2)}$ couples a pair of opposite nodes $(x_{1,k}, 0)$ and $(x_{1,k}, \varepsilon)$ of the horizontal edges. For the second preconditioner (66), the 2×2 matrix $\varLambda_i^{(0)}$ is diagonal with two equal nonzero entries. Due to the pointed out property the system with the matrix $\mathscr{Y}^{\rm E}$

can be solved in $\mathscr{O}((n_1+n_2)\log\overline{n})$ arithmetical operations. The Schur complement $\mathbf{Y}_E$ and the preconditioners $\mathscr{Y}^{\mathrm{E}}$ can be calculated in the trigonometric basis for $n_1 \times n_2$ arithmetical operations, whereas matrix vector multiplications by $\mathscr{Y}^{\mathrm{E}}$ require $\mathscr{O}((n_1+n_2)\log\overline{n})$ arithmetical operations, see, e.g., Korneev [31] and Rytov [46]. Costs of some of these operations can be considerably reduced, if some up to date techniques are applied, such as $\mathscr{H}$-matrices and tensor-train decompositions Hackbusch et al. [25], Khoromskij & Wittum [28], Dolgov et al. [16]. For instance, the $\mathscr{H}$-matrix approximation technique provides the cost $\mathscr{O}(n_{\partial\Omega}\log^g n_{\partial\Omega})$ for the computation of the matrix $\mathbf{Y}_E$, $n_{\partial\Omega} = 2(n_1+n_2)$, as well as for storage operations and matrix vector multiplications.

5 Piecewise Orthotropic Discretizations on Domains Composed of Rectangles with Arbitrary Aspect Ratios

5.1 Schur Complement and Domain Decomposition Algorithms

We turn now to the piecewise orthotropic problem (4) and its piecewise orthotropic FE discretization (6) by means of decomposition and discretization meshes (2)-(3). The bilinear form $\alpha_\Omega(u,v)$, $\forall u,v \in \overset{\circ}{\mathscr{V}}(\Omega)$, induces the FE stiffness matrix $\mathbf{K}$ and its inter-subdomain Schur complement $\mathbf{S} = \mathbf{K}_B - \mathbf{K}_{B,I}\mathbf{K}_I^{-1}\mathbf{K}_{I,B}$, which, obviously, may be viewed as assembled from the stiffness matrices $\mathbf{K}_j$ and the corresponding Schur complements $\mathbf{S}_j$ for subdomains Ω_j. Preconditioners for subdomain matrices $\mathbf{K}_j$, $\mathbf{S}_j$, studied in the preceding sections, will be used in the fast solvers for the systems (6) and

$$\mathbf{S}\mathbf{u}_B = \mathbf{F}_B \tag{76}$$

and we start from the Schur complement solver. It is based on the use of the two preconditioners $\mathscr{S}_k$, $k = 1,2$, for the matrix $\mathbf{S}$ with different properties. It is implied that $\mathscr{S}_1$ is as close as possible to $\mathbf{S}$ in the spectrum and is cheap, at least much cheaper than $\mathbf{S}$, for matrix-vector multiplications. The preconditioner $\mathscr{S}_2$ is allowed to be less close to $\mathbf{S}$ in the spectrum, but is cheap for operations $\mathscr{S}_2^{-1}\mathbf{y}$ and at least much cheaper than $\mathbf{S}$ and $\mathscr{S}_1$ for operations $\mathbf{S}^{-1}\mathbf{y}$, $\mathscr{S}_1^{-1}\mathbf{y}$. Under these assumptions, we solve the system (76) by a PCG with the preconditioner $\mathscr{S}_1$, whereas systems $\mathscr{S}_1\mathbf{x} = \mathbf{y}$, arising at each PCG iteration, are solved inexactly by means of the iterative processes

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \sigma_k\mathscr{S}_2^{-1}(\mathscr{S}_1\mathbf{x}^k - \mathbf{y}), \quad k = 1,2,\dots,k_s, \tag{77}$$

with Chebyshev iteration parameters σ_k for some fixed numbers k_s of iterations. In other words, the system $\mathbf{S}\mathbf{u}_B = \mathbf{F}_B$ is solved by a PCG with the preconditioner $\mathscr{S}_{1,\mathrm{it}}$ which inverse is

$$\mathscr{S}_{1,\mathrm{it}}^{-1} = \Big[\mathbf{I} - \prod_{k=1}^{k_s}(\mathbf{I} - \sigma_k\mathscr{S}_2^{-1}\mathscr{S}_1)\Big]\mathscr{S}_1^{-1}. \tag{78}$$

Proposition 1. *Suppose that*

ι) preconditioners $\mathscr{S}_k$ satisfy

$$\underline{\gamma}_1 \mathscr{S}_1 \prec \mathbf{S} \prec \overline{\gamma}_1 \mathscr{S}_1, \quad \underline{\gamma}_2 \mathscr{S}_2 \prec \mathscr{S}_1 \prec \overline{\gamma}_2 \mathscr{S}_2,$$

ιι) matrix-vector multiplications by $\mathbf{S}$ and $\mathscr{S}_1$ spend $\mathscr{N}_{\mathrm{S}}$ and $\mathscr{N}_{\mathscr{S}_1}$ arithmetical operations, respectively, and

ιιι) solving the system $\mathscr{S}_2 \mathbf{v}_B = \mathscr{F}_B$, $\forall \mathscr{F}_B$, requires $\mathscr{N}_{\mathscr{S}_2}$ operations.

Then solving the system (76) with the prescribed accuracy $\varepsilon \in (0,1)$ in the norm $\|\cdot\|_{\mathbf{S}}$ requires not more than

$$c\sqrt{\overline{\gamma}_1/\underline{\gamma}_1}\left[\mathscr{N}_{\mathrm{S}} + \sqrt{\overline{\gamma}_2/\underline{\gamma}_2}(\mathscr{N}_{\mathscr{S}_1} + \mathscr{N}_{\mathscr{S}_2})\right] \log \varepsilon^{-1}$$

arithmetical operations, $c = \mathrm{const.}$

Proof. The proof of these statements can be found in Nepomnyaschikh [40] and many other places, see, e.g., Korneev & Langer [32]. □

Preconditioner $\mathscr{S}_1$

We transform each subdomain Ω_j to $\Omega_{j,\varepsilon}$ and by means of the condensed transformed mesh we define the preconditioner (28), denoted now $\mathbf{C}_{\mathrm{ff,j}}$. Setting $\mathscr{S}_{1,j} = \sqrt{\rho_{1,j}\rho_{2,j}}\,\mathbf{C}_{\mathrm{ff,j}}$, we assemble $\mathscr{S}_1$ from these subdomain matrices. As it follows from these definitions and Corollary 3,

$$\mathscr{S}_1 \prec \mathbf{S} \prec \mathscr{S}_1 \tag{79}$$

and

$$\mathrm{ops}\,[\mathscr{S}_1 \mathbf{v}] \prec (\log \overline{n}) \sum_j (n_{1,j} + n_{2,j}) \prec (J_1 N_2 + J_2 N_1) \log \overline{n}, \quad \forall \mathbf{v}.$$

Preconditioner $\mathscr{S}_2$

For each subdomain Ω_j, we consider the preconditioner $\mathscr{S}_{2,j} = \mathscr{Y}_j$, where $\mathscr{Y}_j = \mathrm{diag}\,[\mathscr{Y}_j^{\mathrm{E}}, \sqrt{\rho_{1,j}\rho_{2,j}}\,\mathbf{B}_{0,j}]$ is defined in the same way as $\mathscr{Y}$ in (67) for the domain $\Omega = \Omega_j$. Then $\mathscr{S}_2$ is assembled from subdomain preconditioners $\mathscr{S}_{2,j}$ and has the block diagonal form $\mathscr{S}_2 = \mathrm{diag}\,[\mathscr{S}_E, \mathbf{K}_V]$, where $\mathbf{K}_V$ is the block of the FE matrix $\mathbf{K}$ for vertices. Obviously, $\mathscr{S}_E$ is assembled from the matrices $\mathscr{S}_{E,j} = \mathscr{Y}_j^{\mathrm{E}}$, and according to Theorem 4

$$\min_j \underline{\mu}_j \mathscr{S}_2 \prec \mathbf{S} \prec \mathscr{S}_2, \tag{80}$$

where $\underline{\mu}_j$ is the value of $\underline{\mu}$ in (69) for a particular subdomain Ω_j.

Taking into account the representation of the FE stiffness matrix in the block form

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_I & \mathbf{K}_{IB} \\ \mathbf{K}_{BI} & \mathbf{K}_B \end{pmatrix}, \tag{81}$$

we define the inverse to the DD preconditioner $\mathscr{K}_{\mathrm{DD}}$ by the expression

$$\mathscr{K}_{\mathrm{DD}}^{-1} = \mathscr{K}_I^{+} + \mathbb{P}\mathscr{S}_{1,\mathrm{it}}^{-1}\mathbb{P}^{\top}. \tag{82}$$

Here the matrix $\mathscr{K}_I$ is related to the interior degrees of freedom for each subdomain and, like the block $\mathbf{K}_I = \mathrm{diag}\,[\mathbf{K}_{I,j}]_{j_1,j_2=1}^{J_1,J_2}$, has the block diagonal structure $\mathscr{K}_I = \mathrm{diag}\,[\mathscr{K}_{I,j}]_{j_1,j_2=1}^{J_1,J_2}$, $\mathbb{P}$ is the prolongation matrix $\mathbb{P}^{\top} = (\mathbb{P}_I^{\top}, \mathbf{I})^{\top}$. It is assumed that

$$\mathscr{K}_{I,j} \prec \mathbf{K}_{I,j} \prec \mathscr{K}_{I,j}, \quad \|\mathbb{P}_{B_j}\mathbf{v}_{B_j}\|_{\mathbf{K}_j} \prec \|\mathbf{v}_{B_j}\|_{\mathbf{S}_j}, \tag{83}$$

where $\mathbb{P}_{B_j}$ is the restriction of the prolongation operator $\mathbb{P}$ to the subdomain $\overline{\Omega}$. It is assumed additionally that

$$\mathrm{ops}\,[\mathscr{K}_I^{-1}\mathbf{f}_I] = N_{\Omega},\ \forall \mathbf{f}, \quad \mathrm{ops}\,[\mathbb{P}\mathbf{v}_B] \prec N_{\Omega},\ \forall \mathbf{v}_B. \tag{84}$$

There is a variety of such preconditioners and prolongation operators in the literature, and we refer only to papers by Oswald [44], Griebel & Oswald [22] and Nepomnyaschikh [41] for examples.

We restrict our considerations to the problem with subdomain-wise constant coefficients $\wp = \rho$. Clearly, the numerical complexity of DD preconditioners $\mathscr{K}_{\mathrm{DD}}$ for (1)-(5) will differ only by a constant depending on μ_1, μ_2. However, an implementation of the DD Schur complement solver will in general differ noticeably, since it requires the calculation of $\mathbf{S}$ and multiplications by it, which can be expensive. For instance, an implementation of $\mathscr{H}$-matrix techniques for the calculation of $\mathbf{S}$ gives the complexity $\mathscr{O}(N_{\Omega}\log N_{\Omega})$, $N_{\Omega} = N_1N_2$. At the same time, the complexity of the same operation and of the matrix vector multiplication by $\mathbf{S}$ is $\mathscr{O}(N_{\Gamma}\log N_{\Gamma})$, $N_{\Gamma} = J_1N_1 + J_2N_2$, if $\wp = \rho$, cf. Hackbusch [24] and Hackbusch et al. [25]. The last estimates are assumed to hold in what follows.

Theorem 5. *Let $H_{k,j_k} \in (0,1)$, $n_{k,j_k} \geq 1$, $\rho_{k,j} > 0$ be arbitrary in the pointed out ranges. Then the total arithmetical costs $Q_{\mathbf{K}}$, $Q_{\mathbf{S}}$ of the DD and Schur complement algorithms satisfy the bounds*

$$\begin{aligned} Q_{\mathbf{K}} &\prec N_1N_2 + [N_{\Gamma}(1+\log\overline{N}) + \Upsilon(J_1J_2)]\sqrt{\overline{N}}(1+\log\overline{N})^{1/2}, \\ Q_{\mathbf{S}} &\prec N_{\Gamma}(1+\log N_{\Gamma}) + [N_{\Gamma}(1+\log\overline{N}) + \Upsilon(J_1J_2)]\sqrt{\overline{N}}(1+\log\overline{N})^{1/2}, \end{aligned} \tag{85}$$

where $\overline{N} = \max_k N_k$ and $\Upsilon(J_1, J_2)$ stands for the cost of the solution of the subsystem with the matrix $\mathbf{K}_V$.

Proof. Let us list the main factors contributing to the complexity of the Schur complement algorithm:

– the number of external PCG iterations $k_{\mathrm{PCG}} = \mathrm{const}$,
– the cost $\mathscr{N}_S \prec N_\Gamma \log N_\Gamma$ arithmetical operations of one matrix-vector multiplication by $\mathbf{S}$,
– the number of secondary iterations (77)

$$k_s \prec 1/\sqrt{\underline{\mu}_{12}} \prec \sqrt{\max_j \overline{n}_j(1+\log \underline{n}_j)} \prec \sqrt{\overline{N}(1+\log \overline{N})},$$

* the cost of the matrix-vector multiplication by $\mathscr{S}_1$ at each secondary iteration

$$\mathscr{N}_{\mathscr{S}_1} \prec N_\Gamma(1+\log \overline{N}),$$

* the cost of solving the system with the preconditioner $\mathscr{S}_2$ at each secondary iteration

$$\mathscr{N}_{\mathscr{S}_2} \prec N_\Gamma(1+\log \overline{N}) + \Upsilon(J_1, J_2).$$

Taking into account the proof of Theorem 4, we conclude that

$$\underline{\mu}_{12}\mathscr{S}_2 \prec \mathscr{S}_1 \prec \mathscr{S}_2, \quad \underline{\mu}_{12} = \min_j \left[\frac{1}{(1+\log \underline{m}_j)} \min\left(\frac{1}{\theta_j}, \frac{1}{(1+\log \underline{m}_j)}\right)\right], \tag{86}$$

and, therefore, the given number k_s of secondary iterations provides the spectral equivalence $\mathscr{S}_{1,\mathrm{it}} \asymp \mathscr{S}_1$.

Now, the last relationship and (79) guarantee that $k_{\mathrm{PCG}} \log \varepsilon^{-1}$ PCG iterations provide the relative error in the norm $\|\cdot\|_{\mathbf{S}}$ bounded by the prescribed $\varepsilon > 0$. The first term in the expression for $\mathscr{N}_{\mathscr{S}_2}$ bounds the arithmetical cost of solving the system with the matrix $\mathscr{S}_E$. Implementing above bounds according to Proposition 1, we come to the bound (85) for $Q_{\mathbf{S}}$.

The DD peconditioner $\mathscr{K}_{\mathrm{DD}}$ is spectrally equivalent to the FE matrix $\mathbf{K}$, what follows from Corollary 3 and (83). The estimate of the DD solver cost is obtained by taking additionally into account the above costs of operations, related to the interface, and (84). □

Suppose that $N_1 = N_2 = N$, $J_1 = J_2 = J$ and the decomposition mesh is fixed. Then

$$Q_{\mathbf{K}} \prec N^2. \tag{87}$$

If the number of subdomains grows with the growth of the numbers N_k of the source mesh lines, the contribution $\Upsilon(J_1 J_2)$ of the solver for the vertex subproblem can compromise this bound. Assuming that a direct solver for systems with the matrix $\mathbf{K}_V$ is sufficiently fast, the bound (87) is retained under the condition

$$J_k \leq N^{1/2}/(1+\log N)^{3/2}.$$

It is worth emphasizing that the above estimates are relatively crude, since we practically made no restrictions on H_{k,j_k}, n_{k,j_k}, and $\rho_{k,j} > 0$ and their change from subdomain to subdomain. It can help to improve the bounds, if the variation of these values can be characterized by some functions.

6 Concluding Remarks

The properties of the interface Schur complement preconditioners in relation to the discretization meshes were discussed in Subsect. 2.1, and, clearly, the Schur complement exhibits the same properties. Therefore, DD algorithms with the same type of Schur complement preconditioning can be efficiently used for a much wider class of discretizations.

Acknowledgements. This research is supported by the grant from the Russian Fund of Basic Research N 08-01-00676-a. The author has been partially supported by the Johann Radon Institute for Computational and Applied Mathematics (RICAM) of the Austrian Academy of Sciences during his research visits at Linz. The author is grateful to the anonymous referee for helpful suggestions and to Professor M. Jung for productive discussions of the problem.

References

[1] Adams, R.A.: Sobolev Spaces. Academic Press, New York (1975)

[2] Andreev, V.B.: Stability of difference schemes for elliptic equations with respect to Dirichlet boundary conditions. Ž. Vyčisl. Mat. i Mat. Fiz. 12, 598–611 (1972) (in Russian)

[3] Andreev, V.B.: Equivalent normalization of mesh functions from $W_2^{1/2}(\gamma)$. In: Issled. Teorii Raznost. Skhem. Ellipt. Parabol., pp. 6–39. Moskow State University Publishing House, Moskow (1973) (in Russian)

[4] Anufriev, I.E., Korneev, V.G.: Fast domain decomposition solver for degenerating PDE of 4th order in 3D domain. Math. Balkanica 20(1), 3–14 (2006)

[5] Aronszajn, N.: Boundary values of functions with finite Dirichlet integral. In: Proc. Conf. Partial Diff. Eqns., Studies in Eigenvalue Problems, University of Kansas (1955)

[6] Babich, V.M., Slobodetski, L.N.: On the boundedness of the Dirichlet integral. Dokl. Akad. Nauk SSSR 106, 604–607 (1956) (in Russian)

[7] Ben Belgacem, F.: Polynomial extensions of compatible polynomial traces in three dimensions. Comput. Methods Appl. Mech. Engrg. 160, 235–241 (1994)

[8] Beuchler, S., Schneider, R., Schwab, C.: Multiresolution weighted norm equivalence and applications. Numer. Math. 98(1), 67–97 (2004)

[9] Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring. I. Math. Comp. 47(175), 103–134 (1986)

[10] Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring. II. Math. Comp. 49(179), 1–16 (1987)

[11] Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring. III. Math. Comp. 51(184), 415–430 (1988)

[12] Bramble, J.H., Pasciak, J.E., Schatz, A.H.: The construction of preconditioners for elliptic problems by substructuring. IV. Math. Comp. 53(187), 1–24 (1989)

[13] Chen, Z., Ewing, R.E., Lazarov, R.D., Maliassov, S., Kuznetsov, Y.A.: Multilevel preconditioners for mixed methods for second order elliptic problems. Numer. Linear Algebra Appl. 3(5), 427–453 (1996)

[14] Cohen, A., Daubechies, I., Vial, P.: Biorthogonal spline wavelets on the interval – Stability and moment conditions. Appl. Comput. Harm. 6(2), 132–196 (1998)

[15] Ciarlet, P.: The finite element method for elliptic problems. North-Holland Publishing Company, Amsterdam (1978)

[16] Dolgov, S.V., Khoromskij, B.N., Oseledets, I., Tyrtyshnikov, E.E.: A reciprocal preconditioner for structured matrices arising from elliptic problems with jumping coeffcients. Linear Algebra Appl. 436(9), 2980–3007 (2012)

[17] Dahmen, W.: Wavelet and multiscale methods for operator equations. Acta Numerica 6, 55–228 (1997)

[18] Dryja, M.: A capacitance matrix method for Dirichlet problems on polygonal regions. Numer. Math. 39(1), 51–64 (1982)

[19] Dryja, M., Widlund, O.B.: Some domain decomposition algorithm for elliptic problems. In: Hayes, L., Kincaid, D. (eds.) Iterative Methods for Large Linear Systems, pp. 273–291. Academic Press, Orlando (1989)

[20] Gagliardo, E.: Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi di funzioni in più variabili. Rend. Sem. Mat. Univ. Padova 27, 284–305 (1957)

[21] Grauschopf, T., Griebel, M., Regler, H.: Additive multilevel-preconditioners based on bilinear interpolation, matrix dependent geometric coarsening and algebraic multigrid coarsening for second order elliptic pdes. Appl. Numer. Math. 23(1), 63–96 (1997)

[22] Griebel, M., Oswald, P.: Tensor product type subspace splittings and multilevel iterative methods for anisothropic problems. Adv. Comput. Math. 4, 171–201 (1995)

[23] Grisvard, P.: Elliptic problems in nonsmooth domains. Pitman Publishing, Boston (1985)

[24] Hackbusch, W.: Direct domain decomposition using the hierarchical matrix technique. In: Herrera, I., Keyes, D.E., Widlund, O.B., Yates, R. (eds.) Proceedings of the Fourteenth International Conference on Domain Decomposition Methods, Domain Decomposition Methods in Science and Engineering, National Autonomous University of Mexico, Mexico City, pp. 39–50 (2003)

[25] Hackbusch, W., Khoromskij, B.N., Kriemann, R.: Direct Schur complement method by domain decomposition based on H-matrix approximation. Comput. Visual. Sci. 8, 179–188 (2005)

[26] Hsiao, G.C., Khoromskij, B.N., Wendland, W.L.: Preconditioning for Boundary Element Methods in Domain Decomposition. Eng. Anal. Bound. Elem. 25, 323–338 (2001)

[27] Khoromskij, B.N., Wittum, G.: Robust Schur complement method for strongly anisotropic elliptic equations. J. Numer. Lin. Alg. Appl. 6, 1–33 (1999)

[28] Khoromskij, B.N., Wittum, G.: Numerical Solution of Elliptic Differential Equations by Reduction to the Interface. Lecture Notes in Computational Science and Engineering, vol. 36. Springer, Heidelberg (1975)

[29] Korneev, V.G.: Schemes of the finite element method for high orders of accuracy. Leningrad State University, Leningrad (1977) (in Russian)

[30] Korneev, V.G.: An almost optimal method for solving Dirichlet problems on decomposition subdomains of the hierarchical hp–version. Differ. Uravn. 37(7), 959–968 (2001) (in Russian)

[31] Korneev, V.G.: Local Dirichlet problems on subdomains of decomposition in *hp* discretizations, and optimal algorithms for their solution. Mat. Model. 14(5), 51–74 (2002)

[32] Korneev, V.G., Langer, U.: Domain Decomposition Methods and Preconditioning. In: Stein, E., de Borst, R., Hughes, T.J.R. (eds.) Encyclopedia of Computational Mechanics, vol. 1, pp. 617–647. John Wiley & Sons (2004)

[33] Korneev, V.G., Poborchi, S.V., Salgado, A.G.: Interface Schur complement preconditioning for piecewise orthotropic discretizations with high aspect ratios. Preprint N37. RICAM, Linz (2006)

[34] Korneev, V.G., Poborchi, S.V., Salgado, A.G.: Interface Schur complement preconditioning for piecewise orthotropic discretizations with high aspect ratios. In: Korneev, V.G. (ed.) Fast Discrete Methods of Computational Mechanics of Continuous Media, pp. 106–159. Publications of Chemistry Research Institute of St. Petersburg University, St. Petersburg (2007)

[35] Korneev, V.G., Xanthis, L., Anufriev, I.E.: Fast adaptive domain decomposition algorithms for *hp*-discretizations of 2-d and 3-d elliptic equations: recent advances. Int. J. Computer Math. Appl. 4, 27–44 (2003)

[36] Korneev, V.G., Xanthis, L., Anufriev, I.E.: Solving finite element *hp* discretizations of elliptic problems by fast DD algorithms. Prikladnaya Matematika, Trudy SPbGPU (Applied Mathematics, Transactions of the St. Petersburg State Polytechnical University) 485, 126–153 (2002)

[37] Korneev, V.G., Xanthis, L., Anufriev, I.E.: Hierarchical and Lagrange *hp* discretizations and fast domain decomposition solvers for them. SFB Report 02-18. Johannes Kepler University, Linz (2002)

[38] Kwak, D.Y., Nepomnyaschikh, S.V., Pyo, H.C.: Domain decomposition for model heterogeneous anisotropic problem. Numer. Lin. Alg. Appl. 10(1-2), 129–157 (2004)

[39] Maz'ya, V., Poborchi, S.: Differentiable functions on bad domains. World Scientific, Singapore (1998)

[40] Nepomnyashchikh, S.V.: Domain decomposition method for elliptic problems with discontinuous coefficients. In: Glowinski, R., et al. (eds.) Domain Decomposition for PDEs, pp. 242–252. SIAM (1990)

[41] Nepomnyaschikh, S.V.: The domain decomposition method for elliptic problems with coefficient jumps in thin strips. Dokl. Akad. Nauk. 323(6), 1034–1037 (1992)

[42] Nepomnyashchikh, S.V.: Domain decomposition methods. Lectures on advanced computational methods in mechanics. Radon Ser. Comput. Appl. Math. 1, 89–159 (2007)

[43] Nepomnyashchikh, S.V.: Metody decompozicii oblasti i fictivnogo prostranstva. Diss. na soiskanie stepeni doktora fiz.-mat. nauk. Novosibirskii gos. universitet (Methods of domain decomposition and ficticious space. Dissertation for sci. degree of doctor of fisical-mathematical sciences), Novosibirsk (2008)

[44] Oswald, P.: Interface preconditioners and multilevel extension operators. In: Lai, C.-H., et al. (eds.) Proceedings of the 11th International Conference on Domain Decomposition Methods in Science and Engineering (London 1998), pp. 97–104 (1999), `ddm.org`

[45] Pflaum, C.: Robust convergence of multilevel algorithms for convection-diffusion equations. SIAM J. Num. Anal. 37(2), 443–469 (2000)

[46] Rytov, A.V.: Numerical testing of the domain decomposition method for deteriorating second order elliptic equation (Chislennoe testirovanie metoda dekompozitsii oblasti dlj vyrozhdayuschegosj ellipticheskogo uravnenij vtorogo porjdka). Scientific-Industrial Bulletin of the St. Petersburg State Polytechnical University (Nauchno-tehnicheskie vedomosti SPbGPU) 4, 1–13 (2006) (in Russian)

[47] Schieweck, N.: A multigrid convergence proof by a strengthened Cauchy inequality for symmetric elliptic boundary value problems. In: Telschow, G. (ed.) Second Multigrid Seminar (Garzau 1985), pp. 49–62 (1985)

[48] Schneider, R.: Multiskalen– und Wavelet–Matrixkompression. B.G. Teubner, Stuttgart (1998)

[49] Scott, L., Zhang, S.: Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math. Comput. 54, 483–493 (1990)

[50] Wittum, G.: Linear iterations as smoothers in multigrid methods: theory with applications to incomplete decompositions. Impact Comput. Sci. Engrg. 1, 180–215 (1989)

[51] Xu, J., Zou, J.: Some nonoverlapping domain decomposition methods. SIAM Rev. 40(4), 857–914 (1998)

[52] Zhang, X.: Multilevel Schwarz methods. Numer. Math. 63(4), 521–539 (1992)

Inexact Additive Schwarz Solvers for *hp*-FEM Discretizations in Three Dimensions

Sven Beuchler

Abstract. In this paper, a boundary value problem of second order in three space dimensions is discretized by means of the *hp*-version of the finite element method. The system of linear algebraic equations is solved by the preconditioned conjugate gradient method with an overlapping domain decomposition preconditioner with inexact subproblem solvers. In addition to a global solver for the low order functions, the ingredients of this preconditioner are local solvers for the patches. Here, a solver is used which utilizes the tensor product structure of the patches. The efficiency in time and iteration numbers of the presented solver is shown in several numerical examples for diffusion like problems as well as for problems in linear elasticity.

1 Introduction

The numerical solution of boundary value problems (BVP) of partial differential equations (PDE) is one of the major challenges in Computational Mathematics. Finite element methods (FEM) are among to the most powerful tools in order to compute an approximate solution of BVPs. For the *h*-version of the FEM, the polynomial degree *p* of the shape functions on the elements is kept constant and the mesh-size *h* is decreased. This is in contrast to the *p*-version of the FEM in which the polynomial degree *p* is increased and the mesh-size *h* is kept constant. Both ideas, mesh refinement and increasing the polynomial degree, can be combined. This is called the *hp*-version of the FEM. The advantage of the *p*-version in comparison to the *h*-version is that the solution converges much faster to the exact solution with respect to the dimension *N* of the approximation space, see e.g. [13, 32, 33], and the references therein as well as [20] for the related spectral element methods.

Sven Beuchler
Institut für Numerische Simulation, Rheinische Friedrich–Wilhelms–Universität Bonn,
Wegelerstraße 6, 53115 Bonn, Germany
e-mail: beuchler@ins.uni-bonn.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 91–108.
springerlink.com

From the literature it is known, see e.g. [34], that preconditioned conjugate gradient (PCG) methods with additive Schwarz preconditioners (ASM) as domain decomposition (DD) are a powerful tool for the development of fast and efficient solvers for the h-version as well as for the p-version of the FEM. One class are nonoverlapping DD preconditioners with inexact subproblem solvers on the subdomains, [17]. This preconditioner requires a solver related to the Dirichlet problem on the elements $\triangle_s$, see [4, 8, 14, 23], a solver related to the Schur complement corresponding to the subdomain boundaries, see [1, 19, 21], and an approximate discrete harmonic extension from $\partial\triangle_s$ to $\triangle_s$, see [3, 5, 12, 16, 26]. This leads to quasioptimal solvers for hp-FEM discretizations in two space dimensions since the coupling between the high order basis functions and the low order basis functions can be removed by paying a $\log p$ term in the condition number estimates, [3].

In the three-dimensional case, the usage of nonoverlapping ASM preconditioners is much more difficult due to the coupling between the different types of basis functions, see e.g. [15, 24, 28, 29] and the references therein. Another approach is using overlapping preconditioners as developed in [27], see also [31] for the tetrahedral case. This decouples the low order basis functions from the high order basis functions used in p-FEM. It remains the solution of high-order systems on patches consisting of about 8 hexahedrons. In [6], fast solvers for the patch structure are proposed. They incorporate the tensor product structure of the patches and are an extension of the results presented in [8]. Condition number estimates on the patches and the construction principle are also presented in [6]. We also refer to the recent publications [2, 11, 22].

The purpose of this paper is the presentation of the performance of the overlapping DD preconditioner [27] combined with the patch preconditioner of [6] as subproblem solver. Our first numerical experiments of this DD preconditioner for the p-version of the FEM are presented for scalar elliptic problems as well as for the system of Lamé equations. In addition, the final condition number estimates are given in the theoretical part of the paper.

The outline of this paper is as follows. Sect. 2 describes the setting of the problem and the discretization. The definition of the preconditioners and the condition number estimates are presented in Sect. 3. The main part of this paper is devoted to Sect. 4. Several numerical experiments show the efficiency of the proposed solvers. Sect. 5 concludes the paper with possible generalizations of the presented results.

Throughout this paper, the integer p denotes the polynomial degree. For two real symmetric and positive definite $n \times n$ matrices A, B, the elation $A \preceq B$ means that $A - cB$ is negative definite, where $c > 0$ is a constant independent of n, or p. The relation $A \sim B$ means $A \preceq B$ and $B \preceq A$, i.e. the matrices A and B are spectrally equivalent. The isomorphism between a function $u = \sum_i u_i \psi_i \in L^2$ and the vector of coefficients $\underline{u} = [u_i]_i$ with respect to the basis $[\Psi] = [\psi_1, \psi_2, \ldots]$ is denoted as $u = [\Psi]\underline{u}$.

2 Setting of the Problem, Discretization

In this paper, we consider the following boundary value problem. Let $\Omega \subset \mathbb{R}^3$ be a bounded Lipschitz domain and $\mathbb{V}$ be a Sobolev space on Ω. Moreover, let $a : \mathbb{V} \times \mathbb{V} \mapsto \mathbb{R}$ be a $\mathbb{V}$ elliptic and bounded bilinear form and $F : \mathbb{V} \mapsto \mathbb{R}$ be a bounded linear functional. Then, we are looking for solutions of

$$\text{Find } u \in \mathbb{V} \quad \text{such that} \quad a(u,v) = F(v) \quad \forall v \in \mathbb{V}. \tag{1}$$

Throughout this paper, the following types of BVPs are investigated:

- Scalar elliptic problems of second order: Here, the corresponding Sobolev space is defined as $\mathbb{V} = H^1_{\Gamma_1}(\Omega) := \{u \in H^1(\Omega), u_{|\Gamma_1} = 0\}$ with $\Gamma_1 \subset \partial\Omega$. We assume that $meas(\Gamma_1) > 0$. The bilinear form and the right hand side are given as

$$\begin{aligned} a(u,v) &= \int\limits_{\Omega} (\nabla u(x) \cdot D(x) \nabla v(x) + c(x) u(x) v(x))\, dx, \\ F(v) &= \int\limits_{\Omega} f(x) v(x)\, dx + \int\limits_{\partial\Omega \setminus \Gamma_1} f_1(x) v(x)\, dS, \end{aligned} \tag{2}$$

  respectively. The functions $D : \overline{\Omega} \mapsto \mathbb{R}^+$ and $c : \overline{\Omega} \mapsto \mathbb{R}^+_0$ are assumed to be bounded and piecewise constant whereas $f \in L^2(\Omega)$, $f_1 \in H^{-1/2}(\partial\Omega)$.
- The system of Lamé equations of linear elasticity: The Sobolev space is defined as $\mathbb{V} = (H^1_{\Gamma_1}(\Omega))^3$ with $meas(\Gamma_1) > 0$. The bilinear form and the right hand side are given as

$$\begin{aligned} a(u,v) &= \int\limits_{\Omega} \frac{E}{1+\nu} \left(\varepsilon(u) : \varepsilon(v) + \frac{\nu}{1-2\nu} \nabla \cdot u(x)\ \nabla \cdot v(x) \right) dx, \\ F(v) &= \int\limits_{\Omega} f(x) \cdot v(x)\, dx + \int\limits_{\partial\Omega \setminus \Gamma_1} v(x) \cdot f_1(x)\, dS, \end{aligned} \tag{3}$$

  respectively. The functions $E : \overline{\Omega} \mapsto \mathbb{R}^+$ and $\nu : \overline{\Omega} \mapsto (0, \frac{1}{2})$ are assumed to be bounded and piecewise constant whereas $f \in (L^2(\Omega))^3$ and $f_1 \in (H^{-1/2}(\partial\Omega))^3$. The strain tensor is defined as $2\varepsilon(u) = \nabla u + (\nabla u)^\top$.

Problem (1) is discretized by means of the *hp*-version of the finite element method using hexahedral elements $\triangle_s$, $s = 1, \ldots, nel$. Let $\hat{\triangle} = (-1,1)^3$ be the reference hexahedron and $F_s : \hat{\triangle} \to \triangle_s$ be the isoparametric mapping to the element $\triangle_s$. We define the space

$$\mathbb{M}_p := \left(\{u \in H^1_{\Gamma_1}(\Omega), u_{|\triangle_s} = \tilde{u}(F_s^{-1}(x,y,z)), \tilde{u} \in \mathbb{Q}_p\} \right)^d$$

with $d = 1$ or $d = 3$, where $\mathbb{Q}_p$ is the space of all polynomials of maximal degree p in each variable. In order to obtain a basis for $\mathbb{M}_p$, let

$$\hat{L}_i(x) = \frac{1}{2}(2i-1)\int_{-1}^{x} L_{i-1}(s)\,ds, \quad i = 1,\dots,p, \quad \hat{L}_0(x) = \frac{1-x}{2} \tag{4}$$

be the i-th integrated Legendre polynomial where

$$L_i(x) = \frac{1}{2^i i!}\frac{\mathrm{d}^i}{\mathrm{d}x^i}(x^2-1)^i$$

denotes the i-th Legendre polynomial. On the reference element $\hat{\triangle} = (-1,1)^3$, the local basis functions

$$\hat{L}_{ijk}(x,y,z) = \hat{L}_i(x)\hat{L}_j(y)\hat{L}_k(z), \quad i,j,k = 0,\dots,p \tag{5}$$

are used. Since $\hat{L}_i(\pm 1) = 0$, $i \geq 2$, the global basis $[\mathrm{H}] = (\zeta_1,\dots,\zeta_N)$ for $\mathbb{M}_p$ is built in the usual way, by using the vertex functions (V), the edge bubble functions (E), face bubble functions (F), and the interior bubble functions (I), locally on each element $\triangle_s$, and globally on Ω. We refer the interested reader to [13] and the references therein concerning the details.

The Galerkin projection of (1) onto the N-dimensional space $\mathbb{M}_p$ leads to the linear system of algebraic finite element equations

$$\mathscr{K}_\zeta \underline{u} = \underline{f}, \quad \text{where} \quad \mathscr{K}_\zeta = [a(\zeta_j,\zeta_i)]_{i,j=1}^{N}, \quad \underline{f} = [F(\zeta_i)]_{i=1}^{N}. \tag{6}$$

Using the vector $\underline{u}$, an approximation $u_p = [\mathrm{H}]\underline{u}$ of the exact solution u of (1) is obtained by the usual finite element isomorphism.

3 The Solution of the Linear System

This section is devoted to the solution of the system of linear algebraic equations (6). It is distinguished between discretizations of scalar elliptic equations as in (2) and the system of Lamé equations as in (3).

3.1 The Scalar Elliptic Case

In this subsection, we consider the bilinear form $a(\cdot,\cdot)$ (2) which is elliptic and bounded on the Sobolev space $\mathbb{V} = H^1_{\Gamma_1}(\Omega)$. It is intended to use an overlapping DD preconditioner which has been developed by Pavarino, [27]. For the definition of the preconditioner, some notation is introduced. Let

$$\mathbb{U}_0 = \left\{u \in H^1_{\Gamma_1}(\Omega), u_{|\triangle_s} = \tilde{u}(F_s^{-1}(x,y,z)), \tilde{u} \in \mathbb{Q}_1\right\} \tag{7}$$

be the space of all finite element functions of maximal polynomial degree 1. For a given node v, let $\Omega_{\mathrm{v}} = \{\cup_s \overline{\triangle}_s, \mathrm{v} \subset \overline{\triangle_s}\}$ be the closed patch associated to a node v of the finite element mesh. Then, for each node v of the finite element mesh, we introduce

$$\mathbb{U}_{\mathrm{v}} = \left\{u \in \mathbb{M}_p, \operatorname{supp} u \subset \Omega_{\mathrm{v}}\right\} \tag{8}$$

as the patch space, cf. an analogous two-dimensional example in Fig. 1.

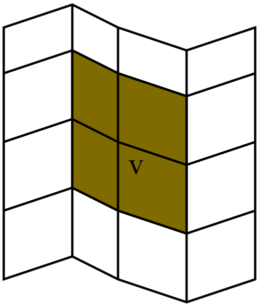


Fig. 1 Patch Ω_{v} of a node v (2D) (marked colored).

Theorem 1. *Let $\mathbb{U}_v$ and $\mathbb{U}_0$ be defined via* (8) *and* (7)*, respectively. Then, for all $u \in \mathbb{M}_p$ there exists a decomposition $u = u_0 + \sum_{\mathrm{v}} u_{\mathrm{v}}$, $u_0 \in \mathbb{U}_0$, $u_{\mathrm{v}} \in \mathbb{U}_{\mathrm{v}}$ such that*

$$a(u,u) \succeq b(u,u) := \inf_{u=u_0+u_{\mathrm{v}}} \left(a(u_0,u_0) + \sum_{\mathrm{v}} a(u_{\mathrm{v}},u_{\mathrm{v}}) \right).$$

Moreover, for all decompositions $u = u_0 + \sum_{\mathrm{v}} u_{\mathrm{v}}, u_0 \in \mathbb{U}_0, u_{\mathrm{v}} \in \mathbb{U}_{\mathrm{v}}$

$$a(u,u) \preceq a(u_0,u_0) + \sum_{\mathrm{v}} a(u_{\mathrm{v}},u_{\mathrm{v}}).$$

The constants depend neither on h nor p.

Proof. This result has been proven by Pavarino [27]. □

Remark 1. The bilinear form $b(\cdot,\cdot)$ in Theorem 1 defines a preconditioner C_ζ for $\mathscr{K}_\zeta$ (6) in the following way. Let $J(\mathrm{v}) = \left[j_1^{\mathrm{v}}, \ldots, j_{n_{\mathrm{v}}}^{\mathrm{v}}\right]$ be the index set of all basis functions ζ_j with $\operatorname{supp}(\zeta_j) \subset \Omega_{\mathrm{v}}$ and $J(0)$ the index set of all vertex functions (V). Due to the partition of $[\mathrm{H}]$ into vertex, edge, face and interior functions, the set $[\zeta_j]_{j\in J(\mathrm{v})}$ forms a n_{v} dimensional basis of the space $\mathbb{U}_{\mathrm{v}}$. Let $P_{\mathrm{v}} \in \mathbb{R}^{n_{\mathrm{v}}\times N}$ be the Boolean matrix with the entries

$$[P_{\mathrm{v}}]_{ij} = \begin{cases} 1 & \text{if } j = j_i^{\mathrm{v}}, 1 \le i \le n_{\mathrm{v}}, \\ 0 & \text{else.} \end{cases}$$

Finally, let

$$C_{\mathrm{V}} = \left[a(\zeta_{j_i^{\mathrm{V}}}, \zeta_{j_k^{\mathrm{V}}})\right]_{i,k=1}^{n_{\mathrm{V}}}. \tag{9}$$

In the same way, P_0 and C_0 corresponding to the set $J(0)$ are introduced. Then, the splitting in Theorem 1 introduces the preconditioner

$$C_{\zeta}^{-1} = P_0^{\top} C_0^{-1} P_0 + \sum_{\mathrm{V}} P_{\mathrm{V}}^{\top} C_{\mathrm{V}}^{-1} P_{\mathrm{V}} \tag{10}$$

with $\mathscr{K}_{\zeta} \sim C_{\zeta}$, see e.g. [34].

Therefore, in the sense of inexact ASM it suffices to develop preconditioners for C_0 and C_{V} where v is running over all non Dirichlet nodes v of the mesh. The system for C_0 corresponds to all low order basis functions. Here, many solvers in the sense of inexact additive Schwarz preconditioners are available for a solution in optimal arithmetical complexity. Examples are multigrid methods, [18], and PCG methods with the BPX-preconditioner, [10], for structured meshes and algebraic multigrid methods (AMG), [30], for unstructured meshes. For the construction of the preconditioner for C_{V} (9), let us make the following mesh

Assumption 1. *Each patch Ω_{V} corresponding to an interior node is the union of eight hexahedrons, two in each space direction. Patches to Neumann nodes on faces of Ω are assumed to be the union of four hexahedrons.*

For interior nodes, the patch Ω_{V} is transformed to the reference patch $\hat{\Omega} = [-2,2]^3$ consisting of eight cubes of length 2. Let $\mathbb{U}_{\mathrm{V}}$ be defined as in (8). This space is equipped with the basis of the integrated Legendre polynomials of (5) denoted as $[\Phi_3] = [\phi_{I,3}]_{I=1}^{M}$, $M = (2p-1)^3$. Using the lexicographic ordering for the local functions, (5) and the structure of $\hat{\Omega}$, the basis functions can be expressed as products of one-dimensional basis functions, e.g.

$$\begin{aligned} \phi_{I,3}(x,y,z) &= \phi_{i,1}(x)\phi_{j,1}(y)\phi_{k,1}(z), \quad 1 \le i,j,k \le 2p-1, \\ I &= (2p-1)^2(k-1) + (2p-1)(j-1) + i, \end{aligned} \tag{11}$$

with the one-dimensional functions $[\Phi_1] := [\phi_{i,1}]_{i=1}^{2p-1}$. These one-dimensional functions are shifted integrated Legendre polynomials (4). More precisely,

$$\begin{aligned} \phi_{1,1}(x) &= \frac{1}{2}\begin{cases} 2+x, & x \in [-2,0], \\ 2-x, & x \in [0,2], \\ 0, & \text{else}, \end{cases} \\ \phi_{i,1}(x+1) &= \begin{cases} \hat{L}_i(x), & |x| \le 1, \\ 0, & \text{else}, \end{cases} \quad \text{for } i = 2,\ldots,p, \\ \phi_{p+i-1,1}(x-1) &= \begin{cases} (-1)^i \hat{L}_i(x), & |x| \le 1, \\ 0, & \text{else} \end{cases} \quad \text{for } i = 2,\ldots,p. \end{aligned} \tag{12}$$

The basis functions (12) for $p = 3$ are displayed in Fig. 2.

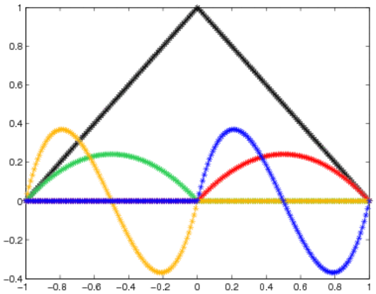


Fig. 2 One dimensional basis functions on the patch for $p = 3$.

The finite element isomorphism introduces the matrix $\hat{K}_{3,p}$ by the relation

$$a(u_p, v_p) = \underline{u}^\top \hat{K}_{3,p} \underline{v} \quad \forall u_p = [\Phi_3]\underline{u}, v_p = [\Phi_3]\underline{v}. \tag{13}$$

For the bilinear form of the Laplacian $a(u,v) = \int_{\hat{\Omega}} \nabla u \cdot \nabla v$, equations (11) and (13) imply

$$\hat{K}_{3,p} = \hat{K}_{1,p} \otimes \hat{M}_{1,p} \otimes \hat{M}_{1,p} + \hat{M}_{1,p} \otimes \hat{K}_{1,p} \otimes \hat{M}_{1,p} + \hat{M}_{1,p} \otimes \hat{M}_{1,p} \otimes \hat{K}_{1,p} \tag{14}$$

with the one-dimensional mass and stiffness matrix

$$\hat{M}_{1,p} = \int_{-2}^{2} [\Phi_1]^\top [\Phi_1]\, dx \quad \text{and} \quad \hat{K}_{1,p} = \int_{-2}^{2} \frac{d}{dx}[\Phi_1]^\top \frac{d}{dx}[\Phi_1]\, dx, \tag{15}$$

respectively.

Concerning the basis $[\Phi_1]$, a wavelet based basis $[\Psi_p]$ has been introduced in [6, (3.46)] by the relation

$$[\Psi_p] = [\Phi_1] W_p \tag{16}$$

where $W_p \in \mathbb{R}^{(2p-1)\times(2p-1)}$ is the corresponding nonsingular basis transformation matrix. This basis is almost stable in $L_2(-2,2)$ and $H_0^1(-2,2)$ as the following theorem states.

Theorem 2. *Let* $[\Psi_p]$ *be defined via* (16)*. Moreover, let* $\chi > 1$*. Then, the relations*

$$c_{m,1}^{-1}(D_M \underline{u}, \underline{u}) \leq \|u\|_{L_2(-2,2)}^2 \leq (\log p \log^\chi \log p) c_{m,2}(D_M \underline{u}, \underline{u}), \tag{17}$$

and

$$(\log p \log^\chi \log p)^{-1} c_{k,1}^{-1}(D_K \underline{u}, \underline{u}) \leq \|u\|_{H^1(-2,2)}^2 \leq c_{k,2}(D_K \underline{u}, \underline{u}) \tag{18}$$

hold for any $u = [\Psi_p]\underline{u}$, $\underline{u} \in \mathbb{R}^{2p-1}$, where D_M and D_K are suitable chosen diagonal matrices. The constants $c_{m,1}$, $c_{k,2}$, $c_{m,2}$ and $c_{k,1}$ are independent of p. Moreover, the operations $W_p\underline{x}$ and $W_p^\top \underline{x}$ require $\mathscr{O}(p)$ floating point operations.

Proof. The constructive proof which includes also the definitions of D_M and D_K has been given in [6]. □

Due to (15) and (16), the relations (17) and (18) are equivalent to the spectral equivalence relations

$$\begin{aligned} c_{m,1}^{-1} D_M \leq W_p^\top \hat{M}_{1,p} W_p &\leq (\log p \log^\chi \log p) c_{m,2} D_M \quad \text{and} \\ (\log p \log^\chi \log p)^{-1} c_{k,1}^{-1} D_K \leq W_p^\top \hat{K}_{1,p} W_p &\leq c_{k,2} D_K. \end{aligned} \tag{19}$$

Hence, we are able to introduce the preconditioner

$$\begin{aligned} \hat{C}_{3,p}^{-1} = &(W_p \otimes W_p \otimes W_p) \\ &(D_K \otimes D_M \otimes D_M + D_M \otimes D_K \otimes D_M + D_M \otimes D_M \otimes D_K)^{-1} (W_p \otimes W_p \otimes W_p)^\top \end{aligned} \tag{20}$$

for $\hat{K}_{3,p}$ (14). Using (14), (20), (19) and the properties of the Kronecker product, we arrive at

$$\frac{1}{\log p \log^\chi \log p} \hat{C}_{3,p} \preceq \hat{K}_{3,p} \preceq (\log p \log^\chi \log p)^2 \hat{C}_{3,p} \tag{21}$$

for any $\chi > 1$.

Remark 2. The results can be extended to Neumann boundary nodes. However, a different wavelets basis, see [6, Theorem 3.9], has to be used into the space directions of one layer of elements instead of (16).

Finally, the ASM preconditioner with inexact subproblem solvers is defined as

$$C_{in,\zeta}^{-1} = P_0^\top C_{BPX}^{-1} P_0 + \sum_{\mathrm{v}} P_{\mathrm{v}}^\top \hat{C}_{3,p}^{-1} P_{\mathrm{v}}, \tag{22}$$

where C_{BPX} denotes the BPX-preconditioner, [10], and $\hat{C}_{3,p}$ is the preconditioner (20). Then, the following result is formulated.

Theorem 3. *Let $a(\cdot,\cdot)$ be the bilinear form (2) and let $C_{in,\zeta}$ be defined by (22). Moreover, let us assume that Assumption 1 is satisfied. Then, the condition number estimate $\kappa(C_{in,\zeta}^{-1}\mathscr{K}_\zeta) \preceq (\log p \log^\chi \log p)^3$ holds for any $\chi > 1$ where the constant is independent on h and p but may depend on D, c and the geometry. Moreover, the action $C_{in,\zeta}^{-1}\underline{r}$ requires $\mathscr{O}(N)$ operations.*

Proof. Due to Assumption 1, we have $\hat{K}_{3,p} \sim C_{\mathrm{v}}$ for all interior nodes v, see also [19]. The result also holds for Neumann nodes with some modifications, cf. Remark 2. Using (21), Theorem 1 and the properties of the BPX-preconditioner, [35], the assertion follows. □

3.2 The Preconditioner for Lamé

In this subsection, we assume that the bilinear form $a(\cdot,\cdot)$ has the form (3) which is the bilinear form for linear elasticity. Using Korn's inequality, it can be proved, see e.g. [9], that

$$c_1\|u\|^2_{H^1(\Omega)} \le a(u,u) \le c_2\|u\|^2_{H^1(\Omega)} \quad \forall u \in (H^1_{\Gamma_1}(\Omega))^3. \tag{23}$$

This suggests to choose the preconditioner

$$C_{Lame,\zeta} = \begin{bmatrix} C_{in,\zeta} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & C_{in,\zeta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & C_{in,\zeta} \end{bmatrix} \tag{24}$$

for $\mathcal{K}_\zeta$. Then, the following result can be proved.

Theorem 4. *Let $a(\cdot,\cdot)$ be the bilinear form* (3) *and let $C_{Lame,\zeta}$ be defined by* (24). *Moreover, let us assume that Assumption 1 is satisfied. Then, $\kappa(C^{-1}_{Lame,\zeta}\mathcal{K}_\zeta) \preceq (\log p \log^\chi \log p)^3$ for any $\chi > 1$ where the constant is independent on h and p but may depend on E, ν and the geometry. The action $C^{-1}_{Lame,\zeta}\underline{r}$ requires $\mathcal{O}(N)$ operations.*

Proof. The result is a direct consequence of (24), (23) and Theorem 3. □

4 Numerical Experiments

In this section, several numerical experiments show the performance of the preconditioners (22) and (24) for scalar elliptic equations and the system of Lamé equations, respectively. The preconditioners have been implemented into the *hp*-version of the program `SPCad3H`, [7]. Note that the stiffness matrix is not assembled, only the action $\mathcal{K}_\zeta\underline{u}$ is implemented. Using sum factorization techniques, [25], this can be performed in at most $\mathcal{O}(p^4)$ operations. Moreover, the cost can be reduced to $\mathcal{O}(p^3)$ if the stiffness matrix is sparse.

Several cases are considered. In each example, a coarse finite element mesh of Level 1 is read from a data file. The computational mesh is obtained by uniform refinement with respect to h and p. In all examples, the corresponding system of linear equations is solved with the pcg-method using the preconditioners (22) and (24) for scalar elliptic problems and linear elasticity, respectively. The relative accuracy of $\varepsilon = 10^{-5}$ is chosen.

For two examples, cf. Tables 4 and 8, the number of unknowns are also displayed in order to give the reader an impression of the size of the problems. For the unit cube problems, the dimension N can easily be computed by the formula

$$N = \mathtt{ndof} \cdot (2^{l-1}p+1)^3,$$

where p is the polynomial degree, l denotes the level of refinement, `ndof` is 1 for scalar elliptic problems and 3 for linear elasticity problems.

4.1 Scalar Elliptic Problems

The first example is the Poisson equation on the unit `cube` with pure Dirichlet boundary conditions. More precisely,

- the bilinear form is chosen as (2) with $D(x) = 1$ and $c(x) = 0$,
- the right hand side is chosen randomly,
- the computational domain is $\Omega = (0,1)^3$ and $\Gamma_1 = \partial\Omega$,
- the coarse mesh consists of one element.

The iteration numbers and the computational time are displayed in Table 1.

Table 1 PCG iterations on the unit `cube` with preconditioner (22).

Levels	2		3		4		5		6	
p	It	Time[sec]	It	Time[sec]	It	Time[sec]	It	Time[sec]	It	Time[sec]
3	2	0.01	14	0.09	16	0.83	18	9.13	18	80.04
5	2	0.02	14	0.41	17	4.82	18	48.54		.
7	11	0.06	25	1.73	25	18.18	23	148.57		.
9	9	0.12	22	3.18	21	34.82	20	296.88		.
11	12	0.28	29	7.51	28	79.26		.		.
13	15	0.59	37	16.55	34	153.59		.		.
15	16	1.02	38	25.00	35	224.81		.		.
17	12	1.12	29	29.77	27	268.38		.		.
19	16	2.16		.		.		.		.
21	17	3.24		.		.		.		.
33	15	15.53		.		.		.		.

For comparison, Table 2 displays the PCG iteration numbers and computational time with a diagonal preconditioner $\mathrm{diag}(\hat{K}_{3,p})$ instead of $\hat{C}_{3,p}$.

From the results, it can be observed that the application of the preconditioner (22) reduces the iteration numbers and computational time dramatically. The numbers of iterations of the PCG method grow moderately with respect to p. One iteration with preconditioner (22) requires about two to three times compared to a pcg-iteration of the unpreconditioned system. With a higher relative accuracy of the pcg method, similar results can be observed. For example, 85 and 130 iterations are required in order to reduce the initial error up to a factor of $\varepsilon = 10^{-10}$ and $\varepsilon = 10^{-15}$, respectively, for $p = 15$ and $Level = 3$.

Due to the Kronecker product structure, the wavelet preconditioner (20) can be made robust against anisotropies of the diffusion coefficient in (2) see [6, Rem. 4.2].

Table 2 PCG iterations on the unit `cube` with diagonal preconditioner.

Levels	2		3		4		5		6	
p	It	Time[sec]	It	Time[sec]	It	Time[sec]	It	Time[sec]	It	Time[sec]
3	11	0.01	64	0.23	94	2.77	96	218.41	93	174.67
5	46	0.08	229	3.16	251	27.31	250	22.08		.
7	128	0.61	443	16.79	472	141.12	464	1115.28		.
9	261	2.56	643	49.92	761	471.69	747	3712.70		.
11	451	8.44	884	131.13	1084	1288.15		.		.
13	662	21.50	1234	314.39	1293	2648.33		.		.
15	882	49.73	1634	730.42	1685	6048.34		.		.
17	1128	89.72	2075	1515.78	2124	10822.79		.		.
19	1430	170.54		.		.		.		.
21	1738	291.68		.		.		.		.
33	4323	4084.91		.		.		.		.

However, this structure is lost by using the preconditioner (21). In order to check this, the next example `cube.aniso` uses

- the diffusion coefficient $D(x) = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 1000 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

The other parameters are chosen as in the previous example `cube`. The results are displayed in Table 3.

Table 3 PCG iteration numbers for `cube.aniso` with anisotropic diffusion.

p	1	3	5	7	9	11	13	15	17	19	21	33
Level												
2	2	2	2	15	11	16	18	21	15	19	21	18
3	9	103	291	492	739	955	1155	1237	1297			
4	28	220	498	793	1094	1357	1667	1710	1745			
5	75	411	826	1232	1656							
6	125	663										

The PCG iteration numbers do not blow up only in Level 2 where the summation in (22) is running over one node v only, e.g. the preconditioner is almost of the form (20). Then, the robustness of the wavelet construction can be observed. Otherwise, no robustness against anisotropies can be observed.

The next example considers the `Fichera` corner. More precisely,

- the bilinear form is chosen as (2) with $D(x) = 1$ and $c(x) = 0$,
- the right hand side is $f = 1$,

- the computational domain is $\overline{\Omega} = [-1,1]^3 \backslash (0,1]^3$ and $\Gamma_1 = \partial\Omega$,
- the coarse mesh consists of seven congruent cubes of volume 1.

Slices of the solution and the coarse mesh are displayed in Fig. 3.

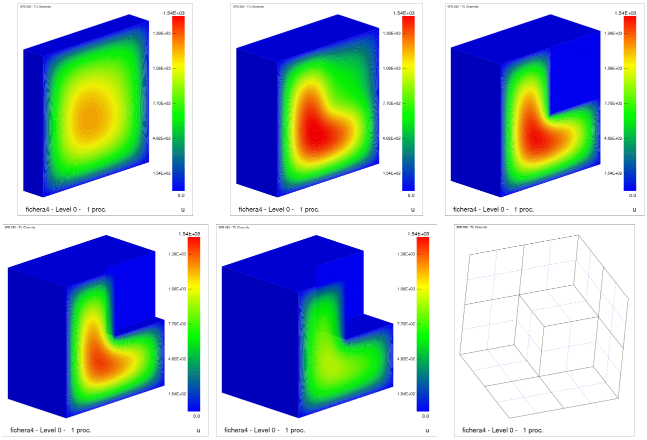


Fig. 3 Solution for the `Fichera` corner (slices at different x-values) and coarse mesh (right, below).

The PCG iteration numbers are displayed in Table 4. The behavior is similar in comparison to the unit `cube`.

Table 4 PCG-iteration numbers for the `Fichera` corner (above) and Number of unknowns N (below).

p	3	5	7	9	11	13	15	17	19
Level									
2	18	19	28	26	31	37	37	31	37
3	30	31	36	34	37	39	39		
4	35	35	39	37					
5	36	37							
6	31								
2	1981	8261	21645	44821	80477	131301	199981	289205	401661
3	13567	60183	161741	339835	615969	1011647	1548373		
4	102601	462077	1255521	2654965					
5	793363	3622725							
6	6241393								

The influence of coefficient `jumps` is investigated in the following coarse mesh consisting of two cubes. More precisely,

- the bilinear form is chosen as (2) with $D(x) = \begin{cases} 1, & z > 0, \\ b, & z \leq 0, \end{cases}$ and $c(x) = 0$,
- the right hand side is chosen as $f(x) = \begin{cases} 1, & z > 0, \\ 0, & z \leq 0, \end{cases}$
- the computational domain is $\Omega = (-1,1)^2 \times (-2,2)$ and $\Gamma_1 = \partial\Omega$,
- the coarse mesh consists of two congruent cubes of volume 2.

As observed in Table 5, the dependence of the jumps of the coefficients in the diffusion results in an increase of the PCG iteration numbers, in particular for higher levels of refinement.

Table 5 PCG iteration numbers for coefficient `jumps` $b = 1$ (left) and $b = 100$ (right).

p	3	5	7	9	11	13	15	17	19	21	33
Level											
2	6	6	16	13	18	22	23	17	23	24	22
3	17	17	26	22	29	39	40	29			
4	18	19	25	21	27	33					
5	19	19									

p	3	5	7	9	11	13	15	17	19	33
Level										
2	7	7	48	35	58	70	77	51	73	71
3	52	51	133	105	150	185	185	145		
4	63	62	113	88	126	145				
5	57	50								

4.2 Lamé Equations of Linear Elasticity

The next examples consider the system of Lamé equations (3). More precisely,

- the bilinear form is chosen as (3) with $E = 10^6$ and $\nu = 0.3$,
- the right hand side is chosen as $f = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^\top$,
- the computational domain is the unit `cube3.030` $\Omega = (0,1)^3$ or the domain $\overline{\Omega} = [-3,3]^3 \backslash A$ with the `holes`, see Fig. 4,
  $A = \{(-1,1) \times (-3,3)^2 \cup (-3,3) \times (-1,1) \times (-3,3) \cup (-3,3)^2 \times (-1,1)\}$.
- all boundary conditions are chosen to be Dirichlet, e.g. $\Gamma_1 = \partial\Omega$,
- the coarse mesh consists of one element or twenty elements of volume eight, respectively.

Now, the preconditioner (24) is used. The results are displayed in Table 6.

The moderate increase of iteration numbers is similar to the Poisson case. In both cases, the absolute numbers are moderately higher than for Poisson which is due to (23).

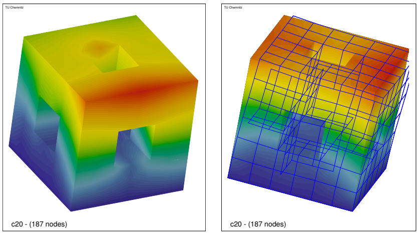


Fig. 4 Computational domain with `holes` and `holesN` (left), Displacements for `holesN` (right).

Table 6 PCG iteration numbers for linear elasticity, unit `cube3.030` (left), domain with `holes` (right).

p	3	5	7	9	11	13	15	17	19	33
Level										
2	10	10	18	18	21	26	27	21	17	25
3	27	29	42	38	50	63	65	51	66	
4	28	29	41	36	47	58	60	66		
5	29	30	39							
6	30									

p	3	5	7	9	11	13	15	17	19	23
Level										
2	27	28	36	32	43	50	53	42	53	59
3	32	34	47	42	56					
4	34									

In all previous examples, the boundary conditions are of Dirichlet type. The next example investigates mixed boundary conditions for three different computational domains. In all cases, the parameters $E = 10^6$ and $\nu = 0.3$ are chosen as in the previous cases. No volume force exists, e.g. $f(x) = 0$. In the first case, the computational domain is a `brick` which is fixed at the bottom face, e.g. there are homogeneous Dirichlet boundary conditions. A surface traction of the form $f_1(x,y,z) = \begin{pmatrix} 0 & 0 & -1 \end{pmatrix}^\top$ acts on the top face whereas no volume forces act on the other faces, e.g. $f_1(x,y,z) = 0$, see Fig. 5. The coarse mesh consists of one element only. In the second case, the computational domain is a `beam` of the form $(0,2)^2 \times (0,10)$, which is fixed at the face $x = 0$. A traction force of the form $f_1(x,y,z) = \begin{pmatrix} 0 & 0 & -0.25 \end{pmatrix}^\top$, acts on the face $x = 10$, see also Fig. 5. In the third case, the computational domain `holesN` is as `holes`. The domain is fixed at the bottom face $z = -3$. A surface traction of the form $f_1(x,y,z) = \begin{pmatrix} 1 & 0 & -1 \end{pmatrix}^\top$, acts on the face $x = 3$, see also Fig. 4.

The PCG iteration numbers are displayed in Table 7. On the one hand, the absolute iteration numbers are higher than for problems with pure Dirichlet boundary conditions, cf. with the examples `cube3.030` and `holes`. On the other hand, the pcg-iteration numbers depend only moderately on the polynomial degree. Another reason for relatively high pcg iteration numbers in the example `brick` is the

Table 7 PCG iteration numbers for mixed boundary conditions: `brick`(left), `beam` (middle) and `holesN` (right).

Level	brick					beam					holesN				
p	1	3	5	7	9	1	3	5	7	9	1	3	5	7	9
2	21	70	83	91	96	68	170	183	189	189	42	71	75	76	75
3	34	83	99	106	110	117	181	193	202	201	49	72	76	78	77
4	45	89	105	113		145	188	197			56	73	76		

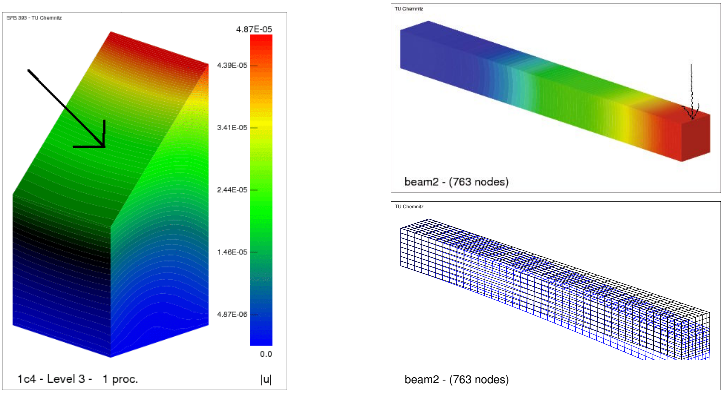


Fig. 5 Setting for the examples `brick` (left) and `beam` (right, top). The total displacement is shown colored. The displacements for `beam` are shown right, below.

Table 8 PCG iteration numbers (above) and number of unknown N (below) for the example the unit `cube3.049` for linear elasticity with $\nu = 0.49$.

p	3	5	7	9	11	13	15	17	19	21	33
Level											
2	25	28	53	46	58	76	81	59	80	87	76
3	78	90	137								
4	85	90	131								
5	89										
6	90										
2	1029	3993	10125	20577	36501	59049	89373	128625	177957	238521	902289
3	6591	27783	73167								
4	46857	206763	555579								
5	352947										
6	2738019										

deformation of the elements. They are not cubes whereas the preconditioner (24) has been developed for the cubes.

Due to (23), the preconditioner (24) is not robust with respect to $\nu \to \frac{1}{2}-$ since the constant c_1 depends on ν. This is observed in the next example `cube3.049`, where a Poisson's ratio $\nu = 0.49$ is chosen, see Table 8. In this example, the computational domain is the unit cube. Pure Dirichlet boundary conditions are used.

5 Conclusions

The paper has presented a quasioptimal solver for systems of linear algebraic equations arising from the discretization of H^1 elliptic problems in three space dimensions using the *hp*-version of the finite element method. The efficiency of the solver has been shown in several numerical examples.

However, this solver seems not be robust against anisotropies in the coefficients, coefficient jumps or a Poisson's ratio ν close to 0.5. The last problem can be resolved by using a mixed formulation for linear elasticity by introducing the hydrostatic pressure p as additional variable. The mathematical details will be presented in a forthcoming paper where solvers for *hp*-fem discretizations of the Stokes problem are investigated.

Acknowledgements. This work has been supported by the FWF projects P20121 and P23484. The basis of the code has been developed during the work at SFB 393 "Numerische Simulation an massive parallelen Rechnern". Further support has been given by the SFB F013 and the Austrian Academy of Sciences.

References

[1] Ainsworth, M.: A preconditioner based on domain decomposition for *h-p* finite element approximation on quasi-uniform meshes. SIAM J. Numer. Anal. 33(4), 1358–1376 (1996)

[2] Antonietti, P.F., Houston, P.: A class of domain decomposition preconditioners for *hp*-discontinuous Galerkin finite element methods. J. Sci. Comput. 46(1), 124–149 (2011)

[3] Babuška, I., Craig, A., Mandel, J., Pitkäranta, J.: Efficent preconditioning for the *p*-version finite element method in two dimensions. SIAM J. Numer. Anal. 28(3), 624–661 (1991)

[4] Beuchler, S.: Multi-grid solver for the inner problem in domain decomposition methods for *p*-FEM. SIAM J. Numer. Anal. 40(3), 928–944 (2002)

[5] Beuchler, S.: A domain decomposition preconditioner for *p*-FEM discretizations of two-dimensional elliptic problems. Computing 74(4), 299–317 (2005)

[6] Beuchler, S.: Wavelet solvers for *hp*-FEM discretizations in 3D using hexahedral elements. Comput. Methods Appl. Mech. Engrg. 198(13-14), 1138–1148 (2009)

[7] Beuchler, S., Meyer, A., Pester, M.: SPC-PM3AdH v1.0-programmers manual. Technical Report SFB393 01-08, Technische Universität Chemnitz (2001)

[8] Beuchler, S., Schneider, R., Schwab, C.: Multiresolution weighted norm equivalences and applications. Numer. Math. 98(1), 67–97 (2004)

[9] Braess, D.: Finite elements. Cambridge University Press, Cambridge (2007)

[10] Bramble, J., Pasciak, J., Xu, J.: Parallel multilevel preconditioners. Math. Comp. 55(191), 1–22 (1991)

[11] Canuto, C., Gervasio, P., Quarteroni, A.: Finite-element preconditioning of G-NI spectral methods. SIAM J. Sci. Comput. 31(6), 4422–4451 (2010)

[12] Costabel, M., Dauge, M., Demkowicz, L.: Polynomial extension operators for H^1, $H(\mathrm{curl})$ and $H(\mathrm{div})$-spaces on a cube. Math. Comp. 77(264), 1967–1999 (2008)

[13] Demkowicz, L.: Computing with *hp* Finite Elements. CRC Press, Taylor and Francis (2006)

[14] Deville, M.O., Mund, E.H.: Finite element preconditioning for pseudospectral solutions of elliptic problems. SIAM J. Sci. Stat. Comp. 18(2), 311–342 (1990)

[15] Guo, B., Gao, W.: Domain decomposition method for the *hp*-version finite element method. Comp. Methods Appl. Mech. Eng. 157, 524–440 (1998)

[16] Guo, B., Zhang, J.: Stable and compatible polynomial extensions in three dimensions and applications to the *p* and *h-p* finite element method. SIAM J. Numer. Anal. 47(2), 1195–1225 (2009)

[17] Haase, G., Langer, U., Meyer, A.: Domain decomposition preconditioners with inexact subdomain solvers. Technical Report 192, TU Chemnitz (1991)

[18] Hackbusch, W.: Multigrid Methods and Applications. Springer, Heidelberg (1985)

[19] Jensen, S., Korneev, V.G.: On domain decomposition preconditioning in the hierarchical $p-$version of the finite element method. Comput. Methods Appl. Mech. Eng. 150(1–4), 215–238 (1997)

[20] Karniadakis, G.M., Sherwin, S.J.: Spectral/HP Element Methods for CFD. Oxford University Press, Oxford (1999)

[21] Korneev, V.G., Langer, U., Xanthis, L.: On fast domain decomposition methods solving procedures for *hp*-discretizations of 3d elliptic problems. Comp. Methods Appl. Math. 3(4), 536–559 (2003)

[22] Korneev, V.G., Rytov, A.: Fast domain decomposition algorithm discretizations of 3-d elliptic equations by spectral elements. Comput. Methods Appl. Mech. Engrg. 197(17-18), 1443–1446 (2008)

[23] Korneev, V.G.: Почти оптимальныи метод решения задач Дирихле на подобластях декомпосиции иерархическои *hp*-версии. Дифферентиальные Уравнения 37(7), 1008–1018 (2001) (An almost optimal method for Dirichlet problems on decomposition subdomains of the hierarchical *hp*-version)

[24] Mandel, J.: Iterative solvers by substructuring for the *p*-version finite element method. Comput. Methods Appl. Mech. Eng. 80(1-3), 117–128 (1990)

[25] Melenk, J.M., Gerdes, K., Schwab, C.: Fully discrete *hp*-finite elements: Fast quadrature. Comput. Methods Appl. Mech. Eng. 190, 4339–4364 (1999)

[26] Munoz-Sola, R.: Polynomial liftings on a tetrahedron and applications to the *h-p* version of the finite element method in three dimensions. SIAM J. Numer. Anal. 34(1), 282–314 (1996)

[27] Pavarino, L.F.: Additive Schwarz methods for the *p*-version finite element method. Numer. Math. 66(4), 493–515 (1994)

[28] Pavarino, L.F., Widlund, O.B.: Iterative substructuring methods for spectral elements in three dimensions. In: Krizek, M., et al. (eds.) Finite Element Methods. 50 Years of the Courant Element, University of Conference held at the Jyväskylä, Finland. Inc. Lect. Notes Pure Appl. Math., vol. 164, pp. 345–355. Marcel Dekker, New York (1994)

[29] Pavarino, L.F., Widlund, O.B.: A polylogarithimc bound for an iterative substructuring method for spectral elements in three dimensions. SIAM J. Numer. Anal. 33(4), 1303–1335 (1996)

[30] Ruge, J.W., Stüben, K.: Algebraic Multigrid. Multigrid methods, ch. 4, pp. 73–130. SIAM, Philadelphia (1987)

[31] Schöberl, J., Melenk, J.M., Pechstein, C., Zaglmayr, S.: Additive Schwarz preconditioning for p-version triangular and tetrahedral finite elements. IMA J. Numer. Anal. 28(1), 1–24 (2008)

[32] Schwab, C.: $p-$ and $hp-$finite element methods. Theory and applications in solid and fluid mechanics. Clarendon Press, Oxford (1998)

[33] Solin, P., Segeth, K., Dolezel, I.: Higher-Order Finite Element Methods. Chapman and Hall, CRC Press (2003)

[34] Toselli, A., Widlund, O.B.: Domain Decomposition Methods - Algorithms and Theory. Springer (2005)

[35] Zhang, X.: Multilevel Schwarz methods. Numer. Math. 63, 521–539 (1992)

A Rigorous Error Analysis of Coupled FEM-BEM Problems with Arbitrary Many Subdomains

Clemens Pechstein and Clemens Hofreither

Abstract. In this article, we provide a rigorous a priori error estimate for the symmetric coupling of the finite and boundary element method for the potential problem in three dimensions. Our theoretical framework allows an arbitrary number of polyhedral subdomains. Our bound is not only explicit in the mesh parameter, but also in the subdomains themselves: the bound is independent of the number of subdomains and involves only the shape regularity constants of a certain coarse triangulation aligned with the subdomain decomposition. The analysis includes the so-called BEM-based FEM as a limit case.

1 Introduction

The coupling of the finite element method (FEM) and the boundary element method (BEM) has a fruitful tradition, see e.g. [5, 7, 15, 17, 30, 38]. The computational domain is split into a finite number of subdomains. On some of the subdomains, a finite element mesh is employed, on the remaining subdomains, a boundary element mesh. Here we assume that the meshes are matching. One of the most successful coupling methods is the symmetric coupling introduced by Costabel [5]. A special case of this method is the BEM domain decomposition (DD) method introduced by Hsiao and Wendland in [13], see also [14] and [16]. An error analysis of the

Clemens Pechstein
Institut für Numerische Mathematik, Johannes Kepler Universität Linz,
Altenberger Str. 69, 4040 Linz, Austria
e-mail: clemens.pechstein@numa.uni-linz.ac.at

Clemens Hofreither
Doctoral Program “Computational Mathematics”, Johannes Kepler Universität Linz,
Altenberger Str. 69, 4040 Linz, Austria
e-mail: clemens.hofreither@dk-compmath.jku.at

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 109–132.
springerlink.com

symmetric FEM-BEM coupling has been provided by Steinbach [30], see also [32] for an analysis of a non-symmetric coupling.

To our best knowledge, in all the available literature on the stability analysis of such FEM-BEM coupling or BEM-DD, it is assumed that the subdomain decomposition is fixed. When considering classes of subdomain decompositions of a fixed computational domain, the a priori error estimates depend not only on the mesh parameters, but on the subdomains themselves.

In the context of pure FEM-DD (see [34], and e.g. the FETI and FETI-DP method introduced in [8, 9]), such error estimates do not depend on the subdomain decomposition at all, because the discretization is never changed when keeping the original domain fixed. On the contrary, for the case of BEM-DD, already by splitting a single subdomain into two subdomains, we change the discretization. To our best knowledge, there is no result available for the symmetric coupling that clarifies the dependence on the subdomains, not even in the simple case where each subdomain is a simplicial coarse element of a coarse mesh. Although in most of the practical applications only a few subdomains are involved, this issue is mathematically unsatisfactory. A desirable error estimate should be explicit in both the fine and coarse mesh parameter.

The first paper towards an explicit analysis is [11], where a so-called BEM-based finite element method is analyzed for the three-dimensional Laplace problem; see also [10]. The BEM-based FEM discretization can be viewed as a special case of the BEM-DD of a domain into polygonal/polyhedral domains whose boundaries are discretized with a few boundary elements, see [3, 4, 10, 37]. If H denotes the typical subdomain diameter, we can express this fact by

$$H \to h.$$

Alternatively, the BEM-based FEM can be viewed as a local Trefftz method [35]. A diagram of all the special cases of FEM-BEM coupling mentioned above is shown in Fig. 1. It is clear that a general analysis in terms of H and h must include the limit case of BEM-based FEM.

The analysis in [11] assumes that each subdomain is the union of a few elements of an auxiliary triangulation with mesh size $H \approx h$. Also, the authors of [11] had to assume that the Poincaré and extension constants of the subdomains and related subregions are uniformly bounded. The theory in [26] yields explicit bounds for the boundary integral operators, at least for three space dimensions. Together with a few more theoretical tools, one obtains "explicit" a priori error estimates.

In the current paper, we provide an analysis for the general symmetric coupling of FEM-BEM with arbitrary subdomains for the potential equation. This includes all the cases sketched in Fig. 1. The assumptions are in their nature less restrictive than in [11]. For the case of three dimensions, we were able to remove all the assumptions on the boundedness of Poincaré and extension constants. We only need that each subdomain is the union of a few elements of a shape regular coarse triangulation and that the exterior angles of each subdomain do not degenerate. Under these

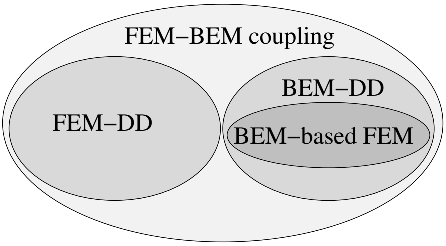


Fig. 1 Diagram of general FEM-BEM coupling and its special cases.

assumptions, we can show explicit bounds for the Poincaré and extension constants. For the bounds of the Poincaré constants we use a result from [28] which builds on [36]. To get the other necessary bounds, we construct an extension operator for polytopes in the spirit of Stein [29] and finally provide an explicit stability estimate.

On the one hand, it is surprising that it took so long to get an analysis with the above (satisfactory) properties, although there are many works available discussing fast solvers for FEM-BEM discretizations with arbitrary many subdomains, see [16, 17, 18, 19, 21, 22, 23, 24, 25, 26]. On the other hand, the analysis below requires some technical tools that were developed only recently.

In the current article, we try to be self-contained up to a certain degree. The remainder is organized as follows. Sect. 2 contains a description of our model problem, the subdomain decomposition, a survey on boundary integral operators, and the symmetric FEM-BEM coupling. In Sect. 3, we present the assumptions and statement of our main result (with the proof postponed). Explicit bounds for boundary integral operators are collected in Sect. 4. This section includes the construction of the explicit extension operator described above (see Sect. 4.2). The proof of the main result is contained in Sect. 5. We conclude with a few remarks on possible extensions.

2 Model Problem and FEM-BEM Coupling

In this section, we describe the model problem and the subdomain decomposition. On each subdomain, we define the harmonic extension operator, the Neumann trace operator, the Steklov-Poincaré operator and a Newton potential. Next, we give a survey on boundary integral operators. In particular, we write the Steklov-Poincaré operator in terms of boundary integral operators. With these ingredients, we formulate the symmetric coupling, which involves a BEM-based approximation of the continuous Steklov-Poincaré operator in the BEM subdomains, and the original bilinear form in the FEM subdomains.

2.1 Model Problem

Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) be a bounded Lipschitz polytope whose boundary $\partial\Omega$ consists of a Dirichlet boundary Γ_D with positive surface measure and a Neumann boundary $\Gamma_N = \partial\Omega \setminus \Gamma_D$. The outward unit normal vector to $\partial\Omega$ is denoted by n. We consider the weak form of the following boundary value problem. For given functions $f \in L^2(\Omega)$, $g_N \in L^2(\Gamma_N)$, and $g_D \in H^{1/2}(\Gamma_D)$,

$$\text{find } u \in H^1(\Omega),\, u_{|\Gamma_D} = g_D : \quad a(u,v) \;=\; \langle \ell, v\rangle \qquad \forall v \in H^1_D(\Omega), \tag{1}$$

where $H^1_D(\Omega) := \{v \in H^1(\Omega) : v_{|\Gamma_D} = 0\}$ and

$$a(u,v) := \int_\Omega \alpha \nabla u \cdot \nabla v\, dx, \qquad \langle \ell, v\rangle := \int_\Omega f\, v\, dx + \int_{\Gamma_N} g_N\, v\, ds.$$

Above, $\langle\cdot,\cdot\rangle$ denotes the duality pairing. We assume that the coefficient $\alpha \in L^\infty(\Omega)$ is uniformly elliptic, i.e.,

$$\alpha(x) \geq \alpha_0 > 0 \qquad \forall x \in \Omega \text{ a.e.}$$

From these assumptions, it follows that the bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is bounded, i.e.,

$$a(v,w) \;\leq\; \|\alpha\|_{L^\infty(\Omega)} \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \qquad \forall v, w \in H^1(\Omega) \tag{2}$$

and $H^1_D(\Omega)$-coercive, in particular

$$a(v,v) \;\geq\; \frac{\alpha_0}{1 + C_F^2} \|v\|^2_{H^1(\Omega)} \qquad \forall v \in H^1_D(\Omega), \tag{3}$$

where C_F is the Friedrichs constant of Ω with respect to the Dirichlet boundary Γ_D. Since $\ell \in H^1_D(\Omega)^*$, the Lax-Milgram theorem delivers the existence of a unique solution.

2.2 Subdomain Decomposition

Let $\{\Omega_i\}_{i=1}^N$ be a non-overlapping decomposition of Ω into open Lipschitz polytopes such that

$$\overline{\Omega} \;=\; \bigcup_{i=1}^N \overline{\Omega}_i, \qquad \Omega_j \cap \Omega_j \;=\; \emptyset \qquad \text{for } i \neq j. \tag{4}$$

The *skeleton* Γ_S is given by

$$\Gamma_S := \bigcup_{i=1}^{N} \partial\Omega_i .$$

Fig. 2 shows a sample domain $\Omega \subset \mathbb{R}^2$ (with two holes) and a subdomain decomposition.

For each subdomain Ω_i, let n_i denote the outward unit normal vector on $\partial\Omega_i$. We assume that the coefficient is piecewise constant with respect to the subdomain decomposition, i.e.,

$$\alpha_{|\Omega_i} = \alpha_i = \text{const} \qquad \forall i = 1,\dots,N.$$

Thanks to the assumptions on f and g_N, we have the splitting property

$$a(u,v) \;=\; \sum_{i=1}^{N} a_i(u_{|\Omega_i}, v_{|\Omega_i}), \qquad \langle \ell, v\rangle \;=\; \sum_{i=1}^{N} \langle \ell_i, v_{|\Omega_i}\rangle, \tag{5}$$

where $a_i : H^1(\Omega_i) \times H^1(\Omega_i) \to \mathbb{R}$ and $\ell_i \in H^1(\Omega_i)^*$ are given by

$$a_i(u,v) \;=\; \alpha_i \int_{\Omega_i} \nabla u \cdot \nabla v \, dx, \qquad \langle \ell_i, v\rangle \;=\; \int_{\Omega_i} f\, v\, dx + \int_{\partial\Omega_i \cap \Gamma_N} g_N\, v\, ds.$$

Note that the theory below can be generalized without any problems to a general functional $\ell \in H^1(\Omega)^*$ that obeys a splitting of the form (5).

2.3 Operators Associated to the Potential Equation

Definition 1 (harmonic extension). For each $i = 1,\dots,N$, let $\mathscr{H}_i : H^{1/2}(\partial\Omega_i) \to H^1(\Omega_i)$ denote the harmonic extension operator such that for $v \in H^{1/2}(\partial\Omega_i)$,

$$(\mathscr{H}_i v)_{|\partial\Omega_i} = v, \qquad a_i(\mathscr{H}_i v, w) = 0 \quad \forall w \in H^1_0(\Omega_i).$$

Due to the Ritz minimum principle, we have that

$$\mathscr{H}_i v \;=\; \operatorname{argmin}\big\{ a_i(\widetilde{v}, \widetilde{v}) : \widetilde{v} \in H^1(\Omega_i),\, \widetilde{v}_{|\partial\Omega_i} = v \big\}. \tag{6}$$

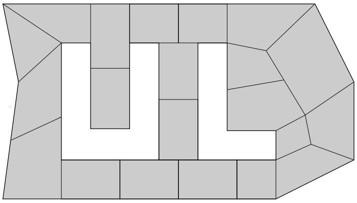


Fig. 2 Example of a subdomain decomposition of a non-convex domain.

Definition 2 (Neumann trace). Let $H_\Delta(\Omega_i) := \{v \in H^1(\Omega_i) : \Delta v \in L^2(\Omega_i)\}$, where Δ is the distributional Laplace operator, and let $\gamma_i^1 : H_\Delta(\Omega_i) \to H^{-1/2}(\partial\Omega_i)$ denote the Neumann trace operator, given by

$$\langle \gamma_i^1 u, v\rangle \;=\; a_i(u, \mathscr{H}_i v) + (\Delta u, \mathscr{H}_i v)_{L^2(\Omega_i)} \qquad \text{for } v \in H^{1/2}(\partial\Omega_i).$$

Note that $\gamma_i^1 u = \alpha_i \frac{\partial u}{\partial n_i}$ for smooth functions u, and that $\Delta \mathscr{H}_i v = 0$ for all functions $v \in H^{1/2}(\partial\Omega_i)$.

Definition 3 (Steklov-Poincaré operator). Let $S_i : H^{1/2}(\partial\Omega_i) \to H^{-1/2}(\partial\Omega_i)$ denote the Steklov-Poincaré operator, given by $S_i := \gamma_i^1 \mathscr{H}_i$.

We have the relation

$$\langle S_i v, w\rangle \;=\; a_i(\mathscr{H}_i v, \mathscr{H}_i w) \qquad \forall v, w \in H^{1/2}(\partial\Omega_i). \tag{7}$$

Definition 4 (Newton potential). For a functional $\psi \in H^1(\Omega_i)^*$, let $u_\psi \in H_0^1(\Omega_i)$ denote the unique solution of

$$a_i(u_\psi, v) \;=\; \langle \psi, v\rangle \qquad \forall v \in H_0^1(\Omega_i).$$

The Newton potential $N_i : H^1(\Omega_i)^* \to H^{-1/2}(\partial\Omega_i)$ is defined by the relation

$$\langle N_i \psi, v_{|\partial\Omega_i}\rangle \;=\; \langle \psi, v\rangle - a_i(u_\psi, v) \qquad \forall v \in H^1(\Omega_i),$$

see also [20].

For any $u \in H^1(\Omega_i)$ and $\psi \in H^1(\Omega_i)^*$ with $a_i(u, v) = \langle \psi, v\rangle$ for all $v \in H_0^1(\Omega_i)$, we have Green's identity

$$a_i(u, v) - \langle \psi, v\rangle \;=\; \langle S_i u_{|\partial\Omega_i} - N_i \psi, v_{|\partial\Omega_i}\rangle \qquad \forall v \in H^1(\Omega_i), \tag{8}$$

such that $S_i u_{|\partial\Omega_i} - N_i \psi$ is the (generalized) conormal derivative of u.

2.4 Boundary Integral Operators

The fundamental solution of the Laplace operator is given by

$$U^*(x, y) \;=\; \begin{cases} -\frac{1}{2\pi} \log|x-y| & \text{if } d = 2, \\ \frac{1}{4\pi} |x-y|^{-1} & \text{if } d = 3. \end{cases}$$

Following, e.g., [31], we define the four boundary integral operators

$$V_i : H^{-1/2}(\partial\Omega_i) \to H^{1/2}(\partial\Omega_i), \qquad K_i : H^{1/2}(\partial\Omega_i) \to H^{1/2}(\partial\Omega_i),$$
$$K_i' : H^{-1/2}(\partial\Omega_i) \to H^{-1/2}(\partial\Omega_i), \qquad D_i : H^{1/2}(\partial\Omega_i) \to H^{-1/2}(\partial\Omega_i),$$

called in turn single layer potential, double layer potential, adjoint double layer potential, and hypersingular operator. For smooth functions, they obey the integral representations

$$(V_i w)(x) \;=\; \int\limits_{\partial\Omega_i} U^*(x,y)\,w(y)\,ds_y, \qquad (K_i v)(x) \;=\; \int\limits_{\partial\Omega_i} \frac{\partial U^*}{\partial n_{i,y}}(x,y)\,v(y)\,ds_y,$$

$$(D_i v)(x) \;=\; -\frac{\partial}{\partial n_{i,x}} \int\limits_{\partial\Omega_i} \frac{\partial U^*}{\partial n_{i,y}}(x,y)\,\big(v(y)-v(x)\big)\,ds_y.$$

Note also that V_i and D_i are self-adjoint and K_i' is the adjoint of K_i. We assume throughout the paper that $\operatorname{diam}(\Omega) \le 1$ if $d = 2$, which ensures that the single layer potential operator is elliptic (see e.g. [12, 31]). From the Caldéron identities (cf. [31, Sect. 6.6]), we get

$$S_i \;=\; V_i^{-1}(\tfrac{1}{2}I + K_i) \;=\; D_i + (\tfrac{1}{2}I + K_i')V_i^{-1}(\tfrac{1}{2}I + K_i). \tag{9}$$

We define the subspaces

$$\begin{aligned} H_*^{-1/2}(\partial\Omega_i) \;&:=\; \{w \in H^{-1/2}(\partial\Omega_i) : \langle w, 1\rangle = 0\},\\ H_*^{1/2}(\partial\Omega_i) \;&:=\; \{v \in H^{1/2}(\partial\Omega_i) : \langle V_i^{-1}v, 1\rangle = 0\}, \end{aligned}$$

cf. [31, Sect. 6.6.1]. Following [33], we have the contraction property

$$(1-c_{K,i})\,\|v\|_{V_i^{-1}} \;\le\; \|(\tfrac{1}{2}I + K_i)v\|_{V_i^{-1}} \;\le\; c_{K,i}\,\|v\|_{V_i^{-1}} \qquad \forall v \in H_*^{1/2}(\partial\Omega_i), \tag{10}$$

with the norm $\|v\|_{V_i^{-1}} := \sqrt{\langle V_i^{-1}v, v\rangle}$ and the contraction constant

$$c_{K,i} = \tfrac{1}{2} + \sqrt{\tfrac{1}{4} - c_{0,i}} \;\in\; \big(\tfrac{1}{2}, 1\big), \quad \text{where} \quad c_{0,i} = \inf_{v \in H_*^{1/2}(\partial\Omega_i)} \frac{\langle D_i v, v\rangle}{\langle V_i^{-1}v, v\rangle} \;\in\; \big(0, \tfrac{1}{4}\big).$$

2.5 Continuous Domain-Skeleton Formulation

Let $I_{\mathrm{BEM}} \subset \{1,\ldots,N\}$ denote the subset of subdomain indices where we want to discretize with the boundary element method, and set $I_{\mathrm{FEM}} = \{1,\ldots,N\} \setminus I_{\mathrm{BEM}}$. We define two subspaces of partially harmonic functions

$$\begin{aligned} V_S \;&:=\; \{v \in H^1(\Omega) : \forall i \in I_{\mathrm{BEM}} : v_{|\Omega_i} = \mathscr{H}_i(v_{|\partial\Omega_i})\},\\ V_{S,D} \;&:=\; \{v \in V_S : v_{|\Gamma_D} = 0\}. \end{aligned}$$

Equipped with the usual H^1-norm, these spaces are Hilbert spaces. We see that the values on $\Gamma_S \cup \big(\bigcup_{i \in I_{\mathrm{FEM}}} \Omega_i\big)$ already determine a function in V_S. Moreover, we have the a-orthogonal splitting

$$H^1(\Omega) \;=\; V_S \oplus \bigcup_{i\in I_{\mathrm{BEM}}} H^1_0(\Omega_i).$$

We consider the variational formulation

$$\text{find } u_S \in V_S,\, u_{S|\Gamma_D} = g_D: \quad a_S(u_S, v) \;=\; \langle \ell_S, v\rangle \qquad \forall v \in V_{S,D}, \tag{11}$$

where

$$\begin{aligned} a_S(u,v) \;&=\; \sum_{i\in I_{\mathrm{BEM}}} \langle S_i u_{|\partial\Omega_i}, v_{|\partial\Omega_i}\rangle + \sum_{i\in I_{\mathrm{FEM}}} a_i(u_{|\Omega_i}, v_{|\Omega_i}),\\ \langle \ell_S, v\rangle \;&=\; \sum_{i\in I_{\mathrm{BEM}}} \langle N_i \ell_i, v_{|\partial\Omega_i}\rangle + \sum_{i\in I_{\mathrm{FEM}}} \langle \ell_i, v_{|\Omega_i}\rangle. \end{aligned}$$

Since V_S and $V_{S,D}$ are subspaces of $H^1(\Omega)$ and $H^1_D(\Omega)$, it follows immediately that the bilinear form $a_S(\cdot,\cdot): V_S \times V_S \to \mathbb{R}$ is bounded and $V_{S,D}$-coercive. The following lemma follows from Green's identity (8).

Lemma 1. *Let u_S be the unique solution of* (11)*, and for $i \in I_{\mathrm{BEM}}$, let $u_i \in H^1_0(\Omega_i)$ be the unique solution of*

$$a_i(u_i, v) \;=\; \langle \ell_i, v\rangle - \langle S_i u_{S|\partial\Omega_i}, v_{|\partial\Omega_i}\rangle \qquad \forall v \in H^1_0(\Omega_i).$$

Then problem (1) *is solved by*

$$u_S + \sum_{i\in I_{\mathrm{BEM}}} u_i.$$

In other words, $u_{S|\Omega_i} + u_i$ solves the Dirichlet problem on Ω_i with Dirichlet data $u_{S|\partial\Omega_i}$.

2.6 Symmetric FEM-BEM Coupling

Let $\mathscr{T}^h(\Gamma_S) = \{\gamma\}$ be a simplicial triangulation of the skeleton Γ_S into line segments if $d = 2$ and into triangular faces if $d = 3$. For each $i \in I_{\mathrm{FEM}}$, let $\mathscr{T}^h(\Omega_i) = \{\tau\}$ be a simplicial triangulation of Ω_i (into triangles if $d = 2$ and tetrahedra if $d = 3$) that matches with $\mathscr{T}^h(\Gamma_S)$ on $\partial\Omega_i$. Our discretization space is given by

$$\begin{aligned} V_S^h := \Big\{ v \in V_S:\; & v_{|\gamma} \in P_1 \quad \forall \gamma \in \mathscr{T}^h(\Gamma_S),\\ & v_{|\tau} \in P_1 \quad \forall \tau \in \mathscr{T}^h(\Omega_i) \quad \forall i \in I_{\mathrm{FEM}} \Big\}, \end{aligned}$$

where P_1 are the polynomials of total degree ≤ 1. Functions in V_S^h are piecewise linear on the skeleton. Restricted to a FEM subdomain Ω_i, they are piecewise linear with respect to $\mathscr{T}^h(\Omega_i)$.

Assumption 1. *The Dirichlet data g_D is piecewise linear with respect to the skeleton triangulation.*

Assumption 1 can always be fulfilled by interpolating or projecting the Dirichlet data. The Galerkin discretization of (11) reads

$$\text{find } u_S^h \in V_S^h,\, u^h_{S|\Gamma_D} = g_D : \ \ a_S(u_S^h, v^h) \ = \ \langle \ell_S, v^h \rangle \quad \forall v^h \in V^h_{S,D}, \tag{12}$$

where

$$V^h_{S,D} := \left\{ v^h \in V_S^h : v^h{}_{|\Gamma_D} = 0 \right\}.$$

With Céa's lemma,

$$\|u_S - u_S^h\|_{H^1(\Omega)} \ \le \ \frac{\|\alpha\|_{L^\infty(\Omega)}}{\alpha_0}\,(1 + C_F^2) \inf_{v^h \in V_S^h} \|u_S - v^h\|_{H^1(\Omega)}.$$

However, computing the stiffness matrix associated to S_i is in general not possible: although we can express S_i via boundary integral operators, we would need the exact inverse V_i^{-1} that appears in the two representations (9).

For $i \in I_{\mathrm{BEM}}$, we use the following approximation of S_i in terms of the boundary integral operators, see [30, Sect. 3.4] and also [5]. Let the space Z_i^h of piecewise constant functions be given by

$$Z_i^h := \{ z \in L^2(\partial\Omega_i) : z_{|\gamma} \in P_0 \quad \forall \gamma \in \mathcal{T}^h(\partial\Omega_i) \} \ \subset \ H^{-1/2}(\partial\Omega_i), \tag{13}$$

where $\mathcal{T}^h(\partial\Omega_i)$ is the restriction of $\mathcal{T}^h(\Gamma_S)$ to $\partial\Omega_i$.

