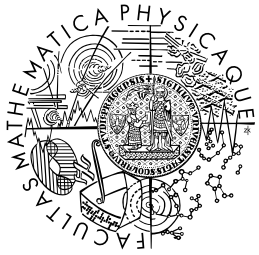


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Jan Papež

Algebraic Error in Matrix Computations in the Context of Numerical Solution of Partial Differential Equations

Department of Numerical Mathematics

Supervisor of the doctoral thesis:	prof. Ing. Zdeněk Strakoš, DrSc.
Study programme:	Mathematics
Specialization:	Scientific and Technical Calculations

Prague 2016

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague November 22, 2016 signature of the author

Title: Algebraic Error in Matrix Computations in the Context of Numerical Solution of Partial Differential Equations

Author: Jan Papež

Department: Department of Numerical Mathematics

Supervisor: prof. Ing. Zdeněk Strakoš, DrSc., Department of Numerical Mathematics

Abstract: Solution of algebraic problems is an inseparable and usually the most time-consuming part of numerical solution of PDEs. Algebraic computations are, in general, not exact, and in many cases it is even principally desirable not to perform them to a high accuracy. This has consequences that have to be taken into account in numerical analysis. This thesis investigates in this line some closely related issues. It focuses, in particular, on spatial distribution of the errors of different origin across the solution domain, backward error interpretation of the algebraic error in the context of function approximations, incorporation of algebraic errors to a posteriori error analysis, influence of algebraic errors to adaptivity, and construction of stopping criteria for (preconditioned) iterative algebraic solvers. Progress in these issues requires, in our opinion, understanding the interconnections between the phases of the overall solution process, such as discretization and algebraic computations.

Keywords: Numerical solution of partial differential equations, algebraic error, spatial distribution of error components, a posteriori error analysis, adaptivity, stopping criteria, preconditioning.

Název práce: Algebraická chyba v maticových výpočtech v kontextu numerického řešení parciálních diferenciálních rovnic

Autor: Jan Papež

Katedra: Katedra numerické matematiky

Vedoucí disertační práce: prof. Ing. Zdeněk Strakoš, DrSc., Katedra numerické matematiky

Abstrakt: Řešení algebraických úloh je neoddělitelnou a často také časově nejnáročnější částí procesu numerického řešení parciálních diferenciálních rovnic (PDR). Algebraické výpočty jsou obecně zatíženy chybami, a v mnoha případech je navíc vysoká přesnost algebraických výpočtů v kontextu celkového řešení dané úlohy nežádoucí. Numerická analýza musí umět pracovat s daným faktem a jeho důsledky. Předložená práce se v daném směru zabývá několika úzce souvisejícími tématy. Jsou to zejména rozložení složek chyby různého původu ve výpočetní oblasti, interpretace algebraických chyb využívající tzv. zpětnou chybu, zahrnutí algebraických chyb do a posteriorní analýzy chyb, vliv algebraických chyb na adaptivitu a konstrukce zastavovacích kritérií pro (předpodmíněné) algebraické řešiče. Dosažení pokroku v těchto otázkách předpokládá, dle našeho názoru, pochopení vzájemných vztahů mezi jednotlivými fázemi celého procesu numerického řešení PDR, jako jsou například diskretizace problému a algebraické výpočty.

Klíčová slova: Numerické řešení parciálních diferenciálních rovnic, algebraická chyba, rozložení dílčích složek chyby, a posteriori analýza chyby, adaptivita, zastavovací kritéria, předpodmínění.

This thesis encloses my studies at the Faculty of Mathematics and Physics, Charles University. Many people helped me to overcome all the difficulties and to make these years fruitful, filled with delight of challenging problems in the interesting field of numerical analysis. Here I would like to thank them.

My deepest gratitude belongs to my supervisor, Prof. Zdeněk Strakoš, who has been revealing to me the beauty of numerical mathematics since my undergraduate studies. He was always patient, kind, and also demanding. I can only hope to adhere, someday, to the thoroughness and consistency he has guided me to. I would like to thank Prof. Martin Vohralík for a warm welcome in Paris and for introducing me to the error estimates based on flux reconstruction. This is now an important part of my thesis and research. Also many other colleagues supported me throughout my studies and were always willing to spend their time in vivid discussions. My thanks go to all members of the "coffee club" at the second floor of the Institute of Computer Science, CAS, and to colleagues from Department of Numerical Mathematics. I am especially grateful to Marie Kubínová, Tomáš Gergelits, and Jan Kuřátko.

Finishing the thesis and my studies would not be possible without the support of my family and friends. Whenever necessary, they encouraged me and helped me to find a motivation for further work. In other times they provided me with a distraction and kept my mind (at least partially) away from formulas and unresolved issues. I would never be where I am without my parents and my dearest Mirka, without their love and firm faith in me. The last but not least word of this acknowledgment therefore belongs to them: děkuji.

The work present in the thesis was supported by the ERC-CZ project LL1202, by the Czech Science Foundation project 13-06684S, by the project IAA100300802 of the Grant Agency of the Academy of Sciences of the Czech Republic, and by the student research grants 695612 and 172915 of the Charles University Grant Agency.

Contents

1. Introduction

Solution of partial differential equations (PDEs) using the (Galerkin) finite element method (FEM) reduces the original *mathematical model*[1]

$$\text{to find } u \in V: \quad a(u,v) = \langle f, v\rangle \qquad \forall v \in V\,, \tag{1.1}$$

where V is a (infinite-dimensional) Hilbert space, $V^\#$ is its dual consisting of linear bounded functionals, $a(\cdot,\cdot): V \times V \to \mathbb{R}$ is a bilinear form (throughout the thesis we further assume that a is bounded and V-elliptic), $f \in V^\#$ is a bounded linear functional on V, and $\langle\cdot,\cdot\rangle : V^\# \times V \to \mathbb{R}$ is the duality pairing, to the *discretized problem*

$$\text{to find } u_h \in V_h: \quad a(u_h,v_h) = \langle f, v_h\rangle \qquad \forall v_h \in V_h\,, \tag{1.2}$$

where the FEM discrete solution u_h is restricted to some finite-dimensional function subspace (discretization subspace) $V_h \subset V$. Given a basis Φ of V_h, the problem (1.2) gives rise to an *algebraic problem*

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.3}$$

that determines the coefficients $\mathbf{x}$ of the discrete solution u_h with respect to the discretization basis Φ, $u_h = \Phi\mathbf{x}$. The FEM discretization basis functions have typically local support. Exact solution of the algebraic system then ensures the global approximation property of the FEM discrete solution u_h.

But in practice we do not solve algebraic problems exactly and only computed approximations $\widehat{\mathbf{x}}$ and $\widehat{u}_h$, $\widehat{u}_h = \Phi\widehat{\mathbf{x}} \in V_h$, are available. In difficult problems we even do not want to achieve a negligible algebraic error, which might be too costly or even out of reach. Then one should ask when to stop the algebraic solver (in iterative computation), and/or whether the spatial distribution of the algebraic error $u_h - \widehat{u}_h$ in the domain can significantly differ from the spatial distribution of the discretization error $u - u_h$. There is no a priori evidence that these distributions are to be analogous. From the nature of algebraic solvers, either direct or iterative, there seems to be no reason for equilibrating the algebraic error over the domain. Including algebraic error into a posteriori error analysis in numerical solution of PDEs and construction of stopping criteria for iterative solvers represent challenging problems. The presented thesis deals with several related questions.

In Chapter 2, which includes the paper Papež et al. [2014] and additional numerical experiments, we demonstrate that the algebraic error $u_h - \widehat{u}_h$ can indeed significantly dominate the total numerical error $u - \widehat{u}_h$ in some part of the domain despite the fact that its norm $\|\!|u_h - \widehat{u}_h|\!\|$ is small in comparison to the norm of the discretization error $\|\!|u - u_h|\!\|$. Here $\|\!|\cdot|\!\|$ stands for an appropriate, problem-related norm on V, e.g., the energy norm $\|\!|v|\!\| \equiv (a(v,v))^{1/2}$.

By the algebraic (forward) error one understands the difference $u_h - \widehat{u}_h$ of the FEM discrete solution and the computed approximation. A standard algebraic error analysis methodology is based on the algebraic backward error that

[1] This is an abstract setting that covers a range of linear elliptic second order boundary value problems; see, e.g., Málek and Strakoš [2015].

interprets the inaccuracies in the algebraic solution of (1.3) as a modification (perturbation) of the data $\mathbf{A}$, $\mathbf{b}$ defining (1.3). In Chapter 3 we relate the algebraic backward error with a modification of the original model (1.1) and its discretization. We interpret the algebraic backward error as transformation of the discretization basis (elaborating further on Gratton et al. [2013]; Papež et al. [2014]), and as a modification of the Green's operator, i.e. the mapping of the source term f in (1.1) to the exact infinite-dimensional solution u. We further use the algebraic backward error for estimating the algebraic (forward) error $u_h - \widehat{u}_h$ via the Fréchet derivative of the matrix inversion.

Information about the error in numerical solution of PDEs should be determined from the computed quantities without hidden assumptions or uncomputable multiplicative factors. Moreover, any appropriate a posteriori error estimator should provide also the information about the local distribution of the error. Historically, most a posteriori analysis in numerical PDEs focuses on estimating the norm of the discretization error $|||u - u_h|||$, which is crucial for adaptive finite element schemes that refines discretization in the parts of the domain where the estimator indicates a large (discretization) error in order to achieve its close-to-uniform spatial distribution over the domain. However, these estimators are often derived for the FEM discrete solution u_h, i.e., assuming the exact solution of the corresponding algebraic system (1.3). In Chapter 4 we consider, as an example, the so-called residual-based error estimator. In the included paper Papež and Strakoš [2016] we study the impact of abandoning the assumption on the exact algebraic solution on the estimator, allowing its evaluation at the presence of the algebraic error. Then we numerically illustrate the effect of the algebraic error on the adaptive finite element discretizations based on the local residual-based error indicators.

Mathematically justified, cheap, and accurate a posteriori error estimator together with properly set stopping criteria for iterative algebraic solvers are key ingredients of an efficient PDE numerical solver. They can significantly reduce the computational cost of the overall solution process preventing oversolving the algebraic problem while, at the same time, guaranteeing that the iterations are not stopped prematurely, i.e., that the error of the computed approximation is below the prescribed tolerance. Paper Papež et al. [2016] included in Chapter 5 presents a methodology for computing upper and lower bounds for both the algebraic and total error norms $|||u_h - \widehat{u}_h|||$, $|||u - \widehat{u}_h|||$. The derived bounds allow for estimating the local distribution of the errors over the computational domain. We also discuss bounds on the norm of the discretization error $|||u - u_h|||$ and their application for constructing stopping criteria balancing the discretization and algebraic errors.

When solving difficult problems, an algebraic preconditioning, i.e., the transformation of the algebraic system that aims at faster convergence behavior of the algebraic solver, is an inherent part of any practical solver. Construction of preconditioners is beyond the scope of the thesis. Following [Málek and Strakoš, 2015, Chapter 8], we show in Chapter 6 that any algebraic preconditioning can be interpreted as transformation of the discretization basis Φ and of the inner product in the infinite-dimensional function space V. This links algebraic preconditioning with the so-called operator preconditioning, i.e., the transformation of the infinite-dimensional problem (1.1). Results in Málek and Strakoš [2015]

and the experiments presented in Chapter 6 of this thesis demonstrate that discretization and preconditioning are tightly coupled.

The results, discussions, and numerical experiments of the thesis consider conforming finite element discretization of linear second-order elliptic (pure diffusion) model problems; in some parts we even restrict ourselves to the Poisson model problem. The observed phenomena pose questions formulated in Chapter 7 that should be taken into account in numerical PDEs in general. While the observations on model problems do not prove the significance of the points in practical problems, one can hardly assume that the observed difficulties disappear. Methodology for numerical solution of PDEs is typically developed on model problems, then extended and applied to real problems. The extension of the results, e.g. of Chapter 5, to difficult problems stemming from real-world applications should be done thoroughly and with a rigorous consideration of possible assumptions and restrictions.

Bibliography

S. Gratton, P. Jiránek, and X. Vasseur. Energy backward error: interpretation in numerical solution of elliptic partial differential equations and behaviour in the conjugate gradient method. *Electron. Trans. Numer. Anal.*, 40:338–355, 2013. ISSN 1068-9613.

J. Málek and Z. Strakoš. *Preconditioning and the conjugate gradient method in the context of solving PDEs*, volume 1 of *SIAM Spotlights*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015. ISBN 978-1-611973-83-9.

J. Papež and Z. Strakoš. Galerkin orthogonality and the multiplicative factors in the residual-based a posteriori error estimator for total error. Preprint MORE/2016/14, Submitted for publication, 2016.

J. Papež, J. Liesen, and Z. Strakoš. Distribution of the discretization and algebraic error in numerical solution of partial differential equations. *Linear Algebra Appl.*, 449:89–114, 2014. ISSN 0024-3795.

J. Papež, Z. Strakoš, and M. Vohralík. Estimating and localizing the algebraic and total numerical errors using flux reconstructions. Preprint MORE/2016/12, Submitted for publication, 2016.

2. Spatial distribution of errors in PDE test problems

The chapter provides a motivation for studying the algebraic error and, in particular, its spatial distribution in numerical solution of partial differential equations. As observed in the paper Papež et al. [2014] that is included in Section 2.1 and in the additional numerical experiments in Section 2.2, the algebraic error can have large local components and it can therefore dominate the total error in some parts of the domain. This demonstrates that the error estimates should provide the information about the local distribution of the errors because global error measures or estimates may not provide sufficient information for constructing reliable stopping criteria in adaptive computations or for iterative algebraic solvers.

The paper Papež et al. [2014] presents also an idea of interpreting the algebraic error as the modification of the discretization basis (in Section 3). It explains the observed behavior of the algebraic error using the spectral decomposition of the problem and the properties of the conjugate gradient method (in Section 4).

2.1 Paper published in Linear Algebra and its Applications

The section includes the paper Papež et al. [2014] published in Linear Algebra and its Applications, vol. 449 (2014).

Linear Algebra and its Applications 449 (2014) 89–114

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Distribution of the discretization and algebraic error in numerical solution of partial differential equations

J. Papež [a,b,1], J. Liesen [c,2], Z. Strakoš [a,*,3]

[a] *Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Prague, Czech Republic*
[b] *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic*
[c] *Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany*

ARTICLE INFO

Article history:
Received 30 April 2013
Accepted 4 February 2014
Available online xxxx
Submitted by E. Tyrtyshnikov

MSC:
65F10
65N15
65N30
65N22
65Y20

Keywords:
Numerical solution of partial differential equations
Finite element method

ABSTRACT

In the adaptive numerical solution of partial differential equations, local mesh refinement is used together with a posteriori error analysis in order to equilibrate the discretization error distribution over the domain. Since the discretized algebraic problems are *not solved exactly*, a natural question is whether the spatial distribution of the algebraic error is analogous to the spatial distribution of the discretization error. The main goal of this paper is to illustrate using standard boundary value model problems that this may not hold. On the contrary, the algebraic error can have large local components which can significantly dominate the total error in some parts of the domain. The illustrated phenomenon is of general significance and it is not restricted to some particular problems or dimensions. To our knowledge, the discrepancy between the spatial

* Corresponding author.
E-mail addresses: papez@cs.cas.cz (J. Papež), liesen@math.tu-berlin.de (J. Liesen), strakos@karlin.mff.cuni.cz (Z. Strakoš).

[1] The research of this author was supported by the GAUK grant 695612, by the ERC-CZ project LL1202 and by the project IAA100300802 of the Grant Agency of the ASCR.
[2] The research of this author was supported by the Heisenberg Program of Deutsche Forschungsgemeinschaft (DFG).
[3] The research of this author was supported by the GACR grant 201/09/0917 and by the ERC-CZ project LL1202.

http://dx.doi.org/10.1016/j.laa.2014.02.009

Adaptivity
A posteriori error analysis
Discretization error
Algebraic error
Spatial distribution of the error

distribution of the discretization and algebraic errors has not been studied in detail elsewhere.

1. Introduction

In numerical solution of partial differential equations, a sufficiently accurate solution (the meaning depends on the particular problem) of the linear algebraic system arising from discretization has to be considered. When the finite element method (FEM) is used for discretization, the system matrix is sparse. The sparsity of the algebraic system matrix is presented as a fundamental advantage of the FEM. It allows to obtain a numerical solution when the problem is hard and the discretized linear system is very large. It is worth, however, to examine some *mathematical* consequences which do not seem to be addressed in the FEM literature.

The FEM generates an approximate solution in form of a linear combination of basis functions with *local* supports. Each basis function multiplied by the proper coefficient thus approximates the desired solution only locally. The *global* approximation property of the FEM discrete solution is then ensured by solving the linear algebraic system for the unknown coefficients; the linear algebraic system links the local approximation of the unknown function in different parts of the domain. If the linear algebraic system is solved *exactly*, then all is fine. But in practice we do not solve exactly. In hard problems we even *do not want* to achieve a small algebraic error. That might be too costly or even impossible to get; see, e.g., [7, Sections 1–3], [24, Sections 1 and 6], [33, Section 2.6], the discussion in [34, pp. 36 and 72], and [38, Section 1]. Then, however, one should naturally ask whether the spatial distribution of the algebraic error in the domain can significantly differ from the distribution of the discretization error. There is no a priori evidence that these distributions are to be analogous. On the contrary, from the nature of algebraic solvers, either direct or iterative, there seems to be no reason for equilibrating the algebraic error over the domain. Numerical results presented in this paper demonstrate that the algebraic error can indeed significantly dominate the total error in some part of the domain. To our knowledge, apart from a brief discussion in [26, Sections 5.1 and 5.9.4], the presented phenomenon has not been studied elsewhere.

In order to avoid misunderstandings, it is worth to point out that the phenomenon described in this paper is not related to the so-called "smoothing properties" of the conjugate gradient (CG) method [23] or to the investigation of smoothing in the multilevel setting (for such analyses see, e.g., [36] or [41, Chapter 9]). Moreover, it is *not* due to the particular iterative solver or due to the specifics of the model problems used in this paper for illustration. Following the standard methodology used in the numerical PDE literature for decades (see, e.g., [8,15,19]), we start by illustrating the phenomenon using the simplest 1D boundary value problem. Furthermore, in order to plot illustrative figures, we use a small number of discretization nodes. In order to avoid the impression

that the simplicity or specifics of the 1D model problem diminish the message, we also present numerical examples with more complicated 2D model problems that illustrate the same phenomenon.

Several other phenomena, in particular the pollution error (see, e.g., [9,31]) and superconvergence of the discretization error in the internal nodes (see, e.g., [42]) are also of interest in the investigation of the spatial error distributions. Investigations of such phenomena are, however, beyond the scope of this paper.

The paper is organized as follows. The 1D model problem and experimental observations for this problem are described in Section 2. In Section 3 the total error is interpreted via a modification of the discretization mesh. Section 4 explains the local behavior of the algebraic error using the spectral analysis and the approximation properties of the algebraic solver (here the CG method). Section 5 presents some numerical results that illustrate the presence of the described phenomenon on 2D model problems and adaptive PDE computations. The paper ends with concluding remarks.

2. 1D model problem

We consider the 1D Poisson boundary value problem

$$-u''(x) = f(x), \quad 0 < x < 1, \qquad u(0) = u(1) = 0, \tag{1}$$

where $f(x)$ is a given (continuous) function, $0 \leqslant x \leqslant 1$. This model problem is frequently used in mathematical literature for illustrations of various analytical as well as numerical phenomena; see, e.g., [15, Section 6.2.2], [19, Section 5.5], [30], [32, Section 3.2.1].

Denoting by $H_0^1(\Omega)$ the standard Sobolev space of functions having square integrable (weak) derivatives in $\Omega \equiv (0,1)$ and vanishing on the end points (in the sense of traces), the weak formulation of (1) looks for $u \in H_0^1(\Omega)$ such that

$$a(u,v) = \ell(v) \quad \text{for all } v \in H_0^1(\Omega), \tag{2}$$

where

$$a(u,v) \equiv \int_0^1 u'v', \qquad \ell(v) \equiv \int_0^1 vf.$$

The bilinear form $a(\cdot,\cdot)$ introduces on $H_0^1(\Omega)$ the *energy norm*

$$\|v'\| = a(v,v)^{1/2}, \quad v \in H_0^1(\Omega). \tag{3}$$

We point out that the energy norm is relevant in many applications; see, e.g., [20, Section 2.2.1].

We discretize the problem (2) by the FEM on the uniform mesh with n inner nodes, i.e. with the mesh size $h = 1/(n+1)$, using the continuous piecewise linear basis functions ϕ_j, $j = 1, \dots, n$, satisfying

$$\begin{aligned} &\phi_j(jh) = 1, \\ &\phi_j(x) = 0, \quad 0 \leqslant x \leqslant (j-1)h \quad \text{and} \quad (j+1)h \leqslant x \leqslant 1. \end{aligned}$$

The discretized problem then looks for $u_h \in V_h \equiv \operatorname{span}\{\phi_1, \dots, \phi_n\}$ such that

$$a(u_h, v_h) = \ell(v_h) \quad \text{for all } v_h \in V_h. \tag{4}$$

The finite-dimensional problem (4) can be equivalently formulated as the system of the linear algebraic equations

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{5}$$

where the *stiffness matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ and the *load vector* $\mathbf{b} \in \mathbb{R}^n$ are given by

$$\mathbf{A} = [A_{ij}], \qquad A_{ij} = a(\phi_j, \phi_i), \tag{6}$$

$$\mathbf{b} = [b_1, \dots, b_n]^T, \qquad b_i = \ell(\phi_i), \quad i, j = 1, \dots, n. \tag{7}$$

The solution $\mathbf{x} = [\xi_1, \dots, \xi_n]^T$ of (5) contains the coefficients of the Galerkin FEM solution u_h of (4) with the respect to the FEM basis $\phi_1, \dots, \phi_n$, i.e.

$$u_h = \sum_{j=1}^{n} \xi_j \phi_j. \tag{8}$$

In the 1D problem (1), the Galerkin FEM solution u_h is known to coincide with the solution u at the nodes of the mesh; see, e.g., [8, Corollary 4.1.1]. Therefore the coefficients ξ_j are equal to the values of u in the nodes,

$$\xi_j = u(jh), \quad j = 1, \dots, n. \tag{9}$$

The stiffness matrix $\mathbf{A}$ has the tridiagonal form

$$\mathbf{A} = h^{-1} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}. \tag{10}$$

The eigenvalues λ_i and eigenvectors $\mathbf{y}_i = [\eta_{1i}, \dots, \eta_{ni}]^T$ of $\mathbf{A}$, $i = 1, \dots, n$, are known analytically (for details and their relationship to the eigenvalues and eigenfunctions of the continuous Laplace operator see, e.g., [10]),

$$\lambda_i = 4h^{-1}\sin^2\left(\frac{i\pi}{2(n+1)}\right), \tag{11}$$

$$\eta_{ji} = \sqrt{\frac{2}{n+1}}\sin\left(\frac{ji\pi}{n+1}\right), \quad j = 1,\dots,n. \tag{12}$$

The approximations w_i to the eigenfunctions of the continuous operator are then given by

$$w_i = \sum_{j=1}^{n} \eta_{ji}\phi_j, \qquad w_i(\ell h) = \eta_{\ell i}. \tag{13}$$

Remark 1. Unlike in the 2D Poisson problem, the stiffness matrix $\mathbf{A}$ in (10) and hence its eigenvalues in (11) depend on the mesh size through the multiplicative factor h^{-1}. This is often avoided by multiplying the system $\mathbf{Ax} = \mathbf{b}$ by h, which does not affect the conditioning of the matrix. However, since the algebraic energy norms $\|\mathbf{z}\|_{\mathbf{A}}$ and $\|\mathbf{z}\|_{(h\mathbf{A})}$ are different, such scaling would later be inconvenient, which is why we prefer to keep the matrix $\mathbf{A}$ as in (10).

We now consider solving the system $\mathbf{Ax} = \mathbf{b}$ using the (unpreconditioned) conjugate gradient (CG) method [23]. As mentioned in the Introduction, our point is to demonstrate on the simplest model problem the possible differences in the distribution of the discretization error and the algebraic error.

Given an initial approximation $\mathbf{x}_0$ and the corresponding initial residual $\mathbf{r}_0 \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_0$, the CG method generates approximations $\mathbf{x}_k \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$, where $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) \equiv \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \dots, \mathbf{A}^{k-1}\mathbf{r}_0\}$ is the kth Krylov subspace generated by $\mathbf{A}$ and $\mathbf{r}_0$. It is well known that these approximations minimize the $\mathbf{A}$-norm of the error, i.e.,

$$\|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}} = \min_{\mathbf{z}\in\mathbf{x}_0+\mathcal{K}_k(\mathbf{A},\mathbf{r}_0)} \|\mathbf{x} - \mathbf{z}\|_{\mathbf{A}};$$

see, e.g., [23, Theorem 4.3].

Writing $\mathbf{x}_k = [\xi_1^{(k)}, \dots, \xi_n^{(k)}]^T$, the resulting approximation of the Galerkin solution u_h in (8) is given by

$$u_h^{(k)} = \sum_{j=1}^{n} \xi_j^{(k)}\phi_j. \tag{14}$$

If u is the exact solution of the model problem (2), then $u - u_h$ is the *discretization error*, $u_h - u_h^{(k)}$ is the *algebraic error*, and $u - u_h^{(k)}$ is the *total error*. As a simple consequence of the Galerkin orthogonality property, the energy norms of these errors satisfy

$$\begin{aligned}\left\|\left(u - u_h^{(k)}\right)'\right\|^2 &= \left\|(u - u_h)'\right\|^2 + \left\|\left(u_h - u_h^{(k)}\right)'\right\|^2 \\ &= \left\|(u - u_h)'\right\|^2 + \|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}}^2;\end{aligned} \tag{15}$$

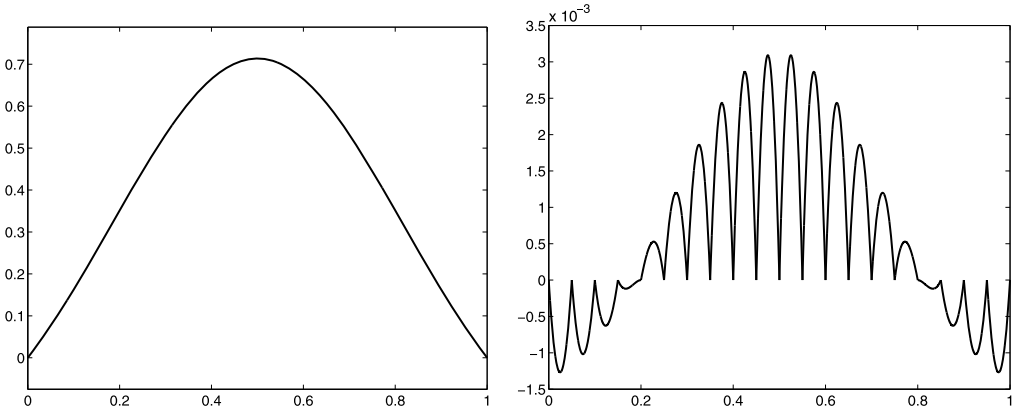


Fig. 1. Left: the exact solution u (see (16)). Right: the discretization error $u - u_h$; the vertical axis is scaled by 10^{-3}.

see, e.g., [14, Theorem 1.3, p. 38]. This means that the CG method leads to an approximation $u_h^{(k)}$ that minimizes the energy norm of the total error over all approximations determined by coefficient vectors from the affine subspace $\mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$.

Remark 2. The equality (15) holds for any vector $\mathbf{x}_k \in \mathbb{R}^n$ and the corresponding approximation of the form (14). In particular, it holds also for the results of finite precision CG computations.

As in [15, p. 120], we consider the exact solution

$$u = \exp\bigl(-5(x-0.5)^2\bigr) - \exp(-5/4) \tag{16}$$

of (1). To obtain the right-hand side $\mathbf{b}$ of the linear algebraic system one may use (9) and hence compute $\mathbf{b} = \mathbf{A}\mathbf{x}$. In order to use an approach analogous to higher dimensions we have chosen to evaluate $\mathbf{b}$ as in (7) using the MATLAB function `quad` (i.e. the adaptive Simpson rule). In comparison with the computation of $\mathbf{b} = \mathbf{A}\mathbf{x}$, the differences are, however, negligible. Furthermore, we have evaluated the error norms by applying the MATLAB function `quad` to the analytic expressions for $(u - u_h)'$ and $u - u_h$ in each subinterval.

Let us now describe our numerical results. We consider the FEM discretization using $n = 19$ inner nodes. This rather small number of nodes allows us to plot illustrative figures, but similar results can be obtained for any choice of n. The resulting solution u and the discretization error $u - u_h$ are shown in Fig. 1. The (squared) energy and L_2 norms of the discretization error are equal to (up to the negligible rounding errors in evaluation of the norms)

$$\bigl\|(u-u_h)'\bigr\|^2 = 6.8078\text{e}{-3} \quad \text{and} \quad \|u-u_h\|^2 = 1.7006\text{e}{-6}. \tag{17}$$

The condition number of the matrix $\mathbf{A}$ is $\kappa(\mathbf{A}) = \lambda_n/\lambda_1 = 161.4$.

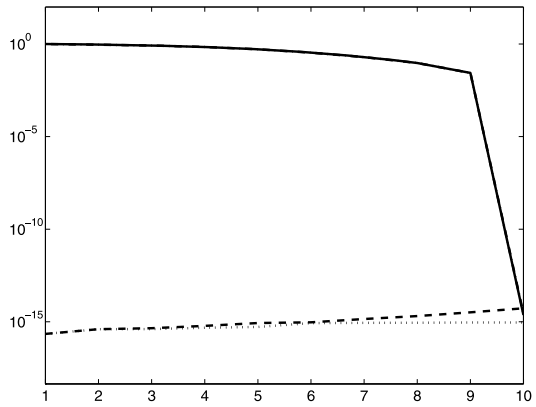


Fig. 2. The relative $\mathbf{A}$-norm of the error $\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}/\|\mathbf{x}-\mathbf{x}_0\|_{\mathbf{A}}$ (solid line), the loss of orthogonality in the standard CG implementation (dashed line) and the loss of orthogonality in the CG implementation with double reorthogonalized residuals (dotted line). In our computations, rounding errors do not play a significant role.

Table 1
Size of the algebraic and total error at several iteration steps for the exact solution (16).

k	$\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}^2$	$\|\mathbf{x}-\mathbf{x}_k\|^2$	$\|(u-u_h^{(k)})'\|^2$	$\|u-u_h^{(k)}\|^2$
7	6.3002e−2	9.9299e−3	6.9810e−2	4.9817e−4
8	1.4505e−2	9.5751e−4	2.1313e−2	4.9570e−5
9	1.2382e−3	2.7011e−5	8.0459e−3	3.0507e−6
10	6.3248e−30	2.2880e−31	6.8078e−3	1.7006e−6

In our experiments we apply the standard implementation of the CG method [23] with $\mathbf{x}_0 = 0$ to $\mathbf{A}\mathbf{x} = \mathbf{b}$. Fig. 2 shows the relative $\mathbf{A}$-norm of the algebraic errors $\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}/\|\mathbf{x}-\mathbf{x}_0\|_{\mathbf{A}}$. In order to show that rounding errors play (almost) no role in our reported results, we also plot the loss of orthogonality among the normalized residual vectors measured in the Frobenius norm for both the standard CG implementation and the CG implementation with double reorthogonalized residuals, which simulates exact arithmetic; see, e.g., [22]. We observe that the loss of orthogonality in the standard CG implementation remains close to the machine precision level, so that the effect of rounding errors indeed is negligible. Taking into account the distribution of the eigenvalues of $\mathbf{A}$ and the choice $\mathbf{x}_0 = \mathbf{0}$, this is to be expected; see [28].

The squared $\mathbf{A}$-norm of the algebraic error $\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}^2$ at the iteration steps $k = 7, 8, 9, 10$ is given in the first column of Table 1. The second column contains, for comparison, the squared Euclidean norm $\|\mathbf{x}-\mathbf{x}_k\|^2$. For the energy and the L_2 norm of the total error $u-u_h^{(k)}$ see the third and the fourth column, respectively.

The algebraic and total errors are visualized for the steps $k = 8, 9$ in Fig. 3. To describe our main point we look at the step $k = 9$. First note that at this step we have

$$\|\mathbf{x}-\mathbf{x}_9\|_{\mathbf{A}}^2 = 1.2382\text{e}{-3} < 6.8078\text{e}{-3} = \left\|(u-u_h)'\right\|^2;$$

cf. (17). In words, the globally measured energy norm of the algebraic error is smaller than the globally measured energy norm of the discretization error. On the other hand,

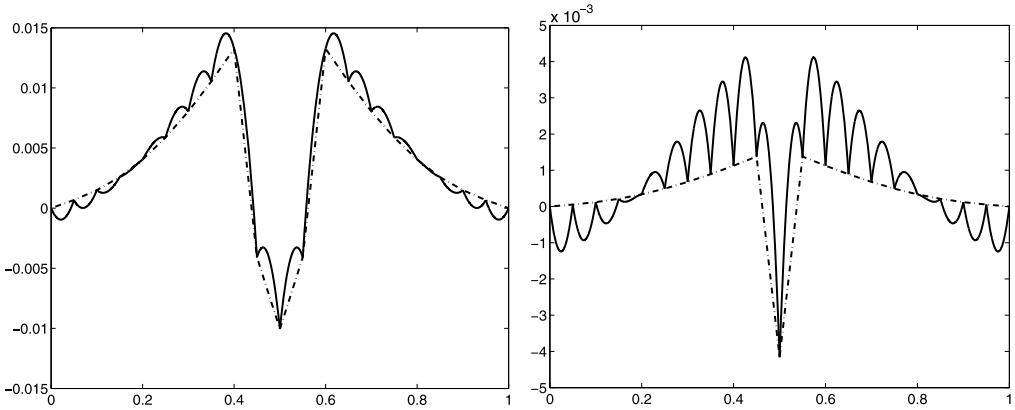


Fig. 3. The algebraic error $u_h - u_h^{(k)}$ (dashed–dotted line) and the total error $u - u_h^{(k)}$ (solid line) at the 8th iteration (left) and at the 9th iteration (right). The vertical axis in the right part of the figure is scaled by 10^{-3}.

as shown in the right part of Fig. 3, the algebraic error is strongly localized in the middle of the domain; here in particular at the component $\xi_{10}^{(9)}$ of $\mathbf{x}_9$, which is much less accurate than the other components. This localization of the algebraic error substantially affects the shape of the total error and leads to the following essential observations:

(1) *The spatial distributions of the discretization error and the algebraic error can be very different from each other.*
(2) *The value of the* (*globally measured*) *energy norm may not be descriptive.*

Similar observations of the error distribution can be made for $k = 8$, which is shown for illustration in the left part of Fig. 3. In this step, however, we have $\|\mathbf{x} - \mathbf{x}_8\|_{\mathbf{A}}^2 > \|(u - u_h)'\|^2$.

The presented example considers the simplest model problem. It does not *prove* that in practical problems the observed phenomenon appears on a catastrophic scale. On the other hand, the presented result is disturbing and poses a question about many commonly used ways of a posteriori error evaluation using global error measures, not distinguishing the sources of error or considering only the discretization error.

One may object that if the error is measured in the L_2 norm instead of the energy norm, one does not see much discrepancy – both $\|\mathbf{x}-\mathbf{x}_9\|_{\mathbf{A}}$ and $\|\mathbf{x}-\mathbf{x}_9\|$ are still relatively large in comparison to $\|u - u_h\|$. This, however, is not an objection against our two points made above. For the given model problem (as well as for a large class of problems with self-adjoint bounded and coercive operators; see, e.g., [19,20]) the energy norm is very natural to consider. With the Galerkin discretization it allows the fundamental Pythagorean identity to be expressed in the form (15), or, more generally, as

$$\left\|\nabla\left(u - u_h^{(k)}\right)\right\|^2 = \left\|\nabla(u - u_h)\right\|^2 + \|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}}^2. \tag{18}$$

This relates in a straightforward way the size of the discretization and algebraic errors. There is no equality analogous to (18) for the L_2 norm of the total, discretization and

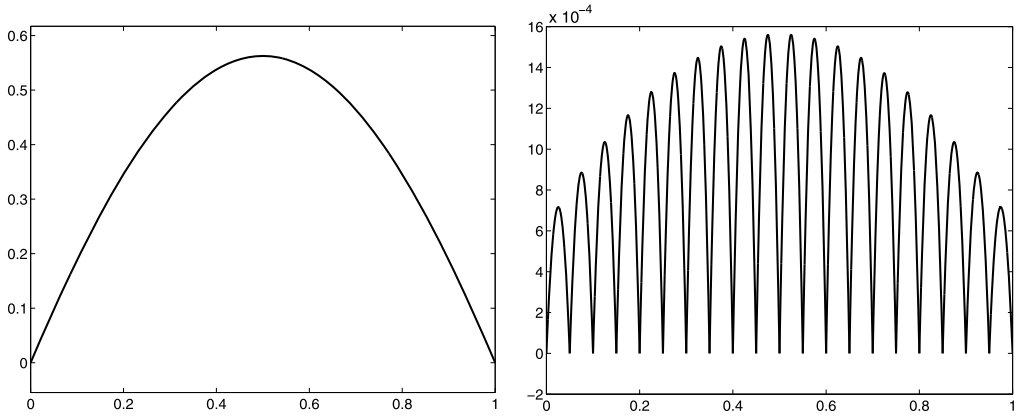


Fig. 4. Left: the exact solution u (see (19)). Right: the discretization error $u - u_h$; the vertical axis is scaled by 10^{-4}.

Table 2
Size of the algebraic and total error at several iteration steps for the exact solution (19).

k	$\|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}}^2$	$\|\mathbf{x} - \mathbf{x}_k\|^2$	$\|(u - u_h^{(k)})'\|^2$	$\|u - u_h^{(k)}\|^2$
7	1.0112e−2	1.1899e−3	1.3612e−2	6.0367e−5
8	2.6905e−3	1.6856e−4	6.1905e−3	9.3021e−6
9	2.5563e−4	5.7123e−6	3.7556e−3	1.1605e−6
10	5.6776e−30	3.8081e−30	3.5000e−3	8.7495e−7

algebraic errors. Moreover, the main point is that evaluation of the algebraic error globally using *any norm* is not sufficient. It should be complemented by investigation of the spatial distribution of the error over the domain or at the local areas of interest.

In order to demonstrate that the above observations are not an artefact of the special solution u in (16), we show also the results for the polynomial exact solution

$$u = (x-2)(x-1)x(x+1). \tag{19}$$

We choose again $n = 19$. The exact solution u and the discretization error $u - u_h$ are given in Fig. 4; the discretization error $u - u_h$ is nonnegative. The squared energy and L_2 norms of the discretization error are equal to

$$\left\|(u - u_h)'\right\|^2 = 3.5000\text{e}{-3} \quad \text{and} \quad \|u - u_h\|^2 = 8.7495\text{e}{-7}.$$

Table 2 and Figs. 5–6 give results analogous to those presented above in Table 1 and Figs. 2–3, respectively.

3. Interpretation of the total error as a modification of the discretization mesh

As argued in [26, p. 9], it is desirable to interpret the inaccuracies in the solution process (including the algebraic errors) in terms of a meaningful modification of the mathematical model; see also [35, pp. 33–35]. This idea can be related to the so-called

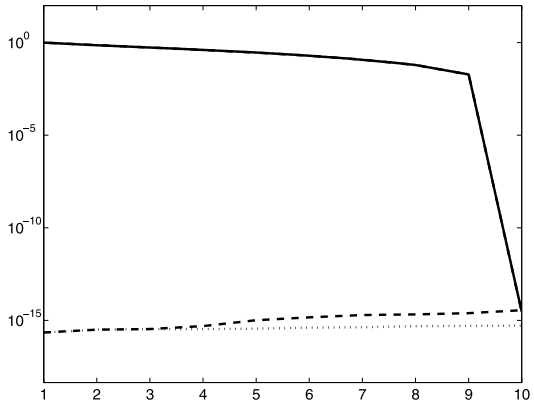


Fig. 5. The relative $\mathbf{A}$-norm of the error $\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}/\|\mathbf{x}-\mathbf{x}_0\|_{\mathbf{A}}$ (solid line), the loss of orthogonality in the standard CG implementation (dashed line) and the loss of orthogonality in the CG implementation with double reorthogonalized residuals (dotted line).

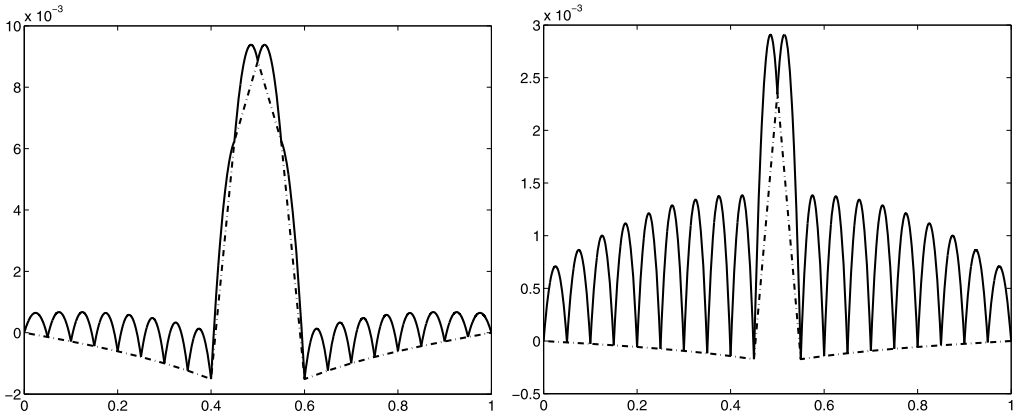


Fig. 6. The algebraic error $u_h - u_h^{(k)}$ (dashed–dotted line) and the total error $u - u_h^{(k)}$ (solid line) at the 8th iteration (left) and at the 9th iteration (right); the vertical axes are scaled by 10^{-3}.

functional backward error by Arioli and others (see, e.g., [6]) where the errors are interpreted as (backward) perturbations of the weak formulation of the problem. This can be appealing in more complicated settings where such perturbation represents a modification of the mathematical model that has some physical interpretation. Within the simple problem setting considered above, an introduction of the functional backward error term counting for inaccurate solving of the discretized algebraic problem into the left-hand side of problem (2) would not satisfy this natural requirement. As pointed out in [6], in the simple case of the Poisson problem (or in similar cases where perturbation of the operator would be difficult to interpret), the operator structure can be preserved by restricting the perturbation to the right-hand side only. This can be relevant, e.g., when the right-hand side is dominated by experimental data and the perturbation is small enough in comparison with experimental errors. In this paper we consider the change of the *discretization*, i.e. the basis functions or the mesh, as an alternative.

Interpreting the algebraic error as a transformation of the FEM basis has been considered in [21, Section 3]. We will use the idea from [21] but present the result in a slightly different way. Let the transformation of the basis $\Phi = [\phi_1, \dots, \phi_n]$ (in our problem the basis of continuous piecewise linear hat functions) to the basis $\widehat{\Phi} = [\widehat{\phi}_1, \dots, \widehat{\phi}_n]$ be represented by a square matrix $\mathbf{D} = [D_{\ell j}] \in \mathbb{R}^{n\times n}$,

$$\widehat{\phi}_j = \phi_j + \sum_{\ell=1}^{n} D_{\ell j}\phi_\ell, \quad j = 1, \dots, n. \tag{20}$$

Please note that unlike the original FEM basis functions ϕ_j, the transformed basis functions $\widehat{\phi}_j$, $j = 1, \dots, n$, need not be of a local support. The relation (20) can be written in the compact form as

$$\widehat{\Phi} = \Phi(\mathbf{I} + \mathbf{D}),$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ denotes the identity matrix.

The transformation matrix $\mathbf{D}$ can be constructed in the following way. An easy calculation shows that an approximate solution $\widehat{\mathbf{x}} = [\widehat{\xi}_1, \dots, \widehat{\xi}_n]^T$ of the algebraic system $\mathbf{Ax} = \mathbf{b}$ represents the *exact* solution of the perturbed system

$$(\mathbf{A} + \mathbf{E})\widehat{\mathbf{x}} = \mathbf{b}, \tag{21}$$

where

$$\mathbf{E} = \frac{(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}})\widehat{\mathbf{x}}^T}{\|\widehat{\mathbf{x}}\|^2}. \tag{22}$$

Let the Galerkin FEM solution u_h (see (4)–(8)) satisfy

$$u_h = \Phi\mathbf{x} = \sum_{j=1}^{n} \xi_j\phi_j = \sum_{j=1}^{n} \widehat{\xi}_j\widehat{\phi}_j = \widehat{\Phi}\widehat{\mathbf{x}} = \Phi(\mathbf{I} + \mathbf{D})\widehat{\mathbf{x}} \tag{23}$$

for some (unknown) matrix $\mathbf{D}$. Then, considering the Galerkin discretization of (2) with $u_h = \widehat{\Phi}\widetilde{\mathbf{x}}$, i.e. the discretization basis $\widehat{\phi}_1, \dots, \widehat{\phi}_n$, and the test functions $\phi_1, \dots, \phi_n$ gives

$$a(u_h, \phi_i) = \ell(\phi_i), \quad i = 1, \dots, n, \tag{24}$$

which can be formulated as the system of the linear algebraic equations

$$\widehat{\mathbf{A}}\widetilde{\mathbf{x}} = \mathbf{b},$$

where

$$\widehat{A}_{ij} = a(\widehat{\phi}_j, \phi_i) = a\left(\phi_j + \sum_{\ell=1}^{n} D_{\ell j}\phi_\ell, \phi_i\right)$$
$$= A_{ij} + \sum_{\ell=1}^{n} A_{i\ell} D_{\ell j}, \tag{25}$$

i.e.

$$\widehat{\mathbf{A}} = \mathbf{A} + \mathbf{AD}. \tag{26}$$

Consequently, knowing the algebraic perturbation matrix $\mathbf{E}$ from (21), we can set

$$\mathbf{AD} = \mathbf{E}, \quad \text{giving} \quad \mathbf{D} = \mathbf{A}^{-1}\mathbf{E}, \tag{27}$$

with $\widehat{\mathbf{x}} = \widetilde{\mathbf{x}}$ the exact algebraic solution of (21) representing the Galerkin solution u_h of (2) in the sense of (24).

Remark 3. Since $\mathbf{E}$ is determined by the algebraic errors in solving $\mathbf{Ax} = \mathbf{b}$, we have no control of the sparsity of the transformation matrix $\mathbf{D} = \mathbf{A}^{-1}\mathbf{E}$, which is, in general, *dense*. Therefore the transformed basis functions $\widehat{\phi}_j$, $j = 1, \ldots, n$, have, in general, *global supports*. This holds also when $\mathbf{E}$ is determined using componentwise backward error with its structure of nonzeros entries determined, e.g., by the structure of nonzeros in $\mathbf{A}$. Since $\mathbf{A}^{-1}$ is, in general, dense, $\mathbf{D} = \mathbf{A}^{-1}\mathbf{E}$ is also dense.

When we set $\widehat{\mathbf{x}} = \mathbf{x}_8$ for our experimental illustration with the exact solution (16), the norms of the perturbation and transformation matrices are

$$\|\mathbf{E}\| = 3.2976\text{e}{-1}, \qquad \|\mathbf{D}\| = 1.4674\text{e}{-2}.$$

Fig. 7 gives the matrices $\mathbf{E}$ (see (22)) and $\mathbf{D}$ (see (27)) visualized using the MATLAB `surf` command. We can see the effect of the multiplication by $\mathbf{A}^{-1}$: the transformation matrix $\mathbf{D}$ has significantly more entries with the size far from zero than the perturbation matrix $\mathbf{E}$. It should be pointed out that our example is on purpose very simple and the mapping from $\mathbf{E}$ to $\mathbf{D} = \mathbf{A}^{-1}\mathbf{E}$ is for the given $\mathbf{A}$ rather benign (the norm $\|\mathbf{D}\|$ is even smaller than $\|\mathbf{E}\|$). In practical problems this may not be the case and $\mathbf{D}$ can have large nonzero elements. The left part of Fig. 8 shows (for the same approximation $\widehat{\mathbf{x}} = \mathbf{x}_8$) the example of the transformed basis function $\widehat{\phi}_j$ (here $\widehat{\phi}_5$; see (20)). Since the entries of the matrix $\mathbf{D}$ are of the order 10^{-3}, $\widehat{\phi}_5$ looks visually the same as ϕ_5. The difference $\widehat{\phi}_5 - \phi_5$ is plotted in the right part of Fig. 8. For other basis functions the situation is analogous. The size of the differences $\widehat{\phi}_j - \phi_j$, $j = 1, \ldots, n$, corresponds to the size of the algebraic error (as well as the discretization error when the algebraic and discretization errors are in balance).

When we consider the approximation $\widehat{\mathbf{x}} = \mathbf{x}_9$ given at the 9th CG iteration step, the norms of the corresponding perturbation and transformation matrices are

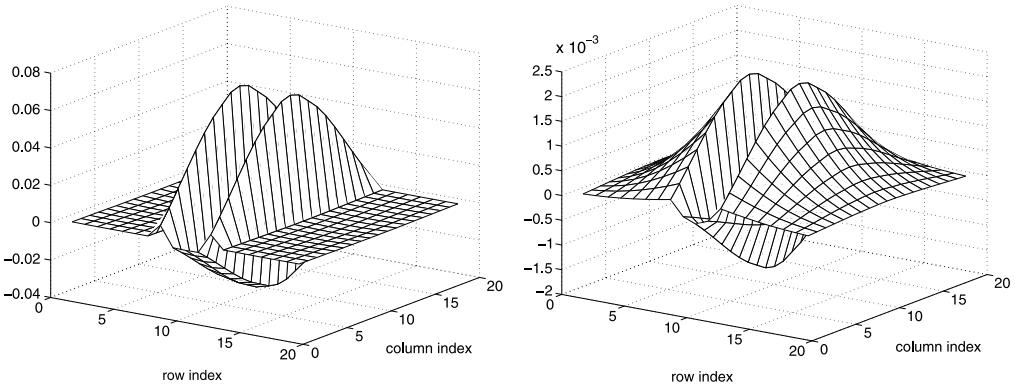


Fig. 7. The perturbation matrix $\mathbf{E}$ (left) and the transformation matrix $\mathbf{D}$ (right) (with the entries visualized using the MATLAB `surf` command) for the approximation $\widehat{\mathbf{x}} = \mathbf{x}_8$ in the example with the exact solution (16). The right vertical axis is scaled by 10^{-3}.

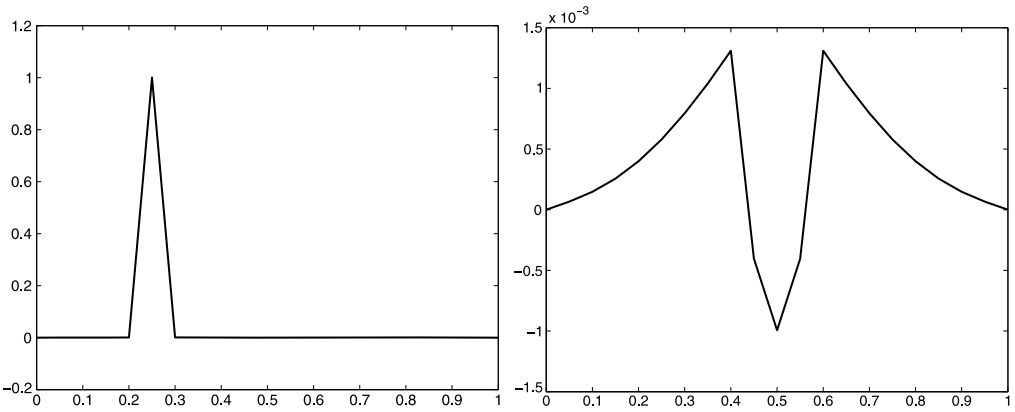


Fig. 8. The transformed basis function $\widehat{\phi}_5$ (left) and the difference $\widehat{\phi}_5 - \phi_5$ (right) for the approximation $\widehat{\mathbf{x}} = \mathbf{x}_8$ in the example with the exact solution (16). For the other basis functions the situation is analogous. The right vertical axis is scaled by 10^{-3}; see the scale in the right part of Fig. 1.

$$\|\mathbf{E}\| = 1.2976\text{e}{-1}, \qquad \|\mathbf{D}\| = 2.4469\text{e}{-3},$$

and the visualization of $\mathbf{E}$, $\mathbf{D}$ and the difference $\widehat{\phi}_j - \phi_j$, $j = 1, \ldots, n$, is analogous.

For the second example with the exact solution (19) and the approximation $\widehat{\mathbf{x}} = \mathbf{x}_9$ given at the 9th CG iteration step, the norms of the perturbation and transformation matrices are

$$\|\mathbf{E}\| = 6.8757\text{e}{-2}, \qquad \|\mathbf{D}\| = 1.3220\text{e}{-3}.$$

Fig. 9 gives the matrix $\mathbf{E}$ and the matrix $\mathbf{D}$. For the transformed basis function $\widehat{\phi}_{11}$ and the difference $\widehat{\phi}_{11} - \phi_{11}$ see Fig. 10.

In the rest of this section we interpret (with some unimportant inaccuracy) the total error $u - u_h^{(9)}$ for the last example (the exact solution u is given by (19) and $u_h^{(9)}$ is

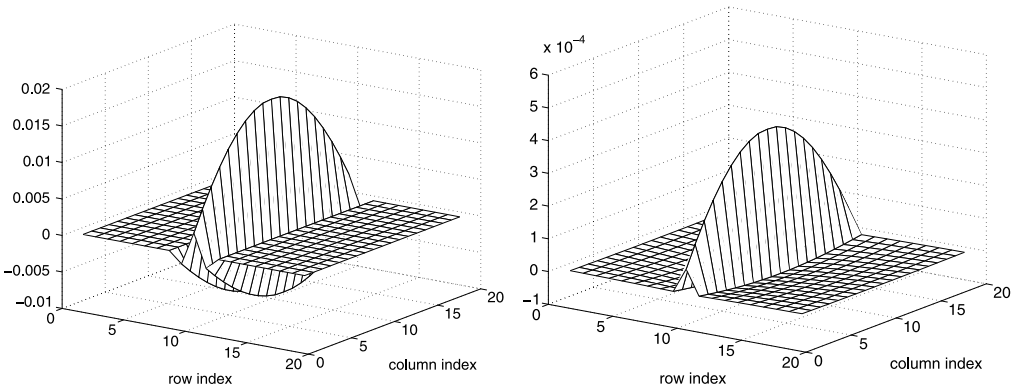


Fig. 9. The perturbation matrix $\mathbf{E}$ (left) and the transformation matrix $\mathbf{D}$ (right) (with the entries visualized using the MATLAB `surf` command) for the approximation $\widehat{\mathbf{x}} = \mathbf{x}_9$ in the example with the exact solution (19). The right vertical axis is scaled by 10^{-4}.

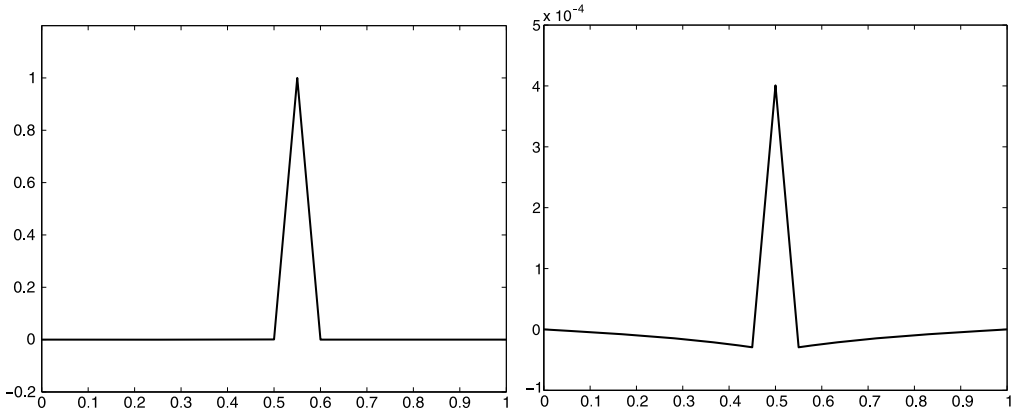


Fig. 10. The transformed basis function $\widehat{\phi}_{11}$ (left) and the difference $\widehat{\phi}_{11} - \phi_{11}$ (right) for the approximation $\widehat{\mathbf{x}} = \mathbf{x}_9$ in the example with the exact solution (19). For the other basis functions the situation is analogous. The right vertical axis is scaled by 10^{-4}; see the scale in the right part of Fig. 4.

determined using the approximation $\mathbf{x}_9$ computed at the 9th CG step) as the discretization error $u - u_H$, where the Galerkin FEM solution u_H corresponds to a *new mesh* and new basis functions which *preserve the locality of their support*. We are aware that this interpretation is here specific for the one-dimensional problem as it is certainly not easily applicable in general, especially for higher-dimensional problems. However, the distortion of the mesh illustrated below shows the possible disturbing effects of the localization of the algebraic error.

The Galerkin FEM solution u_H coincides with the solution u at the nodes of the mesh; see [8, Corollary 4.1.1]. Therefore we construct the new mesh in such a way that the new nodes τ_i are given as the roots of the total error $u - u_h^{(9)}$ (i.e. the discretization error $u - u_H$) and therefore

$$u_H(\tau_i) = u(\tau_i) = u_h^{(9)}(\tau_i).$$

In order to interpret the large total error in the middle of the interval as the discretization error, we replace (with no claim for optimality) the central node 0.5 of the original mesh by two nodes defined as $0.5 \pm 0.7h$, i.e.

$$\begin{aligned} \tau_i,\ i = 1, \dots, 18 &= \text{roots of } u - u_h^{(9)} \text{ for } 0 < x < 0.5, \\ \tau_{19} &= 0.5 - 0.7h, \\ \tau_{20} &= 0.5 + 0.7h, \\ \tau_i,\ i = 21, \dots, 38 &= \text{roots of } u - u_h^{(9)} \text{ for } 0.5 < x < 1. \end{aligned} \tag{28}$$

The new mesh now consists of $n = 38$ inner nodes, with 36 of them forming 18 close pairs. Please note that the new central element is 1.4 times longer than the elements in the original (uniform) mesh,[4] i.e. $\tau_{20} - \tau_{19} = 1.4h$.

Let ψ_j, $j = 1, \dots, n$, be the continuous piecewise linear FEM basis functions satisfying

$$\begin{aligned} &\psi_j(\tau_j) = 1, \\ &\psi_j(x) = 0, \quad 0 \leqslant x \leqslant \tau_{j-1} \quad \text{and} \quad \tau_{j+1} \leqslant x \leqslant 1. \end{aligned}$$

As mentioned above, the Galerkin solution u_H coincides with the solution u at the nodes of the mesh. We can therefore write

$$u_H = \sum_{j=1}^{n} \xi_j \psi_j, \quad \xi_j = u(\tau_j),\ j = 1, \dots, n.$$

The discretization error $u - u_H$ is nonnegative and the squared energy and L_2 norms of the discretization error $u - u_H$ are close to the analogous quantities for $u - u_h^{(9)}$,

$$\left\| (u - u_H)' \right\|^2 = 3.4224\text{e}{-3} \quad \text{respectively} \quad \|u - u_H\|^2 = 9.8141\text{e}{-7},$$

while

$$\left\| \left(u - u_h^{(9)}\right)' \right\|^2 = 3.7556\text{e}{-3} \quad \text{respectively} \quad \left\| u - u_h^{(9)} \right\|^2 = 1.1605\text{e}{-6}.$$

The comparison of the discretization error $u - u_H$ with the total error $u - u_h^{(9)}$ is given in the left part of Fig. 11. With our choice of the nodes (28), the positive values of $u - u_h^{(9)}$ coincide, except for $\tau_{18} < x < \tau_{21}$, with the error $u - u_H$; see the detail of the comparison in the right part of Fig. 11. There is a slight discrepancy between $u - u_H$ and $u - u_h^{(9)}$ for $\tau_{18} < x < \tau_{21}$.

Interpretation of the total error as the error of the *exact* discretized solution using a modified discretization mesh can rise, as illustrated above, interesting points. First, the

[4] This is the reason for denoting the Galerkin FEM solution corresponding to the new mesh with the subscript H commonly used for denoting the quantities corresponding to a coarser mesh.