Definition 5 (Approximate Steklov-Poincaré operator). The approximate Steklov-Poincaré operator

$$\widetilde{S}_i : H^{1/2}(\partial\Omega_i) \to H^{-1/2}(\partial\Omega_i)$$

is defined by

$$\widetilde{S}_i v \ := \ D_i v + (\tfrac{1}{2} I + K_i) w_i^h(v),$$

where $w_i^h(v) \in Z_i^h$ is the unique solution of the variational problem

$$\langle z^h, V_i w_i^h(v) \rangle \ = \ \langle z^h, (\tfrac{1}{2} I + K_i) v \rangle \qquad \forall z^h \in Z_i^h.$$

Let $w_i(v) \in H^{-1/2}(\partial\Omega_i)$ be given by

$$w_i(v) \ := \ V_i^{-1}(\tfrac{1}{2} I + K_i) v \ = \ S_i v.$$

By the Galerkin orthogonality and an energy argument,

$$\langle \widetilde{S}_i v, v \rangle \ = \ \langle D_i v, v \rangle + \langle w_i^h(v), V_i w_i^h(v) \rangle \ \le \ \langle D_i v, v \rangle + \langle w_i(v), V_i w_i(v) \rangle \ = \ \langle S_i v, v \rangle.$$

Using Cauchy's inequality and the contraction properties (10), we obtain that for $v \in H_*^{1/2}(\partial\Omega_i)$,

$$\begin{aligned}\langle S_i v, v\rangle &= \langle V_i^{-1}(\tfrac{1}{2}I + K_i)v, v\rangle \le \|(\tfrac{1}{2}I + K_i)v\|_{V_i^{-1}} \|v\|_{V_i^{-1}} \le c_{K,i} \|v\|^2_{V_i^{-1}} \\ &\le c_{K,i} c_{0,i}^{-1} \langle D_i, v, v\rangle \le c_{K,i} c_{0,i}^{-1} \Big(\langle D_i, v, v\rangle + \langle V_i w_i^h(v), w_i^h(v)\rangle\Big).\end{aligned}$$

Since the first and last term are invariant to adding a constant, we can summarize that

$$\frac{c_{0,i}}{c_{K,i}} \langle S_i v, v\rangle \le \langle \widetilde{S}_i v, v\rangle \le \langle S_i v, v\rangle \qquad \forall v \in H^{1/2}(\partial\Omega_i), \tag{14}$$

see also [6], [30], and [25, Lemma 1.33]. Using the approximations $\widetilde{S}_i \approx S_i$ for $i \in I_{\mathrm{BEM}}$, we define the modified bilinear form

$$\widetilde{a}_S(v, w) := \sum_{i \in I_{\mathrm{BEM}}} \langle \widetilde{S}_i v, w\rangle + \sum_{i \in I_{\mathrm{FEM}}} a_i(v, w) \qquad \text{for } v, w \in V_S.$$

For simplicity, we assume that there are no volume sources given in the BEM subdomains.

Assumption 2. *For all $i \in I_{\mathrm{BEM}}$, we have $f_{|\Omega_i} = 0$.*

Under Assumption 2, the evaluation of the Newton potential $N_i \ell_i$ simplifies to integrating g_N against a test function over $\partial\Omega_i \cap \Gamma_N$, and so no approximation of N_i is necessary.

The inexact Galerkin formulation corresponding to (11) reads

$$\text{find } u_S^h \in V_S^h, u^h_{S|\Gamma_D} = g_D : \quad \widetilde{a}_S(u_S^h, v^h) = \langle \ell_S, v^h\rangle \qquad \forall v^h \in V^h_{S_D}. \tag{15}$$

3 Main Result

In this section, we state our main result: an a-priori error estimate for the formulation (15). Not only will this estimate be explicit in the discretization parameters, but it will in a certain sense be independent of the subdomain decomposition. In order to parameterize the subdomain decomposition, we could assume that each subdomain is an element of a coarse mesh. To be more general and to allow at least for subdomains that are polytopes, we use the following assumption which is standard in the theory of iterative substructuring methods, cf. [34, Assumption 4.3].

Assumption 3. *Each subdomain Ω_i is the union of a few simplicial elements of a global shape regular triangulation $\mathscr{T}^H(\Omega)$ such that the number of coarse elements per subdomain is uniformly bounded.*

Let $H_i = \operatorname{diam}(\Omega_i)$ denote the subdomain diameters. The above assumption implies that $H_i \simeq H_j$ if $\partial\Omega_i \cap \partial\Omega_j \neq \emptyset$, and that each subdomain boundary $\partial\Omega_i$ splits into a uniformly bounded number of coarse facets (cf. [11, Assumption 4.4]).

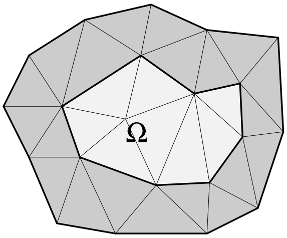


Fig. 3 Sketch of triangulation $\mathscr{T}^H(\widehat{\Omega})$. In dark: $\widehat{\Omega} \setminus \Omega$, thin lines indicate the coarse elements of $\mathscr{T}^H(\widehat{\Omega})$.

The next assumption essentially states that the exterior angles of all subdomains (including those touching the outer boundary $\partial\Omega$) are bounded away from zero, see also Sect. 6.

Assumption 4. *The coarse triangulation $\mathscr{T}^H(\Omega)$ from Assumption 3 can be extended to a shape regular triangulation $\mathscr{T}^H(\widehat{\Omega})$ of a larger domain $\widehat{\Omega} \supset \overline{\Omega}$.*

For an illustration see Fig. 3. Our final assumption concerns the fine triangulations used for the FEM and BEM.

Assumption 5. *The triangulations $\mathscr{T}^h(\Gamma_S)$ and $\mathscr{T}^h(\Omega_i)$, $i \in I_{\text{FEM}}$, are shape regular.*

We define the local mesh parameters

$$h_i := \begin{cases} \max\limits_{\gamma \in \mathscr{T}^h(\partial\Omega_i)} \operatorname{diam}(\gamma) & \text{if } i \in I_{\text{BEM}}, \\ \max\limits_{\tau \in \mathscr{T}^h(\Omega_i)} \operatorname{diam}(\tau) & \text{if } i \in I_{\text{FEM}}, \end{cases}$$

and set $h := \max_{i=1}^N h_i$.

Theorem 1. *Let $d = 3$, let Assumptions 1–5 hold, and suppose that the solution u of* (1) *satisfies $u \in H^2(\Omega)$. Then for the solution u_S^h of* (15)*,*

$$\|u_S - u_S^h\|_{H^1(\Omega)} \leq C \Big(\sum_{i=1}^N h_i^2 \, |u|^2_{H^2(\Omega_i)} \Big)^{1/2} \leq C h |u|_{H^2(\Omega)} .$$

The constant C depends only on the coefficient α, on the Friedrichs constant C_F, and on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$, $\mathscr{T}^h(\Gamma_S)$ and $\mathscr{T}^h(\Omega_i)$, $i \in I_{\text{FEM}}$.

Proof. The proof is postponed to Sect. 5.4. □

4 Explicit Bounds for the Constants $c_{0,i}$

In this subsection, we work out an explicit lower bound for the constants $c_{0,i}$ from Sect. 2.4 in three dimensions which depends only on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$. We heavily use the results from [27], where a series of constants related to the boundary integral operators V_i and D_i are bounded in terms of Poincaré and extension constants. Throughout the rest of the paper, C denotes a generic constant.

4.1 Explicit Bounds for Poincaré Constants

Definition 6. For a bounded Lipschitz domain $D \subset \mathbb{R}^3$, the Poincaré constant is defined as the smallest constant $C_P(D)$ such that

$$\|v - \overline{v}^D\|_{L^2(D)} \;\leq\; C_P(D)\,\mathrm{diam}(D)\,|v|_{H^1(D)} \qquad \forall v \in H^1(D),$$

where $\overline{v}^D = |D|^{-1}\int_D v\,dx$ is the mean value of v.

The following lemma is a direct consequence of [28, Lemma 4.1], see also [36].

Lemma 2. *Let Assumption 3 hold and let m be a fixed integer. Then there exists a constant C that depends only on m and on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$ such that for any connected union D of at most m coarse elements of $\mathscr{T}^H(\widehat{\Omega})$,*

$$C_P(D) \;\leq\; C.$$

4.2 An Extension Operator for Polytopes

In this subsection, we define a Sobolev extension operator for Lipschitz polytopes in the spirit of Stein [29] and provide an explicit estimate in terms of shape regularity constants only.

Let D be the connected union of a few elements from $\mathscr{T}^H(\Omega)$ and let its open surrounding D' be defined by

$$\overline{D'} \;=\; \bigcup\{\overline{T} : T \in \mathscr{T}^H(\widehat{\Omega}),\, T \not\subseteq D,\, \overline{T} \cap \partial D \neq \emptyset\}, \tag{16}$$

see Fig. 4 (right). Let $\mathscr{V}_{\partial D} = \{p\}$ be the set of coarse vertices of $\mathscr{T}^H(\widehat{\Omega})$ that lie on ∂D. For each such coarse vertex, we define the vertex patch ω_p by

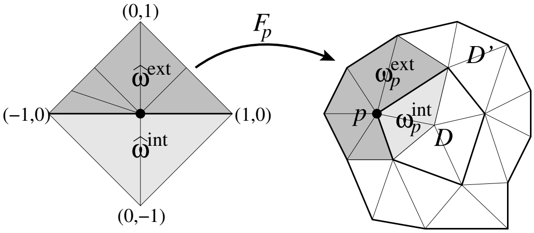
$$\overline{\omega}_p \;=\; \bigcup\{\overline{T} : T \in \mathscr{T}^H(\widehat{\Omega}),\, p \in \overline{T}\},$$

and

$$\omega_p^{\mathrm{int}} \;:=\; \omega_p \cap D, \qquad \omega_p^{\mathrm{ext}} \;:=\; \omega_p \cap D',$$

cf. Fig. 4 (right). Without loss of generality, we assume that ω_p^{int} and ω_p^{ext} each contain at least one coarse node that does not lie on ∂D. This condition can always be fulfilled by formally subdividing some of the coarse elements.

Fig. 4 Mapping of a node patch ω_p in two dimensions.



We define the reference patch

$$\widehat{\omega} := \begin{cases} \text{conv}^\circ(\{(-1,0),(1,0),(0,1),(0,-1)\}) & \text{if } d=2, \\ \text{conv}^\circ(\{(-1,0,0),(1,1,0),(1,-1,0),(0,0,1),(0,0,-1)\}) & \text{if } d=3, \end{cases}$$

where $\text{conv}^\circ(S)$ denotes the interior of the convex hull of the set S. Furthermore, we define the subsets

$$\widehat{\omega}^{\text{int}} := \widehat{\omega} \cap \{x : x_d < 0\}, \qquad \widehat{\omega}^{\text{ext}} := \widehat{\omega} \cap \{x : x_d > 0\},$$

where x_d refers to the d-th component of x.

Let $\mathscr{T}_p(\widehat{\omega})$ be a shape regular simplicial triangulation of $\widehat{\omega}$ such that there exists a bijective continuous mapping $F_p : \widehat{\omega} \to \omega_p$ with the following properties.

- For each element $T \in \mathscr{T}_p(\widehat{\omega})$, the restricted mapping $F_{p|T}$ is affine linear,
- $F_p(0) = p$,
- $F_p(\widehat{\omega} \cap \{x : x_d = 0\}) = \omega_p \cap \partial D$,
- $F_p(\widehat{\omega}^{\text{int}}) = \omega_p^{\text{int}}$ and $F_p(\widehat{\omega}^{\text{ext}}) = \omega_p^{\text{ext}}$,
- for each element $T \in \mathscr{T}_p(\widehat{\omega})$,

$$c_1 H_D^d \le \det(F'_{p|T}) \le c_2 H_D^d,$$
$$\|F'_{p|T}\|_{\ell^2} \le c_3 H_D, \qquad \|(F'_{p|T})^{-1}\|_{\ell^2} \le c_4 H_D^{-1},$$

where $H_D := \text{diam}(D)$ and the constants c_1, c_2, c_3, and c_4 only depend on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$.

For an illustration in two dimensions, see Fig. 4. Under the conditions on $\mathscr{T}^H(\widehat{\Omega})$ stated in Assumption 4, such a triangulation and mapping exists for every coarse vertex $p \in \mathscr{V}_{\partial D}$.

On the reference patch we define

$$\widehat{E} : H^1(\widehat{\omega}^{\text{int}}) \to H^1(\widehat{\omega}^{\text{ext}}), \qquad (\widehat{E}w)(x_1,\ldots,x_d) := w(x_1,\ldots,x_{d-1},-x_d),$$

i.e., the reflection of v across the hyperplane $\{x : x_d = 0\}$, where the above definition first applies to C^∞ functions and is then completed by density (which indeed leads to a bounded operator). For each coarse node $p \in \mathscr{V}_{\partial D}$ we define

$$E_p : H^1(\omega_p^{\text{int}}) \to H^1(\omega_p^{\text{ext}}), \qquad E_p v := \big(\widehat{E}(v \circ F_p)\big) \circ F_p^{-1}.$$

Since F_p is continuous and piecewise affine linear, $E_p v$ is indeed in H^1. Furthermore, we have by construction that

$$(E_p v)_{|\omega_p \cap \partial D} = v_{|\omega_p \cap \partial D}.$$

Finally, we define the extension operator

$$\mathscr{E}_D : H^1(D) \to H^1(\mathbb{R}^d), \qquad (\mathscr{E}_D v)_{|D} := v, \qquad (\mathscr{E}_D v)_{|D'} := \sum_{p \in \mathscr{V}_{\partial D}} \varphi_p \cdot E_p v,$$

where φ_p is the nodal finite element basis function on $\mathscr{T}^H(\widehat{\Omega})$ associated with the coarse node p.

Lemma 3. *Let Assumptions 3 and 4 hold, let D be the connected union of a few elements from $\mathscr{T}^H(\Omega)$, and let the extension operator $\mathscr{E}_D$ be defined as above. Then $\mathscr{E}_D$ indeed maps into $H^1(\mathbb{R}^d)$. Let $\mathscr{D} = \{D\}$ be a collection of subregions of Ω such that every $D \in \mathscr{D}$ is the connected union of at most m elements of $\mathscr{T}^H(\Omega)$. Then there exists a constant $C_{\mathscr{E}}$ depending only on m and on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$ such that for all $D \in \mathscr{D}$,*

$$|\mathscr{E}_D v|^2_{H^1(\mathbb{R}^d)} + H_D^{-2} \|\mathscr{E}_D v\|^2_{L^2(\mathbb{R}^d)} \leq C_{\mathscr{E}} \Big(|v|^2_{H^1(D)} + H_D^{-2} \|v\|^2_{L^2(D)} \Big) \quad \forall v \in H^1(D).$$

Proof. Let $v \in H^1(D)$ be arbitrary but fixed. For each $p \in \mathscr{V}_{\partial D}$, the function $\varphi_p \cdot E_p v$ vanishes on $\mathbb{R}^d \setminus (\overline{D} \cup \omega_p^{\text{ext}})$. Hence,

$$(\mathscr{E}_D v)_{|\mathbb{R}^d \setminus \overline{D}} \in H^1(\mathbb{R}^d \setminus \overline{D}).$$

Since

$$\sum_{p \in \mathscr{V}_{\partial D}} \varphi_p(x) = 1 \qquad \forall x \in \partial D,$$

we have $(\mathscr{E}_D v)_{|\partial D} = v_{|\partial D}$ and hence $\mathscr{E}_D v \in H^1(\mathbb{R}^d)$. With standard finite element techniques (see e.g. [1, 2]), one shows that

$$|E_p v|_{H^1(\omega_p^{\text{ext}})} \leq C |v|_{H^1(\omega_p^{\text{int}})}, \qquad \|E_p v\|_{L^2(\omega_p^{\text{ext}})} \leq C \|v\|_{L^2(\omega_p^{\text{int}})}.$$

The constant C depends only on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$ because there is only a small number of different triangulations $\mathscr{T}_p(\widehat{\omega})$.

Since $\|\varphi_p\|_{L^\infty} = 1$, it follows from the above that

$$\|\varphi_p \cdot E_p v\|^2_{L^2(\omega_p^{\text{ext}})} \leq C \|v\|^2_{L^2(\omega_p^{\text{int}})}.$$

Since $\|\nabla\varphi_p\|_{L^\infty} \leq C H_D^{-1}$, we can conclude from the product rule that

$$\begin{aligned} |\varphi_p \cdot E_p v|^2_{H^1(\omega_p^{\text{ext}})} &\leq C\Big(|E_p v|^2_{H^1(\omega_p^{\text{ext}})} + H_D^{-2}\,\|E_p v\|^2_{L^2(\omega_p^{\text{ext}})}\Big) \\ &\leq C\Big(|v|^2_{H^1(\omega_p^{\text{int}})} + H_D^{-2}\,\|v\|^2_{L^2(\omega_p^{\text{int}})}\Big). \end{aligned}$$

Since the number of coarse elements and coarse nodes in $\overline{D}$ is bounded in terms of m, the desired estimate follows by summing the above estimate over $p \in \mathscr{V}_{\partial D}$. □

Let the operator

$$\mathscr{E}_{D'} : H^1(D') \to H^1(\overline{D} \cup D')$$

be defined analogously to $\mathscr{E}_D$, but exchanging the roles of D and D'.

Lemma 4. *Let $\mathscr{D} = \{D\}$ as in Lemma 3 and let D' denote the surroundings of D as defined in* (16)*. Then there exists a uniform constant $C_{\mathscr{E}'}$ depending only on m and on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$ such that*

$$|\mathscr{E}_{D'} v|^2_{H^1(D)} \leq C_{\mathscr{E}'}\,|v|^2_{H^1(D')} \qquad \forall v \in H^1(D').$$

Proof. The proof follows by combining the proof of Lemma 3 with the Poincaré inequality in D, see Lemma 2. □

4.3 Explicit Bounds for Boundary Integral Operators

Definition 7. For each subdomain Ω_i, we define the seminorm and norm

$$\begin{aligned} |v|_{\star,H^{1/2}(\partial\Omega_i)} &:= |\mathscr{H}_i v|_{H^1(\Omega_i)}, \\ \|v\|_{\star,H^{1/2}(\partial\Omega_i)} &:= \Big(|\mathscr{H}_i v|^2_{H^1(\Omega_i)} + \frac{1}{\operatorname{diam}(\Omega_i)^2}\,\|\mathscr{H}_i v\|^2_{L^2(\Omega_i)}\Big)^{1/2} \end{aligned}$$

(see [27]), which is equivalent to the Sobolev-Slobodeckii norm $\|\cdot\|_{H^{1/2}(\partial\Omega)}$, and the associated dual norm

$$\|w\|_{\star,H^{-1/2}(\partial\Omega_i)} := \sup_{v \in H^{1/2}(\partial\Omega_i)} \frac{\langle w, v\rangle}{\|v\|_{\star,H^{1/2}(\partial\Omega_i)}}.$$

Above and in the following we silently exclude $v = 0$ from the supremum.

In the sequel, we state ellipticity and boundedness results for the boundary integral operators V_i and D_i. In several of the lemmas below, we have to assume that $d = 3$. The two-dimensional case is harder and not further considered in the article at hand. See also [27, Remark 4] and Sect. 6.

Lemma 5. *Let $d = 3$ and let Assumptions 3–4 hold. Then, for each $i = 1, \ldots, N$,*

$$\langle w, V_i w\rangle \geq \tfrac{1}{2} C_{\mathscr{E}}^{-2}\,\|w\|^2_{\star,H^{-1/2}(\partial\Omega_i)} \qquad \forall w \in H^{-1/2}(\partial\Omega_i),$$

i.e., the operators V_i are uniformly elliptic with respect to the norms $\|\cdot\|_{\star,H^{-1/2}(\partial\Omega_i)}$ and the ellipticity constant depends only on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$.

Proof. The statement follows from [27, Lemma 6.1, Corollary 6.2]. The proof there uses the Jones extension, but remains valid for the extension operator $\mathscr{E}_{\Omega_i}$ constructed in Sect. 4.2. □

Lemma 6. *Let Assumptions 3–4 hold and let Ω_i' be the surrounding of Ω_i as defined in* (16)*. Then*

$$\langle D_i v, v\rangle \;\geq\; \tfrac{1}{2} C_{\mathscr{E}'}^{-2}\, |v|^2_{\star,H^{1/2}(\partial\Omega_i)} \qquad \forall v \in H^{1/2}(\partial\Omega_i).$$

Proof. See [27, Lemma 3.8, Lemma 6.4]. □

Lemma 7. *Let $d = 3$ and let Assumptions 3–4 hold. Then*

$$H_i^{-2}\, \|\mathscr{H}_i v\|^2_{L^2(\Omega_i)} \;\leq\; C_P^*\, |\mathscr{H}_i v|^2_{H^1(\Omega_i)} \qquad \forall v \in H_*^{1/2}(\partial\Omega_i),$$

where the constant C_P^* depends only on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$.

Proof. See [27, Lemma 6.7]. □

Lemma 8. *Let $d = 3$ and let Assumptions 3–4 hold. Then*

$$\|V w\|_{H^{-1/2}(\partial\Omega_i)} \;\leq\; C_V^*\, \|w\|_{\star,H^{-1/2}(\partial\Omega_i)} \qquad \forall v \in H^{-1/2}(\partial\Omega_i),$$

where the constant C_V^* depends only on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$.

Proof. See [27, Lemma 6.8]. □

Lemma 9. *For $d = 3$, and each subdomain Ω_i, we have*

$$c_{0,i} \;\geq\; \tfrac{1}{4}\, (C_{\mathscr{E}})^{-2}\, (C_{\mathscr{E}'})^{-2}\, (1 + C_P^*)^{-1},$$

i.e., there is a uniform lower bound for the constants $c_{0,i}$ just in terms of the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$.

Proof. See [27, Corollary 6.10]. □

5 Error Analysis

This section contains the proof of our main theorem. First, we formulate a lemma à la Strang which bounds the total error in terms of the approximation error of the Dirichlet data on the skeleton and the H^1 approximation error in the FEM subdomains, and the approximation error of the Neumann data in the norm induced by the local single layer potentials. Both terms can be estimated explicitly in the fine and coarse mesh parameter.

Since the original domain Ω is fixed, we assume without loss of generality that $\operatorname{diam}(\Omega) = 1$.

5.1 A Lemma à la Strang

Lemma 10. *Let $u_S \in V_S$ and $u_S^h \in V_S^h$ be the solutions of* (11) *and* (15). *For $i \in I_{\mathrm{BEM}}$, let $w_i(u_S) \in H^{-1/2}(\partial\Omega_i)$ be given by*

$$w_i(u_S) := V_i^{-1}(\tfrac{1}{2}I + K_i)u_{S|\partial\Omega_i} = S_i u_{S|\partial\Omega_i}.$$

Then, we have the error estimate

$$\|u_S - u_S^h\|_{H^1(\Omega)} \le \delta \left[\inf_{v^h \in V_S^h} \|u_S - v^h\|_{H^1(\Omega)} + \Big(\sum_{i \in I_{\mathrm{BEM}}} \inf_{z^h \in Z_i^h} \|w_i(u_S) - z^h\|_{V_i}^2 \Big)^{1/2} \right],$$

where

$$\delta = \max(1+\beta, \beta\,\|\alpha\|_{L^\infty(\Omega)}) \max\Big(1, \max_{i \in I_{\mathrm{BEM}}} \frac{c_{K,i}}{\sqrt{1-c_{K,i}}}\Big),$$

and

$$\beta = \frac{1+C_F^2}{\alpha_0} \max\Big(1, \max_{i \in I_{\mathrm{BEM}}} \frac{c_{K,i}}{c_{0,i}}\Big).$$

Proof. First, we homogenize (11) and (15). Let $g \in V_S^h$ be an arbitrary but fixed extension of the Dirichlet datum g_D (i.e., $g_{|\Gamma_D} = g_D$). Then $u_S = g + u_{S,0}$ and $u_{S,h} = g + u_{S,0}^h$ where

$$\begin{aligned} u_{S,0} \in V_{S,D} : \quad & a_S(u_{S,0}, v) = \langle \ell_S, v\rangle - a_S(g, v) && \forall v \in V_{S,D}, \\ u_{S,0}^h \in V_{S,D}^h : \quad & \widetilde{a}_S(u_{S,0}^h, v^h) = \langle \ell_S, v^h\rangle - \widetilde{a}_S(g, v^h) && \forall v^h \in V_{S,D}^h. \end{aligned}$$

From (14), (7), (2), and (3) it follows that

$$\begin{aligned} \widetilde{a}_S(v, v) &\ge \frac{\alpha_0}{1+C_F^2} \min\Big(1, \min_{i \in I_{\mathrm{BEM}}} \frac{c_{0,i}}{c_{K,i}}\Big) \|v\|_{H^1(\Omega)}^2 && \forall v \in V_{S,D}, \\ \widetilde{a}_S(v, w) &\le \|\alpha\|_{L^\infty(\Omega)} \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} && \forall v, w \in V_S. \end{aligned}$$

The Strang lemma from [11, Lemma 4.1] implies that

$$\|u_{S,0} - u_{S,0}^h\|_{H^1(\Omega)} \le \max(1+\beta, \beta\,\|\alpha\|_{L^\infty(\Omega)}) \times \left[\inf_{v^h \in V_{S,D}^h} \|u_{S,0} - v^h\|_{H^1(\Omega)} + \sup_{v^h \in V_{S,D}^h} \frac{\widetilde{a}_S(u_S, v_h) - \langle \ell_S, v_h\rangle}{\|v_h\|_{H^1(\Omega)}} \right].$$

Using that $\langle \ell_S, v_h \rangle \;=\; a_S(u_S, v_h)$, we obtain

$$\widetilde{a}_S(u_S, v_h) - \langle \ell_S, v_h \rangle \;=\; \sum_{i \in I_{\mathrm{BEM}}} \langle \widetilde{S}_i u_S - S_i u_S, v_h \rangle.$$

Following the proof of [11, Lemma 4.2], we get

$$\langle \widetilde{S}_i u_S - S_i u_S, v_h \rangle \;\leq\; \frac{c_{K,i}}{\sqrt{1 - c_{K,i}}} |v_h|_{H^1(\Omega_i)} \, \|w_i(u_S) - w_i^h(u_S)\|_{V_i}.$$

The rest of the proof follows from Cauchy's inequality and the fact that $u_S - u_{S,0} = g \in V_S^h$. □

5.2 Error Estimate for the Dirichlet Data

Theorem 2. *Let Assumptions 1–3 and Assumption 5 hold. Assume further that the solution u of* (1) *satisfies $u \in H^2(\Omega)$. Then there exists a constant C only depending on the shape regularity constants of $\mathscr{T}^h(\Gamma_S)$, $\mathscr{T}^H(\Omega)$, and $\mathscr{T}^h(\Omega_i)$, $i \in I_{\mathrm{FEM}}$, such that*

$$\inf_{v^h \in V_S^h} \|u_S - v^h\|_{H^1(\Omega)} \;\leq\; C \Big(\sum_{i=1}^{N} h_i^2 \, |u|^2_{H^2(\Omega_i)} \Big)^{1/2} \;\leq\; C h |u|_{H^2(\Omega)}.$$

Proof. The proof is analogous to [11, Theorem 4.8]. First, recall that due to Assumption 2, $f_{|\Omega_i} = 0$ for $i \in I_{\mathrm{BEM}}$, and so

$$u_S \;=\; u.$$

From Assumption 3 and Assumption 5 it follows that for each $i \in I_{\mathrm{BEM}}$, the triangulation $\mathscr{T}^h(\partial\Omega_i)$ can be extended to an auxiliary triangulation $\widetilde{\mathscr{T}}^h(\Omega_i)$ with mesh parameter h_i, such that the shape regularity constants of $\widetilde{\mathscr{T}}^h(\Omega_i)$ are bounded in terms of the shape regularity constants of $\mathscr{T}^h(\Gamma_S)$ and $\mathscr{T}^H(\Omega)$. This implies a global triangulation $\widetilde{\mathscr{T}}^h(\Omega)$ of the entire domain Ω. Let

$$\widetilde{V}^h(\Omega) \;:=\; \{ v \in H^1(\Omega) : v_{|T} \in P_1 \quad \forall T \in \widetilde{\mathscr{T}}^h(\Omega) \},$$

and let $I^h u_S \in \widetilde{V}^h(\Omega)$ denote the nodal interpolant of $u_S \in H^2(\Omega)$. Due to the minimizing property (6) of the harmonic extension and a standard interpolation result (see [2]), we obtain

$$\begin{aligned} \inf_{v^h \in V_S^h} \|u_S - v^h\|_{H^1(\Omega)} \;&\leq\; \inf_{v^h \in \widetilde{V}^h(\Omega)} \|u_S - v^h\|_{H^1(\Omega)} \\ &\leq\; \|u_S - I^h u_S\|_{H^1(\Omega)} \;\leq\; C \Big(\sum_{i=1}^{N} h_i^2 \, |u_S|^2_{H^2(\Omega_i)} \Big)^{1/2}, \end{aligned}$$

where C depends only on the mentioned shape regularity constants. □

5.3 Error Estimate for the Neumann Data

Throughout this subsection, assume that $d = 3$ and that Assumptions 3–5 hold. Let $\mathscr{F}_i = \{F\}$ denote the set of triangular coarse faces on $\partial\Omega_i$ (cf. Assumption 3). We define the face seminorms

$$|v|_{H_{\sim}^{1/2}(F)} := \left(\int\limits_F \int\limits_F \frac{|v(x) - v(y)|^2}{|x-y|^3} \, ds_x \, ds_y \right)^{1/2} \qquad \text{for } v \in H^{1/2}(F),\ F \in \mathscr{F}_i,$$

and the piecewise seminorm

$$|v|_{H_{\sim \mathrm{pw}}^{1/2}(\partial\Omega_i)} := \Big(\sum_{F \in \mathscr{F}_i} |v|^2_{H_{\sim}^{1/2}(F)} \Big)^{1/2}.$$

The space $H_{\sim \mathrm{pw}}^{1/2}(\partial\Omega_i)$ is the subspace of $L^2(\partial\Omega_i)$ where the above seminorm is bounded.

Definition 8. For each $i \in I_{\mathrm{BEM}}$, the L^2-projector $Q_i^h : L^2(\partial\Omega_i) \to Z_i^h$ is given by

$$(Q_i^h v, z^h)_{L^2(\partial\Omega)} = (v, z^h)_{L^2(\partial\Omega_i)} \qquad \forall z^h \in Z_i^h,$$

with the space Z_i^h from (13).

Of course, the above equation can be localized and

$$(Q_i^h v)_{|\gamma} = \frac{1}{|\gamma|} \int\limits_\gamma v \, ds \qquad \text{for } \gamma \in \mathscr{T}^h(\partial\Omega_i).$$

Lemma 11. *The operator Q_i^h satisfies, for all $w \in H_{\sim \mathrm{pw}}^{1/2}(\partial\Omega_i)$, the approximation properties*

$$\|w - Q_i^h w\|_{L^2(\partial\Omega_i)} \le C h_i^{1/2} |w|_{H_{\sim \mathrm{pw}}^{1/2}(\partial\Omega_i)},$$

$$\|w - Q_i^h w\|_{\star, H^{-1/2}(\partial\Omega_i)} \le C h_i |w|_{H_{\sim \mathrm{pw}}^{1/2}(\partial\Omega_i)},$$

where the constant C depends only on the shape regularity constants of $\mathscr{T}^H(\Omega)$ and $\mathscr{T}^h(\Gamma_S)$.

Proof. First, we split the local boundary $\partial\Omega_i$ into the (plane) triangular faces $F \in \mathscr{F}_i$. Each such face can be mapped to a reference face. Applying [31, Theorem 10.2] to each face and summing over the faces, we obtain the first estimate (the proof of that theorem is constructed by interpolating estimates in the L^2- and H^1-seminorm at $1/2$).

The second estimate is shown along the lines of [31, Corollary 10.3]: Using the definition of the dual norm, the projection property of Q_i^h, Cauchy's inequality, and the first estimate of the current lemma, we obtain

$$\begin{aligned}
\|w - Q_i^h w\|_{\star,H^{-1/2}(\partial\Omega_i)} &= \sup_{v\in H^{1/2}(\partial\Omega_i)} \frac{(w - Q_i^h w,\, v)_{L^2(\partial\Omega_i)}}{\|v\|_{\star,H^{1/2}(\partial\Omega_i)}} \\
&= \sup_{v\in H^{1/2}(\partial\Omega_i)} \frac{(w - Q_i^h w,\, v - Q_i^h v)_{L^2(\partial\Omega_i)}}{\|v\|_{\star,H^{1/2}(\partial\Omega_i)}} \\
&\leq \|w - Q_i^h w\|_{L^2(\partial\Omega_i)} \sup_{v\in H^{1/2}(\partial\Omega_i)} \frac{\|v - Q_i^h v\|_{L^2(\partial\Omega_i)}}{\|v\|_{\star,H^{1/2}(\partial\Omega_i)}} \\
&\leq C h_i^{1/2} |w|_{H^{1/2}_{\sim\text{pw}}(\partial\Omega_i)} C h_i^{1/2} \sup_{v\in H^{1/2}(\partial\Omega_i)} \frac{|v|_{H^{1/2}_{\sim\text{pw}}(\partial\Omega_i)}}{\|v\|_{\star,H^{1/2}}(\partial\Omega_i)}.
\end{aligned}$$

Using (A9) and (A12) from [11], we can conclude that

$$|v|_{H^{1/2}_{\sim\text{pw}}(\partial\Omega_i)} \leq C\|v\|_{\star,H^{1/2}}(\partial\Omega_i) \qquad \forall v \in H^{1/2}(\partial\Omega_i).$$

The (generic) constants in both estimates depend only on the shape regularity constants of $\mathscr{T}^H(\Omega)$. □

Our last prerequisite is a Neumann trace inequality. For a proof see [11, Theorem 4.10 and Sect. A.2].

Lemma 12. *There exists a constant C depending only on the shape regularity constants of $\mathscr{T}^H(\Omega)$ such that*

$$|\gamma_i^1 v|_{H^{1/2}_{\sim\text{pw}}(\partial\Omega_i)} \leq C|v|_{H^2(\Omega_i)} \qquad \forall v \in H^2(\Omega_i).$$

Combining the tools and estimates above we get the following error estimate.

Theorem 3. *Let $d = 3$ and let Assumptions 3–5 hold. Then there exists a constant C only depending on the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$ and $\mathscr{T}^h(\Gamma_S)$ such that*

$$\inf_{z^h\in Z_i^h} \|\gamma_i^1 v - z^h\|_{V_i} \leq C h_i |v|_{H^2(\Omega_i)} \qquad \forall v \in H^2(\Omega_i).$$

Proof. Using Lemma 8 and Lemma 11, we obtain

$$\begin{aligned}
\inf_{z^h\in Z_i^h} \|\gamma_i^1 v - z^h\|_{V_i} &\leq C_V^* \inf_{z^h\in Z_i^h} \|\gamma_i^1 v - z^h\|_{\star,H^{-1/2}(\partial\Omega_i)} \\
&\leq C_V^* \|\gamma_i^1 v - Q_i^h \gamma_i^1 v\|_{\star,H^{-1/2}(\partial\Omega_i)} \leq C_V^* C h_i |\gamma_i^1 v|_{H^{1/2}_{\sim\text{pw}}(\partial\Omega_i)}.
\end{aligned}$$

An application of Lemma 12 concludes the proof. □

5.4 Proof of Theorem 1

Noticing that $w_i(u_S) = \gamma_i^1 u$ and combining Lemma 10 and Theorem 2, we obtain

$$\|u_S - u_S^h\|_{H^1(\Omega)} \leq C\left[\Big(\sum_{i=1}^{N} h_i^2\, |u|^2_{H^2(\Omega_i)}\Big)^{1/2} + \Big(\sum_{i\in I_{\mathrm{BEM}}} \inf_{z^h\in Z_i^h} \|\gamma_i^1 u - z^h\|^2_{V_i}\Big)^{1/2}\right]$$

Because of Lemma 9 and because $c_{K,i}$ depends monotonically decreasingly on $c_{0,i}$, the constant C above is bounded only in terms of the shape regularity constants of $\mathscr{T}^H(\widehat{\Omega})$, $\mathscr{T}^h(\Gamma_S)$, and $\mathscr{T}^h(\Omega_i)$, $i \in I_{\mathrm{FEM}}$. Applying Theorem 3 on each BEM subdomain concludes the proof of Theorem 1.

6 Conclusion and Extensions

First, we would like to note that we can relax Assumption 4 to the weaker assumption that there exists a shape regular coarse triangulation for the neighborhood of each subdomain (with uniform shape regularity constants). This way, small exterior angles of the computational domain Ω are allowed as long as there are no small exterior angles of the subdomains themselves.

We believe that with careful effort, the above theory can be extended to the two-dimensional case, see [27, Remark 4]. Also, it should be possible to drop Assumption 2 and incorporate an approximation of the Newton potential, see [30].

Using the explicit bounds for the boundary integral operators, it is possible to lift the results in [16, 17] on BETI and coupled FETI/BETI methods to the current setting. Hence, the convergence of these solvers does not depend on the subdomains, but only on the shape regularity of the subdomain decomposition.

For the case of reduced regularity ($u_S \notin H^2(\Omega)$), we first show a stability result. By choosing $v^h = 0$ and $z^h = 0$ in the infima in the statement of Lemma 10, and using [27, Lemma 5.4], one can show that

$$\|u_S - u_S^h\|_{H^1(\Omega)} \leq C\,|u_S|_{H^1(\Omega)},$$

under the minimal assumption that $u_S \in H^1(\Omega)$. Interpolating the H^2 and H^1 error estimate, we immediately get that

$$\|u_S - u_S^h\|_{H^1(\Omega)} \leq Ch^s\, \|u_S\|_{H^{1+s}(\Omega)}$$

if $u_S \in H^{1+s}(\Omega)$.

Acknowledgements. The authors gratefully acknowledge the financial support of the Austrian Science Funds (FWF) under grant DK W1214.

References

[1] Brenner, S.C., Scott, L.R.: The mathematical theory of finite element methods. Texts in Applied Mathematics, vol. 15. Springer, New York (2002)

[2] Ciarlet, P.G.: The finite element method for elliptic problems. Studies in Mathematics and its Applications, vol. 4. North-Holland, Amsterdam (1987)

[3] Copeland, D., Langer, U., Pusch, D.: From the boundary element method to local Trefftz finite element methods on polyhedral meshes. In: Bercovier, M., Gander, M.J., Kornhuber, R., Widlund, O. (eds.) Domain Decomposition Methods in Science and Engineering XVIII. Lecture Notes in Computational Science and Engineering, vol. 70, pp. 315–322. Springer, Heidelberg (2009)

[4] Copeland, D.M.: Boundary-element-based finite element methods for Helmholtz and Maxwell equations on general polyhedral meshes. Int. J. Appl. Math. Comput. Sci. 5(1), 60–73 (2009)

[5] Costabel, M.: Symmetric methods for the coupling of finite elements and boundary elements. In: Brebbia, C., Wendland, W.L., Kuhn, G. (eds.) Boundary Elements IX, pp. 411–420. Springer, Heidelberg (1987)

[6] Costabel, M.: Some historical remarks on the positivity of boundary integral operators. In: Schanz, M., Steinbach, O. (eds.) Boundary Element Analysis - Mathematical Aspects and Applications. LNACM, vol. 29, pp. 1–27. Springer, Berlin (2007)

[7] Costabel, M., Stephan, E.P.: Coupling of finite and boundary element methods for an elastoplastic interface problem. SIAM J. Numer. Anal. 27, 1212–1226 (1990)

[8] Farhat, C., Roux, F.X.: A method of finite element tearing and interconnecting and its parallel solution algorithm. Int. J. Numer. Meth. Engrg. 32, 1205–1227 (1991)

[9] Farhat, C., Lesoinne, M., Le Tallec, P., Pierson, K., Rixen, D.: FETI-DP: A dual-primal unified FETI method I: A faster alternative to the two-level FETI method. Int. J. Numer. Meth. Engrg. 50, 1523–1544 (2001)

[10] Hofreither, C.: L_2 error estimates for a nonstandard finite element method on polyhedral meshes. J. Numer. Math. 19(1), 27–39 (2011)

[11] Hofreither, C., Langer, U., Pechstein, C.: Analysis of a non-standard finite element method based on boundary integral operators. Electron. Trans. Numer. Anal. 37, 413–436 (2010)

[12] Hsiao, G.C., Wendland, W.L.: A finite element method for some integral equations of the first kind. J. Math. Anal. Appl. 58, 449–481 (1977)

[13] Hsiao, G.C., Wendland, W.L.: Domain decomposition in boundary element methods. In: Proceedings of the Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, pp. 41–49. SIAM, Philadelphia (1991)

[14] Hsiao, G.C., Steinbach, O., Wendland, W.L.: Domain decomposition methods via boundary integral equations. J. Comput. Appl. Math. 125, 521–537 (2000)

[15] Langer, U.: Parallel iterative solution of symmetric coupled FE/BE- equations via domain decomposition. Contemporary Mathematics 157, 335–344 (1994)

[16] Langer, U., Steinbach, O.: Boundary element tearing and interconnecting method. Computing 71(3), 205–228 (2003)

[17] Langer, U., Steinbach, O.: Coupled boundary and finite element tearing and interconnecting methods. In: Kornhuber, R., Hoppe, R., Periaux, J., Pironneau, O., Widlund, O.B., Xu, J. (eds.) Domain Decomposition in Science and Engineering XV. Lecture Notes in Computational Sciences and Engineering, vol. 40, pp. 83–97. Springer, Heidelberg (2004)

[18] Langer, U., Steinbach, O.: Coupled finite and boundary element domain decomposition methods. In: Schanz, M., Steinbach, O. (eds.) Boundary Element Analysis - Mathematical Aspects and Applications. LNACM, vol. 29, pp. 61–95. Springer, Berlin (2007)

[19] Langer, U., Of, G., Steinbach, O., Zulehner, W.: Inexact data-sparse boundary element tearing and interconnecting methods. SIAM J. Sci. Comp. 29, 290–314 (2007)

[20] McLean, W.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press, Cambridge (2000)

[21] Of, G.: BETI-Gebietszerlegungsmethoden mit schnellen Randelementverfahren und Anwendungen. PhD thesis, Universität Stuttgart, Germany (2006)

[22] Of, G.: The all-floating BETI method: Numerical results. In: Langer, U., Discacciati, M., Keyes, D.E., Widlund, O.B., Zulehner, W. (eds.) Operating Systems. Lecture Notes in Computational Science and Engineering, vol. 60, pp. 295–302. Springer, Heidelberg (2008)

[23] Of, G., Steinbach, O.: The all-floating boundary element tearing and interconnecting method. J. Numer. Math. 17(4), 277–298 (2009)

[24] Pechstein, C.: BETI-DP methods in unbounded domains. In: Kunisch, K., Of, G., Steinbach, O. (eds.) Numerical Mathematics and Advanced Applications – Proceedings of the 7th European Conference on Numerical Mathematics and Advanced Applications, Graz, Austria, September 2007, pp. 381–388. Springer, Heidelberg (2008)

[25] Pechstein, C.: Finite and boundary tearing and interconnecting methods for multiscale elliptic partial differential equations. PhD thesis, Johannes Kepler University, Linz (2008)

[26] Pechstein, C.: Boundary element tearing and interconnecting methods in unbounded domains. Appl. Numer. Math. 59(11), 2824–2842 (2009)

[27] Pechstein, C.: Shape-explicit constants for some boundary integral operators. Appl. Anal. (to appear, 2012); available online December 2011, doi:10.1080/00036811.2011.643781

[28] Pechstein, C., Scheichl, R.: Weighted Poincaré inequalities. NuMa-Report 2010-10, Institute of Computational Mathematics, Johannes Kepler University Linz, Austria (2010) (submitted)

[29] Stein, E.M.: Singular Integrals and Differentiability Properties of Functions. Princeton Math Series, vol. 30. Princeton University Press, Princeton (1970)

[30] Steinbach, O.: Stability estimates for hybrid coupled domain decomposition methods. Lecture Notes in Mathematics, vol. 1809. Springer, Heidelberg (2003)

[31] Steinbach, O.: Numerical approximation methods for elliptic boundary value problems – Finite and boundary elements. Springer, New York (2008)

[32] Steinbach, O.: A note on the stable one-equation coupling of finite and boundary elements. SIAM J. Numer. Anal. 49, 1521–1531 (2011)

[33] Steinbach, O., Wendland, W.L.: On C. Neumann's method for second-order elliptic systems in domains with non-smooth boundaries. J. Math. Anal. Appl. 262(2), 733–748 (2001)

[34] Toselli, A., Widlund, O.B.: Domain Decoposition Methods – Algorithms and Theory. Series in Computational Mathematics, vol. 34. Springer, Heidelberg (2005)

[35] Trefftz, E.: Ein Gegenstück zum Ritz'schen Verfahren. In: Proc. Ind. Int. Cong. Appl. Mech., Zurich, pp. 131–137 (1926)

[36] Veeser, A., Verfürth, R.: Poincaré constants of finite element stars. IMA. J. Numer. Anal. (2011); published online May 30, doi:10.1093/imanum/drr011

[37] Weißer, S.: Residual error estimate for BEM-based FEM on polygonal meshes. Numer. Math. 118(4), 765–788 (2011)

[38] Zienkiewicz, O.C., Kelly, D.M., Bettess, P.: The coupling of the finite element method and boundary solution procedures. Int. J. Numer. Meth. Eng. 11, 355–375 (1977)

A Review of Anisotropic Refinement Methods for Triangular Meshes in FEM

René Schneider

Abstract. This review gives an overview of current anisotropic refinement methods in finite elements, with the focus on the actual refinement step. In this we highlight strengths and weaknesses of different approaches and hope to stimulate research into closer coupling of the refinement process with efficient solution strategies for the equation systems arising from the discretized equations. A rough overview of different categories of error estimation techniques relevant in the anisotropic setting is also given.

1 Introduction

Anisotropic adaptive meshes in the finite element method have been active area of research since the end of the nineteen eighties, albeit earlier considerations go back to the seventies and even fifties, see Apel [2, Section I.1] for a short summary of this early period. Many engineering and scientific applications are known where anisotropically refined meshes significantly outperform isotropic locally refined meshes, yet the majority of finite element algorithms and software exclude anisotropic meshes.

To spur further research in this interesting field and to ease the entry for young researchers into this field we provide a brief overview of the area in this review paper.

In this we restrict our self to adaptivity based on *a posteriori* analysis of the properties of the solution, which is generally carried out in a loop of the famous structure

$$\texttt{solve} \quad \to \quad \texttt{estimate} \quad \to \quad \texttt{refine}.$$

René Schneider
Fakultät für Mathematik, TU Chemnitz, Reichenhainer Strasse 41,
09107 Chemnitz, Germany
e-mail: rene.schneider@mathematik.tu-chemnitz.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 133–152.
springerlink.com

Thus anisotropy in the solution has to be detected by the estimate step based on the computed solution for the current mesh. The refine step then has to build a suitably anisotropically refined mesh and the whole process is repeated until a satisfactory accuracy of the solution is gained.

In contrast to this there is a large area of research where anisotropic refinement is guided by *a priori* analysis, which means that the information on suitable mesh refinement is based on information that is available without computing even an approximation of the solution. Examples for this are the behavior of the solution of elliptic PDEs in boundary layers (e.g. [31]) or near edges (see e.g. [2]). We do not consider this class in this present work.

We provide only a rough overview of error estimation techniques in the context of anisotropic refinement, which is usually considered most important. The main focus of this review is a systematic overview of strategies to actually execute the anisotropic refinement, i.e. the final part of the adaptive loop, which highlights strengths and weaknesses of the different classes of approaches found in the literature.

2 Background

Standard isotropic error estimates for conforming finite element discretizations of elliptic PDEs are usually based on isotropic approximation error estimates, e.g. [1, Corollary 1.2]

$$\inf_{v\in\mathbb{P}^p}\sum_{m=0}^{p+1} d^m |u-v|_{W^{m,r}(\Omega)} \le C d^{p+1} |u|_{W^{p+1,r}(\Omega)}, \tag{1}$$

where $\mathbb{P}^p$ denotes the space of polynomials of degree at most p, d the diameter of the domain $\Omega \subset \mathbb{R}^n$, i.e. $d := \sup\limits_{x,y\in\Omega} ||x-y||$, C a constant which may depend on the shape of Ω but not on the diameter d and $W^{m,r}(\Omega)$ are the Sobolev spaces with corresponding norm and semi-norm.

As these estimates don't use directional information, they can not take benefit if the solution is anisotropic, i.e. if the behavior is significantly different in different spatial directions, and the size of the domain in these directions matches the solution behavior suitably. The remedy of this situation gives rise to anisotropic approximation error estimates, e.g. [2, (2.5)], [11]

$$\inf_{v\in\mathbb{P}^p} |u-v|_{W^{m,q}(\Omega)} \le C \sum_{\substack{\alpha_1+\ldots+\alpha_n=p+1-m\\ \alpha_1,\ldots,\alpha_n\ge 0}} h_1^{\alpha_1}\cdot\ldots\cdot h_n^{\alpha_n} \left| \frac{\partial^{p+1-m}u}{\partial x_1^{\alpha_1}\ldots\partial x_n^{\alpha_n}} \right|_{W^{m,p}(\Omega)}, \tag{2}$$

where $h_1,\ldots,h_n$ are edge lengths of a axis-aligned minimal bounding box of the domain Ω. In this type of estimate large directional derivatives of the function u may be compensated by a small domain size in this direction. For the whole finite element mesh this small size of the domain in one direction translates into small size of the elements in the direction where the function varies strongly, and possibly larger size in the directions where the solution varies slowly, once the approximation error estimates are applied to the whole mesh.

To illustrate this we supply a simple example. We consider interpolation of the function $u(x,y) := -1/2\,(x+1)(x-1)$ by linear elements on a mesh of axis aligned right triangles of the square $\Omega = (-1,1)^2$. As this function is constant in the y-direction, mesh refinement in this direction has no effect on the error $||u - I_h u||_{H^1}$, where I_h is the nodal interpolation operator for each mesh. Only the mesh resolution in x-direction is responsible for the reduction of this error, see Fig. 1.

Of course, the solution being constant in one direction is in some sense a degenerated case. However, the phenomenon to some extend carries over to the more general anisotropic case, where the functions directional second derivatives are substantially different. There refinement in the dominating direction has a much stronger impact, until directional mesh resolution and solution behavior balance. Once this balance is reached all directions should be refined in the same manner.

In Fig. 2 we give a simple example of this, for the function $u(x,y) := -1/2((1-\alpha)x^2 + \alpha y^2)$. This function has constant curvature in all directions. Curvature in x-directions is $(1-\alpha)$, in y-direction α. The top of the figure shows the function for $\alpha = 1/10$ to give an impression. This function is interpolated on tensor product meshes with n_x nodes in x-direction and n_y nodes in y-direction (as in Fig. 1). Starting with a mesh of $n_x = n_y = 4$ (so 16 nodes in total) the number of nodes in each direction is increased aiming to keep a fixed ratio $n_x/n_y = const$. The special cases of refinement only in x-direction ($n_x/n_y = 0$) and only in y-direction ($n_x/n_y = \infty$) are considered, as well as several values in between, for $\alpha = 1/100$ and $\alpha = 1/1000$.

In both cases the error in x-direction dominates on the coarse meshes. So refinement in y-direction (solid blue lines) achieves almost no error reduction. Up to a certain point refinement in x-direction alone (dotted red line) gives the best error reduction. However, as this refinement in x-direction proceeds, a point is reached where the error in x-direction is reduced so far that the error in the y-direction starts to dominate, hence the reduction tails of.

All cases of fixed ratio $n_x/n_y \in (0,\infty)$ result in parallel lines in these plots, reflecting the same order of error as function of the total number of nodes. This is indeed the optimal order as predicted by the (isotropic) a priori analysis. However, the constants in this behavior are substantially different, with the best values achieved when the ratio of the mesh spacing in x and y direction matches the ratio of the curvatures in these directions, as may be expected from the anisotropic estimate (2). Note that in these examples the error is reduced by one order of magnitude ($\alpha = 1/100$) respectively one and a half ($\alpha = 1/1000$) by applying the correct mesh stretching (dashed red line) compared to uniform meshes with $n_x/n_y = 1$.

During the stage where the correct mesh stretching is not reached yet, i.e. the refinement takes place only in the direction which dominates the error contributions, the improved order of the directional refinement can be observed, see again the dashed red lines.

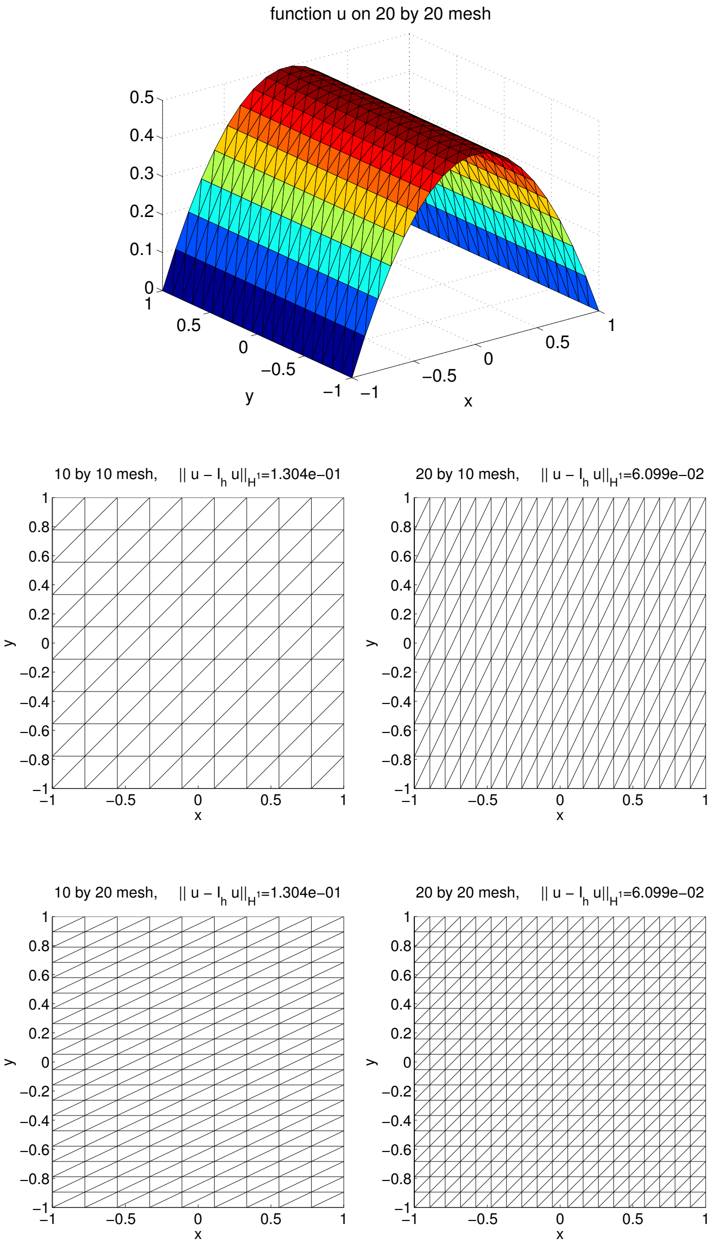


Fig. 1 Different meshes and corresponding $\mathbb{H}^1(\Omega)$ interpolation error for the example $u(x,y) := -1/2\,(x+1)(x-1)$.

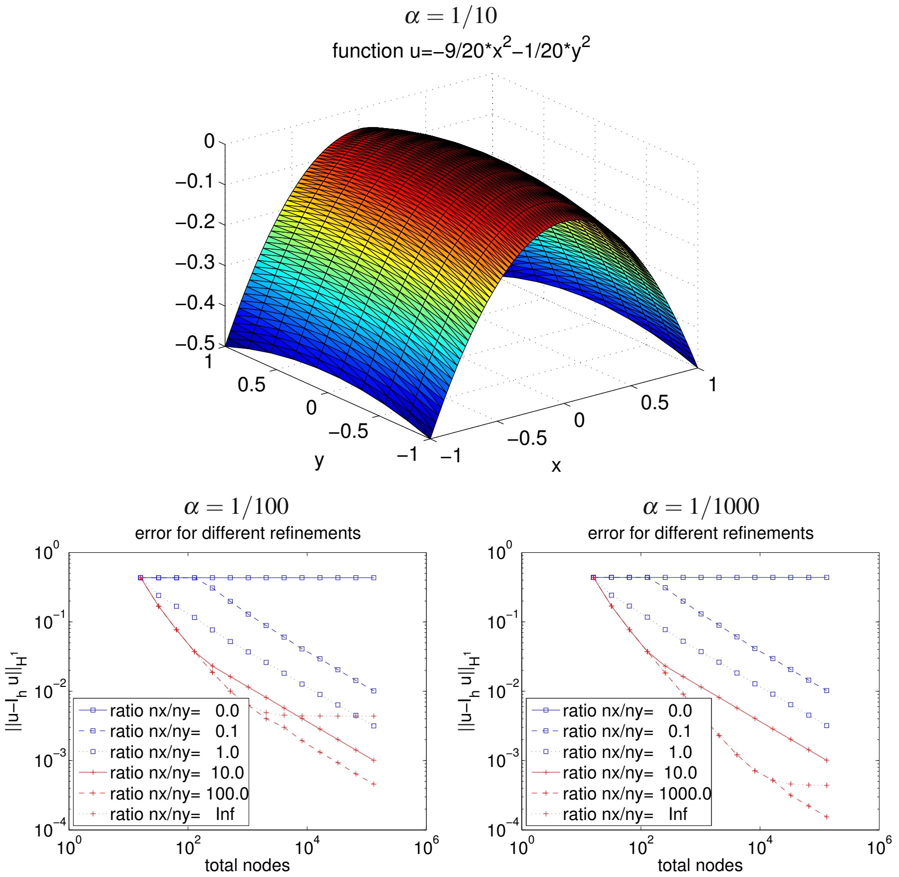


Fig. 2 Example $u(x,y) := -1/2((1-\alpha)x^2 + \alpha y^2)$ and comparison of $\mathbb{H}^1(\Omega)$ interpolation error vs. total number of nodes (=$nx \cdot n_y$) for different ratios of n_x/n_y.

This kind of anisotropic behavior of the solution plays an important role in many applications, e.g. reentrant edges in three dimensional structural mechanics and boundary layers in fluid dynamics. In both these examples singularities of the solution are restricted to small portions of the domain, but have a large impact on the overall approximation of the solution in the whole domain. A substantial part of the computing resources has to be used to resolve the effects in the regions where the singularities affect the solution. As the singularities are of anisotropic character, the overall computational effort can be significantly reduced if the singularities are treated by suitable anisotropic meshes in these regions, see for example [34, Example 3] which we reproduce here to illustrate this point.

In this example the solution to a reaction-diffusion equation is considered which has a boundary layer near the Dirichlet boundary. Fig. 3 shows the convergence histories for different refinement approaches. In this example the error is reduced by

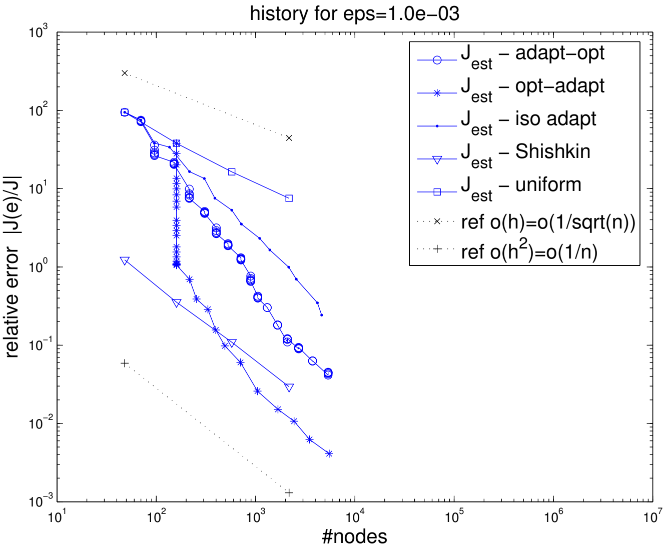


Fig. 3 Convergence histories for Example 3 of [34].

two orders of magnitude by optimizing the node positions in a coarse mesh, which is used as basis for subsequent further local mesh refinement (labeled opt-adapt). This advantage is due to the strong anisotropic refinement towards the boundary layer which this optimization introduces, see Figure 4 for the initial and optimized coarse meshes. In contrast the isotropic local refinement (labeled iso-adapt in Figure 3) leads to strong over-refinement along the layer because the anisotropy of the solution is not exploited.

For comparison uniform refinement (labeled uniform) and Shiskin meshes (labeled Shishkin) which are based on a priori analysis are included in the figure, as well as a further (less successful) algorithm proposed in [34] combining local refinement with node optimization (labeled adapt-opt). See [34] for details on the example.

Before we discuss refinement criteria and refinement strategies we should introduce two concepts which are important in the discussion of anisotropic finite elements. The first is the aspect ratio of an element, which may be defined as the ratio of the largest length scale of an element to the shortest. Thus the aspect ratio of an element is a measure of anisotropy of the element. The precise definitions of largest and shortest length scale vary in the literature, for example they may be the diameter of the element and the diameter of the largest in-circle of a planar element, or just the length of the longest edge of a triangle and the hight perpendicular to this edge.

Further, many sources on anisotropic finite elements consider a maximum angle condition, which goes back to Babuska and Aziz [4]. This condition is an analogue to the minimum angle condition in isotropic elements, required to proof certain

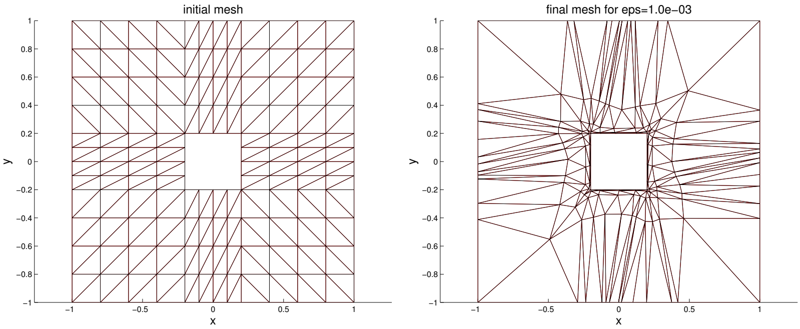


Fig. 4 Initial and optimized coarse mesh for Example 3 of [34].

bounds on the interpolation error. However, more recent work indicates that this condition can be replaced, see e.g. [35] for an illustrative example and discussion, or [9].

3 Refinement Criteria

As the error estimation or indication provides the information where and how the mesh should be refined, this is probably the most crucial step in adaptive methods in general. Thus, even though this is not the focus of this paper, in this section we want to give at least a rough overview of this area. As such this overview will neither be complete nor detailed.

In comparison to error estimation in isotropic adaptive mesh refinement, two major additional issues arise in anisotropic adaptive refinement. The first is that the error estimates or indicators have to perform well on anisotropic meshes in order to be a reliable guide for the refinement. While this is a trivial consequence of the anisotropic meshes, it is a major issue concerning the analysis, as important parts of the standard techniques in error estimation are not robust in this respect. The second issue is that the estimates or indicators have to provide additional information on the directional behavior of the error in order to guide the directional refinement.

There are many papers on anisotropic error estimation. However, not all of them consider both aspects. This should be observed in the search for a suitable error estimate.

Most of the a posteriori estimates in the literature can be grouped into six non-disjoint categories,

- interpolation error estimates,
- post-processing and recovery techniques,
- residual estimates,

- local problem estimates,
- hierarchical estimates and
- dual weighted residual (DWR) estimates.

In the following we will briefly characterize these categories, highlight important properties and issues, and provide pointers to literature.

3.1 Interpolation Error Estimates

Anisotropic interpolation error estimates are the basis for most anisotropic error estimation techniques, as they allow to infer a measure of the distance between the solution and the best approximation on the finite element mesh based on properties of the solution. Which properties are used and how these properties are derived gives rise to the distinct methods in the following subsections.

A very important contribution in this context is the work of Apel [2] which provides a framework for deriving such estimates as well as the resulting estimates for the most common element types of variable polynomial degrees.

For linear elements especially in two dimensions, more details have been revealed regarding the optimal shape and size, the necessity (respectively no-necessity) of the maximum-angle condition in several works, e.g. [9], [11] and [35].

3.2 Post-processing and Recovery Techniques

As the interpolation error estimates reveal, for linear finite elements the quadratic part of the solution gives the dominant terms in the error estimates. Thus a large number of works use the Hessian of the solution to guide the mesh refinement. However, a useful approximation of the Hessian is of course not readily available from the piecewise linear approximation of the solution (as the Hessian is piecewise zero). Thus approximations are defined by post-processing and recovery techniques, see [36, 29] for two recent overview and comparison papers. The frequently discussed variants are:

- computing the derivatives of the gradient by a weak formulation, e.g. for $w_h \approx w := \partial^2 u / \partial x^2$

$$\int\limits_{\Omega} w_h v_h \mathrm{d}\Omega = -\int\limits_{\Omega} \frac{\partial u_h}{\partial x}\frac{\partial v_h}{\partial x}\mathrm{d}\Omega + \int\limits_{\Omega} \frac{\partial u_h}{\partial x} v_h n_x \mathrm{d}\Gamma \qquad \forall v_h \in \mathbb{V}_h,$$

  see e.g. [10],

- averaging the gradient of u_h at the nodal points and using a linear interpolation of those gradient component values of which again gradients (thus approximations of the second derivatives of u) can be computed (Zienkiewicz-Zhu post-processing),
- local least squares approximation of the nodal values by (at least second order) polynomials, using the Hessian of these polynomials as approximation.

The recent overview and numerical study of these approaches in [29] comes to the conclusion:

> Numerical results in 2D and 3D show that the precision of all methods considered depends strongly on the mesh topology and that no convergence can be certified in general. However, there is no blow up and the values obtained are probably accurate enough in order to be used as refinement or coarsening criteria in adaptive algorithms.

3.3 Residual Estimates

The classical residual error estimate for the Poisson equation on linear finite elements bounds the H^1 semi-norm of the error by [1, (2.17)]

$$||\nabla u - \nabla u_h||_{L_2(\Omega)} \leq C \left(\sum_{T\in\mathscr{T}_h} h_T^2 ||r||^2_{L_2(T)} + \sum_{E\in\partial\mathscr{T}_h} h_T ||R||^2_{L_2(E)} \right)^{1/2},$$

with the element residual $r := -\Delta u_h - f$, the edge residual $R := -\left[\frac{\partial u_h}{\partial n}\right]$ (jump of the normal derivative across edge E) and the element diameter h_T. Even though this estimate contains the in general unknown constant C, this class of estimates is quite popular, because it is computationally cheap and easy to implement, especially if the element residual is neglected because it is of higher order than the edge residuals.

Applying this kind of estimate on anisotropic meshes is more difficult, i.e. if one wants to make use of the anisotropy of the mesh, it is not appropriate to just use the diameter h_T of the elements, but one has to use the short length scales of the elements as well. Kunert developed a whole class of error estimates of this kind and showed their efficiency and robustness on anisotropic meshes [22, 24, 23]. However, to do so Kunert introduced so called matching functions, which measure the alignment of the anisotropy in the mesh with the anisotropy of the solution. This in turn requires knowledge of the solution itself, or as Kunert proposed to use Zienkiewicz-Zhu post-processing of the numerical solution to get an approximation for this purpose.

A clear disadvantage of this class of estimators is that they only provide information on the location of the error, not on the directionality of it.

3.4 Local Problem Estimates

As the name *local problem estimates* indicates, these estimates require the solution of the PDE on small parts of the domain with the residual of the PDE as right hand side. Typically these small parts are patches of neighboring elements, and there are variants which use local Dirichlet problems as well as local Neumann problems. In the Neumann variant the local problems are singular, but can be treated if the condition of *equilibrated residuals* (e.g. [1, Chapter 6]) is fulfilled, which gives them a second name.

As the element geometry is direct input into the definition of the local problems, anisotropy is not an issue in the definition. However, Grosman [15, Section 6.1] investigated *equilibrated residual* estimates for a singularly perturbed reaction diffusion problem on anisotropic meshes and found that the matching function of Kunert is also required for these estimates in order to prove reliability on anisotropic meshes.

A nice feature of these local problems is that they can also be used to derive appropriate stretching of the elements, see [3].

3.5 Hierarchical Estimates

The simplest idea for error estimation in numerical methods for differential equations is to use the solution of a better method (higher order or finer mesh) to compute the difference to the current numerical solution, and to use this difference as approximation of the error. This is also done in the PDE context, with more or less effort spent in trying to avoid the computationally expensive solution with the better method.

As in the case of the local problem estimates the approximated error function can be used to guide stretching of the elements, see e.g. [13, 20].

3.6 Dual Weighted Residual (DWR) Estimates

All estimates that we have mentioned so far attempt to reflect properties of $e = u - u_h$, e.g. by considering local and global norms of e. Dual weighted residual estimates take a different view. In many applications not the PDE solution u is of interest, but a quantity derived from u. Thus the primary interest in adaptive methods should not be to make e as small as possible/required, but to make the error in the approximation of this derived quantity as small as possible/required. It turns out that in many cases it is not necessary that e is "small" in the whole domain, but only where the errors affect the approximation of the quantity of interest. The solution of a dual problem provides a measure of this influence, see [6] for a detailed introduction. Adapting the mesh not for *e* but for this quantity of interest allows significant savings in many applications.

Recently a modification of this approach was published by Richter [30] which allows the exploitation of anisotropy in both the solution of the PDE and the solution of the adjoint PDE. Richter uses an interesting concept of semi discretization in the spatial directions, combining the errors of the semi discretizations to estimate the error of the full discretization. Thus the error contributions in different directions are separated. This approach may be useful in context with other error estimation strategies as well.

4 Refinement Strategies

There are three basic approaches to achieve the actual anisotropic mesh refinement:

- node relocation (r-refinement, mesh optimization, moving meshes, mesh smoothing),
- anisotropic re-meshing,
- splitting of elements (bisection, blue refinement).

In the following we will discuss these three approaches in detail, highlighting advantages and shortcomings. We will compare them by asking the following questions (general criteria):

- Does it allow unbounded aspect ratio?
- Does it allow unbounded local resolution?
- Is re-alignment of the mesh with solution features possible?
- Is it possible to utilize actual one-dimensional behavior of the solution?
- What are the computational cost of the refinement?
- Is it suitable for implementation with fast solvers (multigrid or multilevel techniques)?

While most of these criteria are obvious, the second (unbounded local resolution) may require a short explanation. The local solution features that are attempted to be resolved by the adaptive mesh refinement, often require locally extremely fine meshes. At the same time, as we shall see, for some refinement methods lower bounds on the local mesh size are implied by the initial mesh, which is clearly a disadvantage. Thus we introduce this as a criterion.

At the end of this section we will then conclude with a short overview regarding these questions.

4.1 Node Relocation

There are several ways of introducing a node relocation which may result in anisotropic mesh refinement in some parts of the domain. Early work in this direction was motivated by time dependent problems and the desire to let locally refined areas of a mesh move along with the solution features (e.g. shock waves) they are meant to capture. Miller and Miller [28] achieved this by modifying the standard space of linear finite elements to have time dependent node positions, and minimizing the $L_2(\Omega)$-norm of the residual of the PDE under consideration over the space of finite element ansatz functions with the time dependent node positions. This in essence defines a PDE for the node movement which is coupled with the original PDE. Already in [28] modifications of this basic approach where discussed to improve robustness and avoid tangling of the mesh.

Because the number of mesh points and the connectivity are usually fixed in these approaches and adaptivity is achieved by relocating the nodes, this strategy if often referred to as r-refinement, in analogy to the terms h-refinement (local element size) and p-refinement (local degree of the ansatz functions).

The minimization of the residual is one way of defining the node movement, even in steady problems. Another approach is the equidistribution of the error among the elements, which also leads to PDEs describing the optimal node positions. This can be generalized to equidistribution of a monitor function, which should reflect the local error, but can be defined fairly general, thus leading to a large class of approaches, see [5] or [19] for an overview.

Minimization of error estimates can also be considered from different perspectives, leading to finite dimensional optimization problems which can be solved by various methods, e.g. [37, 7, 34, 33].

If the nodes are only relocated, it is of course not possible to reduce the element size in all the domain at once. Thus combination with local h-refinement is frequently discussed, e.g. [25, 37], leading to rh-refinement. Further, keeping the connectivity of the mesh fixed becomes quickly a burden as the meshes are strongly deformed, thus several authors consider modification of the mesh connectivity as well, e.g. [37, 27].

Regarding our criteria it has to be said that this class of refinement strategies in general has the disadvantage that the optimization of the node positions is computationally expensive, even though basically all approaches attempt to keep this expense as low as possible, e.g. by hierarchical techniques [37, 14], by special techniques for optimization [34] or relying on the efficiency of the PDE solvers. Further, at least as long as the connectivity of the mesh is kept fixed, possible one dimensional behavior of the solution can not be exploited. For the same reason the local resolution is in practise bounded, as long as the solution behavior is not trivial in all of the domain but a manifold of measure zero.

On the positive side, combination with fast solvers is easily achieved, as even structured meshes can be used as starting meshes of the optimization, e.g. [14]. For most methods there is in theory no limit to the achievable aspect ratio, albeit the limitation on the local resolution also affect this aspect in practise. The meshes can be re-aligned with solution features, up to the mild restrictions which the topology of the initial mesh may pose.

4.2 Anisotropic Re-meshing

A popular method for constructing anisotropic meshes is to equip the domain Ω with a non-Euclidian metric, defined by a metric tensor, then generating a mesh which is isotropic with respect to this metric. If this metric tensor is chosen suitably, anisotropic meshes can be generated this way [8]. The availability of the software BAMG for two dimensional anisotropic mesh generation by F. Hecht [17] allowed other authors to combine this approach with different choices of the metric defining the anisotropy, which led to a large number of publications in this direction, e.g. [20, 12, 29, 9] to name just a few.

However, utilization of this idea in three dimensions is far less common, as the development of software is far more involved, and availability of software is limited. Further, the computational cost for (global) re-meshing is undoubtedly larger than for local modifications of just a few mesh cells. This disadvantage may not be critical as the cost for the mesh refinement may still be negligible in the context of the overall adaptive solution algorithm. Also this approach allows relatively strong mesh changes in a single step, thus potentially arriving at a (near) optimal adaptive mesh in fewer steps. However, to our knowledge there is no guarantee for this, nor are infinite loops of refinement/de-refinement proven to be ruled out. Thus the relatively high cost for one step of the refinement is used for our criterion.

Another significant disadvantage of this approach is the absence of hierarchical information required by efficient solution strategies for the algebraic systems, as the meshes in a refinement sequence are in no direct relation to each other. Algebraic multigrid methods may be used to reduce this disadvantage, as they don’t require a sequence of meshes. This however is possible for each refinement approach.

The four other criteria of our list are all in favor of this class of refinement strategies: there are no bounds on the aspect ratio or on the local resolution, the mesh can be re-aligned with solution features, and due to the unbounded aspect ratios, one-dimensional behavior can be exploited.

4.3 Splitting of Elements

In isotropic adaptive mesh refinement, the splitting of elements into smaller elements is the canonical way of achieving mesh refinement. There, methods mainly differ in the way they allow or avoid hanging nodes and degeneration of element quality.

In transferring these approaches to anisotropic mesh refinement a number of pitfalls appear, which we want to illustrate. We leave the issue of hanging nodes aside, as this is essentially the same as in isotropic refinement. However, the issue of mesh quality is fundamentally different. First of all, there is obviously no requirement to keep the aspect ratio of the elements bounded. Thus all devices to guarantee this (longest edge bisection, Bänsch green refinement, red-green refinement) are not desirable in this context. Indeed, how to efficiently increase the aspect ratio and fully exploit the flexibility gained by dropping the requirement of bounded aspect ratios is not trivial.

For meshes of quadrangles or hexahedra, it is an intuitive and common approach to split the elements only along one dimension of the reference element, thus generating two elements in the refined mesh, see e.g. [30, 3, 32]. If this process is repeated with alternating the direction of the splitting, the usual uniform refinement patterns can be generated, which makes this a generalization of the usual isotropic refinement of quadrangular meshes. However, this approach has the fundamental disadvantage that the initial mesh of the refinement sequence has to be aligned with the anisotropy of the solution features in order to exploit them, re-alignment is not possible with this approach alone. In this sense it does only allow meshes with pre-defined (by the initial mesh) stretching directions.

Intuitively one might assume the greater geometric flexibility of triangular meshes to be an advantage in this respect. A number of authors thus suggest to introduce anisotropy into the meshes by pre-defined refinement patterns for the reference triangle, deciding which pattern to use by marking individual edges of the triangles for refinement, e.g. [15, 38]. While these refinement patterns do indeed increase the aspect ratio and to some extend allow for re-alignment with solution features, this approach alone is not sufficient for achieving true anisotropic refinement, as it is not possible to exploit one-dimensional behavior of the solution.

To illustrate this we refer to Figure 5, where refinement is only desired in horizontal direction, starting from a very coarse mesh with only two triangles and four nodes. Refinement variant a) in this figure combines two triangles into a rectangle, splits the rectangle into two and the resulting rectangles again into two triangles. This variant is called blue refinement due to Kornhuber and Roitzsch [21]. It achieves the desired refinement in horizontal direction, resulting in a number of nodes proportional to the inverse of the element size in horizontal direction, thus fully exploiting the one dimensional behavior. However, it is not possible to consider this by refinement patterns of individual triangles, it is required to treat pairs of triangles. Further, as this is based on the refinement of quadrangles, it is not suitable for re-aligning the mesh.

Variant b) in the same figure treats the elements individually. There it is assumed that each edge with positive length in horizontal direction will be marked for refinement. The corresponding refinement can be considered as a fixed pattern for the reference element, or as two successive bisections of the triangle. In any case the choice of how to connect the nodes is not unique, but this has no effect on the point we want to make here. While the aspect ratio is increased for some of the resulting elements as the process is repeated, significant refinement in vertical direction is implied by this scheme, destroying the proportionality between the number of nodes and the element size in horizontal direction, thus implying a lower asymptotic order than variant a).

Variant c) assumes that only edges are marked that are in horizontal direction. While the number of nodes shows the desired linear growth, the diagonal edge of the initial mesh remains, thus the element size in horizontal direction is not reduced at all.

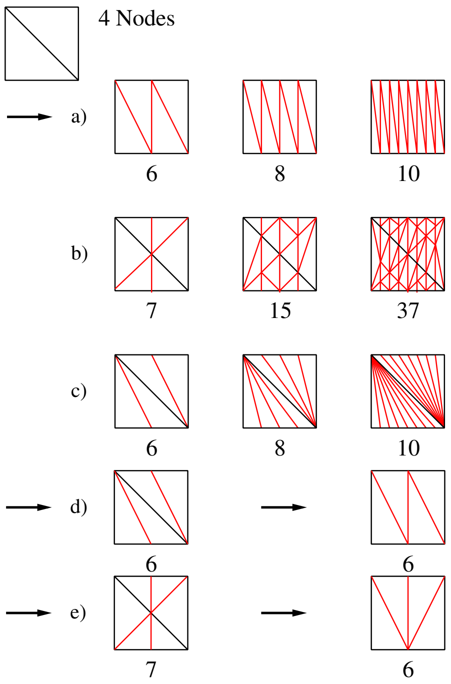


Fig. 5 Refinement options for triangles: a) blue refinement; b) splitting all edges with positive length in horizontal direction; c) splitting only edges in horizontal direction; d) combination of c) with edge swap; e) combination of b) with edge collapse.

We conclude that uni-directional refinement of triangular meshes is not possible by edge oriented refinement patterns for the triangles alone, because at least one of the effects of b) or c) in Figure 5 will render this approach inefficient.

For this reason several authors combine the refinement with other "mesh modification" operations in order to achieve anisotropic refinement, e.g. [10, 13, 16, 26, 27]:

- mesh reconnection (edge swapping),
- mesh coarsening (node removal),
- local node movement.

Parts d) and e) of Fig. 5 demonstrate that a combination of the refinement patterns with either mesh reconnection or coarsening is already sufficient to remove the flaw discussed above, by allowing to express the blue refinement as a combination of mesh modification operations. The node movement may not be necessary in this context, but is used as it accelerates the alignment of the mesh with solution features. In theory any mesh that is obtained by moving the position of one node within the patch of triangles that this nodes belongs to can also be obtained by an (infinite) sequence of refinement and coarsening operations, or approximated by a finite sequence of such operations.

According to [18, Section 4.2] the mesh generator BAMG [17] uses these operations to modify an initial mesh such that it conforms with the metric provided for the re-meshing. The class of meshes that can be obtained by combinations of these modifications is obviously very large. One could even argue in a similar way as above that by an infinite sequence of such modification steps any given mesh can be generated from any given initial triangulation of a domain (subject to constraints on topology and boundary of the meshes).

In contrast to the re-meshing approach the mesh modification approach does allow to keep track of the hierarchy in mesh refinement starting from a coarse mesh, thus offering the potential to use hierarchical techniques (e.g. multigrid or BPX) to develop fast solvers for the systems of equations arising in the discretization of the PDEs. However, the inclusion of coarsening and reconnection, which is necessary to achieve full potential for the anisotropic refinement, complicates the construction of hierarchical solution strategies. Neither of these operations is common in multigrid algorithms, so further research may be required into suitable solver strategies for the resulting class of mesh hierarchies.

To summarize this subsection we consider again our criteria. Refinement by splitting elements into smaller elements puts no bounds on the aspect ratio or local resolution of the mesh, allows re-alignment with solution features and is suitable for implementation with fast solvers. Further, this approach of refinement is computationally cheap. However, full utilization of one-dimensional behavior of the solution is only possible if the splitting is combined with further local modifications (node removal, edge swapping) which also allow more efficient re-alignment with solution features, but complicate implementation of fast solution strategies.

4.4 Summary of Properties of the Three Refinement Strategies

The criteria as given at the begin of this section are summarized for the discussed refinement approaches in Table 1.

It should be noted that the computational cost considered is that per refinement step. The picture may be a bit different if the number of refinement steps to achieve a prescribed tolerance is considered instead. Ultimately this should be considered in combination, with the solver in the loop, increasing the importance of fast solvers.

Further, the summary in Table 1 is in its nature very coarse, certainly missing some (combinations of) aspects, reflecting merely our view of important aspects in this context. Future research may of course overcome some of the deficiencies of the approaches.

Table 1 Summary of properties of the refinement approaches for triangular meshes.

method \ criterion	unbounded aspect ratio	unbounded local resolution	re-alignment	utilize one-dim behavior	computationally cheap	combination with fast solvers
node relocation	±	−	+	−	−	+
anisotropic re-meshing	+	+	+	+	±	−
splitting of elements (alone)	+	+	±	−	+	+
splitting of elements + local modifications	+	+	+	+	+	±

5 Conclusions

We have provided an overview of important aspects of anisotropic adaptive finite element algorithms, including a literature overview of suitable error estimates and refinement strategies.

An ideal anisotropic refinement algorithm strategy should fulfil all six criteria from Section 4. However, none of the three mentioned refinement strategies is completely satisfying, having deficiencies either in the ability for anisotropic refinement or in the suitability for modern hierarchical or multigrid solvers. To this end the concise discussion of the splitting of elements approach as given in Subsection 4.3 is new to our knowledge, and should influence further research.

While the only refinement strategy that satisfies all six criteria, the splitting of elements combined with local mesh modification (coarsening and reconnecting) is already known for some time in the literature, we don't know of any source that takes advantage of the availability of the (non-standard) hierarchical structure of the meshes to combine the approach with fast solvers. So we propose this as an open area of future research.

To put it more general, the close interaction of solver and refinement algorithm should be investigated to facilitate the search for more efficient overall solution strategies.

Acknowledgements. We would like to thank an anonymous referee for helpful remarks which certainly helped to improve this manuscript.

Further we would like to thank Arnd Meyer and Ulrich Langer for their inspiration and their ongoing contributions to the field of efficient numerical methods for PDEs, by adaptivity and efficient solvers for the linear systems. In this special thanks go to Arnd for many fruitful discussions.

References

[1] Ainsworth, M., Oden, J.: A Posteriori Error Estimation in Finite Element Analysis. Wiley (2000)

[2] Apel, T.: Anisotropic Finite Elements: Local Estimates and Applications. Teubner, Leipzig (1999)

[3] Apel, T., Grosman, S., Jimack, P., Meyer, A.: A new methodology for anisotropic mesh refinement based upon error gradients. Appl. Numer. Math. 50, 329–341 (2004)

[4] Babuška, I., Aziz, A.: On the angle condition in the finite element method. SIAM J. Numer. Anal. 13(2), 214–226 (1976)

[5] Baines, M.: Grid adaptation via node movement. Appl. Numer. Math. 26, 77–96 (1998)

[6] Bangerth, W., Rannacher, R.: Adaptive Finite Element Methods for Differential Equations. Birkhäuser, Basel (2003)

[7] Bank, R., Smith, R.: Mesh smoothing using a posteriori error estimates. SIAM J. Numer. Anal. 34(3), 979–997 (1997)

[8] Borouchaki, H., George, P., Hecht, F., Laug, P., Saltel, E.: Delaunay mesh generation governed by metric specifications. part i. algorithms. Finite Elements Anal. Design 25, 61–83 (1997)

[9] Cao, W.: On the error of linear interpolation and the orientation, aspect ratio, and internal angles of a triangle. SIAM J. Numer. Anal. 43, 19–40 (2005)

[10] Dolejsi, V.: Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes. Comput. Vis. Sci. 1, 165–178 (1998)

[11] Formaggia, L., Perotto, S.: New anisotropic a priori error estimates. Numer. Math. 89(4), 641–667 (2001)

[12] Formaggia, L., Micheletti, S., Perotto, S.: Anisotropic mesh adaptation in computational fluid dynamics: Application to the advection-diffusion-reaction and the Stokes problems. Appl. Numer. Math. 51(4), 511–533 (2004)

[13] Fortin, M.: Anisotropic mesh adaptation through hierarchical error estimators. In: Minev, P., Lin, Y. (eds.) Scientific Computing and Applications, vol. 7, pp. 53–65. Nova Science Publishers (2001)

[14] Grajewski, M., Köster, M., Turek, S.: Numerical analysis and implementational aspects of a new multilevel grid deformation method. Appl. Numer. Math. 60(8), 767–781 (2010)

[15] Grosman, S.: Adaptivity in anisotropic finite element calculations. PhD thesis, TU Chemnitz, Chemnitz, Germany (2006)

[16] Habashi, W., Dompierre, J., Bourgault, Y., Ait-Ali-Yahia, D., Fortin, M., Vallet, M.G.: Anisotropic mesh adaptation: towards user-independent, mesh-independent and solver-independant cfd. part i: general principles. Int. J. Numer. Meth. Fluids 32, 725–744 (2000)

[17] Hecht, F.: Bidimensional anisotropic mesh generator. Tech. rep., INRIA, Rocquencourt, software (1997), http://www.ann.jussieu.fr/hecht/ftp/bamg/

[18] Huang, W.: Mathematical principles of anisotropic mesh adaptation. Comm. Comput. Phys. 1(2), 276–310 (2005)

[19] Huang, W., Russell, R.: Adaptive Moving Mesh Methods. Applied Mathematical Sciences. Springer (2011)

[20] Huang, W., Kamenski, L., Lang, J.: A new anisotropic mesh adaptation method based upon hierarchical a posteriori error estimates. J. Comput. Phys. 229(6), 2179–2198 (2010)

[21] Kornhuber, R., Roitzsch, R.: On adaptive grid refinement in the presence of internal or boundary layers. Impact Comput. Sci. Engrg. 2, 40–72 (1990)

[22] Kunert, G.: A posteriori error estimation for anisotropic tetrahedral and triangular finite element meshes. PhD thesis, TU Chemnitz (1999), http://archiv.tu-chemnitz.de/pub/1999/0012/index.html

[23] Kunert, G.: Toward anisotropic mesh construction and error estimation in the finite element method. Numer. Meth. Partial Diff. Eqns. 18(5), 625–648 (2002)

[24] Kunert, G., Verfürth, R.: Edge residuals dominate a posteriori error estimates for linear finite element methods on anisotropic triangular and tetrahedral meshes. Numer. Math. 86, 283–303 (2000)

[25] Lang, J., Cao, W., Huang, W., Russel, R.: A two-dimensional moving finite element method with local refinement based on a posteriori error estimates. Appl. Numer. Math. 46, 75–94 (2003)

[26] Li, X., Shephard, M., Beall, M.: 3d anisotropic mesh adaptation by mesh modification. Comput. Meth. Appl. Mech. Engrg. 194, 4915–4950 (2005)

[27] Mahmood, R., Jimack, P.: Locally optimal unstructured finite element meshes in 3 dimensions. Comput. Struct. 82(23–26), 2105–2116 (2004)

[28] Miller, K., Miller, R.: Moving finite elements 1. SIAM J. Numer. Anal. 18(6), 1019–1032 (1981)

[29] Picasso, M., Alauzet, F., Borouchaki, H., George, P.L.: A numerical study of some Hessian recovery techniques on isotropic and anisotropic meshes. SIAM J. Sci. Comput. 33(3), 1058–1076 (2011)

[30] Richter, T.: A posteriori error estimation and anisotropy detection with the dual-weighted residual method. Int. J. Numer. Meth. Fluids 62(1), 90–118 (2010)

[31] Roos, H.–G.: Layer-adapted grids for singular perturbation problems. ZAMM Z. Angew. Math. Mech. 78(5), 291–309 (1998)

[32] Beuchler, S., Meyer, A.: SPC-PM3AdH v1.0 - Programmer's Manual. Tech. Rep. Preprint SFB393/01-08, TU Chemnitz, Chemnitz (2001), http://www.tu-chemnitz.de/sfb393/

[33] Schneider, R.: Applications of the discrete adjoint method in computational fluid dynamics. PhD thesis, University of Leeds (2006), http://www.comp.leeds.ac.uk/research/pubs/theses/schneider.pdf

[34] Schneider, R., Jimack, P.: Toward anisotropic mesh adaption based upon sensitivity of a posteriori estimates. School of Computing Research Report Series 2005.03, University of Leeds (2005), http://www.engineering.leeds.ac.uk/computing/research/publications/reports/200407/

[35] Shewchuk, J.: What is a good linear finite element? interpolation, conditioning, anisotropy, and quality measures. Preprint, Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA (2002), http://www.cs.berkeley.edu/~jrs/papers/elemj.pdf

[36] Vallet, M.G., Manole, C.M., Dompierre, J., Dufour, S., Guibault, F.: Numerical comparison of some Hessian recovery techniques. Int. J. Numer. Meth. Engrg. 72(8), 987–1007 (2007)

[37] Walkley, M., Jimack, P., Berzins, M.: Anisotropic adaptivity for finite element solutions of 3-d convection-dominated problems. Int. J. Numer. Methods Fluids 40, 551–559 (2002)

[38] Wang, D., Li, R., Yan, N.: An edge-based anisotropic mesh refinement algorithm and its application to interface problems. Comm. Comput. Phys. 8(3), 511–540 (2010)

A Postprocessing Finite Element Strategy for Poisson's Equation in Polygonal Domains: Computing the Stress Intensity Factors

Boniface Nkemzi and Michael Jung

Abstract. We present a new finite element algorithm for computing the stress intensity factors and the solution of boundary value problems for the Poisson equation in two-dimensional domains with corners. The method makes use of an explicit expression for the stress intensity factors in terms of the function of the right hand side, the solution of the boundary value problem and smooth cutoff functions to compute from an initial finite element solution approximations of the stress intensity factors. The computed values of the stress intensity factors are then used for post processing the finite element solution and the approximated stress intensity factors. The algorithm leads to good approximations of both stress intensity factors and the finite element solution of the boundary value problem on quasi regular meshes. The results are illustrated by numerical experiments.

1 Introduction

Solutions of elliptic boundary value problems are known to exhibit singularities in neighborhoods of corners, cracks, edges, conical vertices, or near boundary points with change in boundary conditions, see for example, [13, 17, 18, 19, 23]. In fact, according to the general theory on H^2-regularity, the generalized solution u of any linear elliptic boundary value problem in domains with geometric singularities and with the right hand side datum in $L_2(\Omega)$ can be split as a sum in the form

Boniface Nkemzi
Faculty of Science, Department of Mathematics, University of Buea, Cameroon
e-mail: nkemzi@yahoo.com

Michael Jung
Fakultät Informatik/Mathematik, Hochschule für Technik und Wirtschaft Dresden,
01069 Dresden, Germany
e-mail: mjung@informatik.htw-dresden.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 153–173.
springerlink.com

$$u = \sum_{k=1}^{N} \gamma_k s_k + w \tag{1}$$

where w is a regular part whose behavior is not affected by the presence of the corners or edges and a singular part which is composed of explicitly defined special singular functions s_k that depend on the geometry, the differential operator and the boundary conditions, and of unknown coefficients γ_k, generally referred to as stress intensity coefficients or stress intensity factors in reference to elasticity theory.

It is well known, that the presence of singularities in the solutions of elliptic boundary value problems may reduce severely the accuracy of standard numerical methods of approximations, such as, the finite element and the finite difference methods. In engineering applications one is usually interested in:

1. adaptive strategies that will improve the accuracy of the approximate solutions,
2. strategies for the accurate computation of the coefficients γ_k of the singularities. These coefficients characterize the strength of the singularity and are useful in most engineering computations.

In the case of corners in two-dimensional domains like the problems we consider here, the stress intensity coefficients γ_k are unique scalar constants. It has been shown that, in this case, the optimal rate of convergence of the finite element solutions can be recovered by means of graded mesh refinements, see for example [1, 4, 5, 20, 26, 27, 29, 30]. However, this approach does not lead directly to the computation of the coefficients γ_k. On the other hand there exist several methods of computing directly the coefficients γ_k without necessarily requiring the solution u, see for example [21, 24, 28] and the references cited therein.

Several other methods exist with which both the solution of the boundary value problem and the coefficients γ_k of the singularities can be computed simultaneously. We mention here the following three:

1. The singular function method: This method consists in augmenting the usual finite element space by the known singular functions, see, for example, [16, 31]. This approach leads to an approximation of the corresponding stress intensity factors. However, it has been shown that the approximate stress intensity factors do not converge in many cases, see [6, 14].
2. The dual singular function method: This method consists in calculating a priori the stress intensity factors. This method has been investigated by many authors, see, for example, [6, 8, 9, 10, 11, 15, 25]. A priori error estimates show that optimal accuracy can be recovered by this means.
3. The singular complement method: This method was introduced and investigated in [2, 3] and it consists of adding to the finite element space appropriate singular test functions.

It should be noted that the methods of points 2. and 3. above essentially describe various post-processing strategies for the finite element solutions using the splitting (1). The methods proposed, for example, in [2, 3, 25] have the advantage that no cutoff functions are required. Cutoff functions are known to be a source of instability in the post-processing computations. However, in [2, 3, 25] the stress intensity factors have to be computed from integral expressions defined over the entire domain containing the singular points. These computations definitely can be very expensive and the accuracy polluted by the effect of the singular points.

It is worth noting that there is no obvious extension of the methods listed above to boundary value problems in three-dimensional domains with edges.

The main motivations of this paper are two-fold: (1) To propose a stable, efficient and robust post-processing strategy for the finite element approximation of solutions and associated stress intensity factors for boundary value problems in two-dimensional domains with corners. (2) To develop a method which can easily be adapted for the computation of edge stress intensity functions in three dimensions.

We note that in our method, cutoff functions are not needed in the pre- and post-calculations of the finite element solutions, see Sect. 3, as it is the case with most of the existing methods, see for example [6, 8, 9, 10, 11, 15]. We use cutoff functions only for the computation of the stress intensity factors. A significant difference in our method lies in the derivation of formulas for the stress intensity factors. Here, the formulas are derived from the explicit representation of the solution in the neighborhood of the corner, see Sect. 2. Hence, our method does not require the knowledge of the dual singular solution of the adjoint problem as is the case, for example, in [6, 8, 9, 10, 11, 15]. Our method can be exploited for the derivation of explicit expressions for edge stress intensity functions for boundary value problems in three-dimensional domains with edges. Moreover, numerical experiments show that the postprocessing finite element method defined in Sect. 3 is stable and converges optimally in the H^1–norm.

This paper is organized as follows: In Sect. 2 the boundary value problem and the derivation of expressions for the stress intensity factors are presented. In Sect. 3 we explain the main approximation algorithm and prove a priori error estimates. It is shown that on quasi uniform mesh without any adaptation, the approximation of the stress intensity factors γ_k using formula (11) yields an accuracy of at least $O(h^{2\sigma})$ for some $0 < \sigma < 1$. However, the initially computed values for γ_k can be used to define a post processing strategy for the finite element solution of the boundary value problem based on the splitting (1) and hence compute better approximations for the stress intensity factors γ_k. The new adaption is shown to yield an accuracy of at least the order $O(h^{\min\{1,2\sigma\}})$ for the solution u of (2) in the $H^1(\Omega)$-norm. In Sect. 4, we present several numerical experiments.

2 Analytical Preliminaries

2.1 The Model Boundary Value Problem and Notations

As model problem, we consider for simplicity the mixed boundary value problem with homogeneous boundary conditions

$$-\Delta u = f \quad \text{in}\, \Omega, \qquad u = 0 \quad \text{on}\, \Gamma_{\mathscr{D}}, \qquad \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on}\, \Gamma_{\mathscr{N}}, \tag{2}$$

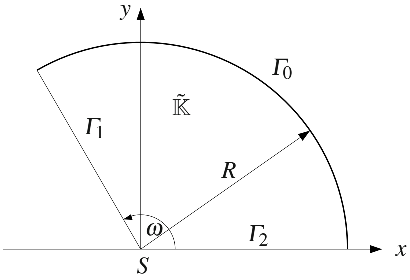
where $f \in L_2(\Omega)$ and $\Omega \subset \mathbb{R}^2$ is a bounded domain with boundary $\Gamma = \bar{\Gamma}_{\mathscr{D}} \cup \bar{\Gamma}_{\mathscr{N}}$ with $\Gamma_{\mathscr{D}} \cap \Gamma_{\mathscr{N}} = \emptyset$, and $\mathbf{n} = (n_1, n_2)^\top$ denotes the outward unit normal on Γ. We assume for the sake of uniqueness of solution of the boundary value problem (2) that meas$(\Gamma_{\mathscr{D}}) > 0$ (Lebesgue measure).

Suppose that Ω has one corner S with interior angle $\omega \in (0, 2\pi]$ and that the boundary Γ is straight line in some neighborhood of the corner S and is sufficiently smooth outside S. Let (r, θ) denote local polar coordinates with respect to S and define a small circular sector neighborhood $\tilde{\mathbb{K}} \subset \Omega$ of S with radius R and angle ω by

$$\tilde{\mathbb{K}} := \{(x,y) \in \Omega : x = r\cos\theta,\ y = r\sin\theta,\ 0 < r < R,\ 0 < \theta < \omega\} \tag{3}$$

with boundary $\partial\tilde{\mathbb{K}} = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2$, see Fig. 1.

Fig. 1 A circular sector.



Define with respect to S a smooth cutoff function $\eta \in C^\infty[0, \infty)$ by

$$\eta(x,y) = \eta(r) := \begin{cases} 1 & \text{for } 0 \le r \le R/3 \\ 0 \le \eta \le 1 & \text{for } R/3 \le r \le 2R/3 \\ 0 & \text{for } r \ge 2R/3 \end{cases} \tag{4}$$

where R is from (3), that is, supp$(\eta) \subset \tilde{\mathbb{K}}$.

Suppose $u \in V_0(\Omega) := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_{\mathscr{D}}\}$ is the unique weak solution of the boundary value problem (2). Then by (4) the function $u_\eta := \eta\, u$ is nonzero only in the circular sector neighborhood $\tilde{\mathbb{K}}$ of the vertex S and it coincides with u near the vertex S. Thus, u_η and u have the same singular behavior near S, since regularity of solutions of elliptic boundary value problem is a local property. Thus, the study of the regularity properties of the solution $u \in V_0(\Omega)$ of (2) can be reduced to the study of the regularity properties of the function u_η. The following result will be helpful.

Lemma 1. *For $f \in L_2(\Omega)$, let $u \in V_0(\Omega)$ be the weak solution of the boundary value problem (2). Then $u_\eta \in V_0(\tilde{\mathbb{K}})$ and u_η is the unique weak solution of the boundary value problem*

$$\begin{aligned} -\Delta u_\eta &= f_\eta && \textit{in} \quad \tilde{\mathbb{K}}\,, \\ u_\eta &= 0 && \textit{on} \quad \Gamma_0\,, \\ u_\eta &= 0 && \textit{on} \quad \Gamma_i \quad \textit{if} \quad \Gamma_i \subset \Gamma_{\mathscr{D}}\,,\ i=1,2\,, \\ \frac{\partial u_\eta}{\partial \mathbf{n}} &= 0 && \textit{on} \quad \Gamma_i \quad \textit{if} \quad \Gamma_i \subset \Gamma_{\mathscr{N}}\,,\ i=1,2 \end{aligned} \tag{5}$$

where $f_\eta \in L_2(\hat{\mathbb{K}})$ *is defined by*

$$f_\eta := \eta f - u\Delta\eta - 2\nabla u \cdot \nabla\eta, \quad \|f_\eta\|_{L_2(\tilde{\mathbb{K}})} \leq C\|f\|_{L_2(\Omega)} \tag{6}$$

The proof is straight forward and so we omit it.

2.2 Singular Decomposition of Solution

We will use the following notations:

$$\phi_k(\theta) = \begin{cases} \sin\lambda_k\theta,\ \lambda_k = \frac{k\pi}{\omega}, & k \in \mathbb{N} \qquad \text{if} \ \ \Gamma_1 \cup \Gamma_2 \subset \Gamma_{\mathscr{D}} \\ \cos\lambda_k\theta,\ \lambda_k = (k-\frac{1}{2})\frac{\pi}{\omega}\,, & k \in \mathbb{N} \qquad \text{if} \ \ \Gamma_1 \subset \Gamma_{\mathscr{D}}\,,\ \Gamma_2 \subset \Gamma_{\mathscr{N}} \\ \sin\lambda_k\theta,\ \lambda_k = (k-\frac{1}{2})\frac{\pi}{\omega}\,, & k \in \mathbb{N} \qquad \text{if} \ \ \Gamma_1 \subset \Gamma_{\mathscr{N}}\,,\ \Gamma_2 \subset \Gamma_{\mathscr{D}} \\ \cos\lambda_k\theta,\ \lambda_k = (k-1)\frac{\pi}{\omega}\,, & k \in \mathbb{N}\setminus\{1\} \ \text{if} \ \ \Gamma_1 \cup \Gamma_2 \subset \Gamma_{\mathscr{N}} \\ \frac{1}{2}, \qquad \lambda_1 = 0\,, & k = 1 \qquad \text{if} \ \ \Gamma_1 \cup \Gamma_2 \subset \Gamma_{\mathscr{N}}\,. \end{cases} \tag{7}$$

Obviously, the system of functions $\{\phi_k : k \in \mathbb{N}\}$ is orthogonal and complete in $L_2(0,\omega)$. Hence, the functions u_η and f_η from (5) can be represented in Fourier series in the form

$$u_\eta(x,y) = u_\eta(r,\theta) = \sum_{k=1}^{\infty} u_k(r)\phi_k(\theta), \quad f_\eta(x,y) = f_\eta(r,\theta) = \sum_{k=1}^{\infty} f_k(r)\phi_k(\theta) \tag{8}$$

with Fourier coefficients defined as usual by

$$u_k(r) = \frac{2}{\omega}\int_0^\omega u_\eta(r,\theta)\phi_k(\theta)d\theta\,, \quad f_k(r) = \frac{2}{\omega}\int_0^\omega f_\eta(r,\theta)\phi_k(\theta)d\theta\,. \tag{9}$$

The main result in this section is the following:

Theorem 1. *For $f \in L_2(\Omega)$, let $u \in V_0(\Omega)$ be the weak solution of (2). Suppose that the domain Ω has only one vertex S with angle ω and let the functions $\{\phi_k : k \in \mathbb{N}\}$ and the real numbers $\{\lambda_k : k \in \mathbb{N}\}$ be as given in (7). Then the unique weak solution $u \in V_0(\Omega)$ of (2) can be split as a sum of a regular and a singular part in the form*

$$u(x,y) = w(x,y) + s(x,y) := w(x,y) + \sum_{0<\lambda_k<1} \gamma_k s_k, \quad w \in H^2(\Omega)\,, \tag{10}$$

where the stress intensity factors γ_k are given explicitly by

$$\gamma_k := \frac{1}{\omega\lambda_k}\int\limits_{\tilde{\mathbb{K}}} f_\eta s_{-k} dx \quad \textit{for} \quad 0<\lambda_k<1 \tag{11}$$

with f_η from (6) and

$$s_k(r,\theta) = r^{\lambda_k}\phi_k(\theta), \qquad s_{-k}(r,\theta) = r^{-\lambda_k}\phi_k(\theta)\,.$$

Moreover, there exists a constant $C>0$ such that

$$|\gamma_k| + \|w\|_{H^2(\Omega)} \le C\|f\|_{L_2(\Omega)}\,.$$

Proof. We give a sketch of the proof.

The solution u_η of the boundary value problem (5) is given explicitly by (8), where the Fourier coefficients u_k are the solutions of the two-point boundary value problems

$$-u_k'' - \frac{u_k'}{r} + \lambda_k^2\frac{u_k}{r^2} = f_k \ \text{ in } \ (0,R)\,,\ |u_k(0)|<\infty,\ u_k(R)=0\,.$$

Hence, the Fourier coefficients u_k are given by

$$u_k(r) = r^{\lambda_k}\frac{1}{2\lambda_k}\int\limits_r^R f_k(\tau)\tau^{1-\lambda_k}d\tau + \frac{r^{-\lambda_k}}{2\lambda_k}\int\limits_0^r f_k(\tau)\tau^{1+\lambda_k}d\tau\,. \tag{12}$$

On the other hand the solution u of (2) can be expressed in the form

$$u(x,y) = (1-\eta)\,u + \eta\,u = (1-\eta)\,u + \sum_{k=1}^{\infty} u_k(r)\phi_k(\theta)\,.$$

By standard regularity principles, see for example [17, 18, 19], we get that

$$(1-\eta)\,u \in H^2(\Omega\setminus\tilde{\mathbb{K}})\,.$$

If $0<\lambda_k<1$ for some $k\in\mathbb{N}$, then the function $u_\eta = \eta\,u$ has singularities. We write in this case formula (12) in the equivalent form

$$u_k(r) = r^{\lambda_k}\frac{1}{2\lambda_k}\left(\int\limits_0^R f_k(\tau)\tau^{1-\lambda_k}d\tau - \int\limits_0^r f_k(\tau)\tau^{1-\lambda_k}d\tau\right) + \frac{r^{-\lambda_k}}{2\lambda_k}\int\limits_0^r f_k(\tau)\tau^{1+\lambda_k}d\tau\,.$$

The first summand in this expression is singular and the other two summands are regular. Therefore, these two summands can be incorporated into the regular part of the solution. Using the definition (9) of the Fourier coefficients f_k we get for the first summand

$$\begin{aligned} r^{\lambda_k}\frac{1}{2\lambda_k}\int\limits_0^R f_k(\tau)\tau^{1-\lambda_k}d\tau &= r^{\lambda_k}\frac{1}{2\lambda_k}\int\limits_0^R \frac{2}{\omega}\int\limits_0^\omega f_\eta(\tau,\theta)\phi_k(\theta)d\theta\tau^{1-\lambda_k}d\tau \\ &= r^{\lambda_k}\frac{1}{\omega\lambda_k}\int\limits_0^R\int\limits_0^\omega f_\eta(\tau,\theta)\tau^{-\lambda_k}\phi_k(\theta)\tau d\theta d\tau \\ &= r^{\lambda_k}\frac{1}{\omega\lambda_k}\int\limits_{\tilde{\mathbb{K}}} f_\eta\tau^{-\lambda_k}\phi_k dx \;=\; r^{\lambda_k}\frac{1}{\omega\lambda_k}\int\limits_{\tilde{\mathbb{K}}} f_\eta s_{-k}dx. \end{aligned}$$

Consequently, we get that the singular part of u which is the same as the singular part of u_η is given by

$$s(r,\theta) = \sum_{0<\lambda_k<1} \gamma_k r^{\lambda_k}\phi_k(\theta)$$

where γ_k is defined by (11). □

Next we show that formula (11) does not depend on the particular choice of the cutoff function η.

Lemma 2. *Formula (11) is independent of the particular cutoff function* η.

Proof. We will need the following notations. For $0 < \varepsilon < R/3$ let

$$B_\varepsilon = \{(x,y)\in\tilde{\mathbb{K}} \,:\, x = r\cos\theta,\; y = r\sin\theta,\; 0<r<\varepsilon,\; 0<\theta<\omega\}$$

and

$$\tilde{\mathbb{K}}_\varepsilon = \tilde{\mathbb{K}}\setminus B_\varepsilon .$$

Taking note that $\Delta s_{-k} = 0$, we get after integrating by parts twice the relations

$$\begin{aligned} \gamma_k &= \frac{1}{\omega\lambda_k}\int\limits_{\tilde{\mathbb{K}}} f_\eta s_{-k}dx = -\frac{1}{\omega\lambda_k}\int\limits_{\tilde{\mathbb{K}}}\Delta(\eta u)s_{-k}dx = -\lim_{\varepsilon\to 0}\frac{1}{\omega\lambda_k}\int\limits_{\tilde{\mathbb{K}}_\varepsilon}\Delta(\eta u)s_{-k}dx \\ &= \lim_{\varepsilon\to 0}\frac{1}{\omega\lambda_k}\int\limits_{\partial\tilde{\mathbb{K}}_\varepsilon\cap\partial B_\varepsilon} u\frac{\partial s_{-k}}{\partial\mathbf{n}}ds - \lim_{\varepsilon\to 0}\frac{1}{\omega\lambda_k}\int\limits_{\partial\tilde{\mathbb{K}}_\varepsilon\cap\partial B_\varepsilon} s_{-k}\frac{\partial u}{\partial\mathbf{n}}ds \end{aligned}$$

where we see that the last equation does no longer contain η. □

Remark 1. We note that formula (11) agrees in the case of Dirichlet boundary condition with the corresponding formula for stress intensity factors as given, for example, in [6].

3 The Finite Element Approximation

In this section we introduce the finite element approach and investigate the error.

3.1 Approximation on Quasi Uniform Mesh

Here we study the approximation of the solution u of problem (2) and the associated stress intensity factors γ_k from (11) by the usual finite element method using piecewise linear polynomials on quasi uniform meshes. Thus, for a discretization parameter $0 < h \le h_0$ (h_0 sufficiently small), let $\mathscr{T}_h := \{T\}$ be a partition of the domain $\bar{\Omega}$ into disjoint triangles T such that the usual assumptions are satisfied, see, for example, [12]. On $\mathscr{T}_h$ we define the finite element space V_{0h} by

$$V_{0h} := \{v_h \in C(\bar{\Omega}) \,:\, v_h\big|_T \in P_1(T),\ v_h = 0 \ \text{ on } \bar{\Gamma}_{\mathscr{D}}\},$$

where $P_1(T)$ denotes the set of polynomials of degree ≤ 1 on T. Then, the finite element approximation $u_h \in V_{0h}$ of the weak solution u of problem (2) is determined by solving formerly the Galerkin equation:

$$\text{Find} \quad u_h \in V_{0h} \,:\, \int\limits_{\Omega} \nabla u_h \cdot \nabla v\, dx = \int\limits_{\Omega} f v\, dx \quad \text{for all} \quad v \in V_{0h}\,. \tag{13}$$

According to the classical theory on the finite element method, see, for example, [12], and standard regularity results, see, for example, [17, 18], the accuracy of the above approximation on a quasi uniform mesh in the Sobolev spaces $H^\ell(\Omega)$ $(\ell = 0,1)$ is

$$\|u - u_h\|_{H^\ell(\Omega)} = O(h^{(2-\ell)\sigma}), \quad \ell = 0,1\,, \tag{14}$$

for every $0 < \sigma < \lambda$, where $\lambda := \min\{\lambda_k : 0 < \lambda_k < 1\}$ with λ_k from (7).

Using the Galerkin solution u_h from (13), we determine an approximation γ_{kh} of the stress intensity factor γ_k from (11) by

$$\gamma_{kh} := \frac{1}{\omega\lambda_k} \int\limits_{\tilde{\mathbb{K}}} f_{\eta h} s_{-k} dx\,, \tag{15}$$

where

$$f_{\eta h} := \eta f - u_h \Delta\eta - 2\nabla\eta \cdot \nabla u_h\,. \tag{16}$$

Theorem 2. *For $f \in L_2(\Omega)$, let u be the weak solution of the boundary value problem (2) and u_h its finite element solution defined according to (13). Suppose that the domain Ω has only one singular corner. Let γ_k denote the stress intensity factors defined according to (11) and let γ_{kh} be their approximations defined according to (15). Then under the assumption that the integral (15) has been accurately computed, $\gamma_k - \gamma_{kh}$ satisfies the estimate*

$$|\gamma_k - \gamma_{kh}| \le C h^{2\sigma} \|f\|_{L_2(\Omega)}\,. \tag{17}$$

Proof. Set

$$B(r_1, r_2) := \{(r,\theta) :\ r_1 < r < r_2,\ 0 < \theta < \omega\} \cap \Omega\,.$$

Considering (11), (6) and taking note that the function $u\Delta\eta - 2\nabla\eta \cdot \nabla u$ is non zero only in $B(\frac{R}{3}, \frac{2R}{3})$, and applying integration by parts and some trivial simplifications we get

$$\begin{aligned}\gamma_k &= \frac{1}{\omega\lambda_k} \int\limits_{\tilde{\mathbb{K}}} \big(\eta f - u\Delta\eta - 2\nabla\eta \cdot \nabla u\big) s_{-k}\, dx \\ &= \frac{1}{\omega\lambda_k} \left(\int\limits_{\tilde{\mathbb{K}}} \eta f s_{-k} dx + \int\limits_{B(\frac{R}{3},\frac{2R}{3})} u\Delta(\eta s_{-k}) dx \right).\end{aligned} \tag{18}$$

Now, taking account of (11), (6), (15), (16) and (14) we get the estimates

$$\begin{aligned}|\gamma_k - \gamma_{kh}| &\leq \int\limits_{B(\frac{R}{3},\frac{2R}{3})} |u - u_h||\Delta(\eta s_{-k})| dx \\ &\leq \|u - u_h\|_{L_2(B(\frac{R}{3},\frac{2R}{3}))} \|\Delta(\eta s_{-k})\|_{L_2(B(\frac{R}{3},\frac{2R}{3}))} \leq Ch^{2\sigma} \|f\|_{L_2(\Omega)}\end{aligned}$$

for every $0 < \sigma < \lambda$, since $\|\Delta(\eta s_{-k})\|_{L_2(B(\frac{R}{3},\frac{2R}{3}))}$ is bounded. □

Remark 2. We observe that the error $\|u - u_h\|_{L_2(B(\frac{R}{3},\frac{2R}{3}))}$ is measured away from the singular point. Hence, the larger we choose the parameter R the better the convergence. In fact, we can obtain $|\gamma_k - \gamma_{kh}| = O(h^2)$ by choosing R sufficiently large.

Remark 3. We observe that if the finite element solution u_h is computed using the graded mesh refinement technique or any other adaptive method that has the convergence order $\|u - u_h\|_{L_2(\Omega)} = O(h^2)$, then we are going to obtain the convergence order of $|\gamma_k - \gamma_{kh}| = O(h^2)$ for the approximated values of the stress intensity factors.

3.2 *Improving the Accuracy of the Approximation*

The accuracy of the finite element solution u_h from (13) and the stress intensity factors γ_{kh} from (15) can be improved as follows.

We split the solution of (2) as before in the form

$$u = w + \sum_{0<\lambda_k<1} \gamma_k s_k, \quad s_k = r^{\lambda_k} \phi_k(\theta), \quad w \in H^2(\Omega).$$

Taking note that the singular parts s_k are harmonic, we see that the regular part w solves the boundary value problem

$$-\Delta w = f \quad \text{in } \Omega, \quad w = -\sum_{0<\lambda_k<1} \gamma_k s_k \quad \text{on } \Gamma_{\mathscr{D}}, \quad \frac{\partial w}{\partial \mathbf{n}} = -\sum_{0<\lambda_k<1} \gamma_k \frac{\partial s_k}{\partial \mathbf{n}} \quad \text{on } \Gamma_{\mathscr{N}}.$$

It follows that the finite element approximation w_h of the regular part w is obtained by solving the Galerkin problem:

Find $w_h \in V_h$ such that

$$\int_{\Omega} \nabla w_h \cdot \nabla v \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma_{\mathscr{N}}} \Big(- \sum_{0<\lambda_k<1} \gamma_k \frac{\partial s_k}{\partial \mathbf{n}} \Big) v \, ds \quad \text{for all} \quad v \in V_{0h} , \qquad (19)$$

where

$$V_h := \{ v_h \in C(\bar{\Omega}) \, : \, v_h\big|_T \in P_1(T) , \; v_h = - \sum_{0<\lambda_k<1} \gamma_k s_{kh} \; \text{ on } \; \bar{\Gamma}_{\mathscr{D}} \} ,$$

$$V_{0h} := \{ v_h \in C(\bar{\Omega}) \, : \, v_h\big|_T \in P_1(T) , \; v_h = 0 \; \text{ on } \; \bar{\Gamma}_{\mathscr{D}} \} ,$$

where s_{kh} is the linear interpolant of s_k.

Using this idea we can perform the following algorithm.

A post-processing iterative procedure for computing γ_{kh} and u_h

1. Set $\gamma_{kh}^{(0)} = 0$.
2. For $j = 1, 2, \ldots$ determine the finite element solution $w_h^{(j)} \in \tilde{V}_h$ such that

$$\int_{\Omega} \nabla w_h^{(j)} \cdot \nabla v \, dx = \int_{\Omega} f v \, dx + \int_{\Gamma_{\mathscr{N}}} \Big(- \sum_{0<\lambda_k<1} \gamma_{kh}^{(j-1)} \frac{\partial s_k}{\partial \mathbf{n}} \Big) v \, ds \quad \forall v \in V_{0h} , \qquad (20)$$

where

$$\tilde{V}_h := \{ v_h \in C(\bar{\Omega}) \, : \, v_h\big|_T \in P_1(T) , \; v_h = - \sum_{0<\lambda_k<1} \gamma_{kh}^{(j-1)} s_{kh} \; \text{ on } \; \bar{\Gamma}_{\mathscr{D}} \} .$$

3. Compute

$$u_h^{(j)} = w_h^{(j)} + \sum_{0<\lambda_k<1} \gamma_{kh}^{(j-1)} s_k . \qquad (21)$$

4. Compute

$$\gamma_{kh}^{(j)} = \frac{1}{\omega \lambda_k} \int_{\tilde{\mathbb{K}}} f_{\eta h} s_{-k} dx \quad \text{with} \quad f_{\eta h} := \eta f - u_h^{(j)} \Delta \eta - 2 \nabla \eta \cdot \nabla u_h^{(j)} . \qquad (22)$$

Note that $u_h^{(1)} = w_h^{(1)} = u_h$, i.e. $u_h^{(1)}$ is the solution of problem (13). The next theorem contains an a priori estimate of the error $u - u_h^{(2)}$ of the solution $u_h^{(2)}$ obtained by our post-processed finite element algorithm.

Theorem 3. *For $f \in L_2(\Omega)$, let $u \in V_0(\Omega)$ be the unique weak solution of the boundary value problem (2) and let $u_h^{(2)}$ be its post-processed finite element approximation defined according to (21) obtained after performing steps 2 till 4 two times. Assume that the domain Ω has only one corner with angle ω such that $0 < \lambda_k < 1$ for some $k \in \mathbb{N}$, where λ_k are defined in (7). Then the error estimate*

$$\|u-u_h^{(2)}\|_{H^1(\Omega)} \le Ch^{\min\{1,2\sigma\}} \tag{23}$$

for every $0<\sigma<\lambda$, *where* $\lambda := \min\{\lambda_k : 0<\lambda_k<1\}$ *holds.*

Proof. The estimate (23) follows by applying the 1st Lemma of Strang and the estimate (17). □

From the estimate (23) we see, that we get in the case of $\sigma \ge 0.5$ the same convergence order as in the case of a smooth solution.

Our numerical examples, see Sect. 4, show that the error of $u_h^{(2)}$ in the L_2 norm has also the same order as in the case of a smooth solution.

4 Numerical Experiments

In this section, we present results of some numerical experiments to illustrate the performance of our method. The first example is a Dirichlet problem for the Laplace equation in a circular sector. Here, the exact solution and the stress intensity factor are known. The second example is the so-called Motz's problem, which is a bench mark problem in mathematics and physics literature for testing various algorithms for the treatment of boundary value problems for two-dimensional Laplace equation with boundary singularities. Though, the exact value for the stress intensity factor is not known, pretty good approximations exist, see, for example, [7, 24].

4.1 Example 1

Let $\Omega \subset \mathbb{R}^2$ be the circular sector given by

$$\Omega = \{(x,y) : x = r\cos\theta,\ y = r\sin\theta,\ 0<r<R,\ 0<\theta<3\pi/2\}$$

with boundary $\Gamma = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2$, where

$$\begin{aligned}
\Gamma_0 &= \{(x,y) : x = R\cos\theta,\ y = R\sin\theta,\ 0<\theta<3\pi/2\},\\
\Gamma_1 &= \{(x,y) : x = r\cos\frac{3\pi}{2},\ y = r\sin\frac{3\pi}{2},\ 0<r<R\},\\
\Gamma_2 &= \{(x,y) : 0\le x\le R,\ y=0\}.
\end{aligned}$$

We consider the Dirichlet problem

$$\begin{cases}
-\Delta u = -\frac{32}{9}\sin\left(\frac{2}{3}\theta\right) & \text{in} \quad \Omega,\\
u = (R^2+4R^{2/3})\sin\left(\frac{2}{3}\theta\right) & \text{on} \quad \Gamma_0,\\
u = 0 & \text{on} \quad \Gamma_1\cup\Gamma_2.
\end{cases} \tag{24}$$

Clearly, the solution of problem (24) has a singularity due to the reentrant corner at the origin. In fact the solution is given by

$$u(r,\theta) = \left(r^2 + 4r^{2/3}\right)\sin\left(\frac{2}{3}\theta\right).$$

That is, the solution u has one singular part and the stress intensity factor is $\gamma_1 = 4$.

Taking $R = 1$ we define the cutoff function

$$\eta(r) = \begin{cases} 1 & \text{if } 0 \le \frac{1}{3}R \\ \frac{15}{16}\left(\frac{8}{15} - \left(\frac{6r}{R} - 3\right) + \frac{2}{3}\left(\frac{6r}{R} - 3\right)^3 - \frac{1}{5}\left(\frac{6r}{R} - 3\right)^5\right) & \text{if } \frac{1}{3}R \le r \le \frac{2}{3}R \\ 0 & \text{if } r \ge \frac{2}{3}R. \end{cases} \tag{25}$$

It is easily verified that $\eta \in C^2[0,1]$. Furthermore, a straight forward computation shows that

$$\gamma_1 = \frac{1}{\pi}\int\limits_{\Omega}\left(-\frac{32}{9}\sin\left(\frac{2}{3}\,\theta\right)\eta - u\Delta\eta - 2\nabla\eta\cdot\nabla u\right)s_{-1}dx = 4\,,\;\; s_{-1} = r^{-2/3}\sin\left(\frac{2}{3}\theta\right).$$

Thus formula (11) is verified.

In our numerical experiments we use starting from the triangulation $\mathscr{T}_{h_1}$ (see Fig. 2) three strategies for the mesh refinement. In the first strategy, starting from the initial triangulation $\mathscr{T}_{h_1}$, the finer triangulations are obtained by successively dividing each triangle into four smaller congruent ones, see Fig. 2.

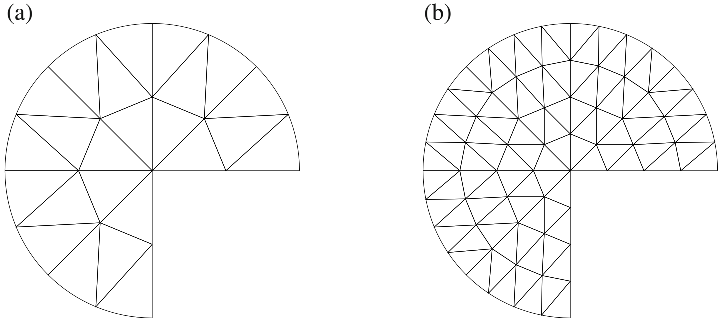


Fig. 2 (a) Initial triangulation $\mathscr{T}_{h_1}$, (b) triangulation $\mathscr{T}_{h_2}$ (uniform refinement).

The second strategy consists of shifting the nodes of the triangulations closer to the singular point at the origin in order to compensate the singularity in the solution, see Fig. 3(a), for $h = h_2$. Here, the grading parameter $\mu = 0.75\lambda = 0.5$ was used (see [26]). The third strategy consists of local graded refinement around the singular

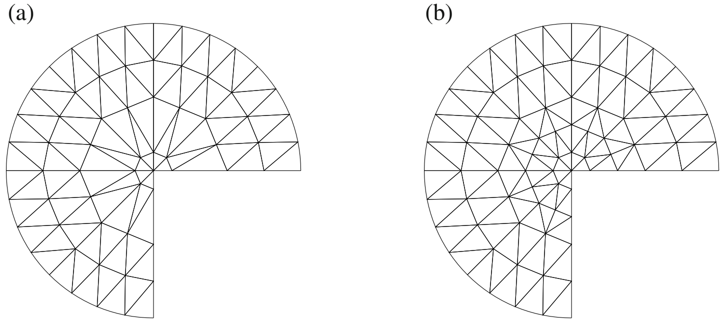


Fig. 3 (a) Triangulation $\mathscr{T}_{h_2}$ with shifted nodes, (b) $\mathscr{T}_{h_2}$ with local graded refinement.

point [22], see Fig. 3(b), for $h = h_2$. Here, we also used the grading parameter $\mu = 0.75\lambda = 0.5$.

In the case of the first two refinement strategies we used a multigrid algorithm to solve the finite element linear system of equations. In the third case the triangulations $\mathscr{T}_{h_q}$, $q = 1, 2, \ldots$, are non-nested. For that reason we did not use a multigrid method for solving the systems of linear finite element equations. We applied a conjugate gradient method with an incomplete Cholesky preconditioner. In all cases we used a high order quadrature to compute the coefficient γ_{1h} and $\gamma_{1h}^{(j)}$ according to (15), (16) as well as (22) with the C^2-smooth cutoff function η defined in (25). Thus, the corresponding errors are negligible in comparison with the principal error $|\gamma_k - \gamma_{kh}^{(j)}|$.

In Table 1 we present the errors $|\gamma - \gamma_{1h}| = |\gamma - \gamma_{1h}^{(1)}|$ and the convergence rates obtained by solving problem (13) and computing $\gamma_{1h}^{(1)}$ by means of formulas (15), (16) (see also (22) with $j = 1$). We compute the rate of convergence α by assuming that

$$|\gamma - \gamma_{1h}^{(1)}| \approx Ch^{\alpha}, \quad \text{that is,} \quad \alpha = \log_2 \frac{|\gamma - \gamma_{1h_i}^{(1)}|}{|\gamma - \gamma_{1h_{i+1}}^{(1)}|}.$$

According Theorem 2 we expect a convergence order $O(h^{2\cdot 2/3}) = O(h^{4/3})$ in the case of uniform mesh refinement. In the case of the other two refinement strategies we expect a convergence order $O(h^2)$ (see Remark 3).

Table 2 contains the errors $|\gamma - \gamma_{1h}^{(2)}|$ and the convergence rates obtained by applying two iterations of the algorithm given in Sect. 3.2. Here, we observe in all cases of mesh refinement the convergence order $O(h^2)$. We see, that we get in the case of uniform mesh refinement an improved convergence order.

In Tables 3–4 we give the errors of the finite element solution and the corresponding convergence rates in the H^1-norm and in Tables 5–6 the errors in the L_2-norm. For the results presented in Tables 3 and 5 the finite element solution is obtained by solving problem (13) and using the three different refinement strategies. In the

Table 1 The computed errors $|\gamma-\gamma_{1h}^{(1)}|$ and rate of convergence α.

		uniform mesh		shifted nodes			graded mesh	
h	#nodes	$\lvert\gamma_1-\gamma_{1h}^{(1)}\rvert$	α	$\lvert\gamma_1-\gamma_{1h}^{(1)}\rvert$	α	#nodes	$\lvert\gamma_1-\gamma_{1h}^{(1)}\rvert$	α
h_2	65	$0.4250\cdot10^{-1}$		$0.7628\cdot10^{-1}$		78	$0.9173\cdot10^{-2}$	
h_3	225	$0.1368\cdot10^{-1}$	1.63	$0.2326\cdot10^{-1}$	1.71	282	$0.2704\cdot10^{-2}$	1.76
h_4	833	$0.5658\cdot10^{-2}$	1.27	$0.5172\cdot10^{-2}$	2.17	1050	$0.9465\cdot10^{-3}$	1.51
h_5	3201	$0.2279\cdot10^{-2}$	1.31	$0.1350\cdot10^{-2}$	1.94	4026	$0.2336\cdot10^{-3}$	2.01
h_6	12545	$0.9108\cdot10^{-3}$	1.32	$0.3462\cdot10^{-3}$	1.96	15738	$0.5819\cdot10^{-4}$	2.00
h_7	49665	$0.3630\cdot10^{-3}$	1.33	$0.8707\cdot10^{-4}$	1.99	62202	$0.1435\cdot10^{-4}$	2.02
h_8	197633	$0.1444\cdot10^{-3}$	1.33	$0.2150\cdot10^{-4}$	2.02	247290	$0.3413\cdot10^{-5}$	2.07
h_9	788481	$0.5741\cdot10^{-4}$	1.33	$0.4966\cdot10^{-5}$	2.11	986106	$0.6648\cdot10^{-6}$	2.36
α_{exp}			4/3		2.00			2.00

Table 2 The computed errors $|\gamma-\gamma_{1h}^{(2)}|$ and rate of convergence α.

		uniform mesh		shifted nodes			graded mesh	
h	#nodes	$\lvert\gamma_1-\gamma_{1h}^{(2)}\rvert$	α	$\lvert\gamma_1-\gamma_{1h}^{(2)}\rvert$	α	#nodes	$\lvert\gamma_1-\gamma_{1h}^{(2)}\rvert$	α
h_2	65	$0.1612\cdot10^{-1}$		$0.5888\cdot10^{-1}$		78	$0.7780\cdot10^{-2}$	
h_3	225	$0.1805\cdot10^{-2}$	3.16	$0.1459\cdot10^{-1}$	2.01	282	$0.1734\cdot10^{-2}$	2.16
h_4	833	$0.5017\cdot10^{-3}$	1.85	$0.2567\cdot10^{-2}$	2.51	1050	$0.6050\cdot10^{-3}$	1.51
h_5	3201	$0.1284\cdot10^{-3}$	1.97	$0.6419\cdot10^{-3}$	2.00	4026	$0.1424\cdot10^{-3}$	2.09
h_6	12545	$0.3211\cdot10^{-4}$	2.00	$0.1611\cdot10^{-3}$	1.99	15738	$0.3463\cdot10^{-4}$	2.04
h_7	49665	$0.7983\cdot10^{-5}$	2.01	$0.3961\cdot10^{-4}$	2.02	62202	$0.8364\cdot10^{-5}$	2.05
h_8	197633	$0.1982\cdot10^{-5}$	2.01	$0.9446\cdot10^{-5}$	2.06	247290	$0.1903\cdot10^{-5}$	2.14
h_9	788481	$0.4861\cdot10^{-6}$	2.03	$0.1925\cdot10^{-5}$	2.29	986106	$0.2855\cdot10^{-6}$	2.73
α_{exp}			2.00		2.00			2.00

case of uniform mesh refinement we expect the convergence order $O(h^{2/3})$ in the H^1-norm and $O(h^{4/3})$ in the L_2-norm. Using the other refinement strategies the expected convergence orders are $O(h)$ in the H^1-norm and $O(h^2)$ in the L_2-norm. Applying two iterations of the algorithm presented in Sect. 3.2, we obtain the errors and convergence rates summarized in Tables 4 and 6. Here, we expect by all refinement strategies the convergence order $O(h)$ in the H^1-norm (see Theorem 3). Again we see that our algorithm leads to an improvement of the convergence order in the case of uniform mesh refinement.

Table 3 The computed errors $\|u-u_h^{(1)}\|_{H^1(\Omega)}$ and rate of convergence α.

	uniform mesh			shifted nodes		graded mesh		
h	#nodes	$\|u-u_{h_i}^{(1)}\|_{H^1(\Omega)}$	α	$\|u-u_{h_i}^{(1)}\|_{H^1(\Omega)}$	α	#nodes	$\|u-u_{h_i}^{(1)}\|_{H^1(\Omega)}$	α
h_2	65	$0.6084\cdot 10^{0}$		$0.5201\cdot 10^{0}$		78	$0.4529\cdot 10^{0}$	
h_3	225	$0.4010\cdot 10^{0}$	0.60	$0.2934\cdot 10^{0}$	0.83	282	$0.2220\cdot 10^{0}$	1.03
h_4	833	$0.2595\cdot 10^{0}$	0.63	$0.1578\cdot 10^{0}$	0.89	1050	$0.1095\cdot 10^{0}$	1.02
h_5	3201	$0.1661\cdot 10^{0}$	0.64	$0.8235\cdot 10^{-1}$	0.94	4026	$0.5434\cdot 10^{-1}$	1.01
h_6	12545	$0.1056\cdot 10^{0}$	0.65	$0.4226\cdot 10^{-1}$	0.96	15738	$0.2707\cdot 10^{-1}$	1.00
h_7	49665	$0.6686\cdot 10^{-1}$	0.66	$0.2147\cdot 10^{-1}$	0.98	62202	$0.1351\cdot 10^{-1}$	1.00
h_8	197633	$0.4210\cdot 10^{-1}$	0.67	$0.1084\cdot 10^{-1}$	0.99	247290	$0.6751\cdot 10^{-2}$	1.00
h_9	788481	$0.2631\cdot 10^{-1}$	0.68	$0.5451\cdot 10^{-2}$	0.99	986106	$0.3374\cdot 10^{-2}$	1.00
α_{exp}			2/3		1.00			1.00

Table 4 The computed errors $\|u-u_h^{(2)}\|_{H^1(\Omega)}$ and rate of convergence α.

	uniform mesh			shifted nodes		graded mesh		
h	#nodes	$\|u-u_{h_i}^{(2)}\|_{H^1(\Omega)}$	α	$\|u-u_{h_i}^{(2)}\|_{H^1(\Omega)}$	α	#nodes	$\|u-u_{h_i}^{(2)}\|_{H^1(\Omega)}$	α
h_2	65	$0.1835\cdot 10^{0}$		$0.2046\cdot 10^{0}$		78	$0.1739\cdot 10^{0}$	
h_3	225	$0.9130\cdot 10^{-1}$	1.01	$0.1037\cdot 10^{0}$	0.98	282	$0.8810\cdot 10^{-1}$	0.98
h_4	833	$0.4555\cdot 10^{-1}$	1.00	$0.5200\cdot 10^{-1}$	0.99	1050	$0.4419\cdot 10^{-1}$	0.99
h_5	3201	$0.2274\cdot 10^{-1}$	1.00	$0.2602\cdot 10^{-1}$	1.00	4026	$0.2211\cdot 10^{-1}$	1.00
h_6	12545	$0.1136\cdot 10^{-1}$	1.00	$0.1301\cdot 10^{-1}$	1.00	15738	$0.1106\cdot 10^{-1}$	1.00
h_7	49665	$0.5680\cdot 10^{-2}$	1.00	$0.6507\cdot 10^{-2}$	1.00	62202	$0.5528\cdot 10^{-2}$	1.00
h_8	197633	$0.2839\cdot 10^{-2}$	1.00	$0.3253\cdot 10^{-2}$	1.00	247290	$0.2764\cdot 10^{-2}$	1.00
h_9	788481	$0.1420\cdot 10^{-2}$	1.00	$0.1626\cdot 10^{-2}$	1.00	986106	$0.1382\cdot 10^{-2}$	1.00
α_{exp}			1.00		1.00			1.00

4.2 Example 2

We consider the Motz's problem, see Fig. 4:

$$-\Delta u = 0 \quad \text{in} \quad \Omega = \{(x,y)\,:\, -1<x<1\,,\, 0<y<1\}$$

Table 5 The computed errors $\|u-u_h^{(1)}\|_{L_2(\Omega)}$ and rate of convergence α.

		uniform mesh		shifted nodes			graded mesh	
h	#nodes	$\|u-u_{h_i}^{(1)}\|_{L_2(\Omega)}$	α	$\|u-u_{h_i}^{(1)}\|_{L_2(\Omega)}$	α	#nodes	$\|u-u_{h_i}^{(1)}\|_{L_2(\Omega)}$	α
h_2	65	$0.5260\cdot 10^{-1}$		$0.4480\cdot 10^{-1}$		78	$0.2715\cdot 10^{-1}$	
h_3	225	$0.2099\cdot 10^{-1}$	1.32	$0.1344\cdot 10^{-1}$	1.74	282	$0.6174\cdot 10^{-2}$	2.14
h_4	833	$0.8305\cdot 10^{-2}$	1.34	$0.3723\cdot 10^{-2}$	1.85	1050	$0.1477\cdot 10^{-2}$	2.06
h_5	3201	$0.3277\cdot 10^{-2}$	1.34	$0.9889\cdot 10^{-3}$	1.91	4026	$0.3614\cdot 10^{-3}$	2.03
h_6	12545	$0.1293\cdot 10^{-2}$	1.34	$0.2564\cdot 10^{-3}$	1.95	15738	$0.8938\cdot 10^{-4}$	2.02
h_7	49665	$0.5105\cdot 10^{-3}$	1.34	$0.6557\cdot 10^{-4}$	1.97	62202	$0.2222\cdot 10^{-4}$	2.00
h_8	197633	$0.2017\cdot 10^{-3}$	1.34	$0.1663\cdot 10^{-4}$	1.98	247290	$0.5542\cdot 10^{-5}$	2.00
h_9	788481	$0.7978\cdot 10^{-4}$	1.34	$0.4193\cdot 10^{-5}$	1.99	986106	$0.1384\cdot 10^{-5}$	2.00
α_{exp}			4/3		2.00			2.00

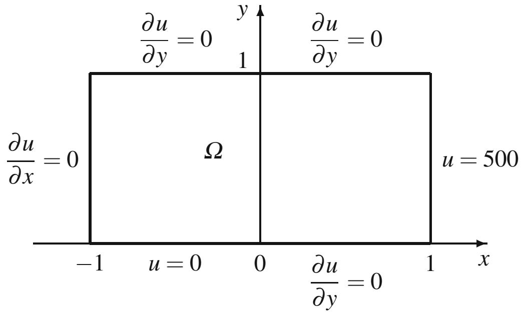
Table 6 The computed errors $\|u-u_h^{(2)}\|_{L_2(\Omega)}$ and rate of convergence α.

		uniform mesh		shifted nodes			graded mesh	
h	#nodes	$\|u-u_{h_i}^{(2)}\|_{L_2(\Omega)}$	α	$\|u-u_{h_i}^{(2)}\|_{L_2(\Omega)}$	α	#nodes	$\|u-u_{h_i}^{(2)}\|_{L_2(\Omega)}$	α
h_2	65	$0.1471\cdot 10^{-1}$		$0.1845\cdot 10^{-1}$		78	$0.1270\cdot 10^{-1}$	
h_3	225	$0.3574\cdot 10^{-2}$	2.04	$0.4856\cdot 10^{-2}$	1.93	282	$0.3076\cdot 10^{-2}$	2.05
h_4	833	$0.8822\cdot 10^{-3}$	2.02	$0.1225\cdot 10^{-2}$	1.99	1050	$0.7521\cdot 10^{-3}$	2.03
h_5	3201	$0.2192\cdot 10^{-3}$	2.00	$0.3066\cdot 10^{-3}$	2.00	4026	$0.1855\cdot 10^{-3}$	2.02
h_6	12545	$0.5467\cdot 10^{-4}$	2.00	$0.7657\cdot 10^{-4}$	2.00	15738	$0.4603\cdot 10^{-4}$	2.01
h_7	49665	$0.1366\cdot 10^{-2}$	2.00	$0.1912\cdot 10^{-4}$	2.00	62202	$0.1146\cdot 10^{-4}$	2.01
h_8	197633	$0.3415\cdot 10^{-5}$	2.00	$0.4778\cdot 10^{-5}$	2.00	247290	$0.2860\cdot 10^{-5}$	2.00
h_9	788481	$0.8540\cdot 10^{-6}$	2.00	$0.1194\cdot 10^{-5}$	2.00	986106	$0.7143\cdot 10^{-6}$	2.00
α_{exp}			2.00		2.00			2.00

with mixed boundary conditions

$$\begin{aligned}
u &= 0\,, && -1 \le x \le 0\,,\, y=0\,,\\
u &= 500\,, && x=1\,,\, 0 \le y \le 1\,,\\
\frac{\partial u}{\partial y} &= 0\,, && 0 < x < 1\,,\, y=0\,,\\
\frac{\partial u}{\partial x} &= 0\,, && x=-1\,,\, 0 < y < 1\,,\\
\frac{\partial u}{\partial y} &= 0\,, && -1 < x < 1\,,\, y=1\,.
\end{aligned}$$

Fig. 4 Motz's Problem.



y
$\frac{\partial u}{\partial y} = 0$
1
$\frac{\partial u}{\partial y} = 0$
$\frac{\partial u}{\partial x} = 0$
Ω
$u = 500$
-1
$u = 0$
0
$\frac{\partial u}{\partial y} = 0$
1
x

The solution u has a singularity near the point $(0,0)$ due to the change in boundary conditions. The solution u admits the expansion

$$u = w + \gamma_1 \eta(r) r^{1/2} \cos\frac{\theta}{2}\,.$$

The exact value of the stress intensity factor γ_1 is not known. However, accurate estimates have been computed, for example, in [24] and the value given is

$$\gamma = 401.162453745234416\,.$$

We consider this value as the exact value in the error estimates.

For the mesh refinement we use the same strategies as presented in the previous section, i.e. in the first case starting from the triangulation $\mathscr{T}_{h_1}$ successively dividing each triangle into four smaller congruent ones, see Fig. 5.

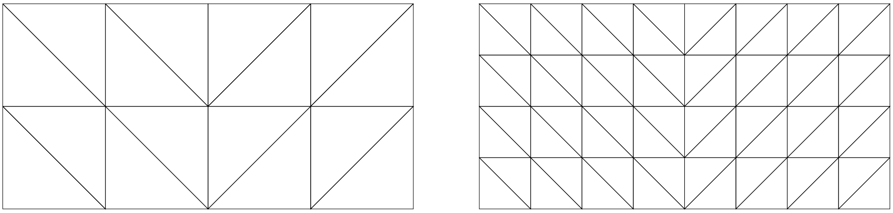


Fig. 5 Initial triangulation $\mathscr{T}_{h_1}$ and triangulation $\mathscr{T}_{h_2}$ (uniform refinement).

In the other two refinement strategies we use the grading parameter $\mu = 0.75\lambda = 0.375$, see Fig. 6 and 7.

For solving the systems of linear finite element equations and computing the coefficient γ_{1h} and $\gamma_{1h}^{(j)}$ according to (15), (16) as well as (22) with the C^2-smooth cutoff function η defined in (25) we used the same methods as in the previous example.

Table 7 shows the absolute errors of $\gamma_{1h} = \gamma_{1h}^{(1)}$ obtained by solving problem (13) and using formulas (15)–(16) in the case of all three mesh refinement strategies.

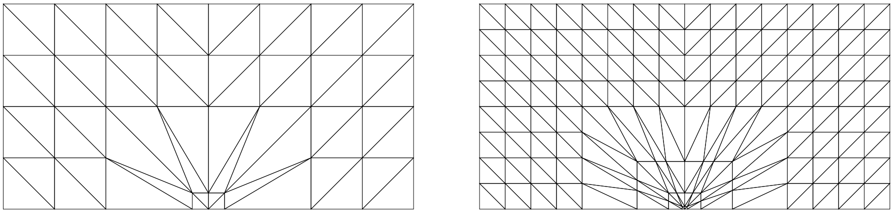


Fig. 6 Triangulations $\mathscr{T}_{h_2}$ and $\mathscr{T}_{h_3}$ with shifted nodes.

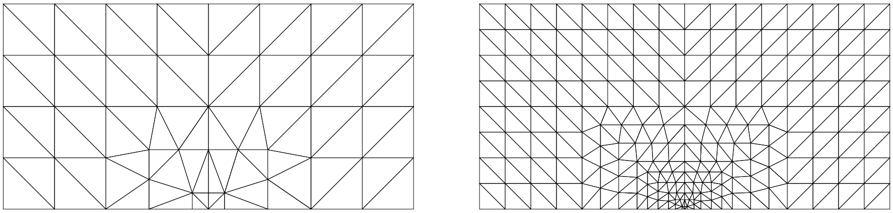


Fig. 7 Triangulations $\mathscr{T}_{h_2}$ and $\mathscr{T}_{h_3}$ with local graded refinement.

In the case of uniform mesh refinement we expect a convergence order $O(h)$ (see Theorem 2 with $\sigma = \frac{1}{2}$) and in the case of the other refinement strategies $O(h^2)$. We observe that in the cases of uniform mesh refinement and the refinement with shifted nodes the rates of convergence in our experiments coincide with the theoretically determined convergence rates. We have no explanation for the somewhat chaotic behavior of the convergence rate in the case of the graded refinement. One reason could be that the triangulations are non-nested.

Table 7 The computed errors $|\gamma - \gamma_{1h}^{(1)}|$ and rate of convergence α.

		uniform mesh		shifted nodes		graded mesh		
h	#nodes	$\lvert\gamma - \gamma_{1h}^{(1)}\rvert$	α	$\lvert\gamma - \gamma_{1h}^{(1)}\rvert$	α	#nodes	$\lvert\gamma - \gamma_{1h}^{(1)}\rvert$	α
h_2	45	$0.1644 \cdot 10^2$		$0.3760 \cdot 10^2$		54	$0.1630 \cdot 10^1$	
h_3	153	$0.5573 \cdot 10^1$	1.56	$0.1309 \cdot 10^2$	1.52	222	$0.1216 \cdot 10^0$	13.4
h_4	561	$0.2644 \cdot 10^1$	1.08	$0.2503 \cdot 10^1$	2.39	813	$0.5904 \cdot 10^{-1}$	1.04
h_5	2145	$0.1283 \cdot 10^1$	1.04	$0.6087 \cdot 10^0$	2.04	3181	$0.1206 \cdot 10^{-1}$	2.23
h_6	8385	$0.6330 \cdot 10^0$	1.02	$0.1523 \cdot 10^0$	2.00	12525	$0.4476 \cdot 10^{-2}$	1.43
h_7	33153	$0.3145 \cdot 10^0$	1.01	$0.3855 \cdot 10^{-1}$	1.98	43932	$0.1556 \cdot 10^{-2}$	1.52
h_8	131841	$0.1567 \cdot 10^0$	1.01	$0.9833 \cdot 10^{-2}$	1.97	167388	$0.4908 \cdot 10^{-3}$	1.66
h_9	525825	$0.7824 \cdot 10^{-1}$	1.00	$0.2580 \cdot 10^{-2}$	1.93	660060	$0.1494 \cdot 10^{-3}$	1.72
α_{exp}			1.00		2.00			2.00

Table 8 contains the absolute errors of $\gamma_{1h}^{(2)}$ obtained by applying two times the steps 2–4 of the algorithm presented in Sect. 3.2. Again, we observe that we get in the case of uniform mesh refinement an improved convergence order. In contrast to the results in Table 7 we see in Table 8 that the absolute errors in the case of uniform mesh refinement are smaller than in the case of the mesh with shifted nodes. The reason is that we have the same convergence orders but in the case of uniform mesh refinement the triangles are better shaped than in the case of the refinement with shifted nodes.

Table 8 The computed errors $|\gamma_1 - \gamma_{1h}^{(2)}|$ and rate of convergence α.

		uniform mesh		shifted nodes			graded mesh	
h	#nodes	$\lvert\gamma_1 - \gamma_{1h}^{(2)}\rvert$	α	$\lvert\gamma_1 - \gamma_{1h}^{(2)}\rvert$	α	#nodes	$\lvert\gamma_1 - \gamma_{1h}^{(2)}\rvert$	α
h_2	45	$0.3549 \cdot 10^{1}$		$0.3665 \cdot 10^{2}$		54	$0.6720 \cdot 10^{-1}$	
h_3	153	$0.3307 \cdot 10^{0}$	3.42	$0.1026 \cdot 10^{2}$	1.84	222	$0.5744 \cdot 10^{0}$	−3.09
h_4	561	$0.9978 \cdot 10^{-1}$	1.73	$0.1657 \cdot 10^{1}$	2.63	813	$0.5286 \cdot 10^{-1}$	3.44
h_5	2145	$0.2408 \cdot 10^{-1}$	2.05	$0.4178 \cdot 10^{0}$	1.99	3181	$0.8674 \cdot 10^{-2}$	2.61
h_6	8385	$0.6075 \cdot 10^{-2}$	1.99	$0.1058 \cdot 10^{0}$	1.98	12525	$0.4989 \cdot 10^{-2}$	0.80
h_7	33153	$0.1525 \cdot 10^{-2}$	1.99	$0.2661 \cdot 10^{-1}$	1.99	43932	$0.5336 \cdot 10^{-3}$	3.22
h_8	131841	$0.3879 \cdot 10^{-3}$	1.98	$0.6765 \cdot 10^{-2}$	1.98	167388	$0.1566 \cdot 10^{-3}$	1.77
h_9	525825	$0.1032 \cdot 10^{-3}$	1.91	$0.1798 \cdot 10^{-2}$	1.91	660060	$0.5418 \cdot 10^{-4}$	1.53
α_{exp}			2.00		2.00			2.00

5 Conclusions

The following observations are evident:

1. Using formula (11) the stress intensity factors γ_k associated with singularities in the solution of boundary value problems for the Laplacian in two-dimensional domains can be approximated in a straight forward way from the finite element solutions u_h according to (15) and (16). The accuracy of the approximation is closely associated to the accuracy of the finite element solution u_h.
2. A new singular function method has been introduced that relies on separately computing the singular part and the regular part of the solution and then adding the two parts to obtain the required solution. Obviously, the new algorithm is different from the traditional dual singular function method as the dual singular solution is not required here. Also, the algorithm is different from that presented in [10, 11].
3. The new algorithm does not show any traces of instability in numerical experiments.

4. It is proved both by theoretical error estimates and by numerical experiments that this algorithm converges optimally and no additional adaptation (graded mesh refinements, shifted nodes, etc.) of the classical finite element strategy is required.
5. As we can observe from the above tables, the number of nodes (this corresponds to the dimension of the linear system to be solved) needed for the new algorithm to yield the same level of error with the graded mesh refinement strategy is far less. Thus, we think that is a good strategy which can be employed in common and industrial calculations.
6. It should finally be noted that for domains with many corners, the formulas for computing the associated stress intensity factors are independent of each other, since they are localized by corresponding cutoff functions. Thus can be computed in parallel.

Acknowledgements. This work was initiated while the author B. Nkemzi was visiting the Department of Computer Science and Applied Mathematics, University of Applied Sciences, Dresden, as an Alexander von Humboldt Fellow in 2008 and 2009. The author expresses sincere thanks to the Alexander von Humboldt Foundation for the fellowship award and to the staff of the Department of Computer Science and Applied Mathematics, University of Applied Sciences, Dresden, for their warm hospitality.

References

[1] Apel, T., Heinrich, B.: Mesh refinement and windowing near edges for some elliptic problem. SIAM J. Numer. Anal. 31(3), 695–708 (1994)

[2] Assous, F., Ciarlet Jr., P., Sonnendrücker, E.: Resolution of the Maxwell equations in domains with reentrant corners. Math. Model. Numer. Anal. 32, 359–389 (1998)

[3] Assous, F., Ciarlet Jr., P., Segré, J.: Numerical solution to the time-dependent Maxwell equations in two-dimensional singular domains: The singular complement method. J. Comput. Phys. 161, 218–249 (2000)

[4] Babuška, I., Kellogg, R.B., Pitkäranta, J.: Direct and inverse error estimates for finite elements with mesh refinements. Numer. Math. 33, 447–471 (1979)

[5] Blum, H.: Numerical treatment of corner and crack singularities. In: Stein, E., Wendland, W.L. (eds.) Finite Element and Boundary Element Techniques from Mathematical and Engineering Point of View. CISM, vol. 301, pp. 172–212. Springer, Wien (1988)

[6] Blum, H., Dobrowolski, M.: On finite element methods for elliptic equations on domains with corners. Computing 28, 53–63 (1982)

[7] Bochniak, M.: The dual singular function method for 2d boundary integral equations. Adv. Comput. Math. 20, 293–310 (2004)

[8] Bourlard, M., Dauge, M., Lubuma, J.M.S., Nicaise, S.: Coefficients of the singularities for elliptic boundary value problems on domains with conical points. III. Finite element methods on polygonal domains. SIAM J. Numer. Anal. 29(11,12), 136–155 (1992)

[9] Brenner, S.C.: Multigrid methods for the computation of singular solutions and stress intensity factors I: Corner singularities. Math. Comput. 226, 559–583 (1999)

[10] Cai, Z., Kim, S.: A finite element method using singular functions for the Poisson equation: Corner singularities. SIAM J. Numer. Anal. 39, 286–299 (2001)

[11] Cai, Z., Kim, S., Shin, B.C.: Solution methods for the Poisson equation with corner singularities: Numerical results. SIAM J. Sci. Comput. 23, 672–682 (2001)

[12] Ciarlet, P.: The finite element method for elliptic problems. North-Holland, Amsterdam (1978)

[13] Dauge, M.: Elliptic boundary value problems on corner domains. Lecture Notes in Mathematics, vol. 1341. Springer, Heidelberg (1988)

[14] Destuynder, P., Djaoua, M.: Estimation de l'erreur sur le coefficient de la singularité de la solution d'un problème elliptic sur un ouvert avec coin. RAIRO Sér Rouge 14, 239–248 (1980)

[15] Dobrowolski, M.: Numerical approximation of elliptic interface and corner problems. Habilitationsschrift, Bonn (1981)

[16] Fix, G., Gulati, S., Wakoff, G.I.: On the use of singular functions with finite element approximations. J. Comp. Phys. 13, 209–238 (1973)

[17] Grisvard, P.: Elliptic problems in nonsmooth domains. Pitman Advanced Publishing Program, Boston (1985)

[18] Grisvard, P.: Singularities in boundary value problems. Springer, Berlin (1992)

[19] Grisvard, P., Wendland, W.L., Whiteman, J.R. (eds.): Singularities and constructive methods of their treatment. Lecture Notes in Mathematics, vol. 1121. Springer, Heidelberg (1985)

[20] Heinrich, B.: The Fourier-finite-element method for Poisson's equation in axisymmetric domains with edges. SIAM J. Numer. Anal. 33, 1885–1911 (1996)

[21] Helsing, J., Jonsson, A.: On the computation of stress fields on polygonal domains with v-notches. Int. J. Numer. Meth. Engng. 53, 433–453 (2002)

[22] Jung, M.: On adaptive grids in multilevel methods. In: Hengst, S. (ed.) GAMM–Seminar on Multigrid–Methods, Gosen, Germany, September 21-25, 1992, pp. 67–80. IAAS, Berlin (1993), Report No. 5

[23] Kozlov, V.A., Maz'ya, V.G., Rossmann, J.: Elliptic boundary value problems in domains with point singularities. Mathematical Surveys and Monographs, vol. 52. American Mathematical Society, Providence (1997)

[24] Lu, T.T., Hu, H.Y., Li, Z.C.: Highly accurate solutions of Motz's and the cracked beam problems. Eng. Anal. Bound. Elem. 28, 1387–1403 (2004)

[25] Moussaoui, M.: Sur l'approximation des solutions de problème de Dirichlet dans un ouvert avec coins. In: Grisvard, P., Wendland, W.L., Whiteman, J.R. (eds.) Singularity and Constructive Methods for their Treatment. Lecture Notes in Mathematics, vol. 1121, pp. 85–103. Springer, Heidelberg (1985)

[26] Oganesyan, L.A., Rukhovets, L.A.: Variational-difference methods for the solution of elliptic equations. Izd. Akad. Nauk Armianskoi SSR, Jerevan (1979) (in Russian)

[27] Raugel, G.: Résolution numérique par une méthode d'éléments finis du problème Dirichlet pour le Laplacien dans un polygone. C. R. Acad. Sci. Paris, Sér. A 286(18), A791–A794 (1978)

[28] Rice, J.R.: A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. 35, 379–386 (1968)

[29] Schatz, A.H., Wahlbin, L.B.: Maximum norm estimates in the finite element method on plane polygonal domains. Part 1. Math. Comput. 32(141), 73–109 (1978)

[30] Schatz, A.H., Wahlbin, L.B.: Maximum norm estimates in the finite element method on plane polygonal domains. Part 2, Refinements. Math. Comput. 33(146), 465–492 (1979)

[31] Strang, G., Fix, G.: An analysis of the finite element method. Prentice-Hall Inc., Englewood Cliffs (1973)

Multilevel Preconditioners for Temporal-Difference Learning Methods Related to Recommendation Engines

Michael Thess

Abstract. In many areas of retail and especially e-business recommendation engines are applied to increase the usability of the store or portal. Advanced recommendation engines use approaches from control theory for adaptive learning. At the forefront of these algorithms reinforcement learning is applied which however requires large transaction numbers to converge. To overcome this problem, we propose a hierarchical approach of reinforcement learning for recommendation engines by combining a multilevel preconditioner with the temporal-difference learning method, the most important algorithm class of reinforcement learning. The multilevel preconditioner works on a combined hierarchy of states and actions. We describe the preconditioner, prove its convergence and present results on real-life data.

1 Introduction

Recommendation engines (REs) have become a major tool to increase customer satisfaction in modern retail systems like e-commerce site, call centers and terminals in supermarkets. Based on the customers transaction behavior they recommend products or more general content. While most REs are still based on the pure analysis of historical transactions to generate recommendations using techniques like basket analysis and clustering [7, 8, 18], there is a trend to adaptive recommendation approaches that learn through communication with customers [1, 4, 5, 11, 17].

Such adaptive approaches are increasingly based on methods of control theory. One of the most promising frameworks is that of reinforcement learning (RL) – a discipline which models adaptive learning using methods of dynamic programming (DP). Probably the most important result of RL was the development of model-free

Michael Thess
prudsys AG, Zwickauer Strasse 16, 09112 Chemnitz, Germany
e-mail: thess@prudsys.com

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 175–195.
springerlink.com

learning methods (i.e. which do not require explicit knowledge of transition probabilities and -rewards of the environment). Here, the class of Temporal-Difference (TD) learning methods proved to be most successful [16].

Although the application of RL to REs seems to look natural, it is thwarted by the enormous complexity of calculations due to the high dimensions and volumes of the real-life data. So most of the early attempts to employ RL for REs [6, 15, 17] are hardly applicable for real-life problems. Especially it turns out that for many practical problems the speed of convergence is too low.

Since the central component of RL, the Bellman equation, can be viewed as a discrete counterpart to the Hamilton-Jacobi-Bellman differential equation, the application of multilevel methods for solving the Bellman equation looks to be very promising. In his pioneering work [19] Omar Ziv studied the application of algebraic multigrid methods to RL. He proposed a multilevel preconditioner for the TD method and was able to prove its convergence. At this, the estimation of the convergence speed is still an open problem but important contributions in this direction have been made by Paprotny [12].

In this paper we are applying Ziv's multilevel preconditioner to TD methods to the area of REs. For most real-live applications the TD method must be applied to the action-value function, instead of the state-value function studied by Ziv. Using a special property of RE formulations, we develop the counterpart of Ziv's preconditioner for action-value functions. We propose some further improvements of the preconditioner for REs. Unlike in the general RL case studied by Ziv which uses coarsening to define the interpolation, recommendation engines usually have access to predefined hierarchies of products which we use to define the interpolation. Finally, we present some computational results on real-life data which indicate the effectiveness of our multilevel approach. At the end, we state the conclusions and propose some ideas of further research.

2 Reinforcement Learning and Recommendation Engines

Reinforcement learning is used amongst other things to control autonomous systems such as robots and also for self-learning games like backgammon or chess. A very promising application area of RL is that of recommendation engines. In this section we present a brief introduction to reinforcement learning and consider its application to REs. For a practical introduction we refer the reader to the standard work ”Reinforcement Learning – An Introduction” by Richard Sutton and Andrew Barton [16], from which some of the illustrations in this section have been taken.

2.1 Reinforcement Learning

RL was based originally on methods of dynamic programming, albeit that in machine learning the theories and terminology have since been developed beyond DP.

Central to this – as is usual in artificial intelligence – is the term agent. Figure 1 shows the interaction between *agent* and environment in reinforcement learning.

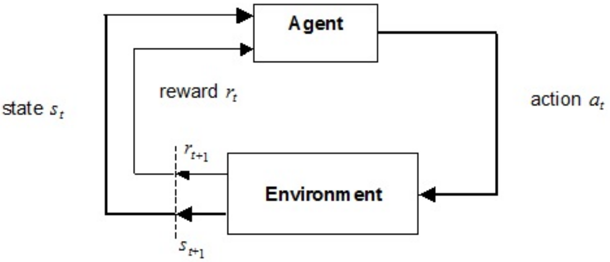


Fig. 1 The interaction between agent and environment in RL.

The agent passes into a new *state* (s), for which it receives a *reward* (r) from the environment, whereupon it decides on a new *action* (a) from the permissible *action set* for s $(A(s))$, by which in most cases it learns, and the environment responds in turn to this action, etc. In such cases we differentiate between *episodic tasks*, which come to an end (as in a game), and *continuing tasks* without any end state (such as a service robot which moves around indefinitely). The goal of the agent consists in selecting the actions in each state so as to maximize the sum of all rewards over the entire episode. The selection of the actions by the agent is referred to as its *policy* π, and that policy which results in maximizing the sum of all rewards is referred to as the *optimal* policy.

Examples

As the first example for RL we can consider a robot, which is required to achieve a goal as quickly as possible. The states are its coordinates, the actions are the selection of the direction of travel and the reward at every step is –1. In order to maximize the sum of rewards over the entire episode, the robot must achieve its goal in the fewest possible steps. A further example is chess, where the positions of the pieces are the states, the moves are the actions and the reward is always 0 except in the final position, at which it is 1 for a win, 0 for a draw and –1 for a loss (this is what we call a *delayed reward*). A final example, to which we will dedicate more intensive study, is recommendation engines. Here for instance the product details views are the states, the recommended products are the actions and the purchases of the products are the rewards.

In order to keep the complexity of determining a good (most nearly optimal) policy within bounds, in most cases it will be assumed that the RL problem satisfies what is called the *Markov* property: In every state the selection of the best action depends only on the current state, and not on its history. A good example of a problem which satisfies the Markov property is once again the game of chess. In order to make the best move in any position, from a mathematical point of view it is totally irrelevant how the position on the board was reached (though when playing the game in practice it is generally helpful).

In case the Markov property is satisfied (*Markov Decision Process – MDP*) RL is based on methods of dynamic programming to solve the Bellman equation. Formally, an MDP is defined as a quadruple $M = (S,A,P,R)$ where S is the set of all states, A is the set of all actions and $A(s)$ the set of all admissible actions in state s, $P = (P^a_{ss'})_{s,s' \in S, a \in A(s)}$ are the transition probabilities from s into s' under action a and $R = (R^a_{ss'})_{s,s' \in S, a \in A(s)}$ the corresponding transition rewards.

We are looking for the optimal policy π. The policy may be stochastic, denoted as $\pi(s,a)$, which implements the stochastic choice of the right action a in the state s or deterministic, denoted as $\pi(s)$, which assigns a unique action to each state: $a = \pi(s)$. A policy $\pi(s,a)$ induces the *Markov chain* (MC) defined as tuple $M_\pi = (S, P^\pi)$ with the state space S and the transition probabilities

$$P^\pi_{ss'} = \sum_{a \in A(s)} \pi(s,a) P^a_{ss'}. \tag{1}$$

For the discrete case the Bellman equation is defined as follows:

$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^\pi(s') \right] \tag{2}$$

where s is a state of S, a an action of $A(s)$ and $\pi(s,a)$ the requested policy. $V^\pi(s)$ is the *state-value function*, which assigns the expected cumulative reward (also called *expected return*) throughout the remainder of the episode to every state s; with the discount rate γ for the weighting of future rewards.

Notice that the Bellman equation has a continuous counterpart, the Hamilton-Jacobi-Bellman (HJB) equation, which we will shortly introduce. For a more detailed discussion of the HJB equation and the relation to other formulations we refer to [9]. Let $s(t)$ be the state and $a(t)$ the action (or control) at time t. The control space is defined by the differential equation

$$\frac{ds(t)}{dt} = g(s(t), a(t))$$

where g is called *state dynamics*. For an initial state s_0 the choice of the action $a(t)$ leads to a unique trajectory $s(t)$. Further, $r(s,a)$ is the reward function. For simplicity, we consider the deterministic policy $\pi(s)$ which assigns a unique action to each state in t: $a(t) = \pi(s(t))$. Then the HJB equation is defined as

$$V^\pi(s) ln\gamma + \nabla V^\pi(s) \cdot g(s, \pi(s)) + r(s, \pi(s)) = 0.$$

Beside the state-value function in RL often the *action-value function* $Q^\pi(s,a)$ is used. It assigns the expected return to every state s and its permissible actions a. The following relations between state- and action-value functions apply:

$$V^{\pi}(s) = \sum_{a} \pi(s,a) Q^{\pi}(s,a), \quad Q^{\pi}(s,a) = \sum_{s'} P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma V^{\pi}(s') \right].$$

The state-value function and action-value function are thus related, and can be converted from one into the other (provided the model of the environment $P^{a}_{ss'}$ and $R^{a}_{ss'}$ is known).

Their combination along the action-value function results in the formulation (2) of the Bellman equation. Conversely, if we substitute the state-value function we obtain the formulation of the Bellman equation for the action-value function:

$$Q^{\pi}(s,a) = \sum_{s'} P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma \sum_{a'} \pi(s',a') Q^{\pi}(s',a') \right]. \tag{3}$$

The solution methods of dynamic programming – especially the policy iteration – provide the basic framework in order to find the requested policy $\pi(s,a)$ in reinforcement learning. This can either be done by solving the Bellman equation directly, supposed the transition probabilities and rewards are known. In this case, which is called *model-based*, the established solution methods of DP can be applied.

For the *model-free case* – the actual reinforcement learning – there exist different methods like Monte Carlo or Temporal-Difference Learning (TD) for the indirect solution of the Bellman equation and the determination of the optimal policy. Notice that for the model-free case usually the action-value function is used because in that case the calculation of the state-value function is not sufficient to determine the optimal policy.

2.2 Solution of the Bellman Equation

The Bellman equation (2) can be written compactly in vector notation as

$$\mathbf{v}^{\pi} = \mathbf{r}^{\pi} + \gamma \mathbf{P}^{\pi} \mathbf{v}^{\pi}, \tag{4}$$

where the vectors of the state values $\mathbf{v}^{\pi}$ and rewards $\mathbf{r}^{\pi}$ together with the matrix of the transition probabilities $\mathbf{P}^{\pi}$ are defined as follows:

$$\begin{aligned} \mathbf{v}^{\pi} &= \left[V^{\pi}(s_1), V^{\pi}(s_2), \ldots, V^{\pi}(s_N) \right]^{\top}, \\ \mathbf{r}^{\pi} &= \left[r^{\pi}(s_1), r^{\pi}(s_2), \ldots, r^{\pi}(s_N) \right]^{\top}, \\ r^{\pi}(s) &= \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} R^{a}_{ss'}, \\ \mathbf{P}^{\pi}_{s,s'}(s) &= \sum_{a} \pi(s,a) P^{a}_{ss'}. \end{aligned} \tag{5}$$

Similarly the Bellman equation for action-value functions (3) can be reformulated as

$$\mathbf{w}^{\pi} = \hat{\mathbf{r}}^{\pi} + \gamma \hat{\mathbf{P}}^{\pi} \mathbf{w}^{\pi}, \tag{6}$$

where the vectors of the action values $\mathbf{w}^{\pi}$ and rewards $\hat{\mathbf{r}}^{\pi}$ together with the matrix of the transition probabilities $\hat{\mathbf{P}}^{\pi}$ are defined as follows:

$$\begin{aligned} \mathbf{w}^{\pi} &= \left[Q^{\pi}(s_1,a_1)Q^{\pi}(s_1,a_2)\ldots Q^{\pi}(s_N,a_M)\right]^{\top}, \qquad (7)\\ \hat{\mathbf{r}}^{\pi} &= \left[\hat{r}^{\pi}(s_1,a_1)\hat{r}^{\pi}(s_1,a_2)\ldots \hat{r}^{\pi}(s_N,a_M)\right]^{\top}, \\ \hat{r}^{\pi}(s,a) &= \sum_{s'} \pi(s,a)P^{a}_{ss'}, \\ \hat{\mathbf{P}}^{\pi}_{s,a,s',a'}(s) &= P^{a}_{ss'}\pi(s',a'). \end{aligned}$$

We now define the *Bellman operator* T^{π} for a policy π as

$$T^{\pi}(\mathbf{v}) = \mathbf{r}^{\pi} + \gamma \mathbf{P}^{\pi}\mathbf{v}$$

and consider the iteration

$$\mathbf{v}^{k+1} = T^{\pi}(\mathbf{v}^{k}). \qquad (8)$$

The following proposition can be easily shown [12].

Proposition 1. *For all MDPs, each policy has a unique and finite state-value function, which is the unique fixed point of the iteration defined in (8). The iteration converges to its fixed point at asymptotic rate of γ for any initial guess. The asymptotic rate is attained in l_{∞}.*

We now turn our attention to the existence and uniqueness of an optimal policy π^{*}, i.e. a policy fulfilling

$$\mathbf{v}^{\pi^{*}} \leq \mathbf{v}^{\pi} \quad \forall \pi \in \Pi_{\mathrm{M}}$$

where Π_{M} is the space of all policies of the Markov Decision Process M. In order to find the best policy π^{*} we consider the *policy iteration* method well known in dynamic programming.

The policy iteration method works as follows: Starting with an initial policy π_0, in each step $0,\ldots,n$ for the current policy π_i we solve the Bellman equation (4) obtaining the state-value function V^{i} (Proposition 1). For V^{i} we calculate the greedy policy

$$\pi_{i+1} = \arg \max_{\mathrm{a}\in \mathrm{A(s)}} Q^{i}(s,a) = \arg \max_{\mathrm{a}\in \mathrm{A(s)}} \sum_{s'} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma V^{i}(s')\right],$$

i.e. the policy π_{i+1} which maximizes the expected return in each state s. For π_{i+1} in turn we calculate the new state-value function V^{i+1}. Thus, we obtain a sequence of policies and state-value functions

$$\pi_0 \to V^{0} \to \pi_1 \to V^{1} \to \pi_2 \to V^{2} \to \ldots$$

The following theorem provides the desired result.

Theorem 1. *The policy iteration terminates after a finite number of iterations with the tuple $(\pi^{*}, \mathbf{v}^{\pi})$.*

2.3 Temporal-Difference Learning

In many applications, the transition probabilities and rewards are not known or their determination is too complex. In these cases model-free RL methods can be applied. The model-free approach is based on learning by iterative adaptation of the action-value function $Q(s,a)$. We begin with the simple $TD(0)$ method. At every step t of the episode the update is performed as follows:

$$Q(s_t,a_t) := Q(s_t,a_t) + \alpha_t\left(r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right), \tag{9}$$

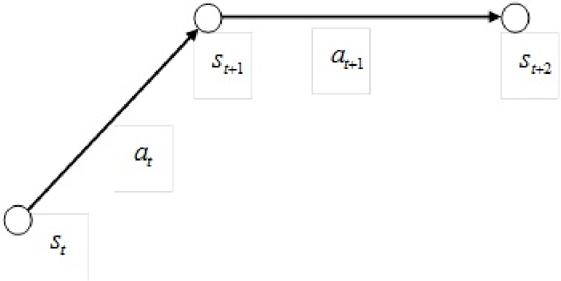


Fig. 2 A sequence of an episode of 2 steps.

where α_t is the step-size parameter at step t. The higher it is, the more quickly the algorithm learns. For the current update d

$$d(s_t,a_t,s_{t+1},a_{t+1}) = r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t) \tag{10}$$

(9) takes the following form:

$$Q(s_t,a_t) := Q(s_t,a_t) + \alpha_t d(s_t,a_t,s_{t+1},a_{t+1}). \tag{11}$$

This means that we calculate the new estimation $\tilde{Q}(s_t,a_t) := r_{t+1} + \gamma Q(s_{t+1},a_{t+1})$ and subtract from it the previous $Q(s_t,a_t)$. If $\tilde{Q}(s_t,a_t)$ is greater than $Q(s_t,a_t)$, then the latter is increased in accordance with (10); if $\tilde{Q}(s_t,a_t)$ is less than $Q(s_t,a_t)$, then the latter is decreased in accordance with (10).

So what does $\tilde{Q}(s_t,a_t) := r_{t+1} + \gamma Q(s_{t+1},a_{t+1})$ mean? We know that $Q(s_t,a_t)$ is the expected return taken across the remainder of the episode. The first term r_{t+1} is the direct reward of the action a_t. The second term $\gamma Q(s_{t+1},a_{t+1})$ is the expected return from the new state s_{t+1}. It follows that there are two possibilities for the reason why $\tilde{Q}(s_t,a_t)$ may be higher than $Q(s_t,a_t)$: either the direct reward r_{t+1} is high or the action a_t has led to a valuable state s_{t+1} with a high action value $Q(s_{t+1}, a_{t+1})$ (or both).

At every step t the $TD(0)$ algorithm that was described modifies the action-value function only in the current state s_t. Since however the action values of the following states are also included in its calculation of the discount, conversely the action values of all preceding states can also be updated at each step, which significantly increases the speed of learning. This is achieved using algorithms of the $TD(\lambda)$ family. This means, at every step all action values are updated as follows:

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha_t d(s_t, a_t, s_{t+1}, a_{t+1}) z_t(s, a), \tag{12}$$

where the weighting function $z_t(s,a)$ describes the relevance of the temporary difference at the time point t for the state-action pair (s,a). The weighting function $z_t(s,a)$ is called *eligibility traces* and awards recently visited states a higher weighting than states which have not been visited for some time. The usual definition of eligibility traces is

$$z_t(s,a) = \begin{cases} \gamma\lambda z_{t-1}(s,a) + 1, & \text{if } (s,a) = (s_t, a_t) \\ \gamma\lambda z_{t-1}(s,a), & \text{else} \end{cases}. \tag{13}$$

For this, $z_t(s,a)$ is initialized as zero for all states. Obviously $z_t(s,a)$ is relatively high if (s,a) is visited often and (s_t,a_t) can be reached from (s,a) in only a few steps. The weighting $z_t(s,a)$ reduces exponentially with the number of steps since the last visit of the time step (s,a).

For calculations, $z_t(s,a)$ can be set to zero if it falls below an epsilon barrier, so that it is given a local support and thus can be implemented asymptotically optimal with respect to computation time and memory. In practice that means that after each update (11) is performed for the current time step t, that is for (s_t, a_t), the update (12) is performed for all preceding time steps in the current episode $t-1, t-2, \ldots, t-m$, where $t-m$ is the last preceding time step where $z_t(s,a)$ lies above the epsilon barrier. At the same time $z_t(s,a)$ must be updated in accordance with (13) for all affected time steps.

In this way the $TD(\lambda)$ algorithm described above performs a continual adaptation of the action-value function $Q(s,a)$ in an intuitive fashion. It can be shown that under some unrestrictive assumptions the $TD(\lambda)$ algorithm converges to the optimal policy π^* and action-value function Q^*.

2.4 Application to Recommendation Engines

An effective approach to using reinforcement learning for recommendation engines is described below. In the simplest case the product detail views form the states, the recommended products the actions and the rewards are the clicks or purchases of the products. The goal consists (depending on the chosen reward) in maximizing the activity (clicks) or the success (sales). In most cases the success is used for optimization.

Fig. 3 illustrates the use of RL for product recommendations in a web shop and shows the interaction between the recommendation engine and the user. Here the optimal proven recommendations are marked with "∗".

In the first and third steps the recommendation engine is following the proven recommendations (exploitation); in the second step a new recommendation is tried out (exploration). The user ignores the first recommendation, but accepts the second and third. The feedback arrows symbolize the updating of the recommendations.

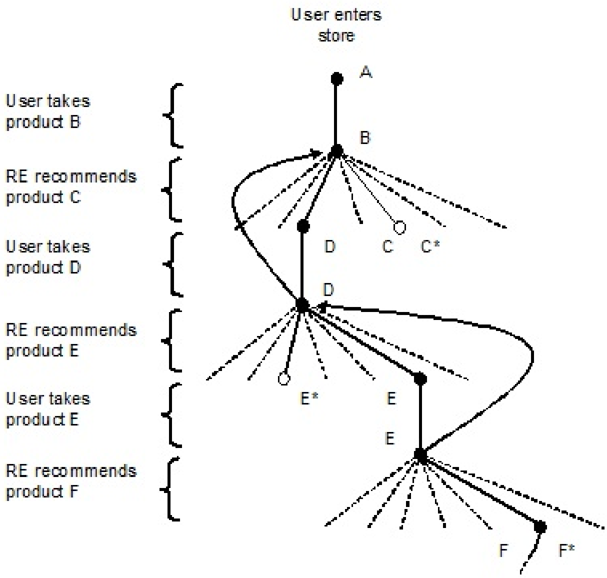


Fig. 3 Example of reinforcement learning for product recommendations in a web shop.

Although this modelling may appear self-evident, it nevertheless represents a highly complex task. Firstly: web shops generally offer very many products; as a rule between a few thousand up to a few million (for instance at a bookshop). Many of these products have virtually no transaction history, i.e. they have scarcely ever been bought, indeed some have never even been clicked on. Furthermore, the existing transactions are mostly clicks, whereas on the other hand placements in the shopping basket (SB) and purchases are far more infrequent. However, as we have stated maximizing sales is the primary goal of REs. Let us summarize these two problems again:

1. High product numbers, a majority of which have minimal transaction history,
2. The vast majority of transactions are clicks, only a fraction are placements in the shopping basket and purchases.

We know however from the theory and practice of RL that high transaction numbers are necessary in order to achieve convergence. The above problems therefore appear to be killer arguments against the direct use of RL for REs. Therefore additional empirical assumptions are usually made, which reduce the complexity of using RL for REs.

We continue with the description of modelling the RE by RL. As described initially, each product view represents a state s and a recommendation of another product represents an action a. Each web session (session for short) forms an episode. The result is that the interaction between user and recommendation engine in each web session can be considered as a sequence of product transitions under the influence of recommendations:

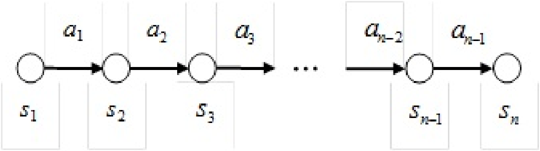


Fig. 4 Sequence of products and recommendations as states and actions.

This permits us to model the most important statistical characteristics such as action values, transition probabilities and rewards using rules $s \to s'$, which can be saved for instance in files or database tables (we will explain the details of this later). Of course not every action will necessarily lead to an accepted recommendation: the user can also ignore the recommendations and go to an entirely unrelated product. However in this case the product transition is added as a new rule to the rule base and thus provides a new potential action.

Since all actions a represent product views, the sets of all states $\mathbf{S}$ and $\mathbf{A}$ are isomorphic:

$$\mathbf{S} \cong \mathbf{A}. \tag{14}$$

It should be noted that for reasons of complexity not all actions $\mathbf{A}$ are considered for each state s of course. Only a proportion $A(s)$, which initially contains all product transitions that have actually occurred, together with actions derived by other means such as hierarchies. This means the action set $A(s)$ expands dynamically in the course of the learning process.

At each step the RE receives a reward r. The sum of all the rewards should be maximized over the complete session. The reward is defined for each step as follows:

if a product s is placed in the shopping basket or is bought, the preceding action (i.e. the “recommendation” a, which led to the product) receives the value of the product s (price, revenue, etc.) as its reward, otherwise it receives a small click reward, close to 0. This r reflects the primary goal of seeking to maximize the shopping basket values or the sales/revenue. Note that orders constitute a delayed reward, since in most cases they appear only at the end of a session. The definition of the correct reward is linked to various refinements which need not be further explored here.

We now introduce the Markov property for our recommendation approach: In every state s the optimal action a, i.e. the best recommendation, depends solely on the current state s, i.e. the product under consideration.

Of course this Markov property for REs is satisfied only incompletely, since the best recommendation also depends on the preceding states of s together with their transactions. Nevertheless, for the evaluation of a recommendation by the user the product currently viewed plays the principal role, so the assumption can be considered to be reasonable. (There is also compelling empirical evidence on this point, namely classic cross-selling, which is described using precisely this form of rules, and whose effectiveness is beyond doubt.)

However, even with the Markov property in place, the complexity to calculate and store the action-value function (and for the model case the transition probabilities and rewards) remains enormous. Thus, in [13] we introduce further *empirical* assumptions which break down the complexity to level that allows the practical handling of real-life RE applications. Moreover, based on adaptive tensor factorizations [13] describes a way to finally skip the simple Markov property (which is replaced by a more general one). For the rest of the paper we assume that we are able to work with the action-value function for the RE case.

Nevertheless, even with all that modifications in place, for big real-life RE applications sometimes the learning of the TD methods turns out to be slow. Thus, we are looking for a way to increase the convergence speed. A natural way to solve this problem is the use of multilevel preconditioners.

3 A Multilevel Preconditioner for the TD Method

Let us define (leaving out the policy notation)

$$\mathbf{A} = \mathbf{I} - \gamma \mathbf{P}^{\pi} \tag{15}$$

$$\mathbf{b} = \mathbf{r}^{\pi} \tag{16}$$

so that equation (4) takes the following form:

$$\mathbf{A}\mathbf{v} = \mathbf{b}. \tag{17}$$

In [19], Ziv investigated the use of classic iteration methods such as the Richardson method for the solution of (17) with the operator (15) and the right-hand side (16).

The transition probability matrix $\mathbf{P}^{\pi}$ is a row-stochastic matrix which means that it is non-negative and each row sums up to one. Hence, the spectral radius of $\mathbf{P}^{\pi}$ is one. Based on the structure of (15) and using the eigenvalue properties of $\mathbf{P}^{\pi}$ it was shown that algebraically smooth error vectors do arise [19]. Hence, multi-level methods are outstandingly suitable for the solution of these problems. Of course everything we have described carries over to the case of action-value functions (6).

Furthermore in [19] and [12] the use of multi-level methods is investigated not only for the problems of dynamic programming (i.e. model-based) but also for model-free methods, in particular for temporal-difference learning. For the latter, two multi-level methods are proposed, the first a multiplicative variant and the second an additive variant. The multiplicative variant is less convincing and it cannot be fully proven to converge.

The additive variant is interesting. It will bear further explanation. In contrast to the above papers, we will consider the action-value function rather than the state-value function.

For this we consider a hierarchy with the levels $0,\ldots,L$, where 0 is the finest grid which has the current state-action values from $(\mathbf{S},\mathbf{A})$ as grid points and L the coarsest grid.

An interpolation $\mathbf{I}_{l+1}^{l}$ of level $l+1$ is given at the level l. For the restriction $\mathbf{I}_{l}^{l+1}$ of level $l+1$ to the level l the transpose or pseudo-inverse of $\mathbf{I}_{l+1}^{l}$ can be used. Furthermore, $\mathbf{I}_{l}^{m}$ is defined as a level l-to-m interpolation

$$\mathbf{I}_{l}^{m} = \mathbf{I}_{m-1}^{m}\mathbf{I}_{m-2}^{m-1}\cdots\mathbf{I}_{l}^{l-1} \tag{18}$$

and $\mathbf{I}_{l}^{l}$ is defined as the unit matrix. Let $\mathbf{C}_{t}^{-1}$ be a preconditioner given by

$$\mathbf{C}_{t}^{-1} = \sum_{l=0}^{L} \beta_{l,t}\mathbf{I}_{l}^{0}\mathbf{I}_{0}^{l}. \tag{19}$$

The scaling factors $\beta_{l,t}$ can be chosen as inverse diagonal of the system matrix on the associated level. Notice that $\mathbf{C}_{t}^{-1}$ can be viewed as an algebraic counterpart of the well-known BPX preconditioner [3].

Now we rewrite the $TD(\lambda)$ method (12) into vector notation:

$$\mathbf{w} := \mathbf{w} + \alpha_t \mathbf{z}_t d \tag{20}$$

where $\mathbf{w}$ in accordance with (7) represents the vector of the action values $Q(s,a)$ and

$$\mathbf{z}_t = \left[z_t(s_1,a_1), z_t(s_1,a_2), \ldots, z_t(s_N,a_M)\right]^T.$$

Then the preconditioned $TD(\lambda)$ method

$$\mathbf{w} := \mathbf{w} + \alpha_t \mathbf{C}_{t}^{-1}\mathbf{z}_t d \tag{21}$$

also converges to the same solution as the $TD(\lambda)$ method (20).

The proof of convergence is based essentially on the following proposition from [19], which is further generalized in [12].

Proposition 2. *Let the convergence preconditions for $TD(\lambda)$ be satisfied. Also let $\mathbf{B}^{-1}$ be a symmetric positive definite $N \times N$ matrix. Then the preconditioned $TD(\lambda)$ method*

$$\mathbf{w} := \mathbf{w} + \alpha_t \mathbf{B}^{-1} \mathbf{z}_t d \tag{22}$$

also converges.

Since $\mathbf{C}_t^{-1}$ is a symmetric and positive definite $N \times N$ matrix, the hierarchically preconditioned $TD(\lambda)$ method is convergent.

In [19] and [12] the construction of the interpolation $\mathbf{I}_{l+1}^{l}$ was considered on an algebraic basis in the course of the algebraic multigrid methods (AMG). In particular the *state aggregation* was considered according to [2]. For this purpose the state space $\mathbf{S}$ is split into K disjoint groups $G_k, k = 1, \dots, K$. The definition of the interpolation is

$$\mathbf{I}_{l+1}^{l} = \begin{cases} 1, \text{ if } i \in G_k, \\ 0, \text{ else,} \end{cases} \tag{23}$$

and the restriction is defined as its pseudo-inverse according to Moore-Penrose:

$$\left(\mathbf{I}_{l}^{l+1}\right)_{ik} = \left[\left(\mathbf{I}_{l+1}^{l}\right)^{T} \mathbf{I}_{l+1}^{l}\right]^{-1} \left(\mathbf{I}_{l+1}^{l}\right)^{T}. \tag{24}$$

The specification of the aggregate groups G_k was performed using AMG methods in combination with RL-specific extensions.

In contrast to this general algebraic case, in most cases product master data are available for recommendation engines, including their assignment to categories. This should be used for the construction of the hierarchies, since it contains important additional information. Thus there are two sources for specification of hierarchies for REs:

1. product hierarchies such as shop taxonomy or product groups,
2. product attributes such as manufacturer, brand or color.

Both data sources are in most practical applications easily accessible from the product master data. Usually there exists a product table (as text file or in a database) whose rows correspond to the different products and the columns to their attributes. In case of the shop taxonomy, there are additional tables which map the products to their parent categories and define the hierarchy of the categories.

The use of 1. seems obvious, however in most cases it requires comprehensive pre-processing so that the pseudo-inverse (24) can be calculated. In the case of 2. this is automatically guaranteed, since every product can be assigned only to one parent state. Notice that only two-grid hierarchies can be constructed from 2., but in most cases this is sufficient.

The preconditioned $TD(\lambda)$ algorithm (22) however works on action values. Therefore the interpolation and restriction operators are required not only for states but also for state-action pairs (s,a). How then can hierarchies for actions be defined? The approach is based on an idea from Alexander Paprotny [13]. Since for the RE approach the spaces **S** and **A** are isomorphic (14), actions can be dealt with just like states and the same interpolations and restrictions used for them.

So whilst the states **S** correspond to the products which obtain recommendations, the actions **A** correspond to the recommended products. Similarly to (23) which operates on the state space, this gives the following definition of the interpolation $\hat{\mathbf{I}}^{l}_{l+1}$ for the state-action space:

$$\left(\hat{\mathbf{I}}^{l}_{l+1}\right)_{ijkp} = \begin{cases} 1, \text{ if } i \in G_k \wedge j \in H_p(i), \\ 0, \text{ else.} \end{cases} \tag{25}$$

Finally we could make a modification to the interpolation $\hat{\mathbf{I}}^{l}_{l+1}$ in order to improve it. This weighted interpolation operator $\bar{\hat{\mathbf{I}}}^{l}_{l+1}$ is defined as follows:

$$\left(\bar{\hat{\mathbf{I}}}^{l}_{l+1}\right)_{ijkp} = \begin{cases} |A(s_i)||A(a_j)|, & \text{if } i \in G_k \wedge j \in H_p(i), \\ 0, & \text{else.} \end{cases} \tag{26}$$

For this $|A(s_i)|$ is the number of all actions in s_i, i.e. all the rules for the associated product as premise and $|A(a_j)|$ is the number of all actions in the state associated with a_j, i.e. the rules with the associated product as conclusion. This weighted interpolation thus prefers rules with "stronger" premise or conclusion products.

Generally, several hierarchies can be derived from the product master data of the RE, for instance by shop hierarchies, product groups and product attributes. Consequently a corresponding preconditioner $\hat{\mathbf{C}}_i^{-1}$ can be derived for every hierarchy in accordance with (22). This leads to the question of whether this too can be used in a combined manner. In fact this is possible, for instance using the following additive preconditioner $\hat{\mathbf{C}}_a^{-1}$:

$$\hat{\mathbf{C}}_a^{-1} = \hat{\mathbf{C}}_1^{-1} + \hat{\mathbf{C}}_2^{-1} + \ldots + \hat{\mathbf{C}}_n^{-1}, \tag{27}$$

where n is the number of all hierarchies that are used. Since all preconditioners $\hat{\mathbf{C}}_i^{-1}$ are symmetric positive definite, this is also $\hat{\mathbf{C}}_a^{-1}$ and from Proposition 2 the convergence of the $TD(\lambda)$ method preconditioned with it also follows.

Description Based on an Example

The use of the hierarchical RL can be illustrated using a simple example. For this let us consider a RE which delivers recommendations for just 3 products. We now obtain a level hierarchy with 3 nodes on the fine grid and 2 nodes on the coarse grid (Fig. 5).

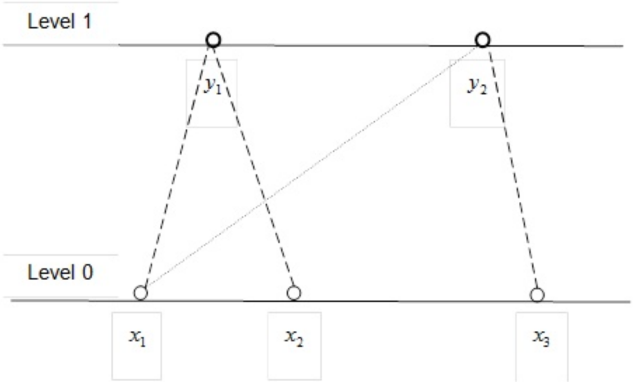


Fig. 5 Interpolation operator for state aggregations. The dotted line would remove the uniqueness of the assignment and is not permitted.

The states are designated on the fine grid as x and on the coarse grid as y. This gives us the interpolation $\mathbf{I}_1^0$ and restriction $\mathbf{I}_0^1$ as follows:

$$\mathbf{I}_1^0\mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \mathbf{I}_0^1\mathbf{x} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

This gives the following preconditioner for the state-value function:

$$\mathbf{C}^{-1}\mathbf{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} 1+\frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 1+\frac{1}{2} & 0 \\ 0 & 0 & 1+1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

For sake of simplicity all scaling factors $\beta_{l,t}$ were set to a constant 1.

The example of the state aggregations should now be extended to include the associated actions, where initially all states are permissible as actions in all states. The result is shown in Fig. 6.

For ease of reading the actions are shown as x at level 0 and y at level 1, where the lower index represents the start nodes and the upper index the target nodes. For instance x_1^2 is the recommendation of product 2 for product 1 at level 0.

Note that on the finest grid – i.e. the product level – the reflexive relation x_i^i is practically meaningless, since a product cannot recommend itself. At levels > 0 these actions are meaningful however, since they are a measure of the strength of product recommendations within the same group relative to one another.