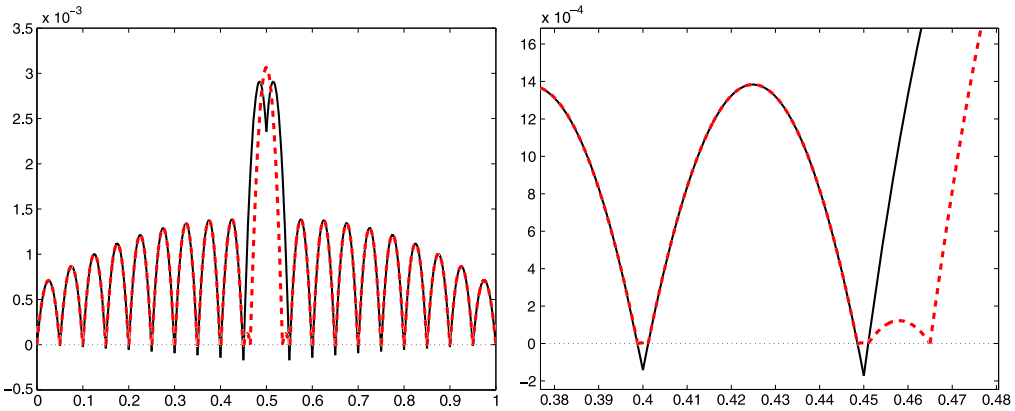


Fig. 11. Left: the total error $u - u_h^{(9)}$ for the original mesh (solid line) and the discretization error $u - u_H$ on the modified mesh (dashed line); the vertical axis is scaled by 10^{-3}. Right: the detail showing the coincidence of the positive values of $u - u_h^{(9)}$ with $u - u_H$ for most of the interval and their slight discrepancy in the middle; the vertical axis is scaled by 10^{-4}.

algebraic error can be interpreted, in the sense described above, as the loss of locality of the support of the modified Galerkin basis functions. Second, the computed approximate solution $u_h^{(k)}$ which includes the error in the solution of the algebraic system can be interpreted (here with a small inaccuracy) as the discrete solution (with the vanishing algebraic error) for a mesh which can possibly have "holes" in the areas where the algebraic error is large (in our construction specific for the 1D problem the mesh has a "hole" in the center of the interval).

4. Spatial distribution of the error in CG computations

In this section we explain the behavior of the algebraic error observed above; see also [26, Section 5.9.4]. In the following we present the experimental illustration with the exact solution (16); see also Figs. 7 and 8. The exposition uses the close relationship between CG and the Lanczos algorithm; for details see the original papers [23,25] and also the survey [28].

Consider the spectral decomposition of the CG error at the kth step,

$$\mathbf{x} - \mathbf{x}_k = \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_k, \mathbf{y}_i)\mathbf{y}_i, \tag{29}$$

where, as above, $\mathbf{y}_i$ denotes the ith normalized eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ_i; see (11)–(12). We denote by $\theta_j^{(k)}$, $j = 1, \ldots, k$, the approximations of the eigenvalues of the matrix $\mathbf{A}$ (*Ritz values*) given at the kth iteration of the Lanczos algorithm applied to the matrix $\mathbf{A}$ and the starting vector $\mathbf{r}_0/\|\mathbf{r}_0\|$. Assuming exact arithmetic, a close approximation of the eigenvalue λ_i by a Ritz value $\theta_j^{(k)}$ means that the size of the ith component $|(\mathbf{x} - \mathbf{x}_k, \mathbf{y}_i)|$ of the error $\mathbf{x} - \mathbf{x}_k$ of the kth CG approximation

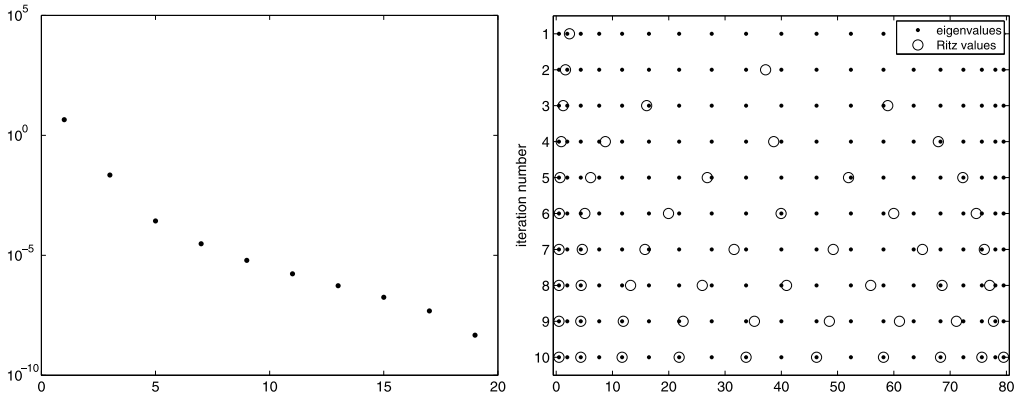


Fig. 12. Left: the squared size of the spectral components $|(\mathbf{x}-\mathbf{x}_0, \mathbf{y}_i)|^2$, $i = 1, \ldots, n$, of the initial error $\mathbf{x}-\mathbf{x}_0$. Right: convergence of the Ritz values (circles) to the eigenvalues of $\mathbf{A}$ (dots) in iterations 1 through 10.

in the direction $\mathbf{y}_i$ becomes small; see, e.g., [28, Theorem 3.3]. As mentioned above, the effect of rounding errors is in our example negligible. Consequently, the previous statement holds also for the presented results of finite precision computations.

Since some eigenvalues of $\mathbf{A}$ are approximated by Ritz values much faster than the others, this fact is reflected in the different behavior of the size of the spectral components $|(\mathbf{x} - \mathbf{x}_k, \mathbf{y}_i)|$, $i = 1, \ldots, n$, as k increases, $k = 0, 1, \ldots$. The individual eigenvectors $\mathbf{y}_i$ have different oscillating patterns; and therefore the individual spectral components of $\mathbf{x} - \mathbf{x}_k$ can develop in a rather nonuniform way as k increases. Using

$$u_h - u_h^{(k)} = \Phi(\mathbf{x} - \mathbf{x}_k) = \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_k, \mathbf{y}_i)\Phi\mathbf{y}_i = \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_k, \mathbf{y}_i) w_i,$$

this can result in a rather nonuniform spatial distribution of the algebraic (and the total) error in Ω. We will illustrate this situation in the following figures.

The squared size of the spectral components $|(\mathbf{x} - \mathbf{x}_0, \mathbf{y}_i)|^2$, $i = 1, \ldots, n$, of the initial error $\mathbf{x}-\mathbf{x}_0$ is given in the left part of Fig. 12. Recall that $\mathbf{x}_0 = \mathbf{0}$ and therefore the initial error is equal to the solution $\mathbf{x}$. Since the solution is symmetric with respect to the center 0.5 of the given interval, the spectral components with even indices vanish (the corresponding projections computed in finite precision arithmetic are on the machine precision level). Since the initial error $\mathbf{x}-\mathbf{x}_0$ is smooth (i.e. nonoscillating), the components of the error with higher indices, which correspond to more oscillating eigenvectors (see (12)), significantly decrease with increasing index i. The Ritz values $\theta_j^{(k)}$, $j = 1, \ldots, k$, are for $k = 1, \ldots, 10$ given in the right part of Fig. 12. The dots represent the eigenvalues of matrix $\mathbf{A}$. As expected, the Ritz values approximate the eigenvalues with odd indices. At the 10th iteration, all such eigenvalues are approximated, all components of the error $\mathbf{x} - \mathbf{x}_{10}$ become very small and the norm of the algebraic error drops to the machine precision level; see Fig. 2 and Table 1. We can observe that the eigenvalues λ_1, λ_3 and partially also λ_5 are approximated much faster (for smaller iteration number) than the others.

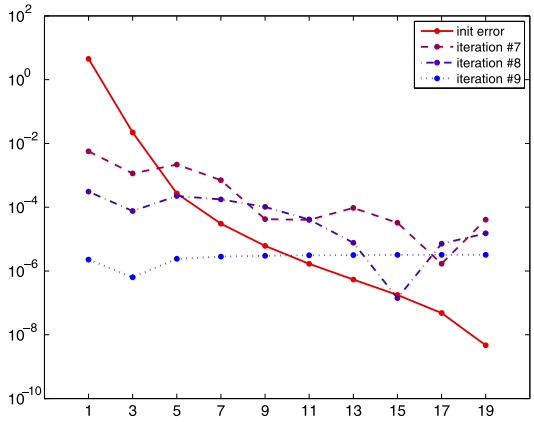


Fig. 13. The development of the squared size of the spectral components of the algebraic error $|(\mathbf{x}-\mathbf{x}_k, \mathbf{y}_i)|^2$, $i = 1, 3, \ldots, 19$, for the iteration steps $k = 0, 7, 8, 9$ (solid, dashed, dashed–dotted and dotted lines respectively). We can observe equilibrating of the size of the spectral components as k increases.

In Fig. 13 the development of the squared size of the spectral components of the algebraic error $\mathbf{x} - \mathbf{x}_k$ is shown for $k = 0, 7, 8, 9$ (only the values with odd indices are plotted; the rest remain at the level 10^{-30}). We can see that the CG method reduces quickly the dominating spectral components of the error which corresponds to the fast approximation of the eigenvalues λ_1 and λ_3 by the Ritz values illustrated above. With increasing k the spectral components of $\mathbf{x}-\mathbf{x}_k$ almost equilibrate. As a consequence, the spatial distribution of the error $\mathbf{x}-\mathbf{x}_k$ changes as k increases and it eventually becomes highly nonuniform in the way substantially different than the spatial distribution of the initial error $\mathbf{x} - \mathbf{x}_0$.

This situation is illustrated in Figs. 14 and 15, where we plot the most dominating approximations w_i to the eigenfunctions of the continuous operator (see (13) and (29)), corresponding to the initial error $\mathbf{x} - \mathbf{x}_0$ and to the error $\mathbf{x} - \mathbf{x}_9$ respectively. The right bottom part of Fig. 14 shows the algebraic part of the initial error in the function space, which is given as the linear combination of the eigenfunction approximations with odd indices

$$u_h - u_h^{(0)} = \Phi(\mathbf{x} - \mathbf{x}_0) = \sum_{i=1}^{10} (\mathbf{x} - \mathbf{x}_0, \mathbf{y}_{2i-1}) w_{2i-1}. \tag{30}$$

(As mentioned above, we use $\mathbf{x}_0 = \mathbf{0}$ and therefore $u_h - u_h^{(0)} = u_h$.) The right bottom part of Fig. 15 shows the algebraic part of the error

$$u_h - u_h^{(9)} = \Phi(\mathbf{x} - \mathbf{x}_9) \approx \sum_{i=1}^{10} (\mathbf{x} - \mathbf{x}_9, \mathbf{y}_{2i-1}) w_{2i-1}; \tag{31}$$

please compare with the algebraic error given in the right part of Fig. 3. Here we neglect the spectral components of $\mathbf{x}-\mathbf{x}_9$ in the direction of even eigenvectors of $\mathbf{A}$ which remain

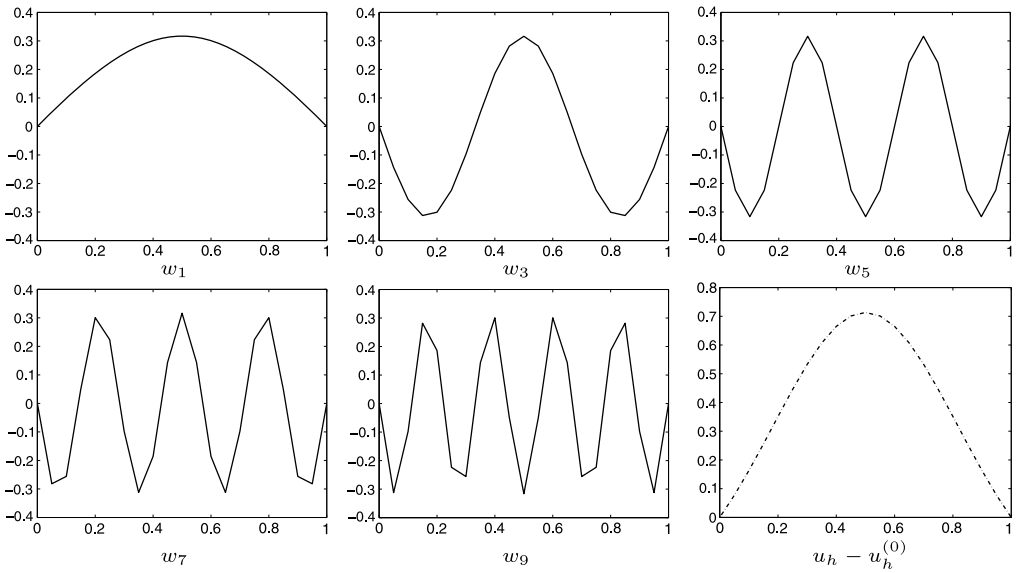


Fig. 14. The approximate eigenfunctions w_i corresponding to the largest components of the initial algebraic error $\mathbf{x} - \mathbf{x}_0$ in the eigenvector basis of the matrix $\mathbf{A}$ and the algebraic part $u_h - u_h^{(0)}$ of the initial error $u - u_h^{(0)}$ (see (30)) (the dashed–dotted line in the right bottom part).

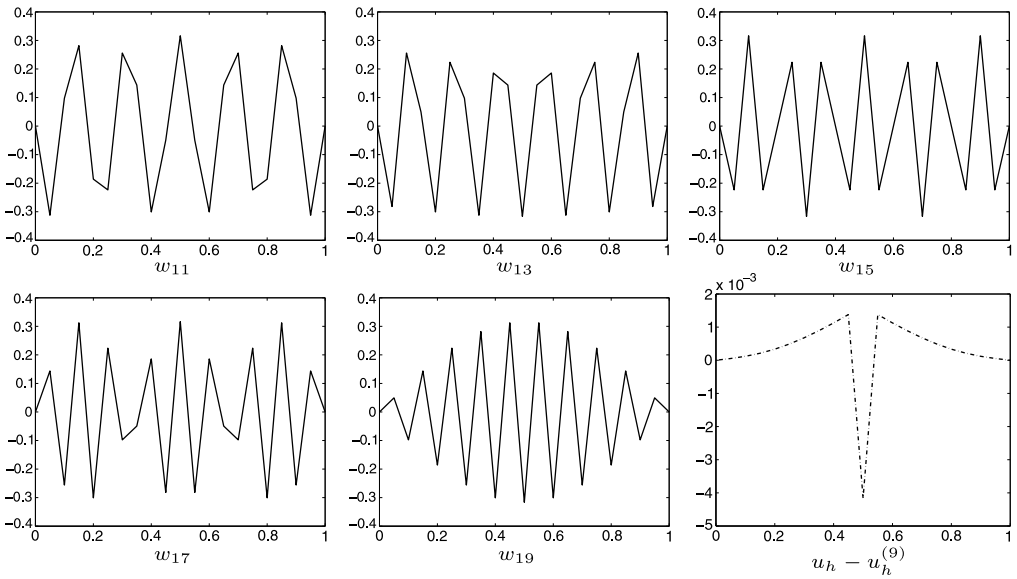


Fig. 15. The approximate eigenfunctions w_i corresponding to the largest components of the algebraic error $\mathbf{x} - \mathbf{x}_9$ in the eigenvector basis of the matrix $\mathbf{A}$ and the algebraic part $u_h - u_h^{(9)}$ of the error $u - u_h^{(9)}$ (see (31)) (the dashed–dotted line in the right bottom part). The vertical axis in the right bottom part of the figure is scaled by 10^{-3}.

at the machine precision level (and therefore we use the approximation instead of the equality).

In the following remark we do not consider the effects of rounding errors (it can easily be shown that for the given point their effects are not important). Since the CG approximate solution $\mathbf{x}_k$ satisfies $\mathbf{x}_k \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$, we have

$$\mathbf{x} - \mathbf{x}_k \in \mathbf{x} - \mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0).$$

The highly irregular spatial distribution of $u_h - u_h^{(9)}$ observed above is caused by *eliminating (to some extent) the spectral components with slowly changing eigenvectors*, which dominate the initial error $u_h - u_h^{(0)}$. As we have seen, all spectral components eventually become almost equal in size and the effect of rapidly changing eigenvectors becomes pronounced. This cannot be explained as one may seemingly suggest and as we have several times experienced during the preparation of this paper, by adding an "oscillatory" vector from $\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$ to $\mathbf{x} - \mathbf{x}_0$.

5. 2D illustrations

Using a simple 1D model problem, we illustrated above that the spatial distribution of the algebraic error can significantly differ from the spatial distribution of the discretization error. Because of its possibly large components in some parts of the domain, the algebraic error can determine the spatial distribution of the total error $u - u_h^{(k)}$ even when its globally measured size (here its energy norm) is smaller than the size of the discretization error. We emphasize that the described phenomenon is of general importance. It cannot be attributed to the specifics of the 1D model problem or the CG method used here for illustration. Of course, its appearance will be different for other problems or algebraic solvers.

In order to illustrate that the same phenomenon can appear also in more complicated settings, we present experiments using two well-known 2D model problems; see, e.g., [1,27].

Peak problem. We consider the 2D Poisson boundary value problem

$$-\Delta u = f \quad \text{in } \Omega \equiv (0,1) \times (0,1), \qquad u = 0 \quad \text{on } \partial\Omega. \tag{32}$$

The right-hand side f is chosen so that the solution u is given by

$$u(x,y) = x(x-1)y(y-1)\exp\left(-100\left(x - \frac{1}{2}\right)^2 - 100\left(y - \frac{117}{1000}\right)^2\right); \tag{33}$$

see the upper left part of Fig. 16.

L-shape problem. We consider the 2D Poisson boundary value problem

$$-\Delta u = 0 \quad \text{in } \Omega, \qquad u = u_D \quad \text{on } \partial\Omega, \tag{34}$$

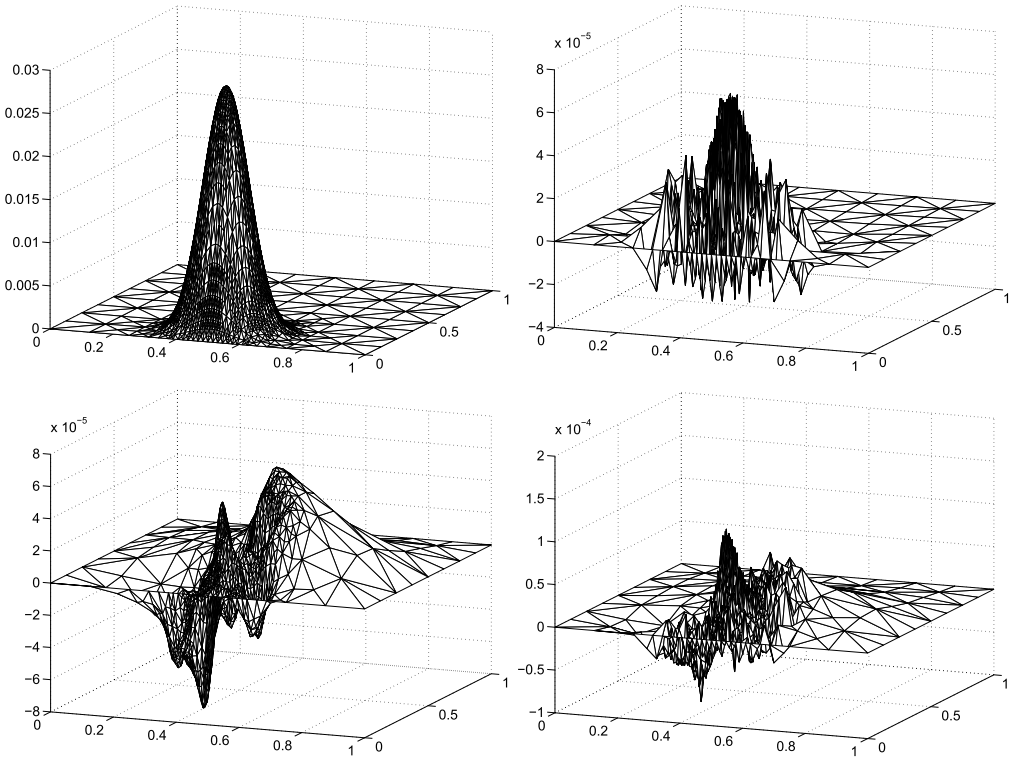


Fig. 16. Peak model problem (32) solved using an adaptively refined mesh with 1486 nodes. Upper left: the solution u (33). Upper right: the discretization error $u - u_h^*$; vertical axis is scaled by 10^{-5}. Bottom left: the algebraic error $u_h^* - u_h^{(k)}$; vertical axis is scaled by 10^{-5}. Bottom right: the total error $u - u_h^{(k)}$; vertical axis is scaled by 10^{-4}. The functions are visualized as piecewise affine functions using the MATLAB `trisurf` command.

where $\Omega \equiv (-1,1) \times (-1,1) \setminus (0,1) \times (-1,0)$. The Dirichlet boundary condition u_D is chosen so that the solution u is given in polar coordinates (r,θ) by

$$u(r,\theta) = r^{2/3} \sin\left(\frac{2}{3}\theta\right); \tag{35}$$

see the upper left part of Fig. 17.

For each model problem we consider a sequence of partitions (meshes) of the domain Ω into the union of non-overlapping, triangular elements such that the non-empty intersection of a distinct pair of elements is a single common node or a single common edge. On a given mesh we discretize the problem, analogously to Section 2, using the piecewise affine finite elements with the basis given by the *hat-functions*, i.e. the piecewise affine functions such that each one corresponds to a node of the partition taking there value 1 and vanishing in all other nodes. The boundary condition u_D is approximated by a piecewise affine function given by the values of u_D in the boundary nodes.

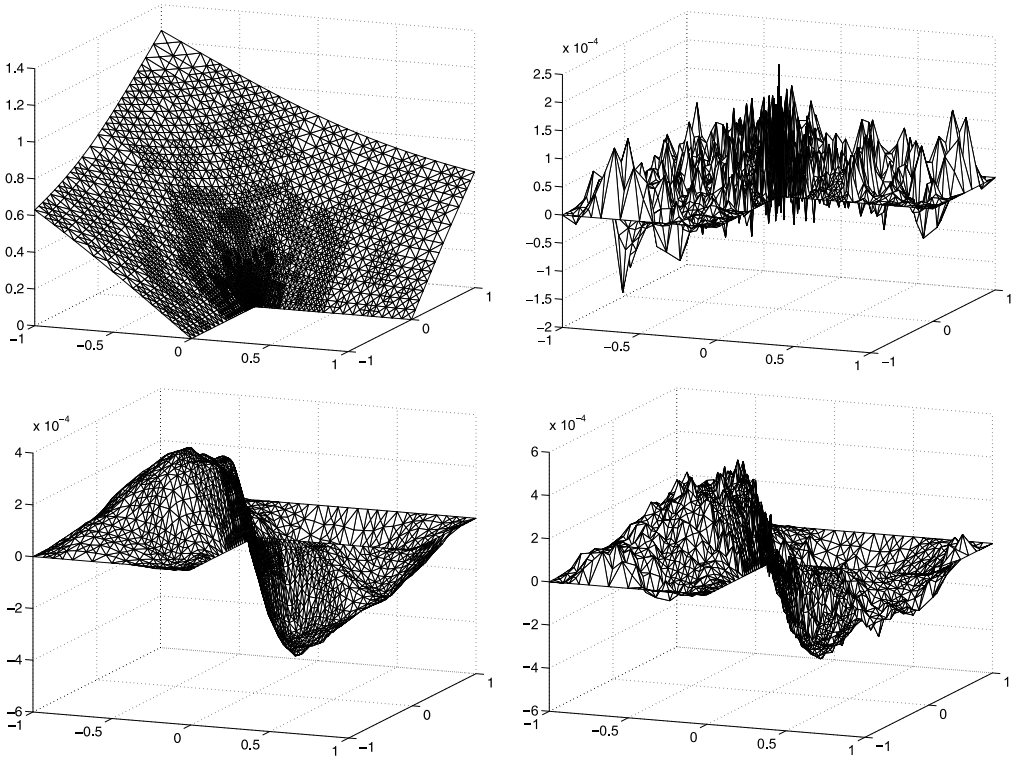


Fig. 17. L-shape model problem (34) solved using an adaptively refined mesh with 3376 nodes. Upper left: the solution u (35). Upper right: the discretization error $u - u_h^*$; vertical axis is scaled by 10^{-4}. Bottom left: the algebraic error $u_h^* - u_h^{(k)}$; vertical axis is scaled by 10^{-4}. Bottom right: the total error $u - u_h^{(k)}$; vertical axis is scaled by 10^{-4}. The functions are visualized as piecewise affine functions using the MATLAB `trisurf` command.

The stiffness matrix and right-hand side are assembled using the MATLAB code listed in [2].

Starting from the regular initial coarse mesh $\mathcal{T}_0$ consisting of 128 congruent triangles for the peak problem and of 192 congruent triangles for the L-shape problem, the sequence of adaptively refined meshes $\mathcal{T}_1, \mathcal{T}_2, \ldots$ is generated using the Adaptive Finite Element Method (AFEM). One iteration of AFEM can schematically be written as follows:

$$\text{SOLVE} \to \text{ESTIMATE} \to \text{MARK} \to \text{REFINE}.$$

Here "SOLVE" means assembling and solving the system of the linear algebraic equations. We solve the systems using the MATLAB backslash operator that gives, for our experiments, sufficiently accurate approximations (i.e. approximations with a normwise relative backward error on the machine precision level). The corresponding piecewise affine approximations are denoted by u_h^*. "ESTIMATE" means the local a posteriori estimation of the error between the exact solution u and its numerical approximation u_h^*.

We consider the residual-based local error estimator (indicator), for an element T of partition $\mathcal{T}_\ell$ and a piecewise affine approximation u_h^*

$$\eta_{R,T}^2(u_h^*) \equiv h_T^2 \|f\|_{L^2(T)}^2 + \sum_{E \subset \partial T} h_E \big\| [\nabla u_h^* \cdot n_E] \big\|_{L^2(E)}^2, \tag{36}$$

where $h_T \equiv \operatorname{diam}(T)$ denotes the diameter of the element T, $h_E \equiv \operatorname{diam}(E)$ denotes the length of an edge $E \subset \partial T$, and $[\nabla u_h^* \cdot n_E]$ denotes the jump of piecewise constant function ∇u_h^* over edge E. In a comparison of 13 a posteriori error estimators on five benchmark problems, the estimator $\eta_{R,T}(u_h^*)$ was found appropriate for practical use in adaptive algorithms in [11, Section 8]. For marking the elements ("MARK") we consider the so-called *greedy algorithm*; see [11, Section 6]. Let the elements of $\mathcal{T}_\ell$ be enumerated such that $\eta_{R,T_1}(u_h^*) \geqslant \eta_{R,T_2}(u_h^*) \geqslant \cdots$ (this enumeration is used here for the sake of a full rigor; practical algorithms use techniques described in literature given below).

For a given $\Theta \in (0,1]$ we find the smallest index m such that

$$\Theta \sum_{T \in \mathcal{T}_\ell} \eta_{R,T}^2(u_h^*) \leqslant \sum_{j=1}^{m} \eta_{R,T_j}^2(u_h^*);$$

see [12, Section 4.2] and for further development [37]. In the experiments we set $\Theta \equiv 0.25$. Finally, "REFINE" stands for the refinement of the elements $T_1, \ldots, T_m$ and the neighboring ones such that the conformity of the mesh is preserved. In the experiments we consider the refinement by Newest-Vertex-Bisection [29] implemented as in [17, Section 5.2].

For the first illustration we consider the peak model problem (32) and we use the mesh $\mathcal{T}_{13}$ given at the 13th AFEM iteration consisting of 1486 nodes. The tightly approximated squared energy norm of the discretization error (computed using the elementwise 16-node Gauss quadrature that is exact for polynomials up to degree 8; see, e.g., [13]) is equal to

$$\big\|\nabla(u - u_h^*)\big\|^2 = 9.5258\text{e}{-6}. \tag{37}$$

The discretization error $u - u_h^*$ visualized as a piecewise affine function (using the MATLAB `trisurf` command) is shown in the upper right part of Fig. 16.

The linear algebraic system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is of order 1436, which is equal to the number of the inner nodes in $\mathcal{T}_{13}$; the condition number is $\kappa(\mathbf{A}) = 1936.8$ (evaluated using the MATLAB `cond` function). Analogously to the 1D case, we apply the CG method with $\mathbf{x}_0 = 0$ to $\mathbf{A}\mathbf{x} = \mathbf{b}$. We stop at the iteration step $k = 67$ when the squared energy norm of the algebraic error $\|\mathbf{x}^* - \mathbf{x}_k\|_{\mathbf{A}}^2$ drops below one percent of the squared energy norm of the discretization error, i.e.

$$\big\|\mathbf{x}^* - \mathbf{x}_k\big\|_{\mathbf{A}}^2 < 0.01 \big\|\nabla(u - u_h^*)\big\|^2, \tag{38}$$

where $\mathbf{x}^*$ denotes the approximation to the solution $\mathbf{x}$ given by the MATLAB backslash operator. The criterion (38) is used here for a maximal rigor of our experimental illustrations. In practice a suitable approximation of $\mathbf{x}$ is not available, $\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}$ is estimated in various ways and incorporating algebraic error estimates into a posteriori error analysis with using it for construction of algebraic stopping criteria requires substantial further investigation; see, e.g., [4, Section 4.1], [28, Section 5.3], [39,40,24,3,6,5]. We denote by $u_h^{(k)}$ the piecewise affine approximation corresponding to the CG approximation $\mathbf{x}_k$. The squared energy norms of the algebraic error and the total error are equal to

$$\begin{aligned} \left\|\mathbf{x}^*-\mathbf{x}_k\right\|_{\mathbf{A}}^2 &= 7.7295\text{e}{-8}, \\ \left\|\nabla\left(u-u_h^{(k)}\right)\right\|^2 &= 9.6018\text{e}{-6}. \end{aligned} \tag{39}$$

Please recall the corresponding energy norm of the discretization error (37) and see the equality (18). The norm of the total error $\|\nabla(u-u_h^{(k)})\|^2$ is (tightly) approximated using elementwise the 16-node Gauss quadrature rule. As we can see in the bottom parts of Fig. 16, the algebraic error $u_h^*-u_h^{(k)}$ substantially affects the shape of the total error $u-u_h^{(k)}$ in the part of the domain Ω where the solution u is (nearly) constant (with small gradients) as well as in the part where u has large gradients.

For the second illustration we consider the L-shape model problem (34) and we use the mesh $\mathcal{T}_{13}$ given at the 13th AFEM iteration consisting of 3376 nodes. The quantities analogous to those presented above in (37) and (39) are

$$\begin{aligned} \left\|\nabla\left(u-u_h^*\right)\right\|^2 &= 2.4512\text{e}{-4}, \\ \left\|\mathbf{x}^*-\mathbf{x}_k\right\|_{\mathbf{A}}^2 &= 2.3873\text{e}{-6}, \\ \left\|\nabla\left(u-u_h^{(k)}\right)\right\|^2 &= 2.4751\text{e}{-4}. \end{aligned} \tag{40}$$

Here the system $\mathbf{A}\mathbf{x}=\mathbf{b}$ is of order 3210, and the condition number is $\kappa(\mathbf{A})=1230.3$ (evaluated using the MATLAB `cond` function). The stopping criterion (38) is satisfied at the iteration step $k=107$. The piecewise affine visualization of the discretization error $u-u_h^*$ is given in the upper right part of Fig. 17. As we can see in the bottom parts of Fig. 17, the algebraic error $u_h^*-u_h^{(k)}$ substantially affects the shape of the total error $u-u_h^{(k)}$ in most of the domain Ω.

6. Concluding remarks

The demonstrated difference between the spatial distributions of the algebraic and the discretization error across the domain (here obtained for the CG method) underlines the importance of constructing reliable stopping criteria for iterative algebraic solvers. In particular, in addition to evaluating parts of the error of different origin (discretization, inaccurate algebraic computations) in appropriate norms, such criteria should take into

account spatial distribution of the total error in the function space. References to the work in this direction can be found in the recent survey [4]; see also, e.g., [24, Section 6] and [16]. One should also recall the goal-oriented adaptivity approach of Rannacher, Becker and their collaborators in the context of duality-based error control, which allows balancing discretization and iteration error in the problem-related areas of interest; see, e.g., the survey papers [33,18] and the references given there. We believe that further developments focusing on the spatial distribution of the algebraic and total errors will be reported in the near future.

Acknowledgements

The authors are grateful to Mario Arioli for his thorough and stimulating report, and to Vít Dolejší, Valeria Simoncini, Gerhard Starke, Endre Süli and Martin Vohralík for useful comments.

References

[1] M. Ainsworth, Robust a posteriori error estimation for nonconforming finite element approximation, SIAM J. Numer. Anal. 42 (2005) 2320–2341.
[2] J. Alberty, C. Carstensen, S.A. Funken, Remarks around 50 lines of Matlab: short finite element implementation, Numer. Algorithms 20 (1999) 117–137.
[3] M. Arioli, E.H. Georgoulis, D. Loghin, Stopping criteria for adaptive finite element solvers, SIAM J. Sci. Comput. 35 (2013) A1537–A1559.
[4] M. Arioli, J. Liesen, A. Międlar, Z. Strakoš, Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems, GAMM-Mitt. 36 (2013) 102–129.
[5] M. Arioli, D. Loghin, A.J. Wathen, Stopping criteria for iterations in finite element methods, Numer. Math. 99 (2005) 381–410.
[6] M. Arioli, E. Noulard, A. Russo, Stopping criteria for iterative methods: applications to PDE's, Calcolo 38 (2001) 97–112.
[7] I. Babuška, Numerical stability in problems of linear algebra, SIAM J. Numer. Anal. 9 (1972) 53–77.
[8] I. Babuška, T. Strouboulis, The Finite Element Method and Its Reliability, Numer. Math. Sci. Comput., Oxford University Press, New York, 2001.
[9] I. Babuška, T. Strouboulis, A. Mathur, C.S. Upadhyay, Pollution-error in the h-version of the finite-element method and the local quality of a posteriori error estimators, Finite Elem. Anal. Des. 17 (1994) 273–321.
[10] D. Boffi, Finite element approximation of eigenvalue problems, Acta Numer. 19 (2010) 1–120.
[11] C. Carstensen, C. Merdon, Estimator competition for Poisson problems, J. Comput. Math. 28 (2010) 309–330.
[12] W. Dörfler, A convergent adaptive algorithm for Poisson's equation, SIAM J. Numer. Anal. 33 (1996) 1106–1124.
[13] D.A. Dunavant, High degree efficient symmetrical Gaussian quadrature rules for the triangle, Internat. J. Numer. Methods Engrg. 21 (1985) 1129–1148.
[14] H.C. Elman, D.J. Silvester, A.J. Wathen, Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics, Numer. Math. Sci. Comput., Oxford University Press, New York, 2005.
[15] K. Eriksson, D. Estep, P. Hansbo, C. Johnson, Computational Differential Equations, Cambridge University Press, Cambridge, 1996.
[16] A. Ern, M. Vohralík, Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs, SIAM J. Sci. Comput. 35 (2013) A1761–A1791.
[17] S. Funken, D. Praetorius, P. Wissgott, Efficient implementation of adaptive P1-FEM in Matlab, Comput. Methods Appl. Math. 11 (2011) 460–490.
[18] M.B. Giles, E. Süli, Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality, Acta Numer. 11 (2002) 145–236.

[19] M.S. Gockenbach, Partial Differential Equations: Analytical and Numerical Methods, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.
[20] M.S. Gockenbach, Understanding and Implementing the Finite Element Method, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
[21] S. Gratton, P. Jiránek, X. Vasseur, Energy backward error: interpretation in numerical solution of elliptic partial differential equations and behaviour in the conjugate gradient method, Electron. Trans. Numer. Anal. 40 (2013) 338–355.
[22] A. Greenbaum, Z. Strakoš, Predicting the behavior of finite precision Lanczos and conjugate gradient computations, SIAM J. Matrix Anal. Appl. 13 (1992) 121–137.
[23] M.R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, J. Res. Natl. Bur. Stand. 49 (1952) 409–436.
[24] P. Jiránek, Z. Strakoš, M. Vohralík, A posteriori error estimates including algebraic error and stopping criteria for iterative solvers, SIAM J. Sci. Comput. 32 (2010) 1567–1590.
[25] C. Lanczos, Solution of systems of linear equations by minimized iterations, J. Res. Natl. Bur. Stand. 49 (1952) 33–53.
[26] J. Liesen, Z. Strakoš, Krylov Subspace Methods: Principles and Analysis, Numer. Math. Sci. Comput., Oxford University Press, Oxford, 2013.
[27] R. Luce, B.I. Wohlmuth, A local a posteriori error estimator based on equilibrated fluxes, SIAM J. Numer. Anal. 42 (2004) 1394–1414.
[28] G. Meurant, Z. Strakoš, The Lanczos and conjugate gradient algorithms in finite precision arithmetic, Acta Numer. 15 (2006) 471–542.
[29] P. Morin, R.H. Nochetto, K.G. Siebert, Convergence of adaptive finite element methods, SIAM Rev. 44 (2002) 631–658, revised reprint of Data oscillation and convergence of adaptive FEM, SIAM J. Numer. Anal. 38 (2000) 466–488.
[30] A.E. Naiman, I. Babuška, H.C. Elman, A note on conjugate gradient convergence, Numer. Math. 76 (1997) 209–230.
[31] J.T. Oden, Y. Feng, Local and Pollution Error Estimation for Finite Element Approximations of Elliptic Boundary Value Problems, in: TICAM Symposium, Austin, TX, 1995, J. Comput. Appl. Math. 74 (1996) 245–293.
[32] A. Quarteroni, Numerical Models for Differential Problems, MS&A. Model. Simul. Appl., vol. 2, Springer-Verlag Italia, Milan, 2009, translated from the 4th (2008) Italian edition by Silvia Quarteroni.
[33] R. Rannacher, A short course on numerical simulation of viscous flow: discretization, optimization and stability analysis, Discrete Contin. Dyn. Syst. Ser. S 5 (2012) 1147–1194.
[34] P.J. Roache, Verification and Validation in Computational Science and Engineering, Hermosa Publishers, Albuquerque, NM, 1998.
[35] P.J. Roache, Building PDE codes to be verifiable and validatable, Comput. Sci. Eng. 6 (2004) 30–38.
[36] V.V. Shaidurov, Some estimates of the rate of convergence for the cascadic conjugate-gradient method, in: Selected Topics in Numerical Methods, Miskolc, 1994, Comput. Math. Appl. 31 (1996) 161–171.
[37] R. Stevenson, Optimality of a standard adaptive finite element method, Found. Comput. Math. 7 (2007) 245–269.
[38] Z. Strakoš, J. Liesen, On numerical stability in large scale linear algebraic computations, ZAMM Z. Angew. Math. Mech. 85 (2005) 307–325.
[39] Z. Strakoš, P. Tichý, On error estimation in the conjugate gradient method and why it works in finite precision computations, Electron. Trans. Numer. Anal. 13 (2002) 56–80.
[40] Z. Strakoš, P. Tichý, Error estimation in preconditioned conjugate gradients, BIT 45 (2005) 789–817.
[41] P.S. Vassilevski, Lecture notes on multigrid methods, Technical report LLNL-TR-439511, Lawrence Livermore National Laboratory, Livermore, CA, 2010.
[42] L.B. Wahlbin, Superconvergence in Galerkin Finite Element Methods, Lecture Notes in Math., vol. 1605, Springer-Verlag, Berlin, 1995.

2.2 Additional numerical experiments

In order to further demonstrate the possible difference between the spatial distribution of the discretization and algebraic errors across the solution domain we provide additional numerical experiments using two model problems with inhomogeneous diffusion tensor. The problems are from the class

$$\begin{aligned} -\nabla \cdot (\mathbf{S}\nabla u) &= 0 \qquad && \text{in } \Omega \equiv (-1,1)\times(-1,1)\,, \\ u &= u_D && \text{on } \partial\Omega\,, \end{aligned} \tag{2.1}$$

where the domain Ω is divided into four subdomains Ω_i corresponding to the axis quadrants numbered counterclockwise and $\mathbf{S}$ is a piecewise constant multiple of the identity matrix, $\mathbf{S}|_{\Omega_i} \equiv s_i\mathbf{I}$, $s_i > 0$; see, e.g., [Morin et al., 2002, Section 5.3]. The energy norm corresponding to the problem (2.1) is $\|\mathbf{S}^{1/2}\nabla v\|$, $v \in H^1_0(\Omega)$.

Inhomogeneous tensor problem I. We consider the problem (2.1) with the exact solution u given in polar coordinates (r,θ) by

$$u(r,\theta)|_{\Omega_i} = r^\alpha(a_i \sin(\alpha\theta) + b_i \cos(\alpha\theta))\,, \tag{2.2}$$

where

$$\begin{aligned} &s_1 = s_3 = 5, \quad s_2 = s_4 = 1 \\ &\alpha = 0.53544095 \\ &a_1 = 0.44721360 \qquad b_1 = 1.00000000 \\ &a_2 = -0.74535599 \qquad b_2 = 2.33333333 \\ &a_3 = -0.94411759 \qquad b_3 = 0.55555556 \\ &a_4 = -2.40170264 \qquad b_4 = -0.48148148 \end{aligned}$$

This choice of parameters was considered, e.g., in Luce and Wohlmuth [2004]; Jiránek et al. [2010]; Carstensen and Merdon [2010]. The corresponding solution u (2.2) is given in the left part of Figure 2.1.

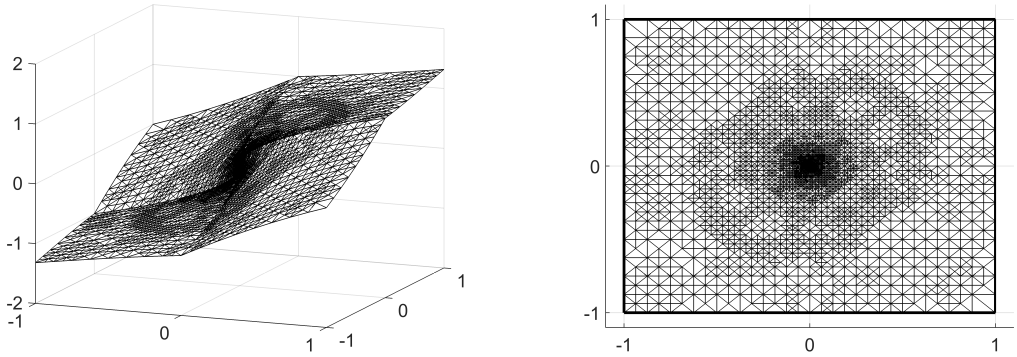


Figure 2.1: Left: the solution (2.2). Right: an example of adaptively refined mesh.

Inhomogeneous tensor problem II. We consider the problem (2.1) with the exact solution u given in polar coordinates (r, θ) by

$$u(r,\theta)|_{\Omega_i} = r^\gamma \mu_i(\theta)\,, \tag{2.3}$$

where

$$\begin{aligned}
s_1 &= s_3 = 161.4476387975881\\
s_2 &= s_4 = 1\\
\gamma &= 0.1\\
\mu_1(\theta) &= \cos((\pi/2-\sigma)\gamma)\cdot\cos((\theta-\pi/2+\rho)\gamma)\\
\mu_2(\theta) &= \cos(\rho\gamma)\cdot\cos((\theta-\pi+\sigma)\gamma)\\
\mu_3(\theta) &= \cos(\sigma\gamma)\cdot\cos((\theta-\pi-\rho)\gamma)\\
\mu_4(\theta) &= \cos((\pi/2-\rho)\gamma)\cdot\cos((\theta-3\pi/2-\sigma)\gamma)\\
\rho &= \pi/4\\
\sigma &= -14.92256510455152
\end{aligned}$$

see [Morin et al., 2002, Section 5.3]. The solution u (2.3) is given in the left part of Figure 2.2; the figure is rotated in comparison with the previous one.

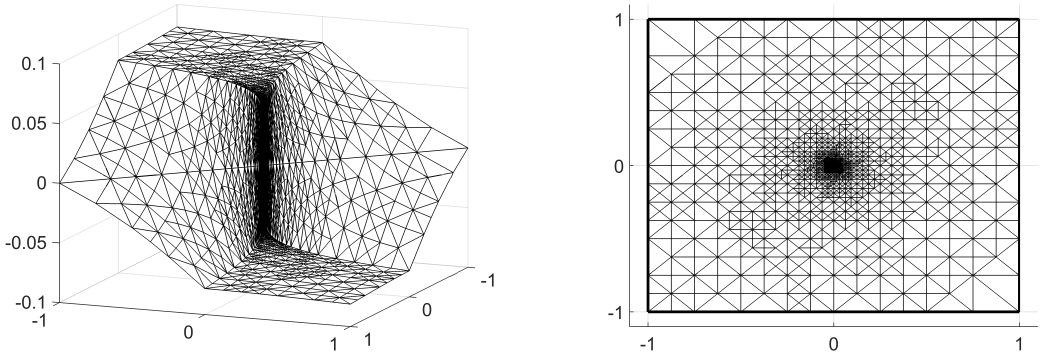


Figure 2.2: Left: the solution (2.3). Right: an example of adaptively refined mesh.

The initial mesh is the same for both model problems I and II and it consists of 128 congruent triangles. The setting of the AFEM procedure generating sequences of the adaptively refined meshes and the notation is adopted from [Papež et al., 2014, Section 5]. The residual-based local error estimator for the problem (2.1) reads as (see, e.g., [Carstensen and Merdon, 2010, Section 2])

$$\eta^2_{R,T}(u_h^*) \equiv \frac{h_T^2}{s_T}\|f\|^2_{L^2(T)} + \sum_{E\subset\partial T}\frac{h_E}{s_E}\|[\mathbf{S}\nabla u_h^*\cdot n_E]\|^2_{L^2(E)}\,, \tag{2.4}$$

where $s_T = s_i$ for $T \subset \Omega_i$, and $s_E \equiv \max\{s_T \mid E \in \partial T\}$.

Recall that $\mathbf{x}^*$ denotes the MATLAB backslash approximation to the solution of the algebraic system corresponding to the discretization of (2.1) on the given mesh using piecewise affine finite elements and u_h^* stands for the piecewise

affine approximation determined by $\mathbf{x}^*$. As in [Papež et al., 2014, Section 5], MATLAB backslash operator gives for our experiments sufficiently accurate approximations and, neglecting the associated numerical error, we will identify u_h^* with the Galerkin FEM solution.

For numerical illustrations we consider (for each model problem) three adaptively refined meshes with the number of vertices exceeding 3000, 10 000, and 20 000, respectively. For the corresponding discrete algebraic systems we apply the (unpreconditioned) conjugate gradient method with zero initial guess. We stop the iteration when the k-th CG approximation $\mathbf{x}_{CG(k)}$ satisfies

$$\|\mathbf{x}^* - \mathbf{x}_{CG(k)}\|_{\mathbf{A}}^2 < 0.01\, \|\mathbf{S}^{1/2}\nabla(u - u_h^*)\|^2 \tag{2.5}$$

and denote by u_h^{CG} the piecewise affine function corresponding to the coefficient vector $\mathbf{x}_{CG(k)}$. We also apply the aggregation-based algebraic multigrid (AGMG, Notay [2010, 2012]; Napov and Notay [2012]) using the MATLAB implementation Notay [2010–2013] with the default choice of parameters. This means that the zero initial guess is used, maximal number of (multigrid) iterations is 100 and the tolerance on the relative residual norm is set to 10^{-6}, i.e. we stop the iteration when the ℓ-th AGMG approximation $\mathbf{x}_{AGMG(\ell)}$ satisfies

$$\frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}_{AGMG(\ell)}\|}{\|\mathbf{b}\|} \leq 10^{-6}\,. \tag{2.6}$$

The piecewise affine approximation determined by the coefficient vector $\mathbf{x}_{AGMG(\ell)}$ is denoted by u_h^{AGMG}.

2.2.1 Results for Inhomogeneous tensor problem I

The problem exhibits singularity at the origin due to the discontinuity of diffusion coefficient. The error $u_h^* - u_h^{CG}$ affects the total error $u - u_h^{CG}$ in most of the domain for all three meshes considered. The error $u_h^* - u_h^{AGMG}$ remains on (or below) the level 10^{-6} and we observe its oscillatory behavior.

Mesh with 3625 nodes

The adaptively refined mesh generated at the 15th step of AFEM has 3625 nodes. The corresponding algebraic system is of the size 3537 with the condition number $\kappa(\mathbf{A}) = 1.39 \times 10^4$. The squared energy norms of the errors are

$$\begin{aligned}
\|\mathbf{S}^{1/2}\nabla(u - u_h^*)\|^2 &= 1.33 \times 10^{-2},\\
\|\mathbf{x}^* - \mathbf{x}_{CG(107)}\|_{\mathbf{A}}^2 &= 1.32 \times 10^{-4},\\
\|\mathbf{x}^* - \mathbf{x}_{AGMG(11)}\|_{\mathbf{A}}^2 &= 1.04 \times 10^{-10}.
\end{aligned}$$

The errors are depicted in Figure 2.3.

Mesh with 10 082 nodes

We consider the mesh with 10 082 nodes generated at the 18th step of AFEM. The corresponding algebraic system is of the size 9927 with the condition number

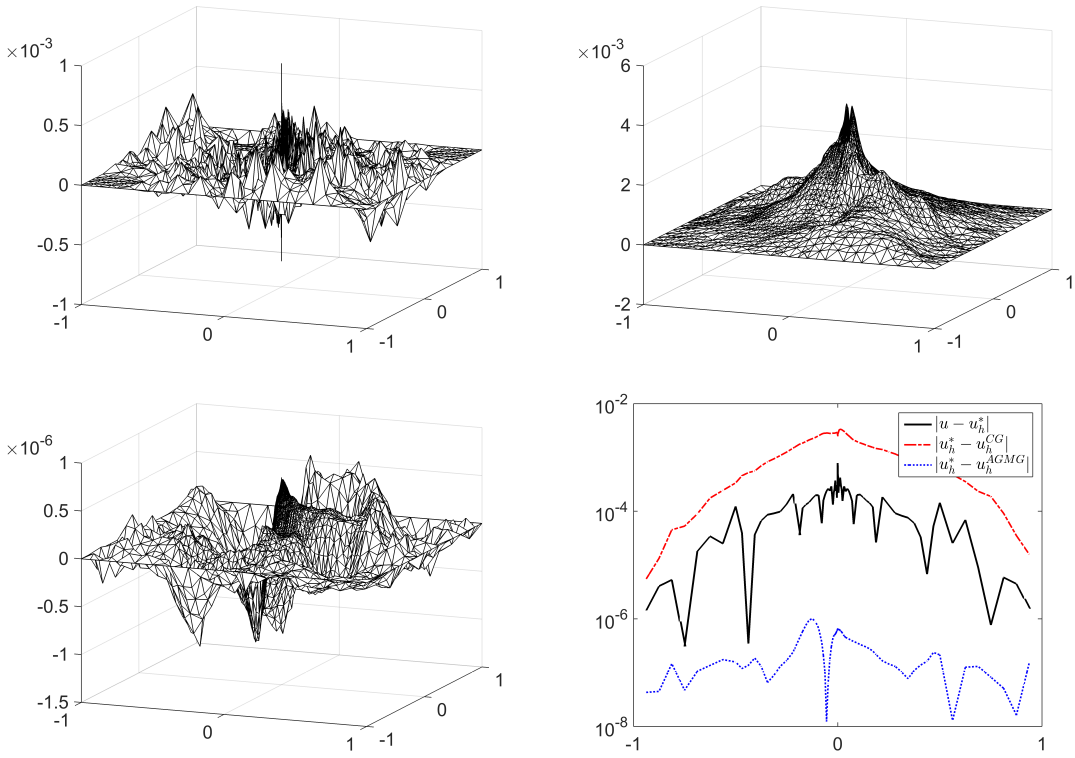


Figure 2.3: Inhomogeneous tensor problem I, mesh with 3625 nodes: discretization error $u - u_h^*$ (upper left), algebraic error in CG $u_h^* - u_h^{CG}$ (upper right), algebraic error in AGMG $u_h^* - u_h^{AGMG}$ (bottom left). Vertical axes are scaled by 10^{-3}, 10^{-3}, and 10^{-6}, respectively. Bottom right: absolute values of errors on the line $y = x$.

$\kappa(\mathbf{A}) = 3.97 \times 10^4$. The squared energy norms of the errors are

$$\begin{aligned} \|\mathbf{S}^{1/2}\nabla(u - u_h^*)\|^2 &= 4.68 \times 10^{-3}, \\ \|\mathbf{x}^* - \mathbf{x}_{CG(208)}\|_{\mathbf{A}}^2 &= 4.54 \times 10^{-5}, \\ \|\mathbf{x}^* - \mathbf{x}_{AGMG(11)}\|_{\mathbf{A}}^2 &= 7.72 \times 10^{-10}, \end{aligned}$$

and the errors are given in Figure 2.4.

Mesh with 25 780 nodes

At the 21st step of AFEM the adaptively refined mesh has 25 780 nodes, the corresponding algebraic system is of the size 25 532 with $\kappa(\mathbf{A}) = 1.01 \times 10^5$ and

$$\begin{aligned} \|\mathbf{S}^{1/2}\nabla(u - u_h^*)\|^2 &= 1.85 \times 10^{-3}, \\ \|\mathbf{x}^* - \mathbf{x}_{CG(469)}\|_{\mathbf{A}}^2 &= 1.85 \times 10^{-5}, \\ \|\mathbf{x}^* - \mathbf{x}_{AGMG(15)}\|_{\mathbf{A}}^2 &= 1.89 \times 10^{-9}. \end{aligned}$$

Because the mesh is too fine for plotting the associated errors over the whole domain, Figure 2.5 depicts only the cross-section error plot over the line $y = x$.

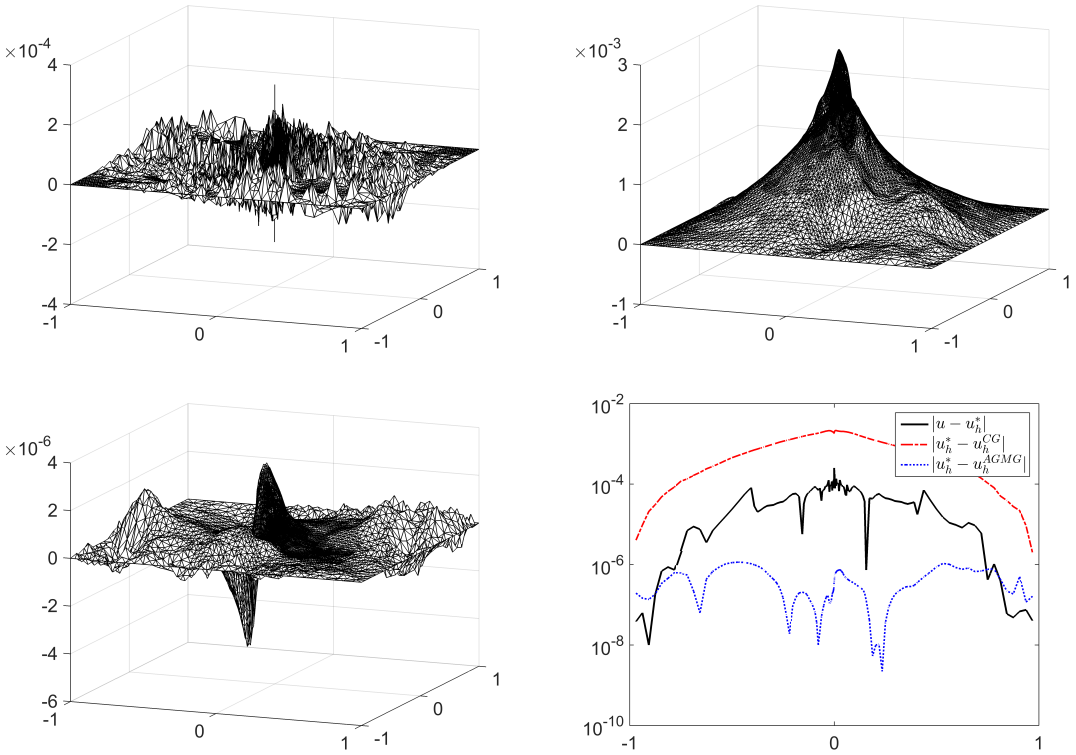


Figure 2.4: Inhomogeneous tensor problem I, mesh with 10 082 nodes: discretization error $u - u_h^*$ (upper left), algebraic error in CG $u_h^* - u_h^{CG}$ (upper right), algebraic error in AGMG $u_h^* - u_h^{AGMG}$ (bottom left). Vertical axes are scaled by 10^{-4}, 10^{-3}, and 10^{-6}, respectively. Bottom right: absolute values of errors on the line $y = x$.

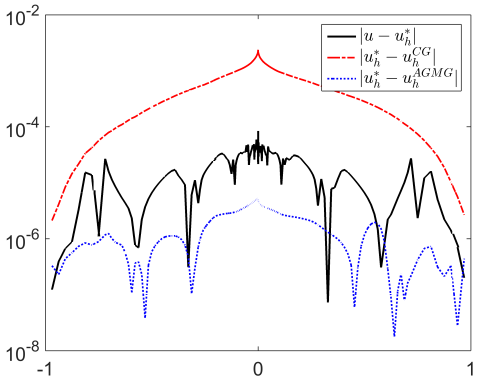


Figure 2.5: Inhomogeneous tensor problem I, mesh with 25 780 nodes: absolute values of errors on the line $y = x$.

2.2.2 Results for Inhomogeneous tensor problem II

The problem exhibits strong singularity at the origin, where the solution (2.3) has a steep gradient. Consequently, the meshes are refined only near the origin, the size of the smallest elements of the mesh with 22 648 nodes (used in the third experiment) is approximately 10^{-13}. The different size of the elements leads to slightly higher condition numbers in comparison with previous problem. However, the conditioning ($\kappa(\mathbf{A}) = 4.81 \times 10^5$ for the mesh with 22 648 nodes) is still moderate.

In the experiments, the discretization error $u - u_h^*$ exhibits strong peak near the origin. In the rest of the domain, $|u - u_h^*|$ is below 10^{-4}. The algebraic error in CG $u_h^* - u_h^{CG}$ is localized in the quadrants where the diffusion tensor is equal to $\mathbf{I}$ and it is several orders of magnitude smaller in the quadrants where diffusion tensor is large. This is a nice demonstration of the fact that CG method minimizes the energy norm of the error. However, the error $u_h^* - u_h^{CG}$ dominates the total error $u - u_h^{CG}$ in most of the domain. The error $u_h^* - u_h^{AGMG}$ grows to the level 10^{-7} which is still in most of the domain smaller than the discretization error.

Mesh with 3247 nodes

The mesh generated at the 35th step of AFEM has 3247 nodes. The corresponding algebraic system is of the size 3200 with the condition number $\kappa(\mathbf{A}) = 6.05 \times 10^4$. The squared energy norms of the errors are

$$\begin{aligned} \|\mathbf{S}^{1/2}\nabla(u - u_h^*)\|^2 &= 4.31 \times 10^{-1}, \\ \|\mathbf{x}^* - \mathbf{x}_{CG(161)}\|_{\mathbf{A}}^2 &= 3.98 \times 10^{-3}, \\ \|\mathbf{x}^* - \mathbf{x}_{AGMG(11)}\|_{\mathbf{A}}^2 &= 4.12 \times 10^{-12}. \end{aligned}$$

The errors are depicted in Figure 2.6.

Mesh with 10 856 nodes

The mesh generated at the 43rd step of AFEM has 10 856 nodes. The corresponding algebraic system is of the size 10 790 with the condition number $\kappa(\mathbf{A}) = 2.27 \times 10^5$. The squared energy norms of the errors are

$$\begin{aligned} \|\mathbf{S}^{1/2}\nabla(u - u_h^*)\|^2 &= 1.51 \times 10^{-1}, \\ \|\mathbf{x}^* - \mathbf{x}_{CG(343)}\|_{\mathbf{A}}^2 &= 1.50 \times 10^{-3}, \\ \|\mathbf{x}^* - \mathbf{x}_{AGMG(11)}\|_{\mathbf{A}}^2 &= 1.09 \times 10^{-11}, \end{aligned}$$

and the errors are given in Figure 2.7.

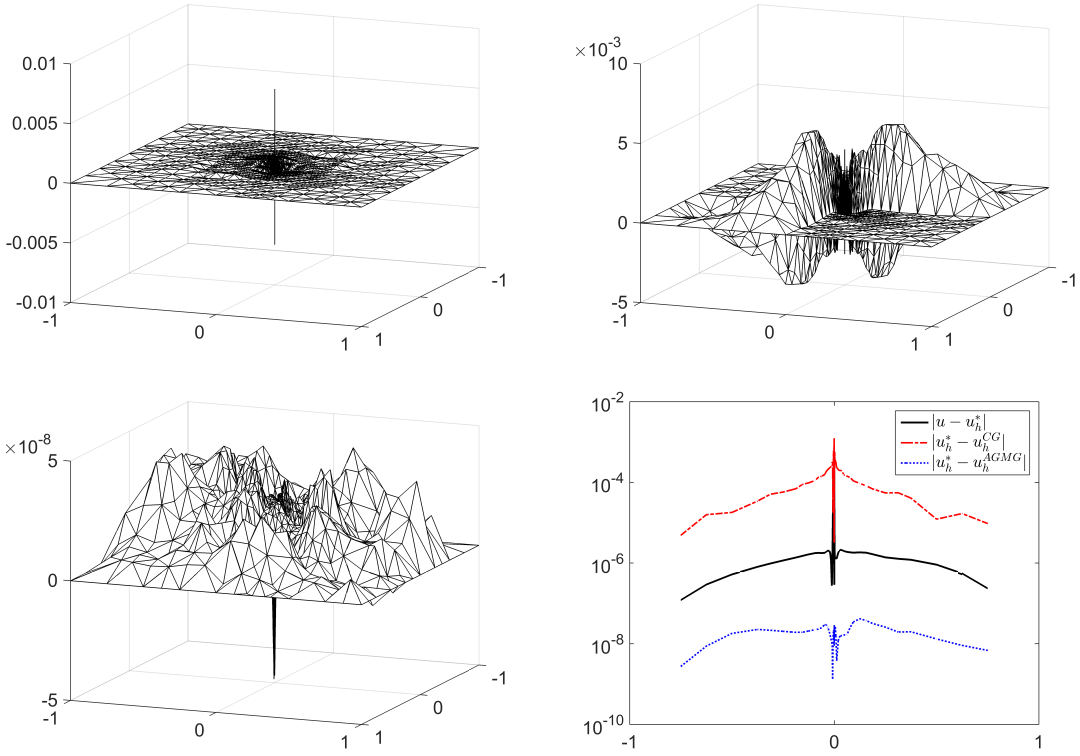


Figure 2.6: Inhomogeneous tensor problem II, mesh with 3247 nodes: discretization error $u - u_h^*$ (upper left), algebraic error in CG $u_h^* - u_h^{CG}$ (upper right), algebraic error in AGMG $u_h^* - u_h^{AGMG}$ (bottom left). Vertical axes are scaled by 10^0, 10^{-3}, and 10^{-8}, respectively. Bottom right: absolute values of errors on the line $y = -x$.

Mesh with 22 648 nodes

At the 48th step of AFEM the adaptively refined mesh has 22 648 nodes, the corresponding algebraic system is of the size 22 538 with $\kappa(\mathbf{A}) = 4.81 \times 10^5$ and

$$\begin{aligned} \|\mathbf{S}^{1/2}\nabla(u - u_h^*)\|^2 &= 7.64 \times 10^{-2}, \\ \|\mathbf{x}^* - \mathbf{x}_{CG(520)}\|_{\mathbf{A}}^2 &= 7.42 \times 10^{-4}, \\ \|\mathbf{x}^* - \mathbf{x}_{AGMG(11)}\|_{\mathbf{A}}^2 &= 2.82 \times 10^{-11}. \end{aligned}$$

The mesh is too fine for plotting the associated errors over the whole domain, and therefore Figure 2.8 depicts only the cross-section error plot over the line $y = -x$.

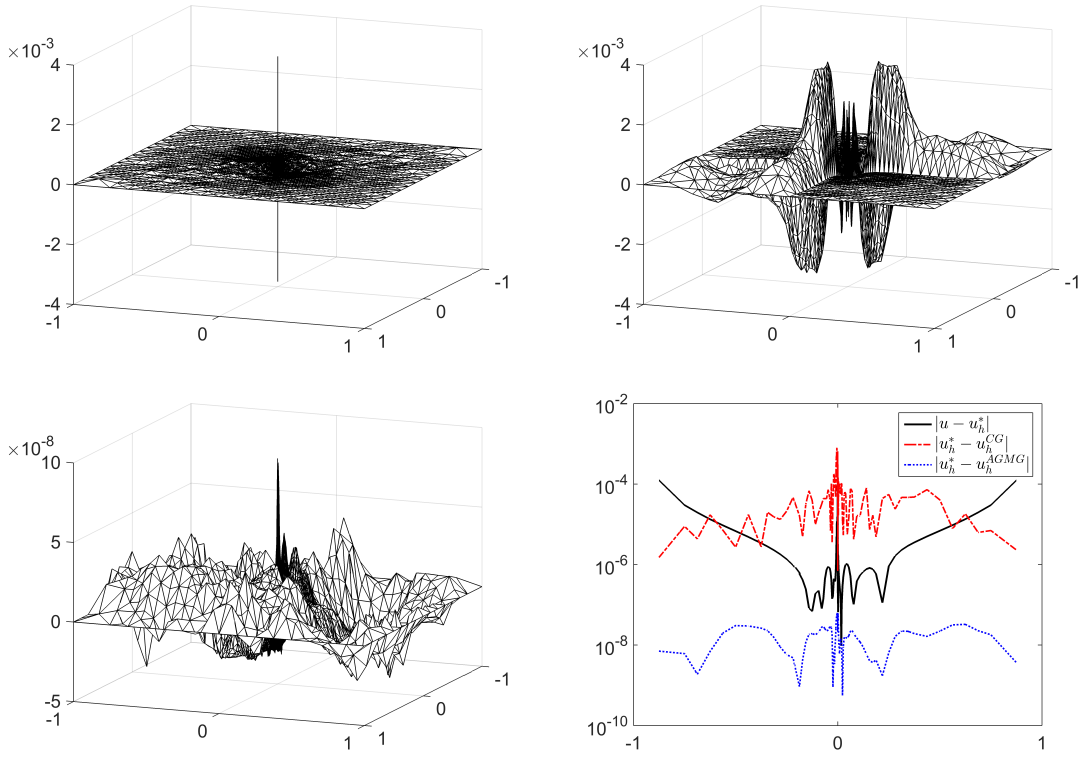


Figure 2.7: Inhomogeneous tensor problem II, mesh with 10 856 nodes: discretization error $u - u_h^*$ (upper left), algebraic error in CG $u_h^* - u_h^{CG}$ (upper right), algebraic error in AGMG $u_h^* - u_h^{AGMG}$ (bottom left). Vertical axes are scaled by 10^{-3}, 10^{-3}, and 10^{-8}, respectively. Bottom right: absolute values of errors on the line $y = -x$.

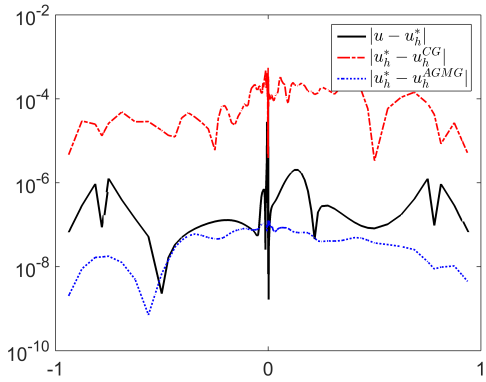


Figure 2.8: Inhomogeneous tensor problem II, mesh with 22 648 nodes: absolute values of errors on the line $y = -x$.

Bibliography

C. Carstensen and C. Merdon. Estimator competition for Poisson problems. *J. Comput. Math.*, 28(3):309–330, 2010. ISSN 0254-9409.

P. Jiránek, Z. Strakoš, and M. Vohralík. A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. *SIAM J. Sci. Comput.*, 32(3):1567–1590, 2010. ISSN 1064-8275.

R. Luce and B. I. Wohlmuth. A local a posteriori error estimator based on equilibrated fluxes. *SIAM J. Numer. Anal.*, 42(4):1394–1414, 2004. ISSN 0036-1429.

P. Morin, R. H. Nochetto, and K. G. Siebert. Convergence of adaptive finite element methods. *SIAM Rev.*, 44(4):631–658 (2003), 2002. ISSN 0036-1445. Revised reprint of "Data oscillation and convergence of adaptive FEM" [SIAM J. Numer. Anal. **38** (2000), no. 2, 466–488].

A. Napov and Y. Notay. An algebraic multigrid method with guaranteed convergence rate. *SIAM J. Sci. Comput.*, 34(2):A1079–A1109, 2012. ISSN 1064-8275.

Y. Notay. An aggregation-based algebraic multigrid method. *Electron. Trans. Numer. Anal.*, 37:123–146, 2010. ISSN 1068-9613.

Y. Notay. AGMG software and documentation. http://homepages.ulb.ac.be/ ynotay/AGMG, 2010–2013. ver. 3.2.0-aca.

Y. Notay. Aggregation-based algebraic multigrid for convection-diffusion equations. *SIAM J. Sci. Comput.*, 34(4):A2288–A2316, 2012. ISSN 1064-8275.

J. Papež, J. Liesen, and Z. Strakoš. Distribution of the discretization and algebraic error in numerical solution of partial differential equations. *Linear Algebra Appl.*, 449:89–114, 2014. ISSN 0024-3795.

3. Backward error interpretation in numerical PDEs

Backward error interprets the inaccuracies in the solution process as a meaningful modification of the mathematical model. So-called functional backward error by Arioli and others (see, e.g., Arioli et al. [2001]) interprets the errors as (backward) perturbations of the weak formulation of the problem. In Nordbotten and Bjørstad [2008]; Keilegavlen and Nordbotten [2015], the error is interpreted as the perturbation of the diffusion tensor, which is in the applications considered therein a subject of uncertainty. These ideas are appealing in more complicated settings where such perturbations represent a modification of the mathematical model that has some physical interpretation. Within the simple problem setting considered in the thesis, we will proceed in a different way.

The algebraic backward error is a standard tool in error analysis that interprets inaccuracies in the solution of an algebraic problem as a perturbation of the data defining the problem; see, e.g., [Higham, 2002, Chapter 7]. We relate the algebraic backward error with transformation of the discretization basis in Section 3.2 and with a modification of the so-called Green's function in Section 3.3. We further use the algebraic backward error for estimating the algebraic forward error via the Fréchet derivative of the matrix inversion; see Section 3.4. The aim of these interpretation is to provide a new perspective to estimating the algebraic error, especially in the context of solving boundary value problems. We are aware that the interpretations in their present form seem not to be beneficial from a computational viewpoint because of their evaluation cost.

Consider a problem given in the weak form: Find $u \in V$ such that

$$a(u, v) = \ell(v) \qquad \forall v \in V, \tag{3.1}$$

where V is a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a bilinear form, and $\ell(\cdot) : V \to \mathbb{R}$ is a linear bounded functional on V. We assume that (3.1) admits the unique solution u. We keep the exposition as general as possible; however, in some parts of the chapter we restrict ourselves to a linear second order elliptic PDEs that generate self-adjoint operators, i.e., to $a(\cdot,\cdot)$ symmetric, bounded and V-elliptic.

Let now $V_h \subset V$ be a finite-dimensional subspace. The function $u_h \in V_h$ satisfying

$$a(u_h, v_h) = \ell(v_h) \qquad \forall v_h \in V_h \tag{3.2}$$

is called the *Galerkin solution* of (3.1). Let $\Phi = \{\phi_1, \dots, \phi_N\}$ be a basis of V_h. The coefficients $\mathbf{x}$ of the Galerkin solution $u_h = \Phi\mathbf{x} = \sum_{i=1}^{N} (\mathbf{x})_i \, \phi_i$ with respect to the basis Φ are given as the solution of the linear algebraic system

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,, \tag{3.3}$$

where

$$\begin{aligned} (\mathbf{A})_{ij} &= a(\phi_j, \phi_i)\,, \\ (\mathbf{b})_i &= \ell(\phi_i)\,, \qquad i, j = 1, \dots, N\,. \end{aligned}$$

Throughout the chapter we assume that $\mathbf{A}$ is non-singular. For $a(\cdot,\cdot)$ symmetric, bounded, and V-elliptic, the matrix $\mathbf{A}$ is symmetric positive definite (SPD).

The ideas of the chapter were partially developed in discussions with Zlatko Drmač, Jörg Liesen, Samuel Relton, Zdeněk Strakoš, Mattia Tani and Tomáš Vejchodský.

3.1 Algebraic backward error

Given an approximation $\widehat{\mathbf{x}}$ to the solution $\mathbf{x}$ of (3.3), in (algebraic) backward error analysis we look for the perturbations $\mathbf{E}$ and $\mathbf{f}$, smallest in a certain sense, such that $\widehat{\mathbf{x}}$ satisfies

$$(\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b}+\mathbf{f}\,. \tag{3.4}$$

The most familiar are the *normwise relative backward error*

$$\min_{\mathbf{E},\mathbf{f}} \{\epsilon \mid (\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b}+\mathbf{f},\ \|\mathbf{E}\| \le \epsilon\|\mathbf{A}\|,\ \|\mathbf{f}\| \le \epsilon\|\mathbf{b}\|\}$$

that can be analogously defined also for other norms, and the *componentwise relative backward error*

$$\min_{\mathbf{E},\mathbf{f}} \{\epsilon \mid (\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b}+\mathbf{f},\ |\mathbf{E}| \le \epsilon|\mathbf{A}|,\ |\mathbf{f}| \le \epsilon|\mathbf{b}|\}\,,$$

where $|\cdot|$ stands for the (elementwise) absolute value. The minimizers of these relative backward errors are presented in Rigal and Gaches [1967], respectively in Oettli and Prager [1964]. We will also consider the special case of (3.4) with $\mathbf{f}=0$, i.e. we look for the perturbation $\mathbf{E}$, such that

$$(\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b}\,. \tag{3.5}$$

Throughout the chapter we will refer several times to the following backward error perturbations. The minimizer of the normwise relative backward error is

$$\frac{\widehat{\mathbf{r}}\,\widehat{\mathbf{x}}^T}{\|\widehat{\mathbf{x}}\|^2} = \arg\min_{\mathbf{E}} \{\epsilon \mid (\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b},\ \|\mathbf{E}\| \le \epsilon\|\mathbf{A}\|\}\,, \tag{3.6}$$

where $\widehat{\mathbf{r}}$ is the residual $\mathbf{b}-\mathbf{A}\widehat{\mathbf{x}}$. For a symmetric positive definite $\mathbf{A}$, we have

$$\frac{\widehat{\mathbf{r}}\,\widehat{\mathbf{x}}^T\mathbf{A}}{\|\widehat{\mathbf{x}}\|_{\mathbf{A}}^2} = \arg\min_{\mathbf{E}} \{\epsilon \mid (\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b},\ \|\mathbf{E}\|_{\mathbf{A}} \le \epsilon\|\mathbf{A}\|_{\mathbf{A}}\}\,; \tag{3.7}$$

see Rigal and Gaches [1967]. The rank-1 minimizers in (3.6), (3.7) are generally nonsymmetric. Following [Bunch et al., 1989, Theorem 2], the symmetric perturbation matrix that has the minimal Frobenius norm among all symmetric perturbation matrices satisfying (3.5) with a symmetric matrix $\mathbf{A}$ is

$$\frac{\widehat{\mathbf{r}}\,\widehat{\mathbf{x}}^T + \widehat{\mathbf{x}}\,\widehat{\mathbf{r}}^T}{\|\widehat{\mathbf{x}}\|^2} - \frac{(\widehat{\mathbf{x}}^T\widehat{\mathbf{r}})}{\|\widehat{\mathbf{x}}\|^4}\widehat{\mathbf{x}}\,\widehat{\mathbf{x}}^T =$$
$$\arg\min_{\mathbf{E}} \left\{\epsilon \mid (\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b},\ \mathbf{A}=\mathbf{A}^T,\ \mathbf{E}=\mathbf{E}^T,\ \|\mathbf{E}\|_F \le \epsilon\|\mathbf{A}\|_F\right\}. \tag{3.8}$$

The standard use of the algebraic backward error consists in bounding the norm of the forward error $\mathbf{x}-\widehat{\mathbf{x}}$. We now present an example of such bound; see, e.g., [Higham, 2002, Section 7.1]. Consider (3.4) with $\|\mathbf{E}\| \le \epsilon\|\mathbf{A}\|$, $\|\mathbf{f}\| \le \epsilon\|\mathbf{b}\|$ and assume that $\epsilon\|\mathbf{A}^{-1}\|\,\|\mathbf{A}\| < 1$. Then simple manipulations show that

$$\begin{aligned}\mathbf{x}-\widehat{\mathbf{x}} &= \mathbf{A}^{-1}\big(\mathbf{E}\mathbf{x}-\mathbf{f}-\mathbf{E}(\mathbf{x}-\widehat{\mathbf{x}})\big),\\ \|\mathbf{x}-\widehat{\mathbf{x}}\| &\le \|\mathbf{A}^{-1}\|\big(\|\mathbf{E}\|\,\|\mathbf{x}\|+\|\mathbf{f}\|+\|\mathbf{E}\|\,\|\mathbf{x}-\widehat{\mathbf{x}}\|\big),\\ &\le \epsilon\|\mathbf{A}^{-1}\|\big(\|\mathbf{A}\|\,\|\mathbf{x}\|+\|\mathbf{b}\|+\|\mathbf{A}\|\,\|\mathbf{x}-\widehat{\mathbf{x}}\|\big),\end{aligned}$$

and

$$\|\mathbf{x}-\widehat{\mathbf{x}}\| \le \frac{\epsilon}{1-\epsilon\|\mathbf{A}^{-1}\|\,\|\mathbf{A}\|}\left(\|\mathbf{A}^{-1}\|\,\|\mathbf{A}\|\,\|\mathbf{x}\|+\|\mathbf{A}^{-1}\|\,\|\mathbf{b}\|\right).$$

Using $\|\mathbf{b}\| = \|\mathbf{A}\mathbf{x}\| \le \|\mathbf{A}\|\|\mathbf{x}\|$ and denoting by $\kappa(\mathbf{A}) \equiv \|\mathbf{A}^{-1}\|\,\|\mathbf{A}\|$, we have

$$\frac{\|\mathbf{x}-\widehat{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \frac{2\epsilon\kappa(\mathbf{A})}{1-\epsilon\kappa(\mathbf{A})}.$$

However, the highly desirable *componentwise* forward error analysis that would allow to bound the individual components of the algebraic error and, consequently, to estimate the distribution of the error in the solution of the original PDE problem over the domain is not available.

3.2 Backward error and transformation of the discretization bases

In this section we interpret the perturbations $\mathbf{E}, \mathbf{f}$ (constructed for a given vector $\widehat{\mathbf{x}}$) as transformations of the discretization basis Φ, such that the system (3.4) corresponds to the algebraic formulation of the problem (3.2) with transformed discretization basis and test functions; cf. [Gratton et al., 2013, Section 3] and [Papež et al., 2014, Section 3].

Consider a transformed discretization basis functions $\Psi = \{\psi_1, \dots, \psi_N\} \subset V_h$, with the transformation from the original basis functions $\phi_1, \dots, \phi_N$ given by a square matrix $\mathbf{D}$,

$$\psi_j = \phi_j + \sum_{\ell=1}^{N} (\mathbf{D})_{\ell j}\phi_\ell\,, \quad j = 1, \dots, N\,, \qquad \Psi = \Phi(\mathbf{I}+\mathbf{D})\,, \tag{3.9}$$

and the transformed test functions $\mathcal{X} = \{\chi_1, \dots, \chi_N\}$ with the transformation given by a square matrix $\mathbf{G}$,

$$\chi_i = \phi_i + \sum_{k=1}^{N} (\mathbf{G})_{ki}\phi_k\,, \quad i = 1, \dots, N\,, \qquad \mathcal{X} = \Phi(\mathbf{I}+\mathbf{G})\,.$$

It is natural to assume that $\mathbf{I}+\mathbf{D}$ and $\mathbf{I}+\mathbf{G}$ are nonsingular, so that Ψ and $\mathcal{X}$ are bases of V_h.

Given a vector $\mathbf{y}$, let the Galerkin solution u_h of (3.2) satisfy

$$u_h = \Phi\mathbf{x} = \Psi\mathbf{y} = \Phi(\mathbf{I}+\mathbf{D})\mathbf{y} \tag{3.10}$$

for some (unknown) matrix $\mathbf{D}$. Consider the Galerkin discretization (3.2) with the basis functions Ψ, $u_h = \Psi\mathbf{y}$, and with the test functions $\mathcal{X}$ determined by some matrix $\mathbf{G}$. This results in the system of linear algebraic equations

$$\overline{\mathbf{A}}\mathbf{y} = \overline{\mathbf{b}}\,, \tag{3.11}$$

where

$$\begin{aligned}(\overline{\mathbf{A}})_{ij} &= a(\psi_j, \chi_i) = a(\phi_j + \sum_{\ell=1}^{N}(\mathbf{D})_{\ell j}\phi_\ell\,,\ \phi_i + \sum_{k=1}^{N}(\mathbf{G})_{ki}\phi_k)\\ &= a(\phi_j,\phi_i) + a(\sum_{\ell=1}^{N}(\mathbf{D})_{\ell j}\phi_\ell\,,\ \phi_i) + a(\phi_j,\ \sum_{k=1}^{N}(\mathbf{G})_{ki}\phi_k)\\ &\quad +\ a(\sum_{\ell=1}^{N}(\mathbf{D})_{\ell j}\phi_\ell\,,\ \sum_{k=1}^{N}(\mathbf{G})_{ki}\phi_k)\\ &= (\mathbf{A})_{ij} + \sum_{\ell=1}^{N}(\mathbf{A})_{i\ell}(\mathbf{D})_{\ell j} + \sum_{k=1}^{N}(\mathbf{G})_{ki}(\mathbf{A})_{kj} + \sum_{\ell=1}^{N}\sum_{k=1}^{N}(\mathbf{G})_{ki}(\mathbf{A})_{k\ell}(\mathbf{D})_{\ell j}\,,\end{aligned}$$

i.e.

$$\overline{\mathbf{A}} = \mathbf{A} + \mathbf{A}\mathbf{D} + \mathbf{G}^T\mathbf{A} + \mathbf{G}^T\mathbf{A}\mathbf{D} = (\mathbf{I}+\mathbf{G})^T\mathbf{A}(\mathbf{I}+\mathbf{D})\,.$$

For the right-hand side,

$$(\overline{\mathbf{b}})_i = \ell(\chi_i) = \ell(\phi_i + \sum_{k=1}^{N}(\mathbf{G})_{ki}\phi_k) = (\mathbf{b})_i + \sum_{k=1}^{N}(\mathbf{G})_{ki}(\mathbf{b})_k\,,$$

i.e.

$$\overline{\mathbf{b}} = (\mathbf{I}+\mathbf{G})^T\mathbf{b}\,.$$

Given an approximation $\widehat{\mathbf{x}}$ to $\mathbf{x}$ and perturbations $\mathbf{E}, \mathbf{f}$ satisfying (3.4), we will associate the *exact* solution $\widehat{\mathbf{x}}$ of (3.4) with the Galerkin solution u_h of (3.2). For this purpose we need to determine transformation matrices $\mathbf{D}$, $\mathbf{G}$ by setting

$$\mathbf{A} + \mathbf{E} = \overline{\mathbf{A}} = (\mathbf{I}+\mathbf{G})^T\mathbf{A}(\mathbf{I}+\mathbf{D})\,, \tag{3.12}$$

$$\mathbf{b} + \mathbf{f} = \overline{\mathbf{b}} = (\mathbf{I}+\mathbf{G})^T\mathbf{b}\,, \tag{3.13}$$

with $\widehat{\mathbf{x}} = \mathbf{y}$ (see (3.10), (3.11)) the exact algebraic solution of (3.4) representing the coefficients of the Galerkin solution u_h of (3.2). By rearranging (3.12) and (3.13),

$$\mathbf{E} = \mathbf{A}\mathbf{D} + \mathbf{G}^T\mathbf{A} + \mathbf{G}^T\mathbf{A}\mathbf{D}\,, \tag{3.14}$$

$$\mathbf{f} = \mathbf{G}^T\mathbf{b}\,. \tag{3.15}$$

Note that if (3.4), (3.12) and (3.13) hold, then

$$(\mathbf{I}+\mathbf{D})\widehat{\mathbf{x}} = \mathbf{x}, \qquad \text{or equivalently} \qquad \mathbf{D}\widehat{\mathbf{x}} = \mathbf{x} - \widehat{\mathbf{x}}\,. \tag{3.16}$$

We now focus on the existence and the uniqueness of transformations $\mathbf{D}, \mathbf{G}$ satisfying (3.12) and (3.13) for the given perturbations $\mathbf{E}, \mathbf{f}$ such that

- $\mathbf{E}, \mathbf{f}$ are general perturbations, or
- $\mathbf{f} = 0$, or
- $\mathbf{A}$ and $\mathbf{A} + \mathbf{E}$ are symmetric positive definite.

3.2.1 General perturbations

Given $\widehat{\mathbf{x}}$ and $\mathbf{E}, \mathbf{f}$ satisfying (3.4), there are infinitely many couples of matrices $\mathbf{D}, \mathbf{G}$ satisfying (3.12)–(3.13). Indeed, $\mathbf{G}$ can be any matrix such that $\mathbf{G}^T\mathbf{b} = \mathbf{f}$ and $(\mathbf{I} + \mathbf{G})$ is invertible. Once such $\mathbf{G}$ is chosen, from (3.14) we get

$$\mathbf{D} = \mathbf{A}^{-1} \left(\mathbf{I} + \mathbf{G}^T\right)^{-1} \left(\mathbf{E} - \mathbf{G}^T\mathbf{A}\right).$$

Given a norm $\|\cdot\|_\square$ in the matrix space $\mathbb{R}^{N\times N}$, $\mathbf{G}$ can be taken as a matrix with the minimum norm mapping $\mathbf{b}$ to $\mathbf{f}$, i.e.

$$\mathbf{G}^T = \arg\min_{\mathbf{H}} \{\|\mathbf{H}\|_\square \mid \mathbf{H}\,\mathbf{b} = \mathbf{f}\}.$$

The formula for the minimizer will be given in Section 3.2.5; see Theorem 3.2.

Following the discussion in [Papež et al., 2014, Remark 3 on p. 100], we note that the inverse $\mathbf{A}^{-1}$ is, in general, dense, and therefore we have no control on the sparsity of the transformation matrix $\mathbf{D}$. Consequently, the transformed basis functions $\Psi = \Phi(\mathbf{I} + \mathbf{D})$ have global supports even for the locally supported functions represented by Φ.

3.2.2 Perturbations with $\mathbf{f} = 0$

In the special case when the right-hand side is considered unchanged, i.e. $\mathbf{f} = 0$, it is natural to set $\mathbf{G} = 0$ and

$$\mathbf{A} + \mathbf{E} = \mathbf{A}\,(\mathbf{I} + \mathbf{D})\,.$$

The unique matrix $\mathbf{D}$ that satisfies this equality is $\mathbf{D} = \mathbf{A}^{-1}\mathbf{E}$. This case is considered in Gratton et al. [2013]; Papež et al. [2014], where the inexact solution of the algebraic system arising from the discretization is interpreted as transformation of the discretization basis functions with the test functions unchanged, $\mathcal{X} = \Phi$.

For the rank-1 perturbation matrices $\mathbf{E}$ given by (3.6) or (3.7), the transformation matrix $\mathbf{D} = \mathbf{A}^{-1}\mathbf{E}$ is also rank-1 with the columns equal to the scaled algebraic error,

$$\mathbf{D} = \frac{(\mathbf{x} - \widehat{\mathbf{x}})\,\widehat{\mathbf{x}}^T}{\|\widehat{\mathbf{x}}\|^2} \qquad \text{and} \qquad \mathbf{D} = \frac{(\mathbf{x} - \widehat{\mathbf{x}})\,\widehat{\mathbf{x}}^T\mathbf{A}}{\|\widehat{\mathbf{x}}\|_{\mathbf{A}}^2}\,, \tag{3.17}$$

respectively. For a symmetric perturbation $\mathbf{E} = \mathbf{E}^T$, e.g. given by (3.8), the associated transformation matrix $\mathbf{D}$ is in general not symmetric (see the numerical examples in Section 3.5), but, assuming that $\mathbf{A}$ is SPD, it is symmetric with respect to the inner product induced by $\mathbf{A}$,

$$(\mathbf{D}\mathbf{x}, \mathbf{y})_{\mathbf{A}} = \mathbf{y}^T\mathbf{A}\mathbf{D}\mathbf{x} = \mathbf{y}^T\mathbf{E}\mathbf{x} = \mathbf{y}^T\mathbf{E}^T\mathbf{x} = \mathbf{y}^T\mathbf{D}^T\mathbf{A}\mathbf{x} = (\mathbf{x}, \mathbf{D}\mathbf{y})_{\mathbf{A}}\,.$$

3.2.3 Perturbations preserving symmetry

In this section we restrict ourselves to the case where the bilinear form $a(\cdot,\cdot)$ in (3.1) is symmetric, bounded and V-elliptic, and therefore $\mathbf{A}$ in (3.3) is a symmetric positive definite matrix. For $\mathbf{A}$ SPD it is natural to consider perturbations $\mathbf{E}$

such that $\mathbf{A}+\mathbf{E}$ remains also SPD. In order to preserve symmetry, it makes sense to set $\mathbf{G} = \mathbf{D}$, i.e. to have the same discretization and test bases $\Psi = \mathcal{X}$, and search for $\mathbf{D}$ satisfying

$$(\mathbf{I}+\mathbf{D})^T\,\mathbf{A}\,(\mathbf{I}+\mathbf{D}) = \mathbf{A}+\mathbf{E}\,, \tag{3.18}$$

$$(\mathbf{I}+\mathbf{D})^T\mathbf{b} = \mathbf{b}+\mathbf{f}\,. \tag{3.19}$$

As we show below, the existence of $\mathbf{D}$ satisfying (3.18)–(3.19) is not guaranteed.

Multiplying (3.18) by $(\mathbf{A}+\mathbf{E})^{-1/2}$ from the left and from the right gives

$$(\mathbf{A}+\mathbf{E})^{-1/2}(\mathbf{I}+\mathbf{D})^T\mathbf{A}(\mathbf{I}+\mathbf{D})(\mathbf{A}+\mathbf{E})^{-1/2} = \mathbf{I}.$$

Denoting $\mathbf{U} \equiv (\mathbf{A}+\mathbf{E})^{-1/2}(\mathbf{I}+\mathbf{D})^T\mathbf{A}^{1/2}$, this gives $\mathbf{U}\mathbf{U}^T = \mathbf{I}$, i.e. $\mathbf{U}$ must be orthogonal. Equivalently, $\mathbf{D}$ satisfies (3.18) if and only if

$$\mathbf{I}+\mathbf{D} = \mathbf{A}^{-1/2}\mathbf{U}^T(\mathbf{A}+\mathbf{E})^{1/2}\,,$$

where $\mathbf{U}$ is any orthogonal matrix. Plugging the expression for $\mathbf{I}+\mathbf{D}$ into (3.19) gives

$$(\mathbf{A}+\mathbf{E})^{1/2}\mathbf{U}\mathbf{A}^{-1/2}\mathbf{b} = \mathbf{b}+\mathbf{f}.$$

Multiplying by $(\mathbf{A}+\mathbf{E})^{-1/2}$ from the left,

$$\mathbf{U}\mathbf{A}^{-1/2}\mathbf{b} = (\mathbf{A}+\mathbf{E})^{-1/2}\,(\mathbf{b}+\mathbf{f})\,.$$

Clearly, an orthogonal matrix $\mathbf{U}$ satisfying the last equation exists if and only if the vectors $\mathbf{A}^{-1/2}\mathbf{b}$ and $(\mathbf{A}+\mathbf{E})^{-1/2}(\mathbf{b}+\mathbf{f})$ have the same Euclidean norm, i.e., if and only if

$$\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} = (\mathbf{b}+\mathbf{f})^T(\mathbf{A}+\mathbf{E})^{-1}(\mathbf{b}+\mathbf{f})\,.$$

This is equivalent to

$$\mathbf{b}^T\mathbf{x} = (\mathbf{b}+\mathbf{f})^T\widehat{\mathbf{x}}\,, \quad \text{or, equivalently,} \quad \|\mathbf{x}\|_{\mathbf{A}} = \|\widehat{\mathbf{x}}\|_{\mathbf{A}+\mathbf{E}}\,,$$

which, in general, does not hold. Moreover, existence of $\mathbf{D}$ satisfying (3.18)–(3.19) for given perturbations $\mathbf{E}, \mathbf{f}$ cannot be verified without knowledge of $\mathbf{b}^T\mathbf{x}$ or $\|\mathbf{x}\|_{\mathbf{A}}$.

If, on the other hand, there exists a matrix $\mathbf{D}$ satisfying (3.18) and (3.19), then it is not unique. It is of the form $\mathbf{D} = \mathbf{A}^{-1/2}\mathbf{U}^T(\mathbf{A}+\mathbf{E})^{1/2} - \mathbf{I}$ where $\mathbf{U}$ is an orthogonal matrix satisfying $\mathbf{U}\mathbf{A}^{-1/2}\mathbf{b} = (\mathbf{A}+\mathbf{E})^{-1/2}(\mathbf{b}+\mathbf{f})$.

3.2.4 Symmetric perturbation minimal with respect to the energy norm

In the previous subsection it was shown that not every perturbations $\mathbf{E}, \mathbf{f}$ in (3.4) such that $\mathbf{A}+\mathbf{E}$ is SPD can be interpreted as stemming from the transformation of the bases with $\mathbf{D} = \mathbf{G}$. For a given approximation $\widehat{\mathbf{x}}$ there exists a set of perturbations $\mathbf{E}, \mathbf{f}$ satisfying (3.4) and (3.18)–(3.19). In this subsection we look for a perturbation, minimal in an appropriate sense, over this set. Hereafter, we assume that $\mathbf{A}$ is symmetric positive definite.

In [Gratton et al., 2013, Section 3], the authors propose to measure the perturbation $\mathbf{E}$ in (3.5) using the norm $\|\mathbf{E}\|_{\mathbf{A},\mathbf{A}^{-1}} \equiv \max_{\|\mathbf{v}\|_{\mathbf{A}}=1} \|\mathbf{E}\mathbf{v}\|_{\mathbf{A}^{-1}}$. When the

same transformation of both the discretization and test bases is considered and $\mathbf{E}$ satisfies (3.18), the quantity $\|\mathbf{E}\|_{\mathbf{A},\mathbf{A}^{-1}}$ can be interpreted as

$$\|\mathbf{E}\|_{\mathbf{A},\mathbf{A}^{-1}} = \max_{\mathbf{v},\mathbf{w}\neq 0} \frac{\left|\mathbf{w}^T\left[(\mathbf{I}+\mathbf{D})\mathbf{A}(\mathbf{I}+\mathbf{D})-\mathbf{A}\right]\mathbf{v}\right|}{\|\mathbf{v}\|_{\mathbf{A}}\|\mathbf{w}\|_{\mathbf{A}}} = \max_{\mathbf{v},\mathbf{w}\neq 0} \frac{|a(\Psi\mathbf{v},\Psi\mathbf{w})-a(\Phi\mathbf{v},\Phi\mathbf{w})|}{\|\mathbf{v}\|_{\mathbf{A}}\|\mathbf{w}\|_{\mathbf{A}}},$$

with $\Psi = \Phi(\mathbf{I}+\mathbf{D})$. Therefore $\|\mathbf{E}\|_{\mathbf{A},\mathbf{A}^{-1}}$ represents the distance between the bilinear forms $a(\Phi\mathbf{v},\Phi\mathbf{w})$ and $a(\Psi\mathbf{v},\Psi\mathbf{w})$ on $\left(\mathbb{R}^N, \|\cdot\|_{\mathbf{A}}\right)$. Following Gratton et al. [2013] we refer to $\|\mathbf{E}\|_{\mathbf{A},\mathbf{A}^{-1}}$ as the *energy norm* of $\mathbf{E}$. Simple algebraic manipulation shows that $\|\mathbf{E}\|_{\mathbf{A},\mathbf{A}^{-1}} = \|\mathbf{A}^{-1/2}\mathbf{E}\mathbf{A}^{-1/2}\|$.

The minimal norm $\|\mathbf{E}\|_{\mathbf{A},\mathbf{A}^{-1}}$ over the set of perturbations $\mathbf{E},\mathbf{f}$ satisfying (3.4) and (3.18)–(3.19) is determined as

$$\min_{\mathbf{D}\in\mathcal{D}(\widehat{\mathbf{x}},\mathbf{x}-\widehat{\mathbf{x}})} \|(\mathbf{I}+\mathbf{D})^T\mathbf{A}(\mathbf{I}+\mathbf{D})-\mathbf{A}\|_{\mathbf{A},\mathbf{A}^{-1}},$$

where $\mathcal{D}(\mathbf{u},\mathbf{v}) \equiv \left\{\mathbf{D}\in\mathbb{R}^{N\times N} \,|\, \mathbf{D}\mathbf{u}=\mathbf{v}\right\}$. The minimization problem is solved in the following theorem.

Theorem 3.1. *Let* $\widehat{\mathbf{x}}$ *and* $\mathbf{x}$ *be arbitrary and let* $\mathbf{A}$ *be symmetric positive definite. Define*

$$\mathbf{x}_1 \equiv \mathbf{x} - \frac{\widehat{\mathbf{x}}^T\mathbf{A}\mathbf{x}}{\|\widehat{\mathbf{x}}\|_{\mathbf{A}}^2}\widehat{\mathbf{x}}, \qquad \widehat{\mathbf{x}}_1 \equiv \widehat{\mathbf{x}} - \frac{\widehat{\mathbf{x}}^T\mathbf{A}\mathbf{x}}{\|\mathbf{x}\|_{\mathbf{A}}^2}\mathbf{x}.$$

Providing $\mathbf{x}_1\neq 0$, $\widehat{\mathbf{x}}_1\neq 0$,

$$\mu(\widehat{\mathbf{x}},\mathbf{x}) \equiv \min_{\mathbf{D}\in\mathcal{D}(\widehat{\mathbf{x}},\mathbf{x}-\widehat{\mathbf{x}})} \|(\mathbf{I}+\mathbf{D})^T\mathbf{A}(\mathbf{I}+\mathbf{D})-\mathbf{A}\|_{\mathbf{A},\mathbf{A}^{-1}} = \left|\frac{\|\mathbf{x}\|_{\mathbf{A}}^2}{\|\widehat{\mathbf{x}}\|_{\mathbf{A}}^2}-1\right|, \tag{3.20}$$

and the minimum is achieved for

$$\widehat{\mathbf{D}} \equiv \frac{(\mathbf{x}-\widehat{\mathbf{x}})\widehat{\mathbf{x}}^T\mathbf{A}}{\|\widehat{\mathbf{x}}\|_{\mathbf{A}}^2} + \left(\frac{\widehat{\mathbf{x}}_1}{\|\widehat{\mathbf{x}}_1\|_{\mathbf{A}}} - \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|_{\mathbf{A}}}\right)\frac{\mathbf{x}_1^T\mathbf{A}}{\|\mathbf{x}_1\|_{\mathbf{A}}}. \tag{3.21}$$

Moreover, $\mathbf{I}+\widehat{\mathbf{D}}$ *is nonsingular.*

The proof of Theorem 3.1 is rather technical and it is given in Appendix 3.A. The perturbation $\widehat{\mathbf{E}}$, $\widehat{\mathbf{f}}$ satisfying (3.4) and $\|\widehat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}} = \mu(\widehat{\mathbf{x}},\mathbf{x})$ is given by plugging (3.21) into (3.18) and (3.19).

3.2.5 Minimal transformation of the discretization basis

For transformation of the discretization basis we derived the condition (3.16), i.e. $\mathbf{D}\widehat{\mathbf{x}} = \mathbf{x}-\widehat{\mathbf{x}}$ or, equivalently, $\mathbf{D}\in\mathcal{D}(\widehat{\mathbf{x}},\mathbf{x}-\widehat{\mathbf{x}})$. A natural question is to find a minimal transformation such that (3.16) is satisfied. This is resolved using the following theorem.

Theorem 3.2. *Let* $\mathbf{u}$, $\mathbf{v}$ *be arbitrary, let* $\|\cdot\|_{\square}$ *be a vector norm with the associated matrix norm denoted in the same way. Then the minimizer of* $\|\mathbf{C}\|_{\square}$ *over all matrices* $\mathbf{C}$ *satisfying* $\mathbf{C}\mathbf{u}=\mathbf{v}$, *i.e.,* $\mathbf{C}\in\mathcal{D}(\mathbf{u},\mathbf{v})$, *is given by*

$$\widetilde{\mathbf{C}} \equiv \operatorname*{arg\,min}_{\mathbf{C}\in\mathcal{D}(\mathbf{u},\mathbf{v})} \|\mathbf{C}\|_{\square} = \mathbf{v}\mathbf{z}^T, \tag{3.22}$$

where $\mathbf{z}$ *is a vector dual to* $\mathbf{u}$ *with respect to the norm* $\|\cdot\|_\square$*, i.e.,*

$$\mathbf{z}^T\mathbf{u} = \|\mathbf{z}\|_D\|\mathbf{u}\|_\square = 1, \quad \textit{where} \quad \|\mathbf{z}\|_D = \max_{\mathbf{w}\neq 0} \frac{|\mathbf{z}^T\mathbf{w}|}{\|\mathbf{w}\|_\square}\,. \tag{3.23}$$

Proof. Let $\mathbf{C} \in \mathcal{D}(\mathbf{u},\mathbf{v})$. Then

$$\|\mathbf{C}\|_\square = \max_{\mathbf{w}\neq 0} \frac{\|\mathbf{C}\mathbf{w}\|_\square}{\|\mathbf{w}\|_\square} \geq \frac{\|\mathbf{C}\mathbf{u}\|_\square}{\|\mathbf{u}\|_\square} = \frac{\|\mathbf{v}\|_\square}{\|\mathbf{u}\|_\square}\,.$$

Now, let $\mathbf{z}$ satisfy (3.23). Existence of such $\mathbf{z}$ is guaranteed by the Hahn–Banach theorem; for the proof in finite-dimensional setting see, e.g., [Horn and Johnson, 2013, Theorem 5.5.9]. It holds

$$\|\widetilde{\mathbf{C}}\|_\square = \max_{\mathbf{w}\neq 0} \frac{\|\widetilde{\mathbf{C}}\mathbf{w}\|_\square}{\|\mathbf{w}\|_\square} = \max_{\mathbf{w}\neq 0} \frac{|\mathbf{z}^T\mathbf{w}|}{\|\mathbf{w}\|_\square}\|\mathbf{v}\|_\square = \frac{\|\mathbf{v}\|_\square}{\|\mathbf{u}\|_\square}\,.$$

Hence, $\|\widetilde{\mathbf{C}}\|_\square = \dfrac{\|\mathbf{v}\|_\square}{\|\mathbf{u}\|_\square} \leq \|\mathbf{C}\|_\square$ for any $\mathbf{C} \in \mathcal{D}(\mathbf{u},\mathbf{v})$ and $\widetilde{\mathbf{C}}$ satisfies (3.22). □

In addition to Theorem 3.2, it was shown in [Dennis and Schnabel, 1996, Theorem 8.1.1] that the unique minimizer (3.22) for the Frobenius norm $\|\cdot\|_\square = \|\cdot\|_F$ is given by

$$\widetilde{\mathbf{C}} = \frac{\mathbf{v}\mathbf{u}^T}{\|\mathbf{u}\|^2}\,. \tag{3.24}$$

Since $\mathbf{z} = \mathbf{u}/\|\mathbf{u}\|^2$ is the vector dual to $\mathbf{u}$ with respect to the Euclidean norm $\|\cdot\|$, (3.24) is also the solution of (3.22) when $\|\cdot\|_\square = \|\cdot\|$.

Applying Theorem 3.2 for finding the minimal transformation $\mathbf{D} \in \mathcal{D}(\widehat{\mathbf{x}}, \mathbf{x}-\widehat{\mathbf{x}})$ of the discretization basis is straightforward. Setting $\mathbf{u} = \widehat{\mathbf{x}}$, $\mathbf{v} = \mathbf{x} - \widehat{\mathbf{x}}$ in (3.24) gives

$$\widetilde{\mathbf{D}} = \frac{(\mathbf{x} - \widehat{\mathbf{x}})\widehat{\mathbf{x}}^T}{\|\widehat{\mathbf{x}}\|^2}\,;$$

cf. (3.17). The perturbation $\mathbf{E}$ given by (3.6) therefore satisfies $\mathbf{E} = \mathbf{A}\widetilde{\mathbf{D}}$, which means that it corresponds to the transformation of the discretization basis that has the minimal Euclidean norm and the minimal Frobenius norm among all transformation matrices in $\mathcal{D}(\widehat{\mathbf{x}}, \mathbf{x} - \widehat{\mathbf{x}})$. Similarly, one can show that the perturbation matrix $\mathbf{E}$ given by (3.7) corresponds to the transformation of the basis minimal with respect to the norm $\|\cdot\|_\mathbf{A}$.

Remark: When a nonzero perturbation $\mathbf{f}$ of the right-hand side in (3.4) is given and one looks for the matrix $\widetilde{\mathbf{G}}$ with the minimal norm satisfying $\widetilde{\mathbf{G}}^T\mathbf{b} = \mathbf{f}$ (see Section 3.2.1), Theorem 3.2 gives

$$\widetilde{\mathbf{G}}^T = \operatorname*{arg\,min}_{\mathbf{H}\in\mathcal{D}(\mathbf{b},\mathbf{f})} \|\mathbf{H}\|_\square = \mathbf{f}\,\mathbf{g}^T\,,$$

where $\mathbf{g}$ is a vector dual to $\mathbf{b}$ with respect to the norm $\|\cdot\|_\square$. Moreover,

$$\operatorname*{arg\,min}_{\mathbf{H}\in\mathcal{D}(\mathbf{b},\mathbf{f})} \|\mathbf{H}\|_F = \operatorname*{arg\,min}_{\mathbf{H}\in\mathcal{D}(\mathbf{b},\mathbf{f})} \|\mathbf{H}\| = \mathbf{f}\,\frac{\mathbf{b}^T}{\|\mathbf{b}\|^2}\,.$$

3.3 Inexact discrete Green's function

The Green's function and the discrete Green's function map the right-hand side of the considered PDE to its (weak) solution, respectively to the Galerkin solution of the discretized problem. They are used in literature for proving maximum and discrete maximum principles of the associated operator; see, e.g., [Protter and Weinberger, 1984, Section 7] and Ciarlet [1970]. In this section we introduce the inexact discrete Green's function that maps the right-hand side to the computed approximation and that can be expressed using the algebraic backward error.

We restrict ourselves to the Poisson problem with homogeneous Dirichlet boundary condition given in the weak form (cf. (3.1)): Find $u \in V \equiv H_0^1(\Omega)$ such that

$$\begin{aligned} a(u,v) &= \ell(v) \qquad \forall v \in V, \\ a(u,v) &\equiv \int_\Omega \nabla u \cdot \nabla v \,, \quad \ell(v) \equiv \int_\Omega f v \,, \end{aligned} \tag{3.25}$$

where $f \in L^2(\Omega)$ and Ω is a domain with a Lipschitz boundary. The results can be extended to more general linear second-order elliptic problems imposing further assumptions on the data; see, e.g., [Protter and Weinberger, 1984, Section 7].

By the Lax–Milgram lemma (Lax and Milgram [1954]), the solution $u \in V$ of (3.25) exists and it is unique, and we can define the *Green's operator* $\widetilde{G}\colon L^2(\Omega) \to V$, $\widetilde{G}(f) = u$. The existence, uniqueness and other properties of this operator are described and proved, e.g., in Nečas [1967]. The Green's operator can often (see, e.g., [Protter and Weinberger, 1984, Section 7]) be expressed using the *Kirchhoff–Helmholtz representation*, for $y \in \Omega$,

$$u(y) = \int_\Omega f(x)\, G(x,y)\, \mathrm{d}x \,, \tag{3.26}$$

where the kernel $G(x,y)$ is called the *Green's function*. Existence of the representation (3.26) requires, in general, additional requirement on the regularity of the source term f.

The discrete Green's function is defined in analogy with the infinite-dimensional case. Let V_h be a finite-dimensional subspace of V. We assume that V_h is a subspace of continuous functions, $V_h \subset C(\Omega)$, which holds for all common conforming discretizations of (3.25). We recall that $u_h \in V_h$ denotes the Galerkin solution of (3.25); see (3.2). The *discrete Green's function* $G_{h,y}(x) \equiv G_h(x,y)$ satisfies, for all $y \in \Omega$, $G_{h,y} \in V_h$, and

$$u_h(y) = \int_\Omega f(x)\, G_h(x,y)\, \mathrm{d}x \,; \tag{3.27}$$

see, e.g., Ciarlet [1970]. Equivalently, the discrete Green's function $G_{h,y} \in V_h$ can be defined (see, e.g., Drăgănescu et al. [2005]) as the solution of

$$a(v_h, G_{h,y}) = v_h(y) \qquad \forall v_h \in V_h \,. \tag{3.28}$$

The discrete Green's function is independent of the chosen discretization basis of V_h. Given a basis Φ and the associated stiffness matrix $\mathbf{A}$ defined in (3.3),

$G_h(x,y)$ satisfies

$$G_h(x,y) = \sum_{i=1}^{N}\sum_{j=1}^{N} \phi_i(y)(\mathbf{A}^{-1})_{ij}\,\phi_j(x)\,. \tag{3.29}$$

Indeed, since $G_{h,y} \in V_h$, for some coefficients $d_j(y)$

$$G_{h,y}(x) = \sum_{j=1}^{N} d_j(y)\phi_j(x)\,.$$

From (3.28),

$$\begin{aligned}\phi_i(y) &= a(\phi_i(x), G_{h,y}(x)) \\ &= a(\phi_i(x), \sum_{j=1}^{N} d_j(y)\phi_j(x)) \\ &= \sum_{j=1}^{N} d_j(y)(\mathbf{A})_{ji} \qquad\qquad i = 1,\ldots,N\,.\end{aligned}$$

Multiplying this equality by $\mathbf{A}^{-1}$ gives

$$d_k(y) = \sum_{i=1}^{N} \phi_i(y)(\mathbf{A}^{-1})_{ik}\,.$$

For ease of notation we will write symbolically

$$G_h(x,y) = \Phi(y)\mathbf{A}^{-1}\Phi^T(x) \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} \phi_i(y)(\mathbf{A}^{-1})_{ij}\,\phi_j(x)\,.$$

Similarly, one can show that when two different bases $\Psi, \mathcal{X}$ of V_h are used in the discretization, giving the stiffness matrix $\overline{\mathbf{A}}$, $(\overline{\mathbf{A}})_{ij} = a(\psi_j, \chi_i)$, $i,j = 1,\ldots,N$, (see Section 3.2), the discrete Green's function has the form

$$G_h(x,y) = \Psi(y)\big(\overline{\mathbf{A}}\big)^{-1}\mathcal{X}^T(x)\,. \tag{3.30}$$

Now, consider an approximation $\widehat{\mathbf{x}} \approx \mathbf{x}$ to the solution of the stiffness system (3.3) and the associated approximate function

$$\widehat{u}_h = \Phi\widehat{\mathbf{x}} \approx u_h = \Phi\mathbf{x}\,.$$

Analogously to (3.27), we look for the *inexact discrete Green's function* $\widehat{G}_h(x,y)$ such that

$$\widehat{u}_h(y) = \int_\Omega f(x)\,\widehat{G}_h(x,y)\,\mathrm{d}x\,. \tag{3.31}$$

Please note, that such function is not uniquely determined, which is in contrast to the uniqueness of the discrete Green's function $G_h(x,y)$. Indeed, similarly to (3.29), a function $\widehat{G}_h(x,y)$ satisfying (3.31) is given by

$$\widehat{G}_h(x,y) = \Phi(y)(\mathbf{A}+\mathbf{E})^{-1}\Phi^T(x)\,, \tag{3.32}$$

for any perturbation $\mathbf{E}$ satisfying $(\mathbf{A}+\mathbf{E})\,\widehat{\mathbf{x}} = \mathbf{b}$, where $\mathbf{E}$ is not unique.

Then (3.31) and (3.27) give

$$(\widehat{u}_h - u_h)(y) = \int_\Omega f(x)\,\left(\widehat{G}_h(x,y) - G_h(x,y)\right)\mathrm{d}x$$

and using (3.32) and (3.29), respectively,

$$\widehat{G}_h(x,y) - G_h(x,y) = \Phi(y)\left((\mathbf{A}+\mathbf{E})^{-1} - \mathbf{A}^{-1}\right)\Phi^T(x)\,. \tag{3.33}$$

We discuss in the next section how to express the term $(\mathbf{A}+\mathbf{E})^{-1} - \mathbf{A}^{-1}$ and use the expression in estimating the algebraic error.

3.4 Estimating the algebraic error via the Fréchet derivative

In this section we consider, for a given approximation $\widehat{\mathbf{x}}$, the algebraic backward error with the perturbation of the system matrix only; see (3.5). Backward error with perturbation of matrix as well as the right-hand side is discussed in the remark at the end of the section.

The (forward) algebraic error can be written as

$$\widehat{\mathbf{x}} - \mathbf{x} = (\mathbf{A}+\mathbf{E})^{-1}\mathbf{b} - \mathbf{A}^{-1}\mathbf{b} \tag{3.34}$$

Assuming the convergence of the (Neumann) series,

$$(\mathbf{A}+\mathbf{E})^{-1} = (\mathbf{I}+\mathbf{A}^{-1}\mathbf{E})^{-1}\mathbf{A}^{-1} = \sum_{k=0}^{\infty}(-\mathbf{A}^{-1}\mathbf{E})^k\mathbf{A}^{-1}\,.$$

Then, for some $\ell \geq 1$,

$$\widehat{\mathbf{x}} - \mathbf{x} \approx \sum_{k=0}^{\ell}(-\mathbf{A}^{-1}\mathbf{E})^k\mathbf{A}^{-1}\mathbf{b} \;-\mathbf{A}^{-1}\mathbf{b} = \sum_{k=1}^{\ell}(-\mathbf{A}^{-1}\mathbf{E})^k\mathbf{A}^{-1}\,\mathbf{b}\,. \tag{3.35}$$

Especially, for $\ell = 1$,

$$\widehat{\mathbf{x}} - \mathbf{x} = -\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b} + O(\|\mathbf{E}\|), \tag{3.36}$$

where $L_g(\mathbf{A},\mathbf{E}) = -\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}$ is the derivative[1] of the function $g(z) = z^{-1}$ at $\mathbf{A}$ in the direction $\mathbf{E}$.

First, consider the rank-1 perturbation matrix given by (3.6), i.e.

$$\mathbf{E} = \frac{(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}})\widehat{\mathbf{x}}^T}{\|\widehat{\mathbf{x}}\|^2}\,.$$

[1] Assuming the differentiability of a matrix function g, the Fréchet derivative of g at $\mathbf{A}$ is a bounded linear mapping $\mathbf{E} \mapsto L_g(\mathbf{A},\mathbf{E})$ such that

$$g(\mathbf{A}+\mathbf{E}) - g(\mathbf{A}) = L_g(\mathbf{A},\mathbf{E}) + O(\|\mathbf{E}\|).$$

Then $-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b}$ is nothing but the scaling of the error $\widehat{\mathbf{x}}-\mathbf{x}$,

$$-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b} = (\widehat{\mathbf{x}}-\mathbf{x})\frac{\widehat{\mathbf{x}}^T\mathbf{x}}{\|\widehat{\mathbf{x}}\|^2}\,. \tag{3.37}$$

Similarly, for any rank-1 matrix $\mathbf{E} = \mathbf{u}\mathbf{v}^T$ in (3.5) we can use the Sherman–Morrison formula (see Sherman and Morrison [1950]) to write

$$\begin{aligned}\widehat{\mathbf{x}}-\mathbf{x} &= \left[(\mathbf{A}+\mathbf{u}\mathbf{v}^T)^{-1}-\mathbf{A}^{-1}\right]\mathbf{b}\\ &= -\frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}\mathbf{b}}{1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}} = \frac{-1}{1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}\cdot\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}\mathbf{b}\,.\end{aligned}$$

Therefore, for a rank-1 perturbation $\mathbf{E}$, we can express the algebraic error $\widehat{\mathbf{x}}-\mathbf{x}$ using the Fréchet derivative as

$$\widehat{\mathbf{x}}-\mathbf{x} = \alpha\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b}$$

with some scaling factor α. For $\mathbf{E}$ given by (3.7), the factor

$$\alpha = -\frac{\|\widehat{\mathbf{x}}\|_{\mathbf{A}}^2}{\widehat{\mathbf{x}}^T\mathbf{b}}$$

is computable without the knowledge of the exact solution $\mathbf{x}$; cf. (3.37).

Evaluation of the term $-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b}$ requires, for a general $\mathbf{E}$, two solutions of linear systems with the matrix $\mathbf{A}$. The simplest idea reducing the evaluation cost is to replace $\mathbf{A}^{-1}\mathbf{b}$ by the computed vector $\widehat{\mathbf{x}}\approx\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$, giving

$$-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\,\mathbf{b} \approx -\mathbf{A}^{-1}\mathbf{E}\,\widehat{\mathbf{x}}\,.$$

From (3.5), $\mathbf{E}\,\widehat{\mathbf{x}} = \mathbf{b}-\mathbf{A}\widehat{\mathbf{x}} = \widehat{\mathbf{r}}$, and

$$-\mathbf{A}^{-1}\mathbf{E}\,\widehat{\mathbf{x}} = -\mathbf{A}^{-1}\widehat{\mathbf{r}}\,.$$

Therefore, approximating $-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\,\mathbf{b}$ in this way leads to the solution of the system $\mathbf{A}\mathbf{y} = -\widehat{\mathbf{r}}$.

Another idea is to compute $-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b}$ using the approaches described in the literature for computing the Fréchet derivative of logarithm or exponential function. The Fréchet derivative can be expressed via power series. Assuming that function g has a power series expansion $g(z) = \sum_{k=0}^{\infty} a_k z^k$ with radius of convergence r, then the Fréchet derivative of g at $\mathbf{A}$, $\|\mathbf{A}\| < r$, in the direction $\mathbf{E}$

$$L_g(\mathbf{A},\mathbf{E}) = \sum_{k=0}^{\infty} a_k \sum_{j=0}^{k} \mathbf{A}^{j-1}\mathbf{E}\mathbf{A}^{k-j}\,;$$

see [Al-Mohy and Higham, 2009, Thm. 3.1]. A recurrence for computing the series above is proposed in [Al-Mohy and Higham, 2009, Thm. 3.2]. However, power series for $g(z) = z^{-1}$ have typically very small radii of convergence and they converge slowly.

Alternatively one can use a block matrix approach taken from [Mathias, 1996, Thm. 2.1]. For a non-singular matrix $\mathbf{A}$

$$\begin{bmatrix}\mathbf{A} & \mathbf{E}\\ 0 & \mathbf{A}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\\ 0 & \mathbf{A}^{-1}\end{bmatrix}.$$

Multiplying by the vector $\left[\begin{smallmatrix}0\\ \mathbf{b}\end{smallmatrix}\right]$ from the right

$$\begin{bmatrix}\mathbf{A} & \mathbf{E}\\ 0 & \mathbf{A}\end{bmatrix}^{-1}\begin{bmatrix}0\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b}\\ \mathbf{A}^{-1}\mathbf{b}\end{bmatrix},$$

or equivalently

$$\begin{bmatrix}\mathbf{A} & \mathbf{E}\\ 0 & \mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}0\\ \mathbf{b}\end{bmatrix}, \qquad \mathbf{y} = -\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b}\,, \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}\,.$$

Since the matrix $\mathbf{E}$ is determined by the computed approximation $\widehat{\mathbf{x}}$ to the solution $\mathbf{x}$, this system cannot be solved all at once. Using its block upper triangular structure one can use back substitution to get the following algorithm:

1. Solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for an approximation $\widehat{\mathbf{x}} \approx \mathbf{x}$;
2. Construct $\mathbf{E}$ such that $(\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b}$;
3. Solve $\mathbf{A}\mathbf{y} + \mathbf{E}\,\widehat{\mathbf{x}} = 0$ for an approximation $\widehat{\mathbf{y}} \approx \widehat{\mathbf{x}} - \mathbf{x}$.

Since $\mathbf{E}\,\widehat{\mathbf{x}} = \widehat{\mathbf{r}}$, the construction of the perturbation matrix $\mathbf{E}$ can be avoided, giving the following algorithm that is nothing but a single step of iterative refinement proposed in [Wilkinson, 1963, p. 121]:

1. Solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for an approximation $\widehat{\mathbf{x}} \approx \mathbf{x}$;
2. Compute the residual $\widehat{\mathbf{r}} = \mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}$;
3. Solve $\mathbf{A}\mathbf{y} = -\widehat{\mathbf{r}}$ for an approximation $\widehat{\mathbf{y}} \approx \widehat{\mathbf{x}} - \mathbf{x}$.