The following interpolation and restriction matrix applies to the example under consideration:

$$\hat{\mathbf{I}}_1^0\hat{\mathbf{y}} = \begin{bmatrix} \begin{bmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{bmatrix} \\ \begin{bmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{bmatrix} \\ \begin{bmatrix} x_3^1 \\ x_3^2 \\ x_3^3 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} & 0 \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} y_1^1 \\ y_1^2 \end{bmatrix} \\ \begin{bmatrix} y_2^1 \\ y_2^2 \end{bmatrix} \end{bmatrix},$$

$$\hat{\mathbf{I}}_0^1\hat{\mathbf{x}} = \begin{bmatrix} \begin{bmatrix} y_1^1 \\ y_1^2 \end{bmatrix} \\ \begin{bmatrix} y_2^1 \\ y_2^2 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} & \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} & 0 \\ 0 & 0 & \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{bmatrix} \\ \begin{bmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{bmatrix} \\ \begin{bmatrix} x_3^1 \\ x_3^2 \\ x_3^3 \end{bmatrix} \end{bmatrix}$$

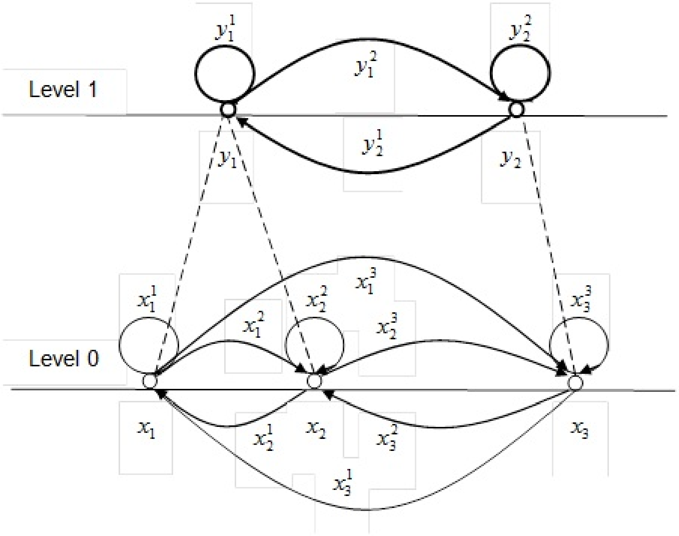


Fig. 6 Interpolation operator for state-action aggregations.

where the preconditioner $\hat{\mathbf{C}}^{-1}$ is derived as follows:

$$\hat{\mathbf{C}}^{-1}\hat{\mathbf{x}} = \begin{bmatrix} \begin{bmatrix} \tilde{x}_1^1 \\ \tilde{x}_1^2 \\ \tilde{x}_1^3 \end{bmatrix} \\ \begin{bmatrix} \tilde{x}_2^1 \\ \tilde{x}_2^2 \\ \tilde{x}_2^3 \end{bmatrix} \\ \begin{bmatrix} \tilde{x}_3^1 \\ \tilde{x}_3^2 \\ \tilde{x}_3^3 \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{bmatrix} 1+\frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & 1+\frac{1}{4} & 0 \\ 0 & 0 & 1+\frac{1}{2} \end{bmatrix} & \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} & 0 \\ \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} & \begin{bmatrix} 1+\frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & 1+\frac{1}{4} & 0 \\ 0 & 0 & 1+\frac{1}{2} \end{bmatrix} & 0 \\ 0 & 0 & \begin{bmatrix} 1+\frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 1+\frac{1}{2} & 0 \\ 0 & 0 & 1+1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \end{bmatrix} \\ \begin{bmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \end{bmatrix} \\ \begin{bmatrix} x_3^1 \\ x_3^2 \\ x_3^3 \end{bmatrix} \end{bmatrix}$$

We should now take a brief look at the method of operation of the preconditioner. Thus an update via the action x_1^2 leads to an update of the actions x_1^2 themselves and also x_2^1 in accordance with

$$\tilde{x}_1^2 = \left(1+\frac{1}{4}\right)x_1^2,\ \tilde{x}_2^1 = \frac{1}{4}x_1^2.$$

This results from the reflexive coarse grid action y_1^1 of the group y_1 on itself. Of course the reflexive update for x_1^2 is especially strong here.

An update via the action x_1^3 leads to an update of the actions x_1^3 themselves and also x_2^3:

$$\tilde{x}_1^3 = \left(1+\frac{1}{2}\right)x_1^3,\ \tilde{x}_2^3 = \frac{1}{2}x_1^3.$$

It is the coarse grid action y_1^2 of the group y_1 on y_2 that is responsible for this. Figure 7 illustrates the general logic of the updates using the example of the action (s_0, a_0).

An update of the rule (s_0, a_0) therefore leads not only to the update of the rule itself but also to the update of all rules in the same state group G of the initial product s_0 into the same action group H of the recommended product a_0.

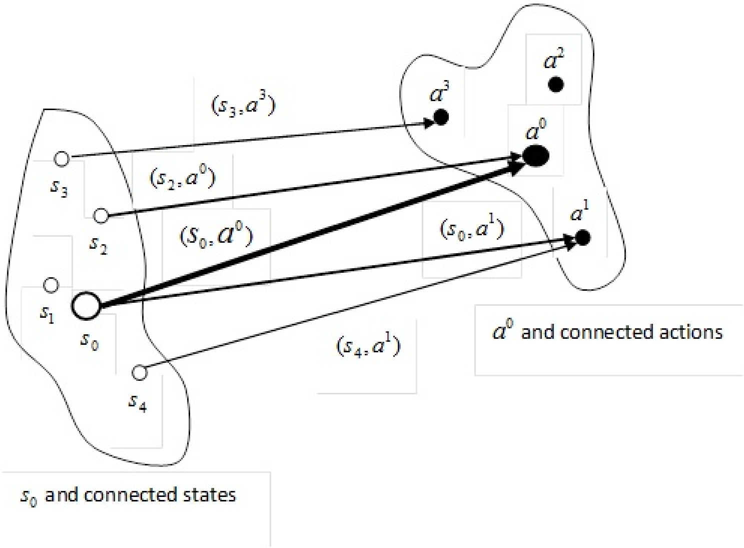


Fig. 7 Illustration of the update logic of the preconditioner $\hat{\mathbf{C}}^{-1}$.

4 Evaluation on Real-Life Data

In the following we present some first results on a real-life data set. We are using the following method for evaluating the prediction quality of an adaptive learning algorithm.

The recommendation engine prudsys RDE stores all relevant transactions of a web shop into daily transaction files. The transaction files contain the time, session ID, product ID, transaction type (product clicked, added to basket, purchased and purchase price), and additional data of all transactions (user ID, channel ID, recommendations actually delivered, control group assignment, etc.). Table 1 summarizes the main columns of a transaction file.

Table 1 Description of the main columns of a transaction file of the prudsys RDE.

Column name	Description
date-time	Date and time of transaction
transactID	transaction ID, i.e. the sessionID
itemID	product ID of the transaction
itemPrice	net unit price of product
transactType	transaction type (0 – clicked, 1 – added to basket, 2 – ordered)

An example of a row of a transaction file is

```
2009-04-18 11:36:21,386AC17893,0045322,17.48,1
```

This means that on April 18, 2009 at 11:36:21 in the session 386AC17893 the product 0045322, having the price of 17.48 EUR, was added to the shopping basket.

Now for the evaluation of a recommendation algorithm the historic data is read session-wise and the transactions are delivered in their actual order. For example, in a session first product A was clicked, than product B, then B was added to the basket and finally ordered. At each product click (i.e. the product detail view was called), the recommendations of the RE are requested. Then these recommendations are sub sequentially compared with the actual product clicks (Clicks), baskets (BK), orders (Ord), and revenue (Rev), and the statistics of the correct predictions is aggregated. Table 2 illustrates this procedure for one session.

Table 2 Illustration of simulation.

			real				right prediction			
Step	TA	REs	Clicks	BK	Ord	Rev	Clicks	BK	Ord	Rev
1	A	C, B	1	0	0	0	0	0	0	0
2	D	E, A	2	0	0	0	0	0	0	0
3	D in BK		2	1	0	0	0	0	0	0
4	A	C, E	3	1	0	0	1	0	0	0
5	A in BK		3	2	0	0	1	1	0	0
6	A ord.		3	2	1	35.00	1	1	1	35.00
Prediction rate							33%	50%	100%	100%

For evaluation, we use a transaction set from a web shop of a mail-order company. It contains a total of about 75,000 different products. The parameters of the RL algorithm are as follows: $TD(\lambda)$ algorithm with $\alpha = 0.1,\ \gamma = 0.5,\ \lambda = 0.9$. We always requested three recommendations for each click. The prediction rates for different transaction numbers are shown in Table 3.

For comparison, we use a multilevel preconditioner (19) with interpolations (25). So far, we use a two-level preconditioner for reasons of computational complexity. For the hierarchy we used the shop hierarchy of the products with a special preprocessing. Especially, the preprocessing excluded multiple parent categories from products and has balanced parent categories. The results of the preconditioned $TD(\lambda)$ algorithm are listed in Table 3, too.

The results show a better prediction rate of the multilevel preconditioner with respect to all transaction types. This means, with the same statistical data the preconditioned TD algorithm learns faster than the one without preconditioner.

Table 3 Prediction rates for TD and preconditioned TD algorithms.

	$TD(\lambda)$				$TD(\lambda)$ with $\mathbf{C}_t^{-1}$			
Sessions	Clicks	BK	Ord	Rev	Clicks	BK	Ord	Rev
1,000	0.77	0.00	0.00	0.00	0.77	0.00	0.00	0.00
5,000	1.25	0.95	2.50	1.48	1.25	0.95	2.50	1.48
10,000	1.66	2.19	2.12	5.36	1.66	2.19	2.12	5.36
50,000	2.73	2.98	4.86	8.77	2.71	3.01	5.03	9.40
100,000	3.71	4.13	5.93	7.24	3.81	4.26	6.02	7.80
500,000	5.81	5.98	7.56	8.11	6.43	6.67	8.41	9.34
1,000,000	7.14	7.09	9.06	10.66	8.36	8.40	10.74	12.21

5 Summary

We have described a multilevel preconditioner for TD of REs which works on a combined hierarchy of states and actions. The convergence of the methods could be proved. First experimental results indicate that this technique increases the convergence speed. An open work remains the estimation of the convergence speed. A further direction of research is the construction of the interpolation operator. One way, as done by Ziv in [19], is the use of algebraic techniques like coarsening to automatically extract the hierarchy. In case of REs, predefined hierarchies already exist and could be used. However, these hierarchies have been designed for marketing and logistics but not for the usage in multilevel techniques. On the other hand, their information is undoubted valuable. So it looks to be promising to combine both approaches in order to obtain an optimal hierarchy.

References

[1] Balabanovic, M.: An Adaptive Web Page Recommendation Service. CACM (1997)

[2] Bertsekas, D.P., Castanon, D.A.: Adaptive Aggregation Methods for Infinite Horizon Dynamic Programming. IEEE Trans. Automatic Control 34(6) (1989)

[3] Bramble, J., Pasciak, J., Xu, J.: Parallel multilevel preconditioners. Math. Comp. 55, 1–12 (1990)

[4] Brand, M.E.: Fast online svd revisions for lightweight recommender systems. In: SIAM International Conference on Data Mining, SDM (2003)

[5] Burke, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12(4) (2002)

[6] Golovin, N., Rahm, E.: Reinforcement Learning Architecture for Web Recommendations. In: Proc. ITCC 2004. IEEE (2004)

[7] Herlocker, J.L.: Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems 22(1) (2004)

[8] Linden, G., Smith, B., York, J.: Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Internet Computing (2003)

[9] Munos, R.: A study of reinforcement learning in the continuous case by the means of viscosity solutions. Machine Learning 40 (2000)

[10] Oswald, P.: Multilevel Finite Element Approximation. B.G. Teubner, Stuttgart (1994)

[11] Paprotny, A.: Praktikumsbericht zum Fachpraktikum bei der Firma prudsys AG. Report. TU Hamburg-Harburg (2009) (in German)

[12] Paprotny, A.: Hierarchical methods for the solution of dynamic programming equations arising from optimal control problems related to recommendation. Diploma thesis, TU Hamburg-Harburg (2010)

[13] Paprotny, A., Thess, M.: A stepwise approach to a self-learning recommendation engine. prudsys documentation, Chemnitz (2011)

[14] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, New Jersey (2002)

[15] Rojanavasu, P., Phaitoon, S., Pinngern, O.: New Recommendation System Using Reinforcement Learning. In: Proceedings of the Fourth International Conference on eBusiness, Bangkok, Thailand, November 19-20 (2005)

[16] Sutton, R.S., Barto, A.G.: Reinforcement Learning. An Introduction. MIT Press, Cambridge (1998)

[17] Shani, G., Heckerman, D., Brafman, R.I.: An MDP-based recommender system. Journal of Machine Learning Research 6 (2005)

[18] Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of Recommendation Algorithms for E-Commerce. In: EC 2000, Minneapolis, Minnesota, October 17-20 (2000)

[19] Ziv, O.: Algebraic Multigrid for Reinforcement Learning. Master thesis, Technion (2005)

Efficient Solvers for Saddle Point Problems with Applications to PDE–Constrained Optimization

Walter Zulehner

Abstract. We review some of the recent work on preconditioners for saddle point problems. In particular, we discuss preconditioners that are constructed based on exact or inexact Schur complements and on interpolation theory. These preconditioners are used within Krylov subspace methods, for which it is shown that the total number of iterations is bounded by global constants. The described techniques are applied to two model problems from optimal control.

1 Introduction

In this article we consider linear systems of equations in saddle point form: Find $u \in \mathbb{R}^n$ and $p \in \mathbb{R}^m$ such that

$$\mathscr{M} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \quad \text{with} \quad \mathscr{M} = \begin{bmatrix} A & B^\top \\ B & -C \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}, \tag{1}$$

where $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times m}$ are symmetric matrices, $B \in \mathbb{R}^{m\times n}$, and $B^\top$ denotes the transpose of B. Of special interest are problems of the form (1) with large and sparse matrices A, B, and C. Such systems arise by discretizing optimization problems in infinite-dimensional function spaces with constraints in form of partial differential equations (PDEs), see, e.g., [15, 21, 36] for typical problems of this form.

Krylov subspace methods are an appropriate class of iterative methods for solving (1) in this situation, we refer to the survey article [3] and the references cited there for these as well as other solution techniques for saddle point problems. See,

Walter Zulehner
Institut für Numerische Mathematik, Johannes Kepler Universität Linz,
Altenberger Str. 69, 4040 Linz, Austria
e-mail: zulehner@numa.uni-linz.ac.at

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 197–216.
springerlink.com

e.g., [20] for early work on Uzawa-type methods for saddle point problems. The linear systems considered in this paper are symmetric and indefinite, for which the minimal residual method (MINRES), see [29], is a prominent representative from the class of Krylov subspace methods. Without preconditioning the convergence rate of MINRES and other similar methods deteriorates if the mesh size, say h, of the underlying discretization method approaches 0. For optimization problems which contain some regularization parameter, say ν, a similar behavior can be observed for ν approaching 0. For some class of model problems we will discuss techniques for constructing preconditioners which eventually lead to convergence rates of the preconditioned MINRES method which do not show these deficiencies.

A well-known classical approach to construct preconditioners for (1) is based on knowledge about A and the (negative) Schur complement S associated with $\mathscr{M}$, given by

$$S = C + BA^{-1}B^{\top}.$$

In such an approach it is typically required that both A and S are positive definite. By interchanging the role of the variables, one can analogously construct Schur complement based preconditioners involving C and the Schur complement T, given by

$$T = A + B^{\top}C^{-1}B,$$

instead, provided C and T are positive definite. If A and C are positive definite, both approaches are available. Then we will see that interpolation theory can be used to construct a whole family of preconditioners, from which one can choose a candidate having some advantage compared to the original Schur complement based preconditioners. We refer to similar ideas in [25], where interpolation theory was used to construct efficient and robust preconditioners for several parameter-dependent problems including saddle point problems.

If neither A nor C is positive definite, none of these Schur complement based approaches is available. We will demonstrate a possible alternative approach in this situation by introducing an indefinite inner product such that A becomes positive definite with respect to this nonstandard inner product. This allows the construction of a preconditioner by using an inexact Schur complement.

The paper is organized as follows. Sect. 2 contains the variational framework for the problems considered here and a short summary of some estimates about condition numbers of the involved matrices, mainly taken from [18]. In Sect. 3, which closely follows the results presented in [40], block diagonal preconditioners are discussed and a few elements from interpolation theory are included, which are used to derive other preconditioners. Sect. 4 is devoted to the case that A is an indefinite matrix. The application of the abstract results of the previous two sections to model problems from optimal control are discussed in Sect. 5, followed by a few numerical experiments reported in Sect. 6.

2 The General Framework

For the analysis we formulate the linear system (1) as a variational problem: The matrices A, B, and C uniquely determine bilinear forms a, b, and c on $\mathbb{R}^n \times \mathbb{R}^n$, $\mathbb{R}^n \times \mathbb{R}^m$, and $\mathbb{R}^m \times \mathbb{R}^m$, respectively, given by

$$a(u,v) = \langle Au, v\rangle, \quad b(v,q) = \langle Bv, q\rangle = \langle B^\top q, v\rangle, \quad c(p,q) = \langle Cp, q\rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Then (1) can be written as a mixed variational problem: For given $f \in V = \mathbb{R}^n$ and $g \in Q = \mathbb{R}^m$, find $u \in V$ and $p \in Q$ such that

$$\begin{aligned} a(u,v) + b(v,p) &= \langle f, v\rangle \quad \text{for all } v \in V, \\ b(u,q) - c(p,q) &= \langle g, q\rangle \quad \text{for all } q \in Q. \end{aligned} \tag{3}$$

Or, equivalently, (1) can be written as a variational problem in the product space $X = V \times Q = \mathbb{R}^{n+m}$: Find $x \in X$ such that

$$\mathscr{B}(x,y) = \langle \mathscr{F}, y\rangle \quad \text{for all } y \in X \tag{4}$$

with

$$\begin{aligned} \mathscr{B}(x,y) = \langle \mathscr{M}x, y\rangle &= \langle Au, v\rangle + \langle B^\top p, v\rangle + \langle Bu, q\rangle - \langle Cp, q\rangle \\ &= a(u,v) + b(v,p) + b(u,q) - c(p,q) \end{aligned}$$

and

$$\mathscr{F} = \begin{bmatrix} f \\ g \end{bmatrix} \quad \text{for} \quad x = \begin{bmatrix} u \\ p \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} v \\ q \end{bmatrix}.$$

Next we introduce inner products and norms. Inner products in $V = \mathbb{R}^n$ and $Q = \mathbb{R}^m$ are uniquely represented by symmetric and positive definite matrices $P \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$:

$$(u,v)_V = \langle u, v\rangle_P, \quad (p,q)_Q = \langle p, q\rangle_R,$$

where here and in the sequel the following standard notations are used:

Definition 1. For a symmetric and positive definite matrix $M \in \mathbb{R}^{r\times r}$, the associated inner product (M-inner product), is given by $\langle z, w\rangle_M = \langle Mz, w\rangle$ for $z, w \in \mathbb{R}^r$. The norm (of both vectors and matrices) associated with the inner product $\langle \cdot, \cdot \rangle_M$ is denoted by $\|\cdot\|_M$ (M-norm).

With the notations introduced above, an inner product in $X = V \times Q = \mathbb{R}^{n+m}$ is given by

$$\begin{aligned} (x,y)_X &= (u,v)_V + (p,q)_Q = \langle u, v\rangle_P + \langle p, q\rangle_R \\ &= \langle x, y\rangle_{\mathscr{P}} \quad \text{with} \quad \mathscr{P} = \begin{bmatrix} P & 0 \\ 0 & R \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}. \end{aligned} \tag{5}$$

Additionally, we introduce dual norms by

$$\|\mathscr{F}\|_{X^*} = \sup_{0\neq y\in X} \frac{\langle \mathscr{F}, y\rangle}{\|y\|_X}, \quad \|f\|_{V^*} = \sup_{0\neq v\in V} \frac{\langle f, v\rangle}{\|v\|_V}, \quad \|g\|_{Q^*} = \sup_{0\neq q\in Q} \frac{\langle g, q\rangle}{\|q\|_Q}.$$

It is easy to see that

$$\|\mathscr{F}\|_{X^*} = \|\mathscr{F}\|_{\mathscr{P}^{-1}}, \quad \|f\|_{V^*} = \|f\|_{P^{-1}}, \quad \|g\|_{Q^*} = \|g\|_{R^{-1}}$$

and

$$\|\mathscr{F}\|_{X^*}^2 = \|f\|_{V^*}^2 + \|g\|_{Q^*}^2.$$

The norm $\|\mathscr{B}\|$ and the so called inf-sup constant γ associated with the bilinear $\mathscr{B}$ are given by

$$\|\mathscr{B}\| = \sup_{0\neq z\in X} \sup_{0\neq y\in X} \frac{|\mathscr{B}(z,y)|}{\|z\|_X \, \|y\|_X} \quad \text{and} \quad \gamma = \inf_{0\neq z\in X} \sup_{0\neq y\in X} \frac{|\mathscr{B}(z,y)|}{\|z\|_X \, \|y\|_X}.$$

These quantities can also be expressed in the following form:

$$\|\mathscr{B}\| = |\mu_{\max}| \quad \text{and} \quad \gamma = |\mu_{\min}|,$$

where $\mu_{\min}$ and $\mu_{\max}$ are the eigenvalues of the generalized eigenvalue problem

$$\mathscr{M} x = \mu \, \mathscr{P} x$$

of minimal and maximal modulus, respectively. The constants $\|\mathscr{B}\|$ and γ are related to the well-posedness of variational problems in a much more general context, see, e.g., see [1, 2].

Observe that the matrix $\mathscr{P}^{-1}\mathscr{M}$ is self-adjoint in the $\mathscr{P}$-inner product. Therefore, by the standard definition of a condition number in this case we have

$$\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right) = \frac{|\mu_{\max}|}{|\mu_{\min}|} = \frac{\|\mathscr{B}\|}{\gamma}.$$

So, we can gain information on the condition number of $\mathscr{P}^{-1}\mathscr{M}$ by analyzing the constants γ and $\mathscr{B}$ of the associated bilinear form $\mathscr{B}$ with respect to the $\mathscr{P}$-norm. Bounds for the condition number of $\mathscr{P}^{-1}\mathscr{M}$ allow us to estimate the convergence rate of MINRES for (1), preconditioned by the block diagonal matrix $\mathscr{P}$. For example, we have the following estimate for the residual of the $2k$-th iterate $x_{2k} \in \mathbb{R}^{n+m}$ produced by this method starting from an initial guess $x_0 \in \mathbb{R}^{n+m}$, see, e.g., [11]:

$$\|\mathscr{F} - \mathscr{M} x_{2k}\|_{\mathscr{P}^{-1}} \leq \frac{2q^k}{1+q^{2k}} \|\mathscr{F} - \mathscr{M} x_0\|_{\mathscr{P}^{-1}}$$

with

$$q = \frac{\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right) - 1}{\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right) + 1}.$$

Particularly well understood is the case $C = 0$. In this case the constants $\|\mathscr{B}\|$ and γ can be estimated in terms of the following constants:

$$\|a\| = \sup_{0\neq u\in V} \sup_{0\neq v\in V} \frac{|a(u,v)|}{\|u\|_V \|v\|_V}, \quad \|b\| = \sup_{0\neq v\in V} \sup_{0\neq q\in Q} \frac{|b(v,q)|}{\|v\|_V \|q\|_Q},$$

and

$$\alpha = \inf_{0\neq u\in \ker B} \sup_{0\neq v\in \ker B} \frac{|a(u,v)|}{\|u\|_V \|v\|_V}, \quad \beta = \inf_{0\neq q\in Q} \sup_{0\neq v\in V} \frac{|b(v,q)|}{\|v\|_V \|q\|_Q}.$$

Such estimates are already contained in the pioneering paper [7] and have been improved more recently, see, e.g., [10, 18, 32, 38]. Sharp upper bounds for $\|\mathscr{B}\|$ are well-known and easy to obtain:

$$\|\mathscr{B}\| \leq \rho\left(\begin{bmatrix} \|a\| & \|b\| \\ \|b\| & 0 \end{bmatrix}\right) = \frac{1}{2}\left(\|a\| + \sqrt{\|a\|^2 + 4\,\|b\|^2}\right), \tag{6}$$

where $\rho(M)$ denotes the spectral radius of a matrix M. For lower bounds for γ we recall a result from [18]:

Theorem 1. *With the notation introduced above, the following holds:*

1. *Let $\gamma_{opt}(\alpha,\beta,\|a\|)$ be the smallest positive root of the cubic equation*

$$\mu^3 - (\|a\|^2 + \beta^2)\mu + \alpha\beta^2 = 0. \tag{7}$$

   Then $\gamma \geq \gamma_{opt}(\alpha,\beta,\|a\|)$.
2. *We have $\gamma \geq \frac{\alpha}{1+\kappa^2}$ with $\kappa = \frac{\|a\|}{\beta}$.*

It was shown in [18] that the bound $\gamma_{\text{opt}}(\alpha,\beta,\|a\|)$ is sharp. The estimate in the second part of Theorem 1 provides a simpler and more transparent lower bound for the inf-sup constant γ.

3 Preconditioners Based on Schur Complements and Interpolation Theory

We start with a well-known case. Let

$$\mathscr{M} = \begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix}, \tag{8}$$

where A is a symmetric and positive definite matrix and B is of full rank $m \leq n$. For the block diagonal preconditioner

$$\mathscr{P} = \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \quad \text{with} \quad S = BA^{-1}B^\top$$

the spectrum of $\mathscr{P}^{-1}\mathscr{M}$, denoted by $\rho\left(\mathscr{P}^{-1}\mathscr{M}\right)$, consists only of three different values:

$$\rho\left(\mathscr{P}^{-1}\mathscr{M}\right)=\left\{\frac{1-\sqrt{5}}{2},1,\frac{1+\sqrt{5}}{2}\right\},$$

see, e.g., [19, 26]. This leads immediately to the following constants

$$\|\mathscr{B}\|=\frac{1+\sqrt{5}}{2},\quad \gamma=\frac{\sqrt{5}-1}{2}$$

for the bilinear form $\mathscr{B}$ associated with $\mathscr{M}$ with respect to the $\mathscr{P}$-norm, and, therefore, to the condition number $\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right)=\|B\|/\gamma=(\sqrt{5}+3)/2\approx 2.62$.

It is easy to see that a similar result is available for the slightly more general case

$$\mathscr{M}=\begin{bmatrix}A & B^\top\\ B & -C\end{bmatrix}$$

where A is a symmetric and positive definite matrix, B is of full rank $m\leq n$, and C is symmetric and positive semidefinite. One can show that

$$\rho\left(\mathscr{P}^{-1}\mathscr{M}\right)\subset\left[-1,\frac{1-\sqrt{5}}{2}\right]\cup\left[1,\frac{1+\sqrt{5}}{2}\right],$$

which implies that

$$\|\mathscr{B}\|\leq\frac{\sqrt{5}+1}{2},\quad \gamma\geq\frac{\sqrt{5}-1}{2},$$

and, therefore, $\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right)=\|B\|/\gamma\leq(\sqrt{5}+3)/2\approx 2.62$.

If both A and C are symmetric and positive definite, we can easily construct two different block preconditioners $\mathscr{P}_0$ and $\mathscr{P}_1$, given by

$$\mathscr{P}_0=\begin{bmatrix}A & 0\\ 0 & S\end{bmatrix}\quad\text{with}\quad S=C+BA^{-1}B^\top$$

and

$$\mathscr{P}_1=\begin{bmatrix}T & 0\\ 0 & C\end{bmatrix}\quad\text{with}\quad T=A+B^\top C^{-1}B,$$

both of which lead to estimates

$$\|\mathscr{B}\|\leq\frac{\sqrt{5}+1}{2},\quad \gamma\geq\frac{\sqrt{5}-1}{2}.$$

The preconditioner $\mathscr{P}_1$ is obtained along the same line of arguments as before applied to $\mathscr{M}$ after interchanging rows and columns. Therefore, we have the following estimates for $i\in\{0,1\}$:

$$\frac{\sqrt{5}-1}{2}\,\|x\|_{\mathscr{P}_i}\leq\|\mathscr{M}x\|_{\mathscr{P}_i^{-1}}\quad\text{and}\quad\|\mathscr{M}x\|_{\mathscr{P}_i^{-1}}\leq\frac{\sqrt{5}+1}{2}\,\|x\|_{\mathscr{P}_i}$$

for all $x \in \mathbb{R}^{n+m}$. The first inequality can also be written as

$$\|\mathscr{M}^{-1}x\|_{\mathscr{P}_i^{-1}} \leq \frac{\sqrt{5}+1}{2}\,\|x\|_{\mathscr{P}_i} \quad \text{for all} \quad x \in \mathbb{R}^{n+m}.$$

Motivated by results from interpolation theory, see [4] for a detailed introduction, we consider now a whole family of preconditioners $\mathscr{P}_\theta$ for $\theta \in (0,1)$, given by

$$\mathscr{P}_\theta = \mathscr{P}_0^{1/2}\left(\mathscr{P}_0^{-1/2}\mathscr{P}_1\mathscr{P}_0^{-1/2}\right)^\theta \mathscr{P}_0^{1/2}, \tag{9}$$

where M^r denotes the r-th power of a symmetric and positive matrix M for $r \in \mathbb{R}$, see [14] for more details on matrix functions. The choice of this family of matrices is motivated by a particular representation of the $\mathscr{P}_0$-norm and the $\mathscr{P}_1$-norm. Let $\lambda_i > 0$, $i = 1,2,\ldots,n+m$ denote the eigenvalues of the generalized eigenvalue problem

$$\mathscr{P}_1 x = \lambda\, \mathscr{P}_0 x.$$

The associated eigenvectors $e_i \in \mathbb{R}^{n+m}$, $i = 1,2,\ldots,n+m$ are chosen to form a orthonormal basis with respect to the $\mathscr{P}_0$-inner product. Then it is easy to see that

$$\|x\|_{\mathscr{P}_0}^2 = \sum_{i=1}^{n+m} \hat{x}_i^2 \quad \text{and} \quad \|x\|_{\mathscr{P}_1}^2 = \sum_{i=1}^{n+m} \lambda_i \hat{x}_i^2 \quad \text{with} \quad x = \sum_{i=1}^{n+m} \hat{x}_i\, e_i.$$

It is quite natural to consider a new norm $\|.\|_\theta$ for $\theta \in (0,1)$, given by

$$\|x\|_\theta^2 = \sum_{i=1}^{n+m} \lambda_i^\theta\, \hat{x}_i^2,$$

which in some sense lies between the $\mathscr{P}_0$-norm and $\mathscr{P}_1$-norm. By direct calculation one can show that this norm is the norm associated with the matrix $\mathscr{P}_\theta$:

$$\|x\|_\theta = \|x\|_{\mathscr{P}_\theta}.$$

There is an interesting integral representation of this norm, see [22]:

$$\|x\|_{\mathscr{P}_\theta}^2 = c_\theta^{-1} \int_0^\infty t^{-(2\theta+1)} K(t;x)^2\, dx \quad \text{with} \quad c_\theta = \frac{\pi}{2\sin(\theta\pi)}$$

and

$$K(t;x) = \inf_{x = x_0 + x_1}\left(\|x_0\|_{\mathscr{P}_0}^2 + t^2\,\|x_1\|_{\mathscr{P}_1}^2\right)^{1/2}.$$

So $\mathscr{P}_\theta$ represents the inner product of the interpolation space with index θ associated with the two spaces whose inner products are represented by $\mathscr{P}_0$ and $\mathscr{P}_1$ in terms of the so-called K-method, see [4] for details. Therefore, we will use the notation

$$\mathscr{P}_\theta = [\mathscr{P}_0, \mathscr{P}_1]_\theta$$

for the matrix $\mathscr{P}_\theta$ defined by (9) in analogy to the corresponding notation for interpolation spaces. Replacing $\mathscr{P}_0$ and $\mathscr{P}_1$ in definition (9) by $\mathscr{P}_0^{-1}$ and $\mathscr{P}_1^{-1}$, then leads to an interpolation matrix $\left[\mathscr{P}_0^{-1}, \mathscr{P}_1^{-1}\right]_\theta$, for which it is easy to see that

$$\left[\mathscr{P}_0^{-1}, \mathscr{P}_1^{-1}\right]_\theta = [\mathscr{P}_0, \mathscr{P}_1]_\theta^{-1}.$$

This property is called the duality theorem in interpolation theory. The main result in interpolation theory is the interpolation theorem, see [4], which reads for the special norms considered here:

Theorem 2. *Let $T \in \mathbb{R}^{(n+m)\times(n+m)}$ with*

$$\|Tx\|_{\mathscr{P}_0^{-1}} \leq c_0 \, \|x\|_{\mathscr{P}_0} \quad \textit{and} \quad \|Tx\|_{\mathscr{P}_1^{-1}} \leq c_1 \, \|x\|_{\mathscr{P}_1} \quad \textit{for all} \quad x \in \mathbb{R}^{n+m}. \tag{10}$$

Then

$$\|Tx\|_{\left[\mathscr{P}_0^{-1},\mathscr{P}_1^{-1}\right]_\theta} \leq c \, \|x\|_{[\mathscr{P}_0,\mathscr{P}_1]_\theta} \quad \textit{with} \quad c = c_0^{1-\theta} c_1^\theta \quad \textit{for all} \quad x \in \mathbb{R}^{n+m}.$$

If this theorem is applied to $T = \mathscr{M}$, it leads to the following estimate:

$$\|\mathscr{M}x\|_{\mathscr{P}_\theta^{-1}} \leq \frac{\sqrt{5}+1}{2} \, \|x\|_{\mathscr{P}_\theta} \quad \text{for all} \quad x \in \mathbb{R}^{n+m},$$

since the conditions (10) are satisfied for $c_0 = c_1 = (\sqrt{5}+1)/2$. Similarly, we obtain for $T = \mathscr{M}^{-1}$ the estimate

$$\|\mathscr{M}^{-1}x\|_{\mathscr{P}_\theta^{-1}} \leq \frac{\sqrt{5}+1}{2} \, \|x\|_{\mathscr{P}_\theta} \quad \text{for all} \quad x \in \mathbb{R}^{n+m},$$

Therefore,

$$\kappa\left(\mathscr{P}_\theta^{-1}\mathscr{M}\right) \leq \left(\frac{\sqrt{5}+1}{2}\right)^2 = \frac{\sqrt{5}+3}{2}.$$

To summarize, starting from two Schur complement based preconditioners $\mathscr{P}_0$ and $\mathscr{P}_1$ we obtain a whole family of preconditioners $\mathscr{P}_\theta$ with the same estimate for the condition number as for $\mathscr{P}_0$ and $\mathscr{P}_1$.

The application of a preconditioner requires that the matrix-vector product $\mathscr{P}^{-1}y$ can be efficiently done for a given vector y. This is typically not the case for the preconditioners discussed so far, if they are applied to interesting saddle point problems. We will address them as ideal (but not yet realizable) preconditioners. For practical problems an ideal preconditioner $\mathscr{P}$ has to be replaced by a realizable preconditioner $\hat{\mathscr{P}}$, which on the one hand allows us to efficiently perform $\hat{\mathscr{P}}^{-1}y$ and on the other hand keeps the condition number close to the condition number of the ideal preconditioner. The last property requires that

$$c_1 \, \hat{\mathscr{P}} \leq \mathscr{P} \leq c_2 \, \hat{\mathscr{P}}, \tag{11}$$

with positive constants c_1 and c_2 that do not depend on parameters like h or ν. We call such constants global constants. Here, the notation $M \le N$ for symmetric matrices M and N means that $N-M$ is positive semi-definite. We call two symmetric matrices M and N spectrally equivalent if a condition of the form $c_1 M \le N \le c_2 M$ is satisfied with global constants c_1 and c_2. It is easy to see that (11) implies

$$\kappa\left(\hat{\mathscr{P}}^{-1}\mathscr{M}\right) \le \frac{c_2}{c_1}\,\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right).$$

For the Schur complement based ideal preconditioner $\mathscr{P}_0$, such a realizable version is typically obtain by replacing A and S by spectrally equivalent preconditioners $\hat{A}$ and $\hat{S}$, which are constructed based on specific properties of the considered problem. Such strategies were proposed and analyzed in the context of mixed formulations of second-order elliptic equations and the Stokes problem in, e.g., [32, 34, 37].

4 Preconditioners Based on Inexact Schur Complements

We consider now the case that

$$\mathscr{M} = \begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix}, \tag{12}$$

where A is a symmetric but indefinite matrix and B is of full rank $m \le n$. From Brezzi's theory, see [7], is known that $\mathscr{M}$ non-singular if and only if the inf-sup constant α introduced in Sect. 2 is strictly positive. The next result, taken from [18], provides a simple criterion to verify this. Its short proof is included for completeness.

Lemma 1. *Assume there is a non-singular matrix $H \in \mathbb{R}^{n\times n}$ with*

1. $\langle Au, Hu\rangle > 0$ *for all $u \in \mathbb{R}^n$ with $u \ne 0$, and*
2. $\ker B$ *is an invariant subspace of H, i.e., if $u \in \ker B$, then $Hu \in \ker B$.*

Then, for

$$P = \frac{1}{2}(H^\top A + AH) \quad \textit{and} \quad R = BP^{-1}B^\top,$$

we have

$$\alpha \ge \frac{1}{\|H\|_P} > 0, \quad \|a\| = \|P^{-1}A\|_P, \quad \textit{and} \quad \beta = \|b\| = 1,$$

where α, $\|a\|$ and β, $\|b\|$ are the constants for the bilinear forms a and b associated with A and B with respect to the P-norm and the R-norm.

Proof. For $0 \ne u \in \ker B$ and the specific choice $v = Hu \in \ker H$ it follows that

$$\sup_{0\ne v\in\ker B} \frac{\langle Au, v\rangle}{\|v\|_P} \ge \frac{\langle Au, Hu\rangle}{\|Hu\|_P} = \frac{\|u\|_P^2}{\|Hu\|_P} \ge \frac{\|u\|_P^2}{\|H\|_P\|u\|_P} = \frac{1}{\|H\|_P}\|u\|_P,$$

which immediately implies that $\alpha \geq \frac{1}{\|H\|_P} > 0$. The representation for $\|a\|$ easily follows from its definition, since

$$\sup_{0 \neq v \in V} \frac{\langle Au, v\rangle}{\|v\|_P} = \sup_{0 \neq v \in V} \frac{\langle P^{-1}Au, v\rangle_P}{\|v\|_P} = \|P^{-1}Au\|_P.$$

For the second part, observe that

$$\sup_{0 \neq v \in V} \frac{\langle Bv, q\rangle}{\|v\|_P} = \sqrt{\langle BP^{-1}B^T q, q\rangle} = \|q\|_R,$$

which implies that $\beta = \|b\| = 1$. □

Lemma 1 suggests to use the following preconditioner for (12):

$$\mathscr{P} = \begin{bmatrix} P & 0 \\ 0 & BP^{-1}B^\top \end{bmatrix},$$

where $BP^{-1}B^\top$ is usually called an inexact Schur complement. Estimates for $\|\mathscr{B}\|$ and γ, where $\mathscr{B}$ is the bilinear form associated with $\mathscr{M}$, follow from (6), Theorem 1, and the estimates or values for $\|a\|$, $\|b\|$, α, and β in Lemma 1.

Condition 1 in Lemma 1 requires that the transformed matrix $H^\top A$ is positive definite (but not necessarily symmetric), or, equivalently, that the symmetric part of the transformed matrix $H^\top A$ is positive definite. In typical applications, see, e.g., the velocity tracking problem discussed in Subsect. 5.2, the matrix A itself is of saddle point form, say

$$A = \begin{bmatrix} D & E^\top \\ E & -F \end{bmatrix}.$$

Starting with the pioneering article [5], several techniques for transforming a saddle point matrix into a positive definite matrix have been studied, see [39] and, more generally, [35]. Observe, however, that Condition 2 of Lemma 1 constitutes an additional restriction on the possible choices for H.

A different and ad hoc approach is used in Subsect. 5.2 for constructing the transformation matrix H by exploiting the particular structure of the problem.

5 Application to Optimal Control Problems

In this section the strategies developed in Sect. 3 and 4 are applied to two model problems from optimal control. Let Ω be an open and bounded domain in $\mathbb{R}^d$ for $d \in \{1,2,3\}$ with Lipschitz-continuous boundary Γ and let $L^2(\Omega)$, $H^1(\Omega)$, and $H^1_0(\Omega)$ be the usual Lebesgue space and Sobolev spaces of functions on Ω.

5.1 Distributed Optimal Control for Elliptic Equations

First we consider the following model problem: Find the state $y \in H^1(\Omega)$ and the control $u \in L^2(\Omega)$ that minimizes the cost functional

$$J(y,u) = \frac{1}{2}\int_{\Omega} |y(x) - y_d(x)|^2 \, dx + \frac{\nu}{2}\int_{\Omega} |u(x)|^2 \, dx$$

subject to the elliptic state equation

$$\begin{aligned} -\Delta y(x) &= u(x) && \text{in } \Omega, \\ y(x) &= 0 && \text{on } \Gamma. \end{aligned}$$

Here $y_d \in L^2(\Omega)$ is a given target (or desired) state and $\nu > 0$ is a cost or regularization parameter.

Using an appropriate finite element space V_h of dimension n for both y and u, we obtain the following discrete version: Minimize

$$\frac{1}{2}(\underline{y} - \underline{y}_d)^\top M(\underline{y} - \underline{y}_d) + \frac{\nu}{2}\underline{u}^\top M\underline{u}$$

subject to

$$K\underline{y} = M\underline{u}.$$

Here the matrices M and K are the mass matrix, representing the L^2-inner product in V_h, and the discretized negative Laplacian, respectively. The underlined quantities $\underline{y}$, $\underline{y}_d$, and $\underline{u}$ denote the coefficient vectors of the corresponding finite element functions relative to the chosen set of basis functions in V_h.

The Lagrangian functional for this constrained optimization problem is given by

$$\mathscr{L}(\underline{y},\underline{u},\underline{p}) = \frac{1}{2}(\underline{y} - \underline{y}_d)^\top M(\underline{y} - \underline{y}_d) + \frac{\nu}{2}\underline{u}^T M\underline{u} + \underline{p}^\top \left(K\underline{y} - M\underline{u}\right),$$

where $\underline{p}$ denotes the Lagrangian multiplier associated with the constraint.

The first-order necessary optimality conditions, which are also sufficient for the problem considered here, are $\nabla\mathscr{L}(\underline{y},\underline{u},\underline{p}) = 0$, and read in details:

$$\begin{bmatrix} M & 0 & K \\ 0 & \nu M & -M \\ K & -M & 0 \end{bmatrix} \begin{bmatrix} \underline{y} \\ \underline{u} \\ \underline{p} \end{bmatrix} = \begin{bmatrix} M\underline{y}_d \\ 0 \\ 0 \end{bmatrix}. \tag{13}$$

For preconditioners constructed by approximating Schur complement based preconditioners for (13), see, e.g., [31]. In [33] indefinite preconditioners for (13) were developed based on inexact Schur complements, for which it could be shown that the condition number of the preconditioned matrix is bounded pendently of the regularization parameter ν.

An alternative approach is to eliminate $\underline{y}$ and $\underline{p}$ by using the third and the first row of (13). Then the reduced equation for $\underline{u}$ reads

$$\left[\nu M + MK^{-1}MK^{-1}M\right] \underline{u} = MK^{-1}M\underline{y}_d.$$

For a multigrid preconditioner for this equation, see, e.g., [13].

We follow here closely the work in [40] (see also [33]), where yet another approach was chosen. From the second row of (13) it follows that

$$\underline{u} = \frac{1}{\nu}\underline{p}.$$

Therefore, we can eliminate the control $\underline{u}$ and obtain the reduced optimality system, which reads after a simple scaling:

$$\begin{bmatrix} M & \sqrt{\nu}K \\ \sqrt{\nu}K & -M \end{bmatrix} \begin{bmatrix} \underline{y} \\ \frac{1}{\sqrt{\nu}}\underline{p} \end{bmatrix} = \begin{bmatrix} M\underline{y}_d \\ 0 \end{bmatrix}. \qquad (14)$$

The two ideal Schur complement based preconditioners discussed in Sect. 3 for (14) are given by

$$\mathscr{P}_0 = \begin{bmatrix} M & 0 \\ 0 & M + \nu KM^{-1}K \end{bmatrix} \quad \text{and} \quad \mathscr{P}_1 = \begin{bmatrix} M + \nu KM^{-1}K & 0 \\ 0 & M \end{bmatrix}.$$

A realizable preconditioner derived from any of these two ideal preconditioners requires an easy-to-invert approximation for the Schur complement $M + \nu KM^{-1}K$, which can be seen as the discretization matrix of a 4-th order differential operator. We will see that this can be avoided by choosing $\mathscr{P}_\theta$ with $\theta = 1/2$ as new preconditioner:

$$\begin{aligned} \mathscr{P}_{1/2} &= [\mathscr{P}_0, \mathscr{P}_1]_{1/2} \\ &= \begin{bmatrix} \left[M, M + \nu KM^{-1}K\right]_{1/2} & 0 \\ 0 & \left[M + \nu KM^{-1}K, M\right]_{1/2} \end{bmatrix} \end{aligned}$$

This preconditioner consists of two identical diagonal blocks, for which it can be shown that

$$\frac{1}{\sqrt{2}}\left(M + \left[M, \nu KM^{-1}K\right]_{1/2}\right) \le \left[M, M + \nu KM^{-1}K\right]_{1/2} \le M + \left[M, \nu KM^{-1}K\right]_{1/2},$$

see [40]. Using (9) it is easy to see that

$$\left[\nu KM^{-1}K, M\right]_{1/2} = \sqrt{\nu}K.$$

Therefore, the preconditioner $\mathscr{P}$, given by

$$\mathscr{P} = \begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \quad \text{with} \quad P = M + \sqrt{\nu}K,$$

performs about as well as $\mathscr{P}_{1/2}$. It can be shown that $\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right) \leq \sqrt{2}$, see [40].

However, $\mathscr{P}$ is still an ideal preconditioner. A realizable version of it requires an efficient approximate evaluation of $\underline{z} = (M + \sqrt{\nu}K)^{-1}\underline{y}$. The matrix $M + \sqrt{\nu}K$ results from the discretization of a standard second-order differential operator, for which such methods are available. For example, one can choose one step of a V-cycle of a multigrid method (with one step of the symmetric Gauss-Seidel method for pre- and post-smoothing) applied to the equation $(M + \sqrt{\nu}K)\underline{z} = \underline{y}$, starting from the initial guess 0. This corresponds to a preconditioner for $M + \sqrt{\nu}K$, which is known to be spectrally equivalent, see, e.g., [27].

Similar results as presented here for elliptic state equations were derived for state equations describing time-periodic eddy current problems and time-periodic parabolic state equations in [9, 16, 17].

5.2 The Velocity Tracking Problem for Stokes Flow

Next we consider the following model problem: Find the velocity $\mathbf{u} \in H^1(\Omega)^d$, the pressure $p \in L_0^2(\Omega) = \{q \in L^2(\Omega) \colon \int_\Omega q \, dx = 0\}$, and the force $\mathbf{f} \in L^2(\Omega)^d$ that minimizes the cost functional

$$J(\mathbf{u},\mathbf{f}) = \frac{1}{2}\int\limits_\Omega |\mathbf{u}(x) - \mathbf{u}_d(x)|^2 \, dx + \frac{\nu}{2}\int\limits_\Omega |\mathbf{f}(x)|^2 \, dx$$

subject to the Stokes problem

$$\begin{aligned} -\Delta\mathbf{u}(x) + \nabla p(x) &= \mathbf{f}(x) && \text{in } \Omega, \\ \nabla \cdot \mathbf{u}(x) &= 0 && \text{in } \Omega, \\ \mathbf{u}(x) &= 0 && \text{on } \Gamma. \end{aligned}$$

Here $\mathbf{u}_d \in L^2(\Omega)^d$ is a given target velocity, $\nu > 0$ is a cost or regularization parameter, and $|.|$ denotes the Euclidean norm in $\mathbb{R}^d$.

This problem was discussed in [40]. The analysis presented here follows closely the presentation in [18], where the time-periodic variant of the problem is discussed.

Using appropriate finite element spaces $\mathbf{V}_h$ of dimension n and Q_h of dimension m for $\mathbf{u}$ and p, respectively, and the same finite element space $\mathbf{V}_h$ for $\mathbf{f}$ as well, we obtain the following discrete version: Minimize

$$\frac{1}{2}(\underline{\mathbf{u}} - \underline{\mathbf{u}}_d)^\top \mathbf{M}(\underline{\mathbf{u}} - \underline{\mathbf{u}}_d) + \frac{\nu}{2}\underline{\mathbf{f}}^\top \mathbf{M}\underline{\mathbf{f}}$$

subject to

$$\begin{aligned} \mathbf{K}\underline{\mathbf{u}} - \mathbf{D}^\top \underline{p} &= \mathbf{M}\underline{\mathbf{f}}, \\ \mathbf{D}\underline{\mathbf{u}} &= 0. \end{aligned}$$

Here the matrices $\mathbf{M}$, $\mathbf{K}$, and $\mathbf{D}$ are the mass matrix, representing the L^2-inner product in $\mathbf{V}_h$, the discretized negative vector Laplacian, and the discretized divergence, respectively. As before, underlined quantities denote the coefficient vectors of finite element functions relative to a basis.

The optimality system, which characterizes a solution to this constrained optimization problem, can be derived similarly to the optimal control problem discussed in Subsect. 5.2, and leads, again after eliminating the control, here $\mathbf{f}$, and a proper scaling, to the following system:

$$\begin{bmatrix} \mathbf{M} & \sqrt{\nu}\mathbf{K} & 0 & -\sqrt{\nu}\mathbf{D}^\top \\ \sqrt{\nu}\mathbf{K} & -\mathbf{M} & -\sqrt{\nu}\mathbf{D}^\top & 0 \\ 0 & -\sqrt{\nu}\mathbf{D} & 0 & 0 \\ -\sqrt{\nu}\mathbf{D} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \underline{\mathbf{u}} \\ \frac{1}{\sqrt{\nu}}\underline{\mathbf{w}} \\ \underline{p} \\ \frac{1}{\sqrt{\nu}}\underline{r} \end{bmatrix} = \begin{bmatrix} \mathbf{M}\underline{\mathbf{u}}_d \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

which is of the form (1) with

$$A = \begin{bmatrix} \mathbf{M} & \sqrt{\nu}\mathbf{K} \\ \sqrt{\nu}\mathbf{K} & -\mathbf{M} \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -\sqrt{\nu}\mathbf{D} \\ -\sqrt{\nu}\mathbf{D} & 0 \end{bmatrix}, \quad \text{and} \quad C = 0. \tag{15}$$

For preconditioners constructed by approximating Schur complement based preconditioners for the corresponding unreduced optimality system, see, e.g., [30].

Obviously, A is indefinite. In order to apply Lemma 1 we choose the matrix H in the following way:

$$H = \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{I} & -\mathbf{I} \end{bmatrix}.$$

It is easy to verify that $\langle Au, Hu \rangle > 0$, if $u \neq 0$, and that $\ker B$ is an invariant subspace of H. Therefore, according to Lemma 1, the first block P of the preconditioner $\mathscr{P}$ is the symmetric part of $H^\top A = HA$, which leads to:

$$P = \frac{1}{2}(H^\top A + AH) = \begin{bmatrix} \mathbf{P} & 0 \\ 0 & \mathbf{P} \end{bmatrix} \quad \text{with} \quad \mathbf{P} = \mathbf{M} + \sqrt{\nu}\mathbf{K},$$

while the second block R is the inexact Schur complement associated with P:

$$R = BR^{-1}B^\top = \nu \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix} \quad \text{with} \quad S = \mathbf{D}\left[\mathbf{M} + \sqrt{\nu}\mathbf{K}\right]^{-1}\mathbf{D}^\top.$$

The following theorem contains estimates for the constants α, $\|a\|$ and β, $\|b\|$ for the bilinear forms a and b associated with A and B with respect to the P-norm and the R-norm, see [18] for the corresponding result for the time-periodic case.

Theorem 3. *Let $\mathscr{M}$ be given by (1) and (15). Then, for $\mathscr{P}$ given by*

$$\mathscr{P} = \begin{bmatrix} P & 0 \\ 0 & S \end{bmatrix} \quad \textit{with} \quad P = \begin{bmatrix} \mathbf{P} & 0 \\ 0 & \mathbf{P} \end{bmatrix} \quad \textit{and} \quad R = \nu \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix},$$

where

$$\mathbf{P} = \mathbf{M} + \sqrt{\nu}\mathbf{K} \quad \textit{and} \quad S = \mathbf{D}\left[\mathbf{M} + \sqrt{\nu}\mathbf{K}\right]^{-1}\mathbf{D}^\top,$$

the following estimates hold

$$\alpha \geq \frac{1}{\sqrt{2}}, \quad \|a\| \leq 1, \quad \beta = 1, \quad \|b\| = 1.$$

Proof. Observe that $\|H\|_P = \sqrt{2}$. Then the results for α, β, and $\|b\|$ follow directly from Lemma 1.

It is easy to see that $A \leq P$ and $-A \leq P$, which implies that

$$\|a\| = \|P^{-1}A\|_P = \|P^{-1/2}AP^{-1/2}\| \leq 1. \qquad \square$$

If α, $\|a\|$, β, and $\|b\|$ in (6) and (7) are replaced by the corresponding lower or upper bounds provided by Theorem 3, we immediately obtain the following bounds:

$$\|\mathscr{B}\| \leq \frac{1}{2}(1+\sqrt{5}) \quad \text{and} \quad \gamma \geq \mu_3,$$

where μ_3 is the smallest positive root of the cubic equation

$$\mu^3 - 2\mu + \frac{1}{\sqrt{2}} = 0.$$

These bounds read in 3-digit accuracy

$$\|\mathscr{B}\| \leq 1.618 \quad \text{and} \quad \gamma \geq 0.381,$$

leading to $\kappa\left(\mathscr{P}^{-1}\mathscr{M}\right) \leq 4.25$,

From the second part of Theorem 1 we know a simpler lower bound for μ_3:

$$\mu_3 \geq \frac{1}{2\sqrt{2}} \approx 0.354.$$

Applying the ideal preconditioner requires to evaluate

$$\left[\mathbf{M} + \sqrt{\nu}\mathbf{K}\right]^{-1}\underline{\mathbf{v}} \quad \text{and} \quad \left[\mathbf{D}\left(\mathbf{M} + \sqrt{\nu}\mathbf{K}\right)^{-1}\mathbf{D}^\top\right]^{-1}\underline{q} \tag{16}$$

for given vectors $\underline{\mathbf{v}}$ and $\underline{q}$. As proposed in [8] the matrix in the second term in (16) is replaced by the matrix $K_p^{-1} + \sqrt{\nu}M_p^{-1}$, where M_p and K_p denote the mass matrix and the discretized negative Laplacian in the finite element space for the pressure, respectively (Cahouet-Charbard preconditioner). This modification of the original preconditioner requires to evaluate only

$$\left[\mathbf{M}+\sqrt{\nu}\mathbf{K}\right]^{-1}\underline{\mathbf{v}}, \quad K_p^{-1}\underline{q}, \quad \text{and} \quad M_p^{-1}\underline{q}. \tag{17}$$

For eventually obtaining a realizable preconditioner we proceed as described at the end of Subsect. 5.1. For the first and the second term in (17) we use one step of a V-cycle with one step of a symmetric Gauss-Seidel method for pre- and post-smoothing, for the third term we use one step of a symmetric Gauss-Seidel method. We refer to [6, 23, 24, 28] for a discussion of the spectral equivalence of the resulting realizable preconditioner and the original ideal preconditioner.

6 Numerical Experiments

Here we present some numerical experiments for the velocity tracking problem on the unit square domain $\Omega = (0,1)\times(0,1) \subset \mathbb{R}^2$. Following Example 1 in [12] we choose the target velocity $\mathbf{u}_d(x,y) = [(U(x,y),V(x,y)]^T$, given by

$$U(x,y) = 10\frac{\partial}{\partial y}(\varphi(x)\varphi(y)) \quad \text{and} \quad V(x,y) = -10\frac{\partial}{\partial x}(\varphi(x)\varphi(y))$$

with

$$\varphi(z) = \big(1-\cos(0.8\pi z)\big)(1-z)^2.$$

The velocity $\mathbf{u}_d(x,y)$ is divergence free. Note that, contrary to the velocity tracking problem for steady-state Stokes flow considered here, in [12] the velocity tracking problem was discussed for a time-dependent Navier-Stokes flow.

The problem was discretized by the Taylor-Hood pair of finite element spaces consisting of continuous piecewise quadratic polynomials for the velocity $\mathbf{u}(x,y)$ and the force $\mathbf{f}(x,y)$ and continuous piecewise linear polynomials for the pressure $p(x,y)$ on a triangulation of Ω. The initial mesh contains four triangles obtained by intersecting Ω by its two diagonals. The final mesh was constructed by applying k uniform refinement steps to the initial mesh, leading to a mesh size $h = 2^{-k}$.

Numerical experiments are presented for the realizable version of the ideal preconditioner $\mathscr{P}$, as described in Subsect. 5.2.

For comparison, we also consider a preconditioner which is based on a more conventional approach. The infinite-dimensional analog of (14) is a system of partial differential equations of second order. This system is typically formulated for $\mathbf{u},\mathbf{w} \in H_0^1(\Omega)^d$ and $p,r \in L^2(\Omega)$. It is easy to see that, for fixed ν, the infinite-dimensional problem is well-posed in these spaces equipped with the standard Lebesgue and Sobolev norms. The matrix

$$\mathscr{P}_{\mathrm{st}} = \begin{bmatrix} \mathbf{K} & 0 & 0 & 0 \\ 0 & \mathbf{K} & 0 & 0 \\ 0 & 0 & M_p & 0 \\ 0 & 0 & 0 & M_p \end{bmatrix}$$

represents these norms on the corresponding finite element spaces. So, it is quite natural to consider $\mathscr{P}_{\text{st}}$ as a reasonable candidate of a preconditioner. For the numerical experiments we consider a realizable version of $\mathscr{P}_{\text{st}}$, constructed completely analogously as for $\mathscr{P}$.

Table 1 shows the condition number of $\mathscr{P}_{\text{st}}^{-1}\mathscr{M}$ for various values of h and ν, where k denotes the number of refinements and N is the total number of degrees of freedom of the discretized reduced optimality system.

In Table 2 the number of MINRES iterations is shown if using the preconditioner $\mathscr{P}_{\text{st}}$, required for reducing the $\mathscr{P}_{\text{st}}^{-1}$-norm of the initial residual by a factor $\varepsilon = 10^{-8}$. The initial guess was 0.

Table 1 Standard preconditioner – condition numbers.

k	N	ν				
		10^{-3}	10^{-2}	10^{-1}	1	10^{4}
4	9 030	> 1000	74.7	10.2	9.76	9.79
5	36 486	> 1000	75	10.4	10.2	10.2
6	146 694	> 1000	75	10.6	10.5	10.5
7	588 294	> 1000	75	10.7	10.7	10.8

Table 2 Standard preconditioner – number of MINRES iterations.

k	N	ν				
		10^{-3}	10^{-2}	10^{-1}	1	10^{4}
4	9 030	687	152	106	108	108
5	36 486	650	153	114	114	114
6	146 694	650	160	120	120	120
7	588 294	659	164	126	126	126

Table 3 and Table 4 show the corresponding results for the (nonstandard) preconditioner $\mathscr{P}$, described in the Subsect. 5.2. Observe the different range for the parameter ν compared to Table 1 and Table 2.

As expected, the number of iterations is robust with respect to the mesh size h for both preconditioners. However, the nonstandard preconditioner $\mathscr{P}$ shows also a robust behavior if the regularization parameter ν approaches zero, while the condition number and the number of iterations grow significantly for $\nu \leq 10^{-2}$ for the standard preconditioner $\mathscr{P}_{\text{st}}$.

Table 3 Nonstandard preconditioner – condition numbers.

k	N	ν				
		10^{-12}	10^{-6}	10^{-4}	1	10^4
4	9 030	3.43	4.01	7.39	9.52	9.75
5	36 486	3.42	4.83	8.20	9.98	10.2
6	146 694	3.45	6.06	8.88	10.3	10.5
7	588 294	3.80	7.12	9.45	10.6	10.7

Table 4 Nonstandard preconditioner – number of MINRES iterations.

k	N	ν				
		10^{-12}	10^{-8}	10^{-4}	1	10^4
4	9 030	32	45	89	106	108
5	36 486	34	47	95	112	114
6	146 694	34	51	101	118	120
7	588 294	34	55	107	124	126

Acknowledgements. The research was supported by the Austrian Science Fund (FWF): W1214-N15, project DK12. The material presented here is based on joint work with Joachim Schöberl (Vienna, Austria), Valeria Simoncini (Bologna, Italy), Markus Kollmann, and Wolfgang Krendl, both from Linz, Austria. A first preliminary version of the contents of this paper was presented in a talk given by the author at the 24th Chemnitz FEM Symposium 2011, Chemnitz, Germany.

References

[1] Babuška, I.: Error-bounds for finite element method. Numer. Math. 16, 322–333 (1971)

[2] Babuška, I.: The finite element method with Lagrangian multipliers. Numer. Math. 20, 179–192 (1973)

[3] Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numerica 14, 1–137 (2005)

[4] Bergh, J., Löfström, J.: Interpolation Spaces. An Introduction. Die Grundlehren der mathematischen Wissenschaften. Band 223. Springer, Heidelberg (1976)

[5] Bramble, J.H., Pasciak, J.E.: A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. Math. Comp. 50, 1–17 (1988)

[6] Bramble, J.H., Pasciak, J.E.: Iterative techniques for time dependent Stokes problems. Comput. Math. Appl. 33(1-2), 13–30 (1997)

[7] Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. RAIRO 8, 129–151 (1974)

[8] Cahouet, J., Chabard, J.P.: Some fast 3D finite element solvers for the generalized Stokes problem. Int. J. Numer. Methods Fluids 8(8), 865–895 (1988)

[9] Copeland, D., Kolmbauer, M., Langer, U.: Domain decomposition solvers for frequency-domain finite element equations. In: Huang, Y., Kornhuber, R., Widlund, O., Xu, J. (eds.) Domain Decomposition in Science and Engineering XIX. Lecture Notes in Computational Science and Engineering, vol. 78, pp. 301–308. Springer, New York (2011)

[10] Gould, N.I.M., Simoncini, V.: Spectral analysis of saddle point matrices with indefinite leading blocks. SIAM J. Matrix. Anal. Appl. 31(3), 1152–1171 (2010)

[11] Greenbaum, A.: Iterative methods for solving linear systems. Frontiers in Applied Mathematics, vol. 17. SIAM, Philadelphia (1997)

[12] Gunzburger, M.D., Manservisi, S.: Analysis and approximation of the velocity tracking problem for Navier-Stokes flows with distributed control. SIAM J. Numer. Anal. 37(5), 1481–1512 (2000)

[13] Hackbusch, W.: Fast solution of elliptic control problems. J. Optimization Theory Appl. 31, 565–581 (1980)

[14] Higham, N.J.: Functions of Matrices. Theory and Computation. SIAM, Philadelphia (2008)

[15] Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE constraints. Mathematical Modelling: Theory and Applications, vol. 23. Springer, Dordrecht (2009)

[16] Kollmann, M., Kolmbauer, M., Langer, U., Wolfmayr, M., Zulehner, W.: A finite element solver for a multiharmonic parabolic optimal control problem. NuMa-Report 2011-10, Institute of Computational Mathematics, Johannes Kepler University Linz (2011)

[17] Kolmbauer, M., Kollmann, M.: A preconditioned MINRES solver for time-periodic parabolic optimal control problems. NuMa-Report 2011-06, Institute of Computational Mathematics, Johannes Kepler University Linz (2011)

[18] Krendl, W., Simoncini, V., Zulehner, W.: Stability estimates and structural spectral properties of saddle point problems. DK-Report 2012-03, Doctoral Program Computational Mathematics, Johannes Kepler University Linz (2012)

[19] Kuznetsov, Y.A.: Efficient iterative solvers for elliptic finite element problems on nonmatching grids. Russ. J. Numer. Anal. Math. Model. 10(3), 187–211 (1995)

[20] Langer, U., Queck, W.: On the convergence factor of Uzawa's algorithm. J. Comput. Appl. Math. 15, 191–202 (1986)

[21] Lions, J.L.: Optimal Control of Systems Governed by Partial Differential Equations. Springer, Heidelberg (1971)

[22] Lions, J.L., Magenes, E.: Non-homogeneous Boundary Value Problems and Applications, vol. I. Die Grundlehren der mathematischen Wissenschaften, Band 181. Springer, Heidelberg (1972)

[23] Mardal, K.A., Winther, R.: Uniform preconditioners for the time dependent Stokes problem. Numer. Math. 98(2), 305–327 (2004)

[24] Mardal, K.A., Winther, R.: Uniform preconditioners for the time dependent Stokes problem. Numer. Math. 103(1), 171–172 (2006)

[25] Mardal, K.A., Winther, R.: Preconditioning discretizations of systems of partial differential equations. Numer. Linear Algebra Appl. 18, 1–40 (2011)

[26] Murphy, M.F., Golub, G.H., Wathen, A.J.: A note on preconditioning for indefinite linear systems. SIAM J. Sci. Comput. 21(6), 1969–1972 (2000)

[27] Olshanskii, M.A., Reusken, A.: On the convergence of a multigrid method for linear reaction-diffusion problems. Computing 65(3), 193–202 (2000)

[28] Olshanskii, M.A., Peters, J., Reusken, A.: Uniform preconditioners for a parameter dependent saddle point problem with application to generalized Stokes interface equations. Numer. Math. 105(1), 159–191 (2006)

[29] Paige, C.C., Saunders, M.A.: Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal. 12, 617–629 (1975)

[30] Rees, T., Wathen, A.J.: Preconditioning iterative methods for the optimal control of the Stokes equations. SIAM J. Sci. Comput. 33, 2903–2926 (2011)

[31] Rees, T., Dollar, H.S., Wathen, A.J.: Optimal solvers for PDE-constrained optimization. SIAM J. Sci. Comput. 32, 271–298 (2010)

[32] Rusten, T., Winther, R.: A preconditioned iterative method for saddle-point problems. SIAM J. Matrix Anal. Appl. 13, 887–904 (1992)

[33] Schöberl, J., Zulehner, W.: Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems. SIAM J. Matrix Anal. Appl. 29, 752–773 (2007)

[34] Silvester, D., Wathen, A.J.: Fast iterative solution of stabilised Stokes systems. Part II: Using general block preconditioners. SIAM J. Numer. Anal. 31(5), 1352–1367 (1994)

[35] Stoll, M., Wathen, A.J.: Combination preconditioning and the Bramble-Pasciak$^+$ preconditioner. SIAM J. Matrix Anal. Appl. 30(2), 582–608 (2008)

[36] Tröltzsch, F.: Optimal Control of Partial Differential Equations. Theory, Methods and Applications. Graduate Studies in Mathematics, vol. 112. American Mathematical Society, Providence (2010)

[37] Wathen, A.J., Silvester, D.: Fast iterative solution of stabilised Stokes systems. Part I: Using simple diagonal preconditioners. SIAM J. Numer. Anal. 30, 630–649 (1993)

[38] Xu, J., Zikatanov, L.: Some observations on Babuška and Brezzi theories. Numer. Math. 94(1), 195–202 (2003)

[39] Zulehner, W.: Analysis of iterative methods for saddle point problems: a unified approach. Math. Comp. 71, 479–505 (2002)

[40] Zulehner, W.: Nonstandard norms and robust estimates for saddle point problems. SIAM J. Matrix Anal. Appl. 32, 536–560 (2011)

Convergence Orders of Iterative Methods for Nonlinear Eigenvalue Problems

Gerhard Unger

Abstract. The convergence analysis of iterative methods for nonlinear eigenvalue problems is in the most cases restricted either to algebraically simple eigenvalues or to polynomial eigenvalue problems. In this paper we consider two classical methods for general holomorphic eigenvalue problems, namely the nonlinear generalized Rayleigh quotient iteration (NGRQI) and the augmented Newton method. The analysis of the convergence order of both methods is based on the representation of the eigenvalues as poles of the resolvent. This approach was already chosen for the analysis of the NGRQI by Langer in [19] for a more general setting where such a representation of the eigenvalues had to be assumed. The convergence orders of both methods depend on the order which an eigenvalue has as pole of the resolvent. Both methods exhibit a local quadratic convergence order for semi–simple eigenvalues. For defective eigenvalues in general only a local linear convergence is possible. In numerical experiments the theoretical results are confirmed.

1 Introduction

We consider nonlinear eigenvalue problems for holomorphic matrix–valued functions $T : \Lambda \to \mathbb{C}^{n\times n}$, where $\Lambda \subset \mathbb{C}$ is a domain, of the following form: Find $\lambda \in \Lambda$ and $v \in \mathbb{C}^n \setminus \{0\}$ such that

$$T(\lambda)v = 0. \tag{1}$$

A pair (λ, v) which fulfills the eigenvalue problem (1) is called eigenpair, λ is called eigenvalue, and v eigenvector. In the following we assume that $\det T(\cdot) \not\equiv 0$ on Λ. A comprehensive review about applications of nonlinear eigenvalue problems and their numerical solution is presented in [5, 21].

Gerhard Unger
Institut für Numerische Mathematik, TU Graz, Steyrergasse 30, 8010 Graz, Austria
e-mail: gerhard.unger@tugraz.at

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 217–237.
springerlink.com

In this paper we focus on the numerical analysis of iterative methods for the refinement of already existing approximations of eigenpairs of nonlinear eigenvalue problems. For the approximate localization of eigenvalues recently the contour integral method was proposed [4, 6]. This method allows to find approximations of all eigenvalues in a given domain and of related eigenvectors without requiring initial approximations. In the case of general holomorphic eigenvalue problems the combination of the contour integral method with refinement methods is a reasonable approach.

For the refinement of approximations of eigenpairs usually iterative methods are proposed which are based on the solution of a sequence of linearized problems, such as the nonlinear generalized Rayleigh quotient iteration (NGRQI) [17, 18, 19, 30], augmented Newton–type methods [2, 3, 23, 29, 24, 34], the method of successive linear problems [28], or methods which utilize QR decompositions [7, 16, 20]. For comprehensive reviews of these methods we refer to [21, 29]. A block Newton method was presented in [15] which enables a simultaneous refinement of the complete eigenspace of one or several eigenvalues and which is robust with respect to defective eigenvalues.

In this paper we review and extend the convergence analysis for the nonlinear generalized Rayleigh quotient iteration and for the augmented Newton method. As main theoretical tool for our analysis we use the theory of eigenvalue problems for holomorphic Fredholm operator–valued functions [9, 14, 22]. This concept provides an extension of the theory of linear eigenvalue problems and it is based on the characterization of the eigenvalues as poles of the resolvent. We will see that the convergence order of both methods depends on the order which an eigenvalue has as pole of the resolvent.

The NGRQI was introduced for polynomial eigenvalue problems in [18] as generalization of the two–sided Rayleigh quotient iteration [25], and local quadratic convergence was shown for semi–simple eigenvalues. In [17], this result was extended to polynomial eigenvalue problems in arbitrarily dimensional Hilbert spaces. For general holomorphic eigenvalue problems the NGRQI was analyzed in [19] under the assumption that the eigenvalues can be locally characterized as poles of the resolvent. For eigenvalue problems with holomorphic matrix–valued functions $T : \Lambda \to \mathbb{C}^{n\times n}$ this assumption is always satisfied if $\det T(\cdot) \not\equiv 0$ on Λ, see, e.g., [8]. The convergence rate of the NGRQI depends on the order which an eigenvalue has as pole of the resolvent [19]. Since semi–simple eigenvalues are simple poles of the resolvent, the NGRQI converges locally quadratically. In the defective case local linear convergence is obtained. A modified version of the NGRQI was suggested and analyzed in [30, 32] where for algebraically simple eigenvalues local quadratic convergence was shown [30].

There are several variants of augmented Newton–type methods for nonlinear eigenvalue problems available. The classical approach is to apply Newton's method to the system of nonlinear equations consisting of the nonlinear eigenvalue problem and in addition of a normalization condition for the eigenvector [2, 3, 34]. This approach is called the augmented Newton method or the inverse iteration for nonlinear eigenvalue problems. Different modifications of this approach were suggested in order to reduce the cost of the computations [23, 29], and to increase the convergence rate of the iteration [24, 28, 29]. In all of the mentioned publications the convergence results for the augmented Newton–type methods are restricted to algebraically simple eigenvalues. In this case, the derivative of the augmented form of the nonlinear eigenvalue problem is non–singular at an eigenvalue and the classical argument of the proof for the convergence of Newton's method can be applied. If the algebraic multiplicity of an eigenvalue is not simple, this argumentation is not possible since the derivative of the augmented form is singular at the eigenvalue. In this paper we show that for semi–simple eigenvalues local quadratic convergence is still obtained, where we utilize that the eigenvalues are simple poles of the resolvent. An analysis of the defective case will not be done in this paper. In [10], the convergence factors for semi–simple and for double defective eigenvalues are analyzed where, however, the convergence itself is not established but assumed. For the case of defective double eigenvalues it is shown that if the augmented Newton method converges, then the convergence order is linear [10].

This paper is organized as follows. In the next section we outline the concept of eigenvalue problems for holomorphic matrix–valued functions and present some important characterizations and properties of the eigenvalues, the eigenvectors, and the resolvent. In Sect. 3 we review the derivation and the convergence analysis of the NGRQI and discuss the conditioning of the arising linear systems. The augmented Newton method is analyzed in Sect. 4 where it is shown that it converges locally quadratically in the case of semi–simple eigenvalues. Moreover, different modifications of the augmented Newton method are discussed. Finally, we present some numerical experiments which support the theoretical results.

In this paper we will use $(\cdot,\cdot)$ as standard inner product, i.e., $(x,y) := y^H x$ for all $x,y \in \mathbb{C}^n$, which induces the Euclidean norm $\|x\| := \sqrt{(x,x)}$. The norm for matrices will always be the spectral norm.

2 Basics of Holomorphic Eigenvalue Problems

In this section we introduce the notation and properties of eigenvalue problems for holomorphic Fredholm operator–valued functions [14, 22]. Here we restrict our presentation to the case of matrix–valued functions. We denote by $\sigma(T)$ the set of all eigenvalues of T in the domain Λ, and by $\rho(T) = \Lambda \setminus \sigma(T)$ the resolvent set. Recall that we assume $\det T(\cdot) \not\equiv 0$ on Λ which implies that the resolvent set $\rho(T)$ is not empty. The dimension of the null space $\ker T(\lambda)$ of an eigenvalue λ is called the geometric multiplicity of λ. An ordered collection of vectors $v_{0,1}, v_{0,2}, \ldots, v_{0,m}$ in $\mathbb{C}^n$ is a Jordan chain of λ of length m if $v_{0,1}$ is an eigenvector corresponding to λ and if

$$\sum_{j=0}^{k-1} \frac{1}{j!} T^{(j)}(\lambda) v_{0,k-j} = 0 \quad \text{for } k = 1, \ldots, m \tag{2}$$

is satisfied, where $T^{(j)}$ is the j–th derivative. The maximal length of a Jordan chain of an eigenvalue λ is denoted by $\varkappa(T,\lambda)$. An eigenvalue λ is called semi–simple if the maximal length of a Jordan chain of λ is one. If in addition the geometric multiplicity of an eigenvalue is one, then it is called an algebraically simple eigenvalue. This definition of an algebraically simple eigenvalue is equivalent to the common one that $\det T'(\lambda) \neq 0$ [14, Prop. A.6.4]. Other equivalent definitions of the multiplicities are possible by using the Smith form [14] or by using root functions [14, 22].

The first result shows that the resolvent $T(\cdot)^{-1} : \Lambda \setminus \sigma(T) \to \mathbb{C}^{n\times n}$ can be represented as a meromorphic function where the eigenvalues are the poles. The order of the poles coincides with the maximal length of the Jordan chains of the eigenvalues.

Theorem 1. [8, Cor. 8.4] *Let $\Lambda \subset \mathbb{C}$ be open and let $T : \Lambda \to \mathbb{C}^{n\times n}$ be a holomorphic matrix–valued function with* $\det T(\cdot) \not\equiv 0$. *Then, every eigenvalue λ of T is isolated, i.e., there exists some neighborhood U of λ such that $U \setminus \{\lambda\} \subset \rho(T)$. Moreover, the resolvent admits a representation as*

$$T(\mu)^{-1} = \sum_{k=-r}^{-1} (\mu-\lambda)^k B_k + F(\mu), \quad \mu \in U \setminus \{\lambda\}, \tag{3}$$

with $B_{-r} \neq 0$, where $r = \varkappa(T,\lambda)$ and $F : \Lambda \to \mathbb{C}^{n\times n}$ is holomorphic.

A characterization of the matrices B_k of the principal part of the resolvent (3) in terms of generalized eigenvectors of T and of the adjoint matrix function T^H provides the Theorem of Keldysh [13], [14, Thm. A.10.2]. The adjoint function $T^H : \{\lambda : \overline{\lambda} \in \Lambda\} \to \mathbb{C}^{n\times n}$ is defined by

$$T^H(\lambda) := (T(\overline{\lambda}))^H.$$

An eigenvector $w \in \mathbb{C}^n$ corresponding to an eigenvalue $\overline{\lambda}$ of the eigenvalue problem for the adjoint function T^H is also called left eigenvector corresponding to the eigenvalue λ of the eigenvalue problem for the function T, since

$$T^H(\overline{\lambda})w = 0 \quad \Leftrightarrow \quad T(\lambda)^H w = 0 \quad \Leftrightarrow \quad w^H T(\lambda) = 0.$$

A triple (λ, v, w) with $T(\lambda)v = 0$ and $w^H T(\lambda) = 0$ is called eigentriple of the eigenvalue problem for the function T.

We cite now the Theorem of Kelydsh for semi–simple eigenvalues which will be needed in the following. For the general version we refer to [14, Thm. A.10.2].

Theorem 2. [14, Thm. A.10.1] *Let the assumptions of Theorem 1 be satisfied. Suppose that $\lambda \in \sigma(T)$ is semi–simple and that $\{v^1,\ldots,v^J\}$ is a basis of the eigenspace* $\ker T(\lambda)$. *Then there exists a unique basis $\{w^1,\ldots,w^J\}$ of* $\ker T^H(\overline{\lambda})$ *such that in some neighborhood U of λ*

$$T(\mu)^{-1} = \sum_{j=1}^{J} \frac{1}{\mu-\lambda} v^j (w^j)^H + F(\mu), \quad \mu \in U \setminus \{\lambda\}, \tag{4}$$

where $F : U \to \mathbb{C}^{n\times n}$ is holomorphic. Moreover, the biorthogonality relation

$$\frac{1}{\mu-\lambda}(T(\mu)v^k, w^j) = \delta_{kj} + \mathcal{O}(\mu-\lambda) \quad \textit{as } \mu \to \lambda \tag{5}$$

holds for $k, j = 1,\dots,J$.

From the representation (4) of the resolvent and the biorthogonality relation (5) some important properties for the derivative $T'(\lambda)$ and the eigenvectors follow. These results are needed later for the analysis of the augmented Newton method.

Corollary 1. *Let the assumptions of Theorem 1 be satisfied. Suppose that λ is a semi–simple eigenvalue and let*

$$\{v^1,\dots,v^J\} \quad \textit{and} \quad \{w^1,\dots,w^J\}$$

be a basis of the eigenspaces $\ker T(\lambda)$ and $\ker T^H(\overline{\lambda})$, respectively, such that the resolvent $T(\mu)^{-1}$ admits the representation (4). Then:

i. For $k, j = 1,\dots,J$ we have

$$(T'(\lambda)v^k, w^j) = \delta_{kj}. \tag{6}$$

ii.

$$\sum_{j=1}^{J}(T'(\lambda)v, w^j)v^j = v \quad \textit{for all } v \in \ker T(\lambda). \tag{7}$$

iii. For F as given in (4),

$$T(\lambda)F(\lambda)T'(\lambda)v^k = 0, \quad k = 1,\dots,J,$$

holds.

Proof. i. If we insert $T(\lambda)v^k = 0$ in (5), then we get

$$\frac{1}{\mu-\lambda}([T(\mu)-T(\lambda)]v^k, w^j) = \delta_{kj} + \mathcal{O}(\lambda-\mu) \quad \text{as } \mu \to \lambda.$$

The assertion follows now by taking the limit $\mu \to \lambda$.
ii. The identity (7) is an immediate consequence of (6).
iii. By using (4), we have

$$F(\mu) = T(\mu)^{-1} - \sum_{j=1}^{J}\frac{1}{\mu-\lambda}v^j(w^j)^H.$$

Multiplying by $T(\mu)$ and adding $\sum_{j=1}^{J} \frac{1}{\mu-\lambda} T(\lambda) v^j (w^j)^H = 0$, this gives

$$\begin{aligned} T(\mu)F(\mu) &= I_n - \sum_{j=1}^{J} \frac{1}{\mu-\lambda}[T(\mu)-T(\lambda)] v^j (w^j)^H \\ &\to I_n - \sum_{j=1}^{J} T'(\lambda) v^j (w^j)^H \quad \text{as } \mu \to \lambda. \end{aligned}$$

With i. we obtain $T(\lambda)F(\lambda)T'(\lambda)v^k = 0$ for $k = 1,\ldots,J$. □

3 Nonlinear Generalized Rayleigh Quotient Iteration

Lancaster introduced in [18] a generalization of the Rayleigh quotient for polynomial matrix–valued functions T,

$$R(\lambda, v, w) = \lambda - \frac{(T(\lambda)v, w)}{(T'(\lambda)v, w)}, \tag{8}$$

where $\lambda \in \mathbb{C}$ and $v, w \in \mathbb{C}^n$, and denoted it nonlinear generalized Rayleigh quotient. Since it is a generalization of the two-sided Rayleigh quotient it is also called two-sided nonlinear Rayleigh quotient. The corresponding iteration

$$\lambda_{i+1} = \lambda_i - \frac{(T(\lambda_i)v_i, w_i)}{(T'(\lambda_i)v_i, w_i)}, \tag{9}$$

where v_i and w_i are the solutions of

$$T(\lambda_i)v_i = a \quad \text{and} \quad T(\lambda_i)^* w_i = b \tag{10}$$

for given vectors $a, b \in \mathbb{C}^n$, Lancaster called nonlinear generalized Rayleigh quotient iteration (NGRQI). In the case that $(T'(\lambda)v, w) \neq 0$ at an eigentriple (λ, v, w), the generalized Rayleigh quotient $R(\lambda, v, w)$ is stationary [18, 31], i.e., the first order terms in the perturbation expansion of $R(\lambda + \delta\lambda, v + \delta v, w + \delta w)$ are zero. If λ is a semi-simple eigenvalue and v is a corresponding eigenvector, then, by Corollary 1, there always exists an corresponding left eigenvector w such that $(T'(\lambda)v, w) \neq 0$. A local quadratic convergence order of the NGRQI was shown for polynomial eigenvalue problems in the case of semi–simple eigenvalues in [18] where it is utilized that

$$\inf_{v \in \ker T(\lambda)} \left\| \frac{T(\lambda + \delta\lambda)^{-1} a}{\|T(\lambda + \delta\lambda)^{-1} a\|} - v \right\| \leq \mathscr{O}(\delta\lambda), \tag{11}$$

which can be deduced by the representation (4) of $T(\lambda + \delta\lambda)^{-1}$ close to an eigenvalue λ.

In the following we review the analysis of the NGRQI for general holomorphic matrix–valued functions as it is given by Langer [19], where the defective case is covered, too. In this approach the NGQRI is traced back to Newton's method for a scalar function ψ which has the eigenvalues as zeros. Defining

$$\psi(\mu) = \frac{1}{(T(\mu)^{-1}a, b)}, \quad \mu \notin \sigma(T), \tag{12}$$

and by using the representation of

$$\frac{d}{d\mu} T(\mu)^{-1} = -T(\mu)^{-1} T'(\mu) T(\mu)^{-1},$$

see, e.g. [12, p. 32], we have

$$\begin{aligned} \frac{\psi(\lambda_i)}{\psi'(\lambda_i)} &= \frac{(T(\lambda_i)^{-1}a, b)}{(T(\lambda_i)^{-1}T'(\lambda_i)T(\lambda_i)^{-1}a, b)} = \frac{(a, [T(\lambda_i)^{-1}]^H b)}{(T'(\lambda_i)T(\lambda_i)^{-1}a, [T(\lambda_i)^{-1}]^H b)} \\ &= \frac{(T(\lambda_i)v_i, w_i)}{(T'(\lambda_i)v_i, w_i)}. \end{aligned}$$

Hence, the NGRQI is Newton's method applied to ψ. For the analysis of the convergence of the NGQRI, we investigate the properties of the function ψ. Let $\lambda \in \sigma(T)$ be an eigenvalue, then, by Theorem 1, there exists a neighborhood U of λ such that

$$T(\mu)^{-1} = \sum_{k=-r}^{\infty} (\mu - \lambda)^k B_k, \quad \mu \in U \setminus \{\lambda\},$$

with $B_{-r} \neq 0$, where $r = \varkappa(T, \lambda)$. The function ψ is well defined in a neighborhood $U_1 \subset U \setminus \{\lambda\}$ of λ if

$$(B_{-r}a, b) \neq 0.$$

This assumption is fulfilled if

$$a \not\perp \ker T^H(\overline{\lambda}) \quad \text{and} \quad b \not\perp \ker T(\lambda) \tag{13}$$

which follows from the representation of B_{-r} in terms of the eigenvectors, see (4) for the semi-simple case and [14, Thm. A.10.2] for the general case. In the following we will assume that (13) holds for a and b. Under this assumption, the function ψ is holomorphic on U_1 and allows the series expansion

$$\psi(\mu) = \frac{(\mu - \lambda)^r}{(B_{-r}a, b)} - (\mu - \lambda)^{r+1} \frac{(B_{-r+1}a, b)}{(B_{-r}a, b)^2} + \mathscr{O}\left((\mu - \lambda)^{r+2}\right).$$

This shows that ψ can be holomorphically extended for $\mu = \lambda$, and that λ is a zero of ψ with multiplicity r. By using the Banach fixed point theorem, the following general convergence result follows immediately.

Theorem 3 ([19, Satz 3])**.** *Let $\lambda \in \sigma(T)$ and let ψ be defined by (12) with $a, b \in \mathbb{C}^n$ such that assumption (13) is fulfilled. Let $s \in \mathbb{N}$ with $s \leq r = \varkappa(T, \lambda)$. Then, there exists a $\delta > 0$ such that the iteration*

$$\lambda_{i+1} = \lambda_i - s \frac{\psi(\lambda_i)}{\psi'(\lambda_i)} \quad \textit{for } i = 0, 1, 2, \ldots \tag{14}$$

converges for any $\lambda_0 \in U_\delta(\lambda) \setminus \{\lambda\}$ to λ. If $s = r$, then the convergence is quadratic and

$$\frac{\lambda_{i+1} - \lambda}{(\lambda_i - \lambda)^2} \to \frac{(B_{-r+1}a, b)}{r(B_{-r}a, b)} \quad as\ i \to \infty.$$

If $s < r$, then the convergence is linear and

$$\frac{\lambda_{i+1} - \lambda}{\lambda_i - \lambda} \to \frac{r - s}{r} \quad as\ i \to \infty.$$

In general $\varkappa(T, \lambda)$ is not known a priori. In this situation, $s = 1$ should be chosen for the iteration (14), which yields the NGRQI and which ensures at least local linear convergence. For a semi–simple eigenvalue the choice $s = 1$ gives quadratic convergence since $\varkappa(T, \lambda) = 1$.

In each iteration step of the NGRQI the vectors v_i and w_i have to be computed by solving the linear systems (10). These vectors are approximations of a right and a left eigenvector, respectively, corresponding to the eigenvalue λ. Here, we cite an approximation result for the vector v_i. A similar result holds for w_i.

Lemma 1 ([19, Satz 4])**.** *Let the sequence $\{\lambda_i\}_{i=0}^{\infty}$ defined by (14) and let the assumption of Theorem 3 be fulfilled. Let $v_i = T(\lambda_i)^{-1}a$ for $i \in \mathbb{N}_0$, then there exists an $i_0 \in \mathbb{N}$ such that*

$$\inf_{v \in \ker T(\lambda)} \left\| v - \frac{v_i}{\|v_i\|} \right\| \le c\,|\lambda_i - \lambda|$$

holds for all $i \ge i_0$, where $c > 0$ is a constant which is independent of i.

Simplifications of the NGRQI can be obtained for Hermitian and complex symmetric eigenvalue problems. For Hermitian eigenvalue problems, the choice $a = b$ as input vectors for the NGRQI is suggested since then only one system of linear equations has to be solved in each iteration step. A similar simplification of the iteration is obtained in the case of complex symmetric eigenvalue problems as it was also proposed in [29] for the Rayleigh functional iteration. Provided that a and b are chosen such that $b = \overline{a}$, the solution w_i of the second equation in (10) is the complex conjugate of the solution v_i of the first equation since

$$T(\lambda_i)v_i = a \quad \Leftrightarrow \quad T(\lambda_i)^\top v_i = a \quad \Leftrightarrow \quad \overline{T(\lambda_i)^\top v_i} = \overline{a} \quad \Leftrightarrow \quad T(\lambda_i)^H \overline{v_i} = \overline{a}.$$

The systems of linear equations (10) which have to be solved in each iteration step of the NGRQI get ill-conditioned close to an eigenvalue. Therefore, in [30] an equivalent bordered version of the NGRQI was suggested where instead of the systems of linear equations (10) the bordered systems

$$\begin{pmatrix} T(\lambda_i) & a \\ b^H & 0 \end{pmatrix} \begin{pmatrix} s_i \\ \mu_i \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha \end{pmatrix}, \quad \begin{pmatrix} T(\lambda_i)^H & b \\ a^H & 0 \end{pmatrix} \begin{pmatrix} t_i \\ \nu_i \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha \end{pmatrix}, \tag{15}$$

are considered, where $\alpha \in \mathbb{R} \setminus \{0\}$ is a scaling factor. In the case of a simple eigenvalue λ, the bordered systems (15) are uniquely solvable if a and b are sufficient good approximations of the right and left eigenvector of λ even for $\lambda_i = \lambda$.

For multiple eigenvalues this result is not true. Hence, one may question if the theoretical convergence results are visible for multiple eigenvalues in practical computations, too. In the following, we want to consider the original variant (9) of the NGRQI and show that as for the inverse iteration for linear eigenvalue problems [26], the error which is made due to the ill–conditioning of (10) points in the direction of the desired eigenspace. To simplify matters, we restrict our analysis to the case of a semi-simple eigenvalue λ. Let us assume that instead of the linear system $T(\lambda_i)v_i = a$ the perturbed system

$$(T(\lambda_i)+\varepsilon E)\tilde{v}_i = a \tag{16}$$

is solved, where ε is small and $\|E\| = 1$. For the estimate of the error $v_i - \tilde{v}_i$, we consider the representation of the inverse of the perturbed matrix

$$(T(\lambda_i)+\varepsilon E)^{-1} = \sum_{n=0}^{\infty} \varepsilon^n (-T(\lambda_i)^{-1}E)^n T(\lambda_i)^{-1}$$

which one gets by using $(T(\lambda_i)+\varepsilon E)^{-1} = (I+\varepsilon T(\lambda_i)^{-1}E)^{-1}T(\lambda_i)^{-1}$ and by utilizing the Neumann series expansion. Then, we obtain

$$v_i - \tilde{v}_i = T(\lambda_i)^{-1}a - (T(\lambda_i)+\varepsilon E)^{-1}a = -\sum_{n=1}^{\infty} \varepsilon^n (-T(\lambda_i)^{-1}E)^n T(\lambda_i)^{-1}a. \tag{17}$$

Essential for the error is the term $T(\lambda_i)^{-1}E$ which exhibits in each coefficient in the series (17). By Theorem 2, we have

$$T(\lambda_i)^{-1}E = \sum_{j=1}^{J} \frac{1}{\lambda_i - \lambda} v^j (w^j)^H E + F(\lambda_i)E,$$

with appropriate basis vectors $\{v_1,\dots,v_J\}$ of $\ker T(\lambda)$ and $\{w_1,\dots,w_J\}$ of $\ker T^H(\lambda)$ and with an appropriate holomorphic function F. This shows that the error which is made by the perturbed system (16) close to an eigenvalue is dominated by a linear combination of eigenvectors of the desired eigenvalue. A similar statement holds for the error of the perturbation of the adjoint problem $T(\lambda_i)^H w_i = b$. From these facts together with (11) and the stationarity of the generalized nonlinear Rayleigh quotient for semi-simple eigenvalues, it follows that a quadratic convergence order is also achieved for multiple semi-simple eigenvalues in practical computations, see Example 5.1. For defective eigenvalues a similar analysis as above for the error arising from the perturbed linear systems can be done. In Example 5.2, the theoretical convergence results are confirmed for defective eigenvalues in practical computations, too.

In [18, 30, 32] a modified version of the NGRQI was proposed where the vectors a and b are updated in every iteration step by w_i and v_i, respectively. In general a higher convergence order cannot be achieved by this modification [18]. Local quadratic convergence of this variant of the NGRQI was shown in [30].

4 Augmented Newton Method

One of the classical approaches for the numerical solution of the nonlinear eigenvalue problem (1) is to apply Newton's method to the augmented system

$$F(v,\lambda) := \begin{pmatrix} T(\lambda)v \\ d^H v - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{18}$$

where the second equation is a normalization constraint with some chosen vector $d \in \mathbb{C}^n \setminus \{0\}$ for which it is assumed that it is not orthogonal to the eigenspace $\ker T(\lambda)$. In many publications [3, 24, 28, 29, 35], this approach was analyzed for algebraically simple eigenvalues. Utilizing that the derivative of the augmented form

$$F'(v,\lambda) = \begin{pmatrix} T(\lambda) & T'(\lambda)v \\ d^H & 0 \end{pmatrix} \tag{19}$$

for an eigenpair (λ, v) of an algebraically simple eigenvalue is non–singular, the standard convergence result for Newton's method can be applied to show local quadratic convergence. Different modifications are proposed in order to reduce the cost of the computations [23, 29] and to increase the convergence rate of the iteration [24, 28, 29]. However, the convergence analysis in all of these publications are restricted to algebraically simple eigenvalues and it is based on the regularity of the derivative of the augmented form (19), which is for a multiple eigenvalue not regular anymore.

For our theoretical analysis it is suitable to write the augmented Newton method in the form as given in Algorithm 1.

Algorithm 1. Augmented Newton method

```
Input: $\lambda_0, v_0, d$ such that $d^H v_0 = 1$
for $i = 0, 1, 2, \dots$ until convergence do
    solve $T(\lambda_i)s_{i+1} = T'(\lambda_i)v_i$ for $s_{i+1}$
    $\lambda_{i+1} = \lambda_i - (d^H v_i)/(d^H s_{i+1})$
    $v_{i+1} = s_{i+1}/d^H s_{i+1}$
end for
```

In the following we will present a convergence analysis of Algorithm 1 for semi–simple eigenvalues. First we want to derive an error estimate for the new eigenvector approximation of Algorithm 1 for a given approximation (μ, z) of an eigenpair.

Lemma 2. *Let $\lambda \in \sigma(T)$ be a semi–simple eigenvalue and let $d \in \mathbb{C}^n$ such that $d \not\perp \ker T(\lambda)$. Then, there exist $\varepsilon > 0$, $\tau > 0$, and $c_\lambda > 0$ such that for all*

$$z \in \left\{ y \in \mathbb{C}^n : \|y\| = 1,\ \min_{v \in \ker T(\lambda)} \|y - v\| \leq \varepsilon \right\}$$

and for all μ *with* $0 < |\mu - \lambda| \leq \tau$ *the estimate*

$$\min_{v \in \ker T(\lambda)} \left\| \frac{T(\mu)^{-1}T'(\mu)z}{d^H T(\mu)^{-1}T'(\mu)z} - v \right\| \leq c_\lambda |\mu - \lambda| \big(|\mu - \lambda| + \|z - v_z\| \big) \tag{20}$$

holds, where v_z *is the best approximation of* z *in* $\ker T(\lambda)$.

Proof. Let $\lambda \in \sigma(T)$ be a semi–simple eigenvalue and let $\{v^1, \ldots, v^J\}$ be a basis of the eigenspace $\ker T(\lambda)$. By Theorem 2, there exists a neighborhood U of λ such that $T(\mu)^{-1}$ admits a representation by

$$T(\mu)^{-1} = (\mu - \lambda)^{-1} B_{-1} + F(\mu), \qquad \mu \in U \setminus \{\lambda\}, \tag{21}$$

with a holomorphic function $F : U \to \mathbb{C}^{n \times n}$ and with $B_{-1} = \sum_{j=1}^{J} v^j (w^j)^H$, where $\{w^1, \ldots, w^J\}$ is an appropriate basis of $\ker T^H(\overline{\lambda})$. Let us choose $\tau_* > 0$ such that the closed disk $\overline{U}_{\tau_*}(\lambda)$ with center λ and radius τ_* is a subset of U. For $z \in \mathbb{C}^n$ with $\|z\| = 1$ and $\mu \in \overline{U}_{\tau_*}(\lambda) \setminus \{\lambda\}$, we denote by $s(\mu, z)$ the solution of

$$T(\mu)s(\mu, z) = T'(\mu)z.$$

Using the representation (21) of $T(\mu)^{-1}$, we can write

$$s(\mu, z) = T(\mu)^{-1}T'(\mu)z = \frac{1}{(\mu - \lambda)} B_{-1}T'(\mu)z + F(\mu)T'(\mu)z \tag{22}$$

and

$$\frac{s(\mu, z)}{d^H s(\mu, z)} = \frac{B_{-1}T'(\mu)z + (\mu - \lambda)F(\mu)T'(\mu)z}{d^H B_{-1}T'(\mu)z + (\mu - \lambda)d^H F(\mu)T'(\mu)z} \tag{23}$$

for $\mu \in \overline{U}_{\tau_*}(\lambda) \setminus \{\lambda\}$. We first show that the denominator of (23) is well defined in a neighborhood of λ provided that z is sufficiently close to the eigenspace $\ker T(\lambda)$. Let $v_z \in \mathbb{C}^n$ be the best approximation of z in $\ker T(\lambda)$ and let

$$\delta z := z - v_z.$$

Using the Taylor series expansion of $T'(\mu)$ in λ and (7), i.e., $B_{-1}T'(\lambda)v = v$ for all $v \in \ker T(\lambda)$, we can write

$$B_{-1}T'(\mu)z = v_z + (\mu - \lambda)B_{-1}T''(\lambda)v_z + R_1(\mu - \lambda)v_z + B_{-1}T'(\mu)\delta z \tag{24}$$

with $\frac{\|R_1(\mu - \lambda)\|}{\mu - \lambda} \to 0$ as $\mu \to \lambda$. By assumption we have $d \not\perp v_z$, therefore there exist $c_1 > 0$, $0 < \tau < \tau_*$ and $\varepsilon > 0$ such that

$$\left| d^H B_{-1}T'(\mu)z + (\mu - \lambda)d^H F(\mu)T'(\mu)z \right| \geq c_1 > 0 \tag{25}$$

for all μ with $0 < |\mu - \lambda| \leq \tau$ and all $z \in \mathbb{C}^n$ with $\|\delta z\| \leq \varepsilon$.

Next, we want to specify the components of the vector $s(\mu, z)$ which lie in the eigenspace $\ker T(\lambda)$. Due to the construction of B_{-1}, the vector $B_{-1}T'(\mu)z$ is an element of $\ker T(\lambda)$. The Taylor series expansion of $F(\mu)T'(\mu)$ in λ gives

$$F(\mu)T'(\mu)(v_z+\delta z) = F(\lambda)T'(\lambda)v_z \\ +(\mu-\lambda)\frac{d}{d\mu}[F(\mu)T'(\mu)]_{|\mu=\lambda}v_z + R_2(\mu-\lambda)v_z + F(\mu)T'(\mu)\delta z \quad (26)$$

with $\frac{\|R_2(\mu-\lambda)\|}{\mu-\lambda} \to 0$ as $\mu \to \lambda$. By Corollary 1.iii., $F(\lambda)T'(\lambda)v_z$ is an element of $\ker T(\lambda)$. Thus, we get from (23) with (26) that

$$\inf_{v\in\ker T(\lambda)} \left\| \frac{s(\mu,z)}{d^H s(\mu,z)} - v \right\| \\ \leq |\mu-\lambda| \frac{\left\| (\mu-\lambda)\frac{d}{d\mu}[F(\mu)T'(\mu)]_{|\mu=\hat{\mu}}v_z + R_2(\mu-\lambda)v_z + F(\mu)T'(\mu)\delta z \right\|}{|d^H B_{-1}T'(\mu)z + (\mu-\lambda)d^H F(\mu)T'(\mu)z|}.$$

Since $\frac{d}{d\mu}[F(\mu)T'(\mu)]$ and $F(\mu)T'(\mu)$ are bounded in $\overline{U}_\tau(\lambda)$, and because of $\frac{\|R_2(\mu-\lambda)\|}{\mu-\lambda} \to 0$ as $\mu \to \lambda$, and by (25), we conclude that there exists a constant $c_\lambda > 0$ such that

$$\inf_{v\in\ker T(\lambda)} \left\| \frac{s(\mu,z)}{d^H s(\mu,z)} - v \right\| \leq c_\lambda |\mu-\lambda| \big(|\mu-\lambda| + \|\delta z\|\big)$$

holds for all z with $\|\delta z\| \leq \varepsilon$ and for all μ with $0 < |\mu-\lambda| \leq \tau$. □

In the next theorem we show the local quadratic convergence of Algorithm 1 for semi–simple eigenvalues.

Theorem 4. *Let $\lambda \in \sigma(T)$ be a semi–simple eigenvalue and let $d \in \mathbb{C}^n$ such that $d \not\perp \ker T(\lambda)$. Then, there exist an $\varepsilon_0 > 0$ and a $\tau_0 > 0$ such that for all*

$$v_0 \in \left\{ z \in \mathbb{C}^n : \|z\| = 1, \min_{v\in\ker T(\lambda)} \|z-v\| \leq \varepsilon_0 \right\}$$

and for λ_0 with $0 < |\lambda_0 - \lambda| \leq \tau_0$ the sequence $\{\lambda_i\}_{i\in\mathbb{N}_0}$ defined by Algorithm 1 converges to λ. Moreover, there exists a constant $c > 0$ such that

$$|\lambda_{i+1}-\lambda| + \|\delta v_{i+1}\| \leq c|\lambda_i - \lambda| \left(|\lambda_i-\lambda| + \|\delta v_i\| \right)$$

holds for all $i \in \mathbb{N}_0$, where δv_i and δv_{i+1} are the best approximation errors of v_i and v_{i+1}, respectively, in $\ker T(\lambda)$.

Proof. Let $\lambda \in \sigma(T)$ be a semi–simple eigenvalue, then we can choose $\tau_* > 0$ as in the proof of Lemma 2 such that for all $\mu \in \overline{U}_{\tau_*}(\lambda) \setminus \{\lambda\}$ the resolvent $T(\mu)^{-1}$ admits a representation by

$$T(\mu)^{-1} = (\mu-\lambda)^{-1}B_{-1} + F(\mu).$$

For $z \in \mathbb{C}^n$ let $v_z \in \mathbb{C}^n$ again be the best approximation of z in $\ker T(\lambda)$ and let

$$\delta z := z - v_z.$$

Then, using (22) and (24), we can write

$$\frac{d^H z}{d^H T(\mu)^{-1} T'(\mu) z} = (\mu - \lambda) \frac{d^H v_z + d^H \delta z}{d^H v_z + \alpha(\mu, z)} = (\mu - \lambda) \left(1 + \frac{d^H \delta z - d^H a(\mu, z)}{d^H v_z + d^H a(\mu, z)} \right) \tag{27}$$

where

$$a(\mu, z) = (\mu - \lambda) B_{-1} T''(\lambda) v_z + R_1(\mu - \lambda) v_z + B_{-1} T'(\mu) \delta z + (\mu - \lambda) F(\mu) T'(\mu) z$$

with $\frac{\|R_1(\mu-\lambda)\|}{\mu-\lambda} \to 0$ as $\mu \to \lambda$. Hence, for sufficiently small $\tau > 0$ and $\varepsilon > 0$ there exists a constant $\tilde{c} > 0$ such that

$$\begin{aligned} \left| \mu - \lambda - \frac{d^H z}{d^H T(\mu)^{-1} T'(\mu) z} \right| &= \left| (\mu - \lambda) \frac{d^H \delta z - d^H a(\mu, z)}{d^H v_z + d^H a(\mu, z)} \right| \\ &\leq \tilde{c} |\mu - \lambda| \big(|\mu - \lambda| + \|\delta z\| \big) \end{aligned} \tag{28}$$

holds for z with $\|\delta z\| \leq \varepsilon$ and for $\mu \in \overline{U}_\tau(\lambda) \setminus \{\lambda\}$. Let us assume that τ and ε are chosen sufficiently small such that also the estimate (20) holds, i.e.,

$$\inf_{v \in \ker T(\lambda)} \left\| \frac{T(\mu)^{-1} T'(\mu) z - v}{d^H T(\mu)^{-1} T'(\mu) z} \right\| \leq c_\lambda \, |\mu - \lambda| \, \big(|\mu - \lambda| + \|\delta z\| \big). \tag{29}$$

Let us consider now Algorithm 1. The first update λ_1 for an initial pair (λ_0, v_0) is given by

$$\lambda_1 = \lambda_0 - \frac{d^H v_0}{d^H T(\lambda_0)^{-1} T'(\lambda_0) v_0}.$$

Using (28) and (29) we get

$$|\lambda_1 - \lambda| \leq \tilde{c} |\lambda_0 - \lambda| \big(|\lambda_0 - \lambda| + \|\delta v_0\| \big)$$

and

$$\min_{v \in \ker T(\lambda)} \|v_1 - v\| \leq c_\lambda \, |\lambda_0 - \lambda| \, \big(|\lambda_0 - \lambda| + \|\delta v_0\| \big)$$

for v_0 with $\|\delta v_0\| \leq \varepsilon$ and for $\lambda_0 \in \overline{U}_\tau(\lambda) \setminus \{\lambda\}$. Choose $\varepsilon_0 > 0$ and $\tau_0 > 0$ such that

$$\varepsilon_0 < \min\left\{ \frac{1}{3\tilde{c}}, \varepsilon \right\} \quad \text{and} \quad \tau_0 < \min\left\{ \frac{\varepsilon_0}{c_\lambda(\tau + \varepsilon_0)}, \frac{1}{3\tilde{c}}, \tau \right\}.$$

This implies that $\eta := \tilde{c}(\tau_0 + \varepsilon_0) < \tilde{c}\left(\frac{1}{3\tilde{c}} + \frac{1}{3\tilde{c}}\right) < 1$. Hence, we get for λ_0 with $0 < |\lambda_0 - \lambda| \leq \tau_0$ and for v_0 with $\|\delta v_0\| \leq \varepsilon_0$ the estimates

$$\begin{aligned} |\lambda_1 - \lambda| &\leq \tilde{c} |\lambda_0 - \lambda| \big(|\lambda_0 - \lambda| + \|\delta v_0\| \big) \leq |\lambda_0 - \lambda| \, \eta < \tau_0 < \tau, \\ \|\delta v_1\| &\leq c_\lambda \tau_0 (\tau_0 + \varepsilon_0) \leq c_\lambda \tau_0 (\tau + \varepsilon_0) \leq \varepsilon_0 < \varepsilon. \end{aligned}$$

Thus, by induction we obtain with (28) and (29)

$$|\lambda_{i+1}-\lambda| \leq \tilde{c}\,|\lambda_i-\lambda|\left(|\lambda_i-\lambda|+\|\delta v_i\|\right) \leq \eta^i\,|\lambda_0-\lambda| \to 0 \tag{30}$$

as $i \to \infty$ and

$$\min_{v\in\ker T(\lambda)} \|v_{i+1}-v\| \leq c_\lambda\,|\lambda_i-\lambda|\left(|\lambda_i-\lambda|+\|\delta v_i\|\right), \tag{31}$$

which prove the assertions. □

From the error estimates (30) and (31) it follows that the sequence of the errors $\{(\lambda_i-\lambda,\delta v_i)\}_{i\in\mathbb{N}}$ converges quadratically to $0\in\mathbb{C}^{n+1}$ since

$$\begin{aligned}\max\{|\lambda_{i+1}-\lambda|,\|\delta v_{i+1}\|\} &\leq c\left(|\lambda_i-\lambda|^2+|\lambda_i-\lambda|\,\|\delta v_i\|\right)\\ &\leq 2c\max\left\{|\lambda_i-\lambda|^2,\|\delta v_i\|^2\right\}.\end{aligned}$$

Hence, we have shown that Newton's method applied to the augmented form (18) exhibits also for multiple semi-simple eigenvalues a local quadratic convergence order as for algebraically simple eigenvalues.

For defective eigenvalues, numerical examples indicate that the convergence of Algorithm 1 is linear. However, to the best of our knowledge, a proof of this conjecture is not available. By considering the representation of the principal part of the resolvent in terms of generalized eigenvectors, which provides the Theorem of Keldysh, a proof of this conjecture might be possible. In [10], the convergence factor for double defective eigenvalues was analyzed. Provided that the sequence $\{\lambda_i\}_{i\in\mathbb{N}}$ converges to an eigenvalue, it was shown that the convergence is linear and that the convergence factor is $1/2$.

As for the NGRQI, also for the augmented Newton method the linear system which has to be solved in each iteration step is ill–conditioned close to a multiple eigenvalue. However, again this only slightly affects the iteration in practical computations. In numerical experiments still a quadratic convergence order for semi–simple eigenvalues with multiple geometric multiplicity is obtained, see Example 5.1.

4.1 Nonlinear Rayleigh Quotients and Functionals

For the improvement of the convergence behavior of the augmented Newton method different variants of nonlinear Rayleigh quotients and Rayleigh functionals were suggested for the update of λ_{i+1} in step 4 of Algorithm 1. A comprehensive analysis of the approximation properties of nonlinear Rayleigh quotients and functionals is presented in [29, 31]. One distinguishes between one-sided and two-sided Rayleigh quotients and functionals. The use of a two-sided nonlinear Rayleigh quotient or functional as update λ_{i+1} in Algorithm 1 requires in addition to the approximation s_{i+1} of a right eigenvector an approximation t_{i+1} of a left eigenvector, which is usually chosen by $t_{i+1} = T(\lambda_i)^{-H} T'(\lambda_i)^H w_i$, where w_i is a normalized approximation of a left eigenvector in the previous step of the algorithm. In [28], the two-sided nonlinear Rayleigh quotient (8) is suggested for the update, i.e.,

$$\lambda_{i+1} = \lambda_i - \frac{w_{i+1}^H T(\lambda_i) v_{i+1}}{w_{i+1}^H T'(\lambda_i) v_{i+1}}.$$

This modification of Algorithm 1 yields for generalized linear eigenvalue problems a local cubic convergence for semi-simple eigenvalues [1], however, for genuine nonlinear eigenvalue problems numerical examples show that a cubic convergence order in general is not possible, but rather a quadratic one.

In [29] the two-sided Rayleigh functional $p(v_{i+1}, w_{i+1})$, which is implicitly defined by

$$w_{i+1}^H T(p(v_{i+1}, w_{i+1})) v_{i+1} = 0,$$

was suggested for the update λ_{i+1}. This approach is a generalization of the method presented in [27] for real symmetric nonlinear eigenvalue problems, where as update λ_{i+1} the one-sided Rayleigh functional $q(v_{i+1}) := p(v_{i+1}, v_{i+1})$ is used. If (λ, v, w) is an eigentriple with $w^H T'(\lambda) v \neq 0$, then p is locally uniquely defined and has the stationarity property [31]

$$p(v + h_1, w + h_2) - \lambda = \mathscr{O}((\|h_1\| + \|h_2\|)^2).$$

Setting $\lambda_{i+1} = p(v_{i+1}, w_{i+1})$ in Algorithm 1, where $v_{i+1} = s_{i+1}/\|s_{i+1}\|$ and $w_{i+1} = t_{i+1}/\|t_{i+1}\|$, yields the improved convergence rates [29, Sect. 4.2]

$$|\lambda_{i+1} - \lambda| = \mathscr{O}(|\lambda_i - \lambda|^2 \|v_i - v\| \|w_i - w\|),$$
$$\|v_{i+1} - v\| = \mathscr{O}(\|v_i - v\|^2 \|w_i - w\|), \quad \|w_{i+1} - w\| = \mathscr{O}(\|w_i - w\|^2 \|v_i - v\|),$$

when assuming that λ is an algebraically simple eigenvalue. The update λ_{i+1} by the Rayleigh functional $p(v_{i+1}, w_{i+1})$ requires the solution of the scalar nonlinear equation

$$g_{i+1}(\mu) := w_{i+1}^H T(\mu) v_{i+1} = 0$$

which cannot be solved directly for genuine nonlinear eigenvalue problems. Using Newton's method for $g_{i+1}(\mu) = 0$ with initial value $\mu_0 = \lambda_i$ gives for the first update

$$\mu_1 = \lambda_i - \frac{w_{i+1}^H T(\lambda_i) v_{i+1}}{w_{i+1}^H T'(\lambda_i) v_{i+1}},$$

which is the two-sided nonlinear Rayleigh quotient $R(\lambda_i, v_{i+1}, w_{i+1})$. This shows the close relation of different variants of Newton's method for nonlinear eigenvalue problems.

4.2 Simplified Newton Method

A simplified version of the augmented Newton method was proposed in [23] where a fixed shift $\sigma \in \mathbb{C}$ is used for all iteration steps for the update of the eigenvector approximation. This method is called residual inverse iteration and can be described as follows: For a given approximation (λ_i, v_i) of an eigenpair (λ, v) first the update λ_{i+1} is determined as solution of the scalar equation

$$e^H T(\sigma)^{-1} T(\lambda_{i+1}) v_i = 0, \tag{32}$$

where σ is an approximation of the desired eigenvalue λ and $e \in \mathbb{C}^n$ is a normalization vector. The update v_{i+1} is given by

$$v_{i+1} = \alpha_{i+1}[v_i - T(\sigma)^{-1} T(\lambda_{i+1}) v_i],$$

where the normalization constant α_{i+1} is chosen such that $d^H v_{i+1} = 1$ for some $d \in \mathbb{C}^n \setminus \{0\}$. The advantage of the residual inverse iteration is that the matrix $T(\sigma)$ is kept fixed during the iteration and can hence be factorized in advance. If λ is an algebraically simple eigenvalue and v is a corresponding eigenvector with $d^H v = 1$, then the residual inverse iteration converges for all (λ_0, v_0) sufficiently close to (λ, v) and the following error estimates hold [23]

$$\frac{\|v_{i+1} - v\|}{\|v_i - v\|} = \mathscr{O}(|\sigma - \lambda|) \quad \text{and} \quad |\lambda_{i+1} - \lambda| = \mathscr{O}(\|v_i - v\|).$$

A detailed analysis of convergence factors of the residual inverse iteration was presented in [11] which includes also a discussion of the convergence of semi-simple and defective eigenvalues. For the later, examples show that the residual inverse iteration may not convergence.

A two-sided variant of the residual inverse iteration was suggested in [29] where the two-sided nonlinear Rayleigh functional $p(v_i, w_i)$ is used as update λ_{i+1} instead of (32). This modification requires an approximation w_i of a left eigenvector w, but it yields an improved convergence rate for the eigenvalue approximation [29], namely

$$|\lambda_{i+1} - \lambda| = \mathscr{O}(\|v_i - v\| \|w_i - w\|).$$

5 Examples

In this section we present numerical examples for the NGRQI and the augmented Newton method for the approximation of multiple semi–simple eigenvalues and of defective eigenvalues.

Example 5.1

We first consider the nonlinear eigenvalue problem of the form

$$T(\lambda) = e^{\lambda} F D(\lambda) G - \lambda I,$$

where $D(\lambda) = \operatorname{diag}(\sin\lambda, e^{\lambda} - 1, 3, \ldots, n)$ with $n = 100$, and where $F, G \in \mathbb{R}^{n\times n}$ are taken as random with full rank. $\lambda = 0$ is a semi–simple eigenvalue of T with geometric multiplicity 2. We observe a quadratic convergence order of the NGRQI and of the augmented Newton method for the approximation of $\lambda = 0$, see Fig. 1. This confirms the theoretical results of Theorem 3 and of Theorem 4.

Example 5.2

We consider the delay eigenvalue problem [10, Example 2] of the form

$$T(\lambda) = -\lambda I + A_0 + A_1 e^{-\lambda},$$

where

$$A_0 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -b_3 & -b_2 & -b_1 \end{pmatrix},$$

and

$$a_1 = \frac{2}{5}\frac{65\pi + 32}{8 + 5\pi}, \qquad a_2 = \frac{9\pi^2(13 + 5\pi)}{8 + 5\pi}, \; a_3 = \frac{324}{5}\frac{\pi^2(5\pi + 4)}{8 + 5\pi},$$
$$b_1 = \frac{260\pi + 128 + 225\pi 2}{80 + 50\pi}, \; b_2 = \frac{45\pi 2}{8 + 5\pi}, \qquad b_3 = \frac{81(\pi^2(40\pi + 32 + 25\pi^2)}{80 + 50\pi}.$$

This eigenvalue problem has a defective eigenvalue $\lambda = 3\pi i$ with $\varkappa(T, \lambda) = 2$, see [10]. According to Theorem 3, the NGRQI converges linearly with the convergence factor $1/2$ which is confirmed by the computations, see Fig. 1. Also the augmented Newton method converges linearly with a convergence factor of $1/2$, as it was already demonstrated in [10].

Example 5.3

Here we consider the boundary element discretization of the Dirichlet Laplacian eigenvalue problem

$$-\Delta u(x) = \lambda^2 u(x) \quad \text{for}\, x \in \Omega, \quad u(x) = 0 \quad \text{for}\, x \in \partial\Omega, \tag{33}$$

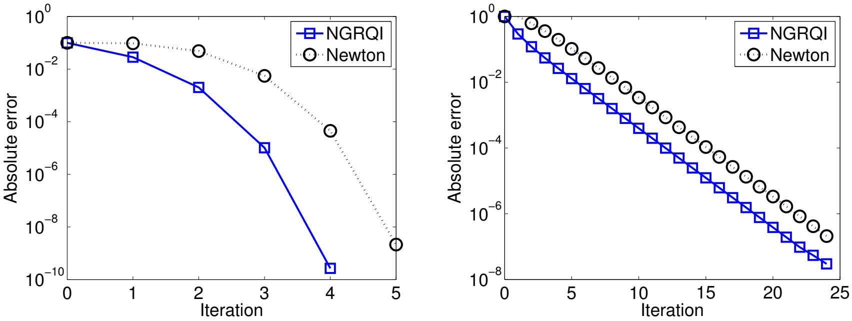


Fig. 1 Convergence of the eigenvalue approximation of Example 5.1 (*left plot*) and Example 5.2 (*right plot*).

where $\Omega = \{x \in \mathbb{R}^3 : \|x\| < 1\}$ is the unit ball. This eigenvalue problem can be represented in terms of the boundary integral equation [33]

$$\frac{1}{4\pi}\int\limits_{\partial\Omega} \frac{e^{i\lambda|x-y|}}{|x-y|}\frac{\partial}{\partial n_y}u(y)ds_y = 0 \quad \text{for } x \in \partial\Omega \tag{34}$$

which yields a nonlinear eigenvalue problem. $\frac{\partial}{\partial n}u$ denotes the normal derivative of u on the boundary $\partial\Omega$. For the approximation of the eigenvalue problem (34) the boundary $\partial\Omega$ is approximated by n_L planar triangles τ_ℓ for different discretization levels L. As ansatz space for the eigenfunctions we chose the space of piecewise constant functions with respect to the boundary triangulation. The Galerkin discretization of the eigenvalue problem (34) takes then the form

$$T_L(\lambda^{(L)})v^{(L)} = 0, \tag{35}$$

where

$$T_L(\lambda^{(L)})[k,\ell] := \frac{1}{4\pi}\int\limits_{\tau_\ell}\int\limits_{\tau_k} \frac{e^{i\lambda^{(L)}|x-y|}}{|x-y|}ds_y ds_x \quad \text{for } k,\ell = 1,\dots,n_L.$$

In order to get coarse approximations for the eigenpairs we have used the contour integral method [6] on the discretization level $L = 2$ with $n_L = 80$ boundary elements. The refinement of the two smallest eigenvalues is done on level $L = 5$ with $n_L = 5120$ boundary elements. On the continuous level the smallest eigenvalue λ_1 is algebraically simple whereas the second smallest one λ_2 is semi-simple with multiplicity three [33]. By the discretization the semi-simple eigenvalue splits into three clustered algebraically simple eigenvalues where the maximal distance between the eigenvalues is smaller than 10^{-5}. As reference solution we have chosen the solution of the contour integral method with 200 quadrature nodes for the contour integration. The convergence behavior of the NGRQI and of the augmented Newton method is given in Fig. 2 where for both methods and for each eigenvalue a quadratic convergence order can be observed.

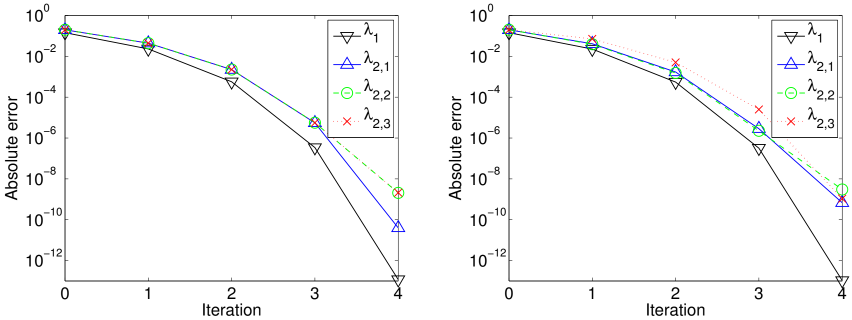


Fig. 2 Numerical results for Example 5.3. Obtained eigenvalue approximations in the course of the NGRQI (*left plot*) and the augmented Newton method (*right plot*) on level $L = 5$.

The results of this example demonstrate that both methods exhibit in practice also in the case of clustered simple eigenvalues a quadratic convergence behavior.

6 Conclusions

In this paper we have reviewed and extended the convergence results of two classical iterative methods for holomorphic eigenvalue problems which are usually restricted either to algebraically simple eigenvalues or to polynomial eigenvalue problems. We have considered the nonlinear generalized Rayleigh quotient iteration and the augmented Newton method which can be used for a more accurate approximation of eigenpairs. In our convergence analysis we have utilized the representation of the resolvent as a meromorphic matrix–valued function which has the eigenvalues as poles. The convergence order of both methods depends on the order of the poles of the resolvent. For semi–simple eigenvalues, which are simple poles of the resolvent, local quadratic convergence has been shown for both methods. This result is novel for the augmented Newton method. In the case of defective eigenvalues, the nonlinear generalized Rayleigh quotient iteration exhibits a local linear convergence order.

The computational cost of both methods differ for non–Hermitian and non–symmetric eigenvalue problems. If the augmented Newton method is used, only one linear system has to be solved per iteration step whereas for the nonlinear generalized Rayleigh quotient iteration additionally an adjoint problem has to be solved.

Numerical experiments support the theoretical convergence results even then when the linear systems get ill–conditioned in the case of semi–simple eigenvalues with multiple geometric multiplicity and in the case of defective eigenvalues.

References

[1] Amiraslani, A., Lancaster, P.: Rayleigh quotient algorithms for nonsymmetric matrix pencils. Numer. Algorithms 51(1), 5–22 (2009)

[2] Andrew, A.L., Chu, K.E., Lancaster, P.: On the numerical solution of nonlinear eigenvalue problems. Computing 55(2), 91–111 (1995)

[3] Anselone, P.M., Rall, L.B.: The solution of characteristic value-vector problems by Newton's method. Numer. Math. 11, 38–45 (1968)

[4] Asakura, J., Sakurai, T., Tadano, H., Ikegami, T., Kimura, K.: A numerical method for nonlinear eigenvalue problems using contour integrals. JSIAM Letters 1, 52–55 (2009)

[5] Betcke, T., Higham, N.J., Mehrmann, V., Schröder, C., Tisseur, F.: NLEVP: A collection of nonlinear eigenvalue problems. MIMS EPrint, Manchester Institute for Mathematical Sciences, University of Manchester (2008)

[6] Beyn, W.J.: An integral method for solving nonlinear eigenvalue problems. Linear Algebra Appl. (2011), doi:10.1016/j.laa.2011.03.030

[7] Dai, H., Lancaster, P.: Numerical methods for finding multiple eigenvalues of matrices depending on parameters. Numer. Math. 76(2), 189–208 (1997)

[8] Gohberg, I., Goldberg, S., Kaashoek, M.A.: Classes of Linear Operators, vol. I. Birkhäuser, Basel (1990)

[9] Gohberg, I.C., Sigal, E.I.: An operator generalization of the logarithmic residue theorem and Rouché's theorem. Math. USSR-Sb. 13, 603–625 (1971)

[10] Jarlebring, E.: Convergence factors of Newton methods for nonlinear eigenvalue problems. Linear Algebra Appl. (2010), doi:10.1016/j.laa. 2010.08.045

[11] Jarlebring, E., Michiels, W.: Analyzing the convergence factor of residual inverse iteration. BIT 51(4), 937–957 (2011)

[12] Kato, T.: Perturbation Theory for Linear Operators. Springer, New York (1966)

[13] Keldysh, M.V.: On the eigenvalues and eigenfunctions of certain classes of non–selfadjoint operators. Dokl. Akad. Nauk SSSR 77, 11–14 (1951)

[14] Kozlov, V., Maz'ya, V.: Differential Equations with Operator Coefficients with Applications to Boundary Value Problems for Partial Differential Equations. Springer Monographs in Mathematics. Springer, Berlin (1999)

[15] Kressner, D.: A block Newton method for nonlinear eigenvalue problems. Numer. Math. 114(2), 355–372 (2009)

[16] Kublanovskaya, V.N.: On an approach to the solution of the generalized latent value problem for λ-matrices. SIAM J. Numer. Anal. 7, 532–537 (1970)

[17] Kummer, H.: Zur praktischen Behandlung nichtlinearer Eigenwertaufgaben abgeschlossener linearer Operatoren. Mitt. Math. Sem. Giessen 62 (1964)

[18] Lancaster, P.: A generalized Rayleigh quotient iteration for lambda-matrices. Arch. Rational Mech. Anal. 8, 309–322 (1961)

[19] Langer, U.: Untersuchungen zum Kummerschen Verfahren zur numerischen Behandlung nichtlinearer Eigenwertaufgaben. Beiträge Numer. Math. 6, 97–110 (1977)

[20] Li, R.C.: Compute multiple nonlinear eigenvalues. J. Comput. Math. 10(1), 1–20 (1992)

[21] Mehrmann, V., Voss, H.: Nonlinear eigenvalue problems: a challenge for modern eigenvalue methods. GAMM Mitt. Ges. Angew. Math. Mech. 27(2), 121–152 (2005)

[22] Mennicken, R., Möller, M.: Non-Self-Adjoint Boundary Eigenvalue Problems. North-Holland Publishing Co., Amsterdam (2003)

[23] Neumaier, A.: Residual inverse iteration for the nonlinear eigenvalue problem. SIAM J. Numer. Anal. 22(5), 914–923 (1985)

[24] Osborne, M.R.: Inverse iteration, Newton's method and nonlinear eigenvalue problems. In: The Contributions of Dr. J. H. Wilkinson to Numerical Analysis, London. Symp. Proc. Series, vol. 19, pp. 21–53 (1978)

[25] Ostrowski, A.M.: On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. IV. (Generalized Rayleigh quotient for nonlinear elementary divisors). Arch. Rational Mech. Anal. 3, 341–347 (1959)

[26] Peters, G., Wilkinson, J.H.: Inverse iteration, ill-conditioned equations and Newton's method. SIAM Rev. 21(3), 339–360 (1979)

[27] Rothe, K.: Lösungsverfahren für nichtlineare Matrixeigenwertaufgaben mit Anwendungen auf die Ausgleichselementmethode. Verlag an der Lottbek, Hamburg (1989)

[28] Ruhe, A.: Algorithms for the nonlinear eigenvalue problem. SIAM J. Numer. Anal. 10, 674–689 (1973)

[29] Schreiber, K.: Nonlinear eigenvale problems: Newton-type methods and nonlinear Rayleigh functionals. PhD thesis, TU Berlin (2008)

[30] Schwetlick, H., Schreiber, K.: A primal-dual Jacobi-Davidson-like method for nonlinear eigenvalue problems. ePrints 2006-011, Institut für Mathematik, TU Berlin (2006)

[31] Schwetlick, H., Schreiber, K.: Nonlinear Rayleigh functionals. Linear Algebra Appl. (2010), doi:10.1016/j.laa.2010.6.048

[32] Spence, A., Poulton, C.: Photonic band structure calculations using nonlinear eigenvalue techniques. J. Comput. Phys. 204(1), 65–81 (2005)

[33] Steinbach, O., Unger, G.: Convergence analysis of a Galerkin boundary element method for the Dirichlet Laplacian eigenvalue problem. SIAM J. Numer. Anal. 50, 710–728 (2012)

[34] Unger, H.: Nichtlineare Behandlung von Eigenwertaufgaben. Z. Angew. Math. Mech. 30, 281–282 (1950)

[35] Voss, H.: Numerical methods for sparse nonlinear eigenvalue problems. In: Proc. of the XVth Summer School on Software and Algorithms of Numerical Mathematics, Hejnice, Czech Republic, pp. 133–160. University of West Bohemia (2004)

Sensitivity Analysis for Maxwell Eigenvalue Problems in Industrial Applications

Stefan Reitzinger, Markus Wabro, and Sabine Zaglmayr

Abstract. In this paper we focus on the sensitivity analysis of Maxwell's eigenvalue problem, where the derivatives of the eigenvalues are calculated with respect to design parameters (i.e., material or geometrical parameters). Utilizing the adjoint approach the derivatives can be calculated at almost no additional cost. The challenge consists in the computation of the required derivatives (i.e., derivatives of bilinear forms with respect to the design parameters) from a higher order, curved finite element discretization. Numerical studies show the application for a real life electromagnetic filter application where the sensitivities of the eigenvalues give a better insight into the characteristics of the underlying filter. The benefit is apparent if the adjoint method is compared to a standard finite difference approach.

1 Introduction

In recent years the term "virtual prototyping" has become more and more important, because development cycles can be reduced and consequently the production costs are decreased. Reliable and efficient numerical methods are the core requirements for such a CAE tool in order to get the required results. In most cases these CAE tools have an underlying parameterized model; changing the (e.g., geometrical or material) parameters allows to study "what if" scenarios of real life applications.

In this paper we concentrate on 3D electromagnetic (EM) simulations, where typically systems of up to millions of unknowns have to be solved for several excitations or eigenvalues and with different parameter sets. Due to the fact that the computational cost grows exponentially for a brute force approach, this would be limited to a few design parameters only. Thus we want to extract as much information as possible from one single EM simulation. Since we are mainly interested

Stefan Reitzinger · Markus Wabro · Sabine Zaglmayr
CST AG, Bad Nauheimerstr. 19, 64289 Darmstadt, Germany
e-mail: {Stefan.Reitzinger,Markus.Wabro,Sabine.Zaglmayr}@cst.com

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 239–256.
springerlink.com

in “functional values” of results of EM simulations (e.g. eigenvalues, transfer functions, objective function, etc.) a linearization of the functional can be applied if the derivative with respect to the design parameters can be computed additionally. This is also of interest for EM-simulation-based optimization or neuronal networks, because with the help of sensitivity information the required large number of EM simulations can be reduced significantly (see e.g. [25, 15]). To speed up the enormous computational effort of an EM simulation, several parallelization strategies (e.g. OpenMP, MPI, GPU) or distributed computing techniques can be applied in addition (c.f., [20]).

In the following we will stay within the Maxwell eigenvalue problem and collect different techniques (e.g., formulation, discretization, solution of the algebraic eigenvalue problem, computation of derivatives) such that the above requirements on a CAE tool are fulfilled. Real life applications range from acceleration physics, electromagnetic filter analysis to photonic band gap structures. The literature on the Maxwell eigenvalue problem is quite large in the mathematical community, e.g., [2, 13, 21, 22, 28, 41], but also in the engineering world, e.g., [6, 26, 43]. In recent times the mathematical analysis of the Maxwell eigenvalue problem with respect to an appropriate discretization was analyzed by several authors, e.g., [9, 10, 12, 13, 21, 28], and especially higher order shape functions with curved elements for the finite element (FE) technique were under consideration, e.g., [2, 17, 34, 41]. Topics which (to our knowledge) are not analyzed yet are eigenvalue problems with arbitrary frequency dependent material distributions (e.g., Debye material model, open boundary conditions) and the a posteriori error estimation (for a general overview on error estimates see, e.g., [42]).

Large scale (lossy) eigenvalue problems and the solution thereof are discussed, e.g., in [6, 26, 43]. For the solution of the algebraic eigenvalue problem the shift invert Arnoldi or Lanczos, LOBPCG, and the Jacobi Davidson are used (see [7, 33, 37] for an overview). A structure preserving Arnoldi method was proposed in [27].

Sensitivity analysis is well established in the structural mechanics community and for network-simulation, see [36] and [8], respectively. For 3D EM simulation a discovery of that method starts with [3, 5] where the adjoint sensivitiy method is used for time and frequency domain applications on structured grids and with the FE method. An error estimator is proposed in [29, 40] for the derivative of a functional (especially for the transfer function). Knowing the sensitivity of some system response, one can start to set up a first order parameterized model which is valid in the surrogate of the nominal value. Applications are discussed for optimization (see [25]) or for the calculation of yield values in order to apply design centering techniques (see [19]).

The rest of the paper is organized as follows: First we review possible formulations of the eigenvalue problem, discuss the discretization and a splitting with respect to the kernel of the eigenvalue problem, which will be used to speed up the final algorithm. The next section is devoted to the calculation of the derivatives. Finally, we present some numerical results and give a short summary and outlook.

2 Problem Formulation, Discretization and Solution Strategy

2.1 Problem Formulation

Let $\Omega \subset \mathbb{R}^3$ be an open bounded connected Lipschitz domain. We consider the time-harmonic electromagnetic resonant cavity problem

$$\begin{aligned} \operatorname{curl}\mathbf{E} + i\omega\mu\mathbf{H} &= \mathbf{0} \quad \text{in } \Omega, && (1)\\ \operatorname{curl}\mathbf{H} - i\omega\varepsilon\mathbf{E} &= \mathbf{0} \quad \text{in } \Omega, && (2)\\ \operatorname{div}(\mu\mathbf{H}) &= 0 \quad \text{in } \Omega, && (3)\\ \operatorname{div}(\varepsilon\mathbf{E}) &= 0 \quad \text{in } \Omega && (4) \end{aligned}$$

for the electric field **E** and the magnetic field **H** with mixed PEC (perfect electric conductor) and PMC (perfect magnetic conductor) boundary conditions

$$\begin{aligned} \mathbf{E} \times \mathbf{n} &= \mathbf{0} \quad \text{on } \Gamma_1 \subset \partial\Omega, && (5)\\ \mathbf{H} \times \mathbf{n} &= \mathbf{0} \quad \text{on } \Gamma_2 := \partial\Omega / \Gamma_1. && (6) \end{aligned}$$

The electric permittivity tensor ε and the permeability tensor μ are assumed to be real-valued and positive-definite a.e. in Ω; Γ_1 is assumed to have a positive measure.

We start with a short summary of the most common variational formulations of the resonant cavity problem. As usual we define the function spaces

$$\begin{aligned} H^1(\Omega) &:= \{w \in L_2(\Omega) \,:\, \nabla w \in L_2(\Omega)\},\\ H^1_{0,\Gamma}(\Omega) &:= \{w \in H^1(\Omega) \,:\, w = 0 \text{ on } \Gamma\},\\ H(\operatorname{curl},\Omega) &:= \{\mathbf{v} \in (L_2(\Omega))^3 \,:\, \operatorname{curl}\mathbf{v} \in (L_2(\Omega))^3\},\\ H_{0,\Gamma}(\operatorname{curl},\Omega) &:= \{\mathbf{v} \in H(\operatorname{curl},\Omega) \,:\, \mathbf{v} \times \mathbf{n} = 0 \text{ on } \Gamma\} \end{aligned}$$

with outer normal vector **n** on $\Gamma \subset \partial\Omega$. Moreover, we refer to the $L_2(\Omega)$-inner product by $(\mathbf{u},\mathbf{v}) := \int_\Omega \mathbf{v}^\top \mathbf{u}\,dx$.

The electric formulation, or briefly the **E**-formulation, originates from eliminating the magnetic field in the cavity problem (1)-(6). In the assumed loss-free case a (constrained) linear eigenvalue problem for the electric field is implied. In weak form it reads: Find non-trivial solutions $(\omega^2, \mathbf{E}) \in \mathbb{R}_0^+ \times H_{0,\Gamma_1}(\operatorname{curl},\Omega)$ such that

$$\begin{aligned} (\mu^{-1}\operatorname{curl}\mathbf{E}, \operatorname{curl}\mathbf{v}) &= \omega^2 (\varepsilon\mathbf{E}, \mathbf{v}) \qquad && \forall \mathbf{v} \in H_{0,\Gamma_1}(\operatorname{curl},\Omega), && (7)\\ (\varepsilon\mathbf{E}, \nabla w) &= 0 && \forall w \in H^1_{0,\Gamma_1}(\Omega). && (8) \end{aligned}$$

The constraint (8) is the divergence-free condition (4) in weak sense. For non-zero frequencies the magnetic field can be recovered by $\mathbf{H} = i\omega^{-1}\operatorname{curl}\mathbf{E}$.

The magnetic formulation, briefly **H**-formulation, is obtained by initially eliminating the electric field, again leading to a (constrained) linear eigenvalue problem:

Find non-trivial solutions $\omega \neq 0$ and $\mathbf{H} \in H_{0,\Gamma_2}(\mathrm{curl}, \Omega)$ such that

$$\begin{aligned} (\varepsilon^{-1}\mathrm{curl}\,\mathbf{H}, \mathrm{curl}\,\mathbf{v}) &= \omega^2 (\mu\mathbf{H}, \mathbf{v}) && \forall \mathbf{v} \in H_{0,\Gamma_2}(\mathrm{curl}, \Omega) \\ (\mu\mathbf{H}, \nabla w) &= 0 && \forall w \in H^1_{0,\Gamma_2}(\Omega). \end{aligned}$$

Remark 1. A mixed formulation in both the electric and the magnetic field is of interest for lossy problem settings. In those cases the electric as well as the magnetic formulation would generally imply non-linear eigenvalue problems. The mixed formulation in **E** and **H** can be formulated as antisymmetric linear eigenvalue problem, as e.g. in [43]. The antisymmetric structure of the eigenvalue problem can be efficiently exploited in numerical solvers by the structure preserving techniques presented in [27].

The electric and magnetic formulations lead to linear constrained eigenvalue problems of the same structure and type, hence it is sufficient to consider only the electric formulation in the sequel. Here, we distinguish in the solution set between static modes and dynamic modes as follows.

Static modes denote all solutions of the **E**-formulation (7)–(8) corresponding to the frequency $\omega = 0$. The number of static modes depends on the topological properties of the Dirichlet boundary, in fact on the number of non-connected parts of Γ_1. If Γ_1 consists of M non-connected parts $\Gamma_{1,k} \subset \Gamma_1$ with $\Gamma_1 = \bigcup_{k=1}^{M} \Gamma_{1,k}$, the cavity problem has $(M-1)$ static modes $\mathbf{E}_k = \nabla\phi_k$ with $k = 1, ..., M-1$ which are uniquely defined by the potential problem (see e.g., [21, 13]):

Find $\phi_k \in \{\phi \in H^1(\Omega) \,:\, \phi = \delta_{kj} \text{ on } \Gamma_{1,j}, j = 1, \dots, M\}$ such that

$$(\varepsilon\nabla\phi_k, \nabla w) = 0 \qquad \forall w \in H^1_{0,\Gamma_1}(\Omega).$$

For multiply-connected computational domains Ω, there are N additional kernel functions which are not in $\nabla H^1(\Omega)$, where N is the number of handles of Ω. This issue can be treated in the context of cohomology spaces, as done, e.g., in [21]. In the sequel, we assume that Ω has no handles.

Dynamic modes refer to solutions of (7)–(8) corresponding to non-zero frequencies $\omega \neq 0$. For dynamic modes, the divergence-free constraint (8) is redundant, since it is implicitly included in testing the first equation (7) with gradients $\mathbf{v} \in \nabla H^1_{0,\Gamma_1}(\Omega)$. For dynamic modes the **E**-formulation reduces to the linear eigenvalue problem:

Find non-trivial solutions $(\omega^2 \neq 0, \mathbf{E}) \in \mathbb{R}^+ \times H_{0,\Gamma_1}(\mathrm{curl}, \Omega)$ such that

$$(\mu^{-1}\mathrm{curl}\,\mathbf{E}, \mathbf{curl}\,\mathbf{v}) = \omega^2 (\varepsilon\mathbf{E}, \mathbf{v}) \qquad \forall \mathbf{v} \in H_{0,\Gamma_1}(\mathrm{curl}, \Omega).$$

In the following, we concentrate on the dynamic modes.

Finally, we cite the equivalent symmetric mixed formulation of the constraint problem as introduced in [23]: Find non-trivial solutions ω^2, $\mathbf{E} \in H_{0,\Gamma_1}(\mathrm{curl}, \Omega)$ and $\phi \in H^1_{0,\Gamma_1}(\Omega)$ such that

$$(\mu^{-1}\mathrm{curl}\,\mathbf{E}, \mathrm{curl}\,\mathbf{v}) + (\varepsilon\nabla\phi, \mathbf{v}) = \omega^2(\varepsilon\mathbf{E}, \mathbf{v}) \qquad \forall\, \mathbf{v} \in H_{0,\Gamma_1}(\mathrm{curl}, \Omega), \tag{9}$$

$$(\varepsilon\mathbf{E}, \nabla\psi) = 0 \qquad \forall\, \psi \in H^1_{0,\Gamma_1}(\Omega), \tag{10}$$

which is often used for numerical analysis of the Maxwell eigenvalue problem, e.g. in [10].

2.2 Discretization

We follow a conforming Galerkin discretization with higher-order, curved finite elements. As is common practice, we use scalar continuous finite elements for discretizing variables in $H^1(\Omega)$ and vector-valued tangentially continuous elements for discretizing variables in $H(\mathrm{curl})$. The latter, so called edge elements, were initially introduced by Nédélec [30, 31]. Hierarchic versions of arbitrary order as well as hp-versions of these are presented, e.g., in [1, 17, 34, 39].

In the sequel, we denote the finite element subspaces by

$$\mathbb{W}_{hp} \subset H^1(\Omega), \qquad \mathbb{V}_{hp} \subset H(\mathrm{curl}, \Omega),$$

and assume the discrete exactness property (see, e.g., [11, 17, 21])

$$\mathbb{V}_{hp}^{\mathrm{ker}} := \ker(\mathrm{curl}, \mathbb{V}_{hp}) = \nabla\mathbb{W}_{hp} \tag{11}$$

which is easily fulfilled by a compatible choice of the polynomial orders in the used edge element discretization, as given in [1, 17, 30, 31, 34]. The corresponding discrete spaces with zero traces on the Dirichlet boundary are denoted by

$$\overline{\mathbb{W}}_{hp} = \mathbb{W}_{hp} \cap H^1_{0,\Gamma_1}(\Omega), \qquad \overline{\mathbb{V}}_{hp} = \mathbb{V}_{hp} \cap H_{0,\Gamma_1}(\mathrm{curl}, \Omega),$$

moreover, for the restriction of the discrete kernel of the curl operator there holds

$$\overline{\mathbb{V}}_{hp}^{\mathrm{ker}} := \ker(\mathrm{curl}, \overline{\mathbb{V}}_{hp}) = \nabla\mathbb{W}_{hp} \cap H_{0,\Gamma_1}(\mathrm{curl}, \Omega). \tag{12}$$

Discretization of the potential problem. The discrete static modes $\mathbf{E}_h^k := \nabla\phi_{h,k} \in \overline{\mathbb{V}}_{hp}$, $k = 1, .., M-1$, are defined by the discretization of the potential problem (2.1) which reads:

Find $\phi_{h,k} \in \{\phi_h \in \mathbb{W}_{hp} \,:\, \phi_h = \delta_{kj} \text{ on } \Gamma_{1,j}, j = 1, \ldots, M\}$ such that

$$(\varepsilon\nabla\phi_{h,k}, \nabla w_h) = 0 \qquad \forall\, w_h \in \overline{\mathbb{W}}_{hp}. \tag{13}$$

There holds $\overline{\mathbb{V}}_{hp}^{\mathrm{ker}} = \nabla\overline{\mathbb{W}}_{hp}^{+}$ with $\overline{\mathbb{W}}_{hp}^{+} := \overline{\mathbb{W}}_{hp} + \mathrm{span}\{\phi_{h,k} \text{ solves } (13) : 1 \leq k \leq \text{M-1}\}$.

Discretization of Maxwell's eigenvalue problem. The set of discrete dynamic modes is characterized by all eigensolutions corresponding to non-zero discrete frequencies $\omega_h \neq 0$ of the discrete electromagnetic eigenvalue problem:

Find non-trivial solutions $(\omega_h^2, \mathbf{E}_h) \in \mathbb{R}^+ \times \overline{\mathbb{V}}_{hp}$ such that

$$(\mu^{-1}\mathrm{curl}, \mathbf{E}_h, \mathrm{curl}\, \mathbf{v}_h) = \omega_h^2\, (\varepsilon \mathbf{E}_h, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \overline{\mathbb{V}}_{hp}. \tag{14}$$

The discrete eigenspace corresponding to $\omega_h = 0$ coincides with the discrete kernel of the curl operator $\overline{\mathbb{V}}_{hp}^{\mathrm{ker}}$. For compatible FE spaces (with standard approximation properties and discrete compactness property) the discrete Maxwell eigenvalue problem is spectrally correct and free of spurious eigenmodes. For details on this issue see the recent review [10] and the references therein.

For the FE basis $\{\varphi_i\}_{i=1}^N$ of $\overline{\mathbb{V}}_{hp}$ we identify the discrete functions

$$\mathbf{E}_h = \sum_{i=1}^{N} (\underline{\mathbf{E}}_h)_i \varphi_i \in \overline{\mathbb{V}}_{hp}$$

with its coefficient vector $\underline{\mathbf{E}}_h \in \mathbb{R}^N$. Then we obtain that the algebraic linear eigenvalue problem to be solved is given by

$$K\underline{\mathbf{E}}_h = \lambda M \underline{\mathbf{E}}_h, \qquad \lambda := \omega_h^2 \tag{15}$$

with symmetric positive semi-definite stiffness matrix $K \in \mathbb{R}^{N\times N}$ with entries $(K)_{ij} = (\mu^{-1}\mathrm{curl}\, \varphi_i, \mathrm{curl}\, \varphi_j)$ and symmetric positive definite mass matrix $M \in \mathbb{R}^{N\times N}$ with $(M)_{ij} = (\varepsilon\varphi_i, \varphi_j)$. Please note that the discrete eigenvalue λ satisfies the Rayleigh quotient representation

$$\lambda = \frac{\langle K\underline{\mathbf{E}}_h, \underline{\mathbf{E}}_h\rangle}{\langle M\underline{\mathbf{E}}_h, \underline{\mathbf{E}}_h\rangle}. \tag{16}$$

2.2.1 FE Spaces with Explicit Discrete Kernel Splitting

In view of efficient solution strategies, we will use an explicit splitting of the discrete spaces into the kernel of the curl operator and a corresponding completion space (denoted by superscript $*$) of the form

$$\mathbb{V}_{hp} = \mathbb{V}_{hp}^* \oplus \mathbb{V}_{hp}^{\mathrm{ker}}, \quad \text{and resp.} \quad \overline{\mathbb{V}}_{hp} = \overline{\mathbb{V}}_{hp}^* \oplus \overline{\mathbb{V}}_{hp}^{\mathrm{ker}}, \tag{17}$$

where $\oplus$ represents a direct, but not necessarily orthogonal sum. We can practically achieve such a splitting by defining

$$\mathbb{V}_{hp}^* := \mathbb{V}_{h,0}^* \oplus \mathbb{V}_p^*, \qquad \mathbb{V}_{hp}^{\mathrm{ker}} = \nabla \mathbb{W}_{hp}$$

where

i. the low-order completion space $\mathbb{V}_{h,0}^*$ is the co-tree space obtained by a tree/co-tree splitting (see, e.g., [4, 11]) of the zero-th order Nédélec space $\mathbb{V}_{h,0}$,

ii. the higher-order completion space $\mathbb{V}_p^*$ is spanned by the higher-order completion basis introduced in [34] for edge-elements of second type and in [41] for those with approximation properties of first type.

The restriction of all subspaces onto $H_{0,\Gamma_1}(\mathrm{curl},\Omega)$ yields the splitting (17) for $\overline{\mathbb{V}}_{hp}$.

Splitted formulation of the algebraic Maxwell eigenvalue problem. For carrying the properties of the space splitting (17) over to the algebraic level, we agree on the following notation. By the superscript 2 we refer to the space $\overline{\mathbb{V}}_{hp}^{\mathrm{ker}}$, while by the superscript 1 to $\overline{\mathbb{V}}_{hp}^*$. Further on, we assume a renumbering of degrees of freedom such that for all $\mathbf{E}_h = \mathbf{E}_h^1 + \mathbf{E}_h^2 \in \overline{\mathbb{V}}_{hp}^* \oplus \overline{\mathbb{V}}_{hp}^{\mathrm{ker}}$ the coefficient vector becomes

$$\underline{\mathbf{E}}_h = \begin{pmatrix} \underline{\mathbf{E}}_h^1 \\ \underline{\mathbf{E}}_h^2 \end{pmatrix}.$$

Then the algebraic eigenvalue problem (15) is rewritten as

$$\begin{pmatrix} K_{11} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \underline{\mathbf{E}}_h^1 \\ \underline{\mathbf{E}}_h^2 \end{pmatrix} = \lambda \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} \underline{\mathbf{E}}_h^1 \\ \underline{\mathbf{E}}_h^2 \end{pmatrix} \tag{18}$$

with symmetric positive-definite K_{11}, M_{11}, and M_{22}. Thanks to the special splitting the discrete divergence-free condition and the orthogonal projection simplify on the algebraic level as follows.

Proposition 1. *Let $\mathbf{E}_h \in \overline{\mathbb{V}}_{hp}$ where $\mathbf{E}_h = \mathbf{E}_h^1 + \mathbf{E}_h^2$ with $\mathbf{E}_h^1 \in \overline{\mathbb{V}}_{hp}^*$ and $\mathbf{E}_h^2 \in \overline{\mathbb{V}}_{hp}^{\mathrm{ker}}$.*

1. *The discrete gradient operator becomes trivial: If there exists a $w_h \in \overline{\mathbb{W}}_{hp}^+$ (with its coefficient vector $\underline{w}_h$) such that $\mathbf{E}_h = \nabla w_h \in \overline{\mathbb{V}}_{hp}$, then*

$$\begin{pmatrix} \underline{\mathbf{E}}_h^1 \\ \underline{\mathbf{E}}_h^2 \end{pmatrix} = \begin{pmatrix} 0 \\ \underline{w}_h \end{pmatrix}$$

and vice versa.

2. *The discrete divergence-free condition, i.e. $(\varepsilon \mathbf{E}_h, \nabla w_h) = 0$ for all $w_h \in \overline{\mathbb{W}}_{hp}^+$, is equivalent to the algebraic condition $M_{21}\underline{\mathbf{E}}_h^1 + M_{22}\underline{\mathbf{E}}_h^2 = 0$.*
3. *The orthogonal projection of $\mathbf{E}_h$ w.r.t. the kernel $\overline{\mathbb{V}}_{hp}^{ker} = \nabla \overline{\mathbb{W}}_{hp}^+$ defined as*

$$\mathbf{E}_h^\perp := \mathbf{E}_h - \nabla w_h \ \text{with}\ w_h \in \overline{\mathbb{W}}_{hp}^+ : \quad (\varepsilon \nabla w_h, \nabla q_h) = 0 \quad \forall q_h \in \overline{\mathbb{W}}_{hp}^+ \tag{19}$$

can be realized on the algebraic level by

$$\underline{\mathbf{E}}_h^\perp = \begin{pmatrix} \underline{\mathbf{E}}_h^1 \\ -M_{22}^{-1} M_{21} \underline{\mathbf{E}}_h^1 \end{pmatrix}.$$

Remark 2. The explicit kernel splitting can also be generalized to periodic boundary conditions if they are properly integrated into the FE space. In addition one needs compatible periodic $H^1(\Omega)$- and $H(\text{curl})$-discretizations and a periodic tree/co-tree splitting.

2.3 Algebraic Eigenvalue Computation

The algebraic eigenvalue problem (15) can be solved with different methods, e.g. shift invert Arnoldi, LOBPCG (Locally Optimal Block Preconditioned Conjugate Gradient), Jacobi Davidson (see e.g. [6, 7, 24, 33, 35, 37]). For all methods the most time consuming parts are the solution of a large, sparse linear equation system (e.g. with a sparse direct solver), which has to be applied several times, and the orthonormalization of the newly generated search basis. Due to the fact that eigenvalues are often clustered in a region of interest, a block version is preferred, which has additionally the advantage of allowing the use of BLAS level 3 routines, see [38].

In the following the shift invert Arnoldi and the LOBPCG are described in more detail, together with the explicit kernel splitting.

2.3.1 Shift Invert Arnoldi Technique

In the algebraic eigenvalue problem (15) we are only interested in eigensolutions corresponding to $\lambda \neq 0$. Hence, the standard shift invert Arnoldi method needs to be extended by orthogonalization w.r.t. to the discrete kernel $\overline{\mathbb{V}}_{hp}^{\text{ker}}$ in the sense of (19). Every iteration step of the shift invert Arnoldi requires in the lines 13-14 the steps (see Alg. 1)

13: Krylov iteration: Solve $(K - \sigma M)\underline{\mathbf{V}} = \underline{\mathbf{Q}}$ with $\underline{\mathbf{Q}} := M\underline{\mathbf{V}}_n$.

14: Orthogonal projection of $\underline{\mathbf{V}}$ w.r.t. kernel $\overline{\mathbb{V}}_{hp}^{\text{ker}}$ as defined in (19) yields $\underline{\mathbf{V}}_{n+1}$.

Assuming the FE splitting with explicit kernel splitting (17) provided in Sect. 2.2.1 the system to be solved in line 13 of Alg. 1 equals

$$\begin{pmatrix} K_{11} - \sigma M_{11} & -\sigma M_{12} \\ -\sigma M_{21} & -\sigma M_{22} \end{pmatrix} \begin{pmatrix} \underline{\mathbf{V}}^1 \\ \underline{\mathbf{V}}^2 \end{pmatrix} = \begin{pmatrix} M_{11}\underline{\mathbf{V}}_n^1 + M_{12}\underline{\mathbf{V}}_n^2 \\ 0 \end{pmatrix} \tag{20}$$

since the preceding iterates $\underline{\mathbf{V}}_n$ are orthogonal to the kernel. Reviewing Proposition 1, we observe that the second line in (20) provides the orthogonalization constraint scaled by σ and the orthogonalization (line 14) is implied. Hence, FE spaces with explicit kernel splitting the Alg. 1 can be simplified by performing the Krylov iteration in line 13 via solving (20) and skipping the orthogonalization step (line 14). This yields a speed up in the practical computation, since the skipped orthogonalization step (line 14) in general requires the solution of a Poisson problem (19) (or at least a sufficiently good approximation) in each iteration.

Algorithm 1. Shift Invert Arnoldi (with orthogonal kernel projection)

```
generalized partial eigenmode decomposition for symmetric $K \in \mathbb{R}^{N \times N}$ and $M \in \mathbb{R}^{N \times N}$
q eigenmodes (with $\lambda \neq 0$) are calculated

choose an appropriate shift value $\sigma \neq 0$, such that $K - \sigma M$ is regular
$A = (K - \sigma M)^{-1} M$
$\mu_i = \frac{1}{\lambda_i - \sigma}$, $i = 1, \dots, q$
$n = 0$
initialize with random $V_n \in \mathbb{R}^{N \times q}$
orthogonalize $V_n$ with respect to the kernel modes
normalize $V_n$

while eigenvalue is not converged do
    $V_{n+1} = AV_n$
    orthogonalize $V_{n+1}$ with respect to the kernel
    orthonormalize $V_{n+1}$ with respect to $V_0, \dots, V_n$
    $V = (V_0, \dots, V_{n+1})$
    solve the projected eigenmode problem $V^\top AVx = \mu x$
    select the best $q$ eigenmodes and check for convergence
    $n = n + 1$
end while
$\lambda_i = \frac{1}{\mu_i} + \sigma$, $i = 1, \dots, q$
```

2.3.2 LOBPCG Method

The idea of the LOBPCG method is to choose a new eigenvector approximation X_{n+1} which minimizes the Rayleigh-quotient within the 3-dimensional subspace $V_n := \text{span}\{X_n, P_n, W_n\}$ with preconditioned residual $W_n := C^{-1}(KX_n - MX_n\Lambda)$, auxiliary search directions $P_n \in \text{span}\{X_{n-1}, X_n\}$, Λ the Ritz values, C^{-1} an appropriate preconditioner and $n = 0, \dots$ the iteration index. In order to avoid convergence towards eigensolutions in the kernel, the preconditioned residual W_n has to be orthogonalized w.r.t. the discrete kernel $\mathbb{V}_{hp}^{\text{ker}}$ in each step. A block version of this method is summarized in Alg. 2.

For the LOPCG method with inexact preconditioning the orthogonal projection cannot be omitted, which is in contrast to the Arnoldi method, even when using FE spaces with explicit kernel splitting. Nevertheless, one can perform the LOBPCG-iteration only in the unknowns of the ‘reduced’ basis $\mathbb{V}_{hp}^*$ and regard the orthogonal projection in an additional correction step, when computing the reduced residual in each iteration step, namely

$$R^1 = (K_{11}X_n^1 - (M_{11} - M_{12}M_{22}^{-1}M_{12})X_n^1\Lambda_n).$$

For the calculation of the preconditioned residual $W^1 = C_{11}^{-1}R^1$ we choose the preconditioner C_{11}^{-1} for the reduced problem K_{11} as presented in [41]. We point out that the kernel functions are not excluded from the $H(\text{curl})$-basis, only the iteration can be written solely in the degrees of freedom of the completion space. The kernel

Algorithm 2. LOBPCG (with orthogonal kernel projection)

```
generalized partial eigenmode decomposition for symmetric $K \in \mathbb{R}^{N\times N}$ and $M \in \mathbb{R}^{N\times N}$
$q$ non-zero eigenmodes are calculated

choose an appropriate shift value $\sigma \neq 0$, such that $K+\sigma M$ is regular
$C^{-1} \approx (K+\sigma M)^{-1}$
initialize eigenvector iterate $X \in \mathbb{R}^{N\times q}$
orthogonalize $X$ with respect to the kernel modes
initialize diagonal matrix $\Lambda \in \mathbb{R}^{q\times q}$ by Rayleigh quotient w.r.t. $X$
$R = KX - MX\Lambda$ and check for convergence
$P = []$

while eigenvalue is not converged do
    $W = C^{-1}R$
    orthogonalize $W$ with respect to the kernel modes
    $V = [W, X, P]$
    solve the projected eigenmode problem $V^\top KVx = \Lambda V^\top MVx$
    select the $q$ best Ritz vectors and values $x$ and $\Lambda$, respectively
    $X = Vx$
    $R = KX - MX\Lambda$ and check for convergence
    $P = [W, 0, P]x$
end while
```

functions get implicitly included by the correction step. Finally, the solution vector is obtained by orthogonalization of the last iterate X^1 via

$$X := \begin{pmatrix} X^1 \\ -M_{22}^{-1}M_{21}X^1 \end{pmatrix}.$$

For further iterative solution strategies and preconditioning techniques see, e.g., [32, 43], for inexact kernel orthogonalization [22, 34].

3 Sensitivity Analysis

Let us consider the parameterized eigenvalue problem from Maxwell's equations (15) where we calculate the derivative of (16) with respect to a design parameter, i.e. we have the form

$$\lambda(p) = \frac{\langle K(p)\underline{\mathbf{E}}_h(p), \underline{\mathbf{E}}_h(p)\rangle}{\langle M(p)\underline{\mathbf{E}}_h(p), \underline{\mathbf{E}}_h(p)\rangle} \tag{21}$$

where $\underline{\mathbf{E}}_h(p)$ and $\lambda(p)$ is the nominal solution of the eigenvalue problem (15) at the parameter set $p = (p_1, \ldots, p_m)$, with m the number of design parameters. For the derivative of an eigenvalue we get (omitting the dependency on p_i)

$$\frac{\partial\lambda}{\partial p_i} = \frac{\langle \frac{\partial K}{\partial p_i}\underline{\mathbf{E}}_h, \underline{\mathbf{E}}_h\rangle - \lambda\langle \frac{\partial M}{\partial p_i}\underline{\mathbf{E}}_h, \underline{\mathbf{E}}_h\rangle}{\langle M\underline{\mathbf{E}}_h, \underline{\mathbf{E}}_h\rangle} \tag{22}$$

by using the fact that $K = K^\top$ and $M = M^\top$, i.e. the system matrices are symmetric and real. Additionally the mass matrix M is positive definite. This formula can be applied to every geometrical boundary where at every point a vector for the directional derivative is defined (see Fig. 1). A continuous description of the direction $v(\cdot)$ is required because curved elements are used.

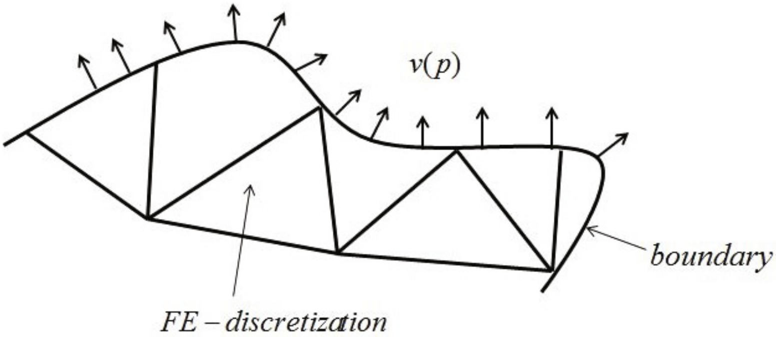


Fig. 1 FE discretization with curved elements and vector field for the directional derivative.

The actual computation of the derivative of the system matrix $\frac{\partial K(p)}{\partial p_i}$ and $\frac{\partial M(p)}{\partial p_i}$ can be done by direct differentiation of the shape functions (see e.g. [3]). For material parameters the assembling process has to be applied with the derivative of the material.

Remark 3.

- The eigenvectors are usually normalized with respect to the M-inner product, i.e., $\langle M\underline{\mathbf{E}}_h, \underline{\mathbf{E}}_h \rangle = 1$.
- The derivation of formula (22) could have been done on the continuous description as well. Based on the bilinear form it can be seen that only an application of the derivatives is required and therefore the matrices never need to be assembled.
- For degenerated eigenmodes formula (22) can be applied if the related subspace can be split up accordingly. Since this case is of no practical interest a deeper analysis is omitted.
- In the case of static modes we assume that PEC regions are properly separated. In that case the zero-eigenvalues of the static modes do not change with a design parameter, and consequently their derivative is always zero.
- If one is interested in the derivative of the eigenmodes $\mathbf{E}_h$, i.e. $\frac{\partial \mathbf{E}_h}{\partial p_i}$ a system of linear equations has to be solved additionally.

4 Numerical Results

For all of the following examples we have used second order Nédélec finite elements of the first type, and (in case of non-polyhedral domains) a second order curving of the elements. For the solution of the algebraic eigenvalue problem the shift invert Arnoldi method of Subsect. 2.3.1 is used.

Our first model is a simple cavity, i.e. a vacuum unit cube with PEC boundary conditions, where one side is varied. Here, analytical results are available, which are required for cross-checking the calculated results. The exact eigenvalues of a brick with spatial dimensions a, b, and c calculate as

$$\omega_i = \frac{1}{2\sqrt{\mu_0\varepsilon_0}}\sqrt{\left(\frac{m}{a}\right)^2+\left(\frac{n}{b}\right)^2+\left(\frac{p}{c}\right)^2},$$

with m, n, p non-negative integers (with at most one equal zero), vacuum permeability $\mu_0 = 4\pi \cdot 10^{-7}$ and permittivity $\varepsilon_0 = \frac{1}{\mu_0 c_0^2}$, and c_0 the speed of light in vacuum (c.f., [14]).

For the unit cube ($a = b = c = 1$m) with one varying side a we will compare the computed sum of the derivatives of the three smallest eigenvalues with the exact result

$$\frac{\partial}{\partial a}(\omega_1+\omega_2+\omega_3)\bigg|_{a=b=c=1} = -\frac{1}{\sqrt{2\mu_0\varepsilon_0}}.$$

The sum is used, as the first three eigenvalues are equal and not distinguishable. Consequently only the sum of the derivatives is uniquely computable.

In Table 1 and Fig. 2 the h-convergence of the method is illustrated.

Table 1 The relative errors of the average of the first three eigenmodes and the sum of the corresponding derivatives.

mesh-size h [m]	1	1/2	1/4	1/8	1/16	1/32
rel. error eigenvalues	1.8E-4	1.1E-4	4.9E-5	1.1E-5	6.2E-7	4.5E-8
rel. error derivatives	2.1E-3	1.2E-3	6.7E-4	4.4E-5	3.2E-6	1.2E-6

As a second example, the method is applied to an iris coupled cavity bandpass filter. The physical layout of this microwave filter is shown in Fig. 3. Here, the engineer has to deal with fourteen geometrical design parameters:

- The heights of the cylinders in the cavity centers hc_1, hc_2, and hc_3, and their radii rc_1, rc_2, and rc_3 are responsible for tuning the frequencies in the cavities.
- A coarse pre-tuning of the required coupling bandwidths is accomplished via the iris widths w_1, w_2, w_3, and w_4.

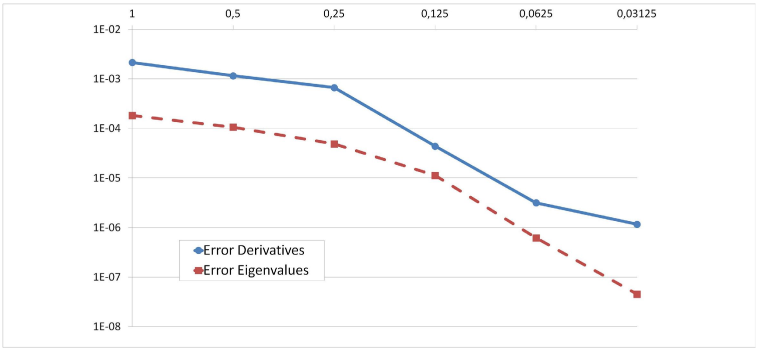


Fig. 2 Comparing the convergence of the eigenvalues (dashed line) with that of the derivatives (solid line).

- Finally the coupling bandwidths are fine-tuned using the screws positioned at the center of the irises with heights hw_1, hw_2, hw_3, and hw_4.

Due to the open design of such a filter, a manual tuning process becomes very complex and delicate, as the filter-characteristics depend on the design parameters in a highly nonlinear way. Especially for small relative bandwidths (less than one percent) an automatic tuning process is proposed.

After six adaptive refinement steps using a Zienkiewicz-Zhu-type error estimator (leading to a mesh consisting of 125k tetrahedrons, a finite element space with 460k vector-valued degrees of freedom, and 4GB computer memory consumption in total; see Fig. 4 for the mesh) the results are considered to be converged. Now arbitrary sensitivities can be calculated without much further ado but just bilinear form-applications. Results for rc_1 and hc_1 can be found in Table 2, a corresponding eigenmode in Fig. 5. In Table 3 we try to ‘predict’ the results of an increased cylinder radius by a Taylor expansion (up to the linear term) and compare with the results computed by a complete simulation run.

Table 2 Results of the filter example. One sensitivity analysis was performed for the radius of the cylinders in the first and the last cavity rc_1 (at $rc_1 = 40$mm), another for their heights hc_1 (at $hc_1 = 27$mm).

eigenvalues [MHz]	derivatives w.r.t. rc_1 [MHz/m]	derivatives w.r.t. hc_1 [MHz/m]
667.2	–1.2	–79
669.0	–3.0	–210
673.1	–5.4	–370
676.0	–5.2	–360
681.1	–19	–1300
681.3	–17	–1200

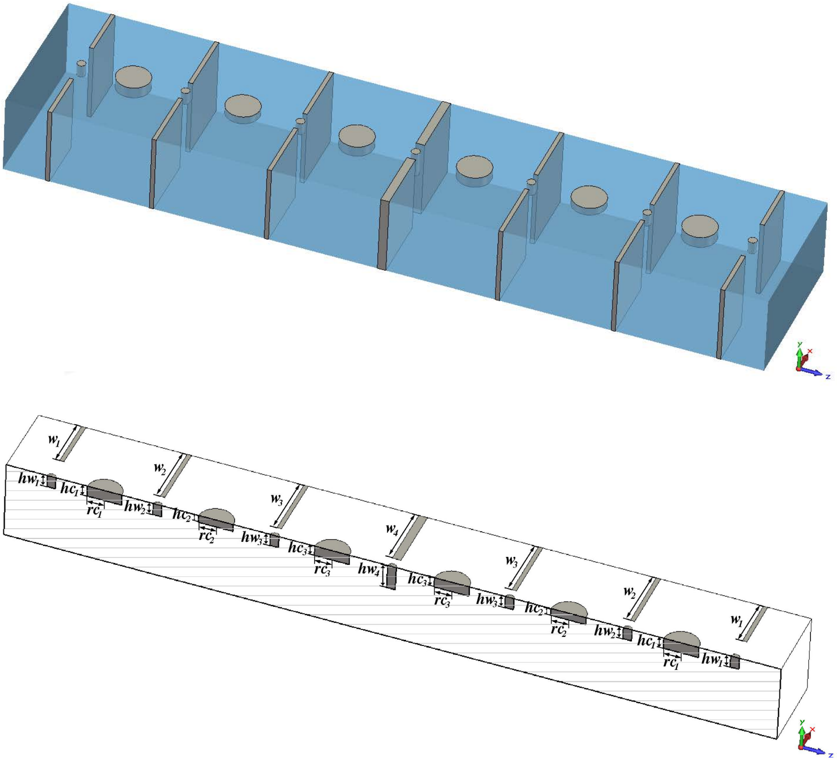


Fig. 3 The filter geometry (complete and cut) with design parameters. The total length of the structure is about 1.8m. The gray areas are PEC parts (screws, cylinders, irises). For technical reasons, some of the design parameters are chosen symmetrically, e.g., the heights of the first and last cavity-cylinder have to be equal hc_1; the corresponding radii rc_1. The same holds for the iris widths and the tuning screws (and for the other cylinder/iris/tuning screw pairs).

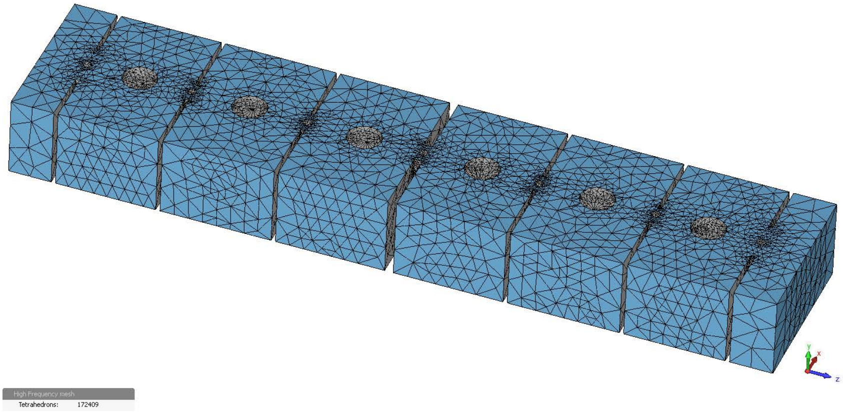
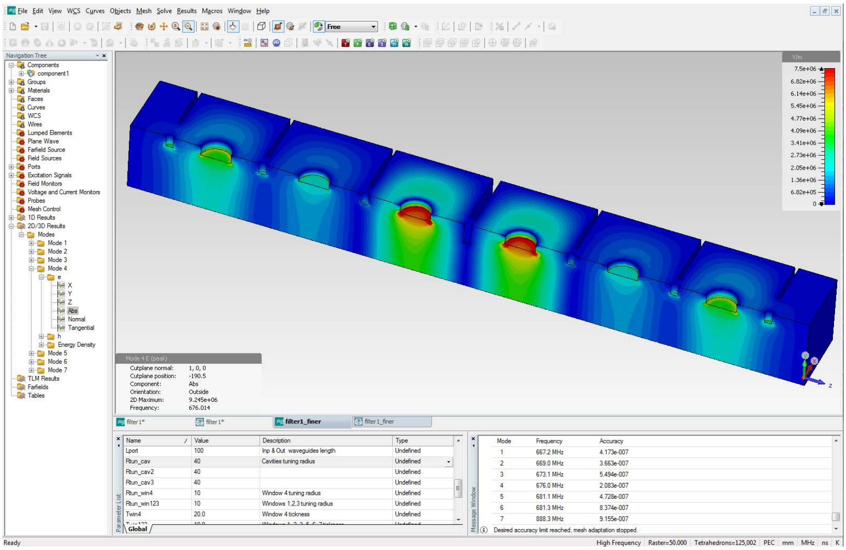


Fig. 4 Mesh of the filter structure after six adaptive refinement steps.



Mode 4 E (peak)

Cutplane normal: 1, 0, 0
Cutplane position: -190.5
Component: Abs
Orientation: Outside
2D Maximum: 9.245e+06
Frequency: 676.014

Name	Value	Description	Type
Lport	100	Inp & Out waveguides length	Undefined
Rtun_cav	40	Cavities tuning radius	Undefined
Rtun_cav2	40		Undefined
Rtun_cav3	40		Undefined
Rtun_win4	10	Window 4 tuning radius	Undefined
Rtun_win123	10	Windows 1,2,3 tuning radius	Undefined
Twin4	20.0	Window 4 tickness	Undefined

Mode	Frequency	Accuracy
1	667.2 MHz	4.173e-007
2	669.0 MHz	3.663e-007
3	673.1 MHz	5.494e-007
4	676.0 MHz	2.083e-007
5	681.1 MHz	4.728e-007
6	681.3 MHz	8.374e-007
7	888.3 MHz	9.155e-007

Desired accuracy limit reached, mesh adaptation stopped.

High Frequency | Raster=50.000 | Tetrahedrons=125,002 | PEC | mm | MHz | ns | K

Fig. 5 Visualization of the fourth eigenmode (peak value; cut through the geometry).

Table 3 We estimate the eigenvalues at $rc_1 = 44$mm by using value and derivative at $rc_1 = 40$mm and linear Taylor expansion $\omega_i(0.044) \approx \omega_i(0.040) + 0.004 \cdot \omega_i'(0.040)$.

eigenvalues computed for $rc_1 = 40$mm [MHz]	eigenvalues estimated for $rc_1 = 44$mm [MHz]	eigenvalues computed for $rc_1 = 44$mm [MHz]	rel. difference
667.2	667.0	667.0	2.9E–5
669.0	668.7	668.7	1.3E–4
673.1	672.5	672.4	1.7E–4
676.0	675.5	675.3	2.1E–4
681.1	679.1	679.3	2.9E–4
681.3	679.5	679.8	4.2E–4

Looking, e.g., at the last two lines of Table 2 one can see that small variations in the body of the structure can have a significant influence on the eigenvalues, i.e., the tuning of the filter.

5 Summary and Outlook

In this paper we presented a collection of reliable and efficient methods in order to solve the loss free Maxwell eigenvalue problem. One major aspect was the calculation of derivatives of eigenvalues with respect to geometrical and material design parameters.

Since the approach is valid for arbitrary parameterized variations we will apply our results to a coupled EM-thermal-mechanical simulation, as sketched in Alg. 3.

Algorithm 3. Coupled simulation of eigenmode sensitivity

```
1: Compute the eigenmode decomposition and the corresponding loss densities (via a perturbation ansatz)
2: Compute the thermal distribution of the loss densities from the eigenmode calculation
3: Compute the structural mechanical displacements stemming from the thermal solution
4: Use the displacement field for the computation of the eigenvalue derivatives
```

Step 4 of Alg. 3 requires no 3D EM simulation, and therefore the sensitivity of eigenvalues can be calculated very efficiently at almost no additional cost if the mechanical displacements are relatively small. Such an application is of great importance for accelerator physics and filter design.

Acknowledgements. The authors would like to thank their colleagues at CST AG for fruitful discussions and pointing out important references.

References

[1] Ainsworth, M., Coyle, J.: Hierarchic finite element bases on unstructured tetrahedral meshes. Int. J. Numer. Meth. Engrg. 58(14), 2103–2130 (2001)

[2] Ainsworth, M., Coyle, J., Ledger, P., Morgan, K.: Computation of Maxwell eigenvalues using higher order edge elements in three dimensions. IEEE Trans. Magnet. 39(5), 2149–2153 (2003)

[3] Akel, H., Webb, J.P.: Design sensitivities for scattering matrix calculation with tetrahedral edge elements. IEEE Trans. Magnet. 36(4), 1043–1046 (2000)

[4] Albanese, R., Rubinacci, G.: Integral formulation for 3D eddy–current computation using edge elements. IEE Proc. A., 457–462 (1988)

[5] Ali, S.M., Nikolova, N.K., Bakr, M.H.: Recent advances in sensitivity analysis with frequency–domain full–wave EM solvers. Appl. Comput. Electromagnet. Soc. J. 19(3), 147–154 (2004)

[6] Arbenz, P., Geus, R.: A comparison of solvers for large eigenvalue problems occuring in the design of resonant cavities. Numer. Linear Alg. Appl. 6, 3–16 (1999)

[7] Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H.: Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM, Philadelphia (2000)

[8] Bandler, J.W., Zhang, Q.J., Biernacki, R.M.: A unified theory for frequency–domain simulation and sensitivity analysis of linear and nonlinear circuits. IEEE Trans. Microwave Theory Techn. 36(12), 1661–1669 (1988)

[9] Boffi, D., Fernandes, P., Gastaldi, L., Perugia, I.: Computational models of electromagnetic resonators: Analysis of edge element approximation. SIAM J. Numer. Anal. 36, 1264–1290 (1999)

[10] Boffi, D.: Finite element approximation of eigenvalue problems. Acta Numerica 19, 1–120 (2010)

[11] Bossavit, A.: Computational Electromagnetism. Academic Press, Boston (1998)

[12] Buffa, A., Perugia, I.: Discontinuous Galerkin approximation of the Maxwell eigenproblem. SIAM J. Numer. Anal. 44, 2198–2226 (2006)

[13] Costabel, M., Dauge, M.: Computation of resonance frequencies for Maxwell equations in non smooth domains. In: Ainsworth, M., Davies, P., Duncan, D., Martin, P., Rynne, B. (eds.) Topics in Computational Wave Propagation. Lecture Notes in Computational Science and Engineering, vol. 31, pp. 125–161. Springer (2003)

[14] Costabel, M., Dauge, M.: Maxwell eigenmodes in tensor product domains (2006), `http://perso.univ-rennes1.fr/monique.dauge/publis/CoDa06MaxTens2.pdf`

[15] Cao, Y., Reitzinger, S., Zhang, Q.J.: Simple and efficient high-dimensional parametric modeling for microwave cavity filters using modular neural network. IEEE Microwave Wireless Components Letters 21(5), 258–260 (2011)

[16] CST MICROWAVE STUDIO ®, ver. 2011, CST AG, Bad Nauheimer Str. 19, D-64289 Darmstadt, Germany (2010), `http://www.cst.com`

[17] Demkowicz, L., Monk, P., Vardapetyan, L., Rachowicz, W.: De Rham diagram for hp finite element spaces. Comput. Math. Appl. 39, 29–38 (2000)

[18] Demkowicz, L., Kurtz, J., Pardo, D.: Computing with hp–Adaptive Finite Elements: Frontiers: three dimensional elliptic and Maxwell problems with applications. Chapman & Hall/CRC (2007)

[19] Graeb, H.E.: Analog Design Centering and Sizing. Springer, The Nederlands (2007)

[20] Haase, G., Kuhn, M., Langer, U.: Parallel multigrid 3D Maxwell solvers. Parallel Comput. 27, 761–775 (2001)

[21] Hiptmair, R.: Finite elements in computational electromagnetrism. Acta Numerica 11, 237–339 (2002)

[22] Hiptmair, R., Neymair, K.: Multilevel method for mixed eigenproblems. SIAM J. Sci. Comput. 23(6), 2141–2164 (2002)

[23] Kikuchi, F.: Mixed and penalty formulations for finite element analysis of an eigenvalue problem in electromagnetism. Comput. Methods Appl. Mech. Eng. 64, 509–521 (1987)

[24] Knyazev, A.V.: Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. SIAM J. Sci. Comput. 23(2), 517–541 (2001)

[25] Koziel, S., Mosler, F., Reitzinger, S., Thoma, P.: Robust microwave design optimization using adjoint sensitivity and trust regions. Int. J. RF Microwave Computer-Aided Engrg. (2011) (submitted)

[26] Lee, L.-Q., Li, Z., Ng, C., Ko, K.: Omega3P: A parallel finite–element eigenmode analysis code for accelerator cavities. SLAC-PUB-13529 (2009)

[27] Mehrmann, V., Schröder, C., Simoncini, V.: An implicitly restarted Krylov subspace method for real symmetric/skew–symmetric eigenproblems. Lin. Algebra Appl. 436(10), 4070–4087 (2012)

[28] Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford Science Publications (2003)

[29] Nair, D., Webb, J.P.: Estimating errors in design sensitivities. IEEE Trans. Magnet. 42(4), 559–562 (2006)

[30] Nédélec, J.C.: Mixed finite elements in R3. Numer. Math. 35, 315–341 (1980)

[31] Nédélec, J.C.: A new family of mixed finite elements in R3. Numer. Math. 50, 57–81 (1986)

[32] Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia (2003)

[33] Saad, Y.: Numerical Methods for Large Eigenvalue Problems. SIAM, Philadelphia (2011)

[34] Schöberl, J., Zaglmayr, S.: High order Nédélec elements with local complete sequence property. Int. J. Comput. Math. Electrical Electronic Engrg. 24(2), 374–384 (2005)

[35] Sleijpen, G.L.G., van der Vorst, H.A.: A Jacobi–Davidson iteration method for linear eigenvalue problems. SIAM J. Matrix Anal. Appl. 17, 401–425 (1996)

[36] Sokolowski, J., Zolesia, J.: Introduction to Shape Optimization: Shape Sensivitiy Analysis. Springer, Berlin (1992)

[37] Stewart, G.W.: Matrix Algorithms II: Eigensystems. SIAM, Philadelphia (2001)

[38] Stathopoulos, A., Wu, K.: A block orthonormalization procedure with constant synchronization requirements. SIAM J. Sci. Comput. 23(6), 2165–2182 (2002)

[39] Vardapetyan, L., Demkowicz, L.: hp-adaptive finite elements in electromagnetics. Comput. Methods Appl. Mech. Eng. 169, 331–344 (1999)

[40] Webb, J.P.: Using adjoint solutions to estimate errors in global quantities. IEEE Trans. Magnet. 41(5), 1728–1731 (2005)

[41] Zaglmayr, S.: High Order Finite Element Methods for Electromagnetic Field Computation. PhD thesis, Johannes Kepler University Linz (2006)

[42] Zienkiewicz, O.C., Taylor, R.L.: Finite Element Method. The Basis, vol. 1. Elsevier (2000)

[43] Zhu, Y., Cangellaris, A.: Multigrid Finite Element Methods for Electromagnetic Field Modeling. IEEE Press Series on Electromagnetic Wave Theory (2006)

Non-sequential Optical Field Tracing

Michael Kuhn, Frank Wyrowski, and Christian Hellmann

Abstract. Optical field tracing methods generalize ray tracing methods by considering harmonic fields instead of ray bundles. This allows the smooth combination of different modeling techniques in different subdomains of the system. Based on tearing and interconnecting ideas, the paper introduces the basic concepts of non-sequential field tracing and derives the corresponding operator equations and a solution formula for the simulation task. The evaluation requires the solution of local Maxwell problems (tearing) and the continuity of the solution across boundaries is achieved along with the convergence of the iterative procedure (interconnecting). The number of local problems to be solved is optimized by a newly introduced light path tree algorithm. Finally some examples for the selection of local Maxwell solvers and numerical results are presented.

1 Introduction

The design of modern optical system requires advanced simulation techniques. In general the simulation task requires the solution of Maxwell's equations in frequency or in time domain. Although the solution of these equations using, e.g., Finite Element Methods (FEM), has been discussed extensively over the last decades,

Michael Kuhn
LightTrans VirtualLab UG, Kahlaische Strasse 4, 07745 Jena, Germany
e-mail: `michael.kuhn@lighttrans.com`

Frank Wyrowski
Institut für Angewandte Physik, Friedrich–Schiller–Universität Jena,
Winzerlaer Strasse 10, 07745 Jena, Germany
e-mail: `frank.wyrowski@uni-jena.de`

Christian Hellmann
LightTrans GmbH, Kahlaische Strasse 4, 07745 Jena, Germany
e-mail: `christian.hellmann@lighttrans.com`

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 257–273.
springerlink.com

the task is still very challenging in the field of optics due to the following main reasons: (1) the wavelength of interest is typically below 1 micrometer, sometimes below 100 nanometers, and, (2) the length scales of one system vary between nanometers and meters. Standard laser systems which are to be designed use a wavelength of 532 nm (green light), use structured surfaces with feature sizes of several micrometers and require the simulation along several centimeters or meters in one system. This indicates that the physical optics simulation, e.g., with standard FEM, is not feasible on standard computers today.

On the other hand, many of those optical systems can be simulated with sufficient accuracy using approximate methods. In particular ray tracing methods are widely used in optics simulation. Several commercial tools based on ray tracing methods have been established in the 1980s together with the emerging PC technology. However, ray tracing methods have some strong limitations, e.g., they break down in the presence of microstructures.

That is why the concept of field tracing has been introduced [6, 12]. Field tracing considers a decomposition of an optical system into subdomains. In contrast to ray tracing, electromagnetic harmonic fields are traced through the system. This approach provides three fundamental advantages of practical concern: (1) field tracing enables unified optical modeling. Its concept allows the utilization of any modeling technique that is formulated for vectorial harmonic fields in different subdomains of the system. (2) The use of vectorial harmonic fields as a basis of field tracing permits a great flexibility in light-source modeling. By propagating sets of harmonic field modes through the system, partially temporally and spatially coherent light as well as ultrashort pulses can be investigated [9]. (3) In system modeling and design, the evaluation of any type of detector function is crucial. The use of vectorially formulated harmonic fields provides unrestricted access to all field parameters and therefore it allows the introduction and evaluation of any type of detector. In field tracing we solve local Maxwell problems for subdomains. These local problems often have properties that give rise to solutions in certain subspaces of all admissible functions. Then, approximate Maxwell solvers are accurate enough and are typically much cheaper than rigorous Maxwell solvers. In this sense, we adapt the main ideas of domain decomposition and tearing and interconnecting methods as they have been used for several applications, see e.g. [3] and [4] and references therein. The goal of field tracing is to construct a problem dependent solver which is as fast as possible and as accurate as needed by combining different subdomain solvers. The solution of the global problem requires a coupling of the local solutions by enforcing continuity conditions. For that purpose we are looking for a generalization of tracing techniques which are well established in optics. An introduction with emphasis on the sequential case has been given in [12]. Here we want to extend the ideas to the non-sequential case and we add more algorithmic modules describing the solver. It is shown how tearing and interconnecting is being applied.

The paper is organized as follows. In Sect. 2, we discuss the definition of local Maxwell solvers. We describe how the 3d Maxwell problem can be formulated using a tearing and interconnecting approach. The solution formula using local operators is derived based on Neumann series resulting in an infinite sum. This sum can be reformulated as an iterative procedure using an update formula which is discussed in Sect. 3. The algorithm itself can be formulated as a light path tree. Field tracing methods for solving local problems are discussed in Sect. 4. Finally we present numerical results in Sect. 5 and conclusions in Sect. 6.

2 The Tearing and Interconnecting Method

Optical system modeling deals with the solution of Maxwell's equations for the electric field **E** and the magnetic field **H** in $\mathbb{R}^3$. In the frequency domain with frequency ω that corresponds to

$$\nabla \times \mathbf{E}(\mathbf{r},\omega) = \mathrm{i}\omega\mu_0\mathbf{H}(\mathbf{r},\omega), \tag{1}$$

$$\nabla \times \mathbf{H}(\mathbf{r},\omega) = -\mathrm{i}\omega\varepsilon_0\hat{\varepsilon}_r(\mathbf{r},\omega)\mathbf{E}(\mathbf{r},\omega), \tag{2}$$

$$\nabla \cdot \Big(\hat{\varepsilon}_r(\mathbf{r},\omega)\mathbf{E}(\mathbf{r},\omega)\Big) = 0, \tag{3}$$

$$\nabla \cdot \Big(\mu_0\mathbf{H}(\mathbf{r},\omega)\Big) = 0 \tag{4}$$

for linear matter equations and isotropic media. The refractive index $\hat{n}(\mathbf{r})$ with $\mathbf{r} = (x,y,z)$ of the system is inhomogeneous and is defined by

$$\hat{n}^2(\mathbf{r},\omega) = \hat{\varepsilon}_r(\mathbf{r},\omega) = \Big(n(\mathbf{r},\omega) + \mathrm{i}\kappa(\mathbf{r},\omega)\Big)^2. \tag{5}$$

The solution for each frequency ω is given by an electromagnetic harmonic field which is specified by the three electric and three magnetic field components. The determination of all field components in the system domain Ω is the most general task to be solved in optical system modeling.

To simplify the notation we summarize all six field components by the field vector **V**:

$$\mathbf{V}(\mathbf{r}) = \Big(E_x(\mathbf{r}), E_y(\mathbf{r}), E_z(\mathbf{r}), H_x(\mathbf{r}), H_y(\mathbf{r}), H_z(\mathbf{r})\Big)^\top. \tag{6}$$

From Maxwell's equations it is clear, that the six components of **V** are not independent. In particular we can calculate the magnetic field always from the electric one. However we use **V** in order to emphasize, that the simulation has to deliver all six field components for providing maximum flexibility to define detectors for the evaluation of field properties. For instance the Poynting vector is of great practical concern in energy considerations. Its definition combines the magnetic and the electric field.

Fig. 1 illustrates the modeling situation of concern. The system is located in the domain $\Omega \subset \mathbb{R}^3$. The J subdomains Ω_j embed *all* locations in which the refractive index $\hat{n}(\mathbf{r})$ with $\mathbf{r} = (x,y,z)$ is inhomogeneous. We denote the boundaries of Ω_j by Γ_j.

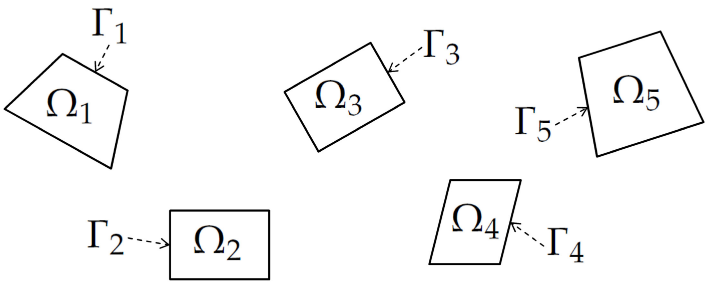


Fig. 1 Formally a system is subdivided into J subdomains Ω_j. All of them are embedded in a homogeneous and isotropic medium of refractive index n. The boundaries of the subdomains are denoted by Γ_j.

From a practical point of view the subdomains are related to components of the system, but that is not essential for the following discussion. In particular it can be of advantage to decompose one component into more than one subdomain. Moreover, it is sometimes helpful to define a subdomain in a homogeneous region of the system. The shape and size of the subdomains can be freely chosen to some extent and depend on the specification of the modeling techniques. All subdomains are embedded in the homogeneous dielectric with refractive index n.

In order to obtain a formulation for the simulation of the entire system we apply some tearing and interconnecting approach. First we define scattering problems for each subdomain Ω_j. Then we state equations that interconnect the solutions of the local scattering problems. Finally, the global problem is described by a equilibrium equation which ensures the continuity of the fields.

In order to define the local scattering problem we denote the field at the boundary Γ_j by

$$\mathbf{V}_j = \mathbf{V}_{|\mathbf{r}\in\Gamma_j} \,. \tag{7}$$

Further, we denote the input (incident) field that hits Ω_j by $\mathbf{V}_j^{\text{in}}$ and the output (scattered) field by $\mathbf{V}_j^{\text{out}}$. The solution of the scattering problems defines a mapping of the input field to the output field by some operator $\mathscr{C}_j$

$$\mathbf{V}_j^{\text{out}} = \mathscr{C}_j \mathbf{V}_j^{\text{in}} \,. \tag{8}$$

The interconnecting problems describe the relation of any pair $(\mathbf{V}_i^{\text{in}}, \mathbf{V}_j^{\text{out}})$ of one input and one output field in a homogeneous medium. For that purpose we introduce the operator $\mathscr{P}_{ji}$ which maps the output field of subdomain i to the incident field of subdomain j with $i \neq j$:

$$\mathbf{V}_j^{\text{in}} = \mathscr{P}_{ji} \mathbf{V}_i^{\text{out}} \,. \tag{9}$$

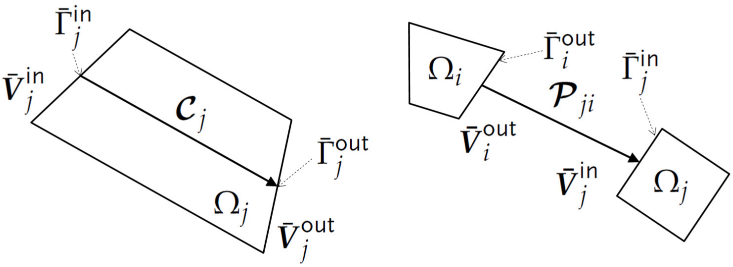


Fig. 2 Notation for the application of field tracing $\mathscr{C}$ through a subdomain between two plane parts of the boundary Γ_j (left) and of the propagation $\mathscr{P}_{ji}$ between plane boundary segments $\bar{\Gamma}$ of two subdomains (right).