Remark: In the previous discussion, we considered only the perturbation of the system matrix as in (3.5). Whenever the both sides of the equation are perturbed,

$$(\mathbf{A}+\mathbf{E})\widehat{\mathbf{x}} = \mathbf{b} + \mathbf{f}\,,$$

the expression for the algebraic error based on the above presented idea is more complicated. Analogously to (3.34) one can write

$$\widehat{\mathbf{x}} - \mathbf{x} = (\mathbf{A}+\mathbf{E})^{-1}(\mathbf{b}+\mathbf{f}) - \mathbf{A}^{-1}\mathbf{b}\,.$$

With a simple manipulation

$$\widehat{\mathbf{x}} - \mathbf{x} = \left[(\mathbf{A}+\mathbf{E})^{-1} - \mathbf{A}^{-1}\right]\mathbf{b} + (\mathbf{A}+\mathbf{E})^{-1}\mathbf{f}\,,$$

or

$$\widehat{\mathbf{x}} - \mathbf{x} = \left[(\mathbf{A}+\mathbf{E})^{-1} - \mathbf{A}^{-1}\right](\mathbf{b}+\mathbf{f}) + \mathbf{A}^{-1}\mathbf{f}\,.$$

Replacing $[(\mathbf{A}+\mathbf{E})^{-1} - \mathbf{A}^{-1}]$ by $-\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}$ then gives two approximations for the algebraic error

$$\widehat{\mathbf{x}} - \mathbf{x} \approx -\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}\mathbf{b} + (\mathbf{A}+\mathbf{E})^{-1}\mathbf{f}\,,$$
$$\widehat{\mathbf{x}} - \mathbf{x} \approx -\mathbf{A}^{-1}\mathbf{E}\mathbf{A}^{-1}(\mathbf{b}+\mathbf{f}) + \mathbf{A}^{-1}\mathbf{f}\,.$$

However, the approximation of the algebraic error using these formulas is even more costly than in the case of perturbation of the system matrix only discussed above.

3.5 Numerical illustrations

For numerical illustrations we consider the one-dimensional test example considered in [Papež et al., 2014, Section 2], i.e., the Poisson problem on the domain $\Omega \equiv (0,1)$ with the exact solution $u = \exp(-5(x-0.5)^2) - \exp(-5/4)$ discretized on a uniform partition with 19 inner nodes. As in Papež et al. [2014] we substitute for $\widehat{\mathbf{x}}$ the approximation given after 8 steps of the CG iteration steps starting with the zero initial vector. Consider the nonsymmetric rank-1 perturbation $\mathbf{E}$ given by (3.6) and the symmetric perturbation given by (3.8) that we denote here by $\mathbf{F}$. Note, that if the approximation $\widehat{\mathbf{x}}$ is orthogonal to its associated residual $\widehat{\mathbf{r}}$, i.e. $\widehat{\mathbf{x}}^T\widehat{\mathbf{r}} = 0$, then $\mathbf{F} = \mathbf{E} + \mathbf{E}^T$. This happens, e.g., when $\widehat{\mathbf{x}}$ is given by the CG method starting with zero initial guess, which is our case.

When we consider the transformation of the discretization basis only (see Section 3.2.2), the transformation matrices are

$$\mathbf{D_E} \equiv \mathbf{A}^{-1}\mathbf{E}\,,$$
$$\mathbf{D_F} \equiv \mathbf{A}^{-1}\mathbf{F}\,.$$

Recall that $\mathbf{D_E}$ is the minimal transformation with respect to the Euclidean and Frobenius norm (see Section 3.2.5) and that $\mathbf{D_F}$ is (generally) nonsymmetric. The norms of the perturbation and transformation matrices are in our example

$$\begin{aligned} \|\mathbf{E}\| &= 3.30 \times 10^{-1}, & \|\mathbf{D_E}\| &= 1.47 \times 10^{-2},\\ \|\mathbf{E}\|_F &= 3.30 \times 10^{-1}, & \|\mathbf{D_E}\|_F &= 1.47 \times 10^{-2},\\ \|\mathbf{F}\| &= 3.30 \times 10^{-1}, & \|\mathbf{D_F}\| &= 6.68 \times 10^{-1},\\ \|\mathbf{F}\|_F &= 4.66 \times 10^{-1}, & \|\mathbf{D_F}\|_F &= 6.68 \times 10^{-1}. \end{aligned}$$

The matrices $\mathbf{E}$, $\mathbf{D_E}$ and $\mathbf{F}$, $\mathbf{D_F}$ are depicted in Figure 3.1 and 3.2, respectively. Figure 3.3 illustrates the difference of a discretization basis function transformed as in (3.9) using $\mathbf{D_E}$, $\mathbf{D_F}$.

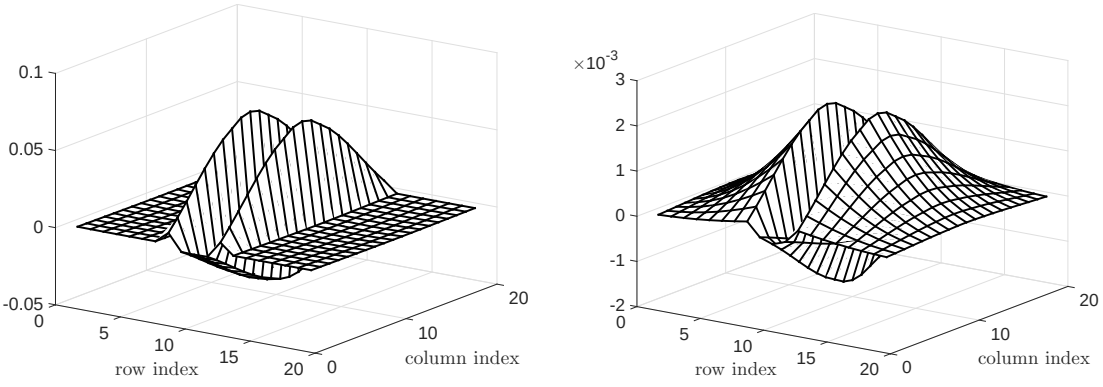


Figure 3.1: The perturbation matrix $\mathbf{E}$ (left) and the transformation matrix $\mathbf{D_E}$ (right) (with the entries visualized using the MATLAB `surf` command). The right vertical axis is scaled by 10^{-3}.

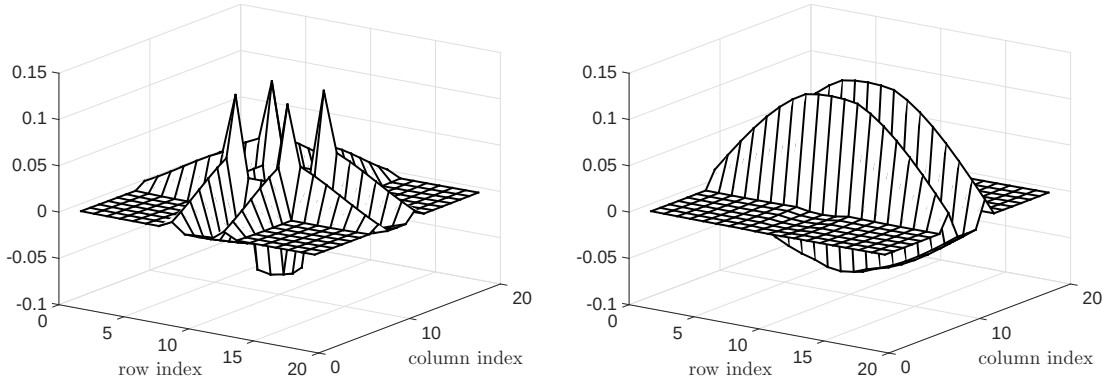


Figure 3.2: The perturbation matrix $\mathbf{F}$ (left) and the transformation matrix $\mathbf{D_F}$ (right) (with the entries visualized using the MATLAB `surf` command).

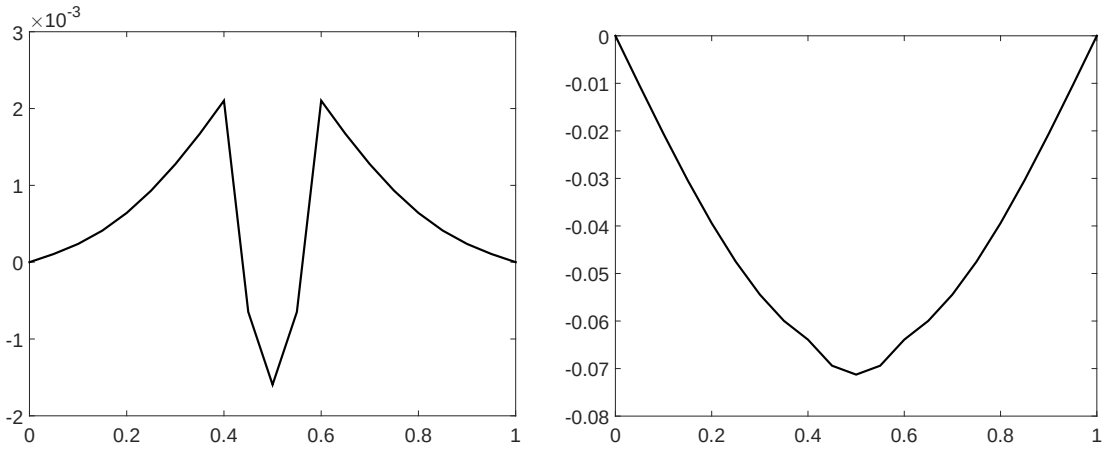


Figure 3.3: The difference $\psi_{10} - \phi_{10}$ for ψ_{10} given by (3.9) with the transformation matrix $\mathbf{D_E}$ (left) and for $\mathbf{D_F}$ (right). The left vertical axis is scaled by 10^{-3}.

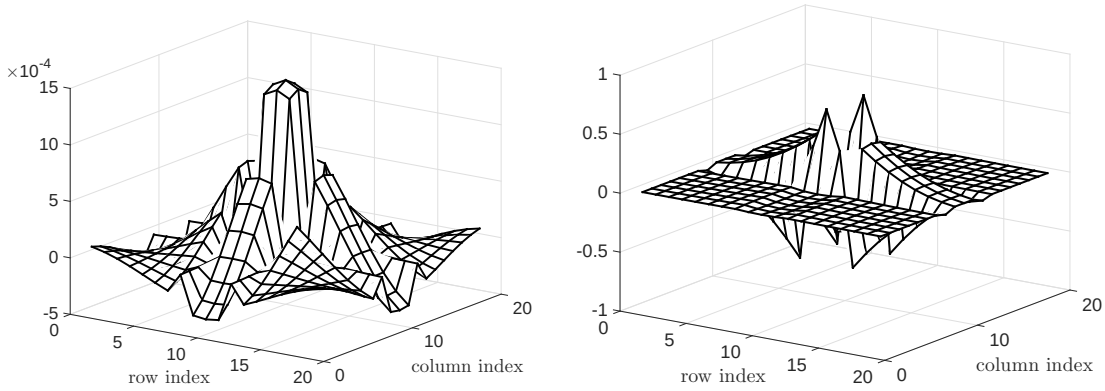


Figure 3.4: The symmetric perturbation matrix $\widehat{\mathbf{E}}$ given by (3.18) minimal with respect to the energy norm (left) and the associated transformation matrix $\widehat{\mathbf{D}}$ (right) given by (3.21). The entries are visualized using the MATLAB `surf` command. The left vertical axis is scaled by 10^{-4}.

Figure 3.4 depicts the symmetric perturbation matrix $\widehat{\mathbf{E}}$ minimal with respect to the energy norm discussed in Section 3.2.4; recall that here the (generally) nonzero perturbation of the right-hand side in (3.4) and the same transformation of discretization basis and test functions are considered, $\Psi = \mathcal{X}$. Note that while the entries of $\widehat{\mathbf{E}}$ are smaller than the entries of the matrices $\mathbf{E}$, $\mathbf{F}$ considered above, some of the entries of the associated transformation matrix $\widehat{\mathbf{D}}$ given by (3.21) are significantly larger than the entries of $\mathbf{D_E}$, $\mathbf{D_F}$.

The Green's function $G(x, y)$ for the one-dimensional Poisson problem is given in Figure 3.5. Figure 3.6 then depicts the difference $G_h(x, y) - \widehat{G}_h(x, y)$ between the discrete Green's function (3.29) and the inexact discrete Green's function (3.32) for the perturbations $\mathbf{E}$, $\mathbf{F}$ considered above.

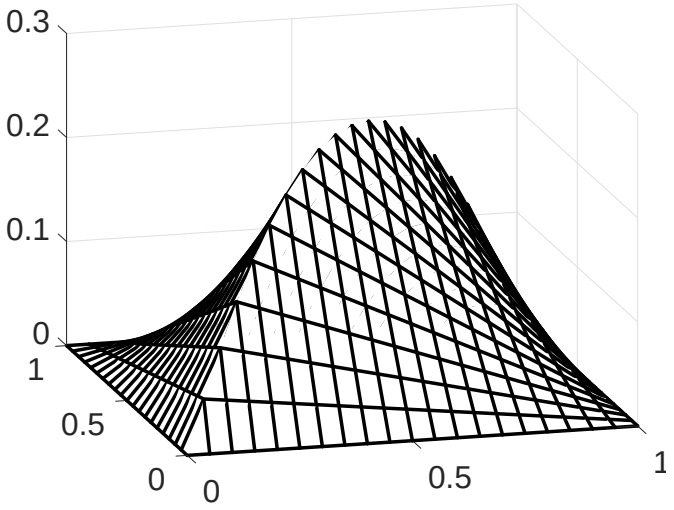


Figure 3.5: The Green's function $G(x, y)$ for the one-dimensional Poisson problem.

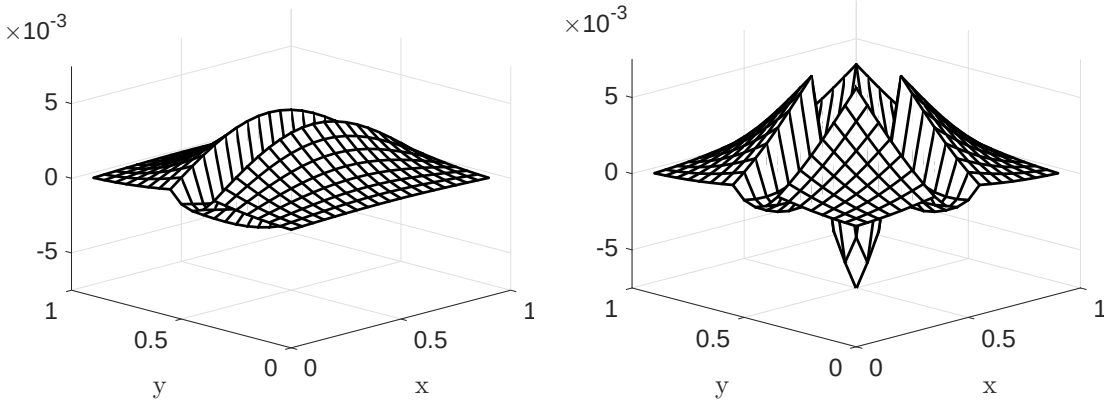


Figure 3.6: The difference $G_h(x, y) - \widehat{G}_h(x, y)$ (see (3.33)) for the perturbation $\mathbf{E}$ (left) and for the symmetric perturbation $\mathbf{F}$ (right).

Bibliography

A. H. Al-Mohy and N. J. Higham. Computing the Fréchet derivative of the matrix exponential, with an application to condition number estimation. *SIAM*

J. Matrix Anal. Appl., 30(4):1639–1657, 2009. ISSN 0895-4798.

M. Arioli, E. Noulard, and A. Russo. Stopping criteria for iterative methods: applications to PDE's. *Calcolo*, 38(2):97–112, 2001. ISSN 0008-0624.

J. R. Bunch, J. W. Demmel, and C. F. Van Loan. The strong stability of algorithms for solving symmetric linear systems. *SIAM J. Matrix Anal. Appl.*, 10 (4):494–499, 1989. ISSN 0895-4798.

P. G. Ciarlet. Discrete variational Green's function. I. *Aequationes Math.*, 4: 74–82, 1970. ISSN 0001-9054.

J. E. Dennis, Jr. and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. ISBN 0-89871-364-1. Corrected reprint of the 1983 original.

A. Drăgănescu, T. F. Dupont, and L. R. Scott. Failure of the discrete maximum principle for an elliptic finite element problem. *Math. Comp.*, 74(249):1–23, 2005. ISSN 0025-5718.

S. Gratton, P. Jiránek, and X. Vasseur. Energy backward error: interpretation in numerical solution of elliptic partial differential equations and behaviour in the conjugate gradient method. *Electron. Trans. Numer. Anal.*, 40:338–355, 2013. ISSN 1068-9613.

N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, second edition, 2002. ISBN 0-89871-521-0.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013. ISBN 978-0-521-54823-6.

E. Keilegavlen and J. M. Nordbotten. Inexact linear solvers for control volume discretizations in porous media. *Comput. Geosci.*, 19(1):159–176, 2015. ISSN 1420-0597.

P. D. Lax and A. N. Milgram. Parabolic equations. In *Contributions to the theory of partial differential equations*, Annals of Mathematics Studies, no. 33, pages 167–190. Princeton University Press, Princeton, N. J., 1954.

R. Mathias. A chain rule for matrix functions and applications. *SIAM J. Matrix Anal. Appl.*, 17(3):610–620, 1996. ISSN 0895-4798.

J. Nečas. *Les méthodes directes en théorie des équations elliptiques*. Masson et Cie, Éditeurs, Paris; Academia, Éditeurs, Prague, 1967.

J. M. Nordbotten and P. E. Bjørstad. On the relationship between the multiscale finite-volume method and domain decomposition preconditioners. *Comput. Geosci.*, 12(3):367–376, 2008. ISSN 1420-0597.

W. Oettli and W. Prager. Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides. *Numer. Math.*, 6:405–409, 1964. ISSN 0029-599X.

J. Papež, J. Liesen, and Z. Strakoš. Distribution of the discretization and algebraic error in numerical solution of partial differential equations. *Linear Algebra Appl.*, 449:89–114, 2014. ISSN 0024-3795.

M. H. Protter and H. F. Weinberger. *Maximum principles in differential equations*. Springer-Verlag, New York, 1984. ISBN 0-387-96068-6. Corrected reprint of the 1967 original.

J.-L. Rigal and J. Gaches. On the compatibility of a given solution with the data of a linear system. *J. Assoc. Comput. Mach.*, 14:543–548, 1967.

J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statistics*, 21:124–127, 1950. ISSN 0003-4851.

J. H. Wilkinson. *Rounding errors in algebraic processes*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

3.A Proof of Theorem 3.1

The simple algebraic manipulation shows that

$$\begin{aligned}\|(\mathbf{I}+\mathbf{D})^T\mathbf{A}(\mathbf{I}+\mathbf{D})-\mathbf{A}\|_{\mathbf{A},\mathbf{A}^{-1}} &= \|\mathbf{A}^{-1/2}(\mathbf{I}+\mathbf{D})^T\mathbf{A}(\mathbf{I}+\mathbf{D})\mathbf{A}^{-1/2}-\mathbf{I}\| \\ &= \|(\mathbf{I}+\mathbf{A}^{1/2}\mathbf{D}\mathbf{A}^{-1/2})^T(\mathbf{I}+\mathbf{A}^{1/2}\mathbf{D}\mathbf{A}^{-1/2})-\mathbf{I}\|\,.\end{aligned}$$

We denote

$$\mathbf{R} \equiv \mathbf{I}+\mathbf{A}^{1/2}\mathbf{D}\mathbf{A}^{-1/2}\,. \tag{3.38}$$

For $\mathbf{D}\in\mathcal{D}(\widehat{\mathbf{x}},\mathbf{x}-\widehat{\mathbf{x}})$, it holds

$$\mathbf{R}\mathbf{A}^{1/2}\,\widehat{\mathbf{x}} = (\mathbf{I}+\mathbf{A}^{1/2}\mathbf{D}\mathbf{A}^{-1/2})\mathbf{A}^{1/2}\,\widehat{\mathbf{x}} = \mathbf{A}^{1/2}(\mathbf{I}+\mathbf{D})\widehat{\mathbf{x}} = \mathbf{A}^{1/2}\,\mathbf{x}\,,$$

i.e. $\mathbf{R}\in\mathcal{D}(\mathbf{A}^{1/2}\,\widehat{\mathbf{x}},\mathbf{A}^{1/2}\,\mathbf{x})$, and we can write (3.20) as

$$\min_{\mathbf{D}\in\mathcal{D}(\widehat{\mathbf{x}},\mathbf{x}-\widehat{\mathbf{x}})}\|(\mathbf{I}+\mathbf{D})^T\mathbf{A}(\mathbf{I}+\mathbf{D})-\mathbf{A}\|_{\mathbf{A},\mathbf{A}^{-1}} = \min_{\mathbf{R}\in\mathcal{D}(\mathbf{A}^{1/2}\,\widehat{\mathbf{x}},\mathbf{A}^{1/2}\,\mathbf{x})}\|\mathbf{R}^T\mathbf{R}-\mathbf{I}\|\,.$$

For ease of notation, write

$$\mathbf{s}\equiv\mathbf{A}^{1/2}\,\widehat{\mathbf{x}}, \qquad \mathbf{y}\equiv\mathbf{A}^{1/2}\,\mathbf{x}\,.$$

Define

$$\mathbf{s}_1 \equiv \mathbf{s}-\frac{\mathbf{s}^T\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}\,, \qquad \mathbf{y}_1\equiv\mathbf{y}-\frac{\mathbf{s}^T\mathbf{y}}{\|\mathbf{s}\|^2}\mathbf{s}\,.$$

There holds $\mathbf{s}_1=\mathbf{A}^{1/2}\,\widehat{\mathbf{x}}_1$, $\mathbf{y}_1=\mathbf{A}^{1/2}\,\mathbf{x}_1$. Therefore, providing $\mathbf{x}_1\neq 0$, $\widehat{\mathbf{x}}_1\neq 0$, the vectors $\mathbf{y}_1$ and $\mathbf{s}_1$ are nonzero. We now show that

$$\min_{\mathbf{R}\in\mathcal{D}(\mathbf{s},\mathbf{y})}\|\mathbf{R}^T\mathbf{R}-\mathbf{I}\| = \left|\frac{\|\mathbf{y}\|^2}{\|\mathbf{s}\|^2}-1\right| = \left|\frac{\|\mathbf{x}\|_{\mathbf{A}}^2}{\|\widehat{\mathbf{x}}\|_{\mathbf{A}}^2}-1\right|$$

and that

$$\arg\min_{\mathbf{R}\in\mathcal{D}(\mathbf{s},\mathbf{y})} \|\mathbf{R}^T\mathbf{R}-\mathbf{I}\| = \widehat{\mathbf{R}} \equiv \mathbf{I} + \frac{(\mathbf{y}-\mathbf{s})\mathbf{s}^T}{\|\mathbf{s}\|^2} + \left(\frac{\mathbf{s}_1}{\|\mathbf{s}_1\|} - \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|}\right)\frac{\mathbf{y}_1^T}{\|\mathbf{y}_1\|}\,. \tag{3.39}$$

Direct calculation shows that $\widehat{\mathbf{R}} \in \mathcal{D}(\mathbf{s},\mathbf{y})$. Indeed, since $\mathbf{y}_1^T\mathbf{s} = 0$, it holds

$$\widehat{\mathbf{R}}\mathbf{s} = \left(\mathbf{I} + \frac{(\mathbf{y}-\mathbf{s})\mathbf{s}^T}{\|\mathbf{s}\|^2} + \left(\frac{\mathbf{s}_1}{\|\mathbf{s}_1\|} - \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|}\right)\frac{\mathbf{y}_1^T}{\|\mathbf{y}_1\|}\right)\mathbf{s} = \left(\mathbf{I} + \frac{(\mathbf{y}-\mathbf{s})\mathbf{s}^T}{\|\mathbf{s}\|^2}\right)\mathbf{s} = \mathbf{y}\,.$$

For any $\mathbf{R} \in \mathcal{D}(\mathbf{s},\mathbf{y})$, there holds

$$\|\mathbf{R}^T\mathbf{R}-\mathbf{I}\| = \max_{\mathbf{v}\neq 0} \frac{\left|\mathbf{v}^T\left(\mathbf{R}^T\mathbf{R}-\mathbf{I}\right)\mathbf{v}\right|}{\|\mathbf{v}\|^2} \geq \frac{\left|\mathbf{s}^T\left(\mathbf{R}^T\mathbf{R}-\mathbf{I}\right)\mathbf{s}\right|}{\|\mathbf{s}\|^2} = \left|\frac{\|\mathbf{y}\|^2}{\|\mathbf{s}\|^2} - 1\right|\,.$$

We proceed by showing that

$$\|\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\| = \left|\frac{\|\mathbf{y}\|^2}{\|\mathbf{s}\|^2} - 1\right|\,.$$

Let $\mathbf{v}$ be an arbitrary vector of $\mathbb{R}^N$. The vectors $\mathbf{s}$ and $\mathbf{y}_1$ form an orthogonal basis of $\mathcal{U} \equiv \operatorname{span}\{\mathbf{s},\mathbf{y}\}$ and one can write $\mathbf{v} = \mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3$, with $\mathbf{u}_1 = \alpha\mathbf{s}$, $\mathbf{u}_2 = \beta\mathbf{y}_1$, and $\mathbf{u}_3$ orthogonal to $\mathcal{U}$. The construction of $\widehat{\mathbf{R}}$ gives

$$\begin{aligned}
&\widehat{\mathbf{R}}\mathbf{u}_1 = \alpha\widehat{\mathbf{R}}\mathbf{s} = \alpha\mathbf{y}\,,\\
&\widehat{\mathbf{R}}\mathbf{u}_2 = \beta\left(\mathbf{I} + \frac{(\mathbf{y}-\mathbf{s})\mathbf{s}^T}{\|\mathbf{s}\|^2} + \left(\frac{\mathbf{s}_1}{\|\mathbf{s}_1\|} - \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|}\right)\frac{\mathbf{y}_1^T}{\|\mathbf{y}_1\|}\right)\mathbf{y}_1 = \beta\frac{\|\mathbf{y}_1\|}{\|\mathbf{s}_1\|}\mathbf{s}_1\,,\\
&\widehat{\mathbf{R}}\mathbf{u}_3 = \mathbf{u}_3\,.
\end{aligned}$$

Then we have

$$\mathbf{u}_2^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)\mathbf{u}_1 = \alpha\beta\frac{\|\mathbf{y}_1\|}{\|\mathbf{s}_1\|}\mathbf{s}_1^T\mathbf{y} = 0\,, \tag{3.40}$$

$$\mathbf{u}_2^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)\mathbf{u}_2 = \beta^2\left(\frac{\|\mathbf{y}_1\|^2}{\|\mathbf{s}_1\|^2}\|\mathbf{s}_1\|^2 - \|\mathbf{y}_1\|^2\right) = 0\,, \tag{3.41}$$

$$\mathbf{u}_3^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)\mathbf{u}_3 = \|\mathbf{u}_3\|^2 - \|\mathbf{u}_3\|^2 = 0\,, \tag{3.42}$$

$$\mathbf{u}_3^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)(\mathbf{u}_1+\mathbf{u}_2) = \mathbf{u}_3^T\left(\widehat{\mathbf{R}}\,(\mathbf{u}_1+\mathbf{u}_2) - (\mathbf{u}_1+\mathbf{u}_2)\right) = 0\,, \tag{3.43}$$

where in the last equality we used the fact that $\widehat{\mathbf{R}}\,(\mathbf{u}_1+\mathbf{u}_2) \in \mathcal{U}$ and $\mathbf{u}_3$ is orthogonal to $\mathcal{U}$. Collecting equalities (3.40)–(3.43), infer

$$\begin{aligned}
\frac{\left|\mathbf{v}^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)\mathbf{v}\right|}{\|\mathbf{v}\|^2} &= \frac{\left|(\mathbf{u}_1+\mathbf{u}_2+\mathbf{u}_3)^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)(\mathbf{u}_1+\mathbf{u}_2+\mathbf{u}_3)\right|}{\|\mathbf{u}_1+\mathbf{u}_2+\mathbf{u}_3\|^2}\\
&= \frac{\left|\mathbf{u}_1^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)\mathbf{u}_1\right|}{\|\mathbf{u}_1\|^2+\|\mathbf{u}_2\|^2+\|\mathbf{u}_3\|^2} \leq \frac{\left|\mathbf{u}_1^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)\mathbf{u}_1\right|}{\|\mathbf{u}_1\|^2}\\
&= \frac{\left|\mathbf{s}^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}}-\mathbf{I}\right)\mathbf{s}\right|}{\|\mathbf{s}\|^2} = \left|\frac{\|\mathbf{y}\|^2}{\|\mathbf{s}\|^2} - 1\right|.
\end{aligned}$$

Thus

$$\|\widehat{\mathbf{R}}^T\widehat{\mathbf{R}} - \mathbf{I}\| = \max_{\mathbf{v}\neq 0} \frac{\left|\mathbf{v}^T\left(\widehat{\mathbf{R}}^T\widehat{\mathbf{R}} - \mathbf{I}\right)\mathbf{v}\right|}{\|\mathbf{v}\|^2} = \left|\frac{\|\mathbf{y}\|^2}{\|\mathbf{s}\|^2} - 1\right|.$$

Moreover, $\widehat{\mathbf{R}}$ is nonsingular. Indeed, let $\mathbf{v} \in \mathbb{R}$ and suppose $\widehat{\mathbf{R}}\mathbf{v} = 0$. Decomposing $\mathbf{v}$ as before,

$$0 = \widehat{\mathbf{R}}\mathbf{v} = \widehat{\mathbf{R}}(\mathbf{u}_1 + \mathbf{u}_2 + \mathbf{u}_3) = \alpha\mathbf{y} + \beta\frac{\|\mathbf{y}_1\|}{\|\mathbf{s}_1\|}\mathbf{s}_1 + \mathbf{u}_3.$$

Since $\mathbf{y}, \mathbf{s}_1$ and $\mathbf{u}_3$ are orthogonal to each other, it holds $\alpha = \beta = 0$ and $\mathbf{u}_3 = 0$, and thus $\mathbf{v} = 0$.

Finally, the formula (3.21) follows from (3.38) and (3.39), and the proof is finished. □

4. Algebraic errors, error indicators and adaptivity

A posteriori error analysis in numerical PDEs often aims at estimating discretization error, which is then used in adaptive finite element schemes to refine discretization in the parts of the domain according to the values of the estimator. Such refinement can reduce the norm of the discretization error at low cost in comparison to uniform mesh refinement, and result in a close-to-uniform spatial distribution of the error over the domain. However, the estimators are often derived for the FEM discrete solution, i.e., assuming the exact solution of the corresponding algebraic system. There is a growing body of work avoiding this assumption; a thorough list of references with the discussion can be found, e.g., in the recent survey [Arioli et al., 2013, Section 4]. We will present one of the possible approaches based on flux reconstruction techniques in Chapter 5 of the thesis.

The impact of abandoning the assumption on the exact algebraic solution onto the so-called residual-based error estimator, allowing its evaluation for a computed approximation is investigated in Papež and Strakoš [2016]. This paper is included in Section 4.1 below. In Section 4.2, we compare convergence of the adaptive finite element numerical solution and the adaptively generated meshes when using the error indicators evaluated for the (tightly approximated) FEM discrete solution and the indicators evaluated for the approximations with nonnegligible algebraic error.

4.1 Paper submitted to IMA Journal of Numerical Analysis

The section includes the paper Papež and Strakoš [2016] submitted to IMA Journal of Numerical Analysis on May 18, 2016. The paper is currently in revision.

Galerkin orthogonality and the multiplicative factors in the residual-based a posteriori error estimator for total error*

J. Papež† Z. Strakoš†

May 18, 2016

Abstract

A posteriori error analysis in numerical PDEs aims at providing sufficiently accurate information about the distance of the numerically computed approximation to the true solution. The information about the error should be determined from the computed quantities without hidden assumptions or uncomputable multiplicative factors. Using the standard Poisson model problem, this short text examines subtleties of the residual-based a posteriori error estimator for the discretization and total error. In particular, we study the impact of abandoning the assumption on the Galerkin orthogonality onto the estimator, allowing its evaluation at the presence of the algebraic error.

Keywords: A posteriori error analysis, residual-based estimator, computable error bound, finite element method, Galerkin orthogonality, inexact algebraic solution.

MSC: 65N15, 65N22, 65N30, 65N50.

1 Introduction

Historically, most a posteriori analysis in numerical PDEs focuses on estimating the discretization error, i.e., on the discrepancy between the exact solution of the original infinite-dimensional formulation of the problem and the *exact solution* of its discretized counterpart. This information is crucial for adaptivity, which refines discretization in the parts of the domain where the estimator indicates a large discretization error in order to achieve its close-to-uniform spatial distribution over the domain. Estimation of the discretization error is, however, linked with the following philosophical as well as practical difficulty: the exact solution of the original problem is unknown, and, unless the algebraic computations providing the coordinates of the discrete solution in the discretization basis are performed exactly or with a negligible algebraic error, the exact solution of the

*This work was supported by the ERC-CZ project LL1202.

†Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Prague, Czech Republic.

discretized problem is also unknown. Near-to-exact algebraic computations can be prohibitive due to extensive computational cost. Reaching exact algebraic results can even be theoretically prohibitive. The eigenvalues of finite matrices, e.g., are in principle (in general) uncomputable by any finite formula due to the Abel–Galois theorem, and they can only be approximated iteratively. Moreover, in case of highly non-normal matrices there is no forward error guarantee of the accuracy of the computed eigenvalue approximations and we can guarantee the backward error only. Due to the inexactness of algebraic computations, the a posteriori error bounds should be based, from their derivation to their application, on the available computed approximations to the solution of the discrete problem. The present text discusses the subtleties one has to deal with while estimating the discretization and the total error using a residual-based a posteriori error estimator; see, e.g., [2, 3, 4]. We focus on the relationship between the estimator and the algebraic error. More specifically, we show that removing the standard Galerkin orthogonality assumption, which is a prerequisite for a mathematically rigorous application of the bound, means a nontrivial revision of the published results.

We will use the following standard model problem. Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain (open, bounded and connected set with a polygonal boundary). We consider the Poisson problem with the homogeneous Dirichlet boundary condition

$$\text{find } u : \Omega \to \mathbb{R} : \qquad -\Delta u = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \partial\Omega\,, \tag{1.1}$$

where $f : \Omega \to \mathbb{R}$ is the source term. Hereafter we use the standard notation for the Sobolev spaces. For $D \subset \Omega$, $L^1(D)$ denotes the space of the (Lebesgue) integrable functions in D, $L^2(D)$ denotes the space of the square integrable functions in D, $(w, v)_D = \int_D v\, w$ denotes the L^2-inner product on $L^2(D)$, and $\|w\|_D = (w, w)_D^{1/2}$ denotes the associated L^2-norm. We omit the subscripts for $D = \Omega$. $H^k(\Omega)$ denotes the Hilbert space of functions in $L^2(\Omega)$ whose weak derivatives up to the order k belong to $L^2(\Omega)$. $H_0^1(\Omega)$ denotes the space of functions in $H^1(\Omega)$ with vanishing trace on the boundary $\partial\Omega$.

Assuming $f \in L^2(\Omega)$, the problem (1.1) can be written in the following weak form

$$\text{find } u \in V \equiv H_0^1(\Omega) : \qquad (\nabla u, \nabla v) = (f, v) \qquad \text{for all } v \in V\,. \tag{1.2}$$

Let $\mathcal{T}$ be a conforming triangulation of the domain Ω, i.e., two distinct and intersecting elements $T_1, T_2 \in \mathcal{T}$ share a common face, edge or vertex. Let $\mathcal{N}$ denote the set of all nodes (i.e. the vertices of the elements of $\mathcal{T}$) while $\mathcal{N}_{\text{int}} \equiv \mathcal{N}\backslash\partial\Omega$ denotes the set of the free nodes. By $\mathcal{E}$ we denote the set of all edges of the elements of $\mathcal{T}$ and, similarly, $\mathcal{E}_{\text{int}} \equiv \mathcal{E}\backslash\partial\Omega$. For any node $z \in \mathcal{N}$, let φ_z be the corresponding hat-function, i.e., the piecewise linear function that takes value 1 at the node z and vanishes at all other nodes. By ω_z we denote the support of φ_z which is equal to the patch $\omega_z = \cup\{T \in \mathcal{T} | z \in T\}$. For an element $T \in \mathcal{T}$ we denote $h_T \equiv \text{diam}(T)$, similarly $h_z \equiv \text{diam}(\omega_z)$ denotes the diameter of $\omega_z, z \in \mathcal{N}$. By $V_h \subset V$ we denote the space of the continuous piecewise linear functions on the triangulation $\mathcal{T}$ vanishing on the boundary $\partial\Omega$, i.e. $V_h \equiv \text{span}\,\{\varphi_z | z \in \mathcal{N}_{\text{int}}\}$. The discrete formulation of (1.2) then reads

$$\text{find } u_h \in V_h : \qquad (\nabla u_h, \nabla v_h) = (f, v_h) \qquad \text{for all } v_h \in V_h\,. \tag{1.3}$$

The solution u_h of (1.3) is called the *Galerkin solution*. Subtracting (1.3) from (1.2) and using $V_h \subset V$, we get the *Galerkin orthogonality*

$$(\nabla(u - u_h), \nabla v_h) = 0 \qquad \text{for all } v_h \in V_h \,. \tag{1.4}$$

The difficulty in estimating the discretization error $u - u_h$ mentioned above can be formulated as follows. Consider any estimator $\mathrm{EST}(\cdot)$ that provides an upper bound

$$|||u - u_h||| \le \mathrm{EST}(u_h) \,, \tag{1.5}$$

where $|||\cdot|||$ denotes an appropriate norm (for the model problem (1.1) typically the energy norm $|||w||| = \|\nabla w\|^{1/2} = (\nabla w, \nabla w)^{1/2}$). In order to evaluate the right-hand side of (1.5) we need u_h that is not available. The common practice is then replacing u_h by the computed approximation u_h^C, giving the seemingly easy solution

$$|||u - u_h||| \le \mathrm{EST}(u_h^C) \,.$$

This inequality is, however, not guaranteed to hold without further justification, that can be highly nontrivial or even impossible to achieve. In particular, provided that

$$\mathrm{EST}(u_h) = \inf_{v_h \in V_h} \mathrm{EST}(v_h) \,, \tag{1.6}$$

the bound (1.5) does indeed lead to a guaranteed upper bound

$$|||u - u_h||| \le \mathrm{EST}(v_h) \quad \text{for all } v_h \in V_h \,. \tag{1.7}$$

Proving (1.6) can, however, represent a challenge. A rigorous incorporation of the algebraic inaccuracy into the a posteriori error analysis that would allow theoretically justified comparison of the discretization and algebraic error is a difficult problem; see, e.g., [8, 12, 13] and [6, Section 7.1]. One can argue that the continuity argument gives $\mathrm{EST}(u_h^C) \approx \mathrm{EST}(u_h)$ for $u_h^C \approx u_h$. However, when solving practical problems one typically needs to estimate the error even for u_h^C far from u_h; see, e.g., [11, Conclusions], [9, 10].

In [3, Lemma 3.1] the bound on the total error is given in the form

$$\|\nabla(u - v_h)\|^2 \le \widetilde{C} \cdot \mathrm{EST}(v_h) + 2\,\|\nabla(u_h - v_h)\|^2 \,, \tag{1.8}$$

where $\widetilde{C}$ is stated to depend only on the minimal angle of the triangulation $\mathcal{T}$, and $v_h \in V_h$ is arbitrary, i.e., it can account for inexact algebraic computations. The proof refers for the case $v_h = u_h$, i.e., for estimating the discretization error, to the paper [4]. The proof is completed by arguing that the general case (1.8) follows via the triangle inequality. In the present paper we critically examine this argument and revisit the derivation of (1.8) in a way that explicitly describes the multiplicative factors, including the worst-case relationship to the infinite-dimensional and the computed approximate solutions u and u_h^C respectively.

In the next section we recall the results from [4] on quasi-interpolation. Section 3 presents the revision of the upper bound (1.8) on the total error and gives its detailed proof that abandons the Galerkin orthogonality assumption. Section 4 comments on the upper bound on the total error obtained by the approach used in [2]. Numerical illustrations are present in Section 5, followed by conclusions.

2 Quasi-interpolation results

This section presents results from [4] used further in the text. We include them here for completeness and self-consistency of the text. Denote by ψ a piecewise linear function taking value 1 at the inner nodes $z \in \mathcal{N}_{\rm int}$ and vanishing on the boundary $\partial\Omega$, $\psi \equiv \sum_{z\in\mathcal{N}_{\rm int}} \varphi_z$. Then φ_z/ψ, $z \in \mathcal{N}_{\rm int}$, represents in Ω a partition of unity, $\sum_{z\in\mathcal{N}_{\rm int}} \varphi_z/\psi = 1$ in Ω. Following [4], for a given $w \in L^1(\Omega)$ define the quasi-interpolation operator $\mathcal{I} : L^1(\Omega) \to V_h$

$$\mathcal{I}w \equiv \sum_{z\in\mathcal{N}_{\rm int}} w_z\varphi_z\,, \quad \text{where} \quad w_z \equiv \frac{(w,\varphi_z/\psi)}{(1,\varphi_z)}\,.$$

The error $w - \mathcal{I}w$ has a vanishing weighted average. Namely, for $w, R \in L^2(\Omega)$ and arbitrary numbers $R_z \in \mathbb{R}, z \in \mathcal{N}_{\rm int}$,

$$\int_\Omega R\,(w - \mathcal{I}w) = \sum_{z\in\mathcal{N}_{\rm int}} \int_\Omega (R - R_z)(w - w_z\psi)(\varphi_z/\psi)\,; \tag{2.1}$$

see [4, Remark 2.4]. Since

$$\int_\Omega (w - w_z\psi)(\varphi_z/\psi) = 0 \qquad \text{for all } z \in \mathcal{N}_{\rm int}\,,$$

the numbers $R_z \in \mathbb{R}$, can be chosen arbitrarily. In particular, R_z can be chosen as the mean value of R on ω_z. Then $\int_{\omega_z} |R - R_z|^2$ is minimal among all $R_z \in \mathbb{R}$.

The following lemmas are stated and proved in [4] for a more general case. Considering the model problem (1.1), we restrict ourselves to the case $w \in H^1_0(\Omega)$. The multiplicative factors in the lemmas then depend on

I. the shape of ω_z ,

II. the shapes of $\omega_{z\partial\Omega} \equiv (\omega_z \cup \omega_\xi \mid z \in \mathcal{N}_{\rm int}, \xi \in \mathcal{N}\backslash\mathcal{N}_{\rm int}, \omega_z \cap \omega_\xi \neq 0)$,

III. the shape coefficients $(\int_{\omega_z} \varphi_z/|\omega_z| \mid z \in \mathcal{N}_{\rm int})$, where $|\omega_z|$ stands for the Lebesgue measure of ω_z,

IV. the overlap

$$M_1 \equiv \max_{z\in\mathcal{N}_{\rm int}} \operatorname{card}\{\xi \in \mathcal{N}\backslash\mathcal{N}_{\rm int} \mid \omega_z \cap \omega_\xi \neq 0\}\,,$$

V. the shape of the elements $T \in \mathcal{T}$,

VI. the value $\max_{z\in\mathcal{N}} h_z\|\nabla\varphi_z\|_\infty$, where $\|\cdot\|_\infty$ denotes the $L^\infty(\Omega)$-norm and $h_z = \operatorname{diam}(\omega_z)$,

VII. the value

$$M_2 \equiv \operatorname*{ess\,sup}_{x\in\Omega} \{h(x)/h_T \mid x \in T \in \mathcal{T}\}\,,$$

where $h(x) \equiv \max\{h_z \mid \varphi_z(x) > 0, z \in \mathcal{N}_{\rm int}\}$, $h_T = \operatorname{diam}(T)$.

The proofs of the lemmas use the Poincaré inequality on ω_z and the Friedrichs inequality on $\omega_{z\partial\Omega}$ defined in **II.** In order to prove Lemma 2.2, the so-called trace theorem (see, e.g., [4, Proposition 4.1]) is used on each element of the triangulation $T \in \mathcal{T}$; the multiplicative factor then depends on the shape of the elements; see **V.** The quantities $\max_{z\in\mathcal{N}} h_z\|\nabla\varphi_z\|_\infty$ and M_2 (see **VI.** and **VII.**) are of the order one on a shape-regular mesh, where $\|\nabla\varphi_z|_T\|_\infty \approx h_T^{-1}$ and $h_z \approx h_T, T \in \omega_z$. They deteriorate on a mesh consisting of triangles with small inner angles, where small and large elements (in the sense of their diameter) adjoint. In order to see the development of the argument, for clarity we present the following two lemmas.

Lemma 2.1 ([4, Theorem 3.1, statement 1.]). *There exists a multiplicative factor $C > 0$ depending on the triangulation $\mathcal{T}$ (more precisely on **I.**–**IV.**), but not on the size of the elements h_T, such that, for all $R \in L^2(\Omega)$, for all $w \in H_0^1(\Omega)$ and for arbitrary numbers $R_z \in \mathbb{R}$, $z \in \mathcal{N}_{\rm int}$,*

$$\int_\Omega R\,(w - \mathcal{I}w) \le C\|\nabla w\| \left\{ \sum_{z\in\mathcal{N}_{\rm int}} h_z^2 \int_{\omega_z} \varphi_z/\psi \; |R - R_z|^2 \right\}^{1/2} .$$

Lemma 2.1 is a consequence of the definition of the quasi-interpolation operator $\mathcal{I}$; see (2.1).

Lemma 2.2 ([4, Theorem 3.2]). *Let $S \subset \mathcal{E}$. There exists a multiplicative factor $C > 0$ depending on the triangulation $\mathcal{T}$ (more precisely on **I.**–**VII.**), but not on the size of the elements h_T, such that for all $J \in L^2(S)$ and for all $w \in H_0^1(\Omega)$,*

$$\int_S J\,(w - \mathcal{I}w) \le C\|\nabla w\| \left(\sum_{T\in\mathcal{T}} h_T \|J\|^2_{S\cap\partial T} \right)^{1/2} .$$

Combining Lemmas 2.1 and 2.2 we get the final inequality.

Lemma 2.3 ([4, Corollary 3.1]). *Let $S \subset \mathcal{E}$. There exists a multiplicative factor $C_1 > 0$ depending on **I.**–**VII.** such that, for all $J \in L^2(S)$, for all $R \in L^2(\Omega)$, for all $w \in H_0^1(\Omega)$, and for arbitrary numbers $R_z \in \mathbb{R}, z \in \mathcal{N}_{\rm int}$,*

$$\begin{aligned} &\int_\Omega R\,(w - \mathcal{I}w) + \int_S J\,(w - \mathcal{I}w) \\ &\qquad \le C_1\|\nabla w\| \left\{ \sum_{z\in\mathcal{N}_{\rm int}} h_z^2 \|R - R_z\|^2_{\omega_z} + \sum_{T\in\mathcal{T}} h_T \|J\|^2_{S\cap\partial T} \right\}^{1/2} . \end{aligned}$$

The following lemma introduces a positive multiplicative factor $C_{\rm intp}$ that plays a key role in our discussion on incorporating the algebraic error into the a posteriori bound on the total error; see Section 3 and the numerical experiments in Section 5.

Lemma 2.4 ([4, Theorem 3.1, statement 3.]). *There exists a multiplicative factor $C_{\rm intp} > 0$ depending on the triangulation $\mathcal{T}$ (more precisely on **I.**–**IV.** and **VI.**) such that, for all $w \in H_0^1(\Omega)$,*

$$\|\nabla\mathcal{I}w\| \le C_{\rm intp}\|\nabla w\| . \tag{2.2}$$

Remark 2.1. *Using the proof of [5, Theorem 2.4] and the discussion in [5, Example 2.3], we can get a better idea about the size of* C_{intp}. *For a shape-regular mesh with* $\max_{z\in\mathcal{N}} h_z\|\nabla\varphi_z\|_\infty \approx 2$ *(see* ***VI.****), there holds* $C_{\mathrm{intp}} \approx 6$. *In general, as stated in [5], it may be very large for small angles in the triangulation.*

For $f \in L^1(\Omega)$ define the mean-value operator $\pi_{\omega_z}(f) \equiv \int_{\omega_z} f/|\omega_z|$. We denote, for $z \in \mathcal{N}$ and for any subset $Z \subset \mathcal{N}$,

$$\mathrm{osc}_z \equiv |\omega_z|^{1/2}\|f - \pi_{\omega_z} f\|_{\omega_z}\,, \qquad \mathrm{osc}(Z) \equiv \left(\sum_{z\in Z} \mathrm{osc}_z^2\right)^{1/2},$$

measuring the oscillations of f, i.e. the variations of the function f from the mean value $\pi_{\omega_z} f$ on the subdomains ω_z. Given $v_h \in V_h$, define for $E \in \mathcal{E}_{\mathrm{int}}$ and any subset $F \subset \mathcal{E}_{\mathrm{int}}$ the edge residuals

$$J_E(v_h) \equiv |E|^{1/2}\left\|\left[\frac{\partial v_h}{\partial n_E}\right]\right\|_E, \qquad J(v_h, F) \equiv \left(\sum_{E\in F} J_E^2(v_h)\right)^{1/2},$$

where $[\cdot]$ denotes the jump of a piecewise continuous function and n_E denotes the unit normal to E (for each $E \in \mathcal{E}_{\mathrm{int}}$ the orientation of the unit normal is set arbitrarily but fixed). The edge residual $J_E(v_h), v_h \in V_h$, measures the jump of the piecewise constant gradient ∇v_h over the inner edge E. We set for brevity $\mathrm{osc} \equiv \mathrm{osc}(\mathcal{N})$ and $J(v_h) \equiv J(v_h, \mathcal{E}_{\mathrm{int}})$. For a given $v_h \in V_h$, we define the jump function $\mathcal{J}(v_h) \in L^2(\mathcal{E}_{\mathrm{int}})$ on the inner edges

$$\mathcal{J}(v_h)|_E \equiv \left[\frac{\partial v_h}{\partial n_E}\right], \quad E \in \mathcal{E}_{\mathrm{int}}\,. \tag{2.3}$$

The Green's formula (see, e.g., [7, p. 14]) gives for a domain D with a Lipschitz continuous boundary ∂D and for $v \in H^2(D)$, $w \in H^1(D)$

$$\int_D \nabla v\cdot\nabla w = -\int_D \Delta v\, w + \int_{\partial D}\left(\frac{\partial v}{\partial n_{\partial D}}\right) w\,, \tag{2.4}$$

where $n_{\partial D}$ denotes the unit normal to ∂D pointing outwards. Let $v_h \in V_h$ and $T \in \mathcal{T}$. Then $v_h|_T$ is a linear function, $v_h|_T \in H^2(T)$ and $\Delta v_h|_T = 0$. Then, applying the Green's formula (2.4) elementwise yields, for any $v_h \in V_h$, $w \in H_0^1(\Omega)$,

$$\begin{aligned}\int_\Omega \nabla v_h\cdot\nabla w &= \sum_{T\in\mathcal{T}}\int_T \nabla v_h\cdot\nabla w = \sum_{T\in\mathcal{T}}\left(-\int_T \Delta v_h\, w + \int_{\partial T}\left(\frac{\partial v_h}{\partial n_{\partial T}}\right) w\right)\\ &= \sum_{E\in\mathcal{E}_{\mathrm{int}}}\int_E \left[\frac{\partial v_h}{\partial n_E}\right] w = \int_{\mathcal{E}_{\mathrm{int}}} \mathcal{J}(v_h)\, w\,.\end{aligned} \tag{2.5}$$

The results recalled in this section are used to prove the upper bound on the total error in the next section.

3 Upper bound on the total error

Now we state the upper bound on the energy norm of the total error using the residual-based a posteriori error estimator.

Theorem 3.1. *There exist triangulation-dependent positive multiplicative factors $C_1, C_{\rm intp}$, and C_2 such that for the solution u of (1.2), the Galerkin solution u_h of (1.3), and an arbitrary $v_h \in V_h$,*

$$\|\nabla(u - v_h)\|^2 \le 2\, C_1^2\, C_2^2 \left(J^2(v_h) + \mathrm{osc}^2\right) + 2\, C_{\rm intp}^2 \|\nabla(u_h - v_h)\|^2 . \tag{3.1}$$

*In particular, C_1 depends on **I.**–**VII.** (see Lemma 2.3), $C_{\rm intp}$ depends on **I.**–**IV.** and **VI.** (see Lemma 2.4), and the factor C_2 depends on the ratios $h_z^2/|\omega_z|$, $z \in \mathcal{N}_{\rm int}$, and $h_T/|E|$, $T \in \mathcal{T}$, $E \in \partial T \cap \mathcal{E}_{\rm int}$.*

Proof. We will use the standard expression for the norm

$$\|\nabla(u - v_h)\| = \sup_{0 \ne w \in H_0^1(\Omega)} \frac{1}{\|\nabla w\|} \int_\Omega \nabla(u - v_h) \cdot \nabla w \,. \tag{3.2}$$

Let $v_h \in V_h$ and $w \in H_0^1(\Omega), w \ne 0$, be arbitrary,

$$\begin{aligned}
\int_\Omega \nabla(u - v_h) \cdot \nabla w &= \int_\Omega \nabla(u - v_h) \cdot \nabla(w - \mathcal{I}w) + \int_\Omega \nabla(u - v_h) \cdot \nabla \mathcal{I}w \\
&= \int_\Omega \nabla(u - v_h) \cdot \nabla(w - \mathcal{I}w) + \int_\Omega \nabla(u - u_h) \cdot \nabla \mathcal{I}w \\
&\quad + \int_\Omega \nabla(u_h - v_h) \cdot \nabla \mathcal{I}w \,.
\end{aligned}$$

It follows from the definition of the interpolation operator that $\mathcal{I}w \in V_h$. The Galerkin orthogonality (1.4) gives

$$\int_\Omega \nabla(u - u_h) \cdot \nabla \mathcal{I}w = 0.$$

Then

$$\begin{aligned}
\int_\Omega \nabla(u - v_h) \cdot \nabla w &= \int_\Omega \nabla(u - v_h) \cdot \nabla(w - \mathcal{I}w) + \int_\Omega \nabla(u_h - v_h) \cdot \nabla \mathcal{I}w \\
&= \int_\Omega \nabla u \cdot \nabla(w - \mathcal{I}w) - \int_\Omega \nabla v_h \cdot \nabla(w - \mathcal{I}w) \\
&\quad + \int_\Omega \nabla(u_h - v_h) \cdot \nabla \mathcal{I}w \,.
\end{aligned}$$

Using the weak formulation (1.2) and the equality (2.5),

$$\int_\Omega \nabla(u - v_h) \cdot \nabla w = \int_\Omega f(w - \mathcal{I}w) - \int_{\mathcal{E}_{\rm int}} \mathcal{J}(v_h)(w - \mathcal{I}w) + \int_\Omega \nabla(u_h - v_h) \cdot \nabla \mathcal{I}w \,.$$

Then Lemma 2.3 with $S = \mathcal{E}_{\rm int}$, $R = f$, $R_z = \pi_{\omega_z} f,\ z \in \mathcal{N}_{\rm int}$, $J = -\mathcal{J}(v_h)$ gives

$$\begin{aligned}\int_\Omega \nabla(u - v_h)\cdot\nabla w &\le C_1\|\nabla w\| \left\{ \sum_{T\in\mathcal{T}} h_T \|\mathcal{J}(v_h)\|^2_{\mathcal{E}_{\rm int}\cap\partial T} + \sum_{z\in\mathcal{N}_{\rm int}} h_z^2 \|f - \pi_{\omega_z} f\|^2_{\omega_z} \right\}^{\frac12} \\ &\quad + \int_\Omega \nabla(u_h - v_h)\cdot\nabla \mathcal{I} w \\ &\le C_1 C_2 \|\nabla w\| \left\{ \sum_{E\in\mathcal{E}_{\rm int}} |E|\, \|\mathcal{J}(v_h)\|^2_E + \sum_{z\in\mathcal{N}_{\rm int}} |\omega_z|\, \|f - \pi_{\omega_z} f\|^2_{\omega_z} \right\}^{\frac12} \\ &\quad + \int_\Omega \nabla(u_h - v_h)\cdot\nabla \mathcal{I} w \\ &= C_1 C_2 \|\nabla w\| \left(J^2(v_h) + \mathrm{osc}^2\right)^{1/2} + \int_\Omega \nabla(u_h - v_h)\cdot\nabla \mathcal{I} w\,.\end{aligned}$$

Using the Cauchy-Schwarz inequality,

$$\int_\Omega \nabla(u - v_h)\cdot\nabla w \le C_1 C_2 \|\nabla w\| \left(J^2(v_h) + \mathrm{osc}^2\right)^{1/2} + \|\nabla \mathcal{I} w\|\, \|\nabla(u_h - v_h)\|\,. \tag{3.3}$$

Dividing (3.3) by $\|\nabla w\|$ and using Lemma 2.4,

$$\begin{aligned}\frac{1}{\|\nabla w\|}\int_\Omega \nabla(u - v_h)\cdot\nabla w &\le C_1 C_2 \left(J^2(v_h) + \mathrm{osc}^2\right)^{1/2} + \frac{\|\nabla \mathcal{I} w\|}{\|\nabla w\|}\|\nabla(u_h - v_h)\| \\ &\le C_1 C_2 \left(J^2(v_h) + \mathrm{osc}^2\right)^{1/2} + C_{\rm intp}\|\nabla(u_h - v_h)\|\,.\end{aligned}$$

Using the representation (3.2) of the energy norm, recall that $w \in H^1_0(\Omega)$ was chosen arbitrarily,

$$\|\nabla(u - v_h)\| \le C_1 C_2 \left(J^2(v_h) + \mathrm{osc}^2\right)^{1/2} + C_{\rm intp}\|\nabla(u_h - v_h)\|\,.$$

Finally, we deduce (3.1) using the inequality $(a+b)^2 \le 2a^2 + 2b^2$. □

We will now give another bound on the total error that will contain information about the exact solution u of (1.2). It is useful for illustration of the role of the factor $C_{\rm intp}$ in (3.1). Setting $w \equiv u - v_h \in H^1_0(\Omega)$ in (3.3), we get

$$\begin{aligned}\|\nabla(u - v_h)\|^2 \le C_1 C_2 \|\nabla(u - v_h)\| &\left(J^2(v_h) + \mathrm{osc}^2\right)^{1/2} \\ &+ \|\nabla(\mathcal{I}u - \mathcal{I}v_h)\|\, \|\nabla(u_h - v_h)\|\,.\end{aligned}$$

Dividing both sides by $\|\nabla(u - v_h)\|$,

$$\|\nabla(u - v_h)\| \le C_1 C_2 \left(J^2(v_h) + \mathrm{osc}^2\right)^{1/2} + \frac{\|\nabla(\mathcal{I}u - \mathcal{I}v_h)\|}{\|\nabla(u - v_h)\|}\ \|\nabla(u_h - v_h)\| \tag{3.4}$$

and therefore (3.1) holds also with the multiplicative factor

$$\widetilde{C}_{\rm intp}(v_h) \equiv \frac{\|\nabla(\mathcal{I}u - \mathcal{I}v_h)\|}{\|\nabla(u - v_h)\|} \tag{3.5}$$

in place of $C_{\rm intp}$. We note that $\widetilde{C}_{\rm intp}(v_h)$ depends on the solution u of (1.2). From the definition of the (solution-independent) factor $C_{\rm intp}$, we have

$\widetilde{C}_{\rm intp}(v) \le C_{\rm intp}$, for all $v \in V$. Therefore $C_{\rm intp}$ represents a worst-case scenario factor and one may expect that most likely $\widetilde{C}_{\rm intp}(v_h) \ll C_{\rm intp}$, $v_h \in V_h$.

The argument in [3, Lemma 3.1] (see (1.8)) seems to be based on the triangle inequality

$$\begin{aligned}\|\nabla(u-v_h)\|^2 &\le \big(\|\nabla(u-u_h)\| + \|\nabla(u_h-v_h)\|\big)^2 \\ &\le 2\,\|\nabla(u-u_h)\|^2 + 2\,\|\nabla(u_h-v_h)\|^2\end{aligned}$$

with the subsequent step $\|\nabla(u-u_h)\|^2 \le \widetilde{C}\cdot \mathrm{EST}(u_h)$, where on the right-hand side of the last inequality $\mathrm{EST}(u_h)$ is replaced by $\mathrm{EST}(v_h)$. As explained in the introduction, a justification for this step seems to be missing.