The evaluation of $\mathscr{P}_{ji}$ requires the solution of a Maxwell problem as before, but now in a half space (with respect to Γ_j) of $\mathbb{R}^3$ with a homogeneous medium and the incident field $\mathbf{V}_j^{\text{out}}$ given at Γ_j. The solution is to be computed at Γ_i only yielding $\mathbf{V}_i^{\text{in}}$.

Finally we have to ensure continuity of the fields. This leads to the equilibrium equation for the multiple interaction problem between all subdomains. The output field on Γ_j must satisfy the equation

$$\mathbf{V}_j^{\text{out}} = \mathbf{V}_j^{\text{source}} + \sum_{i=1, i\neq j}^{J} \mathscr{C}_j \mathscr{P}_{ji} \mathbf{V}_i^{\text{out}} . \tag{10}$$

The optional source field $\mathbf{V}_j^{\text{source}}$ contributes to the output field of subdomain j and is therefore added to the sum over all contributions from other subdomains. With (10) we have derived a set of J equations for the unknowns $\mathbf{V}_j^{\text{out}}$ for $j = 1, \dots, J$.

Next we derive a matrix formulation of equation (10). To this end we define the following vectors and matrices:

$$\underline{\mathbf{V}}^{\text{out}} = \left(\mathbf{V}_j^{\text{out}}\right)_{j=1}^{J} , \tag{11}$$

$$\underline{\mathbf{V}}^{\text{source}} = \left(\mathbf{V}_j^{\text{source}}\right)_{j=1}^{J} , \tag{12}$$

$$\mathbf{C} = \text{diag}\left(\mathscr{C}_j\right)_{j=1}^{J} , \tag{13}$$

$$\mathbf{P} = \left(\mathscr{P}_{ji}\right)_{j,i=1}^{J} , \tag{14}$$

$$\mathbf{I} = \mathrm{diag}\big(\mathscr{I}\big)_{j=1}^{J}. \tag{15}$$

$\mathbf{I}$ is the diagonal matrix of identity operators $\mathscr{I}$. The diagonal elements of $\mathbf{P}$ are always zero, because we do not consider a mapping of the output field of a subdomain to its own input field. With the definitions we can rewrite equation (10) and it follows

$$\big(\mathbf{I} - \mathbf{CP}\big)\underline{\mathbf{V}}^{\mathrm{out}} = \underline{\mathbf{V}}^{\mathrm{source}} \tag{16}$$

which yields

$$\underline{\mathbf{V}}^{\mathrm{out}} = \big(\mathbf{I} - \mathbf{CP}\big)^{-1}\underline{\mathbf{V}}^{\mathrm{source}}. \tag{17}$$

If the condition

$$\|\mathbf{CP}\| < 1 \tag{18}$$

is satisfied then equation (17) is well defined and it can be solved by the Neumann series [7]

$$\underline{\mathbf{V}}^{\mathrm{out}} = \sum_{m=0}^{\infty} \big(\mathbf{CP}\big)^{m}\underline{\mathbf{V}}^{\mathrm{source}}. \tag{19}$$

Condition (18) is satisfied for a broad range of applications. Any absorption process in media, at outer boundaries (infinity) or at boundaries associated with detectors leads to $\|\mathbf{CP}\| < 1$ since we also have $\|\mathbf{C}\| \leq 1$ and $\|\mathbf{P}\| \leq 1$. However, for cavities without any losses we have $\|\mathbf{CP}\| = 1$ and hence, the Neumann series does not converge. In such a case, the tearing and interconnecting approach has to be used within an eigenvalue solver.

The limit of the series in (19) is the solution of the optical simulation problem. An appropriate truncation can be used to approximate the solution. It is obvious that successive summands can be computed by an update formula. This approach results in a so called light path tree algorithm which is discussed in the next section. In order to compute the summands, local Maxwell problems have to be solved for evaluating the operators $\mathbf{C}$ and $\mathbf{P}$. Any rigorous or approximate solver can be used as long as the coupling using the field vector $\mathbf{V}$ is ensured. This approach is called Field Tracing and is being discussed in Sect. 4.

3 The Light Path Tree

In this section, we discuss how the sum in equation (19) can be computed efficiently. The goal is to use update formulas in order to avoid repetitions of identical operations. Let us define an iterative process by truncating the infinite sum. This defines the k-th iterate by

$$(\underline{\mathbf{V}}^{\text{out}})_k = \sum_{m=0}^{k} \big(\mathbf{CP}\big)^m \underline{\mathbf{V}}^{\text{source}}. \tag{20}$$

We introduce an auxiliary variable $(\underline{\mathbf{W}})_k$. Then, by defining the initial conditions

$$(\underline{\mathbf{V}}^{\text{out}})_0 = \underline{\mathbf{V}}^{\text{source}}, \tag{21}$$
$$(\underline{\mathbf{W}})_0 = \underline{\mathbf{V}}^{\text{source}} \tag{22}$$

we obtain the following update formulas

$$(\underline{\mathbf{W}})_{k+1} = \big(\mathbf{CP}\big)(\underline{\mathbf{W}})_k, \tag{23}$$
$$(\underline{\mathbf{V}}^{\text{out}})_{k+1} = (\underline{\mathbf{V}}^{\text{out}})_k + (\underline{\mathbf{W}})_{k+1}. \tag{24}$$

Given a threshold δ, a suitable stopping criterion could be defined by

$$r_k < \delta \tag{25}$$

where r_k is the relative power of the update $(\underline{\mathbf{W}})_k$:

$$r_k = \frac{||(\underline{\mathbf{W}})_k||}{||\underline{\mathbf{V}}^{\text{source}}||}. \tag{26}$$

Altogether we have defined an iterative process in order to compute the solution vector $\underline{\mathbf{V}}^{\text{out}}$. At each iteration step, we have to solve sets of local Maxwell problems: one for each domain Ω_j (application of the operator **C**) and one for each free space region between any Ω_i and any Ω_j (application of the operator **P**). The convergence is guaranteed if (18) holds. Convergence results using the stopping criterion (25) are given in Sect. 5.

It turns out that the update formula (23) requires further discussion. Expanding the matrix notation for some line j formally gives the sum

$$W_j = C_j \sum_{i=1}^{J} P_{ji} W_i.$$

Each summand represents a harmonic field. In order to exploit local properties of the fields for constructing efficient subdomain solvers, it is desirable not to compute the sum but to work with individual summands in further computations.

This situation leads us to the development of the light path tree. This algorithm allows to take into account the sparsity of the iteration vectors $(\underline{\mathbf{W}})_k$. Such a sparsity is typical for optical simulations. Practically this appears for the following reasons: (1) only a single light source is present, (2) the light propagates along a path though components (e.g. through a sequence of lenses in a microscope), (3) the solution is to be computed in one (or a few) planes representing detectors (e.g. a camera) only. In [12] the sequential field tracing (named therein ”Convective single neighbor approximation”) has been discussed which results in $(\underline{\mathbf{W}})_k$ having one non-zero entry for systems with an start(source)-to-end(detector) light path. Here, we extend this approach to the general case which we also refer to as non-sequential field tracing.

Let us discuss the structure of the light path tree using a simple example of an optical system consisting of a light source, two plates and a detector plane where the field is to be computed, see Fig. 3 for the setup.

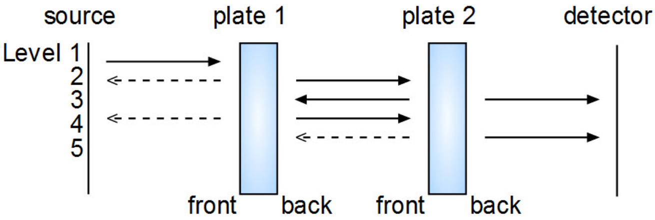


Fig. 3 Example of an optical system with one source, two plates and one detector plane. The arrows show the single summands of the sum, levels stand for iteration steps to compute truncated sums.

In Fig. 3, the arrows indicate that a field is propagated between two plates. Dashed arrows indicate that the operation does not contribute to the final result, i.e. it can be omitted. Further, the arrows are ordered by level 1–5. The index of the level corresponds to the iteration index k. For practical reasons we have introduced a front and a back side for each plate. For the setup described in Fig. 3, the corresponding light path tree is shown in Fig. 4. The nodes of the tree are associated with one entry (summand) of any vector $(\underline{\mathbf{W}})_k$ or with the solution $\mathbf{V}_j^{\text{out}}$ at Γ_j. The connections between the nodes are associated with an operator $\mathscr{C}_j$ or $\mathscr{P}_{ji}$. Without loss of generality we assume that $\mathbf{V}_j^{\text{source}} \neq 0$ for one index j only. Hence, the tree has a single root node.

Neglecting dashed connections, the light path tree is optimal for computing the truncated sum (20) in the sense that it contains only those operations that are required and identical operations (solving the same Maxwell problem twice) do not appear.

Finally we are going to discuss an algorithm for generating the light path tree automatically. In particular we want to detect the sparsity based on a ray tracing approximation using a pilot ray. First let us introduce a data set that describes a pilot ray:

$$\{r\} = (r_p, r_d, r_i) = (position, direction, intensity). \tag{27}$$

Now we define two types of operators for a pilot ray: (i) $\tilde{C}_j$ - for a given pilot ray on Γ_j the effect of the domain Ω_j on the pilot ray is computed. (ii) $\tilde{P}_{ji}$ - for a given pilot ray on Γ_i the effect of the free space between Γ_i and Γ_j is computed. Additionally, we may control the termination by using intensity-rules for the pilot

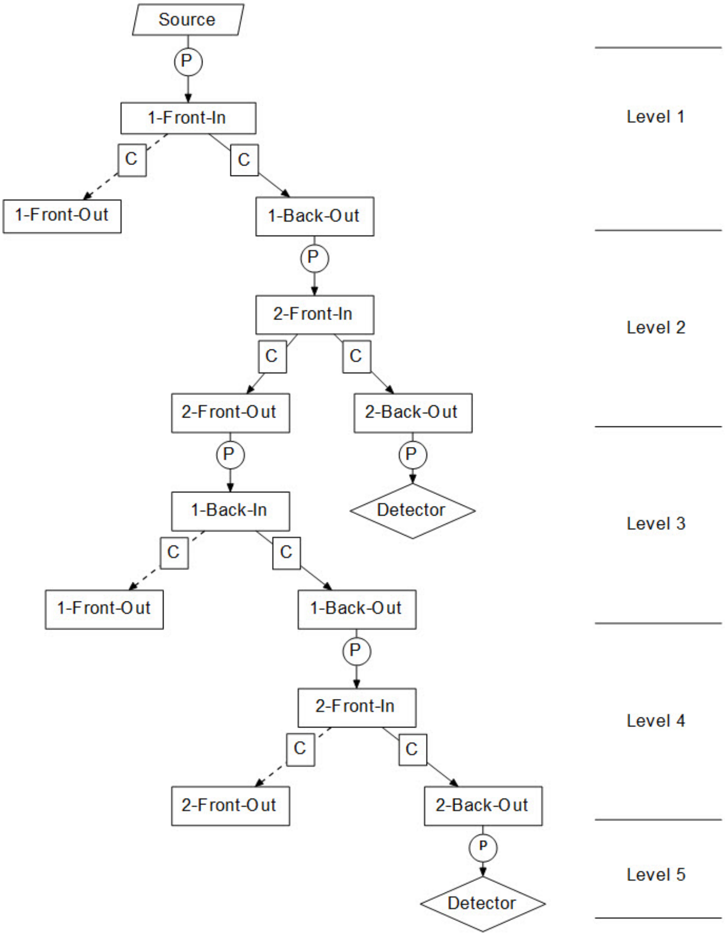


Fig. 4 The light path tree for the 2-plate example and the truncated sum at $k = 5$.

ray. For that purpose, we initialize the intensity at the source with value one. The operators $\tilde{P}_{ji}$ and $\tilde{C}_j$ then manipulate the intensity by considering absorption effects, Fresnel effects at interfaces and any other effects. For the given source we define the root node n^0 of the tree and assign an initial pilot ray with intensity 1. The algorithm *AddNodes* building the tree is called recursively for a list of nodes $\mathscr{S}$ with the initial list $\mathscr{S} = \{n^0\}$.

Again, condition (18) guarantees that the algorithm for the tree generation terminates.

Algorithm 1. AddNodes($\mathscr{S}$)

```
$\hat{\mathscr{S}} := \emptyset$, $\mathscr{S} := \{n^0\}$
for all Nodes $n \in \mathscr{S}$ do
    Let $i$ be the index of the boundary associated with $n$.
    Let $\{r\}$ be the pilot ray associated with $n$.
    Use $\tilde{P}_{ji}$ and apply it to $\{r\}$ in order to find the index $j$ such that the intersection point of the ray $\{r\}$ with $\Gamma_j$ is the closest one for all boundaries.
    if (no intersection is found) then
        BREAK
    end if
    Create a new node $\hat{n}$ associated with $\Gamma_j$. Add the node to the tree. Connect $\hat{n}$ and $n$ by a (ii)-type connection.
    if ($\Gamma_j$ is a detector panel) then
        Assign $\mathbf{V}_j^{\mathrm{source}}$ to $\hat{n}$.
    else
        Compute a new pilot ray $\{\hat{r}\} = \tilde{P}_{ji}\{r\}$ and assign $\{\hat{r}\}$ to $\hat{n}$.
        Apply $\tilde{C}_j$ to $\{\hat{r}\}$. For each output ray $\{\hat{\hat{r}}\}$, create a new node $\hat{\hat{n}}$. Assign the boundary and $\{\hat{\hat{r}}\}$ to $\hat{\hat{n}}$.
        Add $\hat{\hat{n}}$ to the tree and connect $\hat{\hat{n}}$ with $\hat{n}$ by a (i)-type connection.
        Add $\hat{\hat{n}}$ to $\hat{\mathscr{S}}$.
    end if
end for
if ($\|\hat{\mathscr{S}}\| > \delta$) then
    AddNodes($\hat{\mathscr{S}}$)
end if
```

4 Field Tracing Methods

In the previous sections we have described algorithms for the solution of the optical simulation task based on tearing and interconnecting techniques. It has been shown that the algorithm requires two classes of operators. The operators $\mathscr{P}$ describe the free-space propagation between arbitrary scatterers and the operators $\mathscr{C}$ describe the scattering effect of optical components. It remains to define explicit formulations for these operators. The question arises, whether $\mathscr{P}$ and $\mathscr{C}$ have to be rigorous Maxwell solvers. If yes, the methods would be restricted to those known from physical optics, including the most prominent ones as finite and boundary element methods, finite difference and finite integration techniques. However, we know from classical optical modeling and design being used for some decades now, that geometrical optics methods and other approximate methods are extremely powerful techniques to describe free space propagation $\mathscr{P}$ and the effect $\mathscr{C}$ of various important optical components on harmonic fields. Those approximations are often sufficient, since the design of the optical system is such that the solution of local Maxwell problems satisfies certain conditions. A typical example for such a restriction are locally paraxial fields as they occur in classical laser systems. Hence, the practical experience strongly encourages the usage of different rigorous and approximate local

Maxwell solvers for the operators $\mathscr{P}$ and $\mathscr{C}$. Of course, any suitable modeling techniques which is to be used for some subdomain of a system must be formulated for electromagnetic harmonic fields. This is being emphasized here, since it has not been the standard in optical modeling in the past. As a result the system modeling is not based on one technique, but on the smooth combination of techniques, which are accurate enough per subdomain. That is what we call unified optical modeling. In this approach harmonic fields are traced through the system in form of different operators $\mathscr{P}$ and $\mathscr{C}$ according to the equations given before. We refer to this as field tracing, which is the natural generalization of ray tracing, in which ray bundles are traced through all subdomains of a system by geometrical optics. In summary, the tearing and interconnecting algorithm, together with a suitable selection of harmonic field tracing techniques for $\mathscr{P}$ and $\mathscr{C}$ for different subdomains, leads to a unified optical modeling by field tracing. The resulting concept of non-sequential field tracing provides the natural generalization of non-sequential ray tracing in the modeling of optical systems.

Some possible choices for the operators $\mathscr{P}$ and $\mathscr{C}$ have been presented in [12]. Therein, several free space operators are derived and their approximation properties are being discussed. A rigorous version of the operator $\mathscr{P}$ can be directly concluded from the plane wave decomposition of a harmonic field at $z = 0$. This decomposition is described by the Fourier transformation $\mathscr{F}$ for any component $\bar{V}_\ell(x,y)$ of the harmonic field into the k-space by [1]:

$$A_\ell(\kappa) = \mathscr{F}\bar{V}_\ell(\rho) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \bar{V}_\ell(\rho)\,\mathrm{e}^{-\mathrm{i}\kappa\cdot\mathbf{rho}}\,d\rho \tag{28}$$

with $\kappa = (k_x, k_y)$, $\rho = (x,y)$ and $\ell = 1,\ldots,6$. Its inverse version is given by

$$\bar{V}_\ell(\rho) = \mathscr{F}^{-1}A_\ell(\kappa) = \frac{1}{2\pi}\int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} A_\ell(\kappa)\,\mathrm{e}^{\mathrm{i}\kappa\cdot\rho}\,d\kappa\,. \tag{29}$$

For its derivation of the spectrum of plane waves operator $\mathscr{P}_{\mathrm{SPW}}$ we use the fact, that the propagation of each plane wave is described by the multiplication of the phase term $\mathrm{e}^{\mathrm{i}k_z z}$ [2, 5]. The vector component k_z is given by

$$k_z = \sqrt{k_0^2 n^2 - (k_x^2 + k_y^2)}\,. \tag{30}$$

The operator is defined by

$$\begin{aligned}\bar{V}_\ell(\rho,z) &= \mathscr{P}_{\mathrm{SPW}}\bar{V}_\ell(\rho,0)\\ &= \mathscr{F}^{-1}\Big\{\mathscr{F}\{\bar{V}_\ell(\rho,0)\}\exp[\mathrm{i}k_z z]\Big\}\,.\end{aligned} \tag{31}$$

The SPW-operator does not introduce a physical approximation. Let us discuss its numerical properties. The bandwidth of the field component is an invariant of the propagation. That can be concluded from the rigorous SPW operator (29). The spectrum is multiplied with the phase factor $\exp[ikz]$. That step does not change the extent of the spectrum which is nothing else than the bandwidth of the field. According to the sampling theorem [1] an invariant bandwidth directly leads to the conclusion, that the (maximum) sampling period of the field is also an invariant of the propagation. In order to apply (31) to a sampled field, two discrete Fourier transformations are required. Their numerical effort is almost optimal $(O(N\log N))$ with respect to the number of sampling points N. The number of sampling points is defined from the sampling period (does not change by propagation) and the maximum field size between input and output. Hence, the SPW operator has an almost optimal numerical effort if the field size does not change considerably by the propagation. This is the case for paraxial field with a small divergence angle. However, the numerical effort to evaluate the result might become infeasible for non-paraxial fields. In such a case the field size after the propagation might be much larger than the field size at $z = 0$. That is why approximate operators as the Fresnel operator, which is suited for the paraxial case, or the far field operator, which is suited for the far field case, have to be used, if possible. In [12], it has been shown how an automatic procedure for selecting the appropriate operator can be designed yielding an automatic selection free space propagation operator. Also fast boundary element methods could be used instead.

For the component propagation operator $\mathscr{C}$ geometrical optic methods are widely used. A discussion of them can also be found in [12]. In [8], the formulation of scattering problems is discussed for finite element methods which can also be used for $\mathscr{C}$. Let us discuss shortly some efficiency issue. In the framework of field tracing the resulting finite element systems for one subdomain remain constant throughout the iterative process. The iterates enter the boundary condition, i.e. the right hand side of the equations, only. Within the field tracing iteration process one and the same system has to be solved multiple times for different right hand sides. That is, using direct solvers for solving finite element systems, can be very efficient here since the computationally expensive matrix-decomposition can be re-used.

Let us discuss here a special case for the operator $\mathscr{C}$, namely an operator $\mathscr{C}$ for a plane interface component. That is we consider a planar boundary between two homogeneous media of real refractive indices n_i and n_t, located at $z = 0$. We assume that the boundaries are perpendicular to the z axis. A plane wave propagating in the xz plane is assumed incident from the material of refractive index n_i at an angle θ_i. Since the boundary is infinite, a single reflected and a single transmitted plane wave propagating in the xz plane are generated by this interaction. The waves propagate in direction defined by angles θ_r and θ_t.

In the quasi-two-dimensional geometry Maxwell's equation divide into two sets, which can be solved separately. One of the sets involves only the y components of the electric field (and the x and z components of the magnetic field), and one talks about TE polarization. The other set contains only the y component of the magnetic field together with the x and z components of the electric field, and then one talks about TM polarization. Both polarization cases allow a straightforward evaluation of the boundary conditions. Considering, for example, the TE polarization and denoting the complex amplitude of the incident wave by $E_{\mathrm{i}}^{\mathrm{TE}}$, this wave be expressed in the form

$$E_{\mathrm{i}y}(x,z) = E_{\mathrm{i}}^{\mathrm{TE}} \exp\left[\mathrm{i}k_0 n_{\mathrm{i}}\left(x\sin\theta_{\mathrm{i}} + z\cos\theta_{\mathrm{i}}\right)\right]. \tag{32}$$

Similarly, the reflected wave with amplitude $E_{\mathrm{r}}^{\mathrm{TE}}$ is written as

$$E_{\mathrm{r}y}(x,z) = E_{\mathrm{r}}^{\mathrm{TE}} \exp\left[\mathrm{i}k_0 n_{\mathrm{i}}\left(x\sin\theta_{\mathrm{r}} - z\cos\theta_{\mathrm{r}}\right)\right], \tag{33}$$

where we have noted that the wave propagates backwards in the incident medium. Finally, the transmitted wave with complex amplitude $E_{\mathrm{t}}^{\mathrm{TE}}$ has the expression

$$E_{\mathrm{t}y}(x,z) = E_{\mathrm{t}}^{\mathrm{TE}} \exp\left[\mathrm{i}k_0 n_{\mathrm{t}}\left(x\sin\theta_{\mathrm{t}} + z\cos\theta_{\mathrm{t}}\right)\right]. \tag{34}$$

Next the continuity conditions for $E_y, H_z, \partial E_y/\partial x, H_x, \partial E_y/\partial z$ are applied to obtain equations which determine the free parameters of the reflected and transmitted plane waves. Applying them at $(x,z) = (0,0)$ leads to three equations between $E_{\mathrm{i}}^{\mathrm{TE}}$, $E_{\mathrm{r}}^{\mathrm{TE}}$, $E_{\mathrm{t}}^{\mathrm{TE}}$, θ_{i}, θ_{r}, and θ_{t}, from which we obtain straightforwardly the *law of reflection*

$$\theta_{\mathrm{r}} = \theta_{\mathrm{i}}, \tag{35}$$

as well as *Snell's law* of refraction

$$n_{\mathrm{t}}\sin\theta_{\mathrm{t}} = n_{\mathrm{i}}\sin\theta_{\mathrm{i}}, \tag{36}$$

which defines the angle θ_{t}. Moreover, we obtain the relations

$$r_{\mathrm{TE}} = \frac{E_{\mathrm{r}}^{\mathrm{TE}}}{E_{\mathrm{i}}^{\mathrm{TE}}} = \frac{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} - n_{\mathrm{t}}\cos\theta_{\mathrm{t}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} + n_{\mathrm{t}}\cos\theta_{\mathrm{t}}} \tag{37}$$

and

$$t_{\mathrm{TE}} = \frac{E_{\mathrm{t}}^{\mathrm{TE}}}{E_{\mathrm{i}}^{\mathrm{TE}}} = \frac{2n_{\mathrm{i}}\cos\theta_{\mathrm{i}}}{n_{\mathrm{i}}\cos\theta_{\mathrm{i}} + n_{\mathrm{t}}\cos\theta_{\mathrm{t}}} \tag{38}$$

for the reflected and refracted field amplitudes. These are called *Fresnel's equations* for TE polarized light.

The two operators being presented here, are used for the plane interface problem discussed in the next section as a model problem. There we assume a paraxial setting with $\theta_i = 0$.

5 Numerical Examples

The performance of the non-sequential field tracing algorithm can be demonstrated using a setup which is used in a Fabry-Perot interferometer in practice. In particular we consider a sequence of parallel plates, see Fig. 5. We introduce a further decomposition: a plate is splitted into two boundaries. Then we apply the spectrum of plane waves operator in the homogeneous medium (air or medium of the plate) and we apply the scattering operator of (37)–(38) at each boundary.

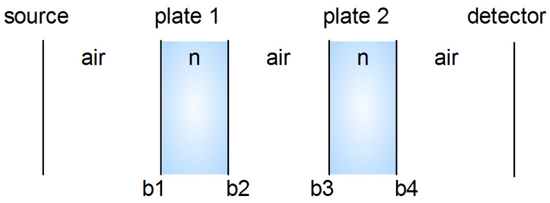


Fig. 5 Experimental setup of multiple plates. We consider air ($n = 1$) between the plates and a variation of the refractive index n of the plates.

The plates are placed in air with refractive index $n = 1.00027$. In the experiments we vary the refractive index of the plates, the number of plates and the thickness of the plates. The distance of plates (if more than one) is 5mm. We use a plane wave light source with a diameter of several millimeters. The wavelength is also varied in the experiments. Absorption is not considered.

In the first series of experiments we investigate the convergence of the algorithm. For that purpose we apply the pilot ray algorithm and observe the convergence of r_k (see (25)).The results are shown in Fig. 6.

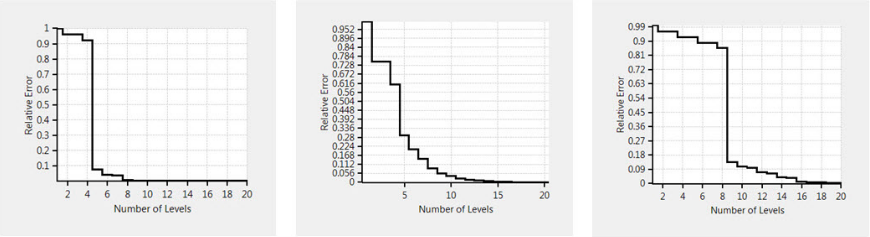


Fig. 6 Convergence results for different setups: 2 plates with $n = 1.5$ (left), 2 plates with $n = 3.0$ (middle) and 4 plates with $n = 1.5$ (right).

We have considered two plates (4 boundaries) with $n = 1.5$, two plates (4 boundaries) with $n = 3.0$ and four plates (8 boundaries) with $n = 1.5$. In order to reach an error below 0.01, the required number of iterations are 8 (2 plates, $n = 1.5$), 13 (2 plates, $n = 3.0$) and 17 (4 plates, $n = 1.5$), respectively.

In the second series of experiments we apply the non-sequential field tracing algorithm to some incident plane wave field, see Fig. 7. We compute the transmission efficiencies for different setups. In order to justify the new method, we compare the results with a rigorous Fourier Modal Method (FMM) [10]. This method considers periodic setups and computes the efficiencies for an infinite incident plane wave. The FMM is well established and we expect a high coincidence of the results computed by the two methods.

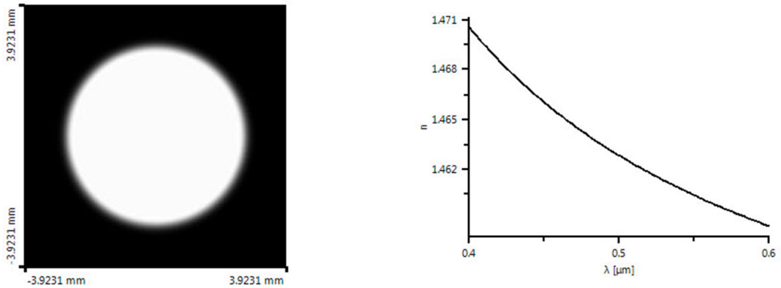


Fig. 7 Amplitude of the incident plane wave (diameter 5mm) on the left. The refractive index *n* of Fused Silica for the wavelength range between 400nm (n=1.4705) and 600nm (n=1.4584) is shown on the right.

In the experiments we also include dispersion effects. These effects occur when the refractive index of materials depends on the wavelength of the light. For that purpose we use Fused Silica as material for the plates. The refractive index is shown in Fig. 7. We vary again some of the system parameters considering a single plate (2 boundaries). In the first series of simulations we vary the thickness of the plate from 1μm to 2μm.

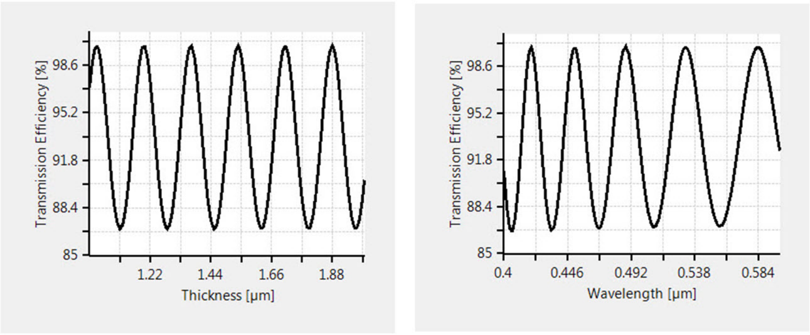


Fig. 8 Transmission efficiency for a single plate of fused silica. Left: wavelength is 500 nm and the thickness is varied between 1μm and 2μm. Right: the thickness is fixed at 2 μm and the wavelength is varied between 400nm (n=1.4705) and 600nm (n=1.4584).

In the second series we vary the wavelength from 400nm to 600nm. The results are shown in Fig. 8. Finally we compare results of the field tracing algorithm with those obtained using FMM.

Table 1 Comparison of transmission efficiencies computed with field tracing (Eff(FT)) and with the Fourier Modal Method (Eff(FMM)).

Wavelength	Thickness	Eff(FT)	Eff(FMM)
500 nm	1.00 μm	97.01%	97.00%
500 nm	1.03 μm	99.90%	99.89%
500 nm	1.11 μm	86.85%	86.87%
500 nm	2.00 μm	91.05%	91.07%
400 nm	2.00 μm	90.94%	90.96%
406 nm	2.00 μm	86.57%	86.60%
420 nm	2.00 μm	99.97%	99.99%
600 nm	2.00 μm	91.98%	92.00%

The results are shown in Table 1. As expected a very good coincidence between all values can be observed.

6 Conclusions

We have presented optical field tracing techniques for the efficient solution of optical simulation problems. The resulting algorithm allows the combination of local Maxwell solvers including rigorous and approximate methods. In optics, local problems are often well behaved and local solvers can be adapted to the local properties in order to speed up calculations. Further research is required to detect and to classify these local properties. Such information can be used to design local solvers appropriately. The solution algorithm presented in the paper can easily run in parallel. In particular all local solvers of one tree level can run in parallel in a distributed computing environment. Then, communication requires the exchange of field data associated with the boundaries of the subdomains only. Though field tracing is basically formulated for propagating harmonic fields through optical systems, it can also be applied to general fields like stationary and pulsed light [9, 11]. To this end the general field can be decomposed into a set of harmonic modes which then can be traced through the system and evaluated by suitable detectors.

Acknowledgements. This is a personal acknowledgment by M. Kuhn, Ph.D. student of Ulrich Langer. I always enjoyed being a student and colleague of Ulrich Langer. I would like to thank him for his marvelous education. Ulrich has an outstanding ability to state scientific problems which are of great practical relevance. He encouraged me to work in the field of Maxwell problems. Finally this has lead me being part of a great development team of the very innovative optical simulation tool VirtualLab^{TM} [6]. I would like to dedicate this paper to Ulrich. Ultimately, his premises ”Set up a mathematical formulation for the problem to be solved!” was the starting point for the research the results of which are being presented in this paper. Last but not least, this paper would have never been possible without the contributions of my co-authors and colleagues at LightTrans.

References

[1] Brigham, E.O.: The fast Fourier transform. Prentice-Hall, Englewood Cliffs (1974)

[2] Goodmann, J.W.: Introduction to Fourier optics. McGraw-Hill, New York (1968)

[3] Langer, U., Steinbach, O.: Coupled boundary and finite element tearing and interconnecting methods. In: Kornhuber, R., Hoppe, R., Periaux, J., Pironneau, O., Widlund, O.B., Xu, J. (eds.) Domain Decomposition Methods in Science and Engineering XV. Lecture Notes in Computational Sciences and Engineering, vol. 40, pp. 83–97. Springer, Heidelberg (2003)

[4] Langer, U., Steinbach, O.: Coupled finite and boundary element domain decomposition methods. In: Schanz, M., Steinbach, O. (eds.) Boundary Element Analysis. Mathematical Aspects and Applications. LNACM, vol. 29, pp. 61–95. Springer, Berlin (2007)

[5] Mandel, L., Wolf, E.: Optical coherence and quantum optics. Cambridge University Press, Cambridge (1995)

[6] LightTrans GmbH: VirtualLab^{TM} - your optical modeling laboratory (2000-2012), `http://www.lighttrans.com`

[7] Meyer, C.D.: Matrix analysis and applied linear algebra. SIAM (2001)

[8] Monk, P.: Finite Element Methods for Maxwell's Equations. Clarendon Press, Oxford (2003)

[9] Tervo, J., Turunen, J., Wyrowski, F.: The Light Cube. In: 5th EOS Topical Meeting on Advanced Imaging Techniques, vol. 3037, European Optical Society (2010)

[10] Turunen, J.: Diffraction theory of microrelief gratings. In: Herzig, H.P. (ed.) Micro-optics Elements, Systems and Applications, pp. 31–52. Taylor & Francis, London (1997)

[11] Wyrowski, F., Hellmann, C., Krieg, R., Schweitzer, H.: Modeling the propagation of ultrashort pulses through optical systems. In: Neev, A., Nolte, J., Trebina, R.P. (eds.) Proc. SPIE, Heisterkamp, San Francisco, vol. 7589 (2010)

[12] Wyrowski, F., Kuhn, M.: Introduction to field tracing. J. Modern Optics 58(5-6), 449–466 (2011)

Boundary Element Method for Linear Elasticity with Conservative Body Forces

Heiko Andrä, Richards Grzhibovskis, and Sergej Rjasanow

Abstract. A boundary integral formulation for a mixed boundary value problem in linear elastostatics with a conservative right hand side is considered. A meshless interpolant of the scalar potential of the volume force density is constructed by means of radial basis functions. An exact particular solution to the Lamé system with the gradient of this interpolant as the right hand side is found. Thus, the need of approximating the Newton potential is eliminated. The procedure is illustrated on numerical examples.

1 Introduction

In presence of a body force (i.e. volume force density) a boundary integral formulation of a boundary value problem (BVP) in linear elastostatics contains not only surface integrals, but also a volume integral over the domain. The volume integration is necessary to evaluate the action of the Newton potential operator on the right hand side of the Lamé system. Thus, the discrete version of the formulation - the boundary element method (BEM) - involves not only a surface mesh, but also some kind of volume discretization. Volume meshing of domains with complex geometry is a difficult task. The corresponding algorithms often lead to meshes with rough approximations of the geometry/boundaries or badly shaped finite elements [24]. If the volume force density has a simple special form (e.g. polynomial or constant) the need for the Newton potential can be eliminated by finding a particular solution to

Heiko Andrä
Fraunhofer ITWM, Fraunhofer–Platz 1, 67663 Kaiserslautern, Germany
e-mail: heiko.andrae@itwm.fraunhofer.de

Richards Grzhibovskis · Sergej Rjasanow
Institut für Angewandte Mathematik, Universität des Saarlandes,
66041 Saarbrücken, Germany
e-mail: richards@num.uni-sb.de, rjasanow@num.uni-sb.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 275–297.
springerlink.com

the non-homogeneous Lamé system, and, thus, reducing the original problem to a BVP without body forces.

In this study, we consider a substantially larger family of the volume force densities, namely, the conservative ones. Despite their special form, such body forces appear in practical applications in a number of cases [3],[13]. Among them are problems in thermoelasticity and poroelasticity (see Sect. 3). Since the body force is conservative, we assume that its scalar potential is given. The task of finding a particular solution to the Lamé system in this case can be reduced to one for the Poisson equation.

Furthermore, we consider settings, in which the scalar potential is given only at some points inside the domain. In this case, we construct a smooth interpolant for the potential by means of radial basis functions (RBFs). A closed form for the exact particular solution is obtained, assuming the right hand side of the Lamé system is the gradient of such interpolant. This solution is then utilized to find the approximate solution to the original BVP.

The article is organized as follows. An overview of techniques for dealing with body forces within the framework of boundary element method is given in Sect. 2. In Sect. 3, we formulate an inhomogeneous three-dimensional system of Lamé equations subjected to a body force and mixed boundary conditions. Some examples of conservative body forces are presented. Next, we give a boundary integral formulation of the problem including the Newton potential. Finally, the particular solution approach is discussed leading to a homogeneous system of Lamé equations. Sect. 4 contains a short description of the standard Galerkin procedure with piecewise constant and piecewise linear basis and test functions. Some comments to the application of the Adaptive Cross Approximation to the resulting system of linear equations follow. A possibility to obtain an exact particular solution in the case where the potential of the conservative body force is a polynomial is presented in Sect. 5. The more complicated case, where the potential is a general function, is the main part of the paper and will be discussed in details in Sect. 6. Here, we give a short description of radial basis functions and describe a possibility to get an approximation for a particular solution. Two numerical examples are presented in the final Sect. 7. The first example shows the numerical results obtained for a problem from thermoelasticity. Here we show the quality of the radial basis functions approximation of the given temperature field. The second example is an application of the method to a body subjected to the gravitational force.

2 Short Review of Techniques for Treating Volume Integrals in BEM

The most simple method for treating volume integrals is the discretization of the volume Ω in so-called cells or interior elements. The numerical integration is performed by using a quadrature rule in each cell. The mesh size in the volume can be chosen independently from the mesh size on the surface and hanging nodes are allowed. This cell integration is convenient for problems with concentrated forces

$$\underline{f}(x) = \sum_{k=1}^{N_f} \underline{F}_k \delta(x - x^{(k)}),$$

at points $x^{(k)} \in \Omega$, where δ denotes Dirac's delta distribution and N_f is a positive number. However, introducing volume discretizations removes one of the most important advantages of BEM.

As a method without using interior cells, a Monte-Carlo method for the computation of volume integrals was proposed by Gipson [16]. The advantage is that the integral can be computed for a very general class of functions $\underline{f}$. The Monte-Carlo method is computationally very expensive.

To avoid the internal discretization, many more or less general and efficient methods have been developed to transform domain integrals into equivalent boundary integrals [31]. One of the first methods of this type is the Galerkin tensor (or vector) method [14], which was used by several authors. This method gives accurate results, but can only be applied to the limited range of body forces $\underline{f}$ which are solutions of elliptic PDEs (e.g. harmonic functions). The Galerkin tensor method can be used successfully for body forces arising from constant gravitational load, rotation about a fixed axis, or steady state thermal loading [15].

An efficient technique, which converts domain integrals to boundary integrals for a wider range of body forces $\underline{f}$, is the dual reciprocity method (DRM) [30, 34, 36, 35]. This method is widely used for various applications in engineering [12]. The DRM is based on a superposition of localized particular solutions. For this purpose the inhomogeneous term $\underline{f}$ of the differential equation is interpolated by using certain trial functions. The accuracy of the DRM depends on the choice of the interpolating functions. Golberg and Chen [18] recognized that the theory of RBFs provides the mathematical basis for the choice of trial functions in the DRM. Their numerical tests have shown that a specific choice of RBFs increase the accuracy and efficiency of the the DRM. One can find only few papers on the convergence analysis of the DRM (see e.g. [17, 19, 25]).

One more technique to transform domain integrals into boundary integrals is the so-called multiple reciprocity method (MRM), which was originally proposed by Nowak [32] and further developed by Nowak and Brebbia [10] as a general tool to solve a wide range of parabolic and hyperbolic problems. The MRM consists of a repeated application of the reciprocity theorem using a sequence of higher order fundamental solutions. The MRM was initially applied to the equations of linear elastostatics by Neves [31].

An alternative method to eliminate the domain integral is the particular solution method (PSM), where approximate particular solutions are used instead of transforming domain integrals into boundary integrals. The PSM was developed by Banerjee and co-workers [2, 21] for free vibration and elasticity problems. The PSM is similar to the DRM in several aspects, which is explained in [23]. In the PSM, the most popular choice of basis function is also the RBF. Since the particular solution is not unique, any complex domain can always be imbedded into a rectangular box. For such rectangular boxes, many fast numerical solvers, e.g. multigrid and FFT-based methods, for finite difference, finite volume, or finite element discretizations on regular meshes are employed. The discussion of this huge class of numerical methods would go beyond the scope of this paper. In this connection it should not be forgotten to mention the fast FFT-based solver for the Lippmann-Schwinger equation in elasticity [27], where a integral equation is solved instead of a partial differential equation. The latter method is highly efficient for volume forces in complex microstructured materials.

Finally, a Fourier expansion technique, i.e. an expansion of the source term in Fourier series, can be used to transform domain into boundary integrals. This method was proposed and used by Tang [41] for potential and elasticity problems.

Four methods for evaluating domain integrals associated with boundary element methods are compared with respect to accuracy and computational costs in [23], where the Poisson equation and the Helmholtz equation are considered as example problems. The DRM and PSM were essentially the same in terms of both CPU time and accuracy. The traditional direct domain integration method with and without multipole acceleration always gave better results with respect to accuracy and computational time for the considered examples. However, subsequent benchmarks, where complex multiply-connected three-dimensional geometries are considered, lead to the conclusion that relative efficiencies of the classical cell-based domain integration and two variants of the DRM is evidently problem dependent [22].

A recent study, where a fast evaluation of the Newton potentials by means of the fast multipole method is described and analyzed, can be found in [33].

3 Description of the Problem

Let $\Omega \in \mathbb{R}^3$ be a open simply connected domain with the Lipschitz boundary $\Gamma = \partial\Omega$. The outward unit normal vector at $x \in \Gamma$ is denoted by $\underline{n}(x)$. The domain Ω is filled with a homogeneous, isotropic, elastic medium. The material parameters of the medium are given by the Young modulus $E > 0$ and the Poisson ratio $\nu \in (0, 1/2)$. In linear isotropic elastostatics, the displacement field

$$\underline{u} : \Omega \to \mathbb{R}^3$$

satisfies the system of Lamé equations

$$-\mu \Delta \underline{u}(x) - (\lambda + \mu)\,\mathrm{grad}\,\mathrm{div}\,\underline{u}(x) = \underline{f}(x) \quad \text{for } x \in \Omega\,, \tag{1}$$

where the Lamé constants λ and μ are defined as

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

In (1), $\underline{f}$ denotes the body force per unit volume (body force density). In Subsect. 3.1, we will give some examples for body forces. For a given displacement field $\underline{u}$, the strain tensor $e(\underline{u},\cdot) \in \mathbb{R}^{3\times 3}$ is defined componentwise as follows

$$e_{ij}(\underline{u},x) = \frac{1}{2}\left(\frac{\partial}{\partial x_i}u_j(x) + \frac{\partial}{\partial x_j}u_i(x)\right) \quad \text{for } i,j = 1,2,3\,.$$

For a homogeneous isotropic material, the stress tensor $\sigma(\underline{u},\cdot) \in \mathbb{R}^{3\times 3}$ is given by the Hooke's law

$$\sigma(\underline{u},x) = \lambda\, \mathrm{tr}\, e(\underline{u},x) I + 2\mu\, e(\underline{u},x) \quad \text{for } x \in \Omega\,. \tag{2}$$

The Lamé equations (1) are subject to the mixed boundary conditions

$$\begin{aligned}(\gamma_0\underline{u})(x) &= \underline{g}_D(x) \quad \text{for } x \in \Gamma_D\,,\\ (\gamma_1\underline{u})(x) &= \underline{g}_N(x) \quad \text{for } x \in \Gamma_N\,,\end{aligned} \tag{3}$$

where $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$. Here, γ_0 denotes the interior Dirichlet trace operator

$$(\gamma_0\underline{u})(x) = \lim_{y\to x}\underline{u}(y)\,, \quad \text{for } y \in \Omega\,,\ x \in \Gamma\,,$$

and γ_1 the interior Neumann trace operator or the boundary stress operator

$$(\gamma_1\underline{u})(x) = (\gamma_0\sigma(\underline{u},\cdot))(x)\,\underline{n}(x) \quad \text{for } x \in \Gamma\,.$$

One might also consider problems, where each component $g_{i,D}$, $i = 1,2,3$ of the displacement field $\underline{g}_D$ is prescribed on its own part of the boundary $\Gamma_{i,D}$, $i = 1,2,3$. The corresponding components $g_{i,N}$, $i = 1,2,3$ of the traction forces are specified on the boundary parts $\Gamma_{i,N}$, $i = 1,2,3$. The boundary parts are relatively open, and the relation $\Gamma = \overline{\Gamma}_{i,N} \cup \overline{\Gamma}_{i,D}$ holds for each $i = 1,2,3$. For the sake of notational simplicity we assume, that the boundary conditions are given in the form (3), i.e. $\Gamma_{1,D} = \Gamma_{2,D} = \Gamma_{3,D} = \Gamma_D$. In these settings, the BVP (1), (3) is uniquely solvable in $\left[H^1(\Omega)\right]^3$ when $meas(\Gamma_D) > 0$, $\underline{g}_D \in \left[H^{1/2}(\Gamma_D)\right]^3$, $\underline{g}_N \in \left[H^{-1/2}(\Gamma_N)\right]^3$, and $\underline{f} \in \left[H^{-1}(\Omega)\right]^3$ (see e.g. [40]).

3.1 Examples of the Body Forces

In contrast to contact forces $\underline{g}_N$ given on the Neumann part Γ_N of the boundary Γ, a body force density $\underline{f}$ in (1) is responsible for a force

$$\underline{F} = \int\limits_V \underline{f}(x)\,dx$$

that acts throughout the part $V \subseteq \Omega$ of the volume of a body. Common examples of body forces include gravity, thermal stress, electric and magnetic forces. Furthermore, inertial forces like centrifugal force and Coriolis force can be considered as body forces.

Conservative and Non-conservative Body Forces

Forces can be classified into two classes by the way they transfer energy from an body. The first class conserves energy in the form of potential energy and transfers the same amount of energy back when the motion is reversed. This class of forces that conserve energy is called conservative forces. The work of a conservative force is path independent [20]. Therefore, conservative forces fulfill the following condition

$$\underline{f} = -\operatorname{grad}\theta,$$

where $\theta : \Omega \to \mathbb{R}$ denotes the scalar potential. Some examples of conservative forces are given below. The second class of forces which do not conserve energy are called non-conservative or dissipative forces. Typical non-conservative forces are friction, plasticity and air drag.

Gravity

Gravitational force appears due to a body's own weight. The particular significance of the gravitational force lies for example in civil engineering applications, like dam and bridge design. The gravitational force density has the form

$$\underline{f}(x) = -\rho\, g\, \underline{e}_3\,, \quad \underline{e}_3 = (0,0,1)^\top,$$

where ρ is the constant mass density of the medium and g denotes the constant gravitational acceleration. Thus, the gravitational force is conservative and

$$\underline{f}(x) = -\operatorname{grad}\theta(x)\,, \quad \theta(x) = \rho\, g x_3 \quad \text{for}\, x = (x_1,x_2,x_3)^\top \in \Omega\,. \tag{4}$$

Inertial Forces

One example of the inertial forces is the centrifugal force. It represents the effects of inertia of a body under rotation and which is a force away from the center of rotation $\hat{x} \in \mathbb{R}^3$. This force density can be written as

$$\underline{f}(x) = \rho\,(2\pi\omega)^2\,(x-\hat{x})\,,$$

where ρ is the constant mass density of the medium and ω denotes the angular frequency. Thus, the centrifugal force is conservative and

$$\underline{f}(x) = -\operatorname{grad}\theta(x), \quad \theta(x) = -\frac{1}{2}\rho\,(2\pi\omega)^2\,|x-\hat{x}|^2 \quad \text{for}\, x \in \Omega\,. \tag{5}$$

Thermoelasticity

We consider an elastic solid occupying the domain Ω that is assumed to be stress free at an uniform reference temperature T_0. For a variable temperature T, there is an additional thermal stress and the constitutive relation (Hooke's law) (2) reads

$$\sigma(\underline{u},x) = \lambda\,\mathrm{tr}\,e(\underline{u},x)\,I + 2\,\mu\,e(\underline{u},x) - (3\lambda+2\mu)\,\alpha(T(x)-T_0)\,I\,, \tag{6}$$

where the additional parameter α is called coefficient of thermal expansion. Thus, the body force density is conservative and given by

$$\underline{f}(x) = -\operatorname{grad}\theta(x), \quad \theta(x) = (3\lambda+2\mu)\,\alpha\,T(x) \quad \text{for}\, x \in \Omega\,. \tag{7}$$

In the uncoupled case, the temperature field can be determined independently of the stress calculations solving an elliptic heat equation. Once the temperature is obtained, elastic stress solver with a non-trivial body force $\underline{f}$ will be employed to complete the problem solution.

Poroelasticity

We consider the fluid flow through a poroelastic medium in a domain Ω. The fluid flow can be modelled by the Stokes-Brinkman equations [11] for the velocity $\underline{v}$

$$\begin{aligned} -\mu\triangle\underline{v} + \kappa^{-1}\underline{v} + \operatorname{grad}p &= 0 \quad \text{in} \quad \Omega, \\ \operatorname{div}\underline{v} &= 0 \quad \text{in} \quad \Omega, \end{aligned}$$

where κ is the permeability and μ is the viscosity. The Stokes-Brinkman system is coupled with the elasticity equations via the pressure p:

$$\operatorname{div}\sigma = -\operatorname{grad}p \quad \text{in} \quad \Omega.$$

The body force $\underline{f} = -\operatorname{grad}p$ in the elasticity equations is conservative, where the potential is the pressure p.

3.2 Boundary Integral Formulation

The solution of the inhomogeneous boundary value problem (1),(3) inside of the domain Ω can be described by the Somigliana identity for $x \in \Omega$

$$\underline{u}(x) = \int\limits_{\Gamma} U^*(x,y)(\gamma_1 \underline{u})(y)\, ds_y - \int\limits_{\Gamma} \gamma_{1,y} U^*(x,y)(\gamma_0 \underline{u})(y)\, ds_y + \int\limits_{\Omega} U^*(x,y) \underline{f}(y)\, dy\,.$$

The problem (1),(3) admits the following symmetric boundary integral formulation (for details, we refer to [37, 39])

$$V\underline{t}(x) - K\underline{g}(x) = \Big(\frac{1}{2}I + K\Big)\underline{\tilde{g}}_D(x) - V\underline{\tilde{g}}_N(x) - N_0\underline{f}(x) \quad \text{for } x \in \Gamma_D\,, \tag{8}$$

$$K'\underline{t}(x) + D\underline{g}(x) = \Big(\frac{1}{2}I - K'\Big)\underline{\tilde{g}}_N(x) - D\underline{\tilde{g}}_D(x) - N_1\underline{f}(x) \quad \text{for } x \in \Gamma_N\,, \tag{9}$$

where $\underline{g} = \gamma_0 \underline{u} - \underline{\tilde{g}}_D$, $\underline{t} = \gamma_1 \underline{u} - \underline{\tilde{g}}_N$, and $\underline{\tilde{g}}_N$,$\underline{\tilde{g}}_D$ are extensions of $\underline{g}_N$ and $\underline{g}_D$ to the whole boundary Γ. The following boundary integral operators are involved in (8),(9)

$$V\underline{w}(x) = \gamma_0 \int\limits_{\Gamma} U^*(x,y)\underline{w}(y)\, ds_y\,, \tag{10}$$

$$K\underline{v}(x) = \frac{1}{2}\underline{v}(x) + \gamma_0 \int\limits_{\Gamma} \gamma_{1,y} U^*(x,y)\underline{v}(y)\, ds_y\,, \tag{11}$$

$$K'\underline{w}(x) = -\frac{1}{2}\underline{w}(x) + \gamma_1 \int\limits_{\Gamma} U^*(x,y)\underline{w}(y)\, ds_y\,, \tag{12}$$

$$D\underline{v}(x) = -\gamma_1 \int\limits_{\Gamma} \gamma_{1,y} U^*(x,y)\underline{v}(y)\, ds_y\,, \tag{13}$$

$$N_0\underline{f}(x) = \gamma_0 \int\limits_{\Omega} U^*(x,y)\underline{f}(y)\, dy\,, \tag{14}$$

$$N_1\underline{f}(x) = \gamma_1 \int\limits_{\Omega} U^*(x,y)\underline{f}(y)\, dy\,. \tag{15}$$

These are the single layer, double layer, adjoint double layer potentials, hypersingular, and Newton potential operator, respectively. For the mapping properties of the above operators, we refer to [37].

3.3 Particular Solution

The main drawback in considering the boundary integral equations (8), (9) is the evaluation of the Newton potential leading to a discretization of the domain Ω.

Besides this direct computation, there exist several more efficient techniques. One of them is the particular solution approach. Let $\underline{u}_p$ be a particular solution of the Lamé equations (1), i.e.

$$-\mu\Delta\underline{u}_p(x)-(\lambda+\mu)\operatorname{grad}\operatorname{div}\underline{u}_p(x)=\underline{f}(x)\quad \text{for } x\in\Omega\,.$$

Then, we decompose the solution in a sum

$$\underline{u}(x)=\underline{u}_0(x)+\underline{u}_p(x)\quad \text{for } x\in\Omega\,,$$

where $\underline{u}_0$ solves the homogeneous problem

$$-\mu\Delta\underline{u}_0(x)-(\lambda+\mu)\operatorname{grad}\operatorname{div}\underline{u}_0(x)=\underline{0}\quad \text{for } x\in\Omega\,, \tag{16}$$

with the modified boundary conditions

$$\begin{aligned}(\gamma_0\underline{u}_0)(x)&=\underline{g}_D(x)-(\gamma_0\underline{u}_p)(x)\quad \text{for } x\in\Gamma_D\,,\\(\gamma_1\underline{u}_0)(x)&=\underline{g}_N(x)-(\gamma_1\underline{u}_p)(x)\quad \text{for } x\in\Gamma_N\,.\end{aligned} \tag{17}$$

4 Boundary Element Method

4.1 Galerkin Procedure

In the Sect. 5 and 6, we will give some exact and approximate formulae for a particular solution. Thus, it is assumed in this section, that the problem is homogeneous. To obtain a boundary element discretization, we approximate

$$\Gamma\approx\Gamma_h=\sum_{\ell=1}^{N}\overline{\tau_\ell}$$

by a conform surface triangulation with N triangles and M nodes. We use the piecewise constant functions (ψ_ℓ is 1 on triangle τ_ℓ and 0 outside τ_ℓ) as basis and test functions for the discretized single layer potential (10). These functions also serve as test functions for the double layer potential (11) and as trial functions for the adjoint double layer potential. The basis functions for the hypersingular and the double layer potentials are chosen to be piecewise linear and continuous with $\varphi_j(x_i)=\delta_{ij}$, φ_j is linear on each τ_ℓ. These functions are also used as test functions for the hypersingular operator (13) and adjoint double layer potential (12). The ansatz for the boundary data is

$$(\gamma_0\underline{u})\,(x)\approx\underline{g}_h(x)=\sum_{i=1}^{M}\mathbf{g}_i\varphi_i(x)\,, \tag{18}$$

$$(\gamma_1 \underline{u})(x) \approx \underline{t}_h(x) = \sum_{j=1}^{N} \mathbf{t}_j \psi_j(x), \tag{19}$$

where it can be assumed with no loss of generality that the coefficient vectors have the form

$$\begin{aligned} \mathbf{g} &= \{\mathbf{g}_i\}_{i=1}^{M} = (\mathbf{g}_N, \mathbf{g}_D)^\top \in \mathbb{R}^{3M}, \quad \mathbf{g}_N \in \mathbb{R}^{3(M-M_D)}, \\ \mathbf{t} &= \left\{\mathbf{t}_j\right\}_{j=1}^{N} = (\mathbf{t}_N, \mathbf{t}_D)^\top \in \mathbb{R}^{3N}, \quad \mathbf{t}_D \in \mathbb{R}^{3(N-N_N)}. \end{aligned}$$

Here the last M_D coefficient vectors of $\mathbf{g}$ approximate the prescribed Dirichlet datum $\underline{g}_D$ on $\Gamma_{h,D}$, and the first N_N coefficient vectors of $\mathbf{t}$ approximate the given Neumann datum $\underline{g}_N$ on $\Gamma_{h,N}$. Thus the Galerkin discretization of (8),(9) leads to the linear system for the unknown coefficients $\mathbf{t}_D$ and $\mathbf{g}_N$

$$\mathbf{P} \begin{pmatrix} V_h & -K_h \\ K_h^T & D_h \end{pmatrix} \mathbf{P}^\top \begin{pmatrix} \mathbf{t}_D \\ \mathbf{g}_N \end{pmatrix} = \mathbf{P} \begin{pmatrix} -V_h & \frac{1}{2}M_h + K_h \\ \frac{1}{2}M_h - K_h^T & -D_h \end{pmatrix} \overline{\mathbf{P}}^\top \begin{pmatrix} \mathbf{t}_N \\ \mathbf{g}_D \end{pmatrix}, \tag{20}$$

where

$$\begin{aligned} \mathbf{P} &= \begin{pmatrix} P_N & 0 \\ 0 & P_D \end{pmatrix}, \quad P_N = \begin{pmatrix} 0 \; I \end{pmatrix}, \quad P_D = \begin{pmatrix} I \; 0 \end{pmatrix}, \\ \overline{\mathbf{P}} &= \begin{pmatrix} \overline{P}_N & 0 \\ 0 & \overline{P}_D \end{pmatrix}, \quad \overline{P}_N = \begin{pmatrix} I \; 0 \end{pmatrix}, \quad \overline{P}_D = \begin{pmatrix} 0 \; I \end{pmatrix}, \\ P_N, \overline{P}_N &\in \mathbb{R}^{3(N-N_N)\times 3N}, \quad P_D, \overline{P}_D \in \mathbb{R}^{3(M-M_D)\times 3M}, \end{aligned}$$

and I denotes a unit matrix of the appropriate dimension. The fully populated matrices in (20) are composed of the following three by three blocks

$$\left(V_h\right)_{k\ell} = \left\langle V\left(\psi_\ell \mathbf{i}\right), \psi_k \mathbf{i}\right\rangle, \quad \left(K_h\right)_{kj} = \left\langle K\left(\varphi_j \mathbf{i}\right), \psi_k \mathbf{i}\right\rangle, \quad \left(D_h\right)_{ij} = \left\langle D\left(\varphi_j \mathbf{i}\right), \varphi_i \mathbf{i}\right\rangle,$$

where $\mathbf{i}$ is a three by three identity matrix, $\langle\cdot,\cdot\rangle$ denotes the scalar product in $L^2\left(\Gamma\right)$, and $k,\ell = 1\ldots N$, $i,j = 1\ldots M$. The sparse mass matrix M_h consists of blocks

$$\left(M_h\right)_{kj} = \mathbf{i} \int_{\tau_k} \varphi_j(x) ds_x.$$

4.2 Adaptive Cross Approximation

As it was mentioned in the Subsect. 4.1, all matrices involved in (20), beside the mass matrix, are dense, i.e. all their entries do not vanish in general, leading to an asymptotically quadratic memory requirement for the whole procedure. Thus, classical boundary element realizations are applicable only for a rather moderate number N of boundary elements. Fortunately, all boundary element matrices can be decomposed into a hierarchical system of blocks which can be approximated by the use of low rank matrices. This approximation can be computed by means of the ACA algorithm. This is nowadays a well established numerical tool in both, scientific and commercial software. The first publications of the ACA are [7] in the mathematical literature and [29] in an engineering journal.

As it was shown in [1] and as we will demonstrate in Sect. 7, the ACA algorithm is also efficient in the approximation of the matrix coming from the interpolation with radial basis functions.

5 Exact Solution for the Polynomial Body Forces

In order to obtain the particular solution for polynomial right hand sides involved in (4), (5), we first remark that if the right hand side of the system of Lamé equations $\underline{f}$ is conservative, i.e.

$$\underline{f}(x) = -\mathrm{grad}\,\theta(x)$$

then a particular solution can be found in the form

$$\underline{u}_p(x) = -\mathrm{grad}\,\psi(x)\,,$$

where the scalar function ψ solves the Poisson equation

$$-\Delta\,\psi(x) = \frac{1}{\lambda+2\mu}\theta(x)\,. \tag{21}$$

A solution to the Poisson equation with a polynomial right hand side can be found in [28]. In particular, when θ is a homogeneous polynomial of degree m, then the solution of the equation (21) is given by

$$\psi(x) = -\frac{1}{\lambda+2\mu}|x|^2\sum_{k=0}^{[m/2]}(-1)^k\frac{|x|^{2k}\Delta^k\theta(x)}{(2,2)_{k+1}(2m+3-2k,2)_{k+1}}\,,$$

where $[\cdot]$ is the integer part of a number, and

$$(a,b)_0 = 1\,,\quad (a,b)_k = a(a+b)\ldots(a+(k-1)b)\quad \text{for}\, k\geq 1$$

denotes the generalized Pochhammer symbol. Thus for low order polynomials, we get

$$\theta(x) = 1\,,\quad \psi(x) = -\frac{1}{\lambda+2\mu}\,\frac{1}{6}\,|x|^2\,,$$

$$\theta(x) = x_i\,,\quad \psi(x) = -\frac{1}{\lambda+2\mu}\,\frac{1}{10}x_i\,|x|^2\,,\ i=1,2,3\,,$$

$$\theta(x) = x_ix_j\,,\quad \psi(x) = -\frac{1}{\lambda+2\mu}\,\frac{1}{14}x_ix_j\,|x|^2\,,\ i,j=1,2,3\,,\ i\neq j\,,$$

$$\theta(x) = x_i^2\,, \quad \psi(x) = -\frac{1}{\lambda+2\mu}\Big(\frac{1}{6}x_i^2 - \frac{1}{140}\,|x|^2\Big)|x|^2\,,\ i=1,2,3\,,$$
$$\theta(x) = x_1x_2x_3\,, \quad \psi(x) = -\frac{1}{\lambda+2\mu}\,\frac{1}{18}x_1x_2x_3\,|x|^2\,,$$
$$\theta(x) = x_ix_j^2, \quad \psi(x) = -\frac{1}{\lambda+2\mu}\Big(\frac{1}{18}x_ix_j^2 - \frac{1}{252}x_i\,|x|^2\Big)|x|^2\,,\ i,j=1,2,3\,,\ i\neq j\,,$$
$$\theta(x) = x_i^3\,, \quad \psi(x) = -\frac{1}{\lambda+2\mu}\Big(\frac{1}{18}x_i^3 - \frac{1}{84}x_i\,|x|^2\Big)|x|^2\,,\ i=1,2,3\,.$$

An exact solution of the Poisson equation (21) for a non-homogeneous polynomial θ is then a linear combination of the above functions.

Gravity

The potential of the gravitational force is given by (4) and, therefore,

$$\underline{u}_p(x) = -\mathrm{grad}\,\psi(x)\,, \quad \psi(x) = -\frac{\rho\,g}{\lambda+2\mu}\,\frac{1}{10}x_3\,|x|^2\,. \tag{22}$$

Inertial Forces

The potential of the centrifugal force is given by (5). Thus, we get

$$\underline{u}_p(x) = -\mathrm{grad}\,\psi(x)\,,$$

where

$$\psi(x) = \frac{\rho\,(2\pi\omega)^2}{2(\lambda+2\mu)}\Big(\frac{1}{6}\,|\hat{x}|^2 + \frac{1}{5}\,(x,\hat{x}) + \frac{61}{420}\,|x|^2\Big)|x|^2\,.$$

6 Approximation of the Body Force

For more general forces, like thermal stress, the analytical form of the particular solution is in general not available. In this section, we describe an approach to approximate a conservative body force density $\underline{f}$ in terms of radial basis functions. The vector valued function $\underline{f}$ can be written in the form

$$\underline{f}(x) = -\mathrm{grad}\,\theta(x) \quad \text{for}\, x\in\Omega\,, \tag{23}$$

where the scalar potential $\theta : \Omega \to \mathbb{R}$ is assumed to be given. For example, in a form of a numerical solution to another boundary value problem, as in the case of thermoelasticity.

6.1 Radial Basis Functions

Radial basis functions are one of the most popular tools for interpolating or approximating scattered data in various applications. A single radial basis function $\phi : \mathbb{R}^3 \to \mathbb{R}$ is a real-valued function whose value depends only on the distance from the origin, i.e.

$$\phi(x) = \phi(|x|) \quad \text{for } x \in \mathbb{R}^3 .$$

Given a set of function values

$$\{\theta_1, \dots, \theta_Q\}$$

at the distinct points

$$\{x_1, \dots, x_Q\} \subset \mathbb{R}^3 ,$$

the approximation problem of the function θ is solved by the interpolant

$$s(x) = \sum_{j=1}^{Q} a_j \phi(|x - x_j|) + p(x) \quad \text{for } x \in \mathbb{R}^3 , \tag{24}$$

where p is an additional low degree ($\deg p \le k \ll N$) polynomial. Let

$$\{p_1, \dots, p_K\}$$

be a basis in the space of three-dimensional polynomials of degree $\deg p \le k$. Then

$$p(x) = \sum_{j=1}^{K} b_j p_j(x) .$$

Thus, Q interpolation conditions

$$s(x_i) = \sum_{j=1}^{Q} a_j \phi(|x_i - x_j|) + \sum_{j=1}^{K} b_j p_j(x_i) = \theta_i \quad \text{for } i = 1, \dots, Q$$

and K orthogonality or side conditions

$$\sum_{j=1}^{Q} a_j p_i(x_j) = 0 \quad \text{for } i = 1, \dots, K$$

lead to a system of $Q + K$ linear equations

$$\begin{pmatrix} A & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix} , \tag{25}$$

with

$$A \in \mathbb{R}^{Q \times Q} , \quad a_{ij} = \phi(|x_i - x_j|) , \quad P \in \mathbb{R}^{Q \times K} , \quad p_{ij} = p_j(x_i) ,$$

and

$$a \in \mathbb{R}^Q\,, \quad b \in \mathbb{R}^K\,, \quad c \in \mathbb{R}^Q\,, \quad c_i = \theta_i\,.$$

Popular choices for the function ϕ include

1. Gaussian
$$\phi(\rho) = \exp(-\beta\rho^2)\,,$$
2. Multiquadric
$$\phi(\rho) = \sqrt{1+\beta\rho^2}\,,$$
3. Inverse quadric
$$\phi(\rho) = \frac{1}{1+\beta\rho^2}\,,$$
4. Polyharmonic splines
$$\phi(\rho) = \rho^k\,, \quad \text{for}\, k = 1,3,\ldots\,,$$
   and
$$\phi(\rho) = \rho^k \ln(\rho)\,, \quad \text{for}\, k = 2,4,\ldots\,,$$

where $\beta > 0$ is a parameter.

One of the important properties of the radial basis functions is that the system of linear equations (25) is guaranteed to be solvable under very mild conditions on the locations of the data points [43]. In particular, radial basis functions do not require that the points lie on any sort of regular grid. The set of points, where the function values are given, has two important characteristics

$$h_{min} = \min_{i=1,\ldots,Q}\min_{j\neq i}|x_i - x_j|\,, \quad h_{max} = \max_{i=1,\ldots,Q}\min_{j\neq i}|x_i - x_j|\,.$$

The accuracy of the interpolation depends on the above separation distance h_{max}

$$||\theta - s||_{L^2(\Omega)} \leq C h_{max}^{5/2} ||\theta||_{H^2(\mathbb{R}^3)}\,,$$

while the estimate on the condition number of the resulting linear system is related to h_{min} as

$$cond(A) \leq C h_{min}^{-2}$$

(see [26] and [43] Table 12.1). Furthermore, the matrices A and P are dense and will require $\mathscr{O}(Q^2)$ words of computer memory. A direct solver will require $\mathscr{O}(Q^3)$ operations. Thus, the direct use of the radial basis functions is suitable for rather small problems with up to a few thousands of points. However, modern data reduction techniques like adaptive "FastRBF" [4, 6] or Adaptive Cross Approximation [1], allow the use of the radial basis functions also for big and even for huge data sets. Similar to the approach from [1], we employ the H-Matrix/ACA approximation technique to reduce the numerical complexity and storage requirements, and construct a sparse preconditioner based on approximate cardinal functions to facilitate convergence of the iterative GMRES solution procedure (see also [5]).

6.2 Exact Solutions

Once the potential θ in (23) is approximated with the radial basis functions (24), the body force density $\underline{f}$ is also approximated by

$$\underline{f}_Q(x) = -\operatorname{grad} s(x) = \sum_{j=1}^{Q} a_j\big(-\operatorname{grad}\phi(|x-x_j|)\big) - \operatorname{grad} p(x) \quad \text{for } x \in \Omega\,.$$

Thus, an approximation for a particular solution can be found in the form

$$\underline{u}_{p,Q}(x) = \sum_{j=1}^{Q} a_j\big(-\operatorname{grad}\psi(|x-x_j|)\big) - \operatorname{grad} q(x)\,,$$

where the functions ψ and q satisfy the Poisson equation (21) with the right hand sides defined by the functions ϕ and p, respectively. The exact solution for the low order polynomials is presented in the Section 5, thus, we discuss the exact solution of the Poisson equation (21) with the right hand side defined by the function ϕ. Since the function ϕ is isotropic, we get in spherical coordinates an ordinary differential equation

$$\psi''(\rho) + \frac{2}{\rho}\psi'(\rho) = -\frac{1}{\lambda+2\mu}\phi(\rho)\,. \tag{26}$$

To avoid singularities at zero, we complete the equation (26) with the initial conditions

$$\psi(0) = \psi'(0) = 0$$

and obtain the following solutions for different radial basis functions

1. for the Gaussian

$$\psi(\rho) = -\frac{1}{\lambda+2\mu}\frac{2\beta^{1/2}\rho - \pi^{1/2}\operatorname{erf}(\beta^{1/2}\rho)}{4\beta^{3/2}\rho}\,,$$

2. for the multiquadric

$$\psi(\rho) = -\frac{1}{\lambda+2\mu}\frac{-8\beta^{1/2}\rho + \beta^{1/2}\rho(5+2\beta\rho^2)\sqrt{1+\beta\rho^2} + 3\operatorname{arcsinh}(\beta^{1/2}\rho)}{24\beta^{3/2}\rho}\,,$$

3. for the inverse quadric

$$\psi(\rho) = -\frac{1}{\lambda+2\mu}\frac{\beta^{1/2}\rho(-2+\ln(1+\beta\rho^2)) + \arctan(\beta^{1/2}\rho)}{2\beta^{3/2}\rho}\,,$$

4. for the polyharmonic splines

$$\psi(\rho) = -\frac{1}{\lambda+2\mu}\frac{1}{(k+2)(k+3)}\rho^{k+2}\,, \quad \text{for}\, k = 1,3,\dots\,,$$

and, finally,

$$\psi(\rho) = -\frac{1}{\lambda+2\mu}\frac{1}{(k+2)(k+3)}\rho^{k+2}\Big(-\frac{2k+5}{(k+2)(k+3)}+\ln(\rho)\Big)\,,$$

for $k = 2,4,\dots$.

Thus, all popular radial basis functions can be used for an approximation of the conservative right hand side of the system of Lamé equations.

7 Numerical Examples

In what follows, we choose the radial basis function to be the polyharmonic spline of order two (also known as the thin plate spline), namely

$$\phi(\rho) = \rho^2 \ln \rho\,.$$

In this case the function ψ is given by

$$\psi(\rho) = \frac{\rho^4\,(20\log\rho - 9)}{400\,(\lambda+2\mu)} \tag{27}$$

the polynomial p in (24) is linear

$$p(x) = (b,x) + b_4\,, \quad b \in \mathbb{R}^3\,, \quad b_4 \in \mathbb{R}\,,$$

and, therefore, $K = 4$. Assuming that Q values of the scalar potential θ of the body force are given at points

$$\{x_j\}_{j=1}^{Q} \subset \overline{\Omega}\,,$$

the method of finding an approximate solution to the BVP (1),(3) can be summarized in the following four steps:

1. Obtain the coefficients $\{a_j\}_{j=1}^{Q}$ and $\{b_i\}_{i=1}^{4}$ from the system (25).
2. Evaluate the particular solution

$$\underline{u}_{p,Q}(x) = \tag{28}$$
$$\frac{1}{\lambda+2\mu}\left[\sum_{j=1}^{Q} a_j \frac{|x-x_j|^2\,(5\log|x-x_j|-1)}{25}\,(x-x_j) + x\left(\frac{b_4}{3}+\frac{(b,x)}{5}\right)+\frac{|x|^2}{10}b\right]$$

on Γ_D and its conormal derivative analytically on Γ_N.

3. Solve the homogeneous BVP (16), (17) numerically by means of the fast Galerkin BEM to get the approximation $\underline{\tilde{u}}_0$ to the function $\underline{u}_0$.
4. The approximate solution to the BVP (1), (3) is $\underline{\tilde{u}}(x) = \underline{\tilde{u}}_0(x) + \underline{u}_{p,Q}(x)$.

Implementation

As mentioned in Subsect. 4.2 and 6.1 the H-Matrix technique in combination with the ACA procedure was utilized to efficiently implement steps 1, and 2 of the method, rendering the complexity of this part to be almost linear in terms of the number of interpolation points Q (for details see [1]). The remaining BEM part of the method is also accelerated by means of the H-Matrix/ACA technique (as in [8]). It has almost linear complexity in terms of the number of boundary elements, which in our experiments is of order $\mathscr{O}(Q^{2/3})$. However, the entries of Galerkin BEM matrices are much more expensive to compute (compared to the entries of the RBF matrix A) as they are expressed as double integrals over the elements. In our numerical tests, the RBF interpolation part of the method took less CPU time than the BEM part for small problems. For problems with about $5 \cdot 10^4$ interpolation nodes and $1.3 \cdot 10^4$ boundary elements these CPU times were similar. As the number of interpolation points gets higher, the ill-conditioning of the interpolation problem demands special treatment. The construction of the sparse preconditioner (based on local cardinal functions) becomes the most time consuming part of the procedure (see Table 2).

Example 1

To verify the accuracy of the proposed solution procedure, we consider a mixed BVP of thermoelasticity with the temperature distribution

$$T(x) = \frac{1}{1 + 5|x - \hat{x}|^2},$$

where the point $\hat{x} = (0.5, 0.5, 0.5)^\top$. We know, that the function

$$\underline{u}_{ex}(x) = \frac{\alpha\,(3\lambda + 2\mu)}{\lambda + 2\mu} \frac{\sqrt{5}|x - \hat{x}| - \arctan\left(\sqrt{5}|x - \hat{x}|\right)}{5^{3/2}|x - \hat{x}|^3}(x - \hat{x})$$

satisfies the equation of thermoelasticity. Prescribing $\underline{u}_{ex}$ and its conormal derivative as boundary conditions, we make it the solution of the BVP. We consider the half-sphere shaped body

$$\Omega = \{x \in \mathbb{R}^3 : x_3 > 0, |x| < 1/2\}$$

with material parameters $E = 1000$, $\nu = 0.24$, and $\alpha = (3\lambda + 2\mu)^{-1}$ (see (7)) and set the boundary conditions

$$(\gamma_0 \underline{u})(x) = \underline{u}_{ex}(x) \quad \text{on } \Gamma_D = \{x \in \Gamma : x_3 = 0, |x| < 1/2\},$$

and

$$(\gamma_1 \underline{u})(x) = (\gamma_1 \underline{u}_{ex})(x) \quad \text{on } \Gamma_N = \{x \in \Gamma \, : \, x_3 > 0\}$$

To study the convergence of the proposed method, we construct a sequence of quasi-uniform tetrahedral meshes (see Table 1).

Table 1 Parameters of the mesh sequence for Example 1.

nr	volume nodes	volume elem.	surface nodes.	surface elem.	discr. $h \times 10^{-2}$	interp. prec. $\varepsilon_2 \times 10^{-4}$	RBF compr.	BEM compr.
1	165	598	107	210	10.59	37.682	0.548	0.420
2	1 032	4 784	422	840	5.296	8.5192	0.507	0.389
3	7 267	38 272	1 682	3 360	2.648	1.4798	0.416	0.245
4	54 485	306 176	6 722	13 440	1.324	0.2218	0.137	0.098
5	454 273	2 630 656	31 362	62 720	0.612	0.0278	0.029	0.028

Table 2 Computational times for the RBF interpolation (Example 1).

nr	interp. nodes	matrix gener. [s]	precond. gener. [s]	iter.	solution [s]
1	165	< 1	< 1	43	< 1
2	1 032	< 1	2	90	< 1
3	7 267	7	14	445	28
4	54 485	108	2801	470	606
5	454 273	1735	74219	1421	25789

The function T is sampled in all nodes and its interpolant in the form (24) is obtained. To illustrate the quality of the interpolant, we evaluate it at Gauss points of each volume element and approximate the L^2 error of the interpolation

$$\varepsilon_2 = \frac{\|T - s\|_{L^2(\Omega)}}{\|T\|_{L^2(\Omega)}} \, .$$

This error is reported in Table 1 (column 7) and plotted in Fig. 1 (right). We observe the convergence rate, which is in agreement with the estimated order $5/2$ found in [26, 38, 42].

The effect of the H-Matrix/ACA acceleration technique on the proposed procedure is shown in Table 1, where the compression rates for the interpolation matrix A (column 8) and the Galerkin discretization of the single layer potential operator V_h (column 9) are reported. Since the condition number of the system (25) can be high, the ACA-accuracy was set to 10^{-8} when approximating the matrix A. The respective parameter for the approximant of the matrix V_h is 10^{-6}. The hierarchical clustering procedure delivers relatively large admissible blocks when applied to a surface (BEM) geometry even for rather low number of elements, which is not the case when clustering a point cloud in Ω. Due to these two reasons, in comparison to the compression of the matrix A the compression of the BEM matrices is more pronounced. It should also be pointed out that the RBF interpolation problem on the set of 454273 data points in $\mathbb{R}^3$ is challenging even on modern hardware. The storage requirement for the fully populated matrix A in this case is about 1.2TB. The blockwise low rank approximant constructed by means of the H-Matrix/ACA technique fits into 33.5GB of computer memory. The CPU time needed for the RBF interpolation procedure is shown in Table 2. We observe the increase of the number of GMRES iterations needed to solve the interpolation problem as well as the cost of preconditioner as the number of points increases. This is due to the widely known problem of high condition number of the interpolation system.

The boundary conditions (17) are composed by evaluating (28) and its conormal derivative at the surface nodes and the corresponding triangles respectively. These geometric entities constitute a surface mesh, which is utilized to obtain the solution $\underline{u}_0$ to the problem (16),(17) by means of the fast Galerkin BEM. The following accuracy measures are plotted in Fig. 1

$$\varepsilon_D = \frac{\|\underline{u}_{ex} - \tilde{\underline{u}}\|_{L^2(\Gamma)}}{\|\underline{u}_{ex}\|_{L^2(\Gamma)}}, \quad \varepsilon_N = \frac{\|\gamma_1 \underline{u}_{ex} - \gamma_1 \tilde{\underline{u}}\|_{L^2(\Gamma)}}{\|\gamma_1 \underline{u}_{ex}\|_{L^2(\Gamma)}}.$$

We observe second order accuracy for the Dirichlet data and the first order - for the Neumann data. Since these are the convergence rates of the Galerkin BEM for the homogeneous mixed type BVP in linear elastostatics [37], the additional error from the RBF interpolation must behave at least as good. To explore this feature further, we perform the RBF interpolation on a coarse mesh but evaluate the particular solution on the finest mesh to obtain the approximate solution to the BVP. The resulting L^2 errors ε_D and ε_N are shown in Table 3. We observe that the reconstruction accuracies for the temperature field and its derivatives are high enough even when about $7 \cdot 10^3$ interpolation nodes are used. This flexibility of choice is possible due to the meshless nature of the interpolation.

Example 2

In the case when the scalar potential θ of the volume force density $\underline{f}$ is a polynomial, it is not necessary to perform the RBF interpolation, since a particular solution is given by (22). Consider a body Ω depicted in Fig. 2 fixed at the bottom and subjected to a gravitational body force $\underline{f}(x) = -\underline{e}_3$. A particular solution in this case is given by

$$\underline{u}_p(x) = \frac{1}{10(\lambda + 2\mu)} \left(2x_3 x + |x|^2 \underline{e}_3\right).$$

Table 3 Changes in solution accuracy caused by choosing less expensive RBF interpolant (Example 1).

nr	interp. nodes	$\varepsilon_D\, 10^{-5}$	$\varepsilon_N\, 10^{-3}$
2	1032	34.165	9.08176
3	7267	6.2860	8.95545
4	54485	4.2192	8.95236
5	454273	4.1597	8.95228

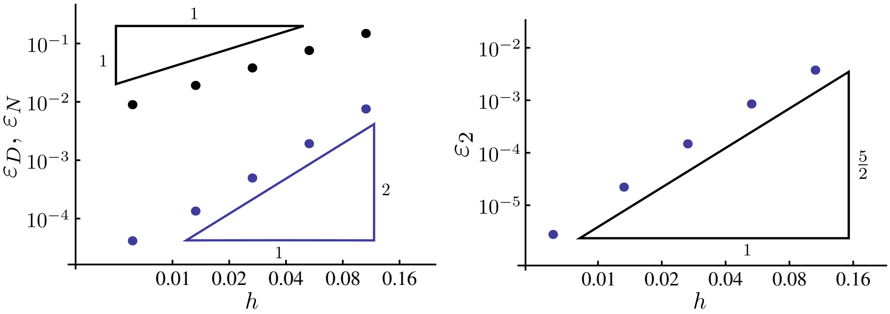


Fig. 1 Example 1. Convergence of the computed Dirichlet and Neumann data to the exact solution (left), and the interpolant of the temperature to the given function (right).

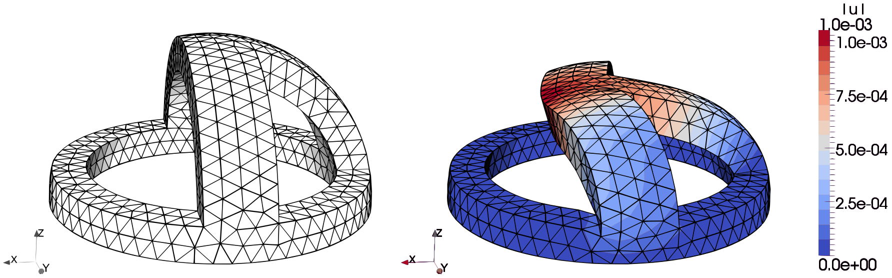


Fig. 2 Example 2. The domain Ω (left) and the resulting deformed configuration (right). The nodes are shifted along the vector field $200\underline{\tilde{u}}$ and the magnitude $|\underline{\tilde{u}}|$ is plotted on the surface.

The approximate solution $\underline{\tilde{u}}$ to the BVP (1), (3) on the surface mesh with 823 nodes and 1654 triangles is shown in Fig. 2.

8 Conclusions

In this paper, we present a possibility to solve the inhomogeneous system of Lamé's equations by the use of the fast Galerkin BEM solver based on a particular solution.

For conservative body forces having a polynomial potential like gravity or inertial forces, the particular solution is obtained in a closed analytical form.

For more general cases of conservative body forces, the RBF interpolation of the scalar potential is used for finding a particular solution in an approximate form.

This particular solution can easily be combined with a fast Galerkin BEM solver to obtain the solution to the original problem without volume discretization of the domain and leading to a meshless method.

The procedure is illustrated on numerical examples. The first example shows the expected convergence of the approximated particular solution to the exact one. We observe the same accuracy of the method as in the case of the Galerkin BEM for the homogeneous problem. The second example illustrates the analytical particular solution in the case of gravity.

Further developments of this approach include applications to viscoelastic and nonlinear problems in solid mechanics.

References

[1] Bambach, M., Grzhibovskis, R., Hirt, G., Rjasanow, S.: Adaptive cross approximation for surface reconstruction based on radial basis functions. J. Eng. Math. 62, 149–160 (2008)

[2] Ahmed, S., Banerjee, P.: Free vibration analysis of bem using particular integrals. J. Eng. Mech. 112, 682–695 (1986)

[3] Barber, J.: Body forces. In: Elasticity, Solid Mechanics and Its Applications, vol. 172, pp. 91–108. Springer, Netherlands (2010)

[4] Beatson, R., Newsam, G.: Fast evaluation of radial basis functions. I. Comput. Math. Appl. 24(12), 7–19 (1992)

[5] Beatson, R., Cherrie, J., Mouat, C.: Fast fitting of radial basis functions: methods based on preconditioned GMRES iteration. Adv. Comput. Math. 11(2-3), 253–270 (1999)

[6] Beatson, R., Newsam, G.: Fast evaluation of radial basis functions: moment-based methods. SIAM J. Sci. Comput. 19(5), 1428–1449 (1998)

[7] Bebendorf, M.: Approximation of boundary element matrices. Numer. Math. 86(4), 565–589 (2000)

[8] Bebendorf, M., Grzhibovskis, R.: Accelerating Galerkin BEM for Linear Elasticity using Adaptive Cross Approximation. Math. Meth. Appl. Sci. 29, 1721–1747 (2006)

[9] Bebendorf, M., Rjasanow, S.: Adaptive Low-Rank Approximation of Collocation Matrices. Computing 70, 1–24 (2003)

[10] Brebbia, C., Nowak, A.: Treatment of domain integrals by using the dual and multiple reciprocity methods. In: Discretization methods in structural mechanics (Vienna, 1989), pp. 13–28. Springer, Berlin (1990)

[11] Brinkman, H.: A calculation of the viscous force exerted by a flowing fluid on a dense swarm of particles. Appl. Sci. Res. A1, 27–34 (1947)

[12] Chen, C., Brebbia, C., Power, H.: Dual reciprocity method using compactly supported radial basis functions. Comm. Numer. Meth. Engrg. 15(2), 137–150 (1999)

[13] Ciarlet, P.: Mathematical elasticity, vol. 1: Three-dimensional elasticity. North-Holland (1988)

[14] Cruse, T.: Boundary integral equation method for three dimension. Tech. rep., AFOSR-TR-75 0813 (1975)

[15] Danson, D.: A boundary element formulation of problems in linear isotropic elasticity with body forces. In: Brebbia, C.A. (ed.) Boundary Element Methods. Springer, Berlin (1981)

[16] Gipson, G.: Boundary Element Fundamentals — Basic Concepts and Recent Developments in the Poisson Equation. Computational Mechanics Publication, Southampton (1987)

[17] Golberg, M.: Recent developments in the numerical evaluation of particular solutions in the boundary element method. Appl. Math. Comput. 75(1), 91–101 (1996)

[18] Golberg, M., Chen, C.: Discrete projection methods for integral equations. Computational Mechanics Publications, Southampton (1997)

[19] Golberg, M.A., Chen, C.S., Bowman, H., Power, H.: Some comments on the use of radial basis functions in the dual reciprocity method. Comput. Mech. 22(1), 61–69 (1998)

[20] Hand, L., Finch, J.: Analytical Mechanics. Cambridge University Press (1998)

[21] Henry, D., Banerjee, P.: A new boundary element formulation for two- and three-dimensional thermoelasticity using particular integrals. Int. J. Numer. Meth. Engrg. 26(9), 2061–2077 (1988)

[22] Hsiao, S., Mammoli, A., Ingber, M.: The evaluation of domain integrals in complex multiply-connected three-dimensional geometries for boundary element methods. Comput. Mech. 32, 226–233 (2003)

[23] Ingber, M., Mammoli, A., Brown, M.: A comparison of domain integral evaluation techniques for boundary element methods. Int. J. Numer. Meth. Engrg. 52(4), 417–432 (2001)

[24] Ivanov, E., Andrä, H., Kudryavtsev, A.: Domain decomposition for automatic parallel generation of tetrahedral meshes. CMAM 6, 178–193 (2006)

[25] Jumarhon, B., Amini, S.: Towards a convergence analysis for the dual reciprocity method. In: Boundary Elements, XXI, Oxford. Int. Ser. Adv. Bound. Elem., vol. 6, pp. 583–592. WIT Press, Southampton (1999)

[26] Light, W., Wayne, H.: On power functions and error estimates for radial basis function interpolation. J. Approx. Theory 92(2), 245–266 (1998)

[27] Moulinec, H., Suquet, P.: A numerical method for computing the overall response of nonlinear composites with complex microstructure. Comput. Meth. Appl. Mech. Engrg. 157(12), 69–94 (1998)

[28] Karachik, V., Antropova, N.: On the solution of the inhomogeneous polyharmonic equation and the inhomogeneous Helmholtz equation. Diff. Eqns. 46(3), 387–399 (2010)

[29] Kurz, S., Rain, O., Rjasanow, S.: The Adaptive Cross Approximation technique for the 3D boundary element method. IEEE Trans. Magn. 38(2), 421–424 (2002)

[30] Nardini, D., Brebbia, C.: A new approach to free vibration analysis using boundary elements. In: Brebbia, C.A. (ed.) Boundary Element Methods in Engineering. Springer, Berlin (1982)

[31] Neves, A., Brebbia, C.: The multiple reciprocity boundary element method in elasticity: A new approach for transforming domain integrals to the boundary. Int. J. Numer. Meth. Engrg. 31(4), 709–727 (1991)

[32] Nowak, A., Brebbia, C.: The multiple reciprocity method: A new approach for transforming BEM domain integrals to the boundary. Engng. Anal. 6(3), 164–167 (1989)

[33] Of, G., Steinbach, O., Urthaler, P.: Fast evaluation of volume potentials in boundary element methods. SIAM J. Sci. Comput. 32(2), 585–602 (2010)

[34] Partridge, P., Brebbia, C.: Computer implementation of the BEM dual reciprocity method for the solution of general field equations. Comm. Appl. Numer. Meth. 6(2), 83–92 (1990)

[35] Partridge, P., Brebbia, C.: The dual reciprocity method. In: Advanced formulations in boundary element methods. Internat. Ser. Comput. Engrg., Comput. Mech., Southampton, pp. 31–75 (1993)

[36] Partridge, P., Brebbia, C., Wrobel, L.: The Dual Reciprocity Boundary Element Method. Elsevier Apllied Science, London (1992)

[37] Rjasanow, S., Steinbach, O.: The Fast Solution of Boundary Integral Equations. Springer Series in Mathematical and Analytical Technology with Applications to Engineering, vol. 12. Springer, New York (2007)

[38] Schaback, R.: Improved error bounds for scattered data interpolation by radial basis functions. Math. Comput. 68(225), 201–216 (1999)

[39] Sirtori, S., Maier, G., Novati, G., Miccoli, S.: A Galerkin symmetric boundary-element method in elasticity - Formulation and implementation. Int. J. Numer. Meth. Engrg. 35(2), 255–282 (1992)

[40] Steinbach, O.: Numerical approximation methods for elliptic boundary value problems. Springer, New York (2008)

[41] Tang, W.: Transforming domain into boundary integrals in BEM: a generalized approach. Lecture Notes in Engineering. Springer (1988)

[42] Wendland, H.: Optimal approximation orders in L^p for radial basis functions. East. J. Approx. 1, 87–102 (2000)

[43] Wendland, H.: Scattered Data Approximation. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press (2005)

Towards the Direct and Inverse Adaptive Mixed Finite Element Formulations for Nearly Incompressible Elasticity at Large Strains

Anke Bucher, Uwe–Jens Görke, and Reiner Kreißig

Abstract. This contribution presents advanced numerical models for the solution of the direct and inverse problems of nearly incompressible hyperelastic processes at large strains. The discussed mixed finite element approach contributes to the numerical simulation of coupled multiphysics problems, including the calibration of appropriate material models (parameter identification). The presented constitutive approach is based on the multiplicative decomposition of the deformation gradient resulting in a two-field formulation with displacement components and hydrostatic pressure as primary variables. The ill-posed inverse problem of parameter identification analyzing inhomogeneous displacement fields is solved using deterministic trust-region optimization techniques. Within this context, a semi-analytical approach for sensitivity analysis represents an efficient and accurate method to determine the gradient of the objective function. The mixed boundary value problem is based on the spatial discretization of the weak formulations of the linear momentum balance and the incompressibility condition. Its linearization serves as basis for the solution of the direct problem, while the implicit differentiation of the weak formulations with respect to material parameters provides the necessary relations for the semi-analytical sensitivity analysis. Adaptive mesh refinement and mesh coarsening are realized controlled by a residual a posteriori error estimator. Efficiency and accuracy of the presented direct and inverse numerical techniques are demonstrated on a typical example.

Anke Bucher
Fakultät Maschinen– und Energietechnik, HTWK Leipzig,
Koburger Str. 62, 04416 Markkleeberg, Germany
e-mail: bucher@me.htwk-leipzig.de

Uwe–Jens Görke
Helmholtz Zentrum für Umweltforschung, Permoserstraße 15, 04318 Leipzig, Germany
e-mail: uwe-jens.goerke@ufz.de

Reiner Kreißig
Fakultät für Maschinenbau, TU Chemnitz, Straße der Nationen 62,
09111 Chemnitz, Germany
e-mail: reiner.kreissig@mb.tu-chemnitz.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 299–323.
springerlink.com

1 Introduction

Real engineering and biological processes are frequently characterized by the simultaneous action of various physical and chemical fields (e. g., mechanical, thermal, electromagnetic). Modeling and simulation of corresponding processes concerns the solution of so-called *multiphysics* problems. Recently, an appropriate numerical treatment of coupled multifield problems has been enabled by improving the efficiency and robustness of numerical methods and by increasing the performance of computational facilities. The solution of these problems plays a central role in the field of high-precision simulation of real processes in a wide variety of applications.

Historically, the modeling of nearly incompressible elastic material behavior is one of the best-investigated multifield problems, whereby the fulfillment of the constraint of volume preservation enforces the definition of an additional primary variable besides the components of the displacement vector, which are usually available for the modeling of deformation processes. Currently, mixed approaches resulting in u/p-c formulations are common standard for commercial finite element codes, in order to simulate nearly incompressible elastic material behavior at small and large strains. The corresponding conceptual basics have been developed about 40 years ago by Hermann [21]. Soon after, Taylor et al. [43] proposed a mixed formulation for the solution of problems of isotropic as well as anisotropic elasticity at small strains. Based on the multiplicative split of the deformation gradient into a volumetric and an isochoric part, Brink and Stein [9], and Simo and Taylor [41] presented mixed model extensions for the case of large strain isotropic elasticity. Le Tallec [30], and Rüter and Stein [40] discussed anisotropic finite element models for nearly incompressible elasticity at large strains. Detailed overviews about theory and numerics of mixed formulations as well as corresponding further references are given, e. g., in the monographs of Brezzi and Fortin [8], and Hughes [23].

The realistic numerical simulation of the mechanical behavior of components, engineering structures, biological tissues and geological formations requires the development and implementation of appropriate constitutive models, independent of the considered length scale. Constitutive models incorporate material parameters and/or material functions, which usually are not measurable but considerably affect numerical results. The process of the determination of these material parameters by adaptation of calculated results to measured data is called *parameter identification* or, more frequently in recent literature, *calibration* of material models. Usually, optimization procedures are used for parameter identification.

In the case of analyses of samples exposed to inhomogeneous stress-strain fields in order to calibrate material models, the value of the objective function appropriately defined for the optimization procedure has to be calculated solving full initial-boundary value problems, e. g., using finite element simulations. Due to their enormous amount of time, gradientless (i. e., stochastic, evolutional) approaches are unsuitable for this purpose, whereas deterministic (gradient-based) optimization procedures have been proved to be successful. For the mathematical foundation of deterministic optimization approaches we refer to the monographs of Dennis and Schnabel [16], and Nocedal and Wright [38]. Their application to problems of parameter identification has been recorded in a variety of relevant papers published in the last two decades. Within this context, the determination of derivatives required for the gradient of the objective function is in the focus of interest.

At the professorship of Solid Mechanics of the Chemnitz University of Technology, the research on topic of the numerical determination of material parameters analyzing inhomogeneous displacement fields is mainly based on early publications of Mahnken and Stein [32] as well as Gelin and Ghouati [18] concerning small inelastic strains. In these studies, the authors present the semi-analytical sensitivity analysis in detail. This method is based on an implicit differentiation of the weak formulation of the equilibrium conditions as well as the fully iterated constitutive relations with respect to material parameters, and results in a global system of linear algebraic equations incorporating the global matrix known from the solution of the direct problem, but being characterized by a modified right-hand side. Further results comprising large strains, specific constitutive models and various formulations of the optimization problem are presented by Mahnken et al. [31, 33], Johansson and Runesson [25], and others. Ogden et al. [39] discussed the application of usually performed analyses of homogeneous physical fields for model calibration to the example of nearly incompressible elastic models. The application of gradient-based optimization techniques including semi-analytical sensitivity analysis to mixed problems of saturated biphasic porous media has been discussed in detail by Mahnken and Steinmann [34], and by Lecampion and Constantinescu [29]. In both cases, geometrically linear deformations have been assumed.