4 A related result

In [2] the authors consider elliptic self-adjoint problems and they use a residual-based error estimator for setting the stopping criterion for the conjugate gradient method. Following their approach, one can easily get an upper bound on the total error. Although the bound is not stated explicitly in [2], it appears in the proof of Theorem 3.3; see the inequality [2, (3.22)]. The derivation proceeds differently from the proof of the bound (3.1) and we present it here to demonstrate that the verification of the assumption (1.6) represents a challenge.

First, [2, Theorem 2.2] recalls the bound on the discretization error: there exists a multiplicative factor $C_{2.2} > 0$ independent of $\mathcal{T}$, h, u, and u_h, such that

$$\|\nabla(u-u_h)\|^2 \le C_{2.2}\,\eta^2(u_h)\,, \qquad \eta(u_h) \equiv \left(\sum_{T\in\mathcal{T}} |T|\,\|f+\Delta u_h\|_T^2 + (J(u_h))^2\right)^{1/2};$$

see, e.g., [14, 1]. Using the standard inverse estimates, [2, Lemma 3.1] yields the inequality

$$\eta^2(w_h) \le (1+\gamma)\,\eta^2(v_h) + C_{3.1}(1+\gamma^{-1})\|\nabla(v_h-w_h)\|^2, \quad \text{for all } w_h, v_h \in V_h, \gamma > 0\,,$$

where the positive factor $C_{3.1}$ depends on the shape-regularity of the mesh. Then, by combining these bounds and the equality

$$\|\nabla(u-v_h)\|^2 = \|\nabla(u-u_h)\|^2 + \|\nabla(u_h-v_h)\|^2$$

that follows from the Galerkin orthogonality (1.4), we get the upper bound on the total error

$$\begin{aligned}\|\nabla(u-v_h)\|^2 &\le C_{2.2}\,\eta^2(u_h) + \|\nabla(u_h-v_h)\|^2 \\ &\le C_{2.2}\,(1+\gamma)\,\eta^2(v_h) + \big(1 + C_{2.2}\,C_{3.1}\,(1+\gamma^{-1})\big)\,\|\nabla(u_h-v_h)\|^2\end{aligned}$$

Finally, by setting $\gamma \equiv 1$,

$$\|\nabla(u-v_h)\|^2 \le 2\,C_{2.2}\,\eta^2(v_h) + (1 + 2\,C_{2.2}\,C_{3.1})\,\|\nabla(u_h-v_h)\|^2\,. \tag{4.1}$$

In the "practical" criteria proposed in [2, Section 5] for numerical experiments the factors are empirically set to $C_{2.2} \equiv 40$, $C_{3.1} \equiv 10$, giving $(1 + 2\,C_{2.2}\,C_{3.1}) = 801$; cf. (1.8). This nicely underlines the subtleties of the residual-based bounds discussed above.

5 Numerical illustrations

We use, *on purpose*, very simple problems to illustrate the possible difference in the values of $\widetilde{C}_{\mathrm{intp}}(v_h)$ and C_{intp}. While $\widetilde{C}_{\mathrm{intp}}(v_h)$ can be, assuming the knowledge of the exact solution u, evaluated up to a negligible quadrature error, for the factor C_{intp} we present a lower bound given by plugging a chosen function into (2.2). The derivation of a more accurate estimate for C_{intp} (see also the discussion in Remark 2.1) is beyond the scope of this paper.

5.1 Numerical illustration in one dimension

We first consider a one-dimensional analogue of the Clément-type quasi-interpolation operator $\mathcal{I}$ to illustrate that C_{intp} can be significantly larger than one.

Consider the domain $\Omega = (0,1)$ with the (non-uniform) partition

$$[0, \beta, 1/3 \pm \beta, 2/3 \pm \beta, 1-\beta, 1]; \quad \beta = 0.01.$$

This partition is adapted to the 1D Laplace problem with the solution

$$\begin{aligned} u(x) \;=\; & \tan^{-1}(cx) - \tan^{-1}(c(x-1/3)) + \tan^{-1}(c(x-2/3)) \\ & - \tan^{-1}(c(x-1)) - \tan^{-1}(c/3) + \tan^{-1}(2c/3) - \tan^{-1}(c)\,, \end{aligned} \tag{5.1}$$

with $c = 1000$. The left part of Figure 1 depicts the solution u and the Clément-type quasi-interpolant $\mathcal{I}u$. For a zero approximate vector, we have

$$\widetilde{C}_{\mathrm{intp}}(0) = \frac{\|(\mathcal{I}u)'\|}{\|u'\|} = 0.77. \tag{5.2}$$

For a quadratic function $w(x) = x(1-x)$, $w \in H^1_0(\Omega)$, we have

$$\frac{\|(\mathcal{I}w)'\|}{\|w'\|} = 3.70;$$

see the right part of Figure 1 for the plot of the function w and the interpolant $\mathcal{I}w$. Consequently, $C_{\mathrm{intp}} \geq 3.70$.

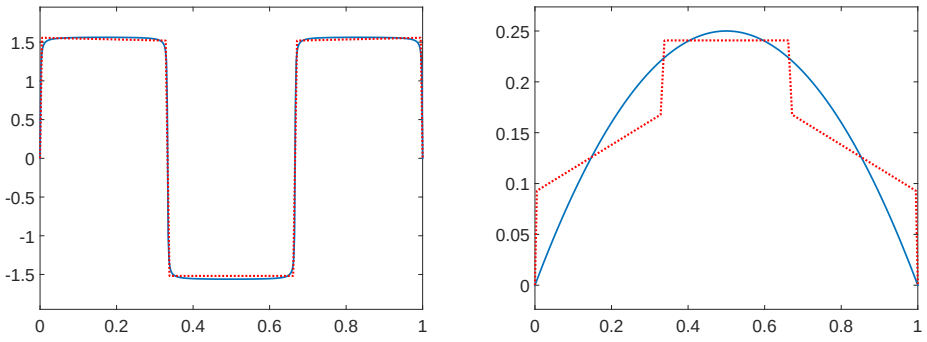


Figure 1: Left: the solution u (5.1) (solid line) and the interpolant $\mathcal{I}u$ (dotted line). Right: the function $w(x) = x(1-x)$ (solid line) and $\mathcal{I}w$ (dotted line).

5.2 Two-dimensional numerical illustration

For two-dimensional numerical illustration we consider the square domain $\Omega \equiv (-1,1) \times (-1,1)$ and the triangulation $\mathcal{T}$ generated by MATLAB[1] command `initmesh('squareg', 'Hmax', 0.1)` that provides a Delaunay triangulation consisting of 1 368 elements with the maximal diameter less than or equal to 0.1. The minimal angle of the mesh is 35.9° and the average of the minimal angles of the elements is 50.3°.

Consider the solution of problem (1.1):

$$u^{(1)}(x,y) = (x-1)(x+1)(y-1)(y+1)\,. \tag{5.3}$$

For the zero approximate solution and the Galerkin solution $u_h^{(1)}$ corresponding to $u^{(1)}$, we have

$$\widetilde{C}_{\mathrm{intp}}(0) = 1.02\,, \qquad \widetilde{C}_{\mathrm{intp}}(u_h^{(1)}) = 0.16\,.$$

Similarly, for the exact solution

$$u^{(2)}(x,y) = (x-1)(x+1)(y-1)(y+1)\exp(-100(x^2+y^2))\,, \tag{5.4}$$

we have

$$\widetilde{C}_{\mathrm{intp}}(0) = 0.76\,, \qquad \widetilde{C}_{\mathrm{intp}}(u_h^{(2)}) = 0.28\,.$$

In Figure 2 we show the values of $\widetilde{C}_{\mathrm{intp}}(v_h)$ for v_h generated in conjugate gradient iterations with zero initial vector for solving the linear algebraic systems corresponding to the discretization of (1.2) with the solutions $u^{(1)}$, $u^{(2)}$ defined above.

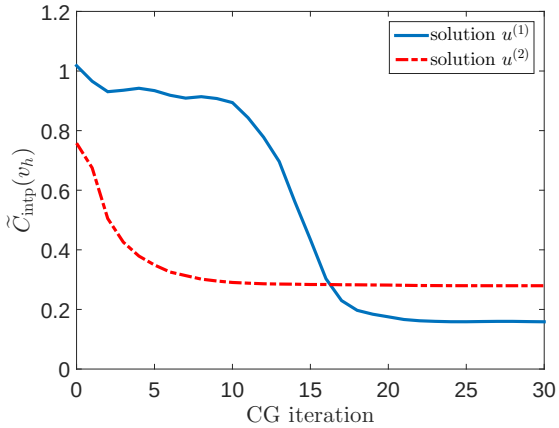


Figure 2: The values of $\widetilde{C}_{\mathrm{intp}}(v_h)$ for v_h generated in conjugate gradient iterations with zero initial vector for solving the linear algebraic systems corresponding to the discretization of (1.2) with the solutions $u^{(1)}$, $u^{(2)}$; see (5.3) and (5.4) respectively.

To bound the constant C_{intp} from below we consider $w_h \in V_h$ such that

$$w_h(z) = 1, \quad z \in \mathcal{N}_{\mathrm{int}}, \qquad w_h = 0 \quad \text{on } \partial\Omega\,. \tag{5.5}$$

[1] using the Partial Differential Equation Toolbox

For this function

$$1.10 = \frac{\|\nabla \mathcal{I} w_h\|}{\|\nabla w_h\|} \leq C_{\mathrm{intp}}\,.$$

Figure 3 gives the difference $w_h - \mathcal{I}w_h$ that is on the machine precision level in most of the domain except patches around the nodes adjacent to the boundary $\partial\Omega$. We recall that the proof of Lemma 2.4 gives for this simple problem and a shape-regular mesh $C_{\mathrm{intp}} \approx 6$; see Remark 2.1 and the original paper [5]. For the value $\widetilde{C}_{\mathrm{intp}}(v_h)$, it can therefore indeed hold $C_{\mathrm{intp}} \gg \widetilde{C}_{\mathrm{intp}}(v_h)$.

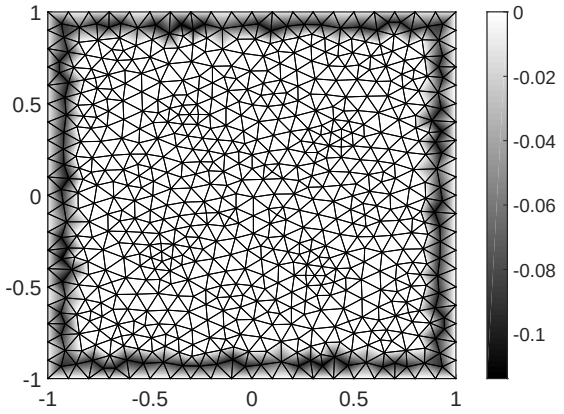


Figure 3: The difference $w_h - \mathcal{I}w_h$ for w_h given by (5.5).

Conclusion

This paper provides the detailed proof of a residual-based upper bound on the total approximation error with focus on the multiplicative factors. In particular, we show that abandoning the Galerkin orthogonality assumption from the derivation leads to an *additional multiplicative factor* in the estimator that scales the contribution of the algebraic error; see (3.1) and (3.4)–(3.5). The factor $\widetilde{C}_{\mathrm{intp}}(v_h)$ depends, besides v_h, also on the unknown infinite-dimensional solution u of (1.2). This generally uncomputable *a posteriori* factor $\widetilde{C}_{\mathrm{intp}}(v_h)$ can be bounded, using the *a priori* information, by the solution-independent factor C_{intp} given by (2.2). The value of C_{intp} can overestimate the value of $\widetilde{C}_{\mathrm{intp}}(v_h)$. In practical problems that are solved using an adaptive mesh refinement algorithm with small angles in the triangulation, the overestimation can be substantial.

References

[1] M. Ainsworth and J. T. Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2000.

[2] M. Arioli, E. H. Georgoulis, and D. Loghin, *Stopping criteria for adaptive finite element solvers*, SIAM J. Sci. Comput., 35 (2013), pp. A1537–A1559.

[3] R. Becker and S. Mao, *Convergence and quasi-optimal complexity of a simple adaptive finite element method*, M2AN Math. Model. Numer. Anal., 43 (2009), pp. 1203–1219.

[4] C. Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1187–1202.

[5] ——, *Clément interpolation and its role in adaptive finite element error control*, in Partial differential equations and functional analysis, vol. 168 of Oper. Theory Adv. Appl., Birkhäuser, Basel, 2006, pp. 27–43.

[6] C. Carstensen, M. Feischl, M. Page, and D. Praetorius, *Axioms of adaptivity*, Comput. Math. Appl., 67 (2014), pp. 1195–1253.

[7] P. G. Ciarlet, *The finite element method for elliptic problems*, vol. 40 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam].

[8] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.

[9] E. Keilegavlen and J. M. Nordbotten, *Inexact linear solvers for control volume discretizations in porous media*, Comput. Geosci., 19 (2015), pp. 159–176.

[10] A. Nissen, P. Pettersson, E. Keilegavlen, and J. M. Nordbotten, *Incorporating geological uncertainty in error control for linear solvers*, in SPE Reservoir Simulation Symposium, Society of Petroleum Engineers, 2015.

[11] J. M. Nordbotten and P. E. Bjørstad, *On the relationship between the multiscale finite-volume method and domain decomposition preconditioners*, Comput. Geosci., 12 (2008), pp. 367–376.

[12] J. Papež, J. Liesen, and Z. Strakoš, *Distribution of the discretization and algebraic error in numerical solution of partial differential equations*, Linear Algebra Appl., 449 (2014), pp. 89–114.

[13] J. Papež, Z. Strakoš, and M. Vohralík, *Estimating and localizing the algebraic and total numerical errors using flux reconstructions.* Preprint MORE/2016/12, Submitted for publication, 2016.

[14] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Advances in Numerical Mathematics Series, WileyTeubner, 1996.

4.2 Algebraic error in the adaptive finite element method

In this section we numerically illustrate the influence of the algebraic error on the residual-based error indicators and on the adaptive mesh refinement. For the illustrations we consider **Inhomogeneous tensor problems I** and **II** of Section 2.2 (see also the references therein) and the **L-shape problem**

$$-\Delta u = 0 \quad \text{in } \Omega \qquad\qquad u = u_D \quad \text{on } \partial\Omega, \tag{4.1}$$

where $\Omega \equiv (-1,1)\times(-1,1)\setminus(0,1)\times(-1,0)$ and the Dirichlet boundary condition u_D is such that the solution is given in polar coordinates (r,θ) by

$$u(r,\theta) = r^{2/3}\sin(2\theta/3)\,;$$

see, e.g., Luce and Wohlmuth [2004]; Ainsworth [2005]; Papež et al. [2014].

Given a discretization mesh $\mathcal{T}$, we denote by $u^*_{\mathcal{T}}$ the piecewise affine approximation corresponding to MATLAB backslash approximation $\mathbf{x}^*$ to the solution of the algebraic system stemming from the discretization of the problem on $\mathcal{T}$. In the experiments, $u^*_{\mathcal{T}}$ provides sufficiently accurate approximation to the corresponding FEM discrete solution.

The sequence $\{\mathcal{T}^*_\ell\}$ of adaptively refined meshes is generated as in [Papež et al., 2014, Section 5] and in Section 2.2 (where the following notation is adopted from). This means that we run adaptive finite element method where we estimate the energy norm of the (discretization) error on the elements using the residual-based local error estimators (indicators) for $u^*_{\mathcal{T}^*_\ell}$,

$$\eta^2_{R,T}(u^*_{\mathcal{T}^*_\ell}) = \frac{h_T^2}{s_T}\|f\|^2_{L^2(T)} + \sum_{E\subset\partial T}\frac{h_E}{s_E}\|[\mathbf{S}\nabla u^*_{\mathcal{T}^*_\ell}\cdot n_E]\|^2_{L^2(E)}\,; \tag{4.2}$$

see, e.g., [Carstensen and Merdon, 2010, Section 2]. Recall that $\mathbf{S} = s_i\mathbf{I}$ on Ω_i (in the L-shape problem, $s_i = 1$ on Ω), $s_T = s_i$ for $T\subset\Omega_i$, $s_E = \max\{s_T \mid E\in\partial T\}$. In the considered model problems the source term $f = 0$ and the first term on the right-hand side of (4.2) vanishes. The second term is the generalization of the jump term $J(u^*_{\mathcal{T}^*_\ell})$ from the paper Papež and Strakoš [2016] included in Section 4.1 for the problems with inhomogeneous diffusion tensor $\mathbf{S}$.

Marking and refinement of the elements is as described in [Papež et al., 2014, Section 5]. In particular, the mesh elements are ordered such that

$$\eta^2_{R,T_1}(u^*_{\mathcal{T}^*_\ell}) \geq \eta^2_{R,T_2}(u^*_{\mathcal{T}^*_\ell}) \geq \cdots \geq \eta^2_{R,T_M}(u^*_{\mathcal{T}^*_\ell}),$$

where M is the number of elements in the triangulation $\mathcal{T}^*_\ell$, and we mark for refinement the elements $T_1,\ldots,T_m$ where m is the smallest index such that

$$\sum_{j=1}^{m}\eta^2_{R,T_j}(u^*_{\mathcal{T}^*_\ell}) \geq \Theta\left(\sum_{j=1}^{m}\eta^2_{R,T}(u^*_{\mathcal{T}^*_\ell}) + \sum_{j=m+1}^{M}\eta^2_{R,T}(u^*_{\mathcal{T}^*_\ell})\right), \qquad \Theta = 0.25.$$

The whole procedure is denoted by AFEM(*).

The second sequence of meshes denoted by $\{\mathcal{T}^{CG}_\ell\}$ is generated by AFEM with the conjugate gradient method applied for solving the system of the linear

algebraic equations. On each level corresponding to the mesh $\mathcal{T}_\ell^{CG}$ we stop the CG iterations (using zero initial guess) when the k-th approximation $\mathbf{x}_{CG(k)}$ satisfies

$$\|\mathbf{x}^* - \mathbf{x}_{CG(k)}\|_{\mathbf{A}}^2 < 0.01\, \|\mathbf{S}^{1/2}\nabla(u - u^*_{\mathcal{T}_\ell^{CG}})\|^2 \tag{4.3}$$

and denote by $u^{CG}_{\mathcal{T}_\ell^{CG}}$ the approximation given by the coefficient vector $\mathbf{x}_{CG(k)}$. The marking strategy and refinement in AFEM(CG) is analogous to AFEM(*), however, it is based on the local error indicators $\eta^2_{R,T}\big(u^{CG}_{\mathcal{T}_\ell^{CG}}\big)$ given by (4.2) with replacing $u^*_{\mathcal{T}_\ell^*}$ by the approximation $u^{CG}_{\mathcal{T}_\ell^{CG}}$.

Remark: We emphasize that the error indicators (4.2) are derived for estimating the energy norm of the *discretization* error $u - u_{\mathcal{T}_\ell}$; see the seminal paper Babuška and Rheinboldt [1978]. There exists an (unspecified) multiplicative factor C such that

$$\begin{aligned}&\|\mathbf{S}^{1/2}\nabla(u - u_{\mathcal{T}_\ell})\| \\ &\qquad \le C\left\{\left(\sum_{K\in\mathcal{T}_\ell}\frac{h_T^2}{s_T}\|f\|^2_{L^2(T)}\right)^{1/2} + \left(\sum_{E\in\mathcal{E}}\frac{h_E}{s_E}\|[\mathbf{S}\nabla u_{\mathcal{T}_\ell}\cdot n_E]\|^2_{L^2(E)}\right)^{1/2}\right\}.\end{aligned} \tag{4.4}$$

The derivation of (4.4) uses the Galerkin orthogonality and the indicators are evaluated for the (unavailable) Galerkin solution $u_{\mathcal{T}_\ell}$. The impact of abandoning the assumption on the Galerkin orthogonality and plugging a computed approximation into (4.4) was thoroughly studied in Papež and Strakoš [2016]. However, the crucial question how the indicators (4.2) estimate the *local* distribution of the energy norm of the total (or discretization) error in the presence of a (nonnegligible) algebraic error is still open.

The simplest way how to compare the sequences of the meshes $\{\mathcal{T}_\ell^*\}$ and $\{\mathcal{T}_\ell^{CG}\}$ is to plot the values $\|\mathbf{S}^{1/2}\nabla(u - u^*_{\mathcal{T}_\ell^*})\|$, $\|\mathbf{S}^{1/2}\nabla(u - u^*_{\mathcal{T}_\ell^{CG}})\|$ that provide sufficiently tight approximations to the energy norm of the corresponding discretization errors. We then visualize also the (local) differences of the meshes given by AFEM(*) and AFEM(CG), respectively. Rather than comparing $\mathcal{T}_\ell^*$ and $\mathcal{T}_\ell^{CG}$ after a given number ℓ of AFEM steps, we show the comparison of the meshes $\mathcal{T}_\ell^*$, $\mathcal{T}_k^{CG}$ with a similar number of vertices. In the figures we depict in black the edges of the triangulation $\mathcal{T}_\ell^*$ that are not in the triangulation $\mathcal{T}_k^{CG}$ (denoted in captions by $\mathcal{T}_\ell^* \setminus \mathcal{T}_k^{CG}$). The edges of $\mathcal{T}_k^{CG}$ that are not in $\mathcal{T}_\ell^*$ are in red color and denoted by $\mathcal{T}_k^{CG} \setminus \mathcal{T}_\ell^*$.

L-shape problem

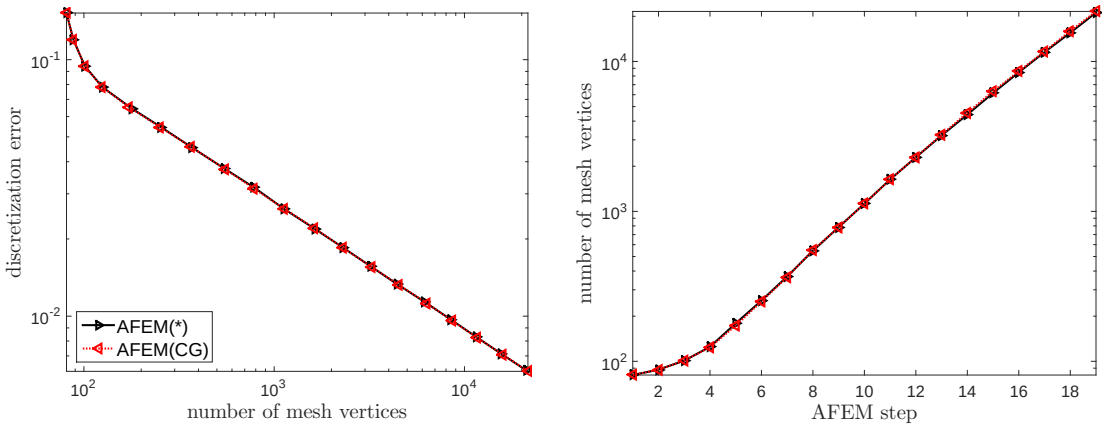


Figure 4.1: Left: the energy norm of the discretization error on the sequence of meshes generated by AFEM(*) and AFEM(CG) respectively. Right: the number of mesh vertices in AFEM steps.

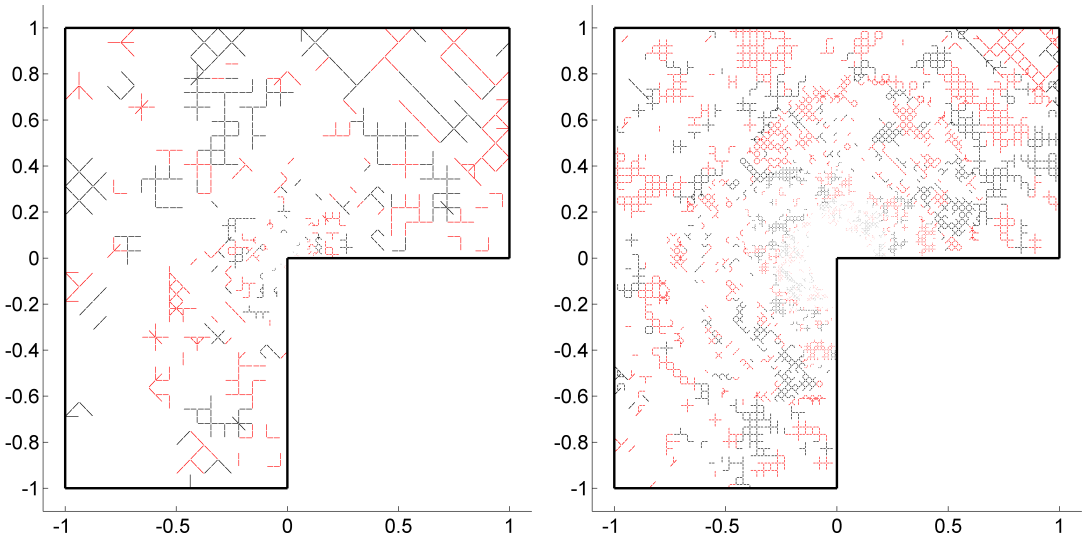


Figure 4.2: The difference of the adaptively refined meshes. Left: $\mathcal{T}_{13}^{*} \setminus \mathcal{T}_{13}^{CG}$ (black), $\mathcal{T}_{13}^{CG} \setminus \mathcal{T}_{13}^{*}$ (red). The mesh $\mathcal{T}_{13}^{*}$ consists of 3376 vertices, $\mathcal{T}_{13}^{CG}$ consists of 3413 vertices. Right: $\mathcal{T}_{19}^{*} \setminus \mathcal{T}_{19}^{CG}$ (black), $\mathcal{T}_{19}^{CG} \setminus \mathcal{T}_{19}^{*}$ (red). The mesh $\mathcal{T}_{19}^{*}$ consists of 21 575 vertices, $\mathcal{T}_{19}^{CG}$ consists of 21 967 vertices.

Inhomogeneous tensor problem I

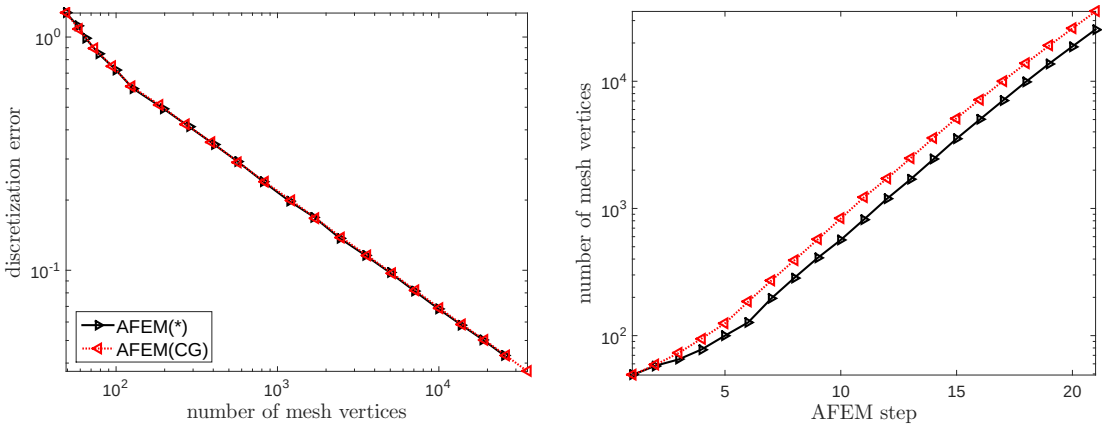


Figure 4.3: Left: the energy norm of the discretization error on the sequence of meshes generated by AFEM(*) and AFEM(CG) respectively. Right: the number of mesh vertices in AFEM steps.

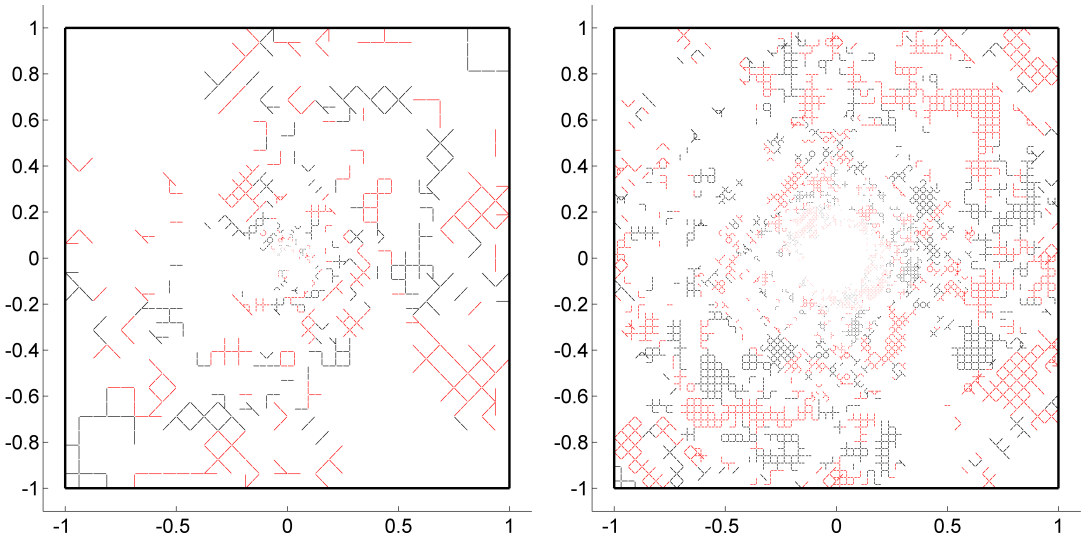


Figure 4.4: The difference of the adaptively refined meshes. Left: $\mathcal{T}_{15}^{*} \setminus \mathcal{T}_{14}^{CG}$ (black), $\mathcal{T}_{14}^{CG} \setminus \mathcal{T}_{15}^{*}$ (red). The mesh $\mathcal{T}_{15}^{*}$ consists of 3625 vertices, $\mathcal{T}_{14}^{CG}$ consists of 3674 vertices. Right: $\mathcal{T}_{21}^{*} \setminus \mathcal{T}_{20}^{CG}$ (black), $\mathcal{T}_{20}^{CG} \setminus \mathcal{T}_{21}^{*}$ (red). The mesh $\mathcal{T}_{21}^{*}$ consists of 25 780 vertices, $\mathcal{T}_{20}^{CG}$ consists of 26 283 vertices.

Inhomogeneous tensor problem II

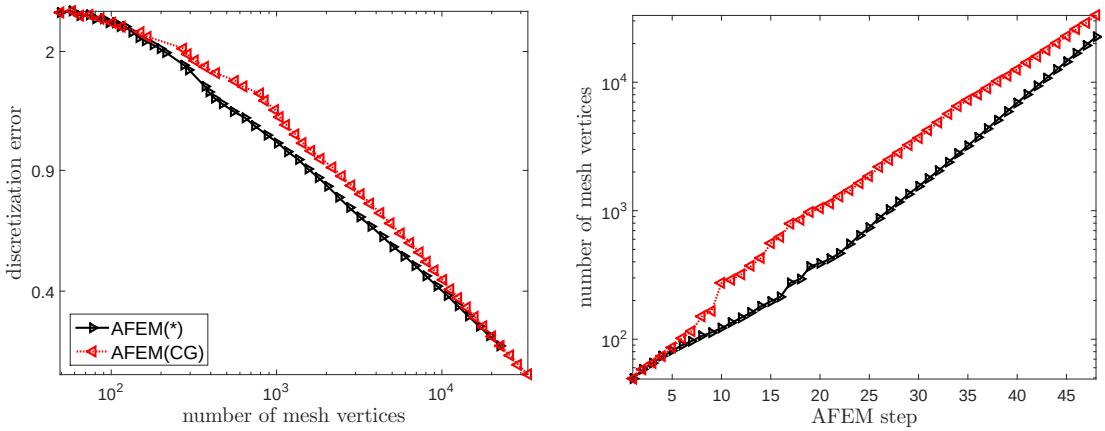


Figure 4.5: Left: the energy norm of the discretization error on the sequence of meshes generated by AFEM(*) and AFEM(CG) respectively. Right: the number of mesh vertices in AFEM steps.

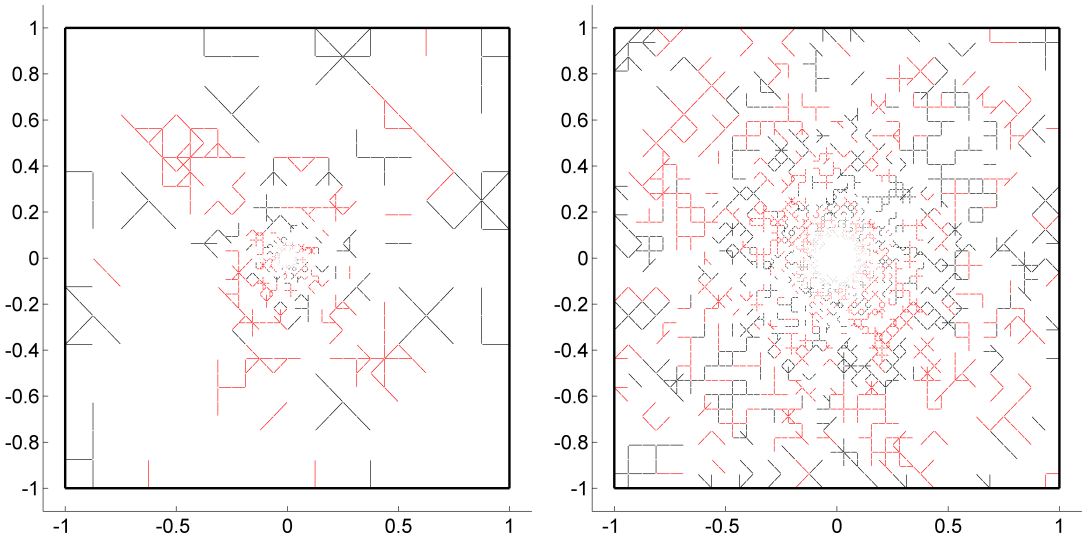


Figure 4.6: The difference of the adaptively refined meshes. Left: $\mathcal{T}_{35}^{*} \setminus \mathcal{T}_{29}^{CG}$ (black), $\mathcal{T}_{29}^{CG} \setminus \mathcal{T}_{35}^{*}$ (red). The mesh $\mathcal{T}_{35}^{*}$ consists of 3247 vertices, $\mathcal{T}_{29}^{CG}$ consists of 3270 vertices. Right: $\mathcal{T}_{48}^{*} \setminus \mathcal{T}_{45}^{CG}$ (black), $\mathcal{T}_{45}^{CG} \setminus \mathcal{T}_{48}^{*}$ (red). The mesh $\mathcal{T}_{48}^{*}$ consists of 22 648 vertices, $\mathcal{T}_{45}^{CG}$ consists of 22 836 vertices.

Observations

The effect of algebraic error on the adaptive refinement procedures is a complex and yet not fully described issue; a recent survey on interplay between the algebraic error and a posteriori error estimates in AFEM can be found, e.g., in Arioli et al. [2013]. The experiments of this section demonstrate that the sequences of meshes constructed adaptively in AFEM can significantly differ when the corresponding algebraic systems are solved with negligible and nonnegligible algebraic error; see the local differences of the meshes in Figures 4.2, 4.4 and 4.6.

Regarding the energy norm of the associated discretization errors (which can be understood as a *global* measure of the mesh quality), the test problems illustrate three possible situations:

1. We refine nearly the same number of elements in each step of AFEM(*) and AFEM(CG) and the elements that differ contribute to the overall norm of the discretization error almost equally. Therefore the norm of the discretization error is decreased in each step of AFEM(CG) very similarly to AFEM(*). This was observed in the experiment with the L-shape test problem; see Figure 4.1.
2. In AFEM(CG) steps we refine more elements than in AFEM(*) and all the refined elements contribute significantly to the overall norm of the discretization error. Then AFEM(CG) can converge as fast as AFEM(*), which means that the reduction of the energy norm of the discretization error against the number of vertices is the same for AFEM(CG) and AFEM(*); see the first AFEM steps for the Inhomogeneous tensor problem I in Figure 4.3. In this (favorable) situation, a smaller number of AFEM(CG) steps with comparison to AFEM(*) is necessary to decrease the discretization error below a prescribed tolerance.
3. The mesh is in AFEM(CG) refined (also) in parts of the domain that do not significantly contribute to the overall norm of discretization error. This slows down the convergence of the AFEM procedure and it makes the solution process less efficient. We observed such behavior in the initial AFEM steps for the Inhomogeneous tensor problem II; see Figure 4.5.

The results of the experiments for these simple model problems are disturbing. They suggest that in practical computations the effect of the algebraic error to adaptivity can be substantial. We believe that a thorough study of the influence of the algebraic error on local a posteriori error indicators and the adaptive refinement procedures is highly desirable. The choice of the indicators and the choice of stopping criteria should avoid the third situation described above, i.e. the refinement of the mesh in inappropriate parts of the domain. Otherwise, the efficiency of the whole adaptive procedure and also reaching a prescribed accuracy can be endangered.

Bibliography

M. Ainsworth. Robust a posteriori error estimation for nonconforming finite element approximation. *SIAM J. Numer. Anal.*, 42(6):2320–2341, 2005. ISSN 0036-1429.

M. Arioli, J. Liesen, A. Międlar, and Z. Strakoš. Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems. *GAMM-Mitt.*, 36(1):102–129, 2013. ISSN 0936-7195.

I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978. ISSN 0036-1429.

C. Carstensen and C. Merdon. Estimator competition for Poisson problems. *J. Comput. Math.*, 28(3):309–330, 2010. ISSN 0254-9409.

R. Luce and B. I. Wohlmuth. A local a posteriori error estimator based on equilibrated fluxes. *SIAM J. Numer. Anal.*, 42(4):1394–1414, 2004. ISSN 0036-1429.

J. Papež and Z. Strakoš. Galerkin orthogonality and the multiplicative factors in the residual-based a posteriori error estimator for total error. Preprint MORE/2016/14, Submitted for publication, 2016.

J. Papež, J. Liesen, and Z. Strakoš. Distribution of the discretization and algebraic error in numerical solution of partial differential equations. *Linear Algebra Appl.*, 449:89–114, 2014. ISSN 0024-3795.

5. Estimating and localizing the algebraic and total numerical errors using flux reconstructions

Paper Papež et al. [2016] included in Section 5.1 presents, using the Poisson model problem, a methodology for computing upper and lower bounds for the energy norm of the algebraic and total errors. The derived bounds do not contain any unspecified constants and allow for estimating the local distribution of the errors over the computational domain. Paper also investigates bounds on the energy norm of the discretization error and their application for constructing stopping criteria balancing the discretization and algebraic errors. An additional comment answering an open question from [Papež et al., 2016, Section 5] and a related numerical experiment is then present in Section 5.2.

The paper Papež et al. [2016] represents the joint work with Martin Vohralík and Zdeněk Strakoš. In Section 5.2, we acknowledge the collaboration with Ivana Pultarová.

5.1 Paper submitted to Numerische Mathematik

The section includes the paper Papež et al. [2016] submitted to Numerische Mathematik on May 3, 2016.

Estimating and localizing the algebraic and total numerical errors using flux reconstructions*

J. Papež[†‡] Z. Strakoš[†] M. Vohralík[§]

May 4, 2016

Abstract

This paper presents a methodology for computing upper and lower bounds for both the algebraic and total errors in the context of the conforming finite element discretization and an arbitrary iterative algebraic solver. The derived bounds are based on the flux reconstruction techniques, do not contain any unspecified constants, and allow estimating the local distribution of both errors over the computational domain. We also discuss bounds on the discretization error, their application for constructing mathematically justified stopping criteria for iterative algebraic solvers, global and local efficiency of the total error upper bound, and the relationship to the previously published estimates on the algebraic error. Theoretical results are illustrated on numerical experiments for higher-order finite element approximations and the preconditioned conjugate gradient method.

Keywords: Numerical solution of partial differential equations, finite element method, a posteriori error estimation, algebraic error, discretization error, stopping criteria, spatial distribution of the error

MSC: 65N15, 65N30, 76M10, 65N22, 65F10.

1 Introduction

Most a posteriori error analyses of numerical approximations of partial differential equations still assume that the discretized algebraic problem is solved *exactly*. This is an unrealistic assumption that cannot be satisfied in large scale numerical computations. There is, fortunately, a growing body of work avoiding it, based on different approaches, see, e.g., [18, 5, 8, 44, 54, 42, 49, 11, 29, 32, 7,

*This work was supported by the ERC-CZ project LL1202. It has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647134 GATIPOR).

[†]Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Prague, Czech Republic.

[‡]Institute of Computer Science, Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic.

[§]INRIA of Paris, 2 rue Simone Iff, 75589 Paris, France.

48, 2, 22], the references given in the survey [3, Section 4], and in the monograph [35, Chapter 12]. Despite this development, a rigorous, mathematically justified, cheap, and accurate estimation of the discretization and algebraic errors that would allow for their comparison in *practical computations* is not, in our opinion, a fully solved problem. On the algebraic side, such comparison should include *localization* of the algebraic error. Since the algebraic computation aims at approximating the inverse of the discrete operator with respect to the given right-hand side, the algebraic error is of global nature and its distribution over the computational domain can be very different from the distribution of the discretization error; see, e.g., [40] and the references therein. To point out challenges that *any* approach that aims at mathematically rigorous incorporation of the algebraic error into a posteriori error analysis must consider, we now discuss several ways of how the algebraic error in numerical PDEs is estimated.

The conjugate gradient (CG) method minimizes the energy norm of the algebraic error over the Krylov subspaces associated with a symmetric positive definite matrix $\mathbf{A}$ and the initial residual; see, e.g., [30], [33, Section 2.2]. The estimates for the error of the CG approximations are widely studied; see, e.g., [26, 12, 50, 38], and the references given there. The estimates can be associated with the relationship of CG to the Gauss quadrature; see, e.g., [33, Section 3.5]. We will briefly discuss the upper bound based on the Gauss–Radau quadrature; see [16, 26, 28, 39] and called in [2, p. A1548] "[t]he only guaranteed upper bound for the $\mathbf{A}$-norm of the CG error". Considering a preassigned node λ, $0 < \lambda < \lambda_{\min}(\mathbf{A})$, where $\lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue of the matrix $\mathbf{A}$, the Gauss–Radau quadrature gives indeed, assuming *exact arithmetic*, an upper bound on the energy norm of the algebraic error. In [2, Section 4.2] the Poincaré inequality adaptive approach for bounding $\lambda_{\min}(\mathbf{A})$ from below and setting the value of λ is proposed.

Numerically, however, the situation is very subtle. In short, if $0 < \lambda \ll \lambda_{\min}(\mathbf{A})$, then the Gauss–Radau quadrature bound may largely overestimate the actual error. On the other hand, for λ very close to $\lambda_{\min}(\mathbf{A})$, which can make the upper bound tight, it might be impossible to compute the upper bound to a sufficient accuracy because of numerical instabilities. The *derivation* of the estimate includes (implicitly or explicitly) inversion of the matrix $\lambda\mathbf{I} - \mathbf{T}_i$, where $\mathbf{I}$ stands for the identity matrix and $\mathbf{T}_i$ is the Jacobi matrix associated with the ith CG iteration. For λ very close to $\lambda_{\min}(\mathbf{A}) \leq \lambda_{\min}(\mathbf{T}_i)$, and, at the same time, $\lambda_{\min}(\mathbf{T}_i)$ very close to $\lambda_{\min}(\mathbf{A})$, the matrix $\lambda\mathbf{I} - \mathbf{T}_i$ may become close to numerically singular. It should be emphasized that the numerical difficulty may not be visible from the final formulas giving the bound; see, e.g., [39]. The numerical stability analysis provided in [28] explained that although the estimates based on the relationship of CG with the Gauss–Radau quadrature can be very useful, they cannot be considered generally applicable guaranteed and computable upper bounds for the energy norm of the algebraic error. The meaning of the terms *guaranteed* and *computable* is within numerical linear algebra restricted only to the cases where the results are justified for all possible input data by a rigorous numerical stability analysis.

Multigrid or, more general, multilevel computations can serve as a second example. Here a standard assumption for a posteriori bounds on the algebraic error, which might require further substantial analysis, is that the algebraic problem on the *coarsest grid* is solved *exactly*; see, e.g., [5, 49]. Moreover, the literature known to the authors does not provide computable upper bounds on

the algebraic and the total errors. This topic has recently been addressed in [41]. Alternatively, in the multilevel context the a priori arguments are often used; see the discussion in Section 3.3.

A remarkable early concept relating the algebraic and discretization errors is represented by the Cascadic Conjugate Gradient method; see [18, 47]. In [18], the algebraic error is estimated assuming the superlinear convergence behavior of the CG method in the subsequent iterations, and using several heuristics and empirically chosen parameters. The analysis of [47] relies on the upper bound for the CG method based on Chebyshev polynomials that is typically not descriptive, and its refined version based on composite polynomials may not hold in finite precision computations; see [25]. The CG iterations can exhibit locally the so-called staircase behavior (see [33, Chapter 5]) that makes the analysis difficult.

The general a posteriori error estimation framework of [46] provides a guaranteed upper bound on the total error independent of the algebraic solver. However, the estimates do not generally allow to distinguish and compare the parts of the error corresponding to different sources and seem not suitable for constructing stopping criteria for iterative solvers.

The widely used residual-based error estimators (see, e.g., [49, 6, 2] and the references in [53]) provide upper bounds on the total error (and possibly on its components) with unspecified generic constants that can be of large value. The proposed practical stopping criteria and algorithms then require an empirical choice of these constants. A review of these and other approaches can be found in the survey [3]; see also the discussion in the Introduction of [32].

The presented paper elaborates further on the ideas used in [32] for finite volume discretizations, and in a more general framework in [22]; see also their application to discontinuous Galerkin finite element discretizations in [20]. Here we consider the conforming finite element setting and derive an upper bound on the total error that will be proved locally efficient and *polynomial-degree-robust* in the spirit of [9, 23]. All results account for the presence of the algebraic error of an arbitrary iterative solver. The paper also presents a guaranteed upper bound on the *algebraic error* and thoroughly discusses its relationship to formulas derived purely algebraically. Fast and reliable numerical computations using iterative algebraic solvers rely on meaningful stopping criteria. The stopping criteria from [32, 22] are modified here in order to avoid a possible early stopping that could invalidate the computed results.

The paper is organized as follows. The diffusion model problem considered in the paper and the notation are described in Section 2. In Section 3 we discuss known results on estimating the algebraic error using algebraic worst-case bounds, a priori arguments, and techniques using additional iteration steps of the algebraic solver. Section 4 gives an upper bound on the *total error* based on a quasi-equilibrated flux reconstruction. In Section 5 we derive an upper bound on the *algebraic error* and discuss its relationship to the bounds presented in Section 3. Section 6 is devoted to estimates of the *discretization error* and to discussion of the stopping criteria. We finally illustrate the obtained results numerically in Section 7 and give a concluding discussion in Section 8. We collect in Appendix A the proofs of the global and local efficiency of the presented total error bound.

2 Setting and notation

Let $\Omega \subset \mathbb{R}^d, d = 2, 3$, be a polygonal (polyhedral) domain (open, bounded, and connected set). We consider the Poisson model problem: find $u : \Omega \to \mathbb{R}$ such that

$$-\nabla\cdot(\nabla u) = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{2.1}$$

that can be equivalently written as the system of two first order equations for the scalar-valued *potential* u and the vector-valued function called *flux* $\boldsymbol{\sigma} \equiv -\nabla u$,

$$\begin{bmatrix} \nabla & I \\ 0 & \nabla\cdot \end{bmatrix} \begin{bmatrix} u \\ \boldsymbol{\sigma} \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix} \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega.$$

Assuming $f \in L^2(\Omega)$, the weak form of the model problem (2.1) is as follows: find $u \in V \equiv H^1_0(\Omega)$ such that

$$(\nabla u, \nabla v) = (f, v) \qquad \forall v \in V, \tag{2.2}$$

where $H^1_0(\Omega)$ denotes the standard Hilbert space of $L^2(\Omega)$ functions whose weak derivatives are in $L^2(\Omega)$ and with trace vanishing on $\partial\Omega$. For $v, w \in L^2(\Omega)$, (v, w) stands for $\int_\Omega v(\mathbf{x})w(\mathbf{x})\,\mathrm{d}\mathbf{x}$ (and similarly in the vector-valued case). Hereafter $\|\cdot\|$ denotes the L^2 norm, $\|w\| \equiv (w, w)^{1/2}$, $w \in L^2(\Omega)$. Owing to (2.2), the flux $\boldsymbol{\sigma}$ is in the space $\mathbf{H}(\mathrm{div}, \Omega)$ of the functions in $[L^2(\Omega)]^d$ with the weak divergence in $L^2(\Omega)$; see, e.g., [15, Section 6.13].

Let $\mathcal{T}_h$ be a simplicial mesh of Ω. We suppose that the mesh is conforming in the sense that, for two distinct elements of $\mathcal{T}_h$, their intersection is either an empty set or a common ℓ-dimensional face, $0 \le \ell \le d-1$. We denote a generic element of $\mathcal{T}_h$ by K and its diameter by h_K. We denote by $\mathbb{P}_p(K)$, $p \ge 0$, the space of pth order polynomial functions on an element K and by $\mathbb{P}_p(\mathcal{T}_h)$ the broken polynomial space spanned by $v_h|_K \in \mathbb{P}_p(K)$ for all $K \in \mathcal{T}_h$.

Let

$$V_h \equiv \left\{ v_h \in \mathbb{P}_p(\mathcal{T}_h) \cap C(\overline{\Omega}) \mid v_h = 0 \text{ on } \partial\Omega \right\} \subset H^1_0(\Omega)$$

be the usual finite element space of continuous, piecewise pth order polynomial functions, $p \ge 1$. The discrete formulation corresponding to the problem (2.2) reads: find $u_h \in V_h$ such that

$$(\nabla u_h, \nabla v_h) = (f, v_h) \qquad \forall v_h \in V_h. \tag{2.3}$$

The (exact) solution u_h of (2.3) satisfies the Galerkin orthogonality

$$(\nabla(u_h - u), \nabla v_h) = 0 \qquad \forall v_h \in V_h. \tag{2.4}$$

Let $\psi_j \in V_h$, $j = 1, \dots, N$, denote a basis of V_h, $\Psi = \{\psi_1, \dots, \psi_N\}$. Employing these functions in (2.3) gives rise to the system of linear algebraic equations

$$\mathsf{A}\mathsf{U} = \mathsf{F}, \tag{2.5}$$

where $u_h = \sum_{j=1}^N \mathsf{U}_j\psi_j = \Psi\mathsf{U}$, $\mathsf{U} = [\mathsf{U}_j]$ is the vector of unknowns, the system matrix $\mathsf{A} = [\mathsf{A}_{j\ell}]$ is symmetric and positive definite, $\mathsf{A}_{j\ell} = (\nabla\psi_\ell, \nabla\psi_j)$, $j, \ell = 1, \dots, N$, and the right-hand side vector $\mathsf{F} = [\mathsf{F}_j]$ is given by $\mathsf{F}_j = (f, \psi_j)$, $j = 1, \dots, N$. Within this model problem setting, we consider an *iterative* algebraic solver approximating the exact solution U of (2.5). At the i-th step,

$i = 0, 1, 2, \ldots$, we obtain the approximation $\mathsf{U}^i = [\mathsf{U}^i_j]$ and the algebraic residual vector $\mathsf{R}^i = [\mathsf{R}^i_j]$ with

$$\mathsf{R}^i \equiv \mathsf{F} - \mathbf{A}\mathsf{U}^i. \tag{2.6}$$

By u^i_h we denote the approximation to the solution u of (2.2) determined by the coefficient vector U^i, $u^i_h \equiv \sum_{j=1}^N \mathsf{U}^i_j \psi_j = \Psi \mathsf{U}^i$. We also rewrite (2.6) in a functional setting. For this purpose, let a function $r^i_h \in L^2(\Omega)$ be a representation of the algebraic residual vector R^i satisfying

$$(r^i_h, \psi_j) = \mathsf{R}^i_j, \qquad j = 1, \ldots, N. \tag{2.7}$$

Two examples are given in Section 5.1 below. Then (2.6) can be rewritten as

$$(r^i_h, \psi_j) = (f, \psi_j) - (\nabla u^i_h, \nabla \psi_j) \qquad \forall j = 1, \ldots, N \tag{2.8}$$

and, together with (2.3), it also implies

$$(r^i_h, v_h) = (f, v_h) - (\nabla u^i_h, \nabla v_h) = (\nabla(u_h - u^i_h), \nabla v_h) \qquad \forall v_h \in V_h. \tag{2.9}$$

This representation, introduced in this paper, will play the key role in the construction of the estimators below.

The total error between the exact solution u and the approximate solution u^i_h is measured in the energy norm $\|\nabla(u - u^i_h)\|$. Analogously, the algebraic energy norm of the error $u_h - u^i_h$ is

$$\begin{aligned} \|\nabla(u_h - u^i_h)\| = \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} &= \big((\mathsf{U} - \mathsf{U}^i), \mathbf{A}(\mathsf{U} - \mathsf{U}^i)\big)^{1/2} \\ &= (\mathbf{A}^{-1}\mathsf{R}^i, \mathsf{R}^i)^{1/2} = \|\mathsf{R}^i\|_{\mathbf{A}^{-1}}, \end{aligned}$$

where (V, U) denotes the standard inner product of the vectors U and V, $\|\mathsf{V}\| \equiv (\mathsf{V}, \mathsf{V})^{1/2}$ stands for the Euclidean norm of V, and $\|\mathbf{A}\|$ is the induced spectral norm of the matrix $\mathbf{A}$.

3 Algebraic bounds

This section presents some well-known algebraic bounds, with a few comments towards the conjugate gradient method and multilevel methods.

3.1 The L^2 (Euclidean) norm residual bound

The simplest algebraic error upper bound consists in

$$\|\nabla(u_h - u^i_h)\| = \|\mathsf{R}^i\|_{\mathbf{A}^{-1}} \le \|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^i\|. \tag{3.1}$$

For a symmetric positive definite matrix, the norm $\|\mathbf{A}^{-1}\|$ is given by the reciprocal of the smallest eigenvalue of the matrix $\mathbf{A}$. It is clear that for $\mathbf{A}$ ill-conditioned, the bound (3.1) can significantly overestimate the algebraic error. Remark that equality is attained for a vector R^i collinear with the eigenvector corresponding to the smallest eigenvalue of $\mathbf{A}$.

Even this simplest worst-case bound may not be easy to compute. The smallest eigenvalue of $\mathbf{A}$ is typically not available, and, if it is close to zero, then the cost of its reliable and accurate approximation may not be negligible; see, e.g., [36, 37]. We derive easily computable L^2 norm residual bounds in Section 5.2 below, based on the residual representation r^i_h in (2.7); see the estimates (5.3), (5.4), and (5.8).

3.2 Bounds using additional algebraic iterations

The following simple idea was to our knowledge first presented for algebraic error estimates in [28, pp. 262–263] for the CG method; see also [50, 38]. For estimating the total error it was then used in [32] and in [22], where an arbitrary algebraic solver was considered.

The triangle inequality gives, at the cost of $\nu > 0$ additional iterations,

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} \leq \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} + \|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}} = \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}}. \quad (3.2)$$

Assuming that for a given parameter $\gamma > 0$, the choice of ν ensures

$$\|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu}\| \leq \gamma \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}, \quad (3.3)$$

we have, using (3.1), an easily computable *upper bound*

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} \leq (1+\gamma)\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}. \quad (3.4)$$

Moreover,

$$\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} + \|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} + \gamma\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}},$$

so that, assuming that $0 < \gamma < 1$, we get the *lower bound*

$$(1-\gamma)\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}. \quad (3.5)$$

Here (3.4) and (3.5) show that the accuracy of the estimate $\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}$ is controlled by the user-specified parameter γ.

We must, however, take into account the following principal issue. If

$$\|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}} = \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}} \ll \|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu}\|,$$

the value of ν satisfying (3.3) can be very large. In the worst case, the value of ν can be even comparable with the size of the problem. Such situation is highly improbable in practical problems where preconditioning is used in order to get a reasonable convergence behavior. Still, for a given parameter γ, the smallest ν_1, respectively ν_2, satisfying

$$\|\mathsf{R}^{i+\nu_1}\|_{\mathbf{A}^{-1}} \leq \gamma\|\mathsf{U}^{i+\nu_1} - \mathsf{U}^i\|_{\mathbf{A}} \quad \text{resp.} \quad \|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu_2}\| \leq \gamma\|\mathsf{U}^{i+\nu_2} - \mathsf{U}^i\|_{\mathbf{A}}, \quad (3.6)$$

where both sides of the inequalities depend on ν_1 respectively ν_2, can significantly differ with $\nu_1 \ll \nu_2$. Section 7.1 below presents a numerical illustration.

Estimating the algebraic error in the CG method in [28, pp. 262-263] considered performing ν additional iterations and using the relation

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}^2 = \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}^2 + \|\mathsf{U} - \mathsf{U}^{i+\nu}\|_{\mathbf{A}}^2 = \|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathbf{A}}^2 + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}}^2 \quad (3.7)$$

that is based on the *global* $\mathbf{A}$*-orthogonality of the CG direction vectors.* The detailed rounding error analysis (see [50, (4.9)], [51, (3.7)] with the reference to the original paper [30]) leads to the following mathematical (exact arithmetic) equivalent of (3.7)

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}^2 = (\mu_{\text{alg}}^{\text{CG},i,\nu})^2 + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}}^2. \quad (3.8)$$

This relation can be derived assuming only *local orthogonality* that is well-preserved also in finite precision CG computations as a consequence of enforcing numerically the orthogonality among the consecutive direction vectors and residuals. Therefore (3.8) holds, apart from a small inaccuracy proportional to machine precision, also for the *computed quantities*. The same, however, has not been proved for (3.7).

In [50, 51], it was shown how to compute $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$ at a negligible cost directly from the coefficients in the CG recurrences; see also [27], [38, Section 5.3]. The resulting lower bound

$$\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu} \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} \tag{3.9}$$

holds until the ratio $\|\mathsf{U}-\mathsf{U}^i\|_{\mathbf{A}}/\|\mathsf{U}-\mathsf{U}^0\|_{\mathbf{A}}$ becomes close to the machine precision (for details see [50, Section 10]), and it is tight providing that the actual energy norm of the error decreases reasonably fast. Analogously to (3.3), assuming (nontrivially) that for a given parameter $\gamma > 0$, the number $\nu > 0$ of additional iteration steps is such that

$$\|\mathbf{A}^{-1}\| \cdot \|\mathsf{R}^{i+\nu}\|^2 \leq \gamma^2 (\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu})^2,$$

then $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$ gives (neglecting the terms proportional to machine precision)

$$(\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu})^2 \leq \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}}^2 \leq (1+\gamma^2)\,(\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu})^2. \tag{3.10}$$

In conclusion, the general bounds in (3.4) and (3.5) do not require any additional assumptions. Their value can be determined directly from the computed quantities $\mathsf{U}^i, \mathsf{U}^{i+\nu}$. The bounds for the CG method in (3.10) can be evaluated at almost no cost, but their validity for numericaly computed approximations $\mathsf{U}^i, \mathsf{U}^{i+\nu}$ had to be proved using a careful numerical stability analysis. As a reward, which is based on the particular properties of the CG method, we get an improved accuracy of the bounds, with the factor characterizing the gap between the lower and the upper bound reduced from $(1+\gamma)/(1-\gamma)$ in (3.4)–(3.5) to $\sqrt{1+\gamma^2}$ in (3.10).

3.3 A priori arguments in multilevel methods

Convergence of multilevel methods is typically proved using the *a priori* contraction argument

$$\|\mathsf{U} - \mathsf{U}^{i+1}\|_{\mathbf{A}} \leq \gamma \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}},$$

where $0 < \gamma < 1$. Then the triangle inequality immediately gives the algebraic error bound

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} \leq \frac{1}{1-\gamma} \|\mathsf{U}^{i+1} - \mathsf{U}^i\|_{\mathbf{A}}.$$

Though such bounds with a priori determined constant γ can be useful (see, e.g., [7, (2.17)–(2.18)] and the references therein), we believe, as discussed in the introduction, that *a posteriori* bounds such as that of [5] or its unknown-constant-free improvement in [41] are preferable.