In recent years, the authors' research on topic of model calibration has been directed to the development of a consistent solution of automated parameter identification for elasto-plastic constitutive models analyzing inhomogeneous displacement fields and global information (e. g., traction-displacement curves). The root of the method resides in a gradient-based optimization procedure (the Levenberg-Marquardt approach has been proved to be particularly successful – cf. [6, 27]) including semi-analytical sensitivity analysis. Appropriate applications to large strain problems have been central since 1996 [19, 28]. The high level of generalization of the developed algorithms enforces their quite simple extension to other material classes.

This paper presents the theoretical foundation and the numerical algorithms for the solution of the direct and inverse problems of nearly incompressible elasticity at large strains, and is structured as follows: The kinematics describing the considered problems is presented in Sect. 2. Furthermore, in this section the thermodynamically consistent definition of specific constitutive relations is discussed. The conceptual basis for optimization techniques applied to model calibration as inverse problem is explained in Sect. 3. Afterwards, Sect. 4 is dedicated to the representation of fundamental variational formulations in Lagrangian description, including their consistent linearization, and to the representation of the semi-analytical sensitivity analysis. A short overview about fundamental aspects of the used adaptive mesh control is given in Sect. 5. The solution of the direct and inverse problems is discussed in Sect. 6 for a specific numerical example.

In the following, tensors as well as their matrix representation (the particular meaning becomes apparent by the specific context) will be denoted by bold-faced characters in direct notation. Their juxtaposition implies the scalar product of two vectors (e. g., $\mathbf{a}\,\mathbf{b} = a_i b^i$), or a single contraction of adjacent indices of two tensors, while double dots indicate the summation product of two second-order tensors (e. g., $\mathbf{a} \cdot\cdot\, \mathbf{b} = a_i{}^j\, b_j{}^i$), or a double contraction of adjacent indices of tensors of rank two and higher. The symbol $\otimes$ denotes the tensor product (e. g., $(\mathbf{a} \otimes \mathbf{b})_{ij} = a_i b_j$). A superposed dot indicates the material rate of a tensor, a superscript $(\cdot)^\top$ the transposed tensor.

2 Nearly Incompressible Elasticity at Large Strains

The property of volume preservation during deformation as geometrical expression of the incompressibility of a material is in particular typical for non-foamed elastomers, but matters also as kinematical constraint for other material classes (e. g., elasto-plasticity). As the usual displacement-related finite element approaches provide solutions for nearly incompressible problems with significantly higher stiffness compared to the real material behavior (*locking* effects), the analysis of mixed variational formulations as fundament for improved finite element models is required.

2.1 Kinematics

The following representations are given in material (i. e., Lagrangian) description, defining all variables as functions of the coordinates of the reference configuration. Within this reference configuration (at time $t = t_0$), the physical body under consideration represents a set $\Omega_0 \subset \mathbb{R}^3$ of material points with a boundary Γ_0 (i. e., a domain in the three-dimensional Euclidean space $\mathbb{E}^3$). The boundary is divided into subdomains Γ_{0D} with Dirichlet type boundary conditions, and Γ_{0N} bearing Neumann type boundary conditions. Within this context, the conditions $\Gamma_0 = \Gamma_{0D} \cup \Gamma_{0N}$ and $\Gamma_{0D} \cap \Gamma_{0N} = \emptyset$ are fulfilled. Material points are uniquely characterized defining their position vectors $\mathbf{X} \in \Omega_0$ and their coordinates (X_1, X_2, X_3), respectively.

At current time t, the physical body under consideration occupies a domain $\Omega_t \subset \mathbb{R}^3$ – the current configuration. Here, the material points are characterized by their position vectors $\mathbf{x}$ and their coordinates (x_1, x_2, x_3), respectively. Defining the law of motion

$$\mathbf{x} = \varphi(\mathbf{X}, t) \tag{1}$$

bijective correlations at each time t are established for material points between their current position in $\mathbb{E}^3$ and their allocation in the reference state.

The deformation gradient $\mathbf{F}$ represents the basis of kinematical considerations within the context of the development of constitutive models at large strains. This fundamental variable provides the mapping of material line elements between the reference and the current configurations. Using the law of motion (1), the deformation gradient can be represented as follows:

$$\mathbf{F} = (\operatorname{Grad}\mathbf{x})^\top = (\operatorname{Grad}\mathbf{U})^\top + \mathbf{I}\,. \tag{2}$$

$\mathbf{U} = \mathbf{U}(\mathbf{X},t)$ is the displacement vector represented as function of material coordinates and time.

Based on the deformation gradient, various strain measures related to the reference configuration or to the current configuration can be defined. As here constitutive relations are formulated in Lagrangian description (i. e., related to the reference configuration), the right Cauchy-Green tensor

$$\mathbf{C} = \mathbf{F}^\top\mathbf{F} = \operatorname{Grad}\mathbf{U} + (\operatorname{Grad}\mathbf{U})^\top + \operatorname{Grad}\mathbf{U}\,(\operatorname{Grad}\mathbf{U})^\top + \mathbf{I} \tag{3}$$

and Green’s strain tensor

$$2\mathbf{E} = \mathbf{C} - \mathbf{I} = \operatorname{Grad}\mathbf{U} + (\operatorname{Grad}\mathbf{U})^\top + \operatorname{Grad}\mathbf{U}\,(\operatorname{Grad}\mathbf{U})^\top \tag{4}$$

serve as relevant material strain measures.

The assumption of nearly incompressible material behavior implies that the volume of a material volume element remains constant during external mechanical loading. The ratio of differential volume elements before and after deformation can be expressed using the determinant of the deformation gradient

$$\frac{dV}{dV_0} = \det\mathbf{F} = J\,. \tag{5}$$

Consequently, the property of ideal incompressibility can be mathematically formulated using the relation $J \equiv 1$.

Following the usual approach (cf., e. g., [40]), a multiplicative split of the deformation gradient into two parts will be defined below, thus enabling the numerical analysis of volume preserving deformation processes:

$$\mathbf{F} = \mathbf{F}_\mathrm{v}\,\mathbf{F}_\mathrm{d}\,. \tag{6}$$

Here, $\mathbf{F}_\mathrm{v}$ characterizes that part of deformation, which results in a pure volume change of a material volume element, whereas the partial deformation gradient $\mathbf{F}_\mathrm{d}$ denotes pure shape changes. This multiplicative split of the deformation gradient has been first proposed by Flory [17], and is therefore usually known from literature as *Flory split*. The volumetric part of the deformation gradient is defined as follows:

$$\mathbf{F}_\mathrm{v} \overset{\mathrm{def}}{=} J^{\frac{1}{3}}\,\mathbf{I} \qquad (J \approx 1)\,. \tag{7}$$

Consequently, for the isochoric part of the deformation gradient follows:

$$\mathbf{F}_\mathrm{d} = J^{-\frac{1}{3}}\,\mathbf{F} \qquad \text{with} \qquad \det\mathbf{F}_\mathrm{d} \equiv 1\,. \tag{8}$$

Based on the above mentioned definitions, an isochoric part of the right Cauchy-Green tensor can be formulated

$$\mathbf{C}_\mathrm{d} = \mathbf{F}_\mathrm{d}^\top \mathbf{F}_\mathrm{d} = J^{-\frac{2}{3}}\,\mathbf{C}\,. \tag{9}$$

2.2 Thermodynamically Consistent Constitutive Equation

The Clausius-Duhem inequality serves as origin for the formulation of constitutive relations considering nearly incompressible material behavior. For the isothermal elastic case this thermodynamic relation is defined as

$$-\rho_0\,\dot{\bar{\psi}}(\mathbf{C}) + \frac{1}{2}\mathbf{T}\cdot\cdot\,\dot{\mathbf{C}} \geq 0 \tag{10}$$

with respect to the reference configuration. Here, ρ_0 is the mass density, $\bar{\psi}$ the free Helmholtz energy density per unit mass, and $\mathbf{T}$ represents the 2nd Piola-Kirchhoff stress tensor. As usual, the dot above a variable indicates its material time derivative.

Elastic deformations represent fully reversible processes. Consequently, for inequality (10) the equal sign has to be applied, and the following hyperelastic constitutive law can be formulated:

$$\mathbf{T} = 2\rho_0\,\frac{\partial\bar{\psi}}{\partial\mathbf{C}} = 2\,\frac{\partial\psi}{\partial\mathbf{C}}\,. \tag{11}$$

Like in classical continuum mechanics, the free Helmholtz energy density is considered as an isotropic tensor function of an appropriate strain measure. In the case of nearly incompressible material behavior an additive split of the free Helmholtz energy density into a deviatoric (i. e., isochoric) part ψ_d depending on the strain tensor $\mathbf{C}_\mathrm{d}$, and a volumetric part ψ_v depending on J is assumed (cf. [40]):

$$\psi = \psi_\mathrm{d}(\mathbf{C}_\mathrm{d}) + \widetilde{\psi}_\mathrm{v}(J)\,. \tag{12}$$

For further considerations it is convenient to define the volumetric part of the free Helmholtz energy density as $\frac{1}{2}\kappa\,\psi_\mathrm{v}^2(J)$, where κ is the bulk modulus. Additionally, ψ_v is characterized by the property $\psi_\mathrm{v}(J=1)=0$. Therewith, for (12) arises the following detailed representation:

$$\psi = \psi_\mathrm{d}(\mathbf{C}_\mathrm{d}) + \frac{1}{2}\kappa\,\psi_\mathrm{v}^2(J) = \psi_\mathrm{d}\left(I(\mathbf{C}_\mathrm{d}), I\!I(\mathbf{C}_\mathrm{d})\right) + \frac{1}{2}\kappa\,\psi_\mathrm{v}^2(J) \tag{13}$$

with the usual definition of main invariants $I(\mathbf{C}_\mathrm{d})$, $I\!I(\mathbf{C}_\mathrm{d})$. Based on the definition

$$p \overset{\mathrm{def}}{=} \kappa\,\psi_\mathrm{v} \tag{14}$$

of the hydrostatic pressure p in terms of a constitutive assumption, and considering (13), from (11) follows the hyperelastic relation

$$\mathbf{T} = 2\frac{\partial \psi}{\partial \mathbf{C}} = 2\frac{\partial \psi_{\mathrm{d}}}{\partial \mathbf{C}_{\mathrm{d}}} \cdot\cdot \frac{\partial \mathbf{C}_{\mathrm{d}}}{\partial \mathbf{C}} + 2\kappa\,\psi_{\mathrm{v}} \frac{\partial \psi_{\mathrm{v}}}{\partial J}\frac{\partial J}{\partial \mathbf{C}} = \mathbf{T}_{\mathrm{d}} + \mathbf{T}_{\mathrm{v}} \tag{15}$$

for the 2nd Piola-Kirchhoff stress tensor in case of nearly incompressible material behavior.

For the modeling of nearly incompressible elastic material behavior the specific representation of the volumetric part of the free Helmholtz energy density

$$\psi_{\mathrm{v}}(J) = \ln J \tag{16}$$

is proposed, which has frequently been discussed in literature resulting in the simple relation $\mathbf{T}_{\mathrm{v}} = p\,\mathbf{C}^{-1}$ within the context of the partial stress $\mathbf{T}_{\mathrm{v}}$. With respect to the deviatoric part of the free Helmholtz energy density the following specific formulations have been studied by the authors

$$\psi_{\mathrm{d}} = c_{10}\left(I(\mathbf{C}_{\mathrm{d}}) - 3\right), \tag{17a}$$

$$\psi_{\mathrm{d}} = c_{10}\left(I(\mathbf{C}_{\mathrm{d}}) - 3\right) + c_{01}\left(II(\mathbf{C}_{\mathrm{d}}) - 3\right), \tag{17b}$$

$$\psi_{\mathrm{d}} = \frac{c_{10}}{\alpha}\left[e^{\alpha(I(\mathbf{C}_{\mathrm{d}})-3)} - 1\right] \tag{17c}$$

known as Neo-Hookean, Mooney-Rivlin and modified Fung models (in the order of presentation). The quantities c_{10}, c_{01} and α represent material parameters.

3 Parameter Identification as Optimization Problem

The identification of material parameters constitutes an inverse problem comprising the analysis of the *effects* of the parameters to be determined onto measurable field variables. As the operator mapping material parameters to mechanical variables (e. g., components of the displacement vector) is usually of implicit, non-linear character with unknown mathematical structure, its explicit closed-form inversion approves to be impossible in this case. Consequently, the problem of model calibration in general results in the solution of an optimization problem: The parameters have to be estimated in such a way that an appropriately defined objective function approaches its minimum.

3.1 Objective Function

A model function will be defined, which characterizes an arbitrary physical variable y depending on a vector of variables $\mathbf{x}$ as well as on a set (vector) of material parameters $\mathbf{c}$

$$y = y(\mathbf{x}, \mathbf{c})\,. \tag{18}$$

In order to analyze a mechanical problem, the corresponding model function can be, for instance, constituted by the displacement field depending on stresses, internal variables and material parameters.

The calibration of constitutive models is aimed at the determination of parameters realizing a sufficiently accurate approximation of measured discrete data $\hat{y}_i$ representing defined conditions for variables $\mathbf{x}_i$. The corresponding parameter set, which is in this narrower sense an optimal one, and thus the best approximation of measured data, can be considered as determined if a least squares norm approaches its minimum

$$\frac{1}{2}\sum_{i=1}^{n}\left[\hat{y}_i - y(\mathbf{x}_i,\mathbf{c})\right]^2 \quad \longrightarrow \quad \min . \tag{19}$$

Based on the definition of a vector of residuals $\mathbf{r}$ between measured and calculated values

$$\mathbf{r}(\mathbf{c}) \;=\; \left\{r_i(\mathbf{c})\right\} \qquad \text{with} \qquad r_i(\mathbf{c}) \;=\; \hat{y}_i - y(\mathbf{x}_i,\mathbf{c}) \tag{20}$$

the objective function Φ can be formulated as the following least squares norm:

$$\Phi(\mathbf{c}) \;=\; \frac{1}{2}\,\mathbf{r}^{\mathrm{T}}(\mathbf{c})\,\mathbf{r}(\mathbf{c}) \;=\; \frac{1}{2}\sum_{i=1}^{n} r_i^2(\mathbf{c}) \quad \longrightarrow \quad \min . \tag{21}$$

The necessary optimality criterion

$$\nabla\Phi(\mathbf{c}^*) \;=\; \mathbf{0} \tag{22}$$

results for models representing linear functions of the parameters $\mathbf{c}$ in a system of linear algebraic equations with the coordinates of the optimal parameter vector $\mathbf{c}^*$ as primary variables. This system is known as *normal equation*

$$\widetilde{\mathbf{J}}^{\top}\widetilde{\mathbf{J}}\,\mathbf{c}^* \;=\; \widetilde{\mathbf{J}}^{\top}\mathbf{r}\,. \tag{23}$$

Here, $\widetilde{\mathbf{J}}$ represents the Jacobian matrix comprising the first derivatives of the residuals with respect to the material parameters.

Based on previous research for elasto-plasticity (cf. [19, 28]) the following specific objective function is defined for the calibration of constitutive models in case of nearly incompressible elasticity at large strains:

$$\Phi(\mathbf{c}) = \sum_{i=1}^{n_{LU}}\sum_{j=1}^{n_T}\sum_{K=1}^{3}\Big(\{U_K(\mathbf{c})\}_{ij} - \{\bar{U}_K\}_{ij}\Big)^2 . \tag{24}$$

Here, U_K, $\bar{U}_K$ denote the calculated/measured displacements, n_{LU} is the number of load steps for the correspondingly evaluated variable, and n_T represents the number of local measuring points for the displacement field. More generalized formulations in terms of (24) considering global variables such as tractions and/or momentums are discussed in [20]. The determination of $y(\mathbf{x},\mathbf{c})$ requires in this case the solution of a complete initial-boundary value problem at each optimization step, for instance using the finite element method.

3.2 Optimization Procedure

For models that are non-linear in the parameters **c**, an iterative approach must be chosen for their determination in the context of deterministic strategies

$$\mathbf{c}_{k+1} = \mathbf{c}_k + \mathbf{s}_k\,. \tag{25}$$

The most frequently cited method for the non-linear approximation based on a least squares norm is the Gauß-Newton method. Within this context, the search step $\mathbf{s}_k$ will be determined from the solution of the linear system of equations

$$\widetilde{\mathbf{J}}_k^\top \widetilde{\mathbf{J}}_k\, \mathbf{s}_k \;=\; -\widetilde{\mathbf{J}}_k^\top\, \mathbf{r}_k\,. \tag{26}$$

This system can be interpreted as the necessary optimality condition for a linear approximation of the objective function $\Phi(\mathbf{c}_k)$. In this sense, the Gauß-Newton procedure represents a special case of the Newton approximation neglecting second order terms of the Hessian matrix (cf. [16, 38] and others). The main advantage of this approximation compared to the classical Newton approach is that only first derivatives in terms of the Jacobian matrix $\widetilde{\mathbf{J}}$ are required.

The Gauß-Newton method converges very fast in a certain neighborhood of the solution $\mathbf{c}^*$. Problems may occur regarding the convergence behavior outside of this surrounding of the solution, and if $\widetilde{\mathbf{J}}$ is rank deficient, or nearly so. These problems will be avoided using so-called *trust-region* type algorithms like the Levenberg-Marquardt procedure, where the search step length will be limited by a properly chosen confidence interval Δ_k

$$\|\mathbf{s}_k\| \;\leq\; \Delta_k\,. \tag{27}$$

A recent and excellent overview of deterministic optimization methods is, for instance, given by Nocedal and Wright [38].

4 Weak Formulation of the Coupled Boundary Value Problem of Nearly Incompressible Elasticity at Large Strains

The boundary value problem of nearly incompressible elasticity is defined in terms of a mixed two-field problem by analogy with approaches discussed in [9, 40] and by other authors. Variational formulations of the linear momentum balance and of a constraint to satisfy the incompressibility serve as starting point for the definition of field equations in Ω_0.

4.1 Linearization of the Weak Formulation – Direct Problem

For physical reasons, the linear momentum balance is originally defined in the current configuration. As explained in detail in [11], this balance relation can be represented after several transformations in the local formulation

$$\mathrm{Div}\left(\mathbf{T}\mathbf{F}^{\top}\right) + \rho_0\left(\mathbf{K} - \mathbf{A}\right) = \mathbf{0} \tag{28}$$

exclusively using variables, that are functions of the coordinates $\mathbf{X}$ of the reference configuration. Here $\mathbf{K}(\mathbf{X})$ are volume forces per unit mass (densities), and $\mathbf{A}(\mathbf{X})$ is the acceleration.

Furthermore, an additional constraint to include the incompressibility condition is considered. Within this context it is assumed, that the hydrostatic pressure at each material point represents a function of the volume change. The local incompressibility condition follows from (14) after some simple transformations

$$\psi_{\mathrm{v}}(J) - \frac{1}{\kappa}p = 0\,. \tag{29}$$

The weak formulations of the linear momentum balance and the incompressibility condition arise after multiplying (28) by an arbitrary vector-valued test function $\mathbf{V} = \mathbf{V}(\mathbf{X})$ (with $\mathbf{V} \in (H^1(\Omega_0))^3$, $\mathbf{V} = \mathbf{0}$ at the boundary Γ_{0D}) and multiplying (29) by the arbitrary scalar-valued test function $q \in L_2(\Omega_0)$ (there are no explicit boundary conditions for q), and integration of both equations over the volume of the undeformed area Ω_0 (i. e., the reference configuration). Considering quasi-static problems, neglecting volume forces, and performing certain transformations that are discussed in detail in [11], for the required weak formulations follows:

$$\int\limits_{\Omega_0} \mathbf{T} \cdot\cdot\, \mathbf{E}\left(\mathbf{U};\mathbf{V}\right) d\Omega_0 = \int\limits_{\Gamma_{0N}} \bar{\mathbf{R}}\mathbf{V}\, d\Gamma_0, \tag{30a}$$

$$\int\limits_{\Omega_0} \left[\psi_{\mathrm{v}}(J) - \frac{1}{\kappa}p\right] q\, d\Omega_0 = 0\,. \tag{30b}$$

Here, $\bar{\mathbf{R}}$ represents the vector of the given external loading at the deformation part of the Neumann boundary. Additionally, a second order tensor $\mathbf{E}\left(\mathbf{U};\mathbf{V}\right)$ was defined to simplify the presentation

$$2\mathbf{E}\left(\mathbf{U};\mathbf{V}\right) \overset{\mathrm{def}}{=} (\mathrm{Grad}\mathbf{V})^{\top} + \mathrm{Grad}\mathbf{V} + \mathrm{Grad}\mathbf{U}(\mathrm{Grad}\mathbf{V})^{\top} + \mathrm{Grad}\mathbf{V}(\mathrm{Grad}\mathbf{U})^{\top}\,. \tag{31}$$

Finally, based on (30a), (30b), and considering the stress decomposition (15), the mixed boundary value problem of nearly incompressible elasticity at large strains comprises the solution of the following nonlinear system of equations

$$a_0(\mathbf{U};\mathbf{V}) + b_0(\mathbf{U};p,\mathbf{V}) = f(\mathbf{V}), \tag{32a}$$

$$b_1(\mathbf{U};q) \;-\; d_0(p,q) \quad = 0 \tag{32b}$$

with respect to the field variables displacement $\mathbf{U}$ and hydrostatic pressure p. For individual parts of this system follows in detail:

$$a_0(\mathbf{U};\mathbf{V}) = \int\limits_{\Omega_0} \mathbf{T}_{\mathrm{d}}(\mathbf{U}) \cdot\cdot\, \mathbf{E}\,(\mathbf{U};\mathbf{V})\, d\Omega_0, \tag{33a}$$

$$b_0(\mathbf{U};p,\mathbf{V}) = \int\limits_{\Omega_0} pJ\, \frac{\partial \psi_{\mathrm{v}}(J(\mathbf{U}))}{\partial J}\, \mathbf{C}^{-1} \cdot\cdot\, \mathbf{E}\,(\mathbf{U};\mathbf{V})\, d\Omega_0, \tag{33b}$$

$$f(\mathbf{V}) = \int\limits_{\Gamma_{0N}} \bar{\mathbf{R}}\mathbf{V}\, d\Gamma_0, \tag{33c}$$

$$b_1(\mathbf{U};q) = \int\limits_{\Omega_0} \psi_{\mathrm{v}}(J(\mathbf{U}))\, q\, d\Omega_0, \tag{33d}$$

$$d_0(p,q) = \int\limits_{\Omega_0} \frac{1}{\kappa} p\, q\, d\Omega_0\,. \tag{33e}$$

To solve the system of equations (32a) and (32b) using Newton's method, its linearization is required. In this context, Taylor series representations with respect to the field variables $\mathbf{U}$ and p, truncated after the first (i. e., linear) term are considered. Following, the solution $(\mathbf{U}_{t+\Delta t}, p_{t+\Delta t}) = (\mathbf{U}+\Delta\mathbf{U}, p+\Delta p)$ of the two-field problem at time $t+\Delta t$ is required, whereas its solution $(\mathbf{U}_t, p_t) = (\mathbf{U}, p)$ at time t is assumed to be known. After suitable transformations (for details, see [11]), the mixed boundary value problem of nearly incompressible elasticity at large strains can be specified in linearized form (individual iteration at current load step) as a linear system of equations

$$a(\mathbf{U};\Delta\mathbf{U},\mathbf{V}) + b_0(\mathbf{U};\Delta p,\mathbf{V}) = f(\mathbf{V}) - a_0(\mathbf{U};\mathbf{V}) - b_0(\mathbf{U};p,\mathbf{V}), \tag{34a}$$

$$b_0(\mathbf{U};q,\Delta\mathbf{U}) \;-\; d_0(\Delta p,q) \quad = d_0(p,q) - b_1(\mathbf{U};q)\,. \tag{34b}$$

Additionally to the terms defined in (33a)-(33e), for the functional $a(\mathbf{U};\Delta\mathbf{U},\mathbf{V})$ follows:

$$\begin{aligned} a(\mathbf{U};\Delta\mathbf{U},\mathbf{V}) = &\int\limits_{\Omega_0} \mathbf{E}\,(\mathbf{U};\mathbf{V}) \cdot\cdot\, \frac{\partial \mathbf{T}(\mathbf{E}(\mathbf{U}),p)}{\partial \mathbf{E}} \cdot\cdot\, \mathbf{E}\,(\mathbf{U};\Delta\mathbf{U})\, d\Omega_0 \\ &+ \int\limits_{\Omega_0} \mathbf{T}\,(\mathbf{E}\,(\mathbf{U})\,,p) \cdot\cdot\, \mathrm{Grad}\,\Delta\mathbf{U}(\mathrm{Grad}\mathbf{V})^{\mathrm{T}}\, d\Omega_0 \end{aligned} \tag{35}$$

with the consistent material matrix $\partial \mathbf{T}/\partial \mathbf{E}$. After spatial discretization, the system (34a), (34b) can be solved for appropriately selected finite element subspaces for $\Delta \mathbf{U}$, $\mathbf{V}$ and Δp, q within the framework of a mixed finite element approach.

4.2 Parameter Derivatives of the Weak Formulation – Inverse Problem

As starting point for the formulation of the numerical relations of the semi-analytical sensitivity analysis, which are required in order to determine the derivatives of primary variables of the direct problem with respect to the material parameters serves the system (30a), (30b) of weak formulations for equilibrium and incompressibility condition. The implicit differentiation of these relations with respect to a single material parameter c_j leads to the following numerical system:

$$\int\limits_{\Omega_0} \left[\frac{d\mathbf{T}}{dc_j} \cdot\cdot\, \mathbf{E}\left(\mathbf{U};\mathbf{V}\right) + \mathbf{T} \cdot\cdot\, \frac{d\mathbf{E}\left(\mathbf{U};\mathbf{V}\right)}{dc_j} \right] d\Omega_0 = \mathbf{0}, \tag{36a}$$

$$\int\limits_{\Omega_0} \left[\frac{d\psi_{\mathrm{V}}(J)}{dc_j} - \frac{1}{\kappa} \frac{dp}{dc_j} \right] q \, d\Omega_0 = 0. \tag{36b}$$

Within this context, it has been assumed that the boundary conditions (i. e., external loads) and the test functions $\mathbf{V}$, q, respectively are independent of the material parameters. Additionally, it should be mentioned that the parameter κ in the case of small strains can be approximated with the bulk modulus. With regard to the presented theory, κ does not represent a material parameter in the proper sense, but is rather a penalty parameter for the approximate fulfillment of the property of volume preservation.

For further considerations, explicit and implicit dependencies of the field variables on material parameters are considered

$$\mathbf{U} = \mathbf{U}(\mathbf{c}), \tag{37a}$$

$$p = p(\mathbf{c}), \tag{37b}$$

$$\mathbf{T} = \mathbf{T}\left(\mathbf{E}(\mathbf{c}), p(\mathbf{c}), \mathbf{c}\right). \tag{37c}$$

Following the definition of corresponding individual derivatives and some additional transformations, a linear system of equations can be formulated in order to calculate the total derivatives of displacement components and hydrostatic pressure with respect to a single material parameter c_j (for more details see [20])

$$a\left(\mathbf{U};\frac{d\mathbf{U}}{dc_j},\mathbf{V}\right)+b_0\left(\mathbf{U};\frac{dp}{dc_j},\mathbf{V}\right)=f_c(\mathbf{U};\mathbf{V}), \tag{38a}$$

$$b_0\left(\mathbf{U};q,\frac{d\mathbf{U}}{dc_j}\right)\ -\ d_0\left(\frac{dp}{dc_j},q\right)\quad =0. \tag{38b}$$

The right-hand side of relation (38a) contains the partial derivatives of the 2nd Piola-Kirchhoff stress tensor with respect to the material parameter c_j

$$f_c(\mathbf{U};\mathbf{V})=-\int\limits_{\Omega_0}\frac{\partial\mathbf{T}}{\partial c_j}\cdot\cdot\,\mathbf{E}(\mathbf{U};\mathbf{V})\,d\Omega_0\,. \tag{39}$$

Consequently, from (11) follows:

$$\frac{\partial\mathbf{T}}{\partial c_j}=2J^{-\frac{2}{3}}\frac{\partial}{\partial c_j}\left(\frac{\partial\psi_{\mathrm{d}}}{\partial\mathbf{C}_{\mathrm{d}}}\cdot\cdot\left[\delta_I^K\delta_L^J\,\mathbf{G}^I\otimes\mathbf{G}_J\otimes\mathbf{G}_K\otimes\mathbf{G}^L-\frac{1}{3}\mathbf{C}_{\mathrm{d}}\otimes\mathbf{C}_{\mathrm{d}}^{-1}\right]\right) \tag{40}$$

with the basis vectors $\mathbf{G}_I$, $\mathbf{G}^I$ of the natural and dual bases, respectively. Here, it has been assumed that the relation for the volumetric part of the free Helmholtz energy density does not contain any material parameters (cf. (16)).

Comparing the linear systems of equations (34a), (34b) and (38a), (38b), it becomes clear that after spatial discretization using suitable finite element spaces the sub-matrices of the global system matrix in both cases have exactly the same numerical structure. Within the context of the semi-analytical sensitivity analysis, the linear system of equations (38a), (38b) must be solved for all material parameters to be identified, and at each load step contributing to the calculation of the objective function. As for this system the variables $\mathbf{U}$ and p are considered at time $t+\Delta t$, the latest global system matrix known from the iterative solution of the direct problem for the current load step (Newton's approximation) can be applied directly for the purpose of sensitivity analysis. Considering the semi-analytical approach for sensitivity analysis, n_c+1 calculations of the complete system per load step with different right-hand sides have to be performed compared to the higher effort for numerical sensitivity analysis requiring $2n_c$ calculations.

5 Adaptive Finite Element Realization

5.1 Stable Element Formulations

The existence and uniqueness of the solution of mixed finite element formulations is associated with the definition of the so-called Inf-Sup condition (also LBB condition) as a stability criterion for finite element approaches in different subspaces (here, $\mathbb{V}_h^3$ and $\mathbb{X}_h$). As described in detail by Bathe [5], the LBB condition has been formulated originally for small strain situations. Studies on the stability of finite element solutions at large isotropic elastic strains can be found, for instance, in [3, 14].

The above presented approach for the finite element solution of coupled boundary value problems of nearly incompressible elasticity at large strains has been realized in the scientific in-house code SPC-PM2AdNlmix. Existing hierarchical data structures, their advanced use by solvers for the global system of equations, and user-friendly generalized material interfaces are applied in context of adaptive strategies for grid modification. The spatial discretization for the mixed problem is realized using so-called Taylor-Hood elements showing a particularly good compatibility with existing hierarchical data structures. First, Taylor and Hood [42] proposed a bi-quadratic/bi-linear rectangular element $\mathscr{Q}_2 - \mathscr{Q}_1$ for the stable numerical solution of the Navier-Stokes equations, and demonstrated the fulfillment of the LBB condition for this case. They used conventional polynomial form functions, with a polynomial degree, which is one order lower referred to the dual variable (hydrostatic pressure) compared to that of the primary variable (here, displacement). This element is characterized by ensuring the continuity of the solution across boundaries of adjacent elements for both the primary and the dual variables (u/p-c formulation, see also [5]). Later on, Taylor and Hood proposed a rectangular element $\mathscr{Q}_2^{(8)} - \mathscr{Q}_1$ differing from the original formulation by omitting the node in the center of the element (see, e. g., [22]). Thus, the form functions for the primary variable are not fully bi-quadratic but of serendipity type. Currently, the combination of finite element spaces proposed by Taylor and Hood rates among the most frequently applied element classes referred to the solution of mixed problems of the considered structure. For that exist also triangular elements for plane problems (e. g., the $\mathscr{P}_2 - \mathscr{P}_1$ element) and corresponding three-dimensional versions, respectively.

Based on an idea of Bramble and Pasciak [7], an iterative method is used in SPC-PM2AdNlmix for the efficient solution of the coupled linear finite element equations. A generalization of Bramble-Pasciak-CG has been proposed in the paper [37] solving mixed finite element schemes for elasticity problems. It became apparent that this solver, which is based on hierarchical structures, interacts very efficiently with the above mentioned elements of Taylor-Hood type.

5.2 Error Controlled Mesh Adaptation

The choice of suitable error estimators and/or error indicators constitutes an important aspect of adaptive numerical methods in order to evaluate the impact of the spatial discretization on simulation results. Considering mixed finite element formulations, local a posteriori error estimators are known from the literature for the Stokes problem (see, e. g., [1, 4, 26, 44]). With regard to the simulation of nearly incompressible elastic materials, Rüter and Stein [40] proposed approaches for local a posteriori Neumann problems in order to get information about upper limits of residual errors without multiplicative constants.

Meyer [36] presented an error estimator with element-oriented parts η_T evaluating the global solution of the mixed problem based on the fulfillment of the local equilibrium conditions. The author demonstrates the analogy to the usual residual a posteriori error estimator for the displacement problem of solid mechanics, and shows its applicability for the case of nearly incompressible elastic material behavior at large strains. Neglecting volume forces, regarding the estimation of the discretization error within an element T with edges E and initial volume Ω_{0T} follows the square value of the a posteriori error estimator η_T

$$\eta_T^2 = \frac{h_T}{\lambda_D}\left(h_T \int\limits_T |\operatorname{Div}\left(\mathbf{T}\mathbf{F}^\top\right)|^2 \, d\Omega_{0T} + \sum_{E\in\partial\Omega_{0T}} \int\limits_E |[\mathfrak{N}_E \mathbf{T}\mathbf{F}^\top]|^2 \, dS_{0E}\right). \quad (41)$$

Here, h_T denotes the characteristic element length, $\mathfrak{N}_E$ is the outward-pointing normal vector of the element edge E, and λ_D represents an interpolation constant, depending on the material. The brackets $[\cdot]$ characterize the jump of a function across an element edge E belonging to adjacent elements. The algorithms for the calculation of edge jumps are described in detail in [12]. In linear elasticity, λ_D is usually approximated by the Young's modulus. Based on corresponding studies addressed to small strains, in the case of the presented constitutive model, it is assumed that the order of magnitude of λ_D is the same as for linear problems, and thus it is approximated by the Young's modulus relevant for the range of small strains.

The maximum element error is detected as comparative value, in order to come to a decision on a modification of the mesh. An element is marked for refinement, if its error estimator exceeds this comparative error by a certain amount. Based on appropriate criteria, also mesh coarsening is allowed. Realizing the mixed boundary value problem of nearly incompressible elasticity at large strains, the algorithms developed for finite elasto-plasticity could be used for error estimation and mapping of field variables on new mesh structures in an effective manner. Additionally, the hierarchical strategies for mesh refinement and coarsening have been adopted, and adapted to stable element classes of multi-field problems. Details of the used adaptive procedure are discussed, e. g., in [12, 13] (see also literature cited therein).

6 Numerical Examples

In the following, the applicability of the presented numerical model for the solution of the direct and inverse problems for nearly incompressible elastic material behavior at large strains will be demonstrated on the example of a typical benchmark. Related to the analyzed finite element grids, mainly stable Taylor-Hood rectangular elements $\mathscr{Q}_2^{(8)} - \mathscr{Q}_1$ have been used (i. e., elements with quadratic approach of serendipity type for displacement degrees of freedom, and bi-linear approach for the hydrostatic pressure). In order to study the solution behavior in principle, occasionally Taylor-Hood triangular elements $\mathscr{P}_2 - \mathscr{P}_1$ have been adopted. Further details on various numerical aspects as well as extensive descriptions of the numerical studies on the example presented below (e. g., comprising adaptive methods) are discussed in [11, 20].

6.1 Adaptive Numerical Solution of Cook's Membrane Problem

This example has been proposed by Cook [15] to perform finite element analyses of the solution behavior of a generalized rectangular element. Due to distinct element distortions to be observed, the discussed benchmark has frequently been studied by various authors since then, in order to analyze numerical properties of mixed formulations under dominant bending load (cf. [2, 24, 35, 40]).

Cook's membrane is fixed at the left boundary, whereas edge-related shear traction is applied at the right boundary. Plane strain conditions are assumed. Fig. 1 (left) shows the geometry and boundary conditions of the discussed structure. Additionally, the deformed geometry at maximum load is presented (right), which has been observed using the scientific in-house code SPC-PM2AdNlMix.

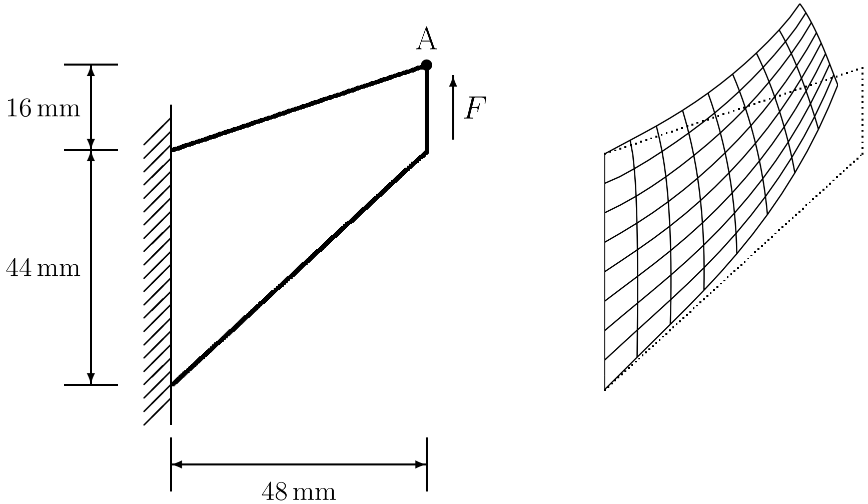


Fig. 1 Cook's membrane problem. Left: Geometry and boundary conditions according to [2] (F = 100 kN). Right: Deformed mesh of rectangular elements.

For validation of the presented constitutive models including their numerical realization, extensive simulations for Cook's membrane problem have been performed under varying conditions. Within this context, the displacement of the upper right corner node (cf. Fig. 1 – vertex A) is of special interest. This value strongly depends on element type and grid length (i. e., discretization level), and has been discussed by many authors before for comparison of various models and software codes.

In the following, selected results will be presented, which have been achieved using the presented constitutive model. The Neo-Hookean approach (17a) has been chosen in order to specify the free Helmholtz energy density. Specific information regarding boundary conditions and material parameters have been taken from Armero [2], indicating values of 240.4 GPa for Young's modulus and of 0.499 for Poisson's ratio. Consequently, material parameters of the nearly incompressible Neo-Hookean model have been calculated as $c_{10} = 40.097\,\text{GPa}$, $\kappa = 40{,}070\,\text{GPa}$.

First comparisons for Cook's membrane problem show a good agreement between results (e. g., for hydrostatic pressure; cf. Fig. 2) achieved with the presented constitutive model and data from Armero as well as from simulations using the commercial finite element code MSC-Marc.

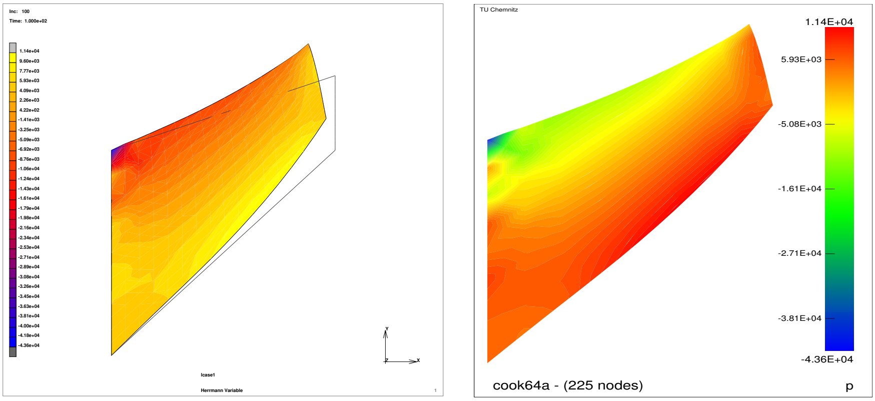


Fig. 2 Cook's membrane problem. Isosurfaces of the hydrostatic pressures distribution at maximum loading. Comparison of simulations performed with the commercial FE code MSC-Marc (left, $p_{max} = 11.437\,\text{GPa}$) with results for the in-house FE code SPC-PM2AdNlMix (right, $p_{max} = 11.434\,\text{GPa}$).

The vertical displacement of the upper right corner node at maximum load has been found as 6.854 mm using SPC-PM2AdNlMix compared to 6.855 mm using MSC-Marc. From a respective figure in [2], the corresponding value obtained by Armero can be determined as approximately 6.8 mm. For comparability of the mentioned studies, a uniform finite element grid discretized with 64 rectangular elements has been used.

It is well-known that the application of purely displacement-related elements in context with the numerical simulation of nearly incompressible material behavior can be extremely error-prone compared to models using mixed finite element spaces. Particularly, in the case of plane strain state assumed for the presented example,

Table 1 Cook's membrane problem. Vertical displacement of the upper right corner node (cf. Fig. 1 – vertex A). Comparison of results for various Taylor-Hood elements with results for displacement type rectangular elements $\mathscr{Q}_2^{(8)}$ (quadratic approach of serendipity type) and $\mathscr{Q}_1$ (bi-linear approach), and displacement type triangular elements $\mathscr{P}_2$ (quadratic approach), $\mathscr{P}_1$ (linear approach) depending on the Poisson's ratio ν.

	Vertical displacement					
	$\mathscr{Q}_2^{(8)} - \mathscr{Q}_1$	Related to element type $\mathscr{Q}_2^{(8)} - \mathscr{Q}_1$				
Poisson's ratio ν	Absolute values	$\mathscr{Q}_2^{(8)}$	$\mathscr{Q}_1$	$\mathscr{P}_2 - \mathscr{P}_1$	$\mathscr{P}_2$	$\mathscr{P}_1$
	mm	%	%	%	%	%
0.35	7.0756	100.01	89.84	100.10	100.69	94.97
0.40	7.2776	97.31	85.70	100.26	98.14	92.28
0.45	7.1391	97.85	81.56	100.46	99.06	92.17
0.49	6.9142	97.74	62.48	100.70	99.85	87.69
0.499	6.8505	96.57	38.48	100.80	99.79	76.77
0.4999	6.8440	95.02	32.44	100.81	99.41	72.86
0.49995	6.8436	94.45	32.01	100.81	99.16	72.58

displacement-related elements prove themselves to be to stiff. Bathe [5] explained this *locking* behavior in detail. Currently, a substantial amount of literature exists analyzing this phenomenon also known as *locking effect*, and discussing methods for its prevention. Table 1 summarizes selected results of element locking analyses in the vicinity of incompressibility using the scientific finite element code SPC-PM2AdNlMix substantiating qualitatively the above mentioned solution behavior. Additionally, quantitatively comparable results using different types of rectangular element are presented like discussed in [35].

Numerical simulations have been performed with different initial finite element grids, in order to investigate the action of adaptive code capabilities. It is exemplarily recognizable from Fig. 3 that even in the case of application of Taylor-Hood elements the material behavior is modeled as being to stiff if a coarse spatial discretization has been generated. With increasing (uniform) grid refinement associated with an increased number of degrees of freedom, the vertical displacement of the upper right corner node of Cook's membrane approaches an asymptotic value. This behavior fully agrees with observations presented in literature. In addition, obviously this asymptotic value can be reached with a smaller number of degrees of freedom using adaptive grid refinement compared to uniform remeshing.

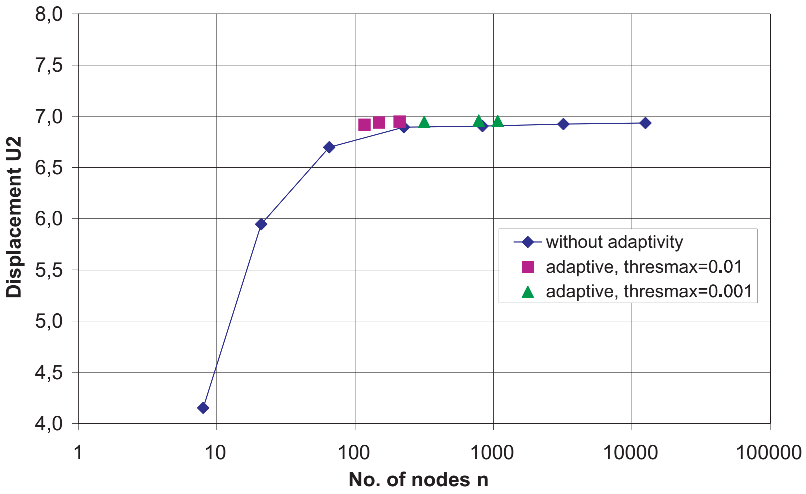


Fig. 3 Cook's membrane problem. Numerical simulation using rectangular elements without as well as with adaptivity. Vertical displacement of node A (cf. Fig. 1) dependent on the node number (semi-logarithmic representation).

6.2 Parameter Identification for Cook's Membrane Problem

This study is aimed to re-identify given parameters based on synthetic (i. e., numerically generated) ”measurements”, and to analyze the numerical behavior of the optimization procedure in the presence of disturbed measurements. For this purpose, the synthetic experimental data has been modified using normally distributed numerically generated ”errors”.

As mentioned above, analyzing the solution behavior of the direct Cook's membrane problem the displacement of the upper right corner node A (cf. Fig. 1) is of special interest. Therefore, parameter identification referred to the problem under consideration has been performed analyzing the components of the displacement vectors of all 225 grid nodes as well as, alternatively, using only the displacement of grid node A. Synthetic measurements have been available for all of the 100 load (i. e., time) increments that have been applied.

In the following, selected results of parameter identification procedures will be presented. Additionally to initial parameter values and optimization results, the number of optimization steps and deviations of numerical results from expected parameter values are given. The essential part of the discussion concerns the analysis of displacement fields, because they carry much more as well as more complex

Table 2 Parameter identification for the constitutive model (17a) (Neo-Hooke, nearly incompressible) analyzing the displacements of all nodes of the FE grid discretizing the Cook's membrane as well as exclusively of the upper right corner node (cf. Fig. 1 – node A). Reidentification of given parameters based on artificial measurements dependent on the measurement error and the chosen starting value for the optimization process (`nopti`: number of optimization steps).

	Optimized value of the parameter (expected: $c_{10} = 3.57\,\text{MPa}$)					
	Without measurement error			Measurement error 10 %		
Starting value c_{10} MPa	Absolute value MPa	Deviation %	`nopti`	Absolute value MPa	Deviation %	`nopti`
Displacements of all nodes						
0.357	3.570	0.00	7	3.570	0.01	6
2.000	3.570	0.00	5	3.570	0.01	5
7.140	3.570	0.00	7	3.570	0.01	7
35.700	3.570	0.00	10	3.570	0.01	10
Displacements of the upper right corner node (node A; cf. Fig. 1)						
0.357	3.570	0.00	7	3.547	0.65	7
2.000	3.570	0.00	5	3.548	0.62	4
7.140	3.570	0.00	7	3.547	0.65	7
35.700	3.570	0.00	10	3.547	0.65	10

information compared to global traction-displacement curves (i. e., various local load paths in individual material points are available).

The Neo-Hookean model is the simplest hyperelastic constitutive model characterized by an only weakly marked non-linearity. In the case of nearly incompressible material behavior this model contains only one material parameter c_{10} to be determined (cf. (17a)).

Numerical studies for Neo-Hookean's model calibration analyzing displacements of all grid nodes and of the upper right corner node of Cook's membrane, respectively have been performed within a wide range of initial parameter values and error arrays referred to synthetic measurements up to a maximum measurement error of 10%. Table 2 and Fig. 4 present selected numerical results showing an unexceptional problem-free progression of the optimization procedures. The predefined material parameter could be re-identified in all cases exactly or displaying deviations, which are much smaller than the maximum measurement error.

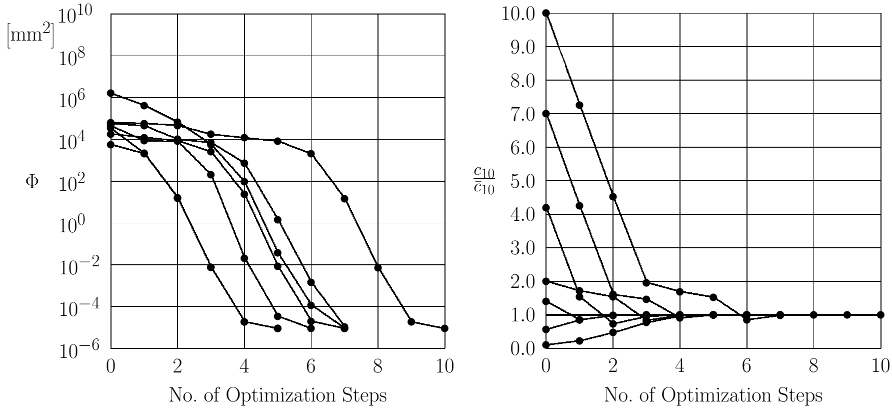


Fig. 4 Parameter identification for the constitutive model (17a) (Neo-Hooke, nearly incompressible) analyzing the displacements of all nodes of the FE grid discretizing the Cook's membrane. History of the optimization process for selected starting values assuming artificial measurements without measurement error. Left: Evolution of the objective function. Right: Evolution of the parameter c_{10} related to its given value $\bar{c}_{10} = 3.57\,\mathrm{MPa}$, which should be re-identified.

7 Conclusions

A constitutive model for the simulation of nearly incompressible elastic material behavior at large strains, including numerical methods for model calibration is reported in this contribution. The formulation of the mixed boundary value problem has been based on the weak forms of the linear momentum balance and of the incompressibility condition, both in material description. With respect to the direct problem, the consistent linearization of the boundary value problem results in association with appropriate spatial discretization techniques in an incremental-iterative solution strategy for he mixed two-field finite element formulation. The solution of the ill-posed inverse problem is based on deterministic optimization procedures of trust-region type evaluating a generalized objective function in terms of a least squares norm, including an efficient and high-precise approach for semi-analytical sensitivity analysis to determine the gradient of the objective function. To define the required relations for the inverse problem, an implicit differentiation of the corresponding weak formulations has been performed with respect to individual material parameters.

The developed numerical algorithms have been implemented into an in-house scientific finite element code. For stable solutions, elements of Taylor-Hood type are used in association with efficient iterative solvers for the global system based on advanced preconditioning techniques. Within the context of the hierarchical solver concept, adaptive methods for history-dependent modifications of the finite element grid can be realized appropriately. The grid adaptivity is mainly controlled by a residual a posteriori error estimator for mixed problems.

The above discussed isotropic hyperelastic constitutive model is based on the multiplicative decomposition of the deformation gradient into a deviatoric part and a volumetric part. For numerical tests, the Neo-Hookean formulation of the free Helmholtz energy density has been studied first. Capability and stability of the presented models and algorithms have been investigated on the frequently discussed literature example of the Cook's membrane problem. Related to the solution of the corresponding direct problem, an excellent agreement between the presented results and results from literature as well as simulations using commercial finite element software could be observed. In addition, the efficiency of hierarchical, adaptive grid modifications has been illustrated in comparison to full remeshing strategies. Within the context of the inverse problem, synthetic measurements have been generated based on predefined material parameters considering global as well as local information. The given material parameters could be re-identified successfully, additionally providing hints for the improvement of the reliability of the solution of the ill-posed inverse problem. Basically, it is recommended, e. g., to determine simultaneously as few parameters as possible based on as much information as possible. This comprises the calibration of parts of the entire set of parameters using different experiments that address the corresponding physical phenomena relevant for individual parameters or parameter subsets. Furthermore, strongly correlated parameters constitute significant problems in parameter identification. If such correlations cannot be avoided it is useful to integrate additional information into the optimization procedure, and to extend the calibration strategy, e. g., by methods of multi-parameter regularization.

The presented research constitutes the basis for further developments in mixed finite elements methods, comprising, e. g., the solution of the direct and inverse problems of hydro-mechanical processes in biphasic porous media at large strains. Within this context, analogies can be exploited in an efficient manner existing with the mixed formulation of nearly incompressible elastic material behavior. Additionally, the fundamental conceptual strategy being at the basis of the presented coupled model can be generalized for application to other multiphysics problems.

Acknowledgements. We express our deep gratitude to Prof. U. Langer (Johannes Kepler University Linz) and Prof. A. Meyer (Chemnitz University of Technology) for strengthening the mathematical foundation of our research in computational mechanics, and for the long-standing fruitful cooperation in common projects, including lots of intense and constructive discussions. This work was financed through DFG (German Research Foundation) Collaborative Research Center SFB 393 and DFG Coordinated Individual Grant PAK 47.

References

[1] Ainsworth, M., Oden, J.T.: A posteriori error estimators for the Stokes and Oseen equations. SIAM J. Numer. Anal. 34, 228–245 (1997)

[2] Armero, F.: On the locking and stability of finite elements in finite deformation plane strain problems. Comput. Struct. 75, 261–290 (2000)

[3] Ball, J.M.: Convexity conditions and existence theorems in nonlinear elasticity. Arch. Rational Mech. Anal. 3, 337–407 (1977)

[4] Bank, R.E., Welfert, B.D.: A posteriori error estimates for the Stokes equations: a comparison. Comput. Meth. Appl. Mech. Engrg. 82, 323–340 (1990)

[5] Bathe, K.-J.: Finite-Elemente-Methoden. Springer, Berlin (2002)

[6] Benedix, U.: Parameterschätzung für elastisch-plastische Deformationsgesetze bei Berücksichtigung lokaler und globaler Vergleichsgrößen (Parameter estimation for elasto-plastic material models considering local and global comparative quantities; in German). Dissertation, Report 4/2000, Institut für Mechanik der TU Chemnitz (2000)

[7] Bramble, J.H., Pasciak, J.E.: A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems. Math. Comp. 50(181), 1–17 (1988)

[8] Brezzi, F., Fortin, M.: Mixed and hybrid Finite Element Methods. Springer, New York (1991)

[9] Brink, U., Stein, E.: On some mixed finite element methods for incompressible and nearly incompressible finite elasticity. Comp. Mech. 19, 105–119 (1996)

[10] Bucher, A., Görke, U.-J., Kreißig, R.: A material model for finite elasto-plastic deformations considering a substructure. Int. J. Plast. 20, 619–642 (2004)

[11] Bucher, A., Görke, U.-J., Steinhorst, P., Kreißig, R., Meyer, A.: Ein Beitrag zur adaptiven gemischten Finite-Elemente-Formulierung der nahezu inkompressiblen Elastizität bei großen Verzerrungen (A contribution to the adaptive mixed finite element formulations for nearly incompressible elasticity at large strains; in German). Preprint CSC/07-06, TU Chemnitz (2007)

[12] Bucher, A., Meyer, A., Görke, U.-J., Kreißig, R.: A contribution to error estimation and mapping algorithms for a hierarchical adaptive FE-strategy in finite elastoplasticity. Comp. Mech. 36(3), 182–195 (2005)

[13] Bucher, A., Meyer, A., Görke, U.-J., Kreißig, R.: A comparison of mapping algorithms for hierarchical adaptive FEM in finite elasto-plasticity. Comp. Mech. 39(4), 521–536 (2007)

[14] Chen, J.S., Han, W., Wu, C.T., Duan, W.: On the perturbed Lagrangian formulation for nearly incompressible and incompressible hyperelasticity. Comput. Meth. Appl. Mech. Engrg. 142, 335–351 (1997)

[15] Cook, R.D.: Improved two-dimensional finite element. J. Struct. Div. ASCE 100, 1851–1863 (1974)

[16] Dennis, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall Inc., Englewood Cliffs (1983)

[17] Flory, P.J.: Thermodynamic relations for high elastic materials. Trans. Faraday Soc. 57, 829–838 (1961)

[18] Gelin, J.-C., Ghouati, O.: An inverse method for material parameter estimation in the inelastic range. Comp. Mech. 16, 143–150 (1995)

[19] Görke, U.-J., Bucher, A., Kreißig, R.: Ein Beitrag zur Materialparameteridentifikation bei finiten elastisch-plastischen Verzerrungen durch Analyse inhomogener Verschiebungsfelder mit Hilfe der FEM (A contribution to the identification of material parameters at large elasto-plastic strains analyzing inhomogeneous displacement fields using the finite element method; in German). Preprint SFB393/01-03, TU Chemnitz (2001)

[20] Görke, U.-J., Bucher, A., Kreißig, R.: Zur Numerik der inversen Aufgabe für gemischte (u/p) Formulierungen am Beispiel der nahezu inkompressiblen Elastizität bei großen Verzerrungen (Towards the numerics of the inverse problem for mixed (u/p) formulations on the example of nearly incompressible elasticity at large strains; in German). Preprint CSC/07-07, TU Chemnitz (2007)

[21] Herrmann, L.R.: Elasticity equations for nearly incompressible materials by a variational theorem. AIAA J. 3, 1896–1900 (1965)

[22] Hood, P., Taylor, C.: Navier-Stokes equations using mixed interpolation. In: Oden, J.T., Gallagher, R.H., Zienkiewicz, O.C., Taylor, C. (eds.) Finite Element Methods in Flow Problems, pp. 121–132. University of Alabama in Huntsville Press (1974)

[23] Hughes, T.J.R.: The finite element method. Dover Publications, New York (2000)

[24] Ibrahimbegovic, A., Taylor, R.L., Wilson, E.L.: A robust quadrilateral membrane finite element with drilling degrees of freedom. Int. J. Num. Meth. Engng. 30, 445–457 (1990)

[25] Johansson, H., Runesson, K.: Parameter identification in constitutive models via optimization with a posteriori error control. Int. J. Numer. Meth. Engng. 62, 1315–1340 (2005)

[26] Kay, D., Silvester, D.: A-posteriori error estimation for stabilized mixed approximations of the Stokes equations. SIAM J. Sci. Comp. 21, 1321–1336 (1999)

[27] Kreißig, R., Benedix, U., Görke, U.-J.: Statistical aspects of the identification of material parameters for elasto-plastic models. Arch. Appl. Mech. 71, 123–134 (2001)

[28] Kreißig, R., Benedix, U., Görke, U.-J., Lindner, M.: Identification and estimation of constitutive parameters for material laws in elastoplasticity. GAMM-Mitteilungen 30(2), 458–470 (2007)

[29] Lecampion, B., Constantinescu, A.: Sensitivity analysis for parameter identification in quasi-static poroelasticity. Int. J. Num. Anal. Meth. Geomech. 29(2), 163–185 (2005)

[30] Le Tallec, P.: Numerical methods for nonlinear three-dimensional elasticity. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. III, pp. 465–622. Elsevier, Amsterdam (1994)

[31] Mahnken, R., Kuhl, E.: Parameter identification of gradient enhanced damage models with the finite element method. Eur. J. Mech. A/Solids 18, 819–835 (1999)

[32] Mahnken, R., Stein, E.: The parameter-identification for visco-plastic models via Finite-Element-Methods and gradient methods. Modelling Simul. Mater. Sci. Eng. 2, 597–616 (1994)

[33] Mahnken, R., Stein, E.: Parameter identification for finite deformation elasto-plasticity in principal directions. Comput. Meth. Appl. Mech. Engrg. 147, 17–39 (1997)

[34] Mahnken, R., Steinmann, P.: Finite element algorithm for parameter identification of material models for fluid saturated porous media. Int. J. Num. Anal. Meth. Geomech. 25(5), 415–434 (2001)

[35] Masud, A., Xia, K.: A stabilized mixed finite element method for nearly incompressible elasticity. J. Appl. Mech. 72, 711–720 (2005)

[36] Meyer, A.: Grundgleichungen und adaptive Finite-Elemente-Simulation bei ”Großen Deformationen” (Basic equations and adaptive finite element simulation at large strains; in German). Preprint CSC/07-02, TU Chemnitz (2007)

[37] Meyer, A., Steidten, T.: Improvements and experiments on the Bramble-Pasciak type CG for mixed problems in elasticity. Preprint SFB393/01-12, TU Chemnitz (2001)

[38] Nocedal, J., Wright, S.J.: Numerical Optimization. Springer (1999)

[39] Ogden, R.W., Saccomandi, G., Sgura, I.: Fitting hyperelastic models to experimental data. Comp. Mech. 34, 484–502 (2004)

[40] Rüter, M., Stein, E.: Analysis, finite element computation and error estimation in transversely isotropic nearly incompressible finite elasticity. Comput. Meth. Appl. Mech. Engrg. 190, 519–541 (2000)

[41] Simo, J.C., Taylor, R.L.: Quasi-incompressible finite elasticity in principal stretches, continuum basis and numerical algorithms. Comput. Meth. Appl. Mech. Engrg. 85, 273–310 (1991)

[42] Taylor, C., Hood, P.: A numerical solution of the Navier Stokes equations using the finite element technique. Comput. Fluids 1, 73–100 (1973)

[43] Taylor, R.L., Pister, K.S., Herrmann, L.R.: On a variational theorem for incompressible and nearly-incompressible orthotropic elasticity. Int. J. Sol. Struct. 4, 875–883 (1968)

[44] Verfürth, R.: A review of a posteriori error estimation and adaptive mesh-refinement techniques. Wiley and Teubner, Chichester and Stuttgart (1996)

Application of the Reciprocity Principle for the Determination of Planar Cracks in Piezoelectric Material

Peter Steinhorst and Barbara Kaltenbacher

Abstract. This paper provides an extension of the reciprocity gap approach for crack detection from electrostatics [1], isotropic [2] and anisotropic linear elasticity [18] to piezoelectric materials. We show unique and stable identifiability of the crack plane from one or two pairs of appropriate Dirichlet-Neumann data and illustrate the approach by numerical tests with simulated data obtained by adaptive finite element computations.

1 Introduction

Piezoelectricity is a coupling between the electrical and the mechanical behavior of certain materials, which is exploited in a wide range of transducer applications. By far most of the piezoelectric devices nowadays consist of ceramics, which are known as brittle material. Therefore, the detection of cracks in piezoelectric ceramics from external measurements in the sense of non-destructive testing is a task which is of high practical interest.

This paper is a further extension of a method using the reciprocity principle, which has been introduced by ANDRIEUX, BEN ABDA et. al in the context of electrostatics [1], and in the case of isotropic linear elasticity [2]. As preliminary work we use a first extension of the method concerning crack plane determination in anisotropic linear elasticity [18], especially including the case of transversal isotropy like in the elastic part of the piezoelectric material tensor. Our extension provides a method which is applicable for crack plane determination in piezoelectrics, using only one or two pairs of Dirichlet and Neumann data on the outer boundary. More

Peter Steinhorst · Barbara Kaltenbacher
Institut für Mathematik, Alpen-Adria-Universität Klagenfurt,
Universitätsstraße 65-67, 9020 Klagenfurt, Austria
e-mail: peter.steinhorst@mathematik.tu-chemnitz.de,
barbara.kaltenbacher@aau.at

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 325–353.
springerlink.com

precisely, we will propose two approaches, based on either electrical or mechanical measurements to determine

- the normal of the crack plane,
- its distance to the boundary, and
- its (approximate) midpoint.

The paper is organized as follows. Sect. 2 provides the mathematical foundations for the inverse problem of crack detection and introduces the reciprocity gap functional in the context of piezoelectricity, along with some of its important properties. In Sect. 3, 4, and 5, we consider determination of crack normal, offset and midpoint, respectively, and discuss the choice of appropriate test functions in the reciprocity gap functional. Finally, in Sect. 6, we provide the results of a 2D-implementation in an existing FEM-software package on some numerical test examples.

2 Mathematical Formulation

2.1 The Forward Problem

Let Ω be a bounded LIPSCHITZ domain in $\mathbb{R}^d$ $(d \in \{2,3\})$, and the crack $\Sigma \subset \Pi$, where Π is a plane in $\mathbb{R}^3$ or a line in $\mathbb{R}^2$ and the distance between Σ and the outer boundary $\partial\Omega$ is strictly positive. Σ needs not necessarily be simply connected. If it consists of several components, all of them have to lie in the same plane Π.

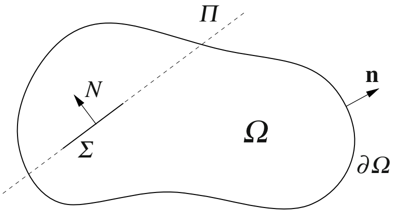


Fig. 1 Geometry setup and notation.

The whole boundary Γ of the cracked domain $\Omega \setminus \Sigma$ consists of the outer boundary $\partial\Omega$ and the crack faces of Σ.

The following derivations are valid in two or three space dimensions. However, for the sake of clarity and brevity we will show explicit formulas only for the 2D case.

The primary fields will be denoted by

- $u : \Omega \to \mathbb{R}^d$ the displacement field,
- $\varphi : \Omega \to \mathbb{R}$ the scalar electric potential field.

Inside $\Omega \setminus \Sigma$ we postulate the absence of volume forces and charges, which corresponds to the source free static piezoelectric differential equations:

$$\begin{aligned} \operatorname{div} \sigma(u,\varphi) &= \mathbf{0}\,, \\ \operatorname{div} D(u,\varphi) &= 0\,. \end{aligned} \tag{1}$$

The stress tensor σ and the dielectric displacement D are coupled via the linear piezoelectric material law:

$$\begin{aligned} \sigma(u,\varphi) &= \mathscr{C} : \varepsilon(u) \quad + \mathscr{E} \cdot \nabla\varphi, \\ D(u,\varphi) &= \mathscr{E}^\top : \varepsilon(u) \, - \, \mathscr{K} \cdot \nabla\varphi. \end{aligned} \tag{2}$$

In here, the strain tensor ε denotes the symmetric part of the displacement gradient:

$$\varepsilon(u) = (\nabla u)^s := \frac{1}{2}\left(\nabla u + (\nabla u)^\top\right), \quad \text{i.e.} \quad \varepsilon_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{3}$$

and the material tensors are

- $\mathscr{C} = \mathbf{c}_{ijkl}$ – the elastic material tensor (of 4th order),
- $\mathscr{K} = \kappa_{ij}$ – the dielectric tensor (of 2nd order),
- $\mathscr{E} = \mathbf{e}_{ijk}$ – the piezoelectric coupling tensor (of 3rd order).

The boundary conditions are of NEUMANN type in both fields, with given mechanical forces and electric charges on the outer boundary:

$$\sigma(u,\varphi)\mathbf{n} = F_m\,, \quad D(u,\varphi)\cdot\mathbf{n} = g_m \qquad \text{on } \partial\Omega\,. \tag{4}$$

Later on we will also use partial DIRICHLET data for crack identification. Nevertheless we formulate the forward problem with pure NEUMANN conditions in order to avoid the distinction of different cases with different partial DIRICHLET data and therewith keep the presentation relatively simple.

Denote the crack faces by Σ^+ and Σ^-. We introduce the following notation for the jumps of u and φ over Σ:

$$[\![u]\!] = u|_{\Sigma^+} - u|_{\Sigma^-}\,, \quad [\![\varphi]\!] = \varphi|_{\Sigma^+} - \varphi|_{\Sigma^-}\,.$$

The following traction free and impermeable boundary conditions are supposed to hold on the crack:

$$\sigma(u,\varphi)N = \mathbf{0}\,, \quad D(u,\varphi)\cdot N = 0 \qquad \text{on } \Sigma\,. \tag{5}$$

Here, $\mathbf{n}$ denotes the outer normal of $\partial\Omega$ and N the normal of the crack plane Π, which is unique up to orientation and without loss of generality can be assumed to be the outer normal of the crack face Σ^+. Both $\mathbf{n}$ and N are normalized to $\|\mathbf{n}\| = \|N\| = 1$.

Remark 1. The assumption (5) of traction-free and charge-free crack faces is often not a very realistic model in piezoelectricity, even in the absence of crack closing. Because of the expected small crack opening, the electric potentials on both sides of the crack influence each other. If a potential difference exists, a strong electric field inside the crack may result. This leads to other more realistic models like traction free and semipermeable conditions

$$\sigma(u,\varphi)N = \mathbf{0}\,, \quad D^{+}(u,\varphi)\cdot N = D^{-}(u,\varphi)\cdot N = -\kappa_{\mathrm{L}}\frac{[\![\varphi]\!]}{[\![u]\!]\cdot N} \qquad \text{on } \Sigma\,, \tag{6}$$

see [6], which are more complicated than the charge-free ones, though. A second effect which might influence the field distribution significantly, is the electrostatic attraction of the crack faces. This can also be considered in a modification of the elastic boundary conditions as compared to the simple traction-free assumption [7, 12]. The case of crack closing and contact between crack faces adds further challenges to the mathematical model and the numerical solution of crack simulation, cf., e.g., [14].

An equivalent matrix-vector-notation of (2), (3) using VOIGT's index mapping is

$$\begin{pmatrix}\underline{\sigma}(u,\varphi)\\ D(u,\varphi)\end{pmatrix} = \begin{pmatrix} C & E\\ E^{\top} & -K\end{pmatrix}\begin{pmatrix}\underline{\varepsilon}(u)\\ \nabla\varphi\end{pmatrix} =: \mathbf{A}\begin{pmatrix}\underline{\varepsilon}(u)\\ \nabla\varphi\end{pmatrix} \tag{7}$$

with

$$\begin{pmatrix}\underline{\varepsilon}(u)\\ \nabla\varphi\end{pmatrix} = \mathscr{B}\mathbf{U}\,.$$

C and K are positive definite matrices and x_d (recall $\Omega \subseteq \mathbb{R}^d$, $d \in \{2,3\}$) is the poling direction.

In the 2D case as electromechanical extension of the plain strain mode we have:

$$C = \begin{pmatrix} c_{11} & c_{12} & 0\\ c_{12} & c_{22} & 0\\ 0 & 0 & c_{33}\end{pmatrix}, \; E = \begin{pmatrix} 0 & e_{12}\\ 0 & e_{22}\\ e_{31} & 0\end{pmatrix}, \; K = \begin{pmatrix}\kappa_{11} & 0\\ 0 & \kappa_{22}\end{pmatrix},$$

$$\mathbf{U} = \begin{pmatrix} u_1\\ u_2\\ \varphi\end{pmatrix}, \; \mathscr{B} = \left(\begin{array}{ccc|cc} \partial_{x_1} & 0 & \partial_{x_2} & & \\ 0 & \partial_{x_2} & \partial_{x_1} & & \\ \hline & & & \partial_{x_1} & \partial_{x_2}\end{array}\right)^{\top},$$

$$\text{and} \qquad \underline{\sigma} = (\sigma_{11}, \sigma_{22}, \sigma_{12})^{\top}, \qquad \underline{\varepsilon} = (\varepsilon_{11}, \varepsilon_{22}, 2\varepsilon_{12})^{\top}.$$

To determine the test functions in the methods proposed below, we will also need the inverse material law

$$\begin{pmatrix}\underline{\varepsilon}(v)\\ \nabla\varphi\end{pmatrix} = \mathbf{A}^{-1}\begin{pmatrix}\underline{\sigma}(v,\psi)\\ D(v,\psi)\end{pmatrix} \qquad \text{with} \quad \mathbf{A}^{-1} = \begin{pmatrix} A & B\\ B^{\top} & -J\end{pmatrix}. \tag{8}$$

The inverse material matrix has the same block structure as the direct one in (7), the pattern of nonzeros in the matrix blocks is also analogous. In the two-dimensional case one obtains by translation formulas from [16] or [17] the constant entries of the matrices

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{12} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}, \; B = \begin{pmatrix} 0 & b_{12} \\ 0 & b_{22} \\ b_{31} & 0 \end{pmatrix}, \; J = \begin{pmatrix} \delta_{11} & 0 \\ 0 & \delta_{22} \end{pmatrix}.$$

As noted in [2] and shortly explained in [18], additional constraints are necessary to ensure existence and uniqueness of a solution in the case of pure NEUMANN problems. Extending these considerations to a piezoelectric material without inner sources (1) in the uncracked domain Ω, the following conditions on the (outer) boundary have to be fulfilled within (4) to ensure the existence of a solution of (1),(4):

$$\int\limits_{\partial\Omega} F_m dS = \mathbf{0}, \qquad \int\limits_{\partial\Omega} x \times F_m dS = \mathbf{0}, \qquad \int\limits_{\partial\Omega} g_m dS = 0. \tag{9}$$

Violation of these conditions would prevent the body from achieving a static equilibrium in a kinematic and electric sense. To ensure uniqueness, the primary fields of the solution also have to be restricted by some conditions

$$\int\limits_{\partial\Omega} u dS = \mathbf{0}, \qquad \int\limits_{\partial\Omega} x \times u dS = \mathbf{0}, \qquad \int\limits_{\partial\Omega} \varphi dS = 0. \tag{10}$$

These conditions fix the body in space in the sense that they eliminate the degrees of freedom for kinematic rigid motions (translations, rotations) and the addition of a constant electric potential.

In order to justify equations (9) and (10), in the 2D case we denote

$$\mathbf{U} = \begin{pmatrix} u \\ \varphi \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} v \\ \psi \end{pmatrix} \quad \text{and } \mathscr{N} = \left(\begin{array}{ccc|cc} n_1 & 0 & n_2 & & \\ 0 & n_2 & n_1 & & \\ \hline & & & n_1 & n_2 \end{array} \right)^{\top}, \tag{11}$$

where $\mathbf{n} = (n_1, n_2)^{\top}$ is the outer normal of Ω, and use the generalized GREEN's formula for linear piezoelectricity

$$-\int\limits_{\Omega} \mathscr{B}^{\top}\mathbf{A}\mathscr{B}\mathbf{U} \cdot \mathbf{V} dx = \int\limits_{\Omega} \mathbf{A}\mathscr{B}\mathbf{U} \cdot \mathscr{B}\mathbf{V} dx - \int\limits_{\partial\Omega} \mathscr{N}^{\top}\mathbf{A}\mathscr{B}\mathbf{U} \cdot \mathbf{V} dS. \tag{12}$$

We have $\mathscr{B}\mathbf{U} = \mathbf{0}$ if and only if $\varepsilon(u) = \mathbf{0}$ and $\nabla\varphi = \mathbf{0}$. The space $\mathscr{R}$ of rigid motions has dimension 4 with basis vectors

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Thus, the solvability and uniqueness conditions follow similarly to the linear elastic case [18].

Variational Formulation

Let $\mathbb{V}_0 \subset [H^1(\Omega\backslash\Sigma)]^{d+1}$ be the space of functions, which satisfy the conditions (10). Furthermore, define test functions v (displacement) and ψ (potential).

The domain $\Omega\backslash\Sigma$ with $\Sigma \neq \emptyset$ is not a LIPSCHITZ domain, so GREEN's formula

$$-\int\limits_{\Omega^*} [\operatorname{div}\sigma(u,\varphi)\cdot v + \operatorname{div} D(u,\varphi)\psi]\,dx \tag{13}$$
$$= \int\limits_{\Omega^*} [\sigma(u,\varphi):\varepsilon(v) + D(u,\varphi)\cdot\nabla\psi]\,dx - \int\limits_{\partial\Omega^*} [\sigma(u,\varphi)n\cdot v + D(u,\varphi)\cdot n\psi]\,dS$$

(see also (12)) is not applicable immediately. According to common practice [3], [9], we split Ω into two subdomains, each of them having the desired boundary smoothness, cf. Fig. 2, and note the relation to domain decomposition methods [10].

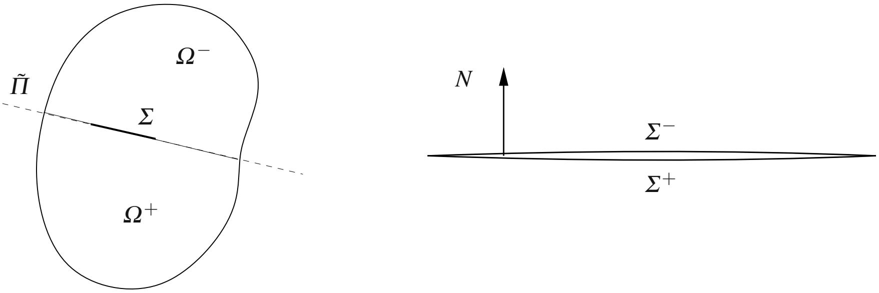


Fig. 2 Splitting of the domain and crack (displayed as slightly opened) with N as outer normal to the crack face Σ^+ and inner normal to Σ^-.

The splitting interface $\tilde{\Pi}$ is constructed as a continuation of Σ by a smooth surface which touches the outer boundary in normal direction so that both subdomains Ω^+, Ω^- are LIPSCHITZ domains. If Π intersects the outer boundary in a non-tangential manner, we may set $\tilde{\Pi} = \Pi$.

Imposing continuity of u, φ, $\sigma(u,\varphi)$ and $D(u,\varphi)$ on $\tilde{\Pi}\setminus\Sigma$, the application of GREEN's formula in each subdomain yields the variational problem:

$$\int\limits_{\Omega\backslash\Sigma} [\sigma(u,\varphi):\varepsilon(v) + D(u,\varphi)\cdot\nabla\psi]\,dx = \int\limits_{\partial\Omega} [F_m\cdot v + g_m\psi]\,dS \quad \forall(v,\psi)\in\mathbb{V}_0\,. \tag{14}$$

We point out a certain symmetry relation: Let $u, v \in [H^1(\Omega)]^d$, $\varphi, \psi \in H^1(\Omega)$. Furthermore, define the interdependence of $\sigma(u,\varphi)$ and $D(u,\varphi)$ by conditions (2) and in the matrix-vector-notation by (7), the same for u replaced by v and φ replaced by ψ. Then the following identity holds:

$$\sigma(u,\varphi) : \varepsilon(v) + D(u,\varphi)\cdot\nabla\psi = \sigma(v,\psi) : \varepsilon(u) + D(v,\psi)\cdot\nabla\varphi . \tag{15}$$

Indeed, by $C = C^\top$, $K = K^\top$, we have that $\mathbf{A} = \mathbf{A}^\top$, and therefore

$$\begin{aligned}\sigma(u,\varphi) : \varepsilon(v) + D(u,\varphi)\cdot\nabla\psi &\overset{(7)}{=} \begin{pmatrix}\underline{\varepsilon}(v)\\ \nabla\psi\end{pmatrix}^\top \mathbf{A} \begin{pmatrix}\underline{\varepsilon}(u)\\ \nabla\varphi\end{pmatrix} = \begin{pmatrix}\underline{\varepsilon}(u)\\ \nabla\varphi\end{pmatrix}^\top \mathbf{A} \begin{pmatrix}\underline{\varepsilon}(v)\\ \nabla\psi\end{pmatrix}\\ &= \sigma(v,\psi) : \varepsilon(u) + D(v,\psi)\cdot\nabla\varphi .\end{aligned}$$

Proposition 1. *For any $(F_m, g_m) \in [H^{-1/2}(\partial\Omega)]^{d+1}$ satisfying (9), there exists a unique solution $(u,\varphi) \in \mathbb{V}_0$ in the sense of (14) of the boundary value problem (1)–(5).*

Proof. Consider problem (14). Changing to matrix-vector-notation and considering (11), we get an equivalent formulation $a(\mathbf{U},\mathbf{V}) = f(\mathbf{V})$, where the bilinear form is defined as

$$a(\mathbf{U},\mathbf{V}) := \int\limits_{\Omega\setminus\Sigma} \Big((\underline{\varepsilon}(v))^\top C \underline{\varepsilon}(u) + (\underline{\varepsilon}(v))^\top E \nabla\varphi - (\underline{\varepsilon}(u))^\top E\nabla\psi + (\nabla\psi)^\top K \nabla\varphi \Big)\, dx .$$

By positivity of C and K we obviously have

$$a(\mathbf{U},\mathbf{U}) \geq c\Big(\|\varepsilon(u)\|_{L^2}^2 + \|\nabla\varphi\|_{L^2}^2 \Big)$$

for some $c > 0$. Coercivity of $a(.,.)$ in $\mathbb{V}_0$ then follows from (10), see, e.g. [18], thus existence and uniqueness of the solution can be concluded from the Lax-Milgram-Lemma. □

2.2 The Inverse Problem

We now assume, that the crack Σ is unknown, because it is completely hidden in the interior of Ω and hence not visible from outside. In order to detect the crack we consider measurements of the DIRICHLET data

$$u_m = u|_{\partial\Omega} \text{ (displacement field)} \quad \text{and/or} \quad \varphi_m = \varphi|_{\partial\Omega} \text{ (electric potential)}$$

on the outer boundary $\partial\Omega$, in addition to the known or just predefined NEUMANN boundary data

$$F_m = \sigma(u,\varphi)\mathbf{n}|_{\partial\Omega} \text{ (stresses)} \quad \text{and} \quad g_m = D(u,\varphi)\cdot\mathbf{n}|_{\partial\Omega} \text{ (charges)} .$$

These overdetermined boundary data provide additional information, which will be used to gain information on the location of the crack, considering it as interior boundary part. Under the application of certain loads (F_m, g_m), at first, the normal vector and subsequently the exact position of the crack plane Π shall be determined. Π is defined by the plane equation $N \cdot x = c$ in the standard orthonormal system $(O; e_1, e_2, (e_3))$ with corresponding coordinate vectors $x = (x_1, x_2, (x_3))^\top$. The inverse problem can be formulated as follows:

Problem 1. Determine Σ such, that the solution of the differential equation (1) in the domain $\Omega \setminus \Sigma$ with the NEUMANN boundary conditions (4) and (5) leads to a displacement field u and an electric potential φ with $u = u_m$ and/or $\varphi = \varphi_m$ on $\partial\Omega$.

2.3 Reciprocity Principle and Reciprocity Gap

In order to construct a reciprocity gap functional like in [18] and combine the approaches of [1] and [2] for use in piezoelectricity, we consider an extension of the reciprocity principle to piezoelectric material behavior. First, we formulate an extension of BETTI's reciprocity theorem to piezoelectric material in the case of divergence free stress fields and dielectric displacements.

Theorem 1. *Let Ω^* a bounded domain with a sufficiently smooth boundary (so that* GREEN*'s formula is valid), Γ the complete boundary of Ω^* and* **n** *the outer normal on Γ. Define*

$$\mathbb{H}(\Omega^*) = \left\{ (v, \psi) \in [H^1(\Omega^*)]^d \times H^1(\Omega^*) : \mathrm{div}\sigma(v, \psi) = \mathbf{0}, \mathrm{div} D(v, \psi) = 0 \textit{ in } \Omega^* \right\}.$$

Then, for all (u, φ), $(v, \psi) \in \mathbb{H}(\Omega^*)$ the following reciprocity relation holds:

$$\int_\Gamma (\sigma(u, \varphi)v + \psi D(u, \varphi)) \cdot \mathbf{n}\, dS = \int_\Gamma (\sigma(v, \psi)u + \varphi D(v, \psi)) \cdot \mathbf{n}\, dS.$$

Proof. First, we compute some divergences:

$$\begin{aligned} \mathrm{div}(\sigma(u, \varphi)v) &\overset{\sigma=\sigma^\top,(3)}{=} \underbrace{v \cdot \mathrm{div}\sigma(u, \varphi)}_{=0 \text{ for } (u,\varphi)\in\mathbb{H}(\Omega^*)} + \sigma(u, \varphi) : \varepsilon(v), \\ \mathrm{div}(\psi D(u, \varphi)) &= \nabla\psi \cdot D(u, \varphi) + \underbrace{\psi \mathrm{div} D(u, \varphi)}_{=0 \text{ for } (u,\varphi)\in\mathbb{H}(\Omega^*)}. \end{aligned}$$

This enables to apply GREEN's formula to prove the Theorem with the help of identity (15)

$$\begin{aligned}\int\limits_{\Gamma} (\sigma(u,\varphi)v + \psi D(u,\varphi)) \cdot \mathbf{n} dS \quad &= \quad \int\limits_{\Omega} [\sigma(u,\varphi) : \varepsilon(v) + \nabla\psi \cdot D(u,\varphi)]\, dx \\ &\overset{(15)}{=} \quad \int\limits_{\Omega} [\sigma(v,\psi) : \varepsilon(u) + \nabla\varphi \cdot D(v,\psi)]\, dx \\ &\overset{(v,\psi)\in\mathbb{H}(\Omega^*)}{=} \int\limits_{\Omega} [\mathrm{div}(\sigma(v,\psi)u) + \mathrm{div}(\varphi D(v,\psi))]\, dx \\ &= \quad \int\limits_{\Gamma} (\sigma(v,\psi)u + \varphi D(v,\psi)) \cdot \mathbf{n} dS. \qquad \square\end{aligned}$$

We will now make use of the spaces $\mathbb{H}(\Omega\backslash\Sigma)$ and $\mathbb{H}(\Omega)$, the latter being contained in $[C(M)]^{d+1}$ for any compact subset M of Ω by interior H^2 regularity of solutions to the defining equations in $\mathbb{H}(\Omega)$ and SOBOLEV's Embedding Theorem. Let $(u,\varphi) \in \mathbb{H}(\Omega\backslash\Sigma)$ be the solution of the problem (1)–(5) in the domain $\Omega\backslash\Sigma$ according to Proposition 1. Then, for each pair of functions $(v,\psi) \in \mathbb{H}(\Omega\backslash\Sigma)$ we define the extended reciprocity gap functional:

$$RG(v,\psi) = \int\limits_{\partial\Omega} [(\sigma(u,\varphi)v) - (\sigma(v,\psi)u) + \psi D(u,\varphi) - \varphi D(v,\psi)] \cdot \mathbf{n} dS \quad (16)$$

$$= \int\limits_{\partial\Omega} [F_m \cdot v - (\sigma(v,\psi)\mathbf{n}) \cdot u_m + \psi g_m - \varphi_m D(v,\psi) \cdot \mathbf{n}]\, dS\,. \quad (17)$$

Due to the choice of $\mathbb{H}(\Omega\backslash\Sigma)$, in an uncracked domain ($\Sigma = \emptyset$) the functional $RG(v,\psi)$ vanishes for each pair $(v,\psi) \in \mathbb{H}(\Omega\backslash\Sigma)$ as a consequence of Theorem 1.

However, if a crack exists, a reciprocity gap arises in the integral (16) because $\partial\Omega$ no longer represents the whole boundary of the domain $\Omega \setminus \Sigma$. The faces of Σ are additional parts of the boundary Γ.

Like in [19] at first proved under additional restrictions, one can show that the following merged extension of the corresponding Lemmas in [2] (also [18]) and [1] holds.

Lemma 1. *Let $\Omega \setminus \Sigma$ be a domain with a crack, $(u,\varphi) \in \mathbb{H}(\Omega\backslash\Sigma)$ the solution of the boundary value problem (1),(4) with one of the conditions (5) or (6). Then,*

$$RG(v,\psi) = \int\limits_{\Sigma} [[\![u]\!] \cdot (\sigma(v,\psi)N) + [\![\varphi]\!](D(v,\psi) \cdot N)]\, dS \qquad \forall (v,\psi) \in \mathbb{H}(\Omega)\,. \quad (18)$$

Note that by Proposition 1, the traces of u, φ in $H^{1/2}(\Sigma)$ are well-defined. Moreover, by interior H^2 regularity and the Trace Theorem, we have $\sigma(v,\psi)N$, $D(v,\psi) \cdot N \in H^{1/2}(\Sigma)$.

Proof. Like in the proof of Proposition 1, we split Ω into two subdomains Ω^+ and Ω^- by an appropriate extension $\tilde{\Pi}$ of the crack surface to the outer boundary of Ω, see the illustration in Fig. 2. The boundaries of Ω^+ and Ω^- exhibit sufficient smoothness, hence GREEN's formula holds in each part separately. Therewith the identity (18) is a direct consequence of (13) in connection with the continuity of the test functions (v,ψ) as well as $\sigma(v,\psi)$ and $D(v,\psi)$ over $\tilde{\Pi}$. □

Remark 2

- It is obvious from the proof of Lemma 1 that the assumption of traction-free crack faces can be replaced by the more general condition

$$\sigma^+(u,\varphi)N = \sigma^-(u,\varphi)N\,.$$

  Likewise for the electric interface conditions, we only need

$$D^+(u,\varphi)\cdot N = D^-(u,\varphi)\cdot N\,.$$

  Therefore, Lemma 1 remains valid for semipermeable interface conditions (6).
- In general, the computation of $RG(.,.)$ via (17) needs the complete elastic and electric CAUCHY data (DIRICHLET- and NEUMANN-) on the outer boundary. As shown in Sections 3–5 below, the choice of special test functions and loads allows for a reduction of the amount of necessary data.

Remark 3. As a matter of fact, the results of this section remain valid for the more general case of sufficiently smooth non-planar surfaces Σ,Π. For the constructions in the following Sections 3-5, planarity will be made use of.

In the next three sections we derive two approaches for identifying the crack normal, offset, and midpoint, based on additional electrostatic or mechanical measurements, respectively. We also discuss construction of the required test fields. To simplify the presentation, we restrict ourselves mainly to the two dimensional case in the remainder of this paper, but wish to emphasize that the following constructions can be extended to the 3D case as well.

3 Identification of the Crack Normal

3.1 Methods of Identification

We focus on two possible variants to identify the crack normal. In each of them exactly one summand vanishes in the integral of (18) and in the integral of (17), respectively, so that only either electrical or mechanical DIRICHLET data are required.

a) Identification from electrical measurements:
   Like in [1], we assume that the load is chosen such that $\int_\Sigma [\![\varphi]\!] dS$ does not vanish. Such loads can be constructed along the lines of the uniqueness proof in [5].
   Choose affinely linear v^i,ψ^i (thus, automatically $(v^i,\psi^i)\in\mathbb{H}(\Omega)$) with

$$D(v^i,\psi^i) = e_i\,, \qquad \sigma(v^i,\psi^i) = \mathbf{0}\,, \qquad i = 1,2(,3)\,. \tag{19}$$

Then the crack normal can be determined analogously to the pure electrostatic case (potential problem), as explained in [1]:

$$L_i = RG(v^i, \psi^i) = \int\limits_{\Sigma} [\![\varphi]\!] N_i dS = N_i \int\limits_{\Sigma} [\![\varphi]\!] dS .$$

Thus,

$$|L| = \sqrt{\sum_{i=1}^{d} RG^2(v^i, \psi^i)} = \left| \int\limits_{\Sigma} [\![\varphi]\!] \, dS \right| \quad \Longrightarrow N_i = \frac{L_i}{|L|} . \tag{20}$$

In 2D, using the inverse material law (8) we get

$$\underline{\varepsilon}(v^1) = \begin{pmatrix} 0 \\ 0 \\ b_{31} \end{pmatrix} , \quad \underline{\varepsilon}(v^2) = \begin{pmatrix} b_{12} \\ b_{22} \\ 0 \end{pmatrix} , \quad \nabla\psi^1 = -\begin{pmatrix} \delta_{11} \\ 0 \end{pmatrix} , \quad \nabla\psi^2 = -\begin{pmatrix} 0 \\ \delta_{22} \end{pmatrix} ,$$

hence it is readily checked that possible test functions satisfying (19) are:

$$v^1 = \frac{b_{31}}{2} \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} , \quad v^2 = \begin{pmatrix} b_{12} x_1 \\ b_{22} x_2 \end{pmatrix} , \quad \psi^1 = -\delta_{11} x_1 , \quad \psi^2 = -\delta_{22} x_2 .$$

b) Identification from mechanical measurements:
Analogously to a), we assume that the load is chosen such that $\int_\Sigma [\![u]\!] \, dS \neq 0$. Additionally we here need a second load with non vanishing displacement jump over Σ, whose normalization gives a direction different from the first one. Choose affinely linear v^{ij}, ψ^{ij} with

$$D(v^{ij}, \psi^{ij}) = \mathbf{0} , \qquad \sigma(v^{ij}, \psi^{ij}) = E^{ij} , \qquad i, j = 1, 2(, 3) , \tag{21}$$

where $E^{ij} \in \mathbb{R}^{d,d}$ and $E^{ij}_{kl} = \frac{1}{2}\left(\delta^i_k \delta^j_l + \delta^i_l \delta^j_k\right)$.

In this case, a determination of the crack normal is possible analogously to the case of linear-elastic material behavior as described in [2, 18].

We construct $R \in \mathbb{R}^{d,d}$ with $R_{ij} = RG(v^{ij}, \psi^{ij})$. It can be shown (cf. Proposition 1 in [18]), that

$$R = \frac{1}{2}(N\mu^\top + \mu N^\top) \qquad \text{with } \mu = \int\limits_{\Sigma} [\![u]\!] \, dS . \tag{22}$$

Furthermore, $\|\mu\|^2 = 2\|R\|_F^2 - (\mathrm{tr}R)^2$ and $\|\mu_n\|^2 = (\mu \cdot N)^2 = (\mathrm{tr}R)^2$, see Proposition 1 in [18]. Defining U as the normalized vector in the direction of μ, we have exactly $N + U$ and $N - U$ as eigenvectors to non vanishing eigenvalues of R (except for the special case $U = \pm N$, where only one non vanishing eigenvalue exists). Hence, N can be determined by computing $R^{(1)}$ and $R^{(2)}$ with two different loads, which lead to $U^{(1)} \neq U^{(2)}$ while N remains unchanged.

We point to Sect. 3.2 for a stability analysis of this procedure.

For the explicit construction of test fields satisfying (21) in 2D, we have to fulfill $\underline{\sigma}(v^{11},\psi^{11}) = (1,0,0)^\top$, $\underline{\sigma}(v^{22},\psi^{22}) = (0,1,0)^\top$ and $\underline{\sigma}(v^{11},\psi^{11}) = (0,0,\frac{1}{2})^\top$. Thus,

$$\underline{\varepsilon}(v^{11}) = \begin{pmatrix} a_{11} \\ a_{12} \\ 0 \end{pmatrix}, \quad \underline{\varepsilon}(v^{22}) = \begin{pmatrix} a_{12} \\ a_{22} \\ 0 \end{pmatrix}, \quad \underline{\varepsilon}(v^{12}) = \frac{1}{2}\begin{pmatrix} 0 \\ 0 \\ a_{33} \end{pmatrix},$$

$$\nabla\psi^{11} = \begin{pmatrix} 0 \\ b_{12} \end{pmatrix}, \quad \nabla\psi^{22} = \begin{pmatrix} 0 \\ b_{22} \end{pmatrix}, \quad \nabla\psi^{12} = \frac{1}{2}\begin{pmatrix} b_{31} \\ 0 \end{pmatrix},$$

and the corresponding test functions:

$$v^{11} = \begin{pmatrix} a_{11}x_1 \\ a_{12}x_2 \end{pmatrix},\ v^{22} = \begin{pmatrix} a_{12}x_1 \\ a_{22}x_2 \end{pmatrix},\ v^{12} = \frac{a_{33}}{4}\begin{pmatrix} x_2 \\ x_1 \end{pmatrix},$$
$$\psi^{11} = b_{12}x_2\,,\ \psi^2 = b_{22}x_2\,,\ \psi^{12} = \frac{b_{31}}{2}x_1\,.$$

Due to the choice of the test functions, in variant a) the summand $(\sigma(v,\psi)\cdot\mathbf{n})\cdot u$ vanishes in the reciprocity integral (16), whereas in variant b) the summand $(D(v,\psi)\cdot\mathbf{n})\cdot\phi$ vanishes. So, one type of DIRICHLET data is dispensable. It is not possible to avoid measurement of the NEUMANN data $\sigma(u,\varphi)\cdot\mathbf{n}$ on $\partial\Omega$ using this kind of test functions. This might be achieved by fixing boundary forces.

To evaluate the boundary integral, we also need the values of v and ψ on $\partial\Omega$. We first use the inverse material law (8) to compute $\varepsilon(v)$, $\nabla\psi$ and then obtain the desired primary fields by integration.

3.2 Stability

Since the given measurements F_m^δ, u_m^δ of F_m, u_m are actually contaminated by noise, we have to make sure that small perturbations in the data only lead to small perturbations in the reconstructed crack. We here mainly investigate the method for crack normal determination based on the displacement jump, i.e., b) in Subsection 3.1, since this involves computation of eigenvectors. The reconstruction of N relies on the fact that the measured matrix of reciprocity gap values $R_{ij} = RG(v^{ij},\psi^{ij})$ lies in the set

$$\mathscr{M} = \{\frac{1}{2}(ab^\top + ba^\top)\ :\ a,b\in\mathbb{R}^d, \|a\| = 1\} = \{\frac{1}{2}(ab^\top + ba^\top)\ :\ a,b\in\mathbb{R}^d\} \quad (23)$$

where $d\in\{2,3\}$, that can be characterized as follows:

Lemma 2. *For the set* $\mathscr{M}$ *as defined in* (23) *we have the identity*

$$\mathscr{M} = \widetilde{\mathscr{M}} := \{R\in\mathbb{R}^{d\times d}\ :\ (\mathrm{tr}R)^2 \le \|R\|_F^2\ \wedge\ \mathrm{rank}(R)\le 2\ \wedge\ R = R^\top\}.$$

Proof. The fact that $\mathscr{M} \subseteq \widetilde{\mathscr{M}}$ is readily checked by just computing the Frobenius norm (using $\|R\|_F^2 = \mathrm{tr}(R^\top R)$) and trace of a matrix of the form given in (23):

$$\|\frac{1}{2}(ab^\top + ba^\top)\|_F^2 - \left[\mathrm{tr}(\frac{1}{2}(ab^\top + ba^\top))\right]^2 = \frac{1}{2}(\|a\|^2\|b\|^2 - (a^\top b)^2) \geq 0\,,$$

where we have used the elementary identity $\mathrm{tr}(ab^\top) = a^\top b$ and the Cauchy-Schwarz inequality.

On the other hand, given an arbitrary $R \in \widetilde{\mathscr{M}}$ we set

$$\bar{R} = \frac{R}{\sqrt{2\|R\|_F^2 - (\mathrm{tr}R)^2}}\,,$$

where the expression under the square root is larger or equal to $\|R\|_F^2$. Abbreviating $\alpha = \mathrm{tr}\bar{R}$ and noting that $\alpha^2 \leq 1$ by $R \in \widetilde{\mathscr{M}}$, one can easily verify that the eigenvalues of $\bar{R}$ are given by (cf. Proposition 2 in [18])

$$\lambda_1 = \frac{\alpha+1}{2}\,, \quad \lambda_2 = \frac{\alpha-1}{2}\,, \quad (\lambda_3 = 0 \text{ if } d = 3)\,. \tag{24}$$

Existence of a vanishing eigenvalue $\lambda_3 = 0$ in case $d = 3$ immediately follows from the fact that $\mathrm{rank}(\bar{R}) = \mathrm{rank}(R) \leq 2$. Denoting by Φ_i a normalized eigenvector for λ_i, $i = 1,2$ and setting (cf. Subsect. 3.3 in [18])

$$z_1 = \sqrt{2(1+\alpha)}\Phi_1\,,\; z_2 = \sqrt{2(1-\alpha)}\Phi_2\,,\; a = \frac{z_1+z_2}{2}\,,\; b = \frac{z_1-z_2}{2}\,, \tag{25}$$

we obtain

$$\frac{1}{2}(ab^\top + ba^\top) = \frac{1}{4}(z_1 z_1^\top - z_2 z_2^\top) = \frac{1+\alpha}{2}\Phi_1\Phi_1^\top - \frac{1-\alpha}{2}\Phi_2\Phi_2^\top = \sum_{i=1}^{d}\lambda_i\Phi_i\Phi_i^\top = \bar{R}\,.$$

□

It can be shown (cf. Proposition 1 in [18]) that the exactly measured matrix entries lead to a matrix R that is contained in $\mathscr{M}$, see (22). However, even arbitrarily small perturbations may lead to a violation of the identity in (22). In order to extract the two directions N,U corresponding to a,b, we therefore have to map the measured data into the set $\mathscr{M}$ before applying the procedure sketched in item b) of Subsection 3.1. Note that plain metric projection is not an option, since $\mathscr{M}$ is not convex. Thus we propose an alternative strategy to construct, for a given R^δ with

$$\|R^\delta - R\|_F \leq \delta\,,$$

which is obviously satisfied for

$$\delta^2 = \sum_{i,j=1}^{d} \|v^{ij}\|^2_{L^q(\partial\Omega)} \|F^\delta_m - F_m\|^2_{L^p(\partial\Omega)} + \sum_{i,j=1}^{d} \|E^{ij}\mathbf{n}\|^2 \mathrm{meas}(\partial\Omega)^{2/q} \|u^\delta_m - u_m\|^2_{L^p(\partial\Omega)},$$

where $\frac{1}{p}+\frac{1}{q}=1$, a nearby $\tilde{R}^\delta \in \mathscr{M}$. Only the upper (lower) triangular part of R^δ is actually computed, the rest of the matrix is completed in a symmetric manner, hence R^δ is symmetric. We subtract a sufficiently large quantity from the diagonal, i.e.

$$\tilde{R}^\delta = R^\delta - \frac{\zeta}{d} I \quad \text{with } \zeta \in [\zeta_-, \zeta_+], \quad \zeta_\pm = \hat{\alpha} \pm \sqrt{\frac{d}{d-1}(\|R^\delta\|^2_F - \frac{1}{d}\hat{\alpha}^2)},$$

where we abbreviate $\hat{\alpha} = \mathrm{tr} R^\delta$. Therewith we obtain

$$(\mathrm{tr}\tilde{R}^\delta)^2 = \hat{\alpha}^2 - 2\hat{\alpha}\zeta + \zeta^2 \leq \|R^\delta\|^2_F + \frac{\zeta^2}{d} - 2\frac{\hat{\alpha}\zeta}{d} = \|\tilde{R}^\delta\|^2_F .$$

On the other hand, ζ should be chosen sufficiently small, namely in case $\|R^\delta\|^2_F \geq \hat{\alpha}^2$ we can set $\zeta = 0$ and in case $\|R^\delta\|^2_F < \hat{\alpha}^2$ we set $\zeta = \zeta_-$ with

$$0 \leq \zeta_- = \hat{\alpha} - \sqrt{\hat{\alpha}^2 - \frac{d}{d-1}(\hat{\alpha}^2 - \|R^\delta\|^2_F)} \leq \frac{d}{d-1}\frac{\hat{\alpha}^2 - \|R^\delta\|^2_F}{\hat{\alpha}} \leq \frac{d}{d-1}\frac{\hat{\alpha}^2 - \|R^\delta\|^2_F}{\|R^\delta\|_F},$$

where by $R \in \mathscr{M}$

$$\hat{\alpha}^2 - \|R^\delta\|^2_F \leq [(\mathrm{tr}R^\delta)^2 - (\mathrm{tr}R)^2] + [\|R\|^2_F - \|R^\delta\|^2_F],$$

hence altogether there exists a $\bar{\delta} > 0$ such that for all $\delta \in (0, \bar{\delta}]$ we have $|\zeta| \leq \hat{C}\delta$ for some $\hat{C} > 0$ and therewith

$$\|\tilde{R}^\delta - R\| \leq \bar{C}\delta \tag{26}$$

for some constant $\bar{C}$ depending only on $\|R\|_F$ and $\mathrm{tr}R$.

It remains to investigate stability of the procedure of finding a, b such that $\tilde{R}^\delta = \frac{1}{2}(ab^\top + ba^\top)$ as described in the proof of Lemma 2 and as used in the construction of N according to item b) in Subsect. 3.1, see also [18] for more details. For this purpose consider $R = (R_{ij})_{ij\in\{1,\dots,d\}}$ (i.e., we omit the tilde and the superscript δ in the matrix entries). It is readily checked that an eigenvector corresponding to $\lambda_i = \frac{\alpha\pm 1}{2}$ is given by

$$\check{\Phi}_i := \begin{pmatrix} -R_{12} \\ \frac{R_{11}-R_{22}\mp 1}{2} \end{pmatrix}, \quad i = 1, 2, \tag{27}$$

in case $d = 2$. Computation of these vectors obviously depends in a Lipschitz stable manner on the matrix entries. The same holds for the computation of the norms of these vectors. Hence, unless $\|\check{\Phi}_i\|$ vanishes, defining $\check{\Phi}^\delta_i$ as the vectors given in (27), with R_{ij} replaced by $\tilde{R}^\delta_{ij}$, we get, for $\Phi^{(\delta)}_i = \check{\Phi}^{(\delta)}_i / \|\check{\Phi}^{(\delta)}_i\|$ the stability estimate

$$\|\Phi_i^\delta - \Phi_i\| \le c\delta$$

for all δ sufficiently small and some constant c depending only on $\|\Phi_i\|$. The problem of vanishing norm $\|\check{\Phi}_i\|$ can be circumvented by using alternative eigenvector representations (that have to be collinear to $\check{\Phi}_i^\delta$ since the eigenspace must have dimension one) such as

$$\bar{\Phi}_i^\delta = \begin{pmatrix} \frac{R_{22}-R_{11}\mp 1}{2} \\ -R_{12} \end{pmatrix}, \quad i = 1,2 \tag{28}$$

in 2D, in case $\|\check{\Phi}_i^\delta\| \le \bar{C}\delta$ with $\bar{C}$ as in (26) (so that $\|\check{\Phi}_i\|$ potentially vanishes). It can be shown that not all of these eigenvector representations vanish simultaneously, so that we can always select a stable representation. We demonstrate this in the 2D case: If $\|\check{\Phi}_i^\delta\| \le \bar{C}\delta$ then we have

$$\|\bar{\Phi}_i^\delta\|^2 = R_{12}^2 + \frac{(R_{22} - R_{11} \mp 1)^2}{4} = \|\check{\Phi}^\delta\|^2 + 1 \pm (R_{11} - R_{22} \mp 1) \ge 1 - \bar{C}\delta \ge \frac{1}{2}$$

for all $\delta \in (0, \min\{\frac{1}{2\bar{C}}, \bar{\delta}\}]$.

The remaining computations according to (25) are obviously again Lipschitz stable, so that the vectors a^1, b^1, a^2, b^2, for the two matrices $R^{(1)}$, $R^{(2)}$ can be stably approximated to within an order of magnitude of δ, i.e.,

$$\|v^\delta - v\| \le \overline{C}\delta, \quad v \in \{a^1, b^1, a^2, b^2\}$$

for all $\delta \in (0, \overline{\delta}]$ with some sufficiently small $\overline{\delta}$ and some constant $\overline{C}$ independent of δ. Assuming that we can choose the two loads such that the corresponding normalized displacement jump vectors $U^{(1)}$, $U^{(2)}$ according to (22), $U^{(i)} = \mu^{(i)}/\|\mu^{(i)}\|$, $i = 1,2$, differ by at least $\overline{C}\delta$ from each other and from N, we therefore end up with an overall estimate

$$\|N^\delta - N\| \le \overline{C}\delta\,.$$

4 Determination of the Crack Plane or -line Π

In order to reconstruct the crack plane it remains to determine its distance to the origin. For this purpose we can again choose between an approach using purely electric or purely mechanical measurements, respectively.

Having obtained the crack normal N according to Sect. 3 we can – like in the electrostatic or in the anisotropic-elastic case – generate a new coordinate system containing the crack normal N as a coordinate direction. Let the new coordinate system be denoted by $(O'; T_1, (T_2), N)$, respectively, with corresponding coordinates X_k, and the origin O' positioned outside the domain Ω in order to obtain uniqueness of the sign of the plane offset. In these new coordinates, the crack plane is defined by the equation $X_d = C$, and the crack normal has the representation $N = (0, (0,)1)^\top$.

Material Law in Rotated Coordinates

For the transformation into the new coordinate system we have to perform a rotation.

As introduced in [18] we define matrices $P(\phi)$ and $Q(\phi)$, which transform vector strains and potential gradients from the X- to the x-coordinates via

$$\underline{\tilde{\varepsilon}}(\tilde{u}) = P(\phi)\,\underline{\varepsilon}(u)\,, \qquad \tilde{\nabla}\varphi = Q(\phi)\nabla\varphi\,.$$

The inverse material law results as

$$\begin{pmatrix}\underline{\varepsilon}(u)\\ \nabla\varphi\end{pmatrix} = \begin{pmatrix}P(-\phi) & \\ & Q(-\phi)\end{pmatrix}\begin{pmatrix}A & B\\ B^\top & -J\end{pmatrix}\begin{pmatrix}P^\top(-\phi) & \\ & Q^\top(-\phi)\end{pmatrix}\begin{pmatrix}\underline{\sigma}(u,\varphi)\\ D(u,\varphi)\end{pmatrix}\,. \tag{29}$$

4.1 Variants for Determining the Plane Offset

Our aim is now to construct appropriate test functions v and ψ in the context of the coupled material law, in order to compute C from the value $RG(v,\psi)$ using Lemma 1. For this purpose, correspondingly to the crack normal determination, we propose two approaches.

a) Identification from electrical measurements:
Choose $(v_E,\psi_E) \in \mathbb{H}(\Omega)$ with

$$D(v_E,\psi_E)\cdot N = X_2\,, \qquad \sigma(v_E,\psi_E)\cdot N = \mathbf{0}\,. \tag{30}$$

Inserting into equation (18) of Lemma 1 we obtain

$$RG(v_E,\psi_E) = \int\limits_\Sigma [\![\varphi]\!]\cdot X_d\,dS = C\int\limits_\Sigma [\![\varphi]\!]\,dS\,.$$

Analogously to the electrostatic case [1] we can now determine C by means of the reciprocity integrals already computed for determination of the crack normal cf. equation (20) in Sect. 3) as follows:

$$C = \frac{|RG(v_E,\psi_E)|}{|L|}\,.$$

To find test functions in $\mathbb{H}(\Omega)$ that satisfy equation (30) in 2D, we first of all consider the divergence free dielectric displacement

$$D(v,\psi) = \begin{pmatrix}-X_1\\ X_2\end{pmatrix}$$

like in the electrostatic case and vanishing stress $\sigma(v,\psi)$. Therewith, the inverse material law (29) would yield

$$\underline{\varepsilon}(v) = P(-\phi)BQ^\top(-\phi)D(v,\psi)\,, \qquad \nabla\psi = -Q(-\phi)JQ^\top(-\phi)D(v,\psi)\,.$$

While it is always possible to construct a displacement field v corresponding to $\underline{\varepsilon}(v)$, a gradient field has to satisfy the integrability condition. With the notation $Q(-\phi)JQ^\top(-\phi) = (d_{ij})$, the expression

$$\nabla\psi = \begin{pmatrix} d_{11}X_1 - d_{12}X_2 \\ d_{12}X_1 - d_{22}X_2 \end{pmatrix}$$

with $d_{12} = (\delta_{11} - \delta_{22})\cos\phi\sin\phi$ for general angle ϕ only defines a gradient field if $d_{12} = -d_{12} = 0$ holds. However, in anisotropic material this is violated, since in general $\delta_{11} \neq \delta_{22}$.

This problem can be solved by adding divergence free terms:

$$D(v_E,\psi_E) = \begin{pmatrix} -X_1 + \gamma_E X_2 \\ X_2 \end{pmatrix}, \quad \sigma(v_E,\psi_E) = \mathbf{0}$$

or

$$D(v_E,\psi_E) = \begin{pmatrix} -X_1 \\ X_2 \end{pmatrix}, \quad \sigma(v_E,\psi_E) = \gamma_{E,\sigma}\begin{pmatrix} X_2 & 0 \\ 0 & 0 \end{pmatrix}.$$

The integrability condition yields

$$\begin{aligned}
\gamma_E &= -2\frac{d_{12}}{d_{11}} = 2\frac{(\delta_{22} - \delta_{11})\cos\phi\sin\phi}{\delta_{11}\cos^2\phi + \delta_{22}\sin^2\phi}\,,\\
\gamma_{E,\sigma} &= 2\frac{d_{12}}{(1{,}0)Q(-\phi)B^\top P^\top(-\phi)(1{,}0{,}0)^\top} = 2\frac{(\delta_{22} - \delta_{11})\cos\phi}{(b_{31} + b_{12})\cos^2\phi + b_{22}\sin^2\phi}\,.
\end{aligned}$$

b) Identification from mechanical measurements:
Choose $(v_T,\psi_T) \in \mathbb{H}(\Omega)$ with

$$D(v_T,\psi_T)\cdot N = 0\,, \qquad \sigma(v_T,\psi_T)\cdot N = \begin{pmatrix} X_2 \\ 0 \end{pmatrix}. \tag{31}$$

Inserting into equation (18) of Lemma 1 we here arrive at

$$RG(v_T,\psi_T) = \int\limits_\Sigma [\![u_T]\!]\cdot X_d\,dS = C\int\limits_\Sigma [\![u_T]\!]\,dS$$

with the possibility of determining C like in the purely elastic case ([2, 18]):

$$C = \frac{|RG(v_T,\psi_T)|}{\sqrt{2(\|R\|_F^2 - (\mathrm{tr}R)^2)}}\,. \tag{32}$$

When looking for test functions in $\mathbb{H}(\Omega)$ that satisfy equation (31) in 2D, the choice

$$D(v,\psi) = \mathbf{0}\,, \qquad \sigma(v,\psi) = \begin{pmatrix} -X_1 & X_2 \\ X_2 & 0 \end{pmatrix}$$

via the inverse material law (29) in general does not lead to a gradient field $\nabla\psi$ either. Similarly to variant a) we therefore choose

$$D(v_T,\psi_T) = \gamma_T \begin{pmatrix} X_2 \\ 0 \end{pmatrix}\,, \quad \sigma(v,\psi) = \begin{pmatrix} -X_1 & X_2 \\ X_2 & 0 \end{pmatrix}$$

or

$$D(v_T,\psi_T) = \mathbf{0}, \quad \sigma(v,\psi) = \begin{pmatrix} -X_1 & X_2 \\ X_2 & 0 \end{pmatrix} + \gamma_{T,\sigma} \begin{pmatrix} X_2 & 0 \\ 0 & 0 \end{pmatrix}$$

and by inserting into the integrability condition get the parameter γ_T or $\gamma_{T,\sigma}$, respectively. By integration we finally obtain test fields for displacement and potential in both cases.

Like in Sect. 3, version a) only requires voltage measurements whereas in version b) only measurements of the mechanical displacement are needed.

5 Approximative Mid Point Determination of a Single Crack

Nearly the same idea as in Sect. 4 can be used to find the center of the crack in a certain sense. Using the introduced new coordinates with the (possibly again moved) origin s.t. $\forall X \in \Omega : X_i \geq 0,\ 1 \leq i \leq d$, it remains to find X_i, $1 \leq i \leq d-1$. Again we evaluate $RG(v,\psi)$ for some appropriate test functions in two variants:

a) Identification from electrical measurements:
For $1 \leq i \leq d-1$ choose $(v_{M,i},\psi_{M,i}) \in \mathbb{H}(\Omega)$ with

$$D(v_{M,i},\psi_{M,i}) \cdot N = X_i\,, \quad \sigma(v_{M,i},\psi_{M,i}) \cdot N = \mathbf{0} \tag{33}$$

by Lemma 1 implies $RG(v_{M,i},\psi_{M,i}) = \int_\Sigma [\![\varphi]\!] \cdot X_i\, dS$. Hence, using $|L|$ from Sect. 3 we define the point $\xi \in \Pi$ with

$$\xi_d = C\,, \quad \xi_i = \frac{|RG(v_{M,i},\psi_{M,i})|}{|L|} \text{ for } 1 \leq i \leq d-1\,.$$

b) Identification from mechanical measurements:
For $1 \leq i \leq d-1$ choose $(v_{m,i},\psi_{m,i}) \in \mathbb{H}(\Omega)$ with

$$D(v_{m,i},\psi_{m,i}) \cdot N = 0\,, \qquad \sigma(v_{m,i},\psi_{m,i}) \cdot N = X_i\, e_d\,. \tag{34}$$

Here, with the matrix R from Sect. 3 and $\|\mu_n\| = |\mathrm{tr}R|$ we get ξ with

$$\xi_i = \frac{|RG(v_{m,i},\psi_{m,i})|}{|\mathrm{tr}R|},\ 1 \leq i \leq d-1\,.$$

In both cases it can be shown, that ξ is in the convex hull of Σ. In the case of a convex single crack with nearly symmetric behavior of the potential jump (a) or of the normal displacement jump (b), ξ gives a good approximation for the center of Σ.

As in Sect. 4, admissible test functions require the gradient property of $\nabla\psi$. We ensure this by adding X_d-parts to $D(v,\psi)$. In the two dimensional case this leads to

$$D(v_M,\psi_M) = \begin{pmatrix} \gamma_{cE} X_2 \\ X_1 \end{pmatrix}, \ \sigma(v_E,\psi_E) = \mathbf{0};$$

$$D(v_m,\psi_m) = \begin{pmatrix} \gamma_{cT} X_2 \\ 0 \end{pmatrix}, \ \sigma(v_E,\psi_E) = \begin{pmatrix} 0 & 0 \\ 0 & X_1 \end{pmatrix};$$

with

$$\gamma_{cE} = \frac{\delta_{11}\sin^2\phi + \delta_{22}\cos^2\phi}{\delta_{11}\cos^2\phi + \delta_{22}\sin^2\phi}, \quad \gamma_{cT} = -\frac{\cos\phi\,\left((b_{12}+b_{31})\sin^2\phi + b_{22}\cos^2\phi\right)}{\delta_{11}\cos^2\phi + \delta_{22}\sin^2\phi}.$$

Again, like in Sect. 3, 4, only either electrical (version a)) or mechanical (version b)) measurements are required.

6 Numerical 2D-Examples

6.1 Setup of Examples

In extension to the purely elastic setting in [18], we consider test problems with piezoelectric material behavior. For this purpose we use the full material parameter set of the ceramic PZT4_b, as can be found, e.g., in [15] or [16].

We always solve the forward problem (1)–(5), that means always vanishing normal stresses and electric impermeability are supposed on the crack. The polarization is in x_2-direction in all tests.

In Example 1 we consider as computational domain $\Omega = [0,8] \times [0,2]$, and deal with a horizontal crack which is split unsymmetrically in two parts, explicitly $\Sigma = \{x \in [1,2] \cup [4,6],\ y = 1\}$. The outer boundary $\partial\Omega$ is charge free ($g_m = 0$) with purely mechanical loads $F_m^{(i)}$ displayed by arrows in Fig. 3.

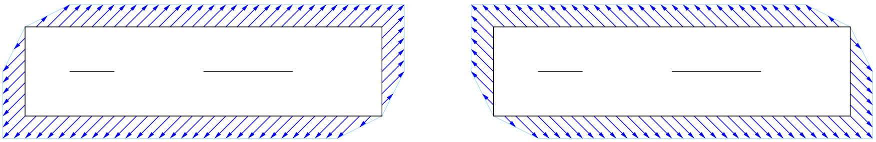


Fig. 3 Illustration of the loads $F_m^{(1)}$ and $F_m^{(2)}$ in Example 1.

In Example 2 we consider $\Omega = [0,4]^2$ and a skew crack $\Sigma = \{(x,x),\ x \in [1,3]\}$. On $\partial\Omega$ we introduce 4 variants of boundary conditions, cf. the illustration in Fig. 4:

- vertical and horizontal loads $F_m^{(1m)}$ and $F_m^{(2m)}$ with $g_m^{(1m)} = g_m^{(2m)} = 0$ (purely mechanical load cases like in Example 1),
- charges $g_m^{(1e)}$ on the upper and lower boundary part and $g_m^{(2e)}$ on the left and right boundary part with $F_m^{(1e)} = F_m^{(2e)} = 0$ (purely electric load cases, which also cause a crack opening).

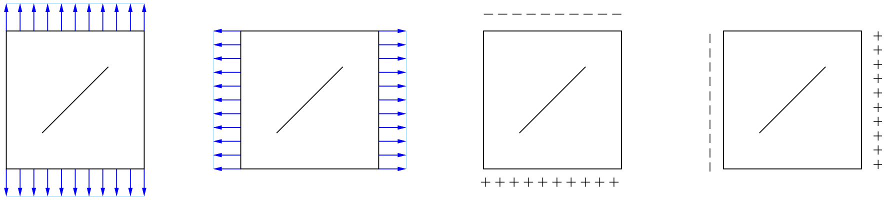


Fig. 4 Illustration of the loads $F_m^{(1\mathrm{m})}$, $F_m^{(2\mathrm{m})}$, $g_m^{(1\mathrm{e})}$ and $g_m^{(2\mathrm{e})}$ in Example 2.

Note that due to the material anisotropy, different states are generated by vertical and horizontal loads, respectively.

To provide simulated boundary data in place of real measurements, we use a forward computation with the variant SPC-PM2AdPiez of the 2D-FEM software package SPC-PM2Ad [13], written on the Chemnitz University of Technology. The inverse computation was implemented as post processing in this software.

Different levels of accuracy of these FEM generated synthetic data provide us with the possibility of assessing the effect of noisy data (in practice due to measurement errors) on the quality of our reconstructions.

Remark 4. Note, that an inverse crime is avoided here by using a completely different method for data simulation (namely FEM) than for the reconstruction (namely the evaluation of boundary integrals with the reciprocity gap approach described in Sect. 3-5).

Remark 5. In this section, convergence is in terms of the mesh size h (or the number of nodes). This corresponds to the noise level δ in Sect. 3.2 in the sense that smaller h corresponds to smaller δ.

6.2 FE-Meshes and Refinement Strategies in the Forward Computation

In each example we start with a coarse mesh of 16 uniform squares (8×2 and 4×4 respectively), where each square is divided once into two triangular elements.

The simplest strategy to compute a more precise solution is uniform mesh refinement, where the next refinement level results from splitting each element into four sub elements (red division). In Example 2, three additional refinement strategies have been tested for comparison. Besides the red division, also a green division (splitting of an element into two sub elements) was allowed to guarantee conforming meshes without hanging nodes.

- **adaptive refinement**
  This strategy applies a residual based error estimator, that means we examine the approximate FEM-solution u,φ on the current mesh refinement level to obtain information where the mesh should be refined [20]. Especially edge jumps of $\sigma(u,\varphi)n$ and $D(u,\varphi)n$ are considered in order to detect edges to be marked for refinement [11, 17].
- **(exclusive) boundary concentrated refinement**
  In each step, only all edges on the outer boundary are marked for refinement. This leads to refinement of all elements on the boundary, with a red-green closure to the inner domain. Some details on the boundary concentrated FEM can be found for e.g., in [4, 8].
- **combined adaptive and boundary concentrated refinement**
  In this combined strategy, all boundary edges are marked for refinement in addition to those marked by the error estimator, in order to achieve a good solution in the whole domain with especially fine resolution on the boundary.

All these strategies follow the intention to obtain an approximate solution, which has a good accuracy, but needs a significantly lower number of nodes and elements than the one obtained from a uniform refinement. The idea of using boundary concentrated strategies is motivated by the fact, that the inverse computation evaluates only boundary integrals.

To illustrate all refinement strategies used here, the deformed meshes in each case are illustrated after refinement up to more than 4000 nodes in Fig. 5.

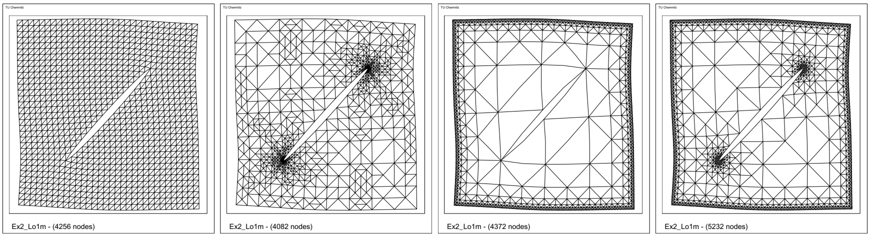


Fig. 5 Deformed meshes obtained by uniform, adaptive, boundary concentrated and a combined adaptive and boundary concentrated mesh refinement.

Computation of the Boundary Integrals

The computation of $RG(v,\psi)$ (16),(17) requires the integration of

$$v^\top F_m\ , \quad \mathbf{n}^\top\sigma(v,\psi)u_m\ , \quad \psi g_m \quad \text{and} \quad \mathbf{n}^\top D(v,\psi)\varphi_m$$

along each boundary edge E. We use a quadratic approximation of u and φ as well as linearly approximated F_m and g_m in each element. Since v,ψ are of polynomial degree less than or equal to two, each summand of the integrand has a maximal polynomial degree of three, so that SIMPSON's rule provides an exact integration of $RG(v,\psi)$ regarding the approximate solution along the edge.

6.3 Results with Uniform Refinement

6.3.1 Development of Computed Integrals during Refinement

To demonstrate the increasing precision of computational results on different refinement levels, Example 2 here is examined using the first load. The computed entries of $R_{ij} = RG(v^{ij},\psi^{ij})$, $L_i = RG(v^i,\psi^i)$ and $\|\mu\|^2 = (\int_\Sigma [\![u]\!] dS)^2$ are displayed in Fig. 6.

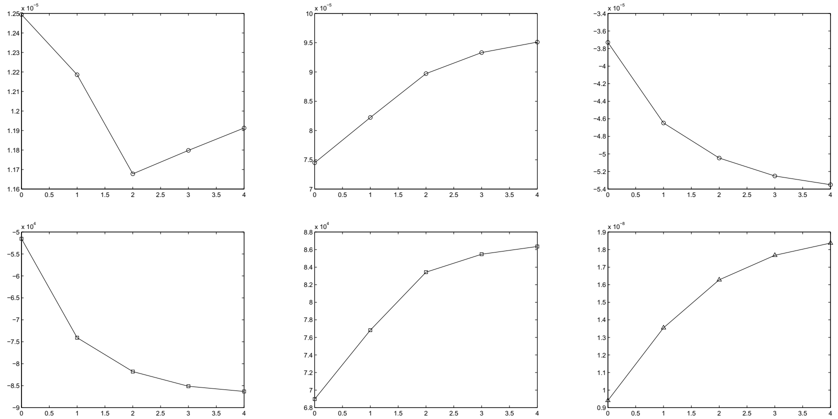


Fig. 6 Development of R_{11}, R_{22}, R_{12} (upper pictures) and L_1,L_2, $\|\mu\|^2$ (lower pictures) in Example 2 with increasing refinement level.

The convergence of these preliminary values during refinement is slow in comparison to the further derived results, which are shown below.

6.3.2 Determination of the Crack Normal

Determination of the crack normal is carried out as described in Sect. 3 according to variants a) and b). In Example 1 both variants yield reasonable results for the angle, as comparison with the correct solution $\phi = 0$ show, see Fig. 7, where we plot the development for increasing refinement in the FEM simulation of the data.

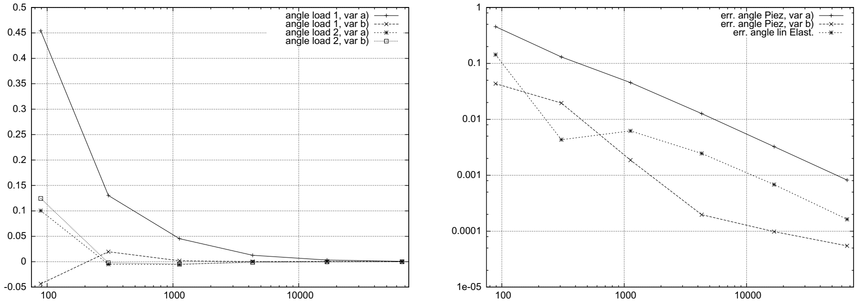


Fig. 7 Reconstruction of crack angle (left) and errors in comparison to purely elastic material (right) for increasing number of nodes in FEM simulation of boundary data, Example 1.

Also the error in the reconstructed angle is displayed in comparison with the purely elastic case. Here we plot the absolute values of the reconstructed angles versus the number of nodes in the first load case on a logarithmic scale along both axes. It can be seen that convergence in case of piezoelectric material as compared to the purely elastic case is not significantly slower.

In Example 2 investigations are somewhat more extended, here four load cases have to be evaluated.

In Fig. 8 the computational results of both variants are displayed and compared with the correct angle $\phi^* = -\frac{\pi}{4}$.

Here it turns out that the loads 1e and 2e yields almost the same angle with each variant.

When using variant a) with the load case 2m a drastic error is observed at a coarser discretization, only after multiple refinement the computed angle approximates the exact one. The reason for this behavior is the relatively small resulting potential jump over the crack, (see Fig. 9) which results in a reduced sensitivity of the inverse method.

The large initial deviation in load case 2m of variant b) results from the fact that due to numerical errors an infeasible matrix R is generated. The diagonal shift modification of R described in Sect. 3.2 enables the computation of the admissible angle $\phi \approx -1.14$, but just leaves some error in the result. After the first refinement step, the errors are in the order of magnitude of the other load cases.

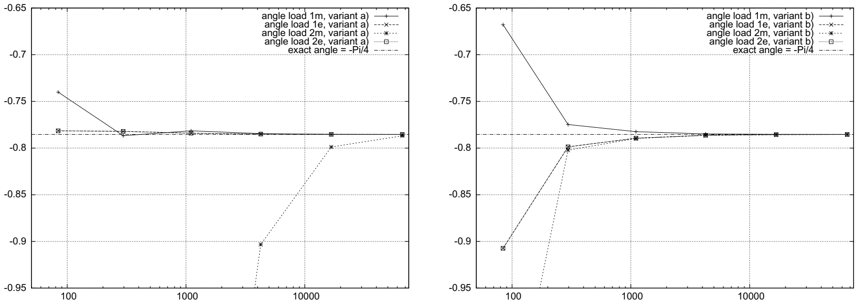


Fig. 8 Reconstruction of crack angle with variants a) (left) and b) (right) of Subsect. 3.1 in Example 2.

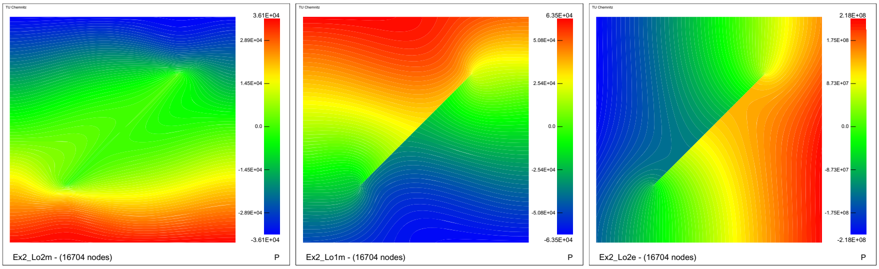


Fig. 9 Potentials in load case 2m (left) in comparison to load cases 1m and 2e in Example 2.

To investigate convergence of the computed angles to the exact one $\phi^* = -\frac{\pi}{4}$ with proceeding refinement we consider the absolute values of the deviations from ϕ^*. In Fig. 10 these are shown in double logarithmic scale with respect to the number of nodes in all four load cases, with variant a) and b).

Here we compare the errors both under the horizontal loads 2m, 2e and under the vertical ones 1m, 1e, with the results for purely elastic materials, where approximately the same convergence behavior can be observed.

6.3.3 Determination of the Crack Plane Distance to the Origin and the Approximate Mid Point Coordinate

Assuming the crack normal to be known, determination of the constant according to Sect. 4 was carried out with both variants a) and b).

Here we use the additional terms with the parameters γ_E and γ_T (cf. Sect. 4.1). The use of $\gamma_{E,\sigma}$- and $\gamma_{T,\sigma}$-terms shows a tendency to less numerical stability in numerical tests. Moreover, for the special case $\phi = 0$ no admissible $\gamma_{T,\sigma}$ exists,

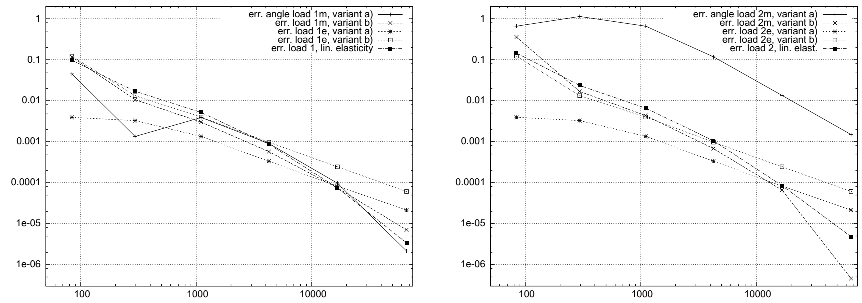


Fig. 10 Errors in crack angle with variants a) and b) of Subsect. 3.1 in Example 2 for piezoelectric material in comparison with the purely elastic case (with a reconstruction method corresponding to variant b)) under horizontal (left) and vertical (right) loads.

while $\gamma_{E,\sigma}$ can be chosen arbitrarily but is not automatically set to zero as is the case for γ_E.

In Example 1 this leads to a good convergence to the exact solution $C^* = 1$ in both variants. In Example 2, computations were carried out for all load cases, the results for each of the two variants a) and b) are shown in Fig. 11.

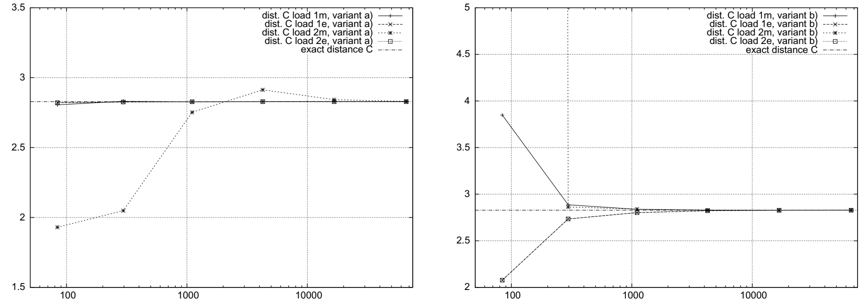


Fig. 11 Reconstructed crack constant C with variants a) (left) and b) (right) in Example 2.

Here again difficulties arise as expected in variant a) in load case 2m, where due to the small potential jump convergence to the exact constant $C = 2\sqrt{2}$ only takes effect for sufficiently fine meshes.

In variant b) with load case 2m we only observe an outlier in the 0th refinement step, due to an infeasible matrix R which leads to a division by zero in (32) after stabilization. After the first refinement step, computations yield good results for all load cases.

To compare the speed of convergence among the different variants and with the purely elastic case, we look at the deviation of the numerically computed crack plane constants to the exact values $C^* = 1$ in Example 1 and $C^* = 2\sqrt{2}$ in Example 2.

Fig. 12 shows the deviations depending on the number of nodes, P1 and P2 denoting the examples with piezo material and LE1, LE2 denoting the corresponding linear elastic examples.

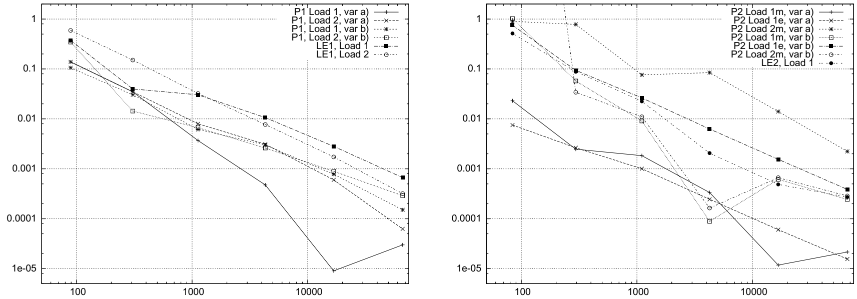


Fig. 12 Comparison of errors in reconstructed plane constant C for piezoelectric and for purely elastic material in Example 1 (left) and 2 (right).

There is basically no difference in the convergence speed. In Example 2 the convergence rate of variant a) is the same as in b) and linear elasticity, convergence just takes effect with a bit of delay in the load case 2m but starts from a slightly lower level in the other load cases.

Furthermore, a mid point approximation using variants a) and b) from Sect. 5 was computed in the second example (with single crack). Fig. 13 shows that the convergence behavior is comparable to the one for the crack plane parameters.

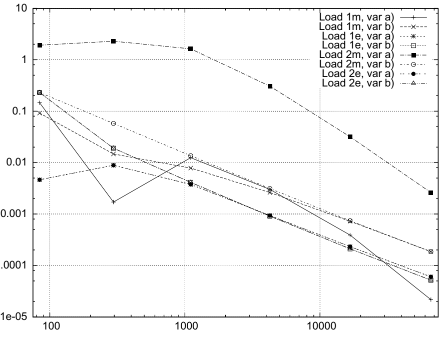


Fig. 13 Deviations of the approximated mid point coordinate from the true value in Example 2.

6.4 Comparison of Results with Different Refinement Strategies

For Example 2 all four refinement strategies were investigated with respect to their results in the inverse computations. The development of the deviations of the crack plane angle and the distance to the origin from the respective exact values are shown in Fig. 14 for the load case 1m.

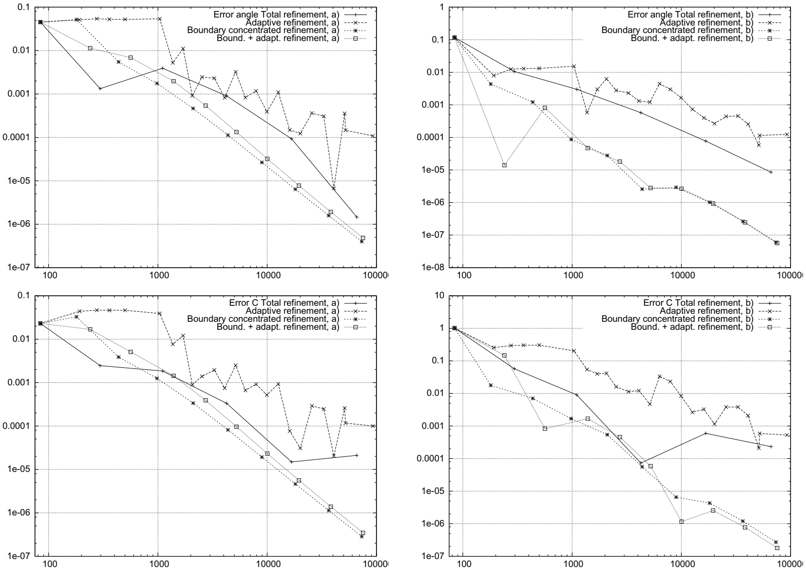


Fig. 14 Comparison of the crack plane angle (upper) and offset (lower) deviation in variant a) (left) and b) (right) for the load case 1m in Example 2.

When using boundary concentrated refinement with or without additional interior adaptivity, the errors are significantly lower with the same number of nodes in sufficient fine meshes. Especially in variant b) the difference is obvious from the figures. The worst results are observed in the case of conventional adaptively refined meshes, where the outer boundary is refined only sparsely. The reason for this effect is probably a concentration of refinement to the surroundings of the crack tips, where stress singularities occur. However, if the crack tips are sufficiently far away from the outer boundary, no pollution effect seems to be visible to the boundary integral, so that this refinement at the crack tips leads to unnecessary additional degrees of freedom.

Acknowledgements. Financial support by the German Science Foundation (DFG) within grant KA 1778/1 and SA 539/4-1 “Risse in Piezokeramiken” is gratefully acknowledged. We also thank the TU Chemnitz for providing the FEM software SPC-PM2Ad. Moreover, we are indebted to A.-M. Sändig, University of Stuttgart as well as A. Ben Abda, Ecole Nationale d'Ingénieurs de Tunis, for valuable suggestions and discussions, and finally the referee for fruitful comments.

References

[1] Andrieux, S., Ben Abda, A.: Identification of planar cracks by complete over determined data: inversion formulae. Inverse Problems 12, 553–563 (1996)

[2] Andrieux, S., Ben Abda, A., Bui, H.D.: Reciprocity principle and crack identification. Inverse Problems 15, 59–65 (1999)

[3] Bamberger, A., Glowinski, R., Tran, Q.H.: A domain decomposition method for the acoustic wave equation with discontinuous coefficients and grid change. SIAM J. Numer. Anal. 34, 603–639 (1997)

[4] Eibner, T.: Randkonzentrierte und adaptive hp-FEM. Dissertation, TU Chemnitz (2006)

[5] Friedman, A., Vogelius, M.: Determining cracks by boundary measurements. Indiana Math. J. 8, 527–556 (1989)

[6] Hao, T.H., Shen, Z.Y.: A new electric boundary condition of electric fracture mechanics and its applications. Engrg. Fracture Mech. 47, 793–802 (1994)

[7] Kemmer, G., Balke, H.: Krafteinwirkungen auf die Flanken nichtleitender Risse in Piezoelektrika. ZAMM 79(S2), 509–510 (1999)

[8] Khoromskij, B.N., Melenk, J.M.: Boundary concentrated FEM. SIAM J. Numer. Anal. 41, 1–36 (2003)

[9] Knees, D.: Regularity results for transmission problems for the Laplace and Lamé operators on polygonal or polyhedral domains. SFB 404, Bericht 2002/10, Universität Stuttgart (2002)

[10] Korneev, V.G., Langer, U.: Domain Decomposition and Preconditioning. In: Stein, E., de Borst, R., Hughes, T.J.R. (eds.) Encyclopedia of Computational Mechanics, Part I, ch.19. John Wiley & Sons (2004)

[11] Kunert, G., Verfürth, R.: Edge residuals dominate a posteriori error estimates for linear finite element methods on anisotropic triangular and tetrahedral meshes. Numer. Math. 86, 283–303 (2000)

[12] Lenk, A.: Elektromechanische Systeme, Band 2 (Systeme mit verteilten Parametern). VEB Verlag Technik, Berlin (1974)

[13] Meyer, A.: Programmer's Manual for Adaptive Finite Element Code SPC-PM 2Ad. Preprintreihe des SFB 393, Preprint 01-18, TU Chemnitz (2001)

[14] Meyer, A., Steinhorst, P.: Modellierung und Numerik wachsender Risse bei piezoelektrischem Material. Preprint CSC/10-01, TU Chemnitz (2010)

[15] Park, S.B., Sun, C.T.: Effect of electric field on fracture of piezoelectric ceramics. Internat. J. Fract. 70, 203–216 (1995)

[16] Scherzer, M., Kuna, M.: Combined analytical and numerical solution of 2D interface corner configurations between dissimilar piezoelectric materials. Internat. J. Fract. 127, 61–99 (2004)

[17] Steinhorst, P.: Anwendung adaptiver FEM für piezoelektrische und spezielle mechanische Probleme. Dissertation, TU Chemnitz (2009)

[18] Steinhorst, P., Sändig, A.–M.: Reciprocity principle for the detection of planar cracks in anisotropic elastic material. Preprint IANS-2010/011, Universität Stuttgart (2010)

[19] Valean, A.: Identifikation von planaren Rissen in der Piezoelektrizität. Diplomarbeit, Universität Stuttgart (2010)

[20] Verfürth, R.: A review of a posteriori error estimation and adaptive mesh-refinement techniques. Wiley-Teubner, Chichester (1996)

PROCRACK: A Software for Simulating Three-Dimensional Fatigue Crack Growth

Frank Rabold, Meinhard Kuna, and Thomas Leibelt

Abstract. In this paper, a finite element software for automated simulation of fatigue crack growth in arbitrarily loaded three-dimensional components is presented. The criterion, direction and amount of crack propagation are controlled by concepts of linear elastic fracture mechanics. The fracture mechanical parameters are calculated by means of a special submodelling technique in combination with the interaction integral or the virtual crack closure technique. In the adaptive crack growth step, the updated crack front position is determined and the mesh in the crack region is automatically adapted. The preprocessing and main FEM-analysis of the cracked structure are done using the commercial software ABAQUS. Two application examples show the capability and performance of the simulation program.

1 Introduction

Fatigue crack growth describes a slowly progressing failure process in a material or structural component as a result of cyclic thermal or mechanical loading. Starting from a defect which already exists or is initiated during service loading, a crack propagates in a subcritical manner until a critical crack size is attained, leading to complete unstable fracture. Therefore, the lifetime and reliability of components under alternating loads are crucially controlled by fatigue crack growth. About sixty percent of all structural failures in engineering practice are caused by fatigue crack growth! Predominantly aircrafts, automotives, wind turbines or railways are affected as the recent problem with axles of ICE train has shown. To ensure safety and durability, the prediction of fatigue crack growth is an important problem, which has to be coped by modern computational techniques.

Frank Rabold · Meinhard Kuna · Thomas Leibelt
Institut für Mechanik und Fluiddynamik, TU Bergakademie Freiberg,
Lampadiusstraße 4, 09596 Freiberg, Germany
e-mail:{Frank.Rabold,Meinhard.Kuna,
Thomas.Leibelt}@imfd.tu-freiberg.de

T. Apel & O. Steinbach (Eds.): Advanced Finite Element Methods and Appl., LNACM 66, pp. 355–374.
springerlink.com

Due to the three-dimensional geometry and complex loading conditions of cracked components, the finite element method (FEM) is the preferred numerical tool to solve the initial boundary value problems of fracture mechanics. The special algorithmic challenges for crack analysis are: (i) to capture the singularities at the crack tip, (ii) to adapt the FEM-discretization to the changing crack size.

While simplified two-dimensional analysis tools are well established (see e.g. [12, 15]), the adaptive treatment of three-dimensional crack growth along curved surfaces is still under current research. The existing approaches can be divided into three classes:

- Extended Finite Element Method (X-FEM) using enriched shape functions for cracks in combination with marching methods [21] or level sets [8, 1] to describe the moving crack fronts. The experience with X-FEM shows that the technique is well suited to simulate the propagating crack discontinuity through a standard FE-mesh, which is adequate to account for loss of stiffness. However to compute the stress intensity factors with the accuracy needed for fatigue crack analysis, the employed enrichment functions are not sufficient.
- Remeshing of a sub-domain around the growing crack whereby only automatic tetraeder element meshing algorithms can be applied [23, 24].
- Replacing the domain around the growing crack by a submodel technique [19], which enables much better accuracy in calculation of fracture parameters.

In the present approach a pragmatic engineering approach of the last type is realized, which uses the submodel technique for accurate crack analysis in combination with an adaptive meshing scheme.

2 Fracture-Mechanical Fundamentals

The three-dimensional fatigue crack growth can be treated in the framework of the classical linear elastic fracture mechanics. The strength of the singular stress fields along the crack front is measured by the three stress intensity factors K_{I}, K_{II} and K_{III}. In doing so, the K-factors are the basic parameters for the analysis of fatigue crack propagation and have to be computed numerically.

2.1 Fatigue Crack Growth

Fig. 1 illustrates the general behavior of fatigue crack growth in dependence of the stress intensity factor range, depending on maximum and minimum load levels

$$\Delta K = K_{\mathrm{max}} - K_{\mathrm{min}}\,. \tag{1}$$

ΔK is equal to ΔK_{I} considering Mode-I crack growth. For crack growth under mixed mode conditions a cyclic *equivalent* stress intensity factor K_{v} is assumed [17]

$$\Delta K = \Delta K_{\mathrm{v}} = \frac{1}{2}\Delta K_{\mathrm{I}} + \frac{1}{2}\sqrt{(\Delta K_{\mathrm{I}})^2 + 5.336 \cdot (\Delta K_{\mathrm{II}})^2 + 4 \cdot (\Delta K_{\mathrm{III}})^2}\,. \tag{2}$$

The crack growth rate is denoted by $\mathrm{d}a/\mathrm{d}N$ as the extension Δa of a crack with the length a occurring in one load cycle. In Fig. 1 the crack growth is subdivided into the regions I, II and III. Stable crack growth is observed in the interval

$$\Delta K_{\mathrm{th}} \le \Delta K < \Delta K_{\mathrm{crit}}\,. \tag{3}$$

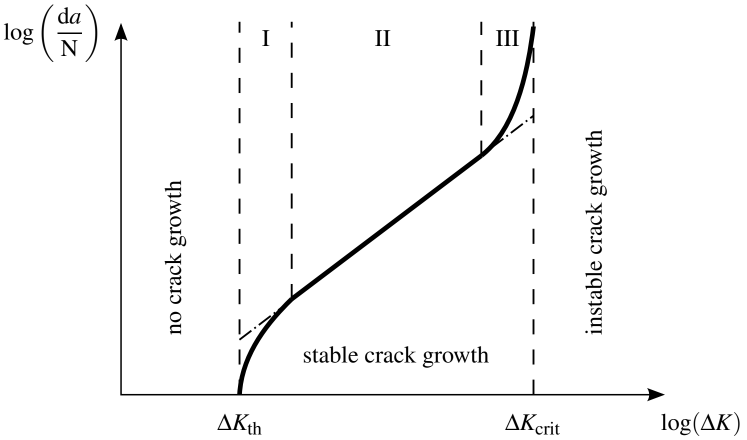


Fig. 1 Macro-crack growth under cyclic loading.

Below the threshold value ΔK_{th} the crack does not propagate at all. This region corresponds to the fatigue resistance of a component with crack. For high values of ΔK the transition to an unstable crack propagation occurs in region III after attaining a critical value ΔK_{crit} with the relation

$$\Delta K_{\mathrm{crit}} = (1 - R_{\mathrm{K}})K_{\mathrm{crit}}\,. \tag{4}$$

Here R_{K} is the loading ratio of the stress intensity factor with

$$R_{\mathrm{K}} = \frac{K_{\mathrm{min}}}{K_{\mathrm{max}}}\,. \tag{5}$$

In general, crack growth under cyclic loading can be described by equations of the type

$$\frac{\mathrm{d}a}{\mathrm{d}N} = g(\Delta K, R_{\mathrm{K}})\,. \tag{6}$$

Several crack growth laws can be found in the literature (e.g. [5]).

For a constant stress intensity ratio R_{K}, the middle part II of the curve can be approximated by a straight line with the so-called PARIS law [14]

$$\frac{\mathrm{d}a}{\mathrm{d}N} = C\,(\Delta K)^m\,. \tag{7}$$

The constants C and m depend on the material, the stress intensity ratio and the environmental ambient conditions. Extending equation (7) to region I and III results in an over- or underestimation of the real crack propagation.

Modern crack growth laws are able to describe all regions I, II and III. The most accepted approach is the so-called NASGRO equation [6, 13]

$$\frac{\mathrm{d}a}{\mathrm{d}N} = C\left[\left(\frac{1-f}{1-R_\mathrm{K}}\right)\Delta K\right]^m \frac{\left(1-\dfrac{\Delta K_\mathrm{th}}{\Delta K}\right)^p}{\left(1-\dfrac{K_\mathrm{max}}{K_\mathrm{crit}}\right)^q} \tag{8}$$

with

$$K_\mathrm{max} = \frac{\Delta K}{1-R_\mathrm{K}}\,. \tag{9}$$

The crack opening function is defined as

$$f = \begin{cases} \max\left(R_\mathrm{K}; A_0 + A_1 R_\mathrm{K} + A_2 R_\mathrm{K}^2 + A_3 R_\mathrm{K}^3\right) & : \quad R_\mathrm{K} \geq 0, \\ A_0 + A_1 R_\mathrm{K} & : \quad -2 \leq R_\mathrm{K} < 0, \\ A_0 - 2A_1 & : \quad R_\mathrm{K} < -2 \end{cases} \tag{10}$$

with

$$\begin{aligned} A_0 &= \left(0.825 - 0.34\alpha + 0.05\alpha^2\right)\left[\cos\left(\frac{\pi}{2}\frac{\sigma_\mathrm{max}}{\sigma_\mathrm{F}}\right)\right]^{\frac{1}{\alpha}}, \\ A_1 &= (0.415 - 0.071\alpha)\frac{\sigma_\mathrm{max}}{\sigma_\mathrm{F}}, \\ A_2 &= 1 - A_0 - A_1 - A_3, \\ A_3 &= 2A_0 + A_1 - 1 \end{aligned} \tag{11}$$

taking into account the dependence of the crack growth curve on the stress intensity ratio R_K. The threshold value ΔK_th is derived from the equation

$$\Delta K_\mathrm{th} = \Delta K_0 \frac{\sqrt{\dfrac{a}{a + a_{0,\mathrm{Eigen}}}}}{\left[\dfrac{1-f}{(1-A_0)(1-R_\mathrm{K})}\right]^{(1+C_\mathrm{th} R_\mathrm{K})}} \tag{12}$$

with

$$C_\mathrm{th} = \begin{cases} C_\mathrm{th}^+ & : R_\mathrm{K} \geq 0, \\ C_\mathrm{th}^- & : R_\mathrm{K} < 0, \end{cases} \tag{13}$$

also accounting for its dependency on R_{K}. The value ΔK_0 comprises the threshold value for $R_{\mathrm{K}} = 0$. For adaptation to the material and the environmental conditions there is a corresponding set of parameters: K_{crit}, C, m, p, q, ΔK_0, C^{+}_{th}, C^{-}_{th}, α, $\sigma_{\max}/\sigma_{\mathrm{F}}$, $a_{0,\mathrm{Eigen}}$.

2.2 Calculation of Stress Intensity Factors

The key point in the simulation of crack propagation is the accurate determination of the stress intensity factors along the crack front. Two methods for calculating the K-factors are presented below.

2.2.1 Interaction Integral

A common numerical technique for analyzing cracks is the J-integral method [3, 16]. But it can not be used directly in mixed mode crack problems to obtain the stress intensity factors in separate form. Based on an extension of the J-integral, the interaction integral method [20, 22, 7, 11] is an indirect way to extract the individual stress intensity factors.

The relationship between the J-integral and the stress intensity factors in three dimensions for homogeneous isotropic materials can be written as

$$J = \frac{(1-\nu^2)}{E}\left(K_{\mathrm{I}}^2 + K_{\mathrm{II}}^2\right) + \frac{1}{2G}K_{\mathrm{III}}^2 = \frac{1}{2}\mathbf{K}^{\top}\cdot\mathbf{Y}\cdot\mathbf{K}$$

with the Young's modulus E, the shear modulus G and the vector of the K-factors $\mathbf{K} = [K_{\mathrm{I}}, K_{\mathrm{II}}, K_{\mathrm{III}}]^{\top}$. The matrix $\mathbf{Y}$ depends on material properties and is called the IRWIN matrix, correlated to the energy release during crack extension.

The relationship between the vector of the stress intensity factors and the interaction integral vector $\mathbf{J}^{\mathrm{int}}$ is given as

$$\mathbf{K} = \mathbf{Y}^{-1}\cdot\mathbf{J}^{\mathrm{int}}, \tag{14}$$

where $\mathbf{J}^{\mathrm{int}} = \left[J_{\mathrm{I}}^{\mathrm{int}}, J_{\mathrm{II}}^{\mathrm{int}}, J_{\mathrm{III}}^{\mathrm{int}}\right]^{\top}$. By superimposing an auxiliary field to the actual field, the components of the interaction integral $\mathbf{J}^{\mathrm{int}}$ are defined by

$$J_{\alpha}^{\mathrm{int}} = \lim_{\Gamma\to 0}\int_{\Gamma}\mathbf{n}\cdot\left[\boldsymbol{\sigma}:\boldsymbol{\varepsilon}_{\alpha}^{\mathrm{aux}}\mathbf{I} - \boldsymbol{\sigma}\cdot\left(\frac{\partial\mathbf{u}_{\alpha}^{\mathrm{aux}}}{\partial\mathbf{x}}\right) - \boldsymbol{\sigma}_{\alpha}^{\mathrm{aux}}\cdot\left(\frac{\partial\mathbf{u}}{\partial\mathbf{x}}\right)\right]\cdot\mathbf{q}\,\mathrm{d}\Gamma. \tag{15}$$

The displacements $\mathbf{u}$, the strains $\boldsymbol{\varepsilon}$ and the stresses $\boldsymbol{\sigma}$ represent the field according to the actual loading. With $\alpha = \mathrm{I}, \mathrm{II}, \mathrm{III}$ an auxiliary field is assumed representing the known near tip solution for the pure Mode I, Mode II or Mode III at the crack front. The auxiliary field is described by the corresponding displacements $\mathbf{u}_{\alpha}^{\mathrm{aux}}$, strains

$\boldsymbol{\varepsilon}_{\alpha}^{\mathrm{aux}}$ and stresses $\boldsymbol{\sigma}_{\alpha}^{\mathrm{aux}}$. The unit vector $\mathbf{q}$ specifies the direction of the virtual crack propagation parallel to the x_1 coordinate.

The contour Γ lies in the x_1-x_2 plane surrounding the crack tip perpendicular to the crack front at the position s. As shown in Fig. 2, the contour begins on the bottom crack surface and ends on the top surface. The vector $\mathbf{n}$ is the outward normal vector to Γ in the x_1-x_2 plane.

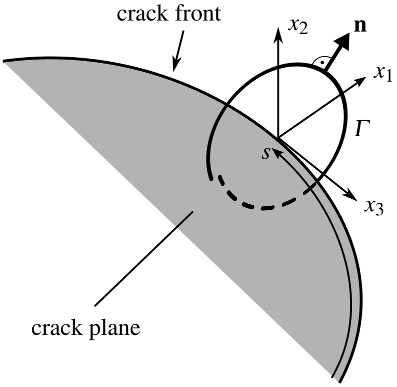


Fig. 2 Definition of local coordinates and the contour Γ at the point s on the crack front.

Thus, once $\mathbf{J}^{\mathrm{int}}$ is obtained, the vector of the stress intensity factors $\mathbf{K}$ can be easily calculated from equation (14).

2.2.2 Modified Crack Closure Integral

The technique of the modified crack closure integral is a simple but efficient method for calculating the energy release rate for certain points along the crack front [18, 2]. Using the finite element method only node displacements on the crack faces and nodal forces at the ligament in front of the crack are needed. Fig. 3 shows the relevant part of the finite element mesh with second-order elements at the crack front. With the equations [11]

$$G_{\mathrm{I}}(s_k) = \frac{1}{2\Delta A}\left[F_2^1\Delta u_2^1 + F_2^2\Delta u_2^2 + \frac{1}{2}F_2^3\Delta u_2^3 + \frac{1}{2}F_2^3\Delta u_2^3\right], \tag{16}$$

$$G_{\mathrm{II}}(s_k) = \frac{1}{2\Delta A}\left[F_1^1\Delta u_1^1 + F_1^2\Delta u_1^2 + \frac{1}{2}F_1^3\Delta u_1^3 + \frac{1}{2}F_1^3\Delta u_1^3\right], \tag{17}$$

$$G_{\mathrm{III}}(s_k) = \frac{1}{2\Delta A}\left[F_3^1\Delta u_3^1 + F_3^2\Delta u_3^2 + \frac{1}{2}F_3^3\Delta u_3^3 + \frac{1}{2}F_3^3\Delta u_3^3\right] \tag{18}$$

and the sum of the considered element areas

$$\Delta A = \frac{1}{2}(b_{k-1} + b_k)\Delta a \tag{19}$$

the energy release rates are easily computed at the nodal coordinate s_k. The required stress intensity factors are calculated by using the relationship between the K-factors and the energy release rate for the separate crack opening modes I, II or III:

$$K_{\mathrm{I}}^2 = \frac{E}{1-\nu^2} G_{\mathrm{I}}, \quad K_{\mathrm{II}}^2 = \frac{E}{1-\nu^2} G_{\mathrm{II}}, \quad K_{\mathrm{III}}^2 = \frac{E}{1+\nu} G_{\mathrm{III}}. \tag{20}$$

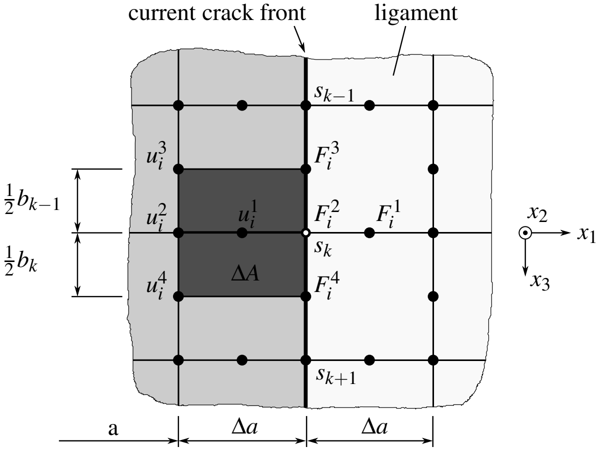


Fig. 3 Modified crack closure integral for three-dimensional second-order elements.

The modified crack closure integral is also applicable for curved crack fronts, see [11]. The equations (16)-(18) provide a good accuracy for the case that the element edges are always perpendicular to the current crack front and the elements prior and behind the crack front have nearly the same size.

2.3 Determination of Crack Deflection

As shown in Fig. 4, a local coordinate systems is attached to the moving crack front point. The x_2 coordinate axis must be perpendicular to the crack plane and the x_3 coordinate axis is the tangent to the crack front. At each point along the crack front the crack always grows in radial direction. The deflection of the crack with respect to its current plane at this point is described using a single crack deflection angle

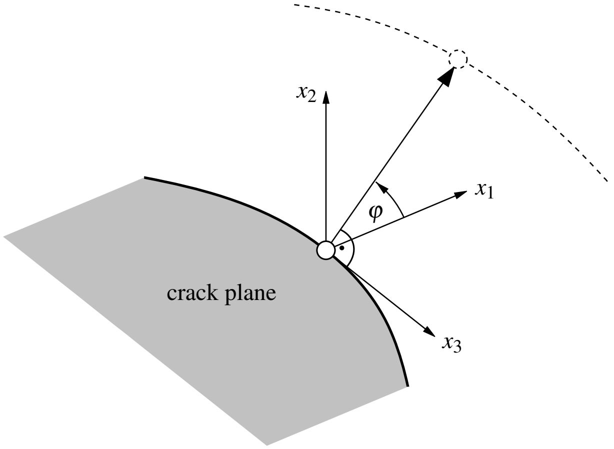


Fig. 4 Definition of the crack deflection angle.

φ. The crack growth direction is assumed to lie always normal to the current crack front tangent.

According to the maximum circumferential stress criterion the crack will propagate into that direction oriented perpendicular to the maximum stress in the local x_1-x_2 plane [4]. Considering load cases with $R \geq 0$ the absolute value of the crack deflection angle φ depends only on the ratio of the ranges of the stress intensity factors $\Delta K_{\mathrm{I}}/\Delta K_{\mathrm{II}}$ and is defined through

$$\varphi = 2\arctan\left[\frac{\operatorname{sign}(1-R_{\mathrm{II}})}{4}\left(\frac{\Delta K_{\mathrm{I}}}{\Delta K_{\mathrm{II}}} - \sqrt{\left(\frac{\Delta K_{\mathrm{I}}}{\Delta K_{\mathrm{II}}}\right)^2 + 8}\right)\right]. \tag{21}$$

Here R_{II} is the loading ratio of the stress intensity factor K_{II}.

2.4 Determination of Crack Growth Increments

If the threshold value ΔK_{th} is exceeded in a load cycle the crack propagates stepwise according to the crack growth law in equation (6). Thus, after every load cycle the position of the crack front has to be updated and a fracture-mechanical analysis at the new crack front is needed. This cycle-by-cycle procedure is computationally very expensive and time consuming, particularly for small values of the crack growth rate.

It is assumed that a minimal change in the crack front geometry induces only a negligible change in the range of the stress intensity factors. To reduce the resulting numerical effort, the crack growth simulation needs not to be repeated for every load cycle. Instead, the crack growth law is integrated for a range of cycles assuming a fixed crack front position. Afterward, the geometry of the crack front is updated, taking into account the internally accumulated crack growth. This crack propagation for the certain number of loading cycles N_{Inc} is then assigned as numerical crack growth increment.

Two control parameters Δa_{min} and Δa_{max} have to be specified for the crack growth increment. Δa_{min} describes the minimal crack growth length per crack growth increment, which is to be achieved along the whole crack front. Contrary, the crack growth must not exceed the maximum value Δa_{max} at any point at the crack front. Otherwise the value for Δa_{min} has to be reduced. The task is now to determine the number of load cycles that meet the above conditions in consideration of the crack growth law and the loading conditions for each crack front point.

First, for each point $i = 1 \ldots n$ at the crack front, the number of load cycles is computed which are required to attain the crack growth length $\Delta a = \Delta a_{\text{min}}$:

$$\Delta N_{\text{i}}^{\text{min}} = \frac{\Delta a_{\text{min}}}{g(\Delta K, R_{\text{K}})_{\text{i}}} \,. \tag{22}$$

Detecting the maximum number

$$\Delta N_{\text{max}}(\Delta a_{\text{min}}) = \max(\Delta N_{\text{i}}^{\text{min}}) \tag{23}$$

gives the number of load cycles which are necessary to get the minimum crack growth along the crack front. The number of load cycles for the crack growth length Δa_{max} for each crack front point and the resulting minimum number are computed in the same way:

$$\Delta N_{\text{i}}^{\text{max}} = \frac{\Delta a_{\text{max}}}{g(\Delta K, R_{\text{K}})_{\text{i}}} \,, \quad \Delta N_{\text{min}}(\Delta a_{\text{max}}) = \min(\Delta N_{\text{i}}^{\text{max}}) \,. \tag{24}$$

The required number of load cycles of the crack growth increment N_{Inc} is obtained by case distinction:

$$N_{\text{Inc}} = \begin{cases} \Delta N_{\text{max}}(\Delta a_{\text{min}}) : & \Delta N_{\text{max}}(\Delta a_{\text{min}}) \leq \Delta N_{\text{min}}(\Delta a_{\text{max}}), \\ \Delta N_{\text{min}}(\Delta a_{\text{max}}) : & \Delta N_{\text{max}}(\Delta a_{\text{min}}) > \Delta N_{\text{min}}(\Delta a_{\text{max}}). \end{cases} \tag{25}$$

Finally, the crack growth length per crack growth increment for each point on the crack front can be calculated:

$$\Delta a_{\text{i}} = g(\Delta K, R_{\text{K}})_{\text{i}} \, N_{\text{Inc}} \,. \tag{26}$$

Fig. 5 shows the crack front of a Mode-I crack after three crack growth increments. The incremental crack growth lengths for the five exemplary points are calculated for the load cycles N_1, N_2 and N_3 according to the above scheme.

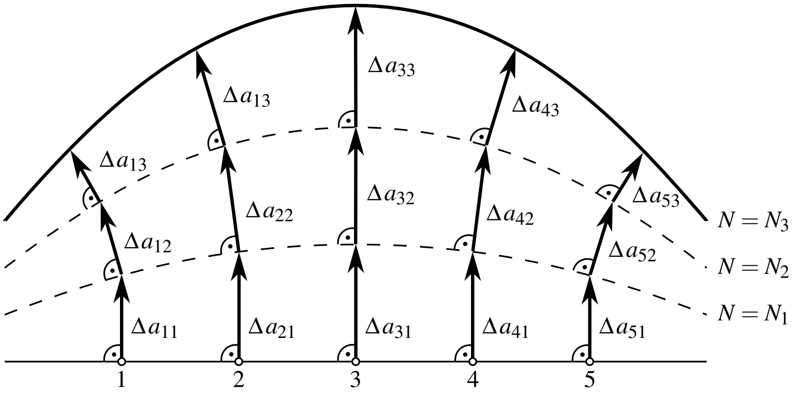


Fig. 5 Crack growth increments for propagating a Mode-I crack.

3 The Simulation Software PROCRACK

The simulation software PROCRACK is a collection of program modules for the automated finite element simulation of fatigue crack growth. The basis of the simulation program is ABAQUS, a software system for finite element analysis and computer-aided engineering. ABAQUS uses the open-source script language PYTHON for scripting and customization. Therefore, several modules of PROCRACK are developed in the computer language PYTHON. This allows an easy and comfortable development and maintenance of the simulation program.

The program modules control the pre- and postprocessing of the simulation. For practical reasons, the computational model is based on the ABAQUS/CAE tool. Thus, existing models with an arbitrary geometry can be loaded and analyzed with respect to fatigue crack propagation. The finite element analysis is carried out with ABAQUS/standard.

3.1 Program Layout

The basic concept of the crack growth simulation is the execution of the following three steps:

1. FE-Analysis of the cracked component under loading,
2. Calculation of the fracture-mechanical parameters,
3. Updating the crack geometry in the cracked component.

These steps are executed repeatedly until the crack growth stops or the crack has grown through the component. The crack growth simulation program PROCRACK is based on this concept.

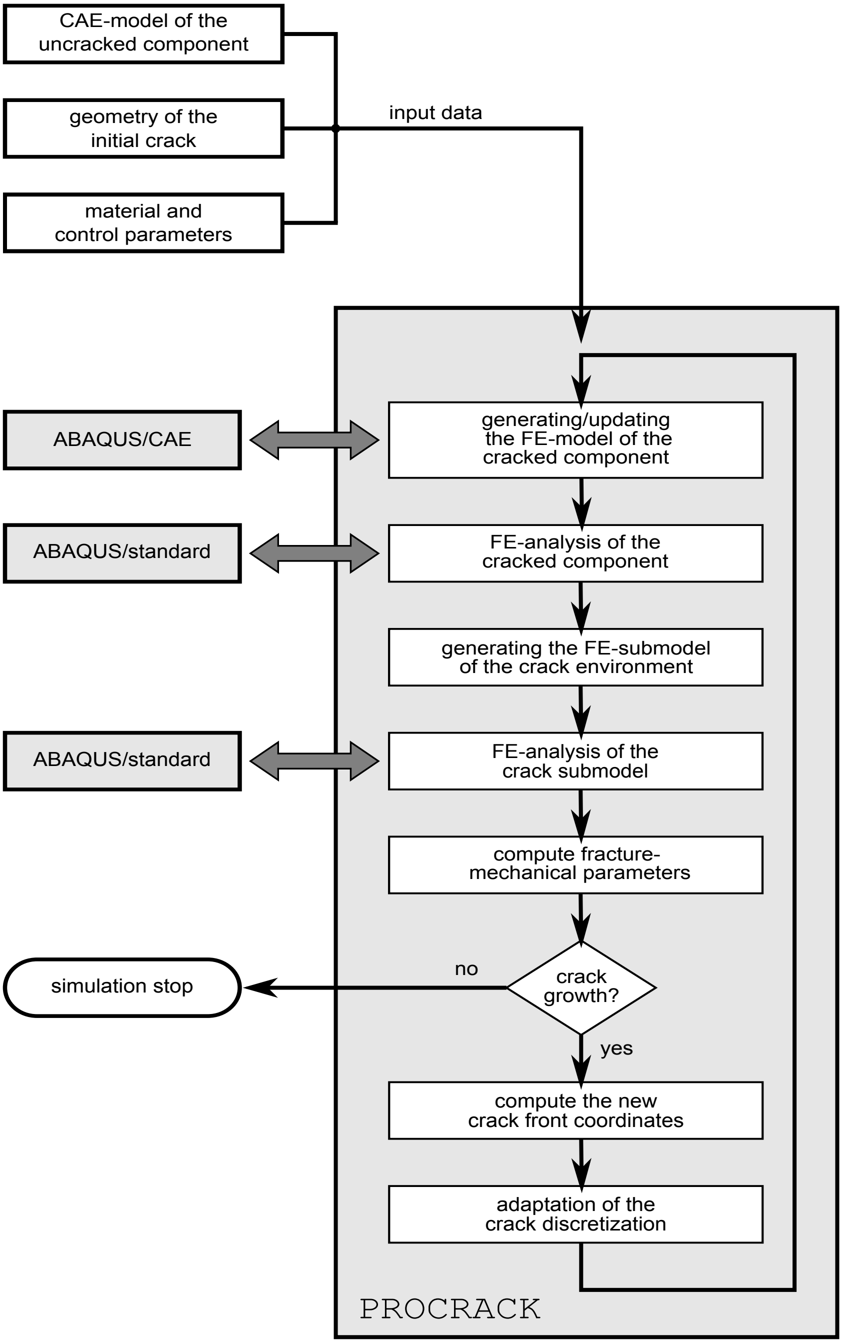

CAE-model of the uncracked component
geometry of the initial crack
material and control parameters
input data
ABAQUS/CAE
ABAQUS/standard
ABAQUS/standard
generating/updating the FE-model of the cracked component
FE-analysis of the cracked component
generating the FE-submodel of the crack environment
FE-analysis of the crack submodel
compute fracture-mechanical parameters
crack growth?
no
yes
simulation stop
compute the new crack front coordinates
adaptation of the crack discretization
PROCRACK

Fig. 6 Flowchart of the crack growth simulation program.

Fig. 6 shows the flowchart of PROCRACK. The computational model of the uncracked component, the coordinates of the initial crack, the parameter for controlling the crack propagation and the material parameters are used as input data. The main routine contains the incremental crack propagation loop where the following steps are executed.

1. Crack generation, meshing with finite elements and writing the input deck using the external program ABAQUS/CAE.
2. FE analysis of the cracked component using ABAQUS.
3. Generating the input deck of the submodel of the crack environment.
4. FE analysis of the submodel using ABAQUS.
5. Calculation of the stress intensity factor range along the crack front.
6. Computation of the crack deflection and the incremental crack growth length at the crack front.
7. Calculation of the new crack front coordinates.
8. Adaptation of the crack discretization.
9. Go to the first step.

Two methods for calculating the stress intensity factors can be selected: either the modified crack closure integral inside of PROCRACK or the ABAQUS-own implementation of the interaction integral. Concerning the crack growth law it can be chosen between the PARIS law or the NASGRO equation, see Sect. 2.1.

If no crack growth occurs and the load remains constant, the simulation of crack propagation is terminated. If the crack has grown through the component, it is assumed that the load bearing capacity is exceeded or the functionality of the component is no longer available. In this case the simulation is also terminated.

The computational model of the component with the current crack geometry is saved after the ABAQUS/CAE preprocessing for using in a subsequent step. The results of each crack propagation loop are stored separately, so a restart is possible at any time.

3.2 Procedure of Crack Generation

The modeling of the initial crack and the incremental crack extension is carried out exclusively in the ABAQUS/CAE tool. The crack is described by its crack area at the geometric level. For the sake of simplicity, the crack front is discretized by a certain number of straight line segments. The segments are connected by geometric points which are used as reference points for the numerical computation of the fracture-mechanical parameters. The crack area is simply approximated by triangles.

A coarse but illustrative discretization of a surface crack after two crack propagation increments is shown in Figure 7. For each incremental extension of the crack area a single layer of triangular elements is added. Due to the special arrangement of the triangles between the old and new crack front it is easily possible to form a three-dimensional curved crack path.

The generation of the crack in the finite element model is done by a special function in ABAQUS/CAE. This function cuts the model at the crack area and generates the crack front as well as the opposite crack faces at the finite element level. The refinement of the mesh is partially controlled by a set of user defined parameters.

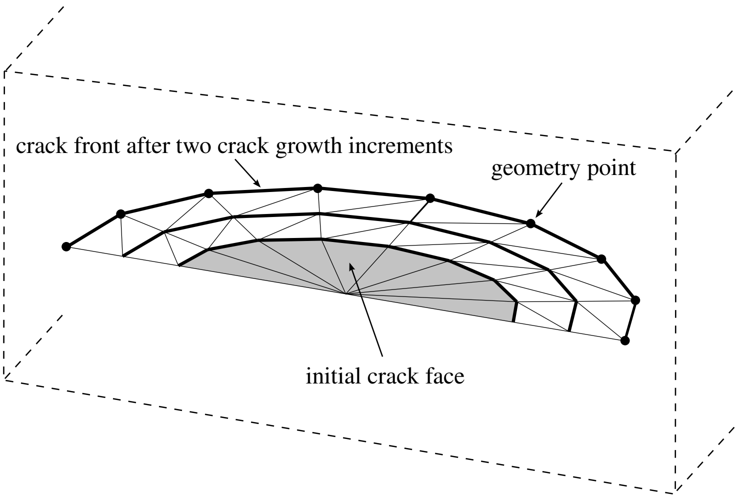


Fig. 7 Geometric discretization of a three-dimensional surface crack.

3.3 Submodel Technique

PROCRACK is using the submodel technique to calculate the fracture-mechanical parameters along the crack front. The cylindrical domain surrounding the current crack front of the component is created as an external part and separately analyzed with the finite element method. The boundary conditions are the corresponding node displacements at the boundary of the cylindrical crack domain. A three-dimensional element mesh of the submodel is automatically created by a subroutine, which is shown in Fig. 8 for an exemplary surface crack.

One benefit of the submodel technique is that the meshing of the global component can be designed independently of the numerical evaluation at the crack front. Furthermore, the finite element mesh of the submodel is appropriately adjusted to the numerical method for determining the stress intensity factors.

In the application of numerical methods, the values of the stress intensity factors at the crack front at the component surface are generally inaccurate for surface cracks or part-through cracks. In addition, the automated modeling of the cylindrical ending of the submodel at a curvilinear components surface is very difficult and leads to a bad numerical solution of the submodel. Therefore, the submodel is not extended to the surface of the component, but ends at a distance s_t ahead of the surface. The values of the stress intensity factors at the components surface will be extrapolated from the adjacent reference points at the crack front.

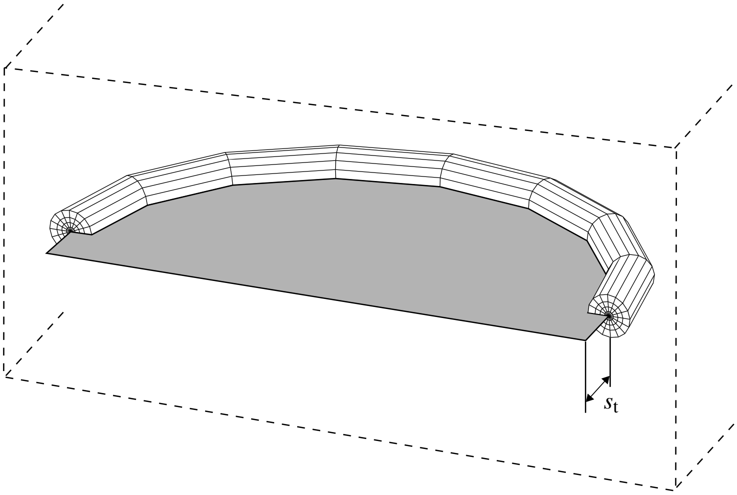


Fig. 8 Submodel for a surface crack.

3.4 *Adaptation of the Crack Discretization*

PROCRACK contains an algorithm for adaptive control of the discretization of the crack front and the crack area. Due to the curved crack propagation, it is possible that the distance l between two geometric points at the crack front is too large and the discretization of the crack front is no longer appropriate. If the distance exceeds the user defined parameter $l_{\max}$, a new geometric point is created in the middle. Otherwise, if the distance falls below the user parameter $l_{\min}$, the geometric point with the shortest distance to its next neighbor point will be removed. The refining and coarsening of the crack front is illustrated in Fig. 9.

The computational effort is increasing with the growing crack. It may be useful to coarsen the fine discretization of the crack area in regions far behind the crack front. Fig. 10 shows the scheme for partial coarsening of the crack area. The procedure is controlled by the parameters for the size of the coarse crack area and the distance to the crack front. With this approach, the time required for the preprocessing and analysis in subsequent crack growth increments can be reduced significantly.

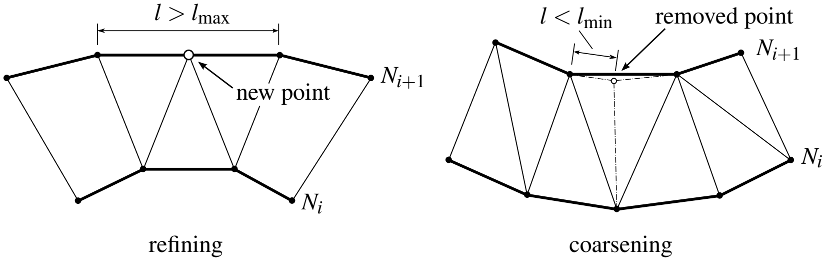


Fig. 9 Adaptation of the crack front.

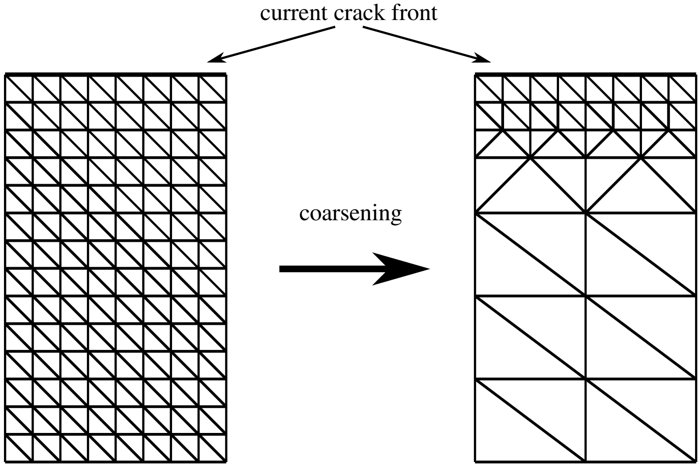


Fig. 10 Adaptation of the crack area.

4 Application Examples

By presenting two application examples, the functionality of the simulation program PROCRACK will be demonstrated below. In both computational models first-order tetrahedral elements are used for the finite element mesh. The steel S355J2G3 is selected as the material. Within joint projects the fracture-mechanical parameters for the NASGRO equation were determined by IWT Freiberg [9, 10]. The parameters for the PARIS law were taken from [5].

4.1 Crack Growth in a Three-Point Bending Specimen

In the first example the fatigue crack growth under cyclic loading ($R = 0.1$) in a three-point bending specimen is investigated. Fig. 11 shows a schematic drawing of the test with boundary conditions and loading. On the opposite side of the loading the specimen contains a initial straight surface crack, which is obliqued by an angle of 60 degrees to the longitudinal edge of the specimen. The calculation of the crack propagation is based on the PARIS law. The initial finite element mesh with 9768 tetrahedral elements is shown in Fig. 12. The part of the specimen, in which the crack will grow, has a finer meshing than the other parts.

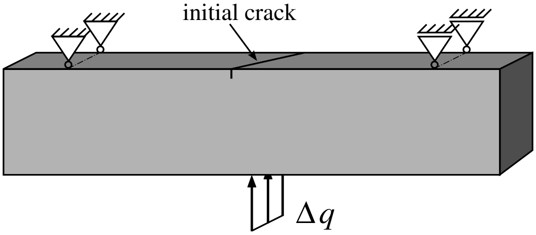


Fig. 11 Schematic of the three-point bending test.

The crack growth simulation[1] is canceled after 42 crack growth increments. At this time the maximum range of the stress intensity factor at the crack front reaches the critical value ΔK_{crit} and the further crack growth is unstable. Fig. 13 illustrates the development of the final crack area and Fig. 14 shows the finite element meshing of the crack at the surface of the specimen. As a result of the initial asymmetric location the crack is growing out of the initial crack plane. Due to the mixed mode at the crack front the crack rotates in a pure Mode-I orientation with increasing length.

4.2 Crack Growth in a Steering Knuckle

The steering knuckle is a part of the chassis of a motor vehicle. It houses the wheel axle and attaches the steering and suspension components to the wheel support assembly. In the following, the fatigue crack growth in the wheel axle under a simple operating load is analyzed. Fig. 15 illustrates the loading ($R = 0$) and boundary conditions at the steering knuckle as well as the location of the initial semi-elliptical surface crack. Due to the symmetry of the component, the three-dimensional computational model is carried out as a halfmodel. The initial finite element mesh with

[1] Computing time: 348 min, CPU: 1x AMD Opteron 8378 (2.4 GHz).

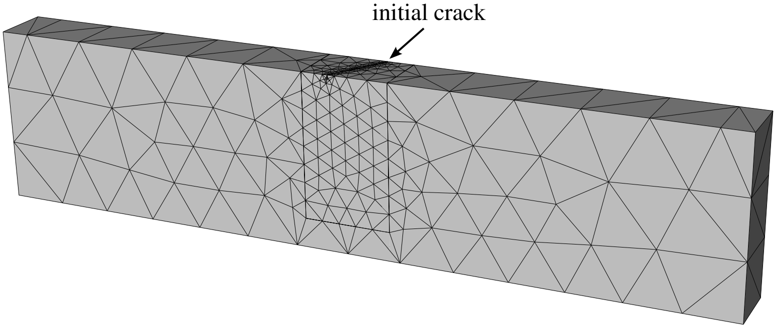


Fig. 12 Initial meshing with finite elements.

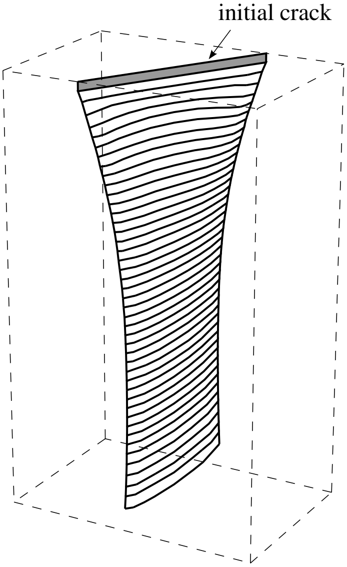


Fig. 13 Final crack area.

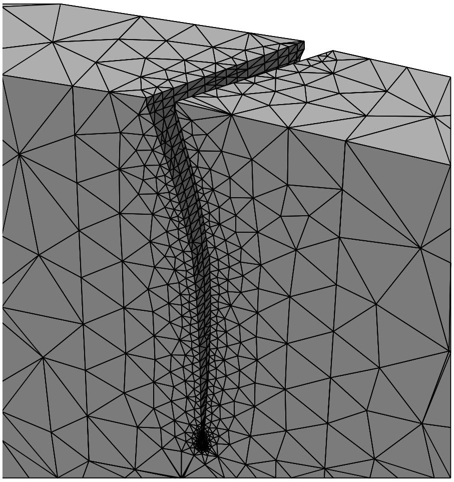


Fig. 14 Detail of the crack.

133137 tetrahedral elements is shown in Fig. 16. The calculation of the fatigue crack propagation is based on the NASGRO equation.

The numerical simulation[2] of the fatigue crack growth in the steering knuckle is stopped after 50 crack growth increments ($4.6 \cdot 10^6$ load cycles). At this point, the maximum range of the stress intensity factor along the crack front reaches the critical value for unstable crack growth. Fig. 17 and 18 show the final crack at the symmetry plane of the computational model and at the surface of the wheel axle. The computed crack area at the base of the wheel axle is illustrated in Fig. 19.

[2] Computing time: 462 min, CPU: 2x AMD Opteron 8378 (2.4 GHz).

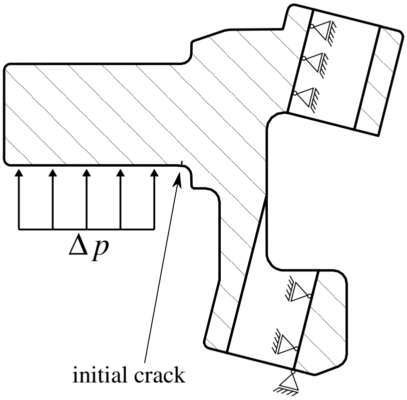


Fig. 15 Loading and boundary conditions for the steering knuckle with an initial surface crack.

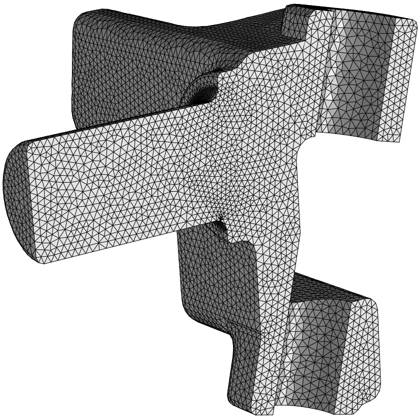


Fig. 16 Finite element mesh for the computational model.

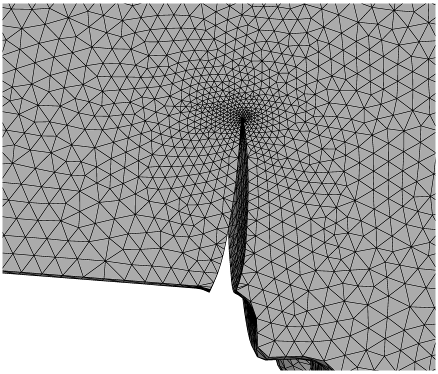


Fig. 17 Detail of the crack at the symmetry plane.

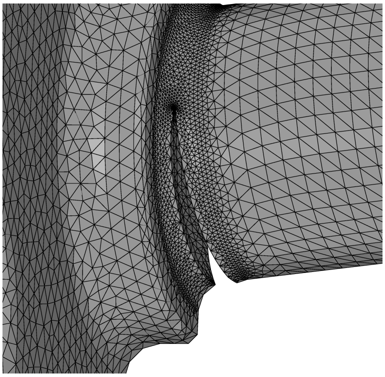


Fig. 18 Detail of the crack at the surface.

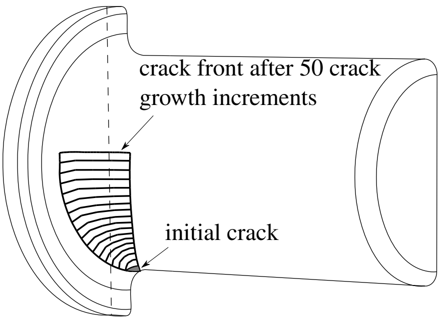


Fig. 19 Crack area at the base of the wheel axle.

5 Conclusions

The simulation program PROCRACK is capable of handling the automated finite element simulation of three-dimensional fatigue crack growth for a component with an initial crack. The preprocessing of the computational model and its numerical analysis is carried out with ABAQUS, a software system for finite element analysis and computer-aided engineering. Therefore, the geometry of the investigated component and its thermomechanical loading can be arbitrary. The method for analyzing the crack within the framework of the linear elastic fracture mechanics is the modified crack closure integral or the interaction integral method. Both the PARIS law and the NASGRO equation are available for selection of the crack growth law. The presented application examples show the capabilities of the simulation program. Due to the use of the open-source computer language PYTHON an easy extension and maintenance of the simulation program is assured.

References

[1] Areias, P., Belytschko, T.: Analysis of three-dimensional crack initiation and propagation using the extended finite element method. Int. J. Numer. Meth. Engrg. 63, 760–788 (2005)

[2] Buchholz, F.: Lokale Formeln höherer Ordnung zur Methode des modifizierten Rißschließungsintegrals. Ingenieur-Archiv. 55, 342–347 (1985)

[3] Cherepanov, G.: Crack propagation in continous media. J. Appl. Math. Mech. 31, 503–512 (1967)

[4] Erdogan, F., Sih, G.: On the crack extension in plates under plane loading and transverse shear. ASME J. Basic Engrg. 85, 519–527 (1963)

[5] FKM-Richtlinie: Bruchmechanischer Festigkeitsnachweis für Maschinenbauteile, 3. Ausgabe. VDMA-Verlag GmbH (2006)

[6] Forman, R., Mettu, S.: Behavior of surface and corner cracks subjected to tensile and bending loads in a Ti-6Al-4V alloy. In: Ernst, H., Saxena, A., McDowell, D. (eds.) Fracture Mechanics: Twenty-Second Symposium ASTM STP 1131, vol. I, pp. 519–546. American Society for Testing and Materials, Philadelphia (1992)

[7] Gosz, M., Moran, B.: An interaction energy integral method for computation of mixed-mode stress intensity factors along non-planar crack fronts in three dimensions. Engrg. Fract. Mech. 69(3), 299–319 (2002)

[8] Gravouil, A., Moes, N., Belytschko, T.: Non-planar 3D crack growth by the extended finite element and level sets – Part II: Level set update. Int. J. Numer. Meth. Engrg. 53, 2569–2586 (2002)

[9] Hübner, P., Pusch, G.: Zyklisches Rißwachstumsverhalten von Baustählen und deren Schweißverbindungen - Analytische Aufbereitung für die Nutzung des Berechnungsprogrammes ESACRACK. In: DVM-Bericht 234: Anwendung der Bruch-und Schädigungsmechanik, DVM, Berlin, pp. 129–138 (2002)

[10] Hübner, P., Pusch, G., Zerbst, U.: Ableitung von Quantilrisswachstumskurven für Restlebensdauerberechnungen. In: DVM-Bericht 236: Fortschritte der Bruch-und Schädigungsmechanik, DVM, Berlin, pp. 121–130 (2004)

[11] Kuna, M.: Numerische Beanspruchungsanalyse von Rissen – FEM in der Bruchmechanik. Vieweg + Teubner (2010)

[12] Meyer, A., Rabold, F., Scherzer, M.: Efficient finite element simulation of crack propagation using adaptive iterative solvers. Comm. Numer. Meth. Engrg. 22, 93–108 (2006)

[13] NASA: Fatigue crack growth computer program "NASGRO" version 3.0 – reference manual. NASA, Lyndon B. Johnson Space Center, Texas, USA, jSC-22267B (2000)

[14] Paris, P., Erdogan, F.: Crack tip stress intensity factors for plane extension and plate bending problems. J. Basic Eng. (Trans. ASME, D) 85, 528–543 (1963)

[15] Rabold, F.: Simulation der Rissausbreitung mit Hilfe adaptiver Finite-Elemente-Verfahren für elastische und plastische Materialien. TU Bergakademie Freiberg, Institut für Mechanik und Fluiddynamik (2009)

[16] Rice, J.: A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. 35, 376–386 (1968)

[17] Richard, H., Schöllmann, M., Fulland, M., Sander, M.: Experimental and numerical simulation of mixed-mode crack growth. In: de Freitas, M. (ed.) Proc. 6th Int. Conference on Biaxial/Multiaxial Fatigue and Fracture, Lisboa, Portugal, June 2001, vol. II, pp. 623–630. Instituto Superior Tecnico, Lisboa (2001)

[18] Rybicki, E., Kanninen, M.: A finite element calculation of stress intensity factors by a modified crack closure integral. Engrg. Fract. Mech. 9(4), 931–938 (1977)

[19] Schöllmann, M., Fulland, M., Richard, H.: Development of a new software for adaptive crack growth simulations in 3d structures. Engrg. Fract. Mech. 70(2), 249–268 (2003)

[20] Stern, M., Becker, E., Dunham, R.: A contour integral computation of mixed-mode stress intensity factors. Int. J. Fracture 12(3), 359–368 (1976)

[21] Sukumar, N., Chopp, D., Moran, B.: Extended finite element method and fast marching method for three-dimensional fatigue crack propagation. Engrg. Fract. Mech. 70, 29–48 (2003)

[22] Yau, J., Wang, S., Corten, H.: A mixed-mode crack analysis of isotropic solids using conservation laws of elasticity. J. Appl. Mech. 47, 335–341 (1980)

[23] Zentech International Limited: Zencrack – A fracture mechanics software for automatic 3D crack propagation analysis (2011), `http://www.zentech.co.uk/zencrack.htm`

[24] Zuo, J., Deng, X., Sutton, M.: Advances in tetrahedral mesh generation for modelling of three-dimensional regions with complex, curvilinear crack shapes. Int. J. Numer. Meth. Engrg. 63, 256–275 (2005)