4 Estimating the total error

We give in this section computable upper and lower bounds on the total error. The upper bound based on flux reconstruction following [17, 10, 32, 22, 23] is derived in a form where the component associated with the algebraic error actually turns out to give its upper bound; see Section 5. The lower bound on the total error is given in Section 4.5 using conforming residual reconstruction. We will frequently use the following representation of the energy norm of the total error

$$\|\nabla(u - u_h^i)\| = \sup_{v\in V,\, \|\nabla v\|=1} (\nabla(u - u_h^i), \nabla v). \tag{4.1}$$

4.1 Concept of the flux reconstructions

The motivation for our approach is to mimic the continuous world, where (using (4.1), (2.2), the Green theorem, and the Cauchy–Schwarz inequality),

$$\begin{aligned}\|\nabla(u - u_h^i)\| &= \inf_{\mathbf{d}\in\mathbf{H}(\mathrm{div},\Omega),\, \nabla\cdot\mathbf{d}=f} \sup_{v\in V,\, \|\nabla v\|=1} \{(f - \nabla\cdot\mathbf{d}, v) - (\nabla u_h^i + \mathbf{d}, \nabla v)\} \\ &= \inf_{\mathbf{d}\in\mathbf{H}(\mathrm{div},\Omega),\, \nabla\cdot\mathbf{d}=f} \|\nabla u_h^i + \mathbf{d}\|;\end{aligned}$$

the equality occurs for $\mathbf{d} = \boldsymbol{\sigma} = -\nabla u$. We also wish to use an upper bound on the algebraic error based on the representation r_h^i. This allows to relate the algebraic and discretization error components.

Practically, a *reconstructed flux* is a piecewise polynomial function in the Raviart–Thomas–Nédélec subspace $\mathbf{V}_h$ of the infinite-dimensional space $\mathbf{H}(\mathrm{div},\Omega)$. It is constructed in an inexpensive *local* way, around each node of the mesh $\mathcal{T}_h$, and it satisfies, on each iteration step $i \geq 1$,

$$\nabla\cdot\mathbf{d}_h^i = f_h - r_h^i. \tag{4.2}$$

Here f_h is a piecewise polynomial approximation of the source term f satisfying

$$(f - f_h, 1)_K = 0 \qquad \forall K \in \mathcal{T}_h. \tag{4.3}$$

The precise definition of the space $\mathbf{V}_h$ and the detailed construction of $\mathbf{d}_h^i$ are given below in Section 4.4.

4.2 Upper bound using the L^2 norm of the algebraic residual representation

Similarly to Section 3.1, to illustrate the ideas, we first present a simple upper bound on the total error following [32, Section 7.1]. It typically yields a large overestimation. It follows from (4.1), the weak formulation (2.2), the construction (4.2), and the Green theorem that

$$\|\nabla(u - u_h^i)\| = \sup_{v\in V,\, \|\nabla v\|=1} \left\{(f - f_h, v) + (r_h^i, v) - (\nabla u_h^i + \mathbf{d}_h^i, \nabla v)\right\}. \tag{4.4}$$

Using (4.3) and the Poincaré inequality on the mesh elements,

$$(f - f_h, v) \leq \eta_{\mathrm{osc}}\|\nabla v\|, \quad \eta_{\mathrm{osc}} \equiv \left(\sum_{K\in\mathcal{T}_h} \eta_{\mathrm{osc},K}^2\right)^{1/2}, \quad \eta_{\mathrm{osc},K} \equiv \frac{h_K}{\pi}\|f - f_h\|_K; \tag{4.5}$$

see, e.g., [22, p. A1767]. The Friedrichs inequality states that there exists a generic constant $0 < C_{\mathrm{F}} \leq 1$ such that

$$\|v\| \leq C_{\mathrm{F}} h_\Omega \|\nabla v\| \qquad \forall v \in V, \tag{4.6}$$

where h_Ω denotes the diameter of the domain Ω. The value of C_{F} can be bounded[1] using, e.g., [45, Chapter 18]. Thus, from the Cauchy–Schwarz inequality and from (4.6),

$$(r_h^i, v) \leq \|r_h^i\| \|v\| \leq \|r_h^i\| C_{\mathrm{F}} h_\Omega \|\nabla v\|, \tag{4.7}$$

$$(\nabla u_h^i + \mathbf{d}_h^i, \nabla v) \leq \|\nabla u_h^i + \mathbf{d}_h^i\| \|\nabla v\|. \tag{4.8}$$

Then (4.4) immediately gives the upper bound on the total error

$$\|\nabla(u - u_h^i)\| \leq \eta_{\mathrm{osc}} + C_{\mathrm{F}} h_\Omega \|r_h^i\| + \|\nabla u_h^i + \mathbf{d}_h^i\|. \tag{4.9}$$

The part η_{osc} measures the oscillations in the right-hand side f and it is often negligible in comparison to the discretization error. The part $C_{\mathrm{F}} h_\Omega \|r_h^i\|$ in (4.9) bounds the algebraic error; see (5.3) below. Finally, we will associate the last term $\|\nabla u_h^i + \mathbf{d}_h^i\|$ with estimating the discretization error as in [22].

4.3 Upper bound using additional algebraic iterations

Following [22], the idea of using $\nu > 0$ additional iterations described in Section 3.2 can be analogously applied here to substantially improve the bound (4.9).

Given the computed approximation u_h^i, we construct the algebraic residual representation r_h^i satisfying (2.7) and a reconstructed flux $\mathbf{d}_h^i \in \mathbf{V}_h$ satisfying (4.2). After $\nu > 0$ additional iterations of the algebraic solver, giving the approximation $u_h^{i+\nu}$, we construct $r_h^{i+\nu}$ satisfying (2.7) with i replaced by $i+\nu$ and a reconstructed flux $\mathbf{d}_h^{i+\nu} \in \mathbf{V}_h$ satisfying $\nabla\cdot \mathbf{d}_h^{i+\nu} = f_h - r_h^{i+\nu}$. Thus,

$$r_h^i = -\nabla\cdot \mathbf{d}_h^i + f_h = -\nabla\cdot \mathbf{d}_h^i + \nabla\cdot \mathbf{d}_h^{i+\nu} + r_h^{i+\nu} \tag{4.10}$$

and we have as above

$$\|\nabla(u - u_h^i)\| = \sup_{v\in V,\, \|\nabla v\|=1} \{(f - f_h, v) + (\mathbf{d}_h^i - \mathbf{d}_h^{i+\nu}, \nabla v) + (r_h^{i+\nu}, v) - (\nabla u_h^i + \mathbf{d}_h^i, \nabla v)\},$$

which immediately leads to, cf. [22, Theorem 3.6]:

Theorem 1 (Upper bound on the total error)**.** *Let u be the weak solution given by (2.2) and let $u_h^i \in V_h$ be its approximation given at the ith algebraic solver iteration with the corresponding algebraic residual representation r_h^i given by (2.8). Let a reconstructed flux $\mathbf{d}_h^i \in \mathbf{V}_h$ satisfy (4.2). Consider $\nu > 0$ additional algebraic iterations, resulting in $r_h^{i+\nu}$ and $\mathbf{d}_h^{i+\nu}$. Then*

$$\|\nabla(u - u_h^i)\| \leq \eta_{\mathrm{total}}^{i,\nu} \equiv \eta_{\mathrm{osc}} + \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\| + \|\nabla u_h^i + \mathbf{d}_h^i\|,$$

where the data oscillation term η_{osc} is given by (4.5) and $C_{\mathrm{F}} h_\Omega$ is the constant from the Friedrichs inequality (4.6).

[1] For example, for a square domain $\Omega \subset \mathbb{R}^2$ we can take $C_{\mathrm{F}} = 1/(2\pi)$, corresponding to the smallest eigenvalue of the Laplace operator; see, e.g., [45, relation (18.48) on p. 196]

Remark 1. The statement of Theorem 1 deserves several comments that point out to the results presented later in the text. We typically choose ν in concordance with the theoretical justification (global efficiency) of Theorem 7 below; see also (7.3c) in the numerical experiments. Local efficiency of $\eta_{\mathrm{total}}^{i,\nu}$ is proved in Appendix A for i and ν based on local stopping criteria. Note that the sum $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|$ gives an upper bound on the algebraic error (see Theorem 3 below), whereas the term $\|\nabla u_h^i + \mathbf{d}_h^i\|$ can be associated, at least in the case of a small algebraic error, with the discretization error; see the further results in Section 6 and Section 7.4.

4.4 Details of the flux reconstruction

We now present the construction of the flux $\mathbf{d}_h^i$. It follows [22, Section 6.2.4] (see also [17, 10]) with the difference in the construction of the algebraic residual representation r_h^i satisfying (2.7). This difference is, however, crucial, as it allows to bound the algebraic error in Theorem 3 below. The construction is rather technical and its detailed study is not substantial for understanding the rest of the paper.

For $K \in \mathcal{T}_h$, let $\mathbf{RTN}_{p'}(K) \equiv [\mathbb{P}_{p'}(K)]^d + \mathbf{x}\mathbb{P}_{p'}(K)$ be the Raviart–Thomas–Nédélec finite element space of order $p' \geq 0$. We set

$$\mathbf{RTN}_{p'}^{-1}(\mathcal{T}_h) \equiv \left\{ \mathbf{v}_h \in [L^2(\Omega)]^d, \mathbf{v}_h|_K \in \mathbf{RTN}_{p'}(K) \quad \forall K \in \mathcal{T}_h \right\}$$

and $\mathbf{RTN}_{p'}(\mathcal{T}_h) \equiv \mathbf{RTN}_{p'}^{-1}(\mathcal{T}_h) \cap \mathbf{H}(\mathrm{div}, \Omega)$. We use a similar notation for these spaces on various patches. Let $\mathcal{V}_h$ denote the set of mesh vertices with subsets $\mathcal{V}_h^{\mathrm{int}}$ for interior vertices and $\mathcal{V}_h^{\mathrm{ext}}$ for boundary ones. Let $\psi_{\mathtt{a}} \in \mathbb{P}_1(\mathcal{T}_h) \cap H^1(\Omega)$ stand for the hat function associated with a vertex $\mathtt{a} \in \mathcal{V}_h$ (i.e., $\psi_{\mathtt{a}}(\mathtt{a}) = 1$, $\psi_{\mathtt{a}}(\mathtt{a}') = 0$ for $\mathtt{a} \neq \mathtt{a}' \in \mathcal{V}_h$). We denote by $\mathcal{T}_{\mathtt{a}}$ the union of elements sharing the vertex $\mathtt{a} \in \mathcal{V}_h$ and by $\omega_{\mathtt{a}}$ the corresponding open subdomain. Let $\mathbf{RTN}_{p'}^{\mathrm{N},0}(\mathcal{T}_{\mathtt{a}})$ be the subspace of $\mathbf{RTN}_{p'}(\mathcal{T}_{\mathtt{a}})$ with zero normal flux through the boundary $\partial\omega_{\mathtt{a}}$ for $\mathtt{a} \in \mathcal{V}_h^{\mathrm{int}}$ and through $\partial\omega_{\mathtt{a}} \backslash \partial\Omega$ for $\mathtt{a} \in \mathcal{V}_h^{\mathrm{ext}}$ (corresponding to a homogeneous Neumann condition). Let $\mathbb{P}_{p'}^*(\mathcal{T}_{\mathtt{a}})$ be spanned by piecewise p'th order polynomials on $\mathcal{T}_{\mathtt{a}}$, with zero mean on $\mathcal{T}_{\mathtt{a}}$ when $\mathtt{a} \in \mathcal{V}_h^{\mathrm{int}}$.

For all vertices $\mathtt{a} \in \mathcal{V}_h$, we first solve the following mixed finite element problems on the patches $\mathcal{T}_{\mathtt{a}}$: find $\mathbf{d}_{h,\mathtt{a}}^i \in \mathbf{RTN}_{p'}^{\mathrm{N},0}(\mathcal{T}_{\mathtt{a}})$ and $q_{h,\mathtt{a}} \in \mathbb{P}_{p'}^*(\mathcal{T}_{\mathtt{a}})$, $p' = p$ or $p' = p+1$, such that

$$(\mathbf{d}_{h,\mathtt{a}}^i, \mathbf{v}_h)_{\omega_{\mathtt{a}}} - (q_{h,\mathtt{a}}, \nabla\cdot \mathbf{v}_h)_{\omega_{\mathtt{a}}} = -(\psi_{\mathtt{a}} \nabla u_h^i, \mathbf{v}_h)_{\omega_{\mathtt{a}}}, \tag{4.11a}$$

$$(\nabla\cdot \mathbf{d}_{h,\mathtt{a}}^i, \phi_h)_{\omega_{\mathtt{a}}} = (f_h \psi_{\mathtt{a}} - \nabla u_h^i \cdot \nabla \psi_{\mathtt{a}}, \phi_h)_{\omega_{\mathtt{a}}} - (r_h^i \psi_{\mathtt{a}}, \phi_h)_{\omega_{\mathtt{a}}} \tag{4.11b}$$

for all $(\mathbf{v}_h, \phi_h) \in \mathbf{RTN}_{p'}^{\mathrm{N},0}(\mathcal{T}_{\mathtt{a}}) \times \mathbb{P}_{p'}^*(\mathcal{T}_{\mathtt{a}})$. Then we set

$$\mathbf{d}_h^i \equiv \sum_{\mathtt{a} \in \mathcal{V}_h} \mathbf{d}_{h,\mathtt{a}}^i. \tag{4.11c}$$

We typically choose f_h to be the $L^2(\Omega)$-orthogonal projection of f onto the space of the piecewise polynomials of degree p', and $r_h^i \in \mathbb{P}_p(\mathcal{T}_h)$; see Section 5.1. Since $\psi_{\mathtt{a}} \in V_h$, (2.8) gives the Neumann compatibility condition of the problem (4.11a)–(4.11b),

$$(\nabla u_h^i, \nabla \psi_{\mathtt{a}})_{\omega_{\mathtt{a}}} = (f, \psi_{\mathtt{a}})_{\omega_{\mathtt{a}}} - (r_h^i, \psi_{\mathtt{a}})_{\omega_{\mathtt{a}}}.$$

Consequently, we can in (4.11b) take all test functions $\phi_h \in \mathbb{P}_{p'}(\mathcal{T}_\mathtt{a})$, which allows to show that $\mathbf{d}_h^i$ given by (4.11) satisfies (4.2), i.e., that $\nabla\cdot\mathbf{d}_h^i = f_h - r_h^i$ holds. Indeed, let $K \in \mathcal{T}_h$ and let $v_h \in \mathbb{P}_{p'}(K)$ be fixed. Since $\sum_{\mathtt{a}\in\mathcal{V}_h} \psi_\mathtt{a}|_K = 1$ and $\sum_{\mathtt{a}\in\mathcal{V}_h} \nabla\psi_\mathtt{a}|_K = 0$ ($\psi_\mathtt{a}$ form a partition of unity on K), we infer

$$\begin{aligned}(\nabla\cdot\mathbf{d}_h^i, v_h)_K &= \sum_{\mathtt{a}\in\mathcal{V}_h} (\nabla\cdot\mathbf{d}_{h,\mathtt{a}}^i, v_h)_K \\ &= \sum_{\mathtt{a}\in\mathcal{V}_h} \left[(f_h\psi_\mathtt{a} - \nabla u_h^i\cdot\nabla\psi_\mathtt{a}, v_h)_K - (r_h^i\psi_\mathtt{a}, v_h)_K\right] \\ &= (f_h, v_h)_K - (r_h^i, v_h)_K,\end{aligned}$$

and (4.2) is proved as $f_h - r_h^i \in \mathbb{P}_{p'}(\mathcal{T}_h)$.

4.5 Lower bound

Following [4, Section 5.1], [46, Section 4.1.1], or [23, Section 3.3], one can use the conforming version of the local Neumann problems (4.11a)–(4.11b) to bound the total error $\|\nabla(u-u_h^i)\|$ from below. For each vertex $\mathtt{a} \in \mathcal{V}_h$, consider the infinite-dimensional space $H^1_*(\omega_\mathtt{a})$

$$H^1_*(\omega_\mathtt{a}) \equiv \begin{cases} v \in H^1(\omega_\mathtt{a});\, (v,1)_{\omega_\mathtt{a}} = 0 & \mathtt{a} \in \mathcal{V}_h^{\mathrm{int}}, \\ v \in H^1(\omega_\mathtt{a});\, v = 0 \text{ on } \partial\omega_\mathtt{a} \cap \partial\Omega & \mathtt{a} \in \mathcal{V}_h^{\mathrm{ext}}. \end{cases} \tag{4.12}$$

For the functions from the space $H^1_*(\omega_\mathtt{a})$ the following Poincaré–Friedrichs-type inequalities hold: there exists a positive constant $C_{\mathrm{PF},\omega_\mathtt{a}}$, depending on the shape of the elements of the patch $\mathcal{T}_\mathtt{a}$ but not on their diameters, and a positive constant $C_{\mathrm{cont,PF},\omega_\mathtt{a}} \equiv 1 + C_{\mathrm{PF},\omega_\mathtt{a}} h_{\omega_\mathtt{a}} \|\nabla\psi_\mathtt{a}\|_{\infty,\omega_\mathtt{a}}$ (see, e.g., [23, inequality (3.29)]) such that

$$\|v\|_{\omega_\mathtt{a}} \le C_{\mathrm{PF},\omega_\mathtt{a}} h_{\omega_\mathtt{a}} \|\nabla v\|_{\omega_\mathtt{a}} \qquad \forall v \in H^1_*(\omega_\mathtt{a}), \tag{4.13}$$

$$\|\nabla(\psi_\mathtt{a} v)\| \le C_{\mathrm{cont,PF},\omega_\mathtt{a}} \|\nabla v\|_{\omega_\mathtt{a}} \qquad \forall v \in H^1_*(\omega_\mathtt{a}). \tag{4.14}$$

For convex patches $\mathcal{T}_\mathtt{a}$ around the interior vertices $\mathtt{a}$ we have $C_{\mathrm{PF},\omega_\mathtt{a}} = 1/\pi$; see, e.g., [43]. For nonconvex patches we refer to [23, 52] and the references therein. For a shape-regular mesh $h_{\omega_\mathtt{a}}\|\nabla\psi_\mathtt{a}\|_{\infty,\omega_\mathtt{a}} = O(1)$ (see, e.g., [14, relation (3.1.43) on p. 124]), giving $C_{\mathrm{cont,PF},\omega_\mathtt{a}} = O(1)$; see the discussion in [23, Remark 3.24].

For each vertex $\mathtt{a} \in \mathcal{V}_h$, let $W_h^\mathtt{a}$ be a finite-dimensional subspace of $H^1_*(\omega_\mathtt{a})$. The simplest choice, which we use in numerical experiments in Section 7.4, is $W_h^\mathtt{a} \equiv \mathbb{P}_p(\mathcal{T}_\mathtt{a}) \cap H^1_*(\omega_\mathtt{a})$. We then have the following bound:

Theorem 2 (Lower bound on the total error)**.** *Let u be the weak solution given by (2.2) and let $u_h^i \in V_h$ be its approximation given at the ith algebraic solver iteration with the corresponding algebraic residual representation r_h^i given by (2.8). For each vertex $\mathtt{a} \in \mathcal{V}_h$, let $m_{h,\mathtt{a}} \in W_h^\mathtt{a}$ be the solution of*

$$(\nabla m_{h,\mathtt{a}}, \nabla v_h)_{\omega_\mathtt{a}} = (f, \psi_\mathtt{a} v_h)_{\omega_\mathtt{a}} - (\nabla u_h^i, \nabla(\psi_\mathtt{a} v_h))_{\omega_\mathtt{a}} \qquad \forall v_h \in W_h^\mathtt{a}.$$

Set $m_h \equiv \sum_{\mathtt{a}\in\mathcal{V}_h} \psi_\mathtt{a} m_{h,\mathtt{a}} \in V$. Then

$$\|\nabla(u-u_h^i)\| \ge \mu_{\mathrm{total}}^i \equiv \frac{\sum_{\mathtt{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathtt{a}}\|^2_{\omega_\mathtt{a}}}{\|\nabla m_h\|}.$$

Proof. Since $m_h \in V$ by construction, we have from (4.1)

$$\begin{aligned}
\|\nabla(u - u_h^i)\| &= \sup_{v\in V,\, \|\nabla v\|=1} (\nabla(u - u_h^i), \nabla v) \\
&\geq \frac{1}{\|\nabla m_h\|}(\nabla(u - u_h^i), \nabla m_h) \\
&= \frac{1}{\|\nabla m_h\|} \sum_{\mathsf{a}\in\mathcal{V}_h} (\nabla(u - u_h^i), \nabla(\psi_{\mathsf{a}} m_{h,\mathsf{a}}))_{\omega_{\mathsf{a}}} \\
&= \frac{1}{\|\nabla m_h\|} \sum_{\mathsf{a}\in\mathcal{V}_h} \{(f, \psi_{\mathsf{a}} m_{h,\mathsf{a}})_{\omega_{\mathsf{a}}} - (\nabla u_h^i, \nabla(\psi_{\mathsf{a}} m_{h,\mathsf{a}}))_{\omega_{\mathsf{a}}}\} \\
&= \frac{1}{\|\nabla m_h\|} \sum_{\mathsf{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathsf{a}}\|_{\omega_{\mathsf{a}}}^2,
\end{aligned}$$

where we have used the fact that $\psi_{\mathsf{a}} m_{h,\mathsf{a}} \in H_0^1(\omega_{\mathsf{a}})$ for all vertices $\mathsf{a} \in \mathcal{V}_h$ and the definition of $m_{h,\mathsf{a}}$. □

Remark 2. The bound μ_{total}^i can further be localized using (4.14) as

$$\mu_{\text{total}}^i \geq \frac{\left\{\sum_{\mathsf{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathsf{a}}\|_{\omega_{\mathsf{a}}}^2\right\}^{1/2}}{(d+1)^{1/2} C_{\text{cont,PF}}},$$

where $C_{\text{cont,PF}} \equiv \max_{\mathsf{a}\in\mathcal{V}_h} C_{\text{cont,PF},\omega_{\mathsf{a}}}$. Denoting by $\mathcal{V}_K$ the vertices of an element K and using the fact that each simplex has $(d+1)$ vertices, this can be seen from

$$\begin{aligned}
\|\nabla m_h\|^2 &= \sum_{K\in\mathcal{T}_h} \left\| \sum_{\mathsf{a}\in\mathcal{V}_K} (\nabla(\psi_{\mathsf{a}} m_{h,\mathsf{a}}))|_K \right\|_K^2 \leq (d+1) \sum_{K\in\mathcal{T}_h} \sum_{\mathsf{a}\in\mathcal{V}_K} \|\nabla(\psi_{\mathsf{a}} m_{h,\mathsf{a}})\|_K^2 \\
&= (d+1) \sum_{\mathsf{a}\in\mathcal{V}_h} \|\nabla(\psi_{\mathsf{a}} m_{h,\mathsf{a}})\|_{\omega_{\mathsf{a}}}^2 \leq (d+1) C_{\text{cont,PF}}^2 \sum_{\mathsf{a}\in\mathcal{V}_h} \|\nabla m_{h,\mathsf{a}}\|_{\omega_{\mathsf{a}}}^2.
\end{aligned}$$

5 Estimating the algebraic error

We will now derive upper bounds on the algebraic error with the help of the representation of the algebraic residual r_h^i satisfying (2.7) and of the flux reconstruction $\mathbf{d}_h^i$ of Section 4.4. We will make links to the bounds of Section 3 derived purely algebraically and to the total error bounds of the previous section. Section 5.4 recalls the lower bounds on the algebraic error of Section 3 and proposes a (function-based) construction of a lower bound analogously to Section 4.5.

5.1 Representation of the algebraic residual

We first propose two piecewise polynomial representations of the algebraic residual r_h^i satisfying (2.7).

The choice $r_h^i \in V_h = \mathbb{P}_p(\mathcal{T}_h) \cap H_0^1(\Omega)$ given by (2.7) requires solving the linear algebraic system with the *global mass matrix*

$$\mathsf{G}\mathsf{C}^i = \mathsf{R}^i, \qquad \mathsf{G}_{j\ell} \equiv (\psi_\ell, \psi_j), \quad j, \ell = 1, \dots, N. \tag{5.1}$$

Then $r_h^i = \Psi \mathsf{C}^i = \Psi \mathbf{G}^{-1} \mathsf{R}^i$.

Equation (5.1) represents a global problem of the same size as (2.5). In order to avoid performing a global solve, we introduce a piecewise *discontinuous* polynomial representation $r_h^i \in \mathbb{P}_p(\mathcal{T}_h)$ using mutually independent local problems. For the ease of notation, the construction below is described for the Lagrangian basis of V_h. Denote by n_j the number of mesh elements forming the support of the basis function ψ_j, $j = 1, \dots, N$. Then, for each element $K \in \mathcal{T}_h$, define $r_h^i|_K \in \mathbb{P}_p(K)$, $r_h^i|_{\partial\Omega} = 0$, such that

$$(r_h^i, \psi_j)_K = \mathsf{R}_j^i / n_j \qquad \text{for } \psi_j \text{ nonvanishing on } K. \tag{5.2}$$

Summing (5.2) over all elements $K \in \mathcal{T}_h$, we see that (2.7) indeed holds. Denoting by R_K^i the vector on the right-hand side of (5.2) and by $\mathbf{G}_K$ the *local mass matrix*

$$(\mathbf{G}_K)_{j\ell} \equiv (\psi_\ell, \psi_j)_K \qquad \text{for } \psi_\ell, \psi_j \text{ nonvanishing on } K,$$

we have

$$r_h^i|_K = \Psi|_K (\mathbf{G}_K^{-1} \mathsf{R}_K^i) \qquad \forall K \in \mathcal{T}_h.$$

Construction (5.2) requires solving the system of the size $\frac{1}{2}(p+1)(p+2)$ separately on each element $K \in \mathcal{T}_h$.

5.2 Bound using the L^2 norm of the residual representation

Similarly to (4.1), using (2.9) and (4.7), the energy norm of the algebraic error satisfies

$$\begin{aligned} \|\nabla(u_h - u_h^i)\| &= \sup_{v_h \in V_h, \|\nabla v_h\| = 1} (\nabla(u_h - u_h^i), \nabla v_h) = \sup_{v_h \in V_h,\, \|\nabla v_h\| = 1} (r_h^i, v_h) \\ &\le C_{\mathrm{F}} h_\Omega \|r_h^i\|. \end{aligned} \tag{5.3}$$

We first discuss the bound (5.3) for the representation r_h^i constructed globally using (5.1). The discussion shows the relationship of (5.3) to the algebraic worst-case bounds of Section 3.1 and the role of the Friedrichs inequality constant $C_{\mathrm{F}} h_\Omega$. In the case (5.1),

$$\|r_h^i\|^2 = (\Psi \mathbf{G}^{-1} \mathsf{R}^i, \Psi \mathbf{G}^{-1} \mathsf{R}^i) = (\mathbf{G}^{-1} \mathsf{R}^i)^T \mathbf{G} (\mathbf{G}^{-1} \mathsf{R}^i) = (\mathsf{R}^i)^T \mathbf{G}^{-1} \mathsf{R}^i = \|\mathsf{R}^i\|_{\mathbf{G}^{-1}}^2,$$

and therefore

$$\|\nabla(u_h - u_h^i)\| = \|\mathsf{U} - \mathsf{U}^i\|_{\mathbf{A}} = \|\mathsf{R}^i\|_{\mathbf{A}^{-1}} \le C_{\mathrm{F}} h_\Omega \|\mathsf{R}^i\|_{\mathbf{G}^{-1}}. \tag{5.4}$$

An analogous estimate for the finite volume method is given in [32, Section 7.1], where it was observed in numerical experiments that this estimate can significantly overestimate the algebraic error. We note that

$$\begin{aligned} \|\mathsf{R}^i\|_{\mathbf{A}^{-1}}^2 &= (\mathsf{R}^i, \mathbf{A}^{-1} \mathsf{R}^i) = (\mathbf{G}^{-1/2} \mathsf{R}^i, \mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2} \mathbf{G}^{-1/2} \mathsf{R}^i) \\ &\le \|\mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2}\| \cdot \|\mathbf{G}^{-1/2} \mathsf{R}^i\|^2 = \|\mathbf{G}^{1/2} \mathbf{A}^{-1} \mathbf{G}^{1/2}\| \cdot \|\mathsf{R}^i\|_{\mathbf{G}^{-1}}^2. \end{aligned} \tag{5.5}$$

Because (5.4) holds also for the special choice of R^i giving the equality in (5.5) (when $\mathsf{G}^{-1/2}\mathsf{R}^i$ is collinear with the eigenvector of $\mathsf{G}^{1/2}\mathsf{A}^{-1}\mathsf{G}^{1/2}$ corresponding to its largest eigenvalue), we have

$$\|\mathsf{G}^{1/2}\mathsf{A}^{-1}\mathsf{G}^{1/2}\| \leq (C_{\mathrm{F}} h_\Omega)^2. \tag{5.6}$$

This means that the reciprocal of the squared Friedrichs inequality constant $(C_{\mathrm{F}} h_\Omega)^{-2}$ (and through that the related smallest eigenvalue of the continuous operator; see, e.g., [45, Section 18]) gives a computable lower bound on the smallest eigenvalue of the (preconditioned) matrix $\mathsf{G}^{-1/2}\mathsf{A}\mathsf{G}^{-1/2}$ (cf. also [31], [2, Section 4.2]),

$$\frac{1}{(C_{\mathrm{F}} h_\Omega)^2} \leq \min_{\lambda \in \mathrm{sp}(\mathsf{G}^{-1/2}\mathsf{A}\mathsf{G}^{-1/2})} \lambda. \tag{5.7}$$

The local construction (5.2) leads to

$$\|\nabla(u_h - u_h^i)\| \leq C_{\mathrm{F}} h_\Omega \left(\sum_{K \in \mathcal{T}_h} \|r_h^i\|_K^2 \right)^{1/2} = C_{\mathrm{F}} h_\Omega \left(\sum_{K \in \mathcal{T}_h} \|\mathsf{R}_K^i\|_{\mathsf{G}_K^{-1}}^2 \right)^{1/2}. \tag{5.8}$$

The detailed relationship between the upper bounds (5.4) and (5.8) remains open.

5.3 Upper bound using additional algebraic iterations

Analogously to Sections 3.2 and 4.3, we can bound the algebraic error using ν additional iteration steps. From (2.9), (4.10), and the Green theorem, for $v_h \in V_h$,

$$(\nabla(u_h - u_h^i), \nabla v_h) = (r_h^i, v_h) = (\mathbf{d}_h^i - \mathbf{d}_h^{i+\nu}, \nabla v_h) + (r_h^{i+\nu}, v_h). \tag{5.9}$$

Thus the following upper bound on the algebraic error immediately follows from (5.3):

Theorem 3 (Upper bound on the algebraic error)**.** *Let the assumptions of Theorem 1 be satisfied. Then*

$$\|\nabla(u_h - u_h^i)\| \leq \eta_{\mathrm{alg}}^{i,\nu} \equiv \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|.$$

Remark 3. The upper bound of Theorem 3 on the algebraic error allows evaluation of the *local indicators* $\eta_{\mathrm{alg},K}^{i,\nu} \equiv \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K$ for the mesh elements $K \in \mathcal{T}_h$, with subsequently using these indicators for estimating the *local distribution* of the algebraic error $\|\nabla(u_h - u_h^i)\|_K$. This can indeed be very useful in localization of the significant components of the algebraic error over the discretization domain Ω, which represents an important problem; see [40] and the numerical illustrations in Section 7.2.

In order to show the relationship between (5.9) and (3.2), we note that, using (2.9),

$$(\mathbf{d}_h^i - \mathbf{d}_h^{i+\nu}, \nabla v_h) = (r_h^i - r_h^{i+\nu}, v_h) = (\nabla(u_h^{i+\nu} - u_h^i), \nabla v_h),$$

so that

$$\|\mathsf{U}^{i+\nu} - \mathsf{U}^i\|_{\mathsf{A}} = \sup_{v_h \in V_h,\, \|\nabla v_h\| = 1} (\nabla(u_h^{i+\nu} - u_h^i), \nabla v_h) \leq \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|.$$

Employing also (5.3) for $i+\nu$ in place of i, the upper bound of Theorem 3 appears weaker than the algebraic bound (3.2),

$$\|\mathsf{U}-\mathsf{U}^i\|_{\mathbf{A}} \le \|\mathsf{U}^{i+\nu}-\mathsf{U}^i\|_{\mathbf{A}} + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}} \le \|\mathbf{d}_h^{i+\nu}-\mathbf{d}_h^i\| + \|\mathsf{R}^{i+\nu}\|_{\mathbf{A}^{-1}}.$$

The fluxes $\mathbf{d}_h^i$ (and $\mathbf{d}_h^{i+\nu}$) are, however, essential for bounding the total error in Theorem 1 and, importantly, the algebraic estimator in Theorem 1 indeed bounds the algebraic error as we see from Theorem 3.

5.4 Lower bound

As seen in Section 3.2 (see (3.3)–(3.5)), a lower bound on the algebraic error is given by

$$(1-\gamma)\|\mathsf{U}^{i+\nu}-\mathsf{U}^i\|_{\mathbf{A}} \le \|\mathsf{U}-\mathsf{U}^i\|_{\mathbf{A}}$$

whenever $C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\| \le \gamma \|\mathsf{U}^{i+\nu}-\mathsf{U}^i\|_{\mathbf{A}}$ with a parameter $0<\gamma<1$. For the CG method, the estimator $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$ of (3.9) should be used instead. Alternatively, we can construct (cf. [41, Theorem 5.2]) a lower bound using homogeneous Dirichlet problems on patches ω_{a}, $\mathsf{a}\in\mathcal{V}_h$, (or larger subdomains of Ω):

Theorem 4 (Lower bound on the algebraic error)**.** *Let the assumptions of Theorem 2 be satisfied. For each vertex $\mathsf{a}\in\mathcal{V}_h$, let $m_{h,\mathsf{a}}\in V_h\cap H_0^1(\omega_{\mathsf{a}})$ be the solution of*

$$(\nabla m_{h,\mathsf{a}},\nabla v_h)_{\omega_{\mathsf{a}}} = (f,v_h)_{\omega_{\mathsf{a}}} - (\nabla u_h^i,\nabla v_h)_{\omega_{\mathsf{a}}} \qquad \forall v_h\in V_h\cap H_0^1(\omega_{\mathsf{a}}).$$

Set $m_h\equiv\sum_{\mathsf{a}\in\mathcal{V}_h} m_{h,\mathsf{a}}\in V_h$. Then

$$\|\nabla(u_h-u_h^i)\| \ge \mu_{\mathrm{alg}}^i \equiv \frac{\sum_{\mathsf{a}\in\mathcal{V}_h}\|\nabla m_{h,\mathsf{a}}\|_{\omega_{\mathsf{a}}}^2}{\|\nabla m_h\|}.$$

Proof. Using (5.3) and the fact that $m_h\in V_h$,

$$\|\nabla(u_h-u_h^i)\| \ge \frac{1}{\|\nabla m_h\|}(\nabla(u_h-u_h^i),\nabla m_h) = \frac{\sum_{\mathsf{a}\in\mathcal{V}_h}\|\nabla m_{h,\mathsf{a}}\|_{\omega_{\mathsf{a}}}^2}{\|\nabla m_h\|}.$$

□

6 Estimating the discretization error and construction of stopping criteria

A posteriori estimation of the discretization error $\|\nabla(u-u_h)\|$ is rather complicated as both u and u_h are unknown. The standard approaches proposed in literature are based on additional assumptions or properly justified heuristics on the algebraic error. Using

$$\|\nabla(u-u_h^i)\|^2 = \|\nabla(u-u_h)\|^2 + \|\nabla(u_h-u_h^i)\|^2 \tag{6.1}$$

that follows from the Galerkin orthogonality (2.4) and the results of the two previous sections, we give upper and lower bounds on the discretization error. We then propose global and local stopping criteria for a linear algebraic solver. In distinction with the previous works [32, Section 6.1] or [22, Section 3.3], the new stopping criteria guarantee that the iterations will not be stopped prematurely.

6.1 Lower bound

The first result follows easily from (6.1) and from the bounds of Theorems 2 and 3:

Theorem 5 (Lower bound on the discretization error)**.** *Let the assumptions of Theorems 2 and 3 hold. Let* $\mu^i_{\text{total}} > \eta^{i,\nu}_{\text{alg}}$*. Then*

$$\|\nabla(u-u_h)\| \geq \mu^{i,\nu}_{\text{discr}} \equiv \left[\left(\mu^i_{\text{total}}\right)^2 - \left(\eta^{i,\nu}_{\text{alg}}\right)^2\right]^{1/2}.$$

In practice the assumption $\mu^i_{\text{total}} > \eta^{i,\nu}_{\text{alg}}$ may not be satisfied in the iterations where $\|\nabla(u_h-u^i_h)\| \approx \|\nabla(u-u^i_h)\|$. The accuracy of the bound in Theorem 5 becomes good from the point where $\eta^{i,\nu}_{\text{alg}}$ gets small enough; see Section 7.4 for numerical illustrations.

6.2 Upper bound

One can similarly combine the upper bound on the total error of Theorem 1 and the lower bound on the algebraic error of Theorem 4 (note that $\eta^{i,\nu}_{\text{total}} \geq \mu^i_{\text{alg}}$):

Theorem 6 (Upper bound on the discretization error)**.** *Let the assumptions of Theorems 1 and 4 hold. Then*

$$\|\nabla(u-u_h)\| \leq \eta^{i,\nu}_{\text{discr}} \equiv \left[\left(\eta^{i,\nu}_{\text{total}}\right)^2 - \left(\mu^i_{\text{alg}}\right)^2\right]^{1/2}.$$

When the CG method is used for solving the algebraic system (2.5), $\mu^{\text{CG},i,\nu}_{\text{alg}}$ of (3.9) is suggested to be used instead of μ^i_{alg} above.

6.3 Stopping criteria balancing the error components

Stopping criteria for algebraic iterative solvers typically aim at stopping the iterations when the algebraic error does not substantially contribute to the total error. Using the (global) energy norm, it seems natural to require that

$$\|\nabla(u_h-u^i_h)\| \leq \gamma_{\text{alg}}\|\nabla(u-u_h)\|\,, \tag{6.2a}$$

where $\gamma_{\text{alg}} > 0$ is a prescribed tolerance. As mentioned above, the spatial distribution of the discretization error and of the algebraic error can be very different from each other and the criterion (6.2a) may not be descriptive; see [40]. Therefore one may rather require that

$$\|\nabla(u_h-u^i_h)\|_{\omega_\mathsf{a}} \leq \gamma_{\text{alg},\omega_\mathsf{a}}\|\nabla(u-u_h)\|_{\omega_\mathsf{a}} \qquad \forall \mathsf{a} \in \mathcal{V}_h. \tag{6.2b}$$

The stopping criteria proposed in [32, Section 6.1] or [22, Section 3.3] replaced $\|\nabla(u_h-u^i_h)\|$ and $\|\nabla(u-u_h)\|$ above by their computable estimates of the form (in the present setting) $\eta^{i,\nu}_{\text{alg}}$ and $\|\nabla u^i_h + \mathbf{d}^i_h\|$. Such criteria seem to work well in practice and allow to prove efficiency of the total error bound (see also Theorem 7 below), but they do not guarantee (6.2a) and there is a danger that the algebraic iterations can be stopped prematurely.

Using the upper bound on the algebraic error $\eta_{\mathrm{alg}}^{i,\nu}$ of Theorem 3 and the lower bound on the discretization error $\mu_{\mathrm{discr}}^{i,\nu}$ of Theorem 5, we propose the stopping criterion

$$\eta_{\mathrm{alg}}^{i,\nu} \leq \gamma_{\mathrm{alg}} \mu_{\mathrm{discr}}^{i,\nu} \tag{6.3}$$

that *guarantees* balancing the error components while implying the validity of (6.2a). Note that (6.3) is equivalent to requesting

$$\eta_{\mathrm{alg}}^{i,\nu} \leq \widetilde{\gamma}_{\mathrm{alg}} \mu_{\mathrm{total}}^{i} \qquad \text{with } \widetilde{\gamma}_{\mathrm{alg}} \equiv \gamma_{\mathrm{alg}} / (1 + \gamma_{\mathrm{alg}}^2)^{1/2} < 1.$$

Following [32, equation (6.3)] or [22, equations (3.13)–(3.15)] a *local stopping criterion* that mimics (6.2b) can be set as

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_{\mathtt{a}}} + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_{\omega_{\mathtt{a}}} \leq \widetilde{\gamma}_{\mathrm{alg},\omega_{\mathtt{a}}} \frac{\|\nabla m_{h,\mathtt{a}}\|_{\omega_{\mathtt{a}}}}{C_{\mathrm{cont,PF},\omega_{\mathtt{a}}}} \qquad \forall \mathtt{a} \in \mathcal{V}_h. \tag{6.4}$$

Unfortunately, the error estimator of Theorem 3 is not guaranteed to locally bound the algebraic error from above, so that (6.2b) may not be, in general, satisfied. Nevertheless, the criterion (6.4) is sufficient to prove the local efficiency of the total error estimator $\eta_{\mathrm{total}}^{i,\nu}$ (see Theorem 8 in Appendix A below) and it seems to ensure the local balance of the algebraic and discretization errors; see numerical experiments in Section 7.5.

7 Numerical illustrations

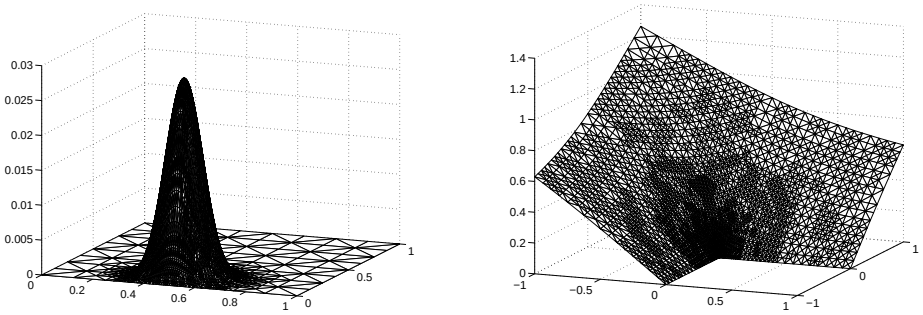


Figure 1: Left: solution (7.1) of the peak problem. Right: solution (7.2) of the L-shape problem.

For numerical illustration we use two test problems that were considered, e.g., in [34, 1].

Peak problem The model problem (2.1) with the square domain $\Omega \equiv (0,1) \times (0,1)$ and the right-hand side f chosen so that the solution u is given by

$$u(x,y) = x(x-1)y(y-1)\exp\left(-100\left(x-\frac{1}{2}\right)^2 - 100\left(y-\frac{117}{1000}\right)^2\right), \tag{7.1}$$

illustrated in Figure 1 (left). In the experiments, we discretize the problem on an adaptively refined mesh with 3 463 nodes using the piecewise quadratic polynomials. The corresponding algebraic system has 13 633 unknowns.

L-shape problem We take $\Omega \equiv (-1,1) \times (-1,1) \setminus [0,1] \times [-1,0]$ and solve

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

where the (inhomogeneous) Dirichlet boundary condition u_D is chosen so that the solution u is in polar coordinates (r, θ) given by

$$u(r,\theta) = r^{2/3} \sin\left(\frac{2}{3}\theta\right), \tag{7.2}$$

illustrated in Figure 1 (right). The extension of our estimates to $u_D \neq 0$ is possible following [19]. In particular, the flux reconstruction of Section 4.4 and the upper bound of Theorem 3 for the algebraic error remain unchanged. In the upper bound (4.9) and in Theorem 1, an additional term corresponding to the approximation of u_D by a piecewise polynomial function is added. This term is neglected in the experiments. We discretize the problem on an adaptively refined mesh with 628 nodes using the piecewise cubic polynomials. The corresponding algebraic system has here 5 098 unknowns.

The experiments are performed in Matlab R2014b with Partial Differential Equation Toolbox. We use our implementation of arbitrary degree conforming finite element method and of Raviart–Thomas–Nédélec spaces. We set $p' = p$, i.e., the reconstructed fluxes $\mathbf{d}_h^i$ are of the same order as the FEM approximation u_h^i. The algebraic system (2.5) is solved using the CG method preconditioned by the incomplete Cholesky decomposition with zero fill-in (Matlab `ichol` command) and starting with the zero initial guess. The exact solutions of the algebraic systems are approximated using the build-in Matlab "backslash" direct solver; in the performed numerical experiments, the algebraic error in this approximate solution is negligible. We point out that the experiments do not aim at the preconditioning tuned to the problem, but at demonstrating fairly the issues that might be encountered in practical use of the presented bounds.

The initial (uniform) meshes are generated using the Matlab Delaunay triangulation (`initmesh` command). For generating the sequence of adaptively refined meshes we, for the reproducibility of the results, refine according to the actual distribution of the *discretization error*, i.e., we compute (up to a quadrature error that is in the given experiments negligible) the discretization error $\|\nabla(u - u_h)\|_K$ on each element of the triangulation (recall that u_h is for the purpose of the experiments sufficiently accurately approximated using the direct solution of the algebraic system). We mark the smallest subset of elements that contributes to the squared energy norm of the discretization error by at least 25%. This requires ordering the elements according to the error size, which is in practice usually avoided, e.g., by proceeding as in [21, Section 5.2] or [49, pp. 10–11]. The refinement of the mesh uses the newest-vertex-bisection algorithm implemented in the Matlab `refinemesh` command.

7.1 Algebraic error: the cost of the additional iterations

We first compare the cost of the upper bounds on the algebraic error of Sections 3.2 and 5.3 in terms of the number ν of the additional algebraic iterations. For the given tolerance $\gamma_{\text{rem}} = 1, 0.5, 0.1$, we identify ν_1, ν_2, and ν_3 as the

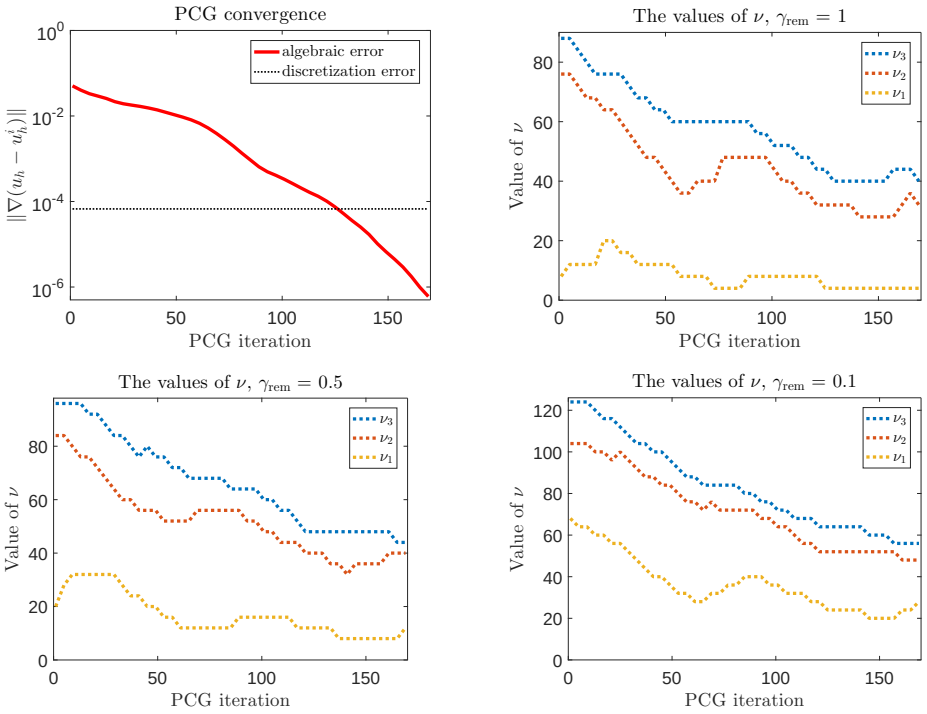


Figure 2: Peak problem: PCG convergence and the values of ν_1, ν_2, ν_3 determined by (7.3) for different choices of $\gamma_{\rm rem}$.

smallest values satisfying

$$\|\mathsf{R}^{i+\nu_1}\|_{\mathbf{A}^{-1}} \le \gamma_{\rm rem}\|\mathsf{U}^{i+\nu_1} - \mathsf{U}^i\|_{\mathbf{A}}, \tag{7.3a}$$

$$\|\mathbf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu_2}\| \le \gamma_{\rm rem}\|\mathsf{U}^{i+\nu_2} - \mathsf{U}^i\|_{\mathbf{A}}, \tag{7.3b}$$

$$C_{\rm F} h_\Omega \|r_h^{i+\nu_3}\| \le \gamma_{\rm rem}\|\mathbf{d}_h^{i+\nu_3} - \mathbf{d}_h^i\|, \tag{7.3c}$$

for each iteration step i. The number of additional iterations ν_1 of (7.3a) is always smaller than ν_2, ν_3. We recall, however, that $\|\mathsf{R}^{i+\nu_1}\|_{\mathbf{A}^{-1}} = \|\mathsf{U} - \mathsf{U}^{i+\nu_1}\|_{\mathbf{A}}$ is not available in practice. The criterion (7.3b) corresponds to the worst-case algebraic bound for $\|\mathsf{R}^{i+\nu_2}\|_{\mathbf{A}^{-1}}$ described in Section 3.1; see (3.6). For the purpose of the present study we (tightly) approximate the norm $\|\mathbf{A}^{-1}\|$ using the Matlab `eigs` command estimating the smallest eigenvalue of $\mathbf{A}$. Finally, the criterion (7.3c) corresponds to the computable upper bound of Theorem 3 on the algebraic error based on the flux reconstruction.

In the experiments (see Figures 2 and 3) we observe relatively large values of ν_2 and ν_3, with $\nu_2 \le \nu_3$. The large value of ν_3 indicates a possible nonnegligible cost of the upper bound of Theorem 3 (and also of the upper bound of Theorem 1 on the total error). The comparison with ν_1 reveals that there may be a room for further improvements. However, as demonstrated below, for the cost of the additional ν_3 iterations, we get in our experiments upper bounds for the total and algebraic errors with very favorable effectivity indices and, in particular, a remarkably accurate information on the *local distribution of these errors.*

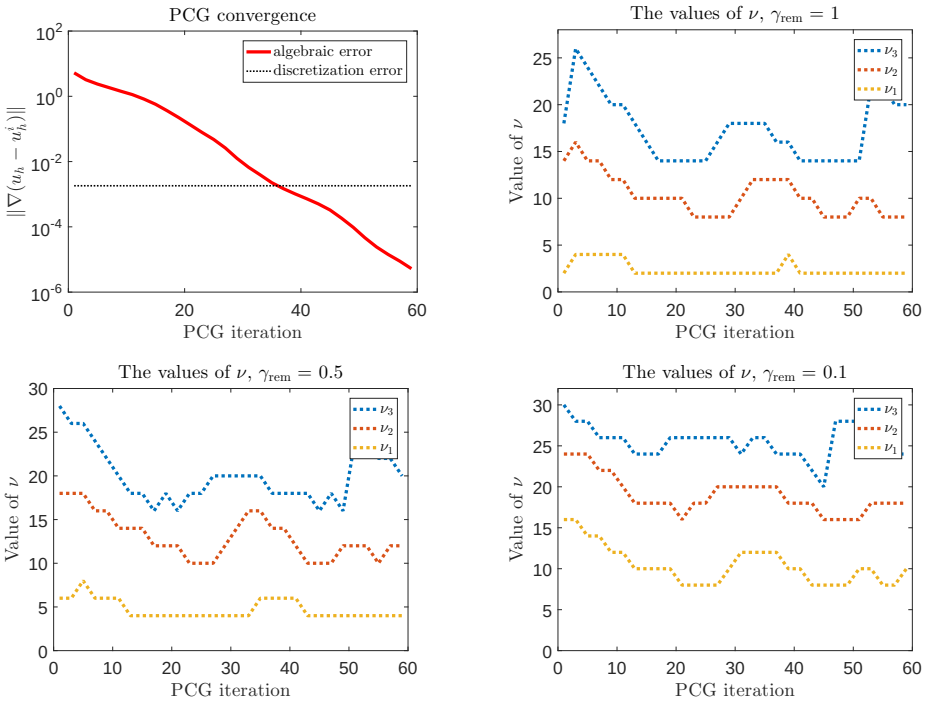


Figure 3: L-shape problem: PCG convergence and the values of ν_1, ν_2, ν_3 determined by (7.3) for different choices of $\gamma_{\rm rem}$.

7.2 Algebraic error: effectivity indices and localization

In this section we study how far the upper bounds on the algebraic error are from the actual error. For the ease of notation, let, corresponding to the bounds of Sections 3.2 and 5.3,

$$\eta_{\mathrm{alg},1}^{i,\nu_1} \equiv \|\mathsf{U}^{i+\nu_1} - \mathsf{U}^i\|_{\mathsf{A}} + \|\mathsf{R}^{i+\nu_1}\|_{\mathsf{A}^{-1}}, \tag{7.4a}$$

$$\eta_{\mathrm{alg},2}^{i,\nu_2} \equiv \|\mathsf{U}^{i+\nu_2} - \mathsf{U}^i\|_{\mathsf{A}} + \|\mathsf{A}^{-1}\|^{1/2} \cdot \|\mathsf{R}^{i+\nu_2}\|, \tag{7.4b}$$

$$\eta_{\mathrm{alg},3}^{i,\nu_3} \equiv \|\mathbf{d}_h^{i+\nu_3} - \mathbf{d}_h^i\| + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu_3}\|. \tag{7.4c}$$

Here ν_1, ν_2, and ν_3 are determined by (7.3). For these bounds, the effectivity indices

$$I_{\mathrm{eff}}^i(\eta_{\mathrm{alg},\bullet}^{i,\nu_\bullet}) \equiv \frac{\eta_{\mathrm{alg},\bullet}^{i,\nu_\bullet}}{\|\mathsf{U} - \mathsf{U}^i\|_{\mathsf{A}}} \tag{7.5}$$

are given in Figures 4–6. They confirm our expectation (see (3.4) and (3.5)) that

$$I_{\mathrm{eff}}^i(\eta_{\mathrm{alg},\bullet}^{i,\nu_\bullet}) \approx 1 + \gamma_{\rm rem},$$

so that, for the cost of $\nu_\bullet$ additional iterations, we get the estimates with the efficiency controlled by the parameter $\gamma_{\rm rem}$. In Figure 5, we give additionally the effectivity index

$$I_{\mathrm{eff}}^i(\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}) \equiv \frac{\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}}{\|\mathsf{U} - \mathsf{U}^i\|_{\mathsf{A}}}$$

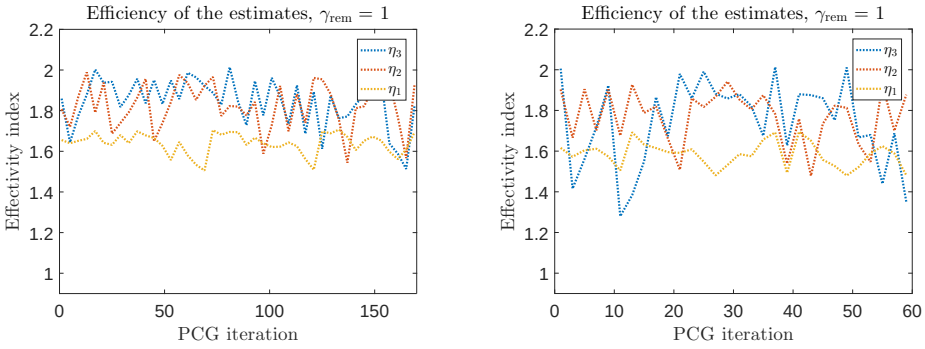


Figure 4: Effectivity indices $I^i_{\rm eff}(\eta^{i,\nu_\bullet}_{{\rm alg},\bullet})$ (7.5) of the algebraic error upper bounds (7.4) in the peak (left) and L-shape problems (right). The values of ν_1, ν_2, ν_3 are determined by (7.3) with $\gamma_{\rm rem} = 1$. Here $\eta^{i,\nu_k}_{{\rm alg},k}$ is simply denoted as η_k.

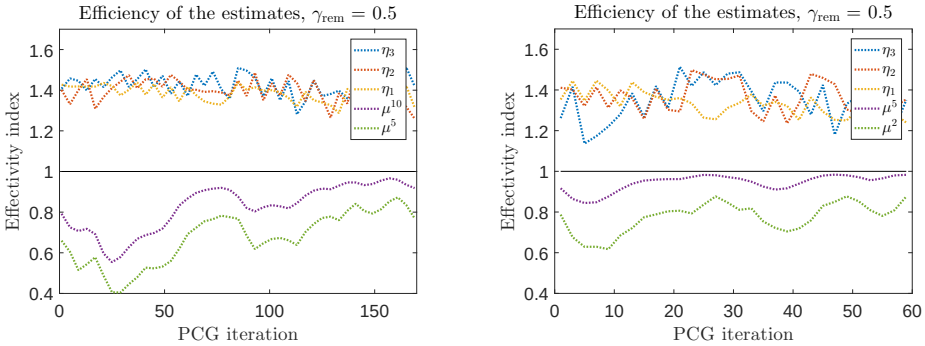


Figure 5: Effectivity indices $I^i_{\rm eff}(\eta^{i,\nu_\bullet}_{{\rm alg},\bullet})$ (7.5) of the algebraic error upper bounds (7.4) and the effectivity index $I^i_{\rm eff}(\mu^{{\rm CG},i,\nu}_{\rm alg})$ of the lower bound $\mu^{{\rm CG},i,\nu}_{\rm alg}$ with the fixed values of ν in the peak (left) and L-shape problems (right). The values of ν_1, ν_2, ν_3 are determined by (7.3) with $\gamma_{\rm rem} = 0.5$. Here $\eta^{i,\nu_k}_{{\rm alg},k}$ and $\mu^{{\rm CG},i,\nu}_{\rm alg}$ are simply denoted as η_k and μ^ν, respectively.

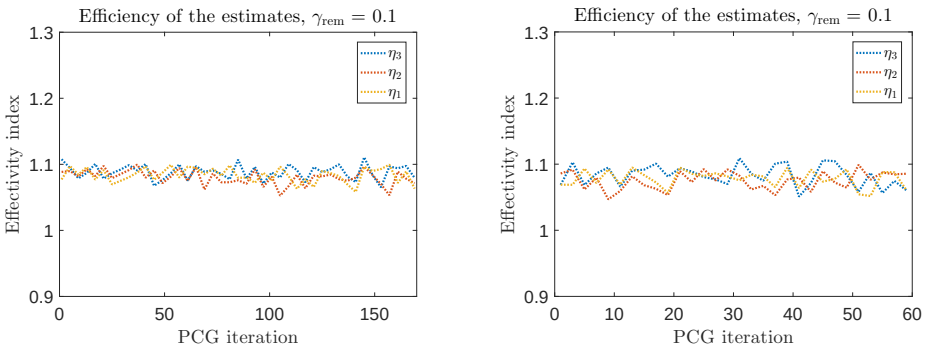


Figure 6: Effectivity indices $I^i_{\rm eff}(\eta^{i,\nu_\bullet}_{{\rm alg},\bullet})$ (7.5) of the algebraic error upper bounds (7.4) in the peak (left) and L-shape problems (right). The values of ν_1, ν_2, ν_3 are determined by (7.3) with $\gamma_{\rm rem} = 0.1$. Here $\eta^{i,\nu_k}_{{\rm alg},k}$ is simply denoted as η_k.

that illustrates the efficiency of the lower bound $\mu_{\text{alg}}^{\text{CG},i,\nu}$ (see (3.9)) from [50, 51], with the values of ν fixed for the peak and the L-shape problems to $\nu = 5$, 10 and 2, 5 respectively. We note that $I_{\text{eff}}^{i}(\mu_{\text{alg}}^{\text{CG},i,\nu})$ strongly depends on the decrease of the energy norm of the algebraic error between the iteration steps i and $i+\nu$. With a more powerful preconditioner resulting in a faster PCG convergence, analogous results will be achieved for much smaller number of additional algebraic iterations.

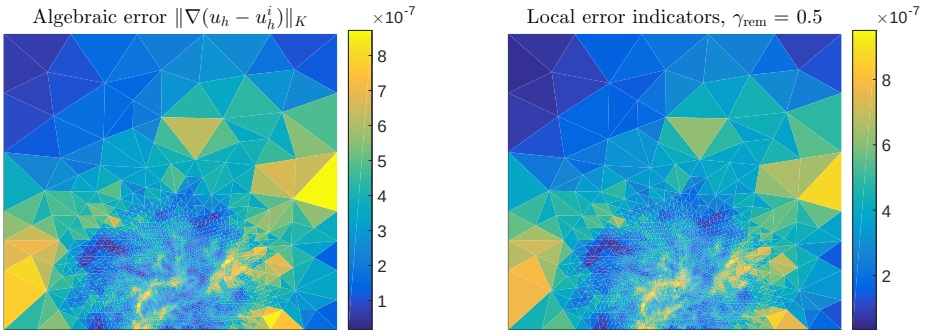


Figure 7: Peak problem, iteration $i = 137$: elementwise distribution of the algebraic error $\|\nabla(u_h - u_h^i)\|_K$ and the local algebraic error indicators $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\text{F}} h_\Omega \|r_h^{i+\nu}\|_K$. The value of ν, $\nu = 48$, is determined by (7.3c) with $\gamma_{\text{rem}} = 0.5$.

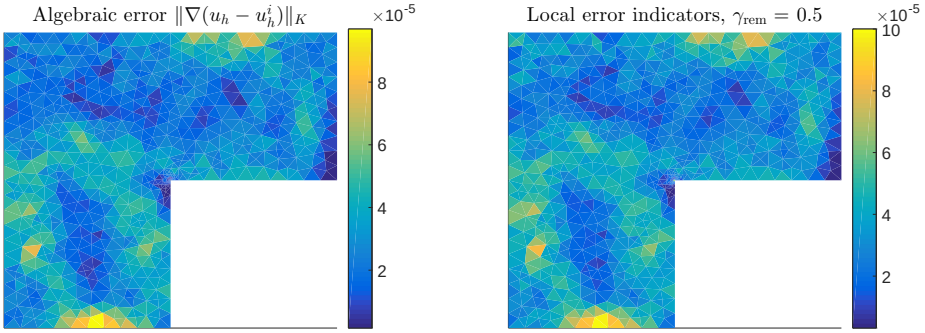


Figure 8: L-shape problem, iteration $i = 39$: elementwise distribution of the algebraic error $\|\nabla(u_h - u_h^i)\|_K$ and the local algebraic error indicators $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\text{F}} h_\Omega \|r_h^{i+\nu}\|_K$. The value of ν, $\nu = 18$, is determined by (7.3c) with $\gamma_{\text{rem}} = 0.5$.

As discussed in Remark 3, the flux-reconstruction-based upper bound of Theorem 3 allows evaluating the local indicators $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\text{F}} h_\Omega \|r_h^{i+\nu}\|_K$ and estimating the *local distribution of the algebraic error* $\|\nabla(u_h - u_h^i)\|_K$. As we can see in Figures 7 and 8, the local indicators provide a remarkably accurate description of the local distribution of the algebraic error. We observed similarly good results also in other iteration steps, choices of $\gamma_{\text{rem}} = 0.1, 1$, and other test problems. Please note that the algebraic error can be localized in parts of the discretization domain Ω where the discretization error can be small, see [40]

and Figures 10 and 11 below. We point out that the algebraic error does not equilibrate over the domain using the adaptive mesh refinement.

7.3 Bounding and localizing the total error

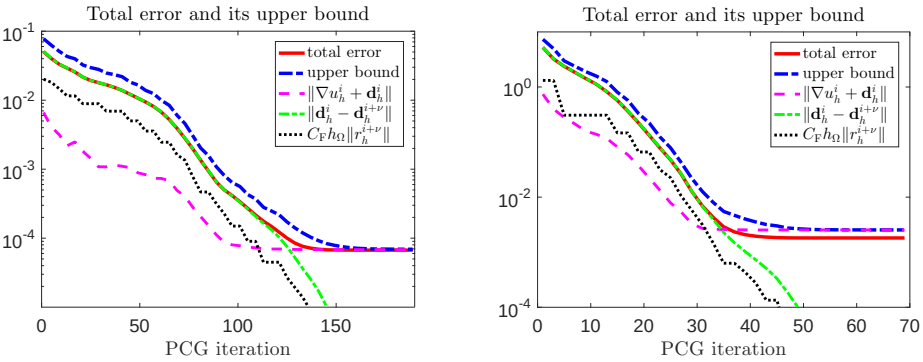


Figure 9: Total error $\|\nabla(u - u_h^i)\|$, the upper bound of Theorem 1, and the error indicators $\|\nabla u_h^i + \mathbf{d}_h^i\|$, $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|$, and $C_F h_\Omega \|r_h^{i+\nu}\|$ in the peak (left) and L-shape problems (right). The value of ν is determined by (7.3c) with $\gamma_{\text{rem}} = 0.5$.

We now illustrate the upper bound $\eta_{\text{total}}^{i,\nu}$ of Theorem 1. Figure 9 depicts the total error $\|\nabla(u - u_h^i)\|$, the upper bound, and the error indicators $\|\nabla u_h^i + \mathbf{d}_h^i\|$, $\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|$, and $C_F h_\Omega \|r_h^{i+\nu}\|$. We observe that $\eta_{\text{total}}^{i,\nu}$ tightly follows the actual value of the error. The parameter γ_{rem} in (7.3c) is set to 0.5.

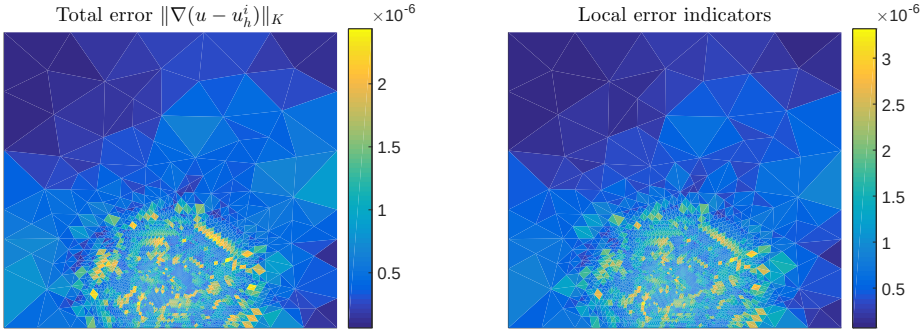


Figure 10: Peak problem: elementwise distribution of the total error $\|\nabla(u - u_h^i)\|_K$ and the local error indicators $\eta_{\text{osc},K} + \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_F h_\Omega \|r_h^{i+\nu}\|_K + \|\nabla u_h^i + \mathbf{d}_h^i\|_K$ in the iteration $i = 137$ with $\nu = 48$.

In Figures 10 and 11 we give the comparison of the local distribution of the total error $\|\nabla(u - u_h^i)\|_K$ and the sum $\eta_{\text{osc},K} + \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_F h_\Omega \|r_h^{i+\nu}\|_K + \|\nabla u_h^i + \mathbf{d}_h^i\|_K$ of the local indicators. Here the iteration step i and the number ν of additional iterations are set as the smallest values determined by the conditions (A.3a)–(A.3b) as described in Appendix A with $\gamma_{\text{alg}} = \gamma_{\text{rem}} = 0.5$.

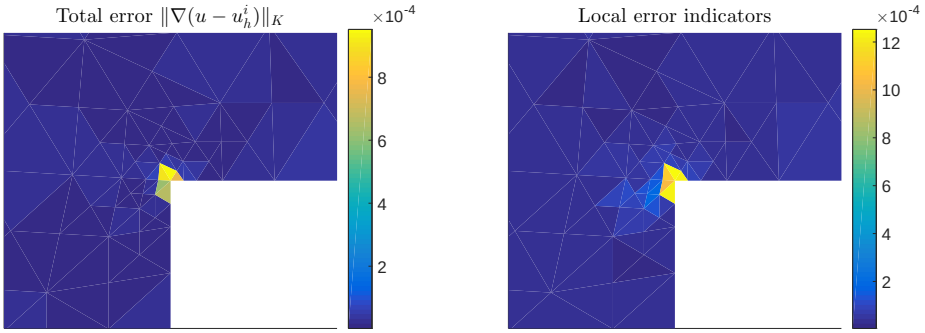


Figure 11: L-shape problem: elementwise distribution of the total error $\|\nabla(u-u_h^i)\|_K$ and the local error indicators $\eta_{\mathrm{osc},K} + \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K + \|\nabla u_h^i + \mathbf{d}_h^i\|_K$ in the iteration $i = 39$ with $\nu = 18$. We plot in both figures the part $[-0.02, 0.02] \times [-0.02, 0.02]$ of the discretization domain Ω.

7.4 Estimating the discretization error

We illustrate the discretization error bounds of Section 6. In Figures 12 and 13 we plot these bounds together with the estimator $\|\nabla u_h^i + \mathbf{d}_h^i\|$ that we have identified with the discretization error in Theorem 1. As in the previous experiments, the number ν of additional iterations is determined by (7.3c) with $\gamma_{\mathrm{rem}} = 0.5$.

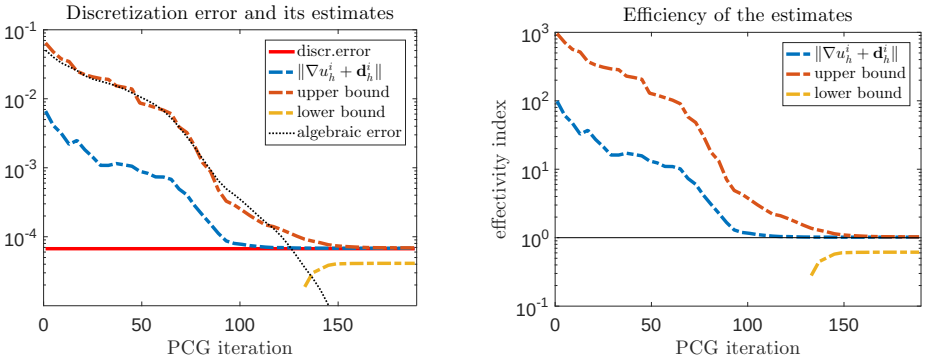


Figure 12: Peak problem: the discretization error $\|\nabla(u-u_h)\|$, the estimate $\|\nabla u_h^i + \mathbf{d}_h^i\|$, the upper bound $\eta_{\mathrm{discr}}^{i,\nu}$ of Theorem 6 with $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$, and the lower bound $\mu_{\mathrm{discr}}^{i,\nu}$ of Theorem 5 (left); the efficiency of the estimates (right).

Estimating the discretization error via Theorems 5 and 6 is naturely inaccurate in the iterations where the energy norm of the total error is mostly dominated by the algebraic error; cf. upper left parts of Figures 2 and 3. When the algebraic error drops below the discretization error, our upper and lower bounds get close to each other and provide a tight estimate for the discretization error.

In all performed experiments (here we present just a small sample), we have observed that $\|\nabla u_h^i + \mathbf{d}_h^i\| > \|\nabla(u-u_h)\|$, i.e., the estimate $\|\nabla u_h^i + \mathbf{d}_h^i\|$ gave an upper bound on the actual discretization error, and this bound was tighter than

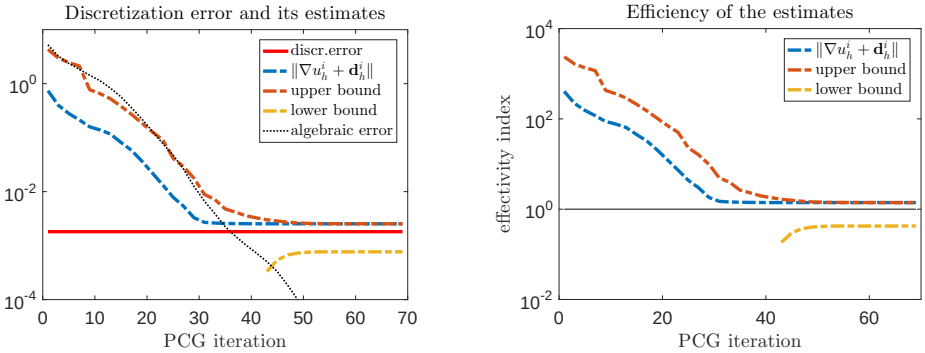


Figure 13: L-shape problem: the discretization error $\|\nabla(u-u_h)\|$, the estimate $\|\nabla u_h^i + \mathbf{d}_h^i\|$, the upper bound $\eta_{\mathrm{discr}}^{i,\nu}$ of Theorem 6 with $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$, and the lower bound of Theorem 5 (left); the efficiency of the estimates (right).

$\eta_{\mathrm{discr}}^{i,\nu}$ of Theorem 6 with $\mu_{\mathrm{alg}}^{\mathrm{CG},i,\nu}$. Therefore the components of the total error bound of Theorem 1 can be, in our test problems, indeed identified with the corresponding discretization and algebraic components of the total error. Consequently, the stopping criteria of [32, 22] (see also (A.3a)–(A.3b) below) seem to behave in practice similarly to the stopping criterion (6.3) that guarantees balancing the algebraic and the discretization error; see (6.2a).

7.5 Local stopping criteria and the spatial distribution of errors

We finally use the L-shape problem to illustrate that the local stopping criterion (6.4) prevents the algebraic error from dominating locally, as it can happen under the global criteria; cf. the numerical experiments of [40]. We consider the approximation u_h^{47} determined by the global stopping criterion (6.3) with $\gamma_{\mathrm{alg}} \equiv 0.5$ (the value of $\nu = 20$ is determined by (7.3c) with $\gamma_{\mathrm{rem}} \equiv 0.5$), and the approximation u_h^{79} satisfying the proposed local stopping criterion (6.4) with $\widetilde{\gamma}_{\mathrm{alg},\omega_{\mathsf{a}}} \equiv \gamma_{\mathrm{alg},\omega_{\mathsf{a}}}/(1+\gamma_{\mathrm{alg},\omega_{\mathsf{a}}}^2)^{1/2}$, $\gamma_{\mathrm{alg},\omega_{\mathsf{a}}} \equiv \gamma_{\mathrm{alg}}$, $\forall \mathsf{a} \in \mathcal{V}_h$ (the number $\nu = 20$ of the additional algebraic iterations is here determined by (A.8a) with $\gamma_{\mathrm{rem},K} \equiv \gamma_{\mathrm{rem}}$, $\forall K \in \mathcal{T}_h$).

Figure 14 depicts the differences $u-u_h^{47}$, $u-u_h^{79}$ and $u_h-u_h^{47}$, $u_h-u_h^{79}$ that visualize the total and algebraic errors respectively. We note that the algebraic part $u_h - u_h^{47}$ substantially affects the shape of $u - u_h^{47}$ in most of the domain Ω. This is not the case for $u - u_h^{79}$ as $|u(\mathbf{x}) - u_h(\mathbf{x})| \geq 10^{-7}$ in most of the domain Ω.

8 Conclusions and open questions

We have exposed in this paper in detail the methodology of $\mathbf{H}(\mathrm{div},\Omega)$-conforming flux and $H_0^1(\Omega)$-conforming residual reconstructions for estimating total, algebraic, and discretization errors for finite element discretizations and iterative algebraic solvers. The proposed upper and lower bounds are guaranteed and they contain no undetermined constants. We have used them for proposing stop-

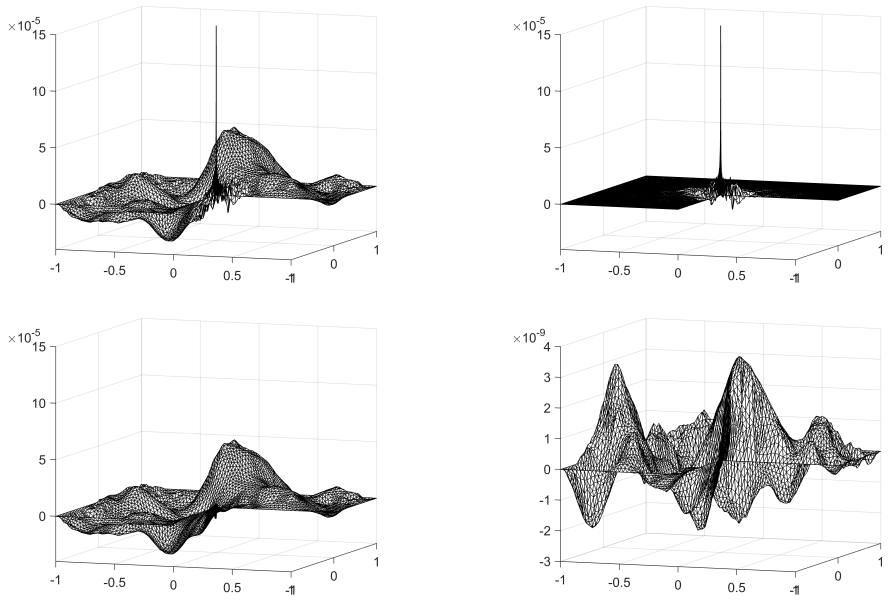


Figure 14: L-shape problem: the difference $u - u_h^{47}$ counting for the total error of the approximation u_h^{47} determined by the *global* stopping criterion (6.3) (upper left), its analogy $u - u_h^{79}$ for the approximation u_h^{79} determined by the *local* stopping criterion (6.4) (upper right), the algebraic part $u_h - u_h^{47}$ (bottom left), and its analogy $u_h - u_h^{79}$ (bottom right). Vertical axes are scaled by 10^{-5}, 10^{-5}, 10^{-5}, and 10^{-9}, respectively.

ping criteria for algebraic solvers that balance the algebraic and discretization errors and avoid stopping the algebraic iterations prematurely. As demonstrated on the model problems, they can practically localize very well the distribution of all errors and they can also avoid a possible local dominance of the algebraic error.

One part of the cost to be paid consists in a possibly nonnegligible amount of additional algebraic iterations that need to be performed. We have studied and reported this cost on two model examples in a rather unfavorable setting without a powerful preconditioner that would ensure very fast convergence and decrease this part of the cost to minimum. We believe that the presented methodology can be useful for many practical problems. Nevertheless, finding less costly alternatives within the presented framework is highly desirable and it represents one of our active research directions.

A Efficiency of the total error bound

We prove in this appendix the global and local efficiency of the upper bound of Theorem 1, which follows and extends the results in [22, 23, 41]. To simplify the presentation, we require that the source term f is piecewise polynomial, $f \in \mathbb{P}_{p'-1}(\mathcal{T}_h)$; see Section 4.4. Consequently, we choose $f_h = f$, and the oscillation term vanishes, $\eta_{\mathrm{osc}} = 0$.

The following lemma extends [13, Theorem 3.1] and [9, p. 1191] (see also [23, Lemma 3.12]) to the inexact algebraic solver case considered in this paper. Recall the space $H^1_*(\omega_\mathtt{a})$ introduced in (4.12).

Lemma 1. *Let $\mathtt{a} \in \mathcal{V}_h$ and let $m_\mathtt{a} \in H^1_*(\omega_\mathtt{a})$ be the solution of*

$$(\nabla m_\mathtt{a}, \nabla v)_{\omega_\mathtt{a}} = (f, \psi_\mathtt{a} v)_{\omega_\mathtt{a}} - \big(\nabla u_h^i, \nabla(\psi_\mathtt{a} v)\big)_{\omega_\mathtt{a}} - \big(r_h^i, \psi_\mathtt{a} v\big)_{\omega_\mathtt{a}} \qquad \forall v \in H^1_*(\omega_\mathtt{a}). \tag{A.1}$$

Then there holds

$$\|\nabla m_\mathtt{a}\|_{\omega_\mathtt{a}} \le C_{\mathrm{cont,PF},\omega_\mathtt{a}} \left(\|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}} + \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}}\right) + C_{\mathrm{PF},\omega_\mathtt{a}} h_{\omega_\mathtt{a}} \|r_h^{i+\nu}\|_{\omega_\mathtt{a}}.$$

Proof. From (A.1) and since, for $v \in H^1_*(\omega_\mathtt{a})$, $\psi_\mathtt{a} v \in H^1_0(\omega_\mathtt{a})$, we have, employing (2.2),

$$(\nabla m_\mathtt{a}, \nabla v)_{\omega_\mathtt{a}} = \big(\nabla(u - u_h^i), \nabla(\psi_\mathtt{a} v)\big)_{\omega_\mathtt{a}} - \big(r_h^i, \psi_\mathtt{a} v\big)_{\omega_\mathtt{a}}.$$

The Cauchy–Schwarz inequality and the bound (4.14) give

$$\big(\nabla(u - u_h^i), \nabla(\psi_\mathtt{a} v)\big)_{\omega_\mathtt{a}} \le \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}} C_{\mathrm{cont,PF},\omega_\mathtt{a}} \|\nabla v\|_{\omega_\mathtt{a}}.$$

Using (4.10), the Cauchy–Schwarz inequality, and (4.13),

$$\begin{aligned}
\big(r_h^i, \psi_\mathtt{a} v\big)_{\omega_\mathtt{a}} &= \big(\nabla\cdot \mathbf{d}_h^{i+\nu} - \nabla\cdot \mathbf{d}_h^i + r_h^{i+\nu}, \psi_\mathtt{a} v\big)_{\omega_\mathtt{a}} \\
&= \big(-\mathbf{d}_h^{i+\nu} + \mathbf{d}_h^i, \nabla(\psi_\mathtt{a} v)\big)_{\omega_\mathtt{a}} + \big(r_h^{i+\nu}, \psi_\mathtt{a} v\big)_{\omega_\mathtt{a}} \\
&\le \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} C_{\mathrm{cont,PF},\omega_\mathtt{a}} \|\nabla v\|_{\omega_\mathtt{a}} + \|r_h^{i+\nu}\|_{\omega_\mathtt{a}} \|\psi_\mathtt{a}\|_{\infty,\omega_\mathtt{a}} \|v\|_{\omega_\mathtt{a}} \\
&\le \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} C_{\mathrm{cont,PF},\omega_\mathtt{a}} \|\nabla v\|_{\omega_\mathtt{a}} + \|r_h^{i+\nu}\|_{\omega_\mathtt{a}} C_{\mathrm{PF},\omega_\mathtt{a}} h_{\omega_\mathtt{a}} \|\nabla v\|_{\omega_\mathtt{a}}.
\end{aligned}$$

Finally, using

$$\|\nabla m_\mathtt{a}\|_{\omega_\mathtt{a}} = \sup_{v \in H^1_*(\omega_\mathtt{a}), \|\nabla v\| = 1} (\nabla m_\mathtt{a}, \nabla v)_{\omega_\mathtt{a}}$$

and combining the above results yields the desired bound. □

The following crucial result has been shown in [9, Theorem 7] (see also [23, Corollary 3.16]) in the two-dimensional case. The three-dimensional proof is in [24].

Lemma 2. *Let $\mathbf{d}_{h,\mathtt{a}}^i$ be given by (4.11) with $p' = p + 1$ and let $m_\mathtt{a}$ be given by (A.1). Let $f \in \mathbb{P}_p(\mathcal{T}_h)$. Then there exists a constant $C_{\mathrm{st},\omega_\mathtt{a}} > 0$ depending only on the shape of elements of the patch $\mathcal{T}_\mathtt{a}$ but not on their diameters such that*

$$\|\psi_\mathtt{a} \nabla u_h^i + \mathbf{d}_{h,\mathtt{a}}^i\|_{\omega_\mathtt{a}} \le C_{\mathrm{st},\omega_\mathtt{a}} \|\nabla m_\mathtt{a}\|_{\omega_\mathtt{a}}. \tag{A.2}$$

The constant $C_{\mathrm{st},\omega_\mathtt{a}}$ is not computable. It can, however, be bounded from above considering a finite-dimensional subspace of $H^1_*(\omega_\mathtt{a})$ and solving the discrete version of the problem (A.1); see [23, Lemma 3.23]. Hereafter we denote

$$C_{\mathrm{cont,PF}} \equiv \max_{\mathtt{a} \in \mathcal{V}_h} C_{\mathrm{cont,PF},\omega_\mathtt{a}}, \quad C_{\mathrm{PF}} \equiv \max_{\mathtt{a} \in \mathcal{V}_h} C_{\mathrm{PF},\omega_\mathtt{a}}, \quad C_{\mathrm{st}} \equiv \max_{\mathtt{a} \in \mathcal{V}_h} C_{\mathrm{st},\omega_\mathtt{a}}.$$

We now state the main result on the *global efficiency* of the estimators of Theorem 1, both for the *global stopping criteria* in the sense of [32, 22] and for the secure stopping criterion in the sense of (6.3), relying on the estimator μ_{total}^i of Theorem 2:

Theorem 7 (Global efficiency)**.** *Let the estimators of Theorem 1 satisfy the global stopping criteria*

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\| \leq \gamma_{\mathrm{rem}} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|, \tag{A.3a}$$

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| \leq \gamma_{\mathrm{alg}} \|\nabla u_h^i + \mathbf{d}_h^i\| \tag{A.3b}$$

with positive parameters γ_{rem}, γ_{alg} *such that*

$$\gamma_{\mathrm{alg}} C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem}} \frac{C_{\mathrm{PF}} \max_{\mathtt{a} \in \mathcal{V}_h} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \right) \leq \frac{1}{2(d+1)}. \tag{A.4}$$

Alternatively, instead of (A.3)–(A.4)*, let*

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\| \leq \gamma_{\mathrm{rem}} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|, \tag{A.5a}$$

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| \leq \frac{\gamma_{\mathrm{alg}}}{(1+\gamma_{\mathrm{alg}}^2)^{1/2}} \, \mu_{\mathrm{total}}^i \tag{A.5b}$$

without any requirement on γ_{rem}, γ_{alg}*, supposing only*

$$\frac{C_{\mathrm{PF}} \max_{\mathtt{a} \in \mathcal{V}_h} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \leq C_{\mathrm{cont,PF}}$$

that is typically satisfied, apart possibly the coarsest meshes. Let the assumptions of Lemma 2 hold. Then the upper bound of Theorem 1 is globally efficient,

$$\eta_{\mathrm{total}}^{i,\nu} \leq C_{\text{glob. eff.}} \|\nabla(u - u_h^i)\|$$

with the global efficiency constant

$$C_{\text{glob. eff.}} \equiv (1 + \gamma_{\mathrm{alg}} + \gamma_{\mathrm{alg}} \gamma_{\mathrm{rem}}) 2(d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}}.$$

Recall that $\mathcal{V}_K$ stands for the vertices of the element K and that the functions $m_{h,\mathtt{a}}$ are specified in Theorem 2. Then the local version of Theorem 7 proving the *local efficiency* under the *local stopping criteria* is as follows:

Theorem 8 (Local efficiency)**.** *Let, for a given element* $K \in \mathcal{T}_h$*, the estimators of Theorem 1 satisfy the local stopping criteria*

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_{K'} \leq \gamma_{\mathrm{rem},K} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{K'} \quad \forall K' \in \mathcal{T}_h \text{ such that } K' \cap K \neq \emptyset, \tag{A.6a}$$

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} \leq \gamma_{\mathrm{alg},K} \|\nabla u_h^i + \mathbf{d}_h^i\|_K \qquad \forall \mathtt{a} \in \mathcal{V}_K \tag{A.6b}$$

with positive parameters $\gamma_{\mathrm{rem},K}$, $\gamma_{\mathrm{alg},K}$ *such that*

$$\gamma_{\mathrm{alg},K} C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem},K} \frac{C_{\mathrm{PF}} \max_{\mathtt{a} \in \mathcal{V}_K} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega} \right) \leq \frac{1}{2(d+1)}. \tag{A.7}$$

Alternatively, instead of (A.6)–(A.7)*, let, for all* $\mathtt{a} \in \mathcal{V}_K$*,*

$$C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_{\omega_\mathtt{a}} \leq \gamma_{\mathrm{rem},K} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}}, \tag{A.8a}$$

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} \leq \frac{\gamma_{\mathrm{alg},K}}{(1+\gamma_{\mathrm{alg},K}^2)^{1/2}} \frac{\|\nabla m_{h,\mathtt{a}}\|_{\omega_\mathtt{a}}}{C_{\mathrm{cont,PF},\omega_\mathtt{a}}}, \tag{A.8b}$$

without any requirement on $\gamma_{\mathrm{rem},K}$, $\gamma_{\mathrm{alg},K}$, *supposing only*

$$\frac{C_{\mathrm{PF}} \max_{\mathsf{a}\in\mathcal{V}_K} h_{\omega_\mathsf{a}}}{C_{\mathrm{F}} h_\Omega} \leq C_{\mathrm{cont,PF}}$$

that is typically satisfied, apart possibly the coarsest meshes. Let the assumptions of Lemma 2 hold. Then we have the local efficiency of the upper bound,

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K + \|\nabla u_h^i + \mathbf{d}_h^i\|_K \leq C_{\text{loc. eff.},K} \sum_{\mathsf{a}\in\mathcal{V}_K} \|\nabla(u - u_h^i)\|_{\omega_\mathsf{a}}$$

with the local efficiency constant

$$C_{\text{loc. eff.},K} \equiv (1 + \gamma_{\mathrm{alg},K} + \gamma_{\mathrm{alg},K}\gamma_{\mathrm{rem},K}) 2 C_{\mathrm{st}} C_{\mathrm{cont,PF}}.$$

Proof of Theorem 7. From the flux construction (4.11) of $\mathbf{d}_h^i$, using (A.2),

$$\begin{aligned}
\|\nabla u_h^i + \mathbf{d}_h^i\|^2 &= \sum_{K\in\mathcal{T}_h} \Big\| \sum_{\mathsf{a}\in\mathcal{V}_K} (\psi_\mathsf{a} \nabla u_h^i + \mathbf{d}_{h,\mathsf{a}}^i) \Big\|_K^2 \\
&\leq (d+1) \sum_{K\in\mathcal{T}_h} \sum_{\mathsf{a}\in\mathcal{V}_K} \|\psi_\mathsf{a} \nabla u_h^i + \mathbf{d}_{h,\mathsf{a}}^i\|_K^2 \\
&= (d+1) \sum_{\mathsf{a}\in\mathcal{V}_h} \|\psi_\mathsf{a} \nabla u_h^i + \mathbf{d}_{h,\mathsf{a}}^i\|_{\omega_\mathsf{a}}^2 \\
&\leq (d+1) C_{\mathrm{st}}^2 \sum_{\mathsf{a}\in\mathcal{V}_h} \|\nabla m_\mathsf{a}\|_{\omega_\mathsf{a}}^2,
\end{aligned}$$

as any element $K \in \mathcal{T}_h$ has $d+1$ vertices. From Lemma 1, we have

$$\begin{aligned}
\left[\sum_{\mathsf{a}\in\mathcal{V}_h} \|\nabla m_\mathsf{a}\|_{\omega_\mathsf{a}}^2\right]^{1/2} &\leq \left[\sum_{\mathsf{a}\in\mathcal{V}_h} C_{\mathrm{cont,PF},\omega_\mathsf{a}}^2 \|\nabla(u - u_h^i)\|_{\omega_\mathsf{a}}^2\right]^{1/2} \\
&+ \left[\sum_{\mathsf{a}\in\mathcal{V}_h} C_{\mathrm{cont,PF},\omega_\mathsf{a}}^2 \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathsf{a}}^2\right]^{1/2} + \left[\sum_{\mathsf{a}\in\mathcal{V}_h} C_{\mathrm{PF},\omega_\mathsf{a}}^2 (h_{\omega_\mathsf{a}})^2 \|r_h^{i+\nu}\|_{\omega_\mathsf{a}}^2\right]^{1/2}.
\end{aligned}$$

Therefore, using $\left[\sum_{\mathsf{a}\in\mathcal{V}_h} \|z\|_{\omega_\mathsf{a}}^2\right]^{1/2} = (d+1)^{1/2}\|z\|$,

$$\begin{aligned}
\|\nabla u_h^i + \mathbf{d}_h^i\| \leq{}& (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\nabla(u - u_h^i)\| \\
&+ (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\| \\
&+ (d+1) C_{\mathrm{st}} C_{\mathrm{PF}} \max_{\mathsf{a}\in\mathcal{V}_h} h_{\omega_\mathsf{a}} \|r_h^{i+\nu}\|.
\end{aligned} \tag{A.9}$$

From the stopping criteria (A.3),

$$\begin{aligned}
\|\nabla u_h^i + \mathbf{d}_h^i\| \leq{}& (d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\nabla(u - u_h^i)\| + (d+1)\gamma_{\mathrm{alg}} C_{\mathrm{st}} \\
&\left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem}} \frac{C_{\mathrm{PF}} \max_{\mathsf{a}\in\mathcal{V}_h} h_{\omega_\mathsf{a}}}{C_{\mathrm{F}} h_\Omega}\right) \|\nabla u_h^i + \mathbf{d}_h^i\|,
\end{aligned}$$

and from (A.4),

$$\|\nabla u_h^i + \mathbf{d}_h^i\| \leq 2(d+1) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \|\nabla(u - u_h^i)\|.$$

Finally, we get the assertion for the stopping criteria (A.3),

$$\begin{aligned}\eta^{i,\nu}_{\text{total}} &= \|\mathbf{d}^{i+\nu}_h - \mathbf{d}^i_h\| + C_{\mathrm{F}} h_\Omega \|r^{i+\nu}_h\| + \|\nabla u^i_h + \mathbf{d}^i_h\| \\ &\leq (1 + \gamma_{\text{alg}} + \gamma_{\text{alg}}\gamma_{\text{rem}})\|\nabla u^i_h + \mathbf{d}^i_h\| \\ &\leq C_{\text{glob. eff.}}\|\nabla(u - u^i_h)\|.\end{aligned}$$

The efficiency under the stopping criteria (A.5) actually does not request any restrictive assumptions of the form (A.4). Using (A.5b) and the bound of Theorem 2,

$$\|\mathbf{d}^{i+\nu}_h - \mathbf{d}^i_h\| \leq \frac{\gamma_{\text{alg}}}{(1+\gamma^2_{\text{alg}})^{1/2}}\|\nabla(u - u^i_h)\|.$$

Now a combination with (A.9) and (A.5a) gives

$$\begin{aligned}\|\nabla u^i_h + \mathbf{d}^i_h\| &\leq (d+1)C_{\text{st}}C_{\text{cont,PF}}\|\nabla(u - u^i_h)\| + (d+1)\frac{\gamma_{\text{alg}}}{(1+\gamma^2_{\text{alg}})^{1/2}}C_{\text{st}} \\ &\quad \left(C_{\text{cont,PF}} + \gamma_{\text{rem}}\frac{C_{\text{PF}}\max_{\mathsf{a}\in\mathcal{V}_h} h_{\omega_\mathsf{a}}}{C_{\mathrm{F}}h_\Omega}\right)\|\nabla(u - u^i_h)\|,\end{aligned}$$

so that the assertion for the stopping criteria (A.5) follows with the constant

$$\begin{aligned}&(d+1)C_{\text{st}}\left(C_{\text{cont,PF}} + \frac{\gamma_{\text{alg}}}{(1+\gamma^2_{\text{alg}})^{1/2}}C_{\text{cont,PF}} + \gamma_{\text{rem}}\frac{\gamma_{\text{alg}}}{(1+\gamma^2_{\text{alg}})^{1/2}}\frac{C_{\text{PF}}\max_{\mathsf{a}\in\mathcal{V}_h} h_{\omega_\mathsf{a}}}{C_{\mathrm{F}}h_\Omega}\right) \\ &\leq (1 + \gamma_{\text{alg}} + \gamma_{\text{alg}}\gamma_{\text{rem}})(d+1)C_{\text{st}}C_{\text{cont,PF}} \leq \frac{C_{\text{glob. eff.}}}{2}.\end{aligned}$$

□

Proof of Theorem 8. For the proof of the local efficiency, we first note that

$$\|\nabla u^i_h + \mathbf{d}^i_h\|_K \leq \sum_{\mathsf{a}\in\mathcal{V}_K}\|\psi_\mathsf{a}\nabla u^i_h + \mathbf{d}^i_{h,\mathsf{a}}\|_{\omega_\mathsf{a}} \leq \sum_{\mathsf{a}\in\mathcal{V}_K} C_{\text{st},\omega_\mathsf{a}}\|\nabla m_\mathsf{a}\|_{\omega_\mathsf{a}}.$$

From Lemma 1,

$$\begin{aligned}\|\nabla u^i_h + \mathbf{d}^i_h\|_K &\leq C_{\text{st}}C_{\text{cont,PF}}\sum_{\mathsf{a}\in\mathcal{V}_K}\|\nabla(u - u^i_h)\|_{\omega_\mathsf{a}} \\ &\quad + C_{\text{st}}C_{\text{cont,PF}}\sum_{\mathsf{a}\in\mathcal{V}_K}\|\mathbf{d}^{i+\nu}_h - \mathbf{d}^i_h\|_{\omega_\mathsf{a}} \\ &\quad + C_{\text{st}}C_{\text{PF}}\max_{\mathsf{a}\in\mathcal{V}_K} h_{\omega_\mathsf{a}}\sum_{\mathsf{a}\in\mathcal{V}_K}\|r^{i+\nu}_h\|_{\omega_\mathsf{a}}.\end{aligned} \tag{A.10}$$

Thus, under the stopping criteria (A.6),

$$\begin{aligned}\|\nabla u^i_h + \mathbf{d}^i_h\|_K &\leq C_{\text{st}}C_{\text{cont,PF}}\sum_{\mathsf{a}\in\mathcal{V}_K}\|\nabla(u - u^i_h)\|_{\omega_\mathsf{a}} + (d+1)C_{\text{st}}\gamma_{\text{alg},K} \\ &\quad \left(C_{\text{cont,PF}} + \gamma_{\text{rem},K}\frac{C_{\text{PF}}\max_{\mathsf{a}\in\mathcal{V}_K} h_{\omega_\mathsf{a}}}{C_{\mathrm{F}}h_\Omega}\right)\|\nabla u^i_h + \mathbf{d}^i_h\|_K.\end{aligned}$$

From (A.7), we further obtain

$$\|\nabla u^i_h + \mathbf{d}^i_h\|_K \leq 2C_{\text{st}}C_{\text{cont,PF}}\sum_{\mathsf{a}\in\mathcal{V}_K}\|\nabla(u - u^i_h)\|_{\omega_\mathsf{a}},$$

so that finally

$$\begin{aligned}\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_K + C_{\mathrm{F}} h_\Omega \|r_h^{i+\nu}\|_K + \|\nabla u_h^i + \mathbf{d}_h^i\|_K \\ \leq (1 + \gamma_{\mathrm{alg},K} + \gamma_{\mathrm{alg},K}\gamma_{\mathrm{rem},K})\|\nabla u_h^i + \mathbf{d}_h^i\| \\ \leq C_{\text{loc. eff.},K} \sum_{\mathtt{a}\in\mathcal{V}_K} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}}.\end{aligned}$$

Let $\widetilde{m}_\mathtt{a} \in H^1_*(\omega_\mathtt{a})$ be the solution of

$$(\nabla \widetilde{m}_\mathtt{a}, \nabla v)_{\omega_\mathtt{a}} = (f, \psi_\mathtt{a} v)_{\omega_\mathtt{a}} - \left(\nabla u_h^i, \nabla(\psi_\mathtt{a} v)\right)_{\omega_\mathtt{a}} \qquad \forall v \in H^1_*(\omega_\mathtt{a}),$$

in the continuous counterpart to $m_{h,\mathtt{a}}$ of Theorem 2 and similarly to (A.1). The fact that $m_{h,\mathtt{a}}$ is a projection of $\widetilde{m}_\mathtt{a}$ from $H^1_*(\omega_\mathtt{a})$ onto $W_h^\mathtt{a}$ gives $\|\nabla m_{h,\mathtt{a}}\|_{\omega_\mathtt{a}} \leq \|\nabla \widetilde{m}_\mathtt{a}\|_{\omega_\mathtt{a}}$. Proceeding as in the proof of Lemma 1 with $r_h^i = 0$, we get the inequality $\|\nabla \widetilde{m}_\mathtt{a}\|_{\omega_\mathtt{a}} \leq C_{\mathrm{cont,PF},\omega_\mathtt{a}} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}}$, so that

$$\|\nabla m_{h,\mathtt{a}}\|_{\omega_\mathtt{a}} \leq C_{\mathrm{cont,PF},\omega_\mathtt{a}} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}}.$$

Thus, under the secure local stopping criterion (A.8b), we obtain

$$\|\mathbf{d}_h^{i+\nu} - \mathbf{d}_h^i\|_{\omega_\mathtt{a}} \leq \frac{\gamma_{\mathrm{alg},K}}{(1 + \gamma_{\mathrm{alg},K}^2)^{1/2}} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}},$$

and, employing (A.10) and (A.8a),

$$\begin{aligned}\|\nabla u_h^i + \mathbf{d}_h^i\|_K \leq C_{\mathrm{st}} C_{\mathrm{cont,PF}} \sum_{\mathtt{a}\in\mathcal{V}_K} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}} + C_{\mathrm{st}} \frac{\gamma_{\mathrm{alg},K}}{(1 + \gamma_{\mathrm{alg},K}^2)^{1/2}} \\ \left(C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem},K} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_K} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega}\right) \sum_{\mathtt{a}\in\mathcal{V}_K} \|\nabla(u - u_h^i)\|_{\omega_\mathtt{a}}.\end{aligned}$$

The claim in this case thus follows from

$$\begin{aligned}C_{\mathrm{st}} \left(C_{\mathrm{cont,PF}} + \frac{\gamma_{\mathrm{alg},K}}{(1+\gamma_{\mathrm{alg},K}^2)^{1/2}} C_{\mathrm{cont,PF}} + \gamma_{\mathrm{rem},K} \frac{\gamma_{\mathrm{alg},K}}{(1+\gamma_{\mathrm{alg},K}^2)^{1/2}} \frac{C_{\mathrm{PF}} \max_{\mathtt{a}\in\mathcal{V}_K} h_{\omega_\mathtt{a}}}{C_{\mathrm{F}} h_\Omega}\right) \\ \leq (1 + \gamma_{\mathrm{alg},K} + \gamma_{\mathrm{alg},K}\gamma_{\mathrm{rem},K}) C_{\mathrm{st}} C_{\mathrm{cont,PF}} \leq \frac{C_{\text{loc. eff.},K}}{2}.\end{aligned}$$

□

References

[1] M. Ainsworth, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

[2] M. Arioli, E. H. Georgoulis, and D. Loghin, *Stopping criteria for adaptive finite element solvers*, SIAM J. Sci. Comput., 35 (2013), pp. A1537–A1559.

[3] M. Arioli, J. Liesen, A. Międlar, and Z. Strakoš, *Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems*, GAMM-Mitt., 36 (2013), pp. 102–129.

[4] I. Babuška and T. Strouboulis, *The finite element method and its reliability*, Numerical Mathematics and Scientific Computation, The Clarendon Press Oxford University Press, New York, 2001.

[5] R. Becker, C. Johnson, and R. Rannacher, *Adaptive error control for multigrid finite element methods*, Computing, 55 (1995), pp. 271–288.

[6] R. Becker and S. Mao, *Convergence and quasi-optimal complexity of a simple adaptive finite element method*, M2AN Math. Model. Numer. Anal., 43 (2009), pp. 1203–1219.

[7] R. Becker, S. Mao, and Z. Shi, *A convergent nonconforming adaptive finite element method with quasi-optimal complexity*, SIAM J. Numer. Anal., 47 (2010), pp. 4639–4659.

[8] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Local error estimates and adaptive refinement for first-order system least squares (FOSLS)*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43. Special issue on multilevel methods (Copper Mountain, CO, 1997).

[9] D. Braess, V. Pillwein, and J. Schöberl, *Equilibrated residual error estimates are p-robust*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1189–1197.

[10] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp., 77 (2008), pp. 651–672.

[11] C. Burstedde and A. Kunoth, *A wavelet-based nested iteration-inexact conjugate gradient algorithm for adaptively solving elliptic PDEs*, Numer. Algorithms, 48 (2008), pp. 161–188.

[12] D. Calvetti, S. Morigi, L. Reichel, and F. Sgallari, *Computable error bounds and estimates for the conjugate gradient method*, Numer. Algorithms, 25 (2000), pp. 75–88.

[13] C. Carstensen and S. A. Funken, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput., 21 (1999/00), pp. 1465–1484.

[14] P. G. Ciarlet, *The finite element method for elliptic problems*, vol. 40 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam].

[15] ——, *Linear and nonlinear functional analysis with applications*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

[16] G. Dahlquist, G. H. Golub, and S. G. Nash, *Bounds for the error in linear systems*, in Semi-infinite programming (Proc. Workshop, Bad Honnef, 1978), vol. 15 of Lecture Notes in Control and Information Sci., Springer, Berlin, 1979, pp. 154–172.

[17] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Math. Comp., 68 (1999), pp. 1379–1396.

[18] P. Deuflhard, *Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results*, in Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993), vol. 180 of Contemp. Math., American Mathematical Society, Providence, RI, 1994, pp. 29–42.

[19] V. Dolejší, A. Ern, and M. Vohralík, *hp-adaptation driven by polynomial-degree-robust a posteriori error estimates for elliptic problems.* HAL Preprint 01165187, 2015.

[20] V. Dolejší, I. Šebestová, and M. Vohralík, *Algebraic and discretization error estimation by equilibrated fluxes for discontinuous Galerkin methods on nonmatching grids*, J. Sci. Comput., 64 (2015), pp. 1–34.

[21] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[22] A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*, SIAM J. Sci. Comput., 35 (2013), pp. A1761–A1791.

[23] ——, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal., 53 (2015), pp. 1058–1081.

[24] ——, *Polynomial-degree-robust flux and potential reconstruction in three space dimensions.* in preparation, 2016.

[25] T. Gergelits and Z. Strakoš, *Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations*, Numer. Algorithms, 65 (2014), pp. 759–782.

[26] G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Numerical analysis 1993 (Dundee, 1993), vol. 303 of Pitman Res. Notes Math. Ser., Longman Sci. Tech., Harlow, 1994, pp. 105–156.

[27] ——, *Matrices, moments and quadrature. II. How to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705.

[28] G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.

[29] H. Harbrecht and R. Schneider, *On error estimation in finite element methods without having Galerkin orthogonality*, Berichtsreihe des SFB 611 457, Universität Bonn, 2009.

[30] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[31] R. Hiptmair, *Operator preconditioning*, Comput. Math. Appl., 52 (2006), pp. 699–706.

[32] P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.

[33] J. Liesen and Z. Strakoš, *Krylov subspace methods: principles and analysis*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[34] R. Luce and B. I. Wohlmuth, *A local a posteriori error estimator based on equilibrated fluxes*, SIAM J. Numer. Anal., 42 (2004), pp. 1394–1414.

[35] J. Málek and Z. Strakoš, *Preconditioning and the conjugate gradient method in the context of solving PDEs*, vol. 1 of SIAM Spotlights, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015.

[36] G. Meurant, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numer. Algorithms, 16 (1997), pp. 77–87 (1998).

[37] ——, *Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm*, Numer. Algorithms, 22 (1999), pp. 353–365 (1999).

[38] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542.

[39] G. Meurant and P. Tichý, *On computing quadrature-based bounds for the A-norm of the error in conjugate gradients*, Numer. Algorithms, 62 (2013), pp. 163–191.

[40] J. Papež, J. Liesen, and Z. Strakoš, *Distribution of the discretization and algebraic error in numerical solution of partial differential equations*, Linear Algebra Appl., 449 (2014), pp. 89–114.

[41] J. Papež, U. Rüde, M. Vohralík, and B. Wohlmuth, *Guaranteed algebraic, discretization, and total error bounds in multigrid finite element solvers*. In preparation, 2016.

[42] A. T. Patera and E. M. Rønquist, *A general output bound result: application to discretization and iteration error estimation and control*, Math. Models Methods Appl. Sci., 11 (2001), pp. 685–712.

[43] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Rational Mech. Anal., 5 (1960), pp. 286–292 (1960).

[44] R. Rannacher, *Error control in finite element computations. An introduction to error estimation and mesh-size adaptation*, in Error control and adaptivity in scientific computing (Antalya, 1998), vol. 536 of NATO Sci. Ser. C Math. Phys. Sci., Kluwer Acad. Publ., Dordrecht, 1999, pp. 247–278.

[45] K. Rektorys, *Variational methods in mathematics, science and engineering*, D. Reidel Publishing Co., Dordrecht-Boston, Mass., second ed., 1980. Translated from the Czech by Michael Basch.

[46] S. Repin, *A posteriori estimates for partial differential equations*, vol. 4 of Radon Series on Computational and Applied Mathematics, Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[47] V. V. Shaidurov, *Some estimates of the rate of convergence for the cascadic conjugate-gradient method*, Comput. Math. Appl., 31 (1996), pp. 161–171.

[48] D. J. Silvester and V. Simoncini, *An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation*, ACM Trans. Math. Software, 37 (2011), pp. Art. 42, 22.

[49] R. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269.

[50] Z. Strakoš and P. Tichý, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electron. Trans. Numer. Anal., 13 (2002), pp. 56–80.

[51] ——, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817.

[52] A. Veeser and R. Verfürth, *Poincaré constants for finite element stars*, IMA J. Numer. Anal., 32 (2012), pp. 30–47.

[53] R. Verfürth, *A posteriori error estimation techniques for finite element methods*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2013.

[54] B. I. Wohlmuth and R. H. W. Hoppe, *A comparison of a posteriori error estimators for mixed finite element discretizations by Raviart-Thomas elements*, Math. Comp., 68 (1999), pp. 1347–1378.

5.2 Further comments

The relationship between the upper bounds (5.4) and (5.8) not resolved in the submitted text Papež et al. [2016] is shown in this section.

The evaluation of the bound (5.4)

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathsf{A}} \leq C_{\mathrm{F}} h_\Omega \|\mathsf{R}^i\|_{\mathsf{G}^{-1}}$$

requires the solve with the global mass matrix G (see (5.1)), while for the evaluation of the bound (5.8)

$$\|\mathsf{U} - \mathsf{U}^i\|_{\mathsf{A}} \leq C_{\mathrm{F}} h_\Omega \left(\sum_{K \in \mathcal{T}_h} \|\mathsf{R}^i_K\|^2_{\mathsf{G}_K^{-1}} \right)^{1/2},$$

only small systems with local mass matrices G_K and the right-hand sides R^i_K have to be solved; see (5.2). We recall that

$$(\mathsf{G}_K)_{j\ell} = (\psi_\ell, \psi_j)_K \qquad \text{for } \psi_\ell, \psi_j \text{ nonvanishing on } K$$

and

$$\mathsf{R}^i_K = \left[\mathsf{R}^i_j / n_j\right] \qquad \text{for } \psi_j \text{ nonvanishing on } K,$$

where n_j denotes the number of mesh elements forming the support of the basis function ψ_j, $j = 1, \ldots, N$. Note that the residual entry R^i_j is distributed equally over the elements forming the support of ψ_j.

Consider an (arbitrary but fixed) ordering of the mesh elements $K_1, \ldots, K_s$. Let $\widetilde{\mathsf{G}} \in \mathbb{R}^{M \times M}$, be the block diagonal symmetric positive definite matrix with matrices G_{K_m} on its diagonal,

$$\widetilde{\mathsf{G}} \equiv \begin{bmatrix} \mathsf{G}_{K_1} & & \\ & \ddots & \\ & & \mathsf{G}_{K_s} \end{bmatrix}$$

and denote by $\widetilde{\mathsf{R}}^i$ the vector of length M consisting of stacked vectors $\mathsf{R}^i_{K_m}$,

$$\widetilde{\mathsf{R}}^i \equiv \begin{bmatrix} \mathsf{R}^i_{K_1} \\ \vdots \\ \mathsf{R}^i_{K_s} \end{bmatrix}.$$

Then

$$\sum_{K \in \mathcal{T}_h} \|\mathsf{R}^i_K\|^2_{\mathsf{G}_K^{-1}} = (\widetilde{\mathsf{R}}^i)^T \widetilde{\mathsf{G}}^{-1} \widetilde{\mathsf{R}}^i.$$

The k-th element $\widetilde{\mathsf{R}}^i_k$ of the vector $\widetilde{\mathsf{R}}^i$, $k = 1, \ldots, M$, is equal to R^i_j / n_j for some index $j \in \{1, \ldots, N\}$, and we can define the mapping $\Theta : \{1, \ldots, M\} \to \{1, \ldots, N\}$, $\Theta(k) = j$. From the definition of n_j, there are n_j indices $k \in \{1, \ldots, M\}$ such that $\Theta(k) = j$. There holds

$$\sum_{k \,|\, \Theta(k) = j} \widetilde{\mathsf{R}}^i_k = n_j \cdot \mathsf{R}^i_j / n_j = \mathsf{R}^i_j,$$

$$\sum_{\substack{k \,|\, \Theta(k) = j \\ g \,|\, \Theta(g) = \ell}} (\widetilde{\mathsf{G}})_{kg} = \sum_{m=1}^{s} (\mathsf{G}_{K_m})_{j\ell} = \sum_{m=1}^{s} (\psi_\ell, \psi_j)_{K_m} = (\psi_\ell, \psi_j) = (\mathsf{G})_{j\ell}.$$

Denote by $\mathbf{P}$ the matrix of size $M \times N$ that has in each row only one non-zero entry, which is equal to 1 and is on the position $(k, \Theta(k))$, $k = 1, \dots, M$. Equivalently, the matrix $\mathbf{P}$ has in its j-th column n_j non-zero entries in the rows with indices k such that $\Theta(k) = j$. Then

$$\mathbf{P}^T \widetilde{\mathsf{R}}^i = \mathsf{R}^i, \qquad \mathbf{P}^T \widetilde{\mathbf{G}} \mathbf{P} = \mathbf{G}.$$

Using

$$\|\mathsf{R}^i\|^2_{\mathbf{G}^{-1}} = (\mathsf{R}^i)^T \mathbf{G}^{-1} \mathsf{R}^i = (\widetilde{\mathsf{R}}^i)^T \mathbf{P} (\mathbf{P}^T \widetilde{\mathbf{G}}^{-1} \mathbf{P})^{-1} \mathbf{P}^T \widetilde{\mathsf{R}}^i,$$

we have

$$\begin{aligned}
\frac{\|\mathsf{R}^i\|^2_{\mathbf{G}^{-1}}}{\sum_{K \in \mathcal{T}_h} \|\mathsf{R}^i_K\|^2_{\mathbf{G}_K^{-1}}} &= \frac{(\widetilde{\mathsf{R}}^i)^T \mathbf{P} (\mathbf{P}^T \widetilde{\mathbf{G}}^{-1} \mathbf{P})^{-1} \mathbf{P}^T \widetilde{\mathsf{R}}^i}{(\widetilde{\mathsf{R}}^i)^T \widetilde{\mathbf{G}}^{-1} \widetilde{\mathsf{R}}^i} \\
&= \frac{(\widetilde{\mathbf{G}}^{-1/2} \widetilde{\mathsf{R}}^i)^T \widetilde{\mathbf{G}}^{1/2} \mathbf{P} (\mathbf{P}^T \widetilde{\mathbf{G}}^{-1} \mathbf{P})^{-1} \mathbf{P}^T \widetilde{\mathbf{G}}^{1/2} (\widetilde{\mathbf{G}}^{-1/2} \widetilde{\mathsf{R}}^i)}{(\widetilde{\mathbf{G}}^{-1/2} \widetilde{\mathsf{R}}^i)^T (\widetilde{\mathbf{G}}^{-1/2} \widetilde{\mathsf{R}}^i)} \\
&\leq \sup_{\mathsf{V} \in \mathbb{R}^M, \mathsf{V} \neq 0} \frac{\mathsf{V}^T \widetilde{\mathbf{G}}^{1/2} \mathbf{P} (\mathbf{P}^T \widetilde{\mathbf{G}}^{-1} \mathbf{P})^{-1} \mathbf{P}^T \widetilde{\mathbf{G}}^{1/2} \mathsf{V}}{\mathsf{V}^T \mathsf{V}} \\
&= \|\widetilde{\mathbf{G}}^{1/2} \mathbf{P} (\mathbf{P}^T \widetilde{\mathbf{G}}^{-1} \mathbf{P})^{-1} \mathbf{P}^T \widetilde{\mathbf{G}}^{1/2}\|.
\end{aligned}$$

Since $\widetilde{\mathbf{G}}^{1/2} \mathbf{P} (\mathbf{P}^T \widetilde{\mathbf{G}}^{-1} \mathbf{P})^{-1} \mathbf{P}^T \widetilde{\mathbf{G}}^{1/2}$ is a symmetric (i.e. orthogonal) nonzero projection,

$$\|\widetilde{\mathbf{G}}^{1/2} \mathbf{P} (\mathbf{P}^T \widetilde{\mathbf{G}}^{-1} \mathbf{P})^{-1} \mathbf{P}^T \widetilde{\mathbf{G}}^{1/2}\| = 1,$$

and therefore

$$\|\mathsf{R}^i\|^2_{\mathbf{G}^{-1}} \leq \sum_{K \in \mathcal{T}_h} \|\mathsf{R}^i_K\|^2_{\mathbf{G}_K^{-1}}.$$

Consequently, the bound (5.8) is weaker than the bound (5.4).

Figures 5.1 and 5.2 depict the comparison of the bounds (5.4) and (5.8) in the test problems from the paper Papež et al. [2016]. We note that the relative overestimation of the bound (5.8) is in these problems quite moderate; below 18% in the peak problem and below 12% in the L-shape problem.

Bibliography

J. Papež, Z. Strakoš, and M. Vohralík. Estimating and localizing the algebraic and total numerical errors using flux reconstructions. Preprint MORE/2016/12, Submitted for publication, 2016.

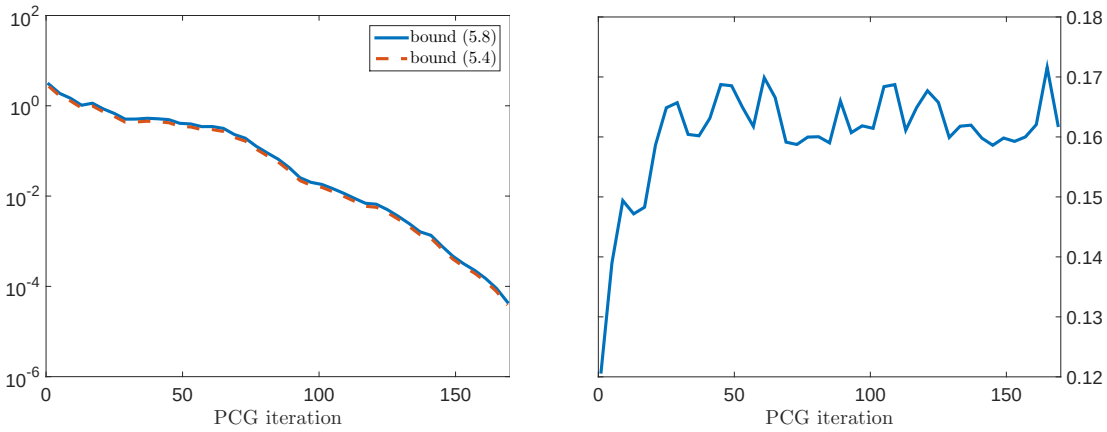


Figure 5.1: Peak problem: upper bounds (5.4) and (5.8) (left); the relative difference $\left(\left(\sum_{K\in\mathcal{T}_h}\|\mathsf{R}_K^i\|^2_{\mathsf{G}_K^{-1}}\right)^{1/2}-\|\mathsf{R}^i\|_{\mathsf{G}^{-1}}\right)/\|\mathsf{R}^i\|_{\mathsf{G}^{-1}}$ (right).

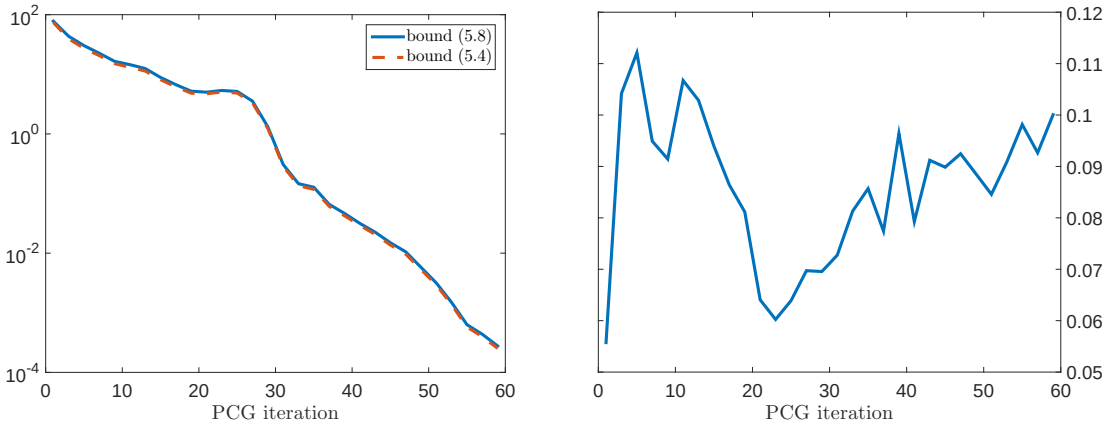


Figure 5.2: L-shape problem: upper bounds (5.4) and (5.8) (left); the relative difference $\left(\left(\sum_{K\in\mathcal{T}_h}\|\mathsf{R}_K^i\|^2_{\mathsf{G}_K^{-1}}\right)^{1/2}-\|\mathsf{R}^i\|_{\mathsf{G}^{-1}}\right)/\|\mathsf{R}^i\|_{\mathsf{G}^{-1}}$ (right).

6. Preconditioning as transformation of discretization basis functions

When solving difficult problems, an inherent part of any practical iterative solver is an algebraic preconditioning, i.e. the transformation of the algebraic system that aims at faster convergence behavior of the algebraic solver. As shown in [Málek and Strakoš, 2015, Chapter 8], algebraic preconditioning in the conjugate gradient method can be regarded as a transformation of the discretization basis and, at the same time, transformation of the inner product in the given Hilbert space. In this chapter we recall the results of [Málek and Strakoš, 2015, Chapter 8] and in numerical examples we study how the transformed discretization basis functions look like.

We also note that it has no reason to measure the quality of computed approximations using the algebraic quantities (such as algebraic residual norm or the Euclidean norm of the algebraic error) with no counterparts in the corresponding function space. In other words, we are interested in convergence of functions, not in convergence of their coordinates in the particular basis.

The text and the notation in the chapter is based on Málek and Strakoš [2015], which provides the references to original results, the complete proofs, and a thorough discussion. The numerical experiments are the result of the joint work with Tomáš Gergelits and Zdeněk Strakoš.

6.1 Notation and setting

Consider a problem given in the weak form

$$\text{to find } u \in V: \quad a(u,v) = \langle f, v\rangle \qquad \forall v \in V, \tag{6.1}$$

where

- V is an infinite-dimensional Hilbert space with the inner product $(\cdot,\cdot)_V$,
- $V^\#$ is the dual space consisting of linear bounded functionals on V,
- $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a bilinear form,
- $f \in V^\#$ is a bounded linear functional on V,
- $\langle \cdot,\cdot\rangle : V^\# \times V \to \mathbb{R}$ is the duality pairing.

Let $\|\cdot\|_V \equiv \sqrt{(\cdot,\cdot)_V}$ be the norm on V induced by the given inner product, and $\|\cdot\|_{V^\#}$ the dual norm on $V^\#$,

$$\|f\|_{V^\#} \equiv \sup_{v\in V; \|v\|_V=1} |\langle f, v\rangle| \,.$$

We further assume that the bilinear form $a(\cdot,\cdot)$ is symmetric, bounded, and V-elliptic, i.e.

$$\begin{aligned} a(w,v) &= a(v,w), && (6.2a)\\ |a(w,v)| &\le C\|w\|_V\|v\|_V, && (6.2b)\\ a(v,v) &\ge \alpha\|v\|_V^2, \qquad \forall v,w\in V, && (6.2c) \end{aligned}$$

where $C>0$, $\alpha>0$. The boundedness and V-ellipticity of $a(\cdot,\cdot)$ assures the existence and uniqueness of the solution of (6.1) for any $f\in V^\#$ through the Lax–Milgram lemma; see, e.g., [Málek and Strakoš, 2015, Section 3.3]. Moreover, the solution u is continuously dependent on the right-hand side functional f,

$$\|u\|_V \le \frac{1}{\alpha}\|f\|_{V^\#},$$

with the coercivity constant α; see (6.2c). Assuming (6.2), the bilinear form $a(\cdot,\cdot)$ is an inner product on V and it induces the *energy norm*

$$\|v\|_a \equiv \sqrt{a(v,v)}, \qquad v\in V. \tag{6.3}$$

Defining

$$\mathcal{A}: V\to V^\#, \qquad \langle \mathcal{A}w, v\rangle = a(w,v) \qquad \forall v,w\in V,$$

we can rewrite (6.1) as

$$\text{find } u\in V: \quad \langle \mathcal{A}u, v\rangle = \langle f, v\rangle \qquad \forall v\in V,$$

or as the problem in the dual space $V^\#$,

$$\mathcal{A}u = f, \qquad u\in V, \quad f\in V^\#, \tag{6.4}$$

with (6.2) assuring that the operator $\mathcal{A}$ is self-adjoint with respect to the duality pairing, i.e.

$$\langle \mathcal{A}v, w\rangle = \langle \mathcal{A}w, v\rangle \qquad \forall v,w\in V,$$

it is bounded and (α-)coercive.

6.2 Riesz map and the operator preconditioning

The *Riesz map* $\tau: V^\#\to V$ determined by the given inner product $(\cdot,\cdot)_V$ provides an isometric isomorphism between V and $V^\#$. For each $f\in V^\#$ there exists unique $\tau f\in V$ such that

$$(\tau f, v)_V = \langle f, v\rangle \qquad \forall v\in V, \tag{6.5}$$

which implies that

$$\|\tau f\|_V = \|f\|_{V^\#}.$$

The existence of the Riesz map follows from the Riesz representation theorem; see, e.g., [Ciarlet, 2013, Section 4.6].

Given an inner product $(\cdot,\cdot)_V$, the associated Riesz map can be interpreted as the transformation of the problem (6.4) in the dual space $V^\#$, which is independent of the choice of the inner product on V, into the equation in the (solution) space V,

$$\tau\mathcal{A}\,u = \tau f, \qquad \tau\mathcal{A} : V \to V, \quad u \in V, \quad \tau f \in V. \tag{6.6}$$

The transformation is commonly called *operator preconditioning*. The operator preconditioning typically aims at providing uniform bounds on the condition numbers of the discretized operators (and their algebraic representations) independently on the discretization parameters; see, e.g., Hiptmair [2006] and the discussion and references in [Málek and Strakoš, 2015, Chapters 4 and 8].

Given an element $\tau r \in V$, we can construct the n-th *Krylov subspace* for $\tau\mathcal{A}$ and τr defined as

$$K_n(\tau\mathcal{A}, \tau r) \equiv \operatorname{span}\{\tau r, (\tau\mathcal{A})(\tau r), \ldots, (\tau\mathcal{A})^{n-1}(\tau r)\} \subset V\,. \tag{6.7}$$

There exists a class of methods (called *Krylov subspace methods*) that use Krylov subspaces for generating approximations to the solution u of (6.6). In the following section we describe the conjugate gradient method (CG) on infinite-dimensional Hilbert space V and the matrix formulation of CG corresponding to solving the discrete version of (6.6) on a finite-dimensional subspace $V_h \subset V$.

6.3 Conjugate gradient method in infinite- and finite-dimensional Hilbert spaces

In many physical applications formulated as the problem (6.4) with self-adjoint, bounded and (α-)coercive operator $\mathcal{A}$, the goal is to minimize the energy norm $\|u-u_n\|_a$ of the error $u-u_n$ where $u_n \in V$ is an approximation to the solution u.

Given an initial approximation $u_0 \in V$ and the corresponding residual $r_0 \equiv f - \mathcal{A}u_0 \in V^\#$, the conjugate gradient method provides in the n-th iteration the approximation $u_n \in u_0 + K_n(\tau\mathcal{A}, \tau r_0) \subset V$ such that

$$\|u - u_n\|_a = \min_{v \in u_0 + K_n(\tau\mathcal{A}, \tau r_0)} \|u - v\|_a\,. \tag{6.8}$$

Algorithm 1 gives the standard formulation of the conjugate gradient method in the Hilbert space V (for the derivation, history, and references see, e.g., [Málek and Strakoš, 2015, Section 5.1]).

Consider now a finite-dimensional subspace $V_h \subset V$ and the Galerkin discretization of (6.1) on V_h

$$\text{to find } u_h \in V_h: \quad a(u_h, v_h) = \langle f, v_h\rangle \qquad \forall v_h \in V_h. \tag{6.9}$$

Define the operator $\mathcal{A}_h : V_h \to V_h^\#$ such that

$$\langle \mathcal{A}_h w_h, v_h\rangle = a(w_h, v_h) = \langle \mathcal{A}w_h, v_h\rangle \qquad \forall v_h, w_h \in V_h\,.$$

By restricting the functional f to $V_h^\#$, i.e., defining

$$f_h \in V_h^\# : \langle f_h, v_h\rangle = \langle f, v_h\rangle \qquad \forall v_h \in V_h\,,$$

Algorithm 1 The CG method for solving (6.4)

Given the inner product $(\cdot,\cdot)_V$ and the associated Riesz map $\tau : V^\# \to V$, $u_0 \in V$, $r_0 = f - \mathcal{A}u_0 \in V^\#$, and $p_0 = \tau r_0 \in V$, compute
for $n = 1, \dots, n_{\max}$ **do**

$$\alpha_{n-1} = \frac{\langle r_{n-1}, \tau r_{n-1}\rangle}{\langle \mathcal{A}p_{n-1}, p_{n-1}\rangle} = \frac{(\tau r_{n-1}, \tau r_{n-1})_V}{(\tau \mathcal{A}p_{n-1}, p_{n-1})_V}$$

$u_n = u_{n-1} + \alpha_{n-1}p_{n-1}$ stop when the stopping criterion is satisfied

$r_n = r_{n-1} - \alpha_{n-1}\mathcal{A}p_{n-1}$

$$\beta_n = \frac{\langle r_n, \tau r_n\rangle}{\langle r_{n-1}, \tau r_{n-1}\rangle} = \frac{(\tau r_n, \tau r_n)_V}{(\tau r_{n-1}, \tau r_{n-1})_V}$$

$p_n = \tau r_n + \beta_n p_{n-1}$

end for

the operator form of (6.9) reads

$$\mathcal{A}_h u_h = f_h, \qquad u_h \in V_h, \quad f_h \in V_h^\# . \tag{6.10}$$

Let $\dim(V_h) = N$ and $\Phi = \{\phi_1, \dots, \phi_N\}$ be a given basis of V_h. Each element $u_h \in V_h$ can be represented by its coordinate vector $\mathbf{u} \in \mathbb{R}^N$ in the basis Φ,

$$u_h = \Phi \mathbf{u} .$$

Consider further the (canonical) dual basis $\Phi^\# = \{\phi_1^\#, \dots, \phi_N^\#\}$ of $V_h^\#$ associated with Φ, i.e.

$$\langle \phi_i^\#, \phi_j\rangle = \begin{cases} 1, & \text{if } i = j\,, \\ 0, & \text{otherwise.} \end{cases}$$

Written symbolically, $\langle \Phi^\#, \Phi\rangle = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We now show the representation in $\mathbb{R}^N$ of the duality pairing $\langle\cdot,\cdot\rangle$, the inner product $(\cdot,\cdot)_V$, and the operator $\mathcal{A}_h$. For any $f = \Phi^\#\mathbf{f} \in V_h^\#$ and $u = \Phi\mathbf{u} \in V_h$, $v = \Phi\mathbf{v} \in V_h$

$$\begin{aligned} \langle f, v\rangle &= \langle \Phi^\#\mathbf{f}, \Phi\mathbf{v}\rangle = \mathbf{v}^*\mathbf{f}, \\ (u, v)_V &= (\Phi\mathbf{u}, \Phi\mathbf{v})_V = \mathbf{v}^*\mathbf{M}\mathbf{u}, \qquad (\mathbf{M})_{ij} = (\phi_j, \phi_i)_V, \\ \mathcal{A}_h u &= \mathcal{A}_h\Phi\mathbf{u} = \Phi^\#\mathbf{A}\mathbf{u}, \qquad (\mathbf{A})_{ij} = \langle \mathcal{A}_h\phi_j, \phi_i\rangle, \quad i,j = 1,\dots,N. \end{aligned} \tag{6.11}$$

The matrix representation of the Riesz map τ is

$$\tau\Phi^\# = \Phi\mathbf{M}^{-1} .$$

Indeed,

$$\mathbf{v}^*\mathbf{f} = \langle f, v\rangle = (\tau f, v)_V = (\tau\Phi^\#\mathbf{f}, \Phi\mathbf{v})_V = (\Phi\mathbf{M}^{-1}\mathbf{f}, \Phi\mathbf{v})_V = \mathbf{v}^*\mathbf{M}\mathbf{M}^{-1}\mathbf{f} .$$

The energy norm (6.3) satisfies, for $v = \Phi\mathbf{v} \in V_h$,

$$\|v\|_a^2 = \langle \mathcal{A}_h v, v\rangle = \langle \mathcal{A}_h\Phi\mathbf{v}, \Phi\mathbf{v}\rangle = \mathbf{v}^*\mathbf{A}\mathbf{v} = \|\mathbf{v}\|_{\mathbf{A}}^2 .$$

Using the above representations, we can rewrite the (finite-dimensional) problem (6.10) as the linear algebraic system

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \qquad \mathbf{A} \in \mathbb{R}^{N\times N}, \quad \mathbf{u}, \mathbf{f} \in \mathbb{R}^N. \tag{6.12}$$

Reformulating Algorithm 1 as an algebraic representation of CG for solving (6.12) gives a standard form of the preconditioned conjugate gradient (PCG) algorithm with the preconditioner $\mathbf{M}$ given in (6.11); see Algorithm 2. The approximation $\mathbf{u}_n$ given in the n-th step of PCG satisfies

$$\|\mathbf{u} - \mathbf{u}_n\|_{\mathbf{A}} = \underset{\mathbf{v}\in\mathbf{u}_0+K_n(\mathbf{M}^{-1}\mathbf{A},\mathbf{M}^{-1}\mathbf{r_0})}{\arg\min} \|\mathbf{u} - \mathbf{v}\|_{\mathbf{A}}\,; \tag{6.13}$$

cf. (6.8). The associated Krylov subspace is given by

$$K_n(\mathbf{M}^{-1}\mathbf{A}, \mathbf{M}^{-1}\mathbf{r_0}) = \operatorname{span}\{\mathbf{M}^{-1}\mathbf{r_0}, \mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-1}\mathbf{r_0}, \dots, (\mathbf{M}^{-1}\mathbf{A})^{n-1}(\mathbf{M}^{-1}\mathbf{r_0})\}\,,$$

analogously to (6.7).

Algorithm 2 The PCG method for solving (6.12)

Given an SPD preconditioner $\mathbf{M}$,
$\mathbf{u}_0$, $\mathbf{r}_0 = \mathbf{f} - \mathbf{A}\mathbf{u}_0$, solve $\mathbf{M}\mathbf{z}_0 = \mathbf{r}_0$, set $\mathbf{p}_0 = \mathbf{z}_0$, and compute
for $n = 1, \dots, n_{\max}$ **do**

$$\alpha_{n-1} = \frac{\mathbf{z}_{n-1}^*\mathbf{r}_{n-1}}{\mathbf{p}_{n-1}^*\mathbf{A}\mathbf{p}_{n-1}}$$
$$\mathbf{u}_n = \mathbf{u}_{n-1} + \alpha_{n-1}\mathbf{p}_{n-1}$$
$$\mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_{n-1}\mathbf{A}\mathbf{p}_{n-1}$$
solve $\mathbf{M}\mathbf{z}_n = \mathbf{r}_n$
$$\beta_n = \frac{\mathbf{z}_n^*\mathbf{r}_n}{\mathbf{z}_{n-1}^*\mathbf{r}_{n-1}}$$
$$\mathbf{p}_n = \mathbf{z}_n + \beta_n\mathbf{p}_{n-1}$$

stop when the stopping criterion is satisfied

end for

To sum up, the solution process using the operator preconditioning (described via the Riesz map τ) in the conjugate gradient method can be illustrated by the scheme

$$\{\mathcal{A}, f, \tau\} \to \{\mathcal{A}_h, f_h, \tau\} \to \left\{\mathbf{A}, \mathbf{f}, \mathbf{M}^{-1}\right\} \to \text{Algorithm 2.}$$

The state-of-the-art literature on using PCG in numerical solution of PDEs proceeds, however, in most of the cases in the following way. First, the algebraic system $\mathbf{A}\mathbf{u} = \mathbf{f}$ is formed by some form of discretization independently of the algebraic method that is then used for its numerical solution. Since the standard unpreconditioned CG, i.e. Algorithm 2 with $\mathbf{M} = \mathbf{I}$, results in most of the cases in a slow convergence of the computed approximation $\mathbf{u}_n$ to the algebraic solution $\mathbf{u}$, preconditioning of the algebraic problem is introduced to accelerate the convergence behavior. Such view separates the preconditioning from the discretization step of the solution process and can be illustrated by the scheme

$$\{\mathcal{A}, f\} \to \{\mathbf{A}, \mathbf{f}\} \to \text{construction of an preconditioner } \widehat{\mathbf{M}} \to \text{Algorithm 2.}$$

6.4 Algebraic preconditioning as transformation of the discretization basis

For a given symmetric positive definite preconditioner $\widehat{\mathbf{M}}$ and its Cholesky decomposition $\widehat{\mathbf{M}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^*$, the common algebraic view on PCG (see, e.g., [Saad, 2003, Section 9.2]) consists in applying the unpreconditioned conjugate gradient method to the transformed *preconditioned system*

$$\underbrace{(\widehat{\mathbf{L}}^{-1}\mathbf{A}(\widehat{\mathbf{L}}^*)^{-1})}_{\mathbf{A}_t}\underbrace{(\widehat{\mathbf{L}}^*\mathbf{u})}_{\mathbf{u}^t} = \underbrace{(\widehat{\mathbf{L}}^{-1}\mathbf{f})}_{\mathbf{f}^t}\,. \tag{6.14}$$

In other words, the preconditioned conjugate gradient method (Algorithm 2) for solving $\mathbf{A}\mathbf{u} = \mathbf{f}$ with the preconditioner $\widehat{\mathbf{M}}$ is regarded[1] as the unpreconditioned conjugate gradient method (Algorithm 2 with setting $\mathbf{M} = \mathbf{I}$) for solving the transformed system $\mathbf{A}_t\mathbf{u}^t = \mathbf{f}^t$. In the following, we show the relationship between algebraic preconditioning and transformation of the discretization basis.

First, consider the PCG with the preconditioner $\mathbf{M}$ as a matrix formulation of the infinite-dimensional CG discretized with the basis Φ, i.e. $\mathbf{M}$ given by (6.11). In this setting, $\mathbf{M} = \mathbf{I}$ means that the discretization basis Φ is orthogonal with respect to the given inner product $(\cdot,\cdot)_V$. This situation is certainly uncommon and we will therefore consider a transformation (i.e. the orthogonalization) of the discretization basis such that this condition is satisfied. Using the fact that $\mathbf{M} = (\Phi,\Phi)_V$ is the Gram matrix of the basis Φ and considering its Cholesky decomposition $\mathbf{M} = \mathbf{L}\mathbf{L}^*$, the orthogonalization coefficients are given in the columns of the (upper triangular) matrix $(\mathbf{L}^*)^{-1}$; see, e.g., [Liesen and Strakoš, 2013, Section 3.6]. Denoting the transformed basis by

$$\Phi_t \equiv \Phi(\mathbf{L}^*)^{-1}$$

we indeed have, writing symbolically,

$$(\Phi_t,\Phi_t)_V = (\Phi(\mathbf{L}^*)^{-1},\Phi(\mathbf{L}^*)^{-1})_V = \mathbf{L}^{-1}(\Phi,\Phi)_V(\mathbf{L}^*)^{-1} = \mathbf{L}^{-1}\mathbf{M}(\mathbf{L}^*)^{-1} = \mathbf{I}.$$

The transformed canonical dual basis associated with Φ_t is

$$\Phi_t^{\#} \equiv \Phi^{\#}\mathbf{L}.$$

Indeed,

$$\langle\Phi_t^{\#},\Phi_t\rangle = \langle\Phi^{\#}\mathbf{L},\Phi(\mathbf{L}^*)^{-1}\rangle = \mathbf{L}^{-1}\langle\Phi^{\#},\Phi\rangle\mathbf{L} = \mathbf{L}^{-1}\mathbf{I}\,\mathbf{L} = \mathbf{I}.$$

Then

$$\begin{aligned}
f_h &= \Phi^{\#}\mathbf{f} = \Phi_t^{\#}\mathbf{L}^{-1}\mathbf{f} = \Phi_t^{\#}\mathbf{f}^t, & \mathbf{f}^t &= \mathbf{L}^{-1}\mathbf{f},\\
u_h &= \Phi\mathbf{u} = \Phi_t\mathbf{L}^*\mathbf{u} = \Phi_t\mathbf{u}^t, & \mathbf{u}^t &= \mathbf{L}^*\mathbf{u},\\
r_n &= \Phi^{\#}\mathbf{r}_n = \Phi_t^{\#}\mathbf{r}_n^t, & \mathbf{r}_n^t &= \mathbf{L}^{-1}\mathbf{r}_n,\\
p_n &= \Phi\mathbf{p}_n = \Phi_t\mathbf{p}_n^t, & \mathbf{p}_n^t &= \mathbf{L}^*\mathbf{p}_n,\\
u_n &= \Phi\mathbf{u}_n = \Phi_t\mathbf{u}_n^t, & \mathbf{u}_n^t &= \mathbf{L}^*\mathbf{u}_n,
\end{aligned}$$

[1] PCG allows various implementations with possibly different behavior in finite precision computations. Discussion on finite precision behavior is, however, beyond the scope of the thesis.

and

$$\begin{aligned}
\tau f = \tau\Phi^{\#}\mathbf{f} &= \tau\Phi_t^{\#}\mathbf{f}^t \\
&= \Phi\mathbf{M}^{-1}\mathbf{f} = \Phi_t\mathbf{L}^*\mathbf{M}^{-1}\mathbf{L}\mathbf{f}^t = \Phi_t\mathbf{f}^t, \qquad \mathbf{M}_t = \mathbf{I}, \\
\mathcal{A}_h u = \mathcal{A}_h\Phi\mathbf{u} &= \mathcal{A}_h\Phi_t\mathbf{u}^t \\
&= \Phi^{\#}\mathbf{A}\mathbf{u} = \Phi_t^{\#}\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\mathbf{u}^t \\
&= \Phi_t^{\#}\mathbf{A}_t\mathbf{u}^t, \qquad\qquad \mathbf{A}_t = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\,.
\end{aligned}$$

Therefore, the matrix representation of Algorithm 1 obtained using the transformed (orthonormal) discretization basis Φ_t applied to the finite-dimensional problem (6.10) gives the unpreconditioned algebraic CG for solving the preconditioned system

$$\mathbf{A}_t\mathbf{u}^t = \mathbf{f}^t, \qquad \text{i.e.} \qquad \left(\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\right)(\mathbf{L}^*\mathbf{u}) = \mathbf{L}^{-1}\mathbf{f}.$$

Observation: Orthogonalization of the discretization basis in the given finite dimensional Hilbert space is equivalent to the algebraic preconditioning of the linear algebraic system associated with the operator preconditioning.

The condition number $\kappa(\mathbf{A}_t) = \|\mathbf{A}_t\|\,\|\mathbf{A}_t^{-1}\|$ of the matrix $\mathbf{A}_t = \mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}$ associated with $\mathbf{M} = \mathbf{L}\mathbf{L}^*$ given by (6.11) satisfies (see, e.g., Hiptmair [2006], [Málek and Strakoš, 2015, Chapter 8])

$$\begin{aligned}
\kappa(\mathbf{A}_t) &= \left(\max_{\mathbf{v}\neq 0}\frac{\mathbf{v}^*\mathbf{A}_t\mathbf{v}}{\mathbf{v}^*\mathbf{v}}\right)\left(\min_{\mathbf{w}\neq 0}\frac{\mathbf{w}^*\mathbf{A}_t\mathbf{w}}{\mathbf{w}^*\mathbf{w}}\right)^{-1} \\
&= \left(\max_{\mathbf{v}\neq 0}\frac{\mathbf{v}^*\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\mathbf{v}}{\mathbf{v}^*\mathbf{L}^{-1}(\mathbf{L}\mathbf{L}^*)(\mathbf{L}^*)^{-1}\mathbf{v}}\right)\left(\min_{\mathbf{w}\neq 0}\frac{\mathbf{w}^*\mathbf{L}^{-1}\mathbf{A}(\mathbf{L}^*)^{-1}\mathbf{w}}{\mathbf{w}^*\mathbf{L}^{-1}(\mathbf{L}\mathbf{L}^*)(\mathbf{L}^*)^{-1}\mathbf{w}}\right)^{-1} \\
&= \left(\max_{\widetilde{\mathbf{v}}\neq 0}\frac{\widetilde{\mathbf{v}}^*\mathbf{A}\widetilde{\mathbf{v}}}{\widetilde{\mathbf{v}}^*\mathbf{M}\widetilde{\mathbf{v}}}\right)\left(\min_{\widetilde{\mathbf{w}}\neq 0}\frac{\widetilde{\mathbf{w}}^*\mathbf{A}\widetilde{\mathbf{w}}}{\widetilde{\mathbf{w}}^*\mathbf{M}\widetilde{\mathbf{w}}}\right)^{-1} \\
&= \left(\max_{v_h\in V_h}\frac{a(v_h,v_h)}{(v_h,v_h)_V}\right)\left(\min_{w_h\in V_h}\frac{a(w_h,w_h)}{(w_h,w_h)_V}\right)^{-1} \\
&\leq \left(\sup_{v\in V}\frac{a(v,v)}{(v,v)_V}\right)\left(\inf_{w\in V}\frac{a(w,w)}{(w,w)_V}\right)^{-1} \qquad (6.15) \\
&\leq \frac{C}{\alpha}\,,
\end{aligned}$$

with the constants C, α from (6.2). This bound is independent of V_h and its basis Φ.

In contrast to the previous development, the algebraic PCG with

$$\widehat{\mathbf{M}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^* \neq \mathbf{M}$$

does not result from the discretization of the infinite-dimensional CG with the (transformed) discretization bases

$$\widehat{\Phi} = \Phi(\widehat{\mathbf{L}}^*)^{-1}, \qquad \widehat{\Phi}^{\#} = \Phi^{\#}\widehat{\mathbf{L}}. \qquad (6.16)$$

Since $\widehat{\mathbf{M}} \neq \mathbf{M}$, the basis $\widehat{\Phi}$ is not orthonormal with respect to the inner product $(\cdot,\cdot)_V$,

$$(\widehat{\Phi},\widehat{\Phi})_V = (\Phi(\widehat{\mathbf{L}}^*)^{-1}, \Phi(\widehat{\mathbf{L}}^*)^{-1})_V = \widehat{\mathbf{L}}^{-1}\mathbf{M}(\widehat{\mathbf{L}}^*)^{-1} \neq \mathbf{I}\,.$$

In order to interpret the algebraic preconditioner $\widehat{\mathbf{M}}$ as transformation (orthonormalization) of the basis $\Phi \mapsto \widehat{\Phi} = \Phi(\widehat{\mathbf{L}}^*)^{-1}$, we have to *change* also *the inner product*. The inner product $(\cdot,\cdot)_{V_h}$ in V_h giving for $u = \Phi\mathbf{u}$, $v = \Phi\mathbf{v}$

$$(u,v)_{V_h} = (u,v)_V = (\Phi\mathbf{u},\Phi\mathbf{v})_V = \mathbf{v}^*\mathbf{M}\mathbf{u},$$

has to be replaced by the inner product $(\cdot,\cdot)_{\mathrm{new},V_h}$ defined by the SPD matrix $\widehat{\mathbf{M}}$, giving with $u = \Phi\mathbf{u} = \widehat{\Phi}\widehat{\mathbf{u}}$, $\widehat{\mathbf{u}} = \widehat{\mathbf{L}}^*\mathbf{u}$, and $v = \Phi\mathbf{v} = \widehat{\Phi}\widehat{\mathbf{v}}$, $\widehat{\mathbf{v}} = \widehat{\mathbf{L}}^*\mathbf{v}$,

$$(u,v)_{\mathrm{new},V_h} = (\widehat{\Phi}\widehat{\mathbf{u}},\widehat{\Phi}\widehat{\mathbf{v}})_{\mathrm{new},V_h} \equiv \widehat{\mathbf{v}}^*\widehat{\mathbf{u}} = \mathbf{v}^*\widehat{\mathbf{L}}\widehat{\mathbf{L}}^*\mathbf{u} = \mathbf{v}^*\widehat{\mathbf{M}}\mathbf{u}.$$

Then $(\widehat{\Phi},\widehat{\Phi})_{\mathrm{new},V_h} = \mathbf{I}$, and the matrix representation of the Riesz map $\widehat{\tau}$ defined by the transformed inner product $(\cdot,\cdot)_{\mathrm{new},V_h}$ is given by

$$\widehat{\tau}\widehat{\Phi}^{\#} = \widehat{\Phi}\widehat{\mathbf{M}}_{\widehat{\tau}}^{-1}$$

with $\widehat{\mathbf{M}}_{\widehat{\tau}} = \mathbf{I}$. Indeed, for $f = \widehat{\Phi}^{\#}\widehat{\mathbf{f}}$, $v = \widehat{\Phi}\widehat{\mathbf{v}}$,

$$\begin{aligned}\widehat{\mathbf{v}}^*\widehat{\mathbf{f}} &= \langle f, v\rangle \\ &= (\widehat{\tau}f, v)_{\mathrm{new},V_h} = (\widehat{\tau}\widehat{\Phi}^{\#}\widehat{\mathbf{f}}, \widehat{\Phi}\widehat{\mathbf{v}})_{\mathrm{new},V_h} = (\widehat{\Phi}\widehat{\mathbf{M}}_{\widehat{\tau}}^{-1}\widehat{\mathbf{f}}, \widehat{\Phi}\widehat{\mathbf{v}})_{\mathrm{new},V_h} = \widehat{\mathbf{v}}^*\widehat{\mathbf{M}}_{\widehat{\tau}}^{-1}\widehat{\mathbf{f}}.\end{aligned}$$

Showing that

$$\mathcal{A}_h u = \widehat{\Phi}^{\#}\widehat{\mathbf{A}}\widehat{\mathbf{u}}, \qquad \widehat{\mathbf{A}} = \widehat{\mathbf{L}}^{-1}\mathbf{A}(\widehat{\mathbf{L}}^*)^{-1}, \quad \widehat{\mathbf{u}} = \widehat{\mathbf{L}}^*\mathbf{u},$$

i.e. that the transformed algebraic system has the form (6.14), is analogous to the previous case with the preconditioner $\mathbf{M}$.

Observation: Algebraic PCG with arbitrary algebraic preconditioning can be interpreted as the matrix formulation of the infinite-dimensional CG discretized with the transformed basis $\widehat{\Phi}$ and, at the same time, using the transformed inner product in V_h. The transformation is such that $\widehat{\Phi}$ is orthonormal with respect to the transformed inner product.

6.5 Numerical illustrations

We now present simple numerical experiments to illustrate the results stated above. For the illustrations we consider **Inhomogeneous tensor problems I** and **II** described in Section 2.2 (see also the references therein) discretized using the piecewise affine finite elements with the basis given by hat-functions, i.e. the piecewise affine functions such that each one corresponds to a node of the mesh taking there value 1 and vanishing in all others. We consider uniform meshes consisting of isosceles right-angled elements and the adaptively refined meshes that are generated by the adaptive procedure described in Section 2.2.

For the considered test problems,

$$V \equiv H_0^1(\Omega), \qquad \Omega \equiv (-1,1)\times(-1,1)\,, \qquad a(u,v) \equiv \int_\Omega \mathbf{S}\nabla u \cdot \nabla v\,,$$

where $\mathbf{S} = s(x)\mathbf{I}$, $x \in \Omega$, is a piecewise constant positive multiple of the identity matrix. We consider the operator preconditioner (see Sections 6.2 and 6.3)

laplace with $(u,v)_V \equiv \int_\Omega \nabla u \cdot \nabla v\,,$

and two algebraic preconditioners

ichol0 using the incomplete Cholesky decomposition of the stiffness matrix with zero fill-in,

ichol(TOL) using the incomplete Cholesky decomposition with threshold dropping, we set the threshold tolerance to 10^{-2};

see, e.g., [Saad, 2003, Chapter 10]. The **ichol0** and **ichol(TOL)** use the Matlab `ichol` command. We will use the notation $\widehat{\mathbf{M}}_{\text{laplace}}$, $\widehat{\mathbf{M}}_{\text{ichol0}}$, and $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ respectively.

Given a discretization basis, **laplace** preconditioning provides the Gram matrix $\widehat{\mathbf{M}}_{\text{laplace}}$ through (6.11) with the (complete) Cholesky decomposition $\widehat{\mathbf{M}}_{\text{laplace}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^*$. In contrast, in **ichol0** and **ichol(TOL)** the lower triangular matrix $\widehat{\mathbf{L}}$ is constructed using the incomplete Cholesky decomposition of the stiffness matrix $\mathbf{A}$ with $\widehat{\mathbf{M}}$ given as $\widehat{\mathbf{M}} \equiv \widehat{\mathbf{L}}\widehat{\mathbf{L}}^*$ (in practical computations $\widehat{\mathbf{M}}$ is not assembled).

The presented experiments involve various computations, which are all subject to numerical errors. In the illustrations below, the round-off errors in inverting matrices, Cholesky decomposition, evaluation of norms and condition numbers are not substantial. For the ease of exposition we will therefore identify the exact results with their computed counterparts.

6.5.1 Convergence of PCG and conditioning of the transformed algebraic problem

The goal of preconditioning is often identified with decreasing the condition number of the transformed matrix. This is misleading, unless the condition number becomes really small, because convergence behavior of CG is not governed by the condition number of the system matrix; see, e.g., the recent summary in Gergelits and Strakoš [2014]. Figures 6.1 and 6.2 depict the energy norm $\|u_h - u_n\|_a$ of the algebraic error of the unpreconditioned CG and PCG with the preconditioners given above. Table 6.1 gives the condition numbers of the stiffness matrix $\mathbf{A}$ and of the transformed (preconditioned) matrices $\mathbf{A}_t$ given by (6.14).

Applying the **laplace** preconditioning is relatively costly. In the setting considered in this chapter, each PCG iteration requires solving[2] the problem with the matrix $\widehat{\mathbf{M}}_{\text{laplace}}$ that is of the same size as the stiffness matrix $\mathbf{A}$. In practice,

[2]Using the Cholesky decomposition $\widehat{\mathbf{M}}_{\text{laplace}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^*$, each iteration of PCG involves one forward and one backward substitution with $\widehat{\mathbf{L}}$ and $\widehat{\mathbf{L}}^*$; see, e.g., [Saad, 2003, Algorithm 9.2].

operator preconditioning is often implemented using an appropriate solver for the Laplace problem, e.g., the multigrid method. As operator-based preconditioning, **laplace** provides a transformed stiffness matrix with the condition number that is bounded independently of the discretization parameters; see (6.15). In the considered test problems with the diffusion tensor $\mathbf{S} = s(x)\mathbf{I}$, $x \in \Omega$, the upper bound reads

$$\begin{aligned}\kappa(\mathbf{A}_t) &\leq \left(\sup_{v\in V} \frac{a(v,v)}{(v,v)_V}\right)\left(\inf_{w\in V} \frac{a(w,w)}{(w,w)_V}\right)^{-1} \\ &= \left(\sup_{v\in V} \frac{\int_\Omega \mathbf{S}\nabla v\cdot\nabla v}{\int_\Omega \nabla v\cdot\nabla v}\right)\left(\inf_{w\in V} \frac{\int_\Omega \mathbf{S}\nabla w\cdot\nabla w}{\int_\Omega \nabla w\cdot\nabla w}\right)^{-1} \\ &\leq \frac{\max_{x\in\Omega} s(x)}{\min_{x\in\Omega} s(x)},\end{aligned}$$

which is equal to 5 and 1.61×10^2, respectively; cf. Table 6.1. Regarding convergence of PCG, the **laplace** preconditioner outperforms the other two considered preconditioners in most of the iterations.

As expected, **ichol(TOL)** is superior to **ichol0** in terms of PCG convergence and the conditioning of the transformed system. Note that for the second test problem discretized using the uniform mesh, **ichol(TOL)** provides problem with an order of magnitude smaller condition number than **laplace** preconditioning. However, the energy norm of the algebraic error decreases substantially faster for the **laplace** preconditioner.

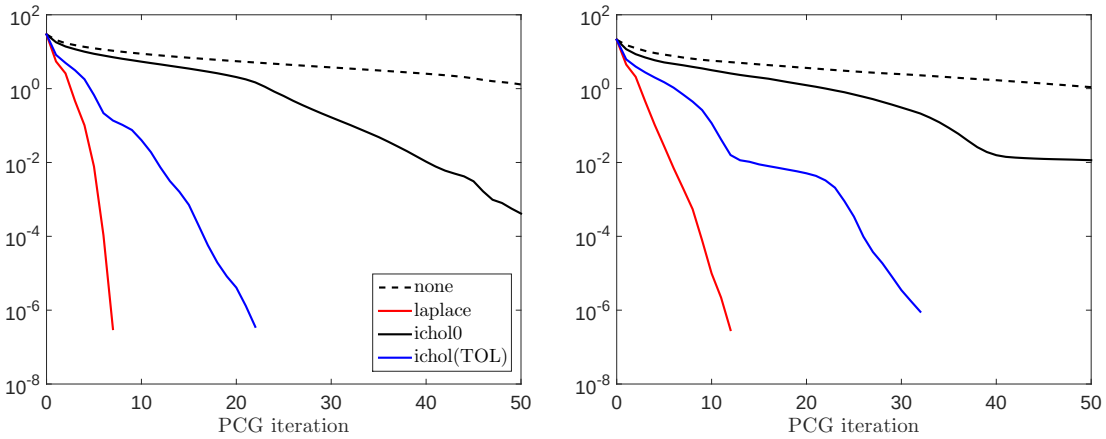


Figure 6.1: Inhomogeneous tensor problem I: convergence of the energy norm of the algebraic error in PCG when the problem is discretized using the uniform (left) and the adaptively refined mesh (right). The algebraic systems are of the size 3969 and 3355, respectively.

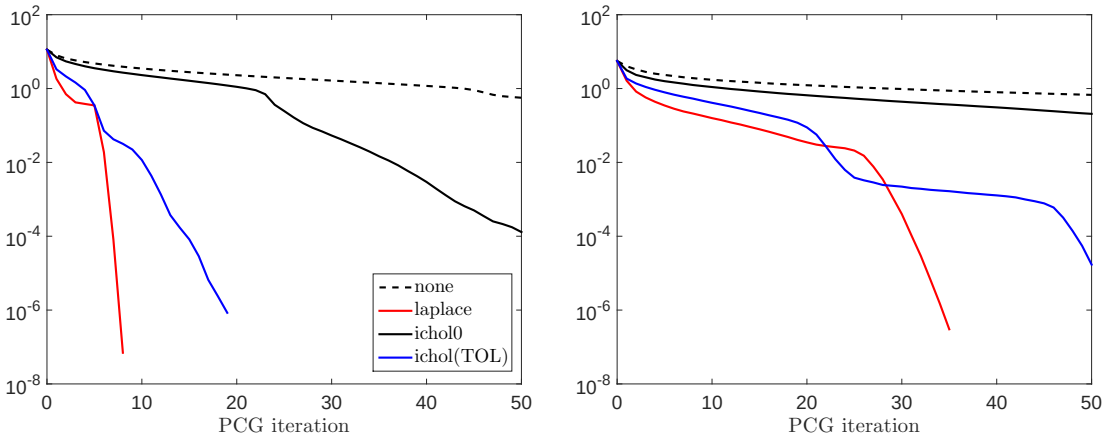


Figure 6.2: Inhomogeneous tensor problem II: convergence of the energy norm of the algebraic error in PCG when the problem is discretized using the uniform (left) and the adaptively refined mesh (right). The algebraic systems are of the size 3969 and 4340, respectively.

problem:	Inhomogeneous tensor I		Inhomogeneous tensor II	
mesh:	uniform	adaptive	uniform	adaptive
size:	3969	3355	3969	4340
none	3.34×10^3	9.22×10^3	6.75×10^4	7.00×10^4
laplace	5	5	1.61×10^2	1.61×10^2
ichol0	4.25×10^2	1.14×10^3	4.31×10^2	6.59×10^3
ichol(TOL)	1.69×10^1	1.19×10^2	1.57×10^1	5.00×10^2

Table 6.1: Condition number $\kappa(\mathbf{A}_t) = \|\mathbf{A}_t\| \, \|\mathbf{A}_t^{-1}\|$ of the preconditioned matrix $\mathbf{A}_t$ given by (6.14) for the chosen preconditioners.

6.5.2 Orthogonalization of the basis and ordering of the degrees of freedom

In Section 6.4 we interpreted algebraic preconditioning as orthogonalization of the discretization basis with respect to the inner product determined by the given preconditioner. The properties such as the support of the resulting orthogonal basis functions depend on the inner product and also on the order in which the basis functions are orthogonalized. We demonstrate this in the following experiments.

We first describe the orderings of degrees of freedom used in the numerical illustrations. Then we discuss the effect of reordering. Finally, for the given preconditioners, we show the sparsity patterns of their Cholesky factors and of the upper triangular matrices that determine the associated transformation of the discretization basis.

The effect of the ordering of the degrees of freedom on convergence of preconditioned conjugate gradient method with incomplete Cholesky decomposition was demonstrated, e.g., in the seminal paper Duff and Meurant [1989], and there is a number of papers seeking an ordering that results in faster PCG convergence; see, e.g., Benzi and Tůma [2000]. Consequently, techniques to reorder degrees of freedom represent an inseparable part of implementations of the PCG.

The order, in which the functions are orthogonalized, can also affect the loss of orthogonality in an orthogonalization process due to round-off. This was a subject of study in so-called Ordered Gram–Schmidt orthogonalization; see, e.g., Štuller [1995]. However, there are very few theoretical results regarding this issue.

Orderings used in numerical illustrations

In the ordering of the degrees of freedom used in the previous subsection (and also in the numerical experiments of Chapters 2 and 4), the new nodes constructed in the mesh refinement steps are ordered progressively; see the illustration in Figure 6.3. We recall that each node corresponds to one degree of freedom.

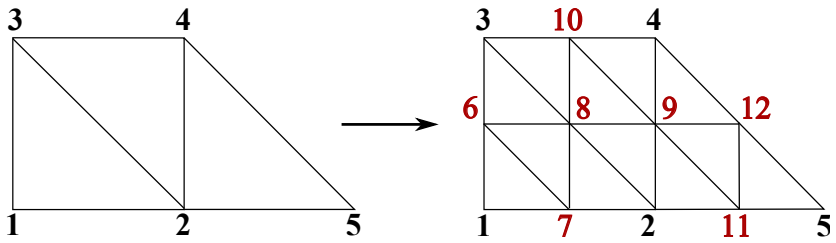


Figure 6.3: Illustration of the ordering of the nodes in the mesh refinement.

For the illustrations we use also two other orderings. Their choice is motivated by an (expected) different sparsity of the upper triangular matrix representing the transformation (orthogonalization) of the discretization basis in the **laplace** preconditioning; see the discussion below. The same orderings are used also in the experiments with **ichol0** and **ichol(TOL)** preconditioners.

We consider the ordering of the nodes of the mesh such that jth node neighbors with the $(j-1)$st node, i.e. they share a common edge in the mesh; see

the illustration in Figure 6.4. Then, in **laplace** preconditioning, ϕ_j must be orthogonalized against all the basis functions $\phi_{j-1}, \ldots, \phi_1$ and the upper triangular matrix $(\widehat{\mathbf{L}}^*)^{-1}$ representing the corresponding transformation (orthogonalization) of the discretization basis (see (6.16)) is full. We will call this ordering **zigzag**.

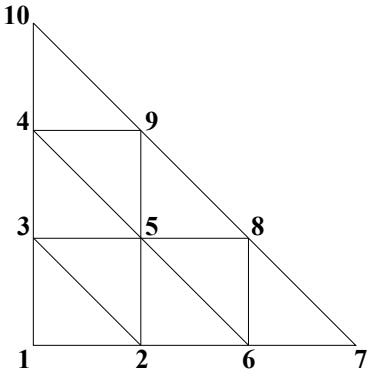


Figure 6.4: Illustration of the **zigzag** ordering of the nodes.

The next ordering is based on a purely algebraic construction. First, the preconditioner $\widehat{\mathbf{M}}_{\text{laplace}}$ is assembled for the original ordering described in the first paragraph. For $\widehat{\mathbf{M}}_{\text{laplace}}$ we construct the permutation matrix[3] $\mathbf{P}$ by the symmetric approximate minimum degree permutation (Amestoy et al. [1996], implemented in Matlab `symamd` command) and change the original ordering of the degrees of freedom accordingly. The permutation aims to provide the matrix $\mathbf{P}\widehat{\mathbf{M}}_{\text{laplace}}\mathbf{P}^T$ that has sparser Cholesky factor than $\widehat{\mathbf{M}}_{\text{laplace}}$. This was observed in our numerical experiments. We also observe that the inverse of the Cholesky factor of $\mathbf{P}\widehat{\mathbf{M}}_{\text{laplace}}\mathbf{P}^T$ is sparser than the inverse of the Cholesky factor of $\widehat{\mathbf{M}}_{\text{laplace}}$, which is an expected behavior; see, e.g., Benzi and Tůma [2000]. We will call this ordering **symamd**. Because of the specific algebraic construction that employs multiple elimination steps similarly as described in detail in Liu [1985], **symamd** is very close to the class of hierarchical reorderings.

Remark: When two bases Φ and $\Phi\mathbf{P}^T$, $V_h = \text{span}\{\Phi\} = \text{span}\{\Phi\mathbf{P}^T\}$, are used for the matrix formulation of the problem (6.10), the corresponding stiffness matrices and **laplace** preconditioners are $\mathbf{A}$ and $\widehat{\mathbf{M}}_{\text{laplace}}$, respectively $\mathbf{P}\mathbf{A}\mathbf{P}^T$ and $\mathbf{P}\widehat{\mathbf{M}}_{\text{laplace}}\mathbf{P}^T$; see (6.11). Let $\widehat{\mathbf{M}}_{\text{ichol0}}$ and $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ denote the preconditioners constructed in **ichol0** and **ichol(TOL)** for $\mathbf{A}$. Then, apart from special cases, the matrices $\mathbf{P}\widehat{\mathbf{M}}_{\text{ichol0}}\mathbf{P}^T$ and $\mathbf{P}\widehat{\mathbf{M}}_{\text{ichol(TOL)}}\mathbf{P}^T$ differ from the **ichol0** and **ichol(TOL)** preconditioners constructed for $\mathbf{P}\mathbf{A}\mathbf{P}^T$.

Convergence of PCG for various orderings

We now discuss and illustrate the effect of reordering of the degrees of freedom on convergence of PCG with the considered preconditioners. The discussion elaborates on Sections 6.1 to 6.4. In the considered model problems, round-off errors

[3] A permutation matrix is a square matrix that has exactly one entry of 1 in each row and each column and zeros elsewhere. Clearly, a permutation matrix $\mathbf{P}$ is an orthogonal matrix, $\mathbf{P}^T\mathbf{P} = \mathbf{I}$.

do not substantially affect convergence behavior. Therefore the discussion of observed behavior does not need to consider effects of round-off errors that can otherwise be very substantial. Results derived assuming exact arithmetic are not, in general, descriptive for finite precision behavior of PCG; see, e.g., Greenbaum [1989], [Meurant and Strakoš, 2006, Section 5], [Liesen and Strakoš, 2013, Section 5.9], and Gergelits and Strakoš [2014].

First, we show that (assuming exact arithmetic) reordering of the degrees of freedom does not affect convergence of the unpreconditioned CG. For that purpose we recall the minimization property of the CG method. The n-th step of CG for solving $\mathbf{A}\mathbf{u} = \mathbf{f}$ with the initial vector $\mathbf{u}_0$ (in our experiments $\mathbf{u}_0 \equiv 0$) and the corresponding residual $\mathbf{r}_0 = \mathbf{f} - \mathbf{A}\mathbf{u}_0$ provides the approximation $\mathbf{u}_n$ such that

$$\mathbf{u}_n = \underset{\mathbf{v} \in \mathbf{u}_0 + K_n(\mathbf{A}, \mathbf{r}_0)}{\arg\min} \|\mathbf{u} - \mathbf{v}\|_{\mathbf{A}} ;$$

see (6.13) and, e.g., [Liesen and Strakoš, 2013, Theorem 2.3.1]. Now let the reordering of the degrees of freedom be represented by a permutation matrix $\mathbf{P}$. Then $\mathbf{P}\mathbf{A}\mathbf{P}^T$ is the stiffness matrix corresponding to the reordered degrees of freedom, $\mathbf{P}\mathbf{f}$ is the permuted right-hand side, and $\widetilde{\mathbf{u}} \equiv (\mathbf{P}\mathbf{A}\mathbf{P}^T)^{-1}(\mathbf{P}\mathbf{f}) = \mathbf{P}\mathbf{u}$. For the n-th Krylov subspace there holds,

$$\begin{aligned} K_n(\mathbf{P}\mathbf{A}\mathbf{P}^T, \mathbf{P}\mathbf{r}_0) &= \operatorname{span}\{\mathbf{P}\mathbf{r}_0, (\mathbf{P}\mathbf{A}\mathbf{P}^T)\mathbf{P}\mathbf{r}_0, \ldots, (\mathbf{P}\mathbf{A}\mathbf{P}^T)^{n-1}\mathbf{P}\mathbf{r}_0\} \\ &= \operatorname{span}\{\mathbf{P}\mathbf{r}_0, \mathbf{P}\mathbf{A}\mathbf{r}_0, \ldots, \mathbf{P}\mathbf{A}^{n-1}\mathbf{r}_0\} \\ &= \{\mathbf{P}\mathbf{v} \mid \mathbf{v} \in K_n(\mathbf{A}, \mathbf{r}_0)\}. \end{aligned}$$

For the ease of notation, we denote

$$\mathbf{P}K_n(\mathbf{A}, \mathbf{r}_0) \equiv \{\mathbf{P}\mathbf{v} \mid \mathbf{v} \in K_n(\mathbf{A}, \mathbf{r}_0)\}.$$

A simple algebraic manipulation shows that the n-th CG approximation $\widetilde{\mathbf{u}}_n$ for solving $(\mathbf{P}\mathbf{A}\mathbf{P}^T)\widetilde{\mathbf{u}} = \mathbf{P}\mathbf{f}$ with the initial approximation $\mathbf{P}\mathbf{u}_0$ satisfies

$$\begin{aligned} \widetilde{\mathbf{u}}_n &= \underset{\mathbf{v} \in \mathbf{P}\mathbf{u}_0 + K_n(\mathbf{P}\mathbf{A}\mathbf{P}^T, \mathbf{P}\mathbf{r}_0)}{\arg\min} \|\widetilde{\mathbf{u}} - \mathbf{v}\|_{\mathbf{P}\mathbf{A}\mathbf{P}^T} \\ &= \underset{\mathbf{v} \in \mathbf{P}\mathbf{u}_0 + \mathbf{P}K_n(\mathbf{A}, \mathbf{r}_0)}{\arg\min} \|\mathbf{P}\mathbf{u} - \mathbf{v}\|_{\mathbf{P}\mathbf{A}\mathbf{P}^T} \\ &= \underset{\mathbf{v} = \mathbf{P}\mathbf{w} \,|\, \mathbf{w} \in \mathbf{u}_0 + K_n(\mathbf{A}, \mathbf{r}_0)}{\arg\min} \|\mathbf{P}\mathbf{u} - \mathbf{v}\|_{\mathbf{P}\mathbf{A}\mathbf{P}^T} \\ &= \mathbf{P} \cdot \underset{\mathbf{w} \in \mathbf{u}_0 + K_n(\mathbf{A}, \mathbf{r}_0)}{\arg\min} \|\mathbf{P}\mathbf{u} - \mathbf{P}\mathbf{w}\|_{\mathbf{P}\mathbf{A}\mathbf{P}^T} \\ &= \mathbf{P} \cdot \underset{\mathbf{w} \in \mathbf{u}_0 + K_n(\mathbf{A}, \mathbf{r}_0)}{\arg\min} \|\mathbf{u} - \mathbf{w}\|_{\mathbf{A}} \\ &= \mathbf{P}\mathbf{u}_n , \end{aligned}$$

and therefore,

$$\|\widetilde{\mathbf{u}} - \widetilde{\mathbf{u}}_n\|_{\mathbf{P}\mathbf{A}\mathbf{P}^T} = \|\mathbf{P}\mathbf{u} - \mathbf{P}\mathbf{u}_n\|_{\mathbf{P}\mathbf{A}\mathbf{P}^T} = \|\mathbf{u} - \mathbf{u}_n\|_{\mathbf{A}} .$$

The convergence of PCG with an operator-based preconditioning (i.e. the **laplace** preconditioning in our illustrations) is also not affected by any reordering of the degrees of freedom. Furthermore, it is independent of the choice of the

discretization basis generating the given subspace V_h, which can be shown using the finite-dimensional analogy to (6.8). Given a finite-dimensional Hilbert space V_h and the Riesz map $\tau : V_h^\# \to V_h$, the CG method for solving the (finite-dimensional) problem (6.10) with the initial approximation $u_0 \in V_h$ provides in the n-th iteration the approximation $u_n \in u_0 + K_n(\tau\mathcal{A}_h, \tau r_0) \subset V_h$ such that

$$\|u_h - u_n\|_a = \min_{v \in u_0 + K_n(\tau\mathcal{A}_h, \tau r_0)} \|u_h - v\|_a \,.$$

The approximation u_n is uniquely determined and $\|u_h - u_n\|_a$ is clearly independent of the basis of V_h. Given a basis Φ with the associated $\mathbf{A}$, $\mathbf{f}$, $\widehat{\mathbf{M}}_{\text{laplace}}$, $\mathbf{u}_0$, the algebraic PCG in Algorithm 2 gives in the n-th iteration the coordinates $\mathbf{u}_n$ of u_n with respect to Φ, $u_n = \Phi\mathbf{u}_n$. Finally, for $\mathbf{u} = \mathbf{A}^{-1}\mathbf{f}$ there holds $u_h = \Phi\mathbf{u}$, and

$$\|u_h - u_n\|_a = \|\mathbf{u} - \mathbf{u}_n\|_{\mathbf{A}} \,.$$

A reordering of the degrees of freedom, however, affects convergence of PCG with **ichol0** and **ichol(TOL)** preconditioners; see, e.g., Duff and Meurant [1989]. Figure 6.5 depicts convergence of PCG with **ichol0** and **ichol(TOL)** in the test problems for the orderings described above. We note that PCG with **ichol0** converges for **symamd** ordering more slowly than for the other two orderings, which was observed also in Duff and Meurant [1989].

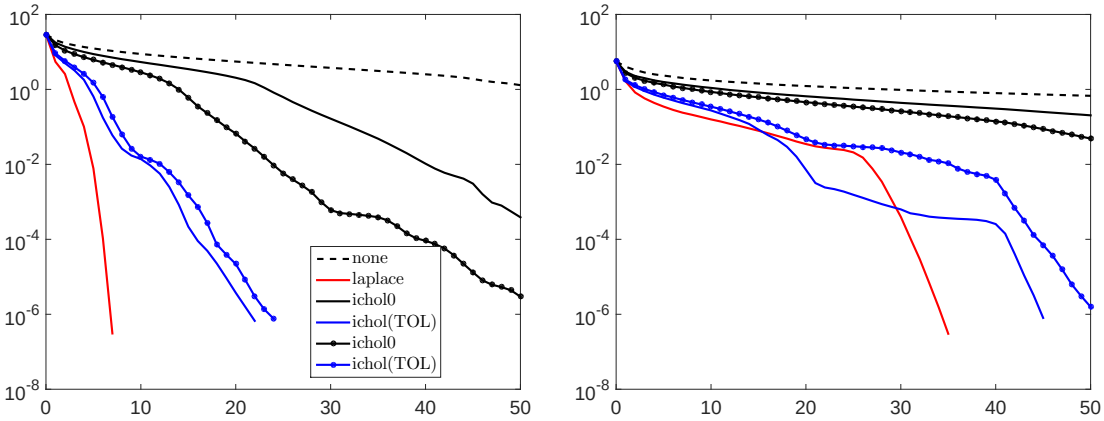
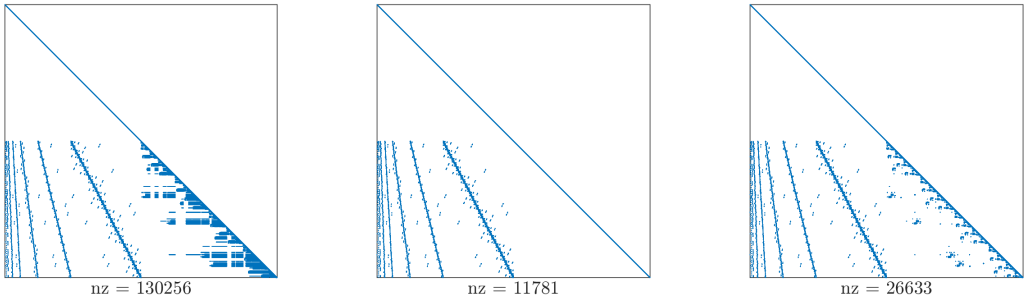


Figure 6.5: Convergence of the energy norm of the algebraic error in PCG for two orderings of degrees of freedom. Convergence is independent of ordering in unpreconditioned CG and in PCG with **laplace** preconditioning. For **ichol0** and **ichol(TOL)** preconditioning the convergence differs. The solid lines with no markers correspond to **symamd** ordering. The lines with dots correspond to **zigzag** ordering. Left: Inhomogeneous tensor problem I on the uniform mesh. Right: Inhomogeneous tensor problem II on the adaptively refined mesh.

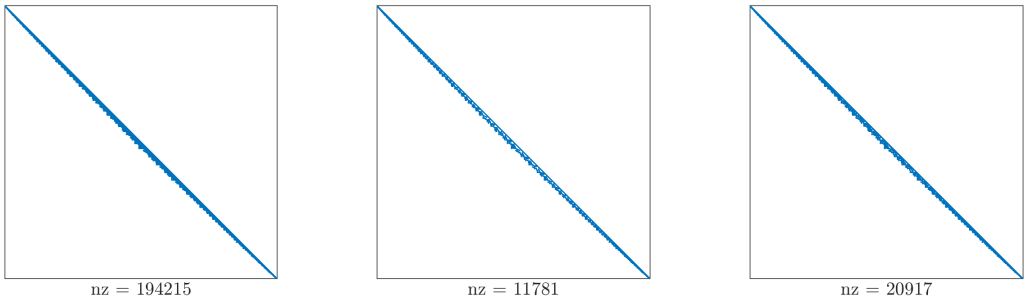
In the previous discussion we focused on the effect of reordering of the degrees of freedom on the PCG convergence, i.e., on the number of iterations that is necessary to drop the error below a given tolerance. However, the ordering also affects the cost of each PCG iteration, which will be illustrated in the following experiment.

Sparsity of the factors

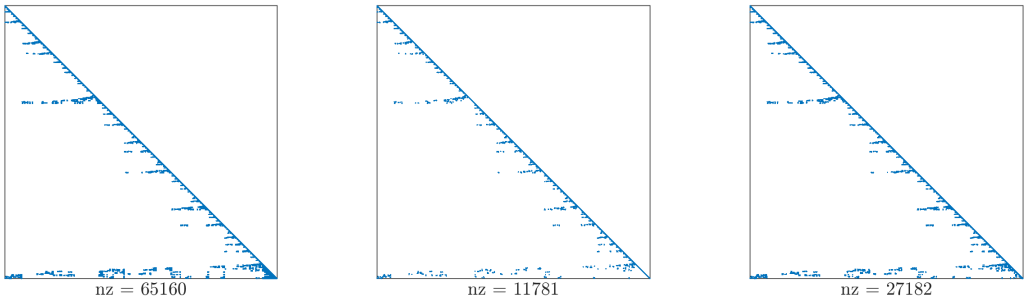
Given a preconditioner $\widehat{\mathbf{M}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^*$, we show the sparsity pattern of its Cholesky factor $\widehat{\mathbf{L}}$ and of the upper triangular matrix $(\widehat{\mathbf{L}}^*)^{-1}$ that provides the orthogonalization coefficients and determines the transformation of the discretization basis; see (6.16). In each figure we also show the number of nonzero elements (denoted as "nz"). This provides an important information about the cost of matrix-vector product and therefore also about the cost of one iteration step of PCG.



(a) Original ordering of the degrees of freedom

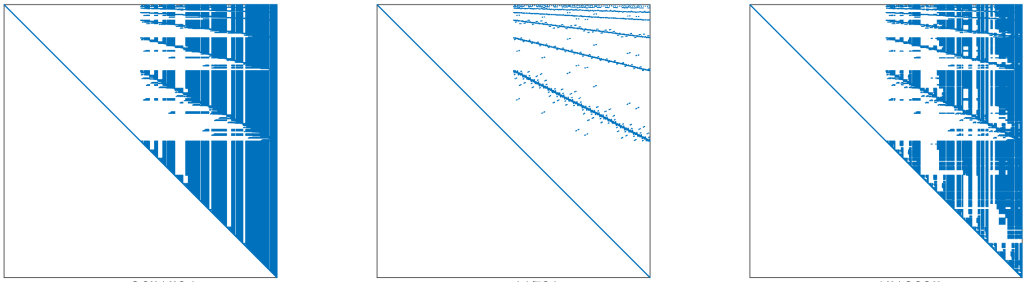


(b) **Zigzag** ordering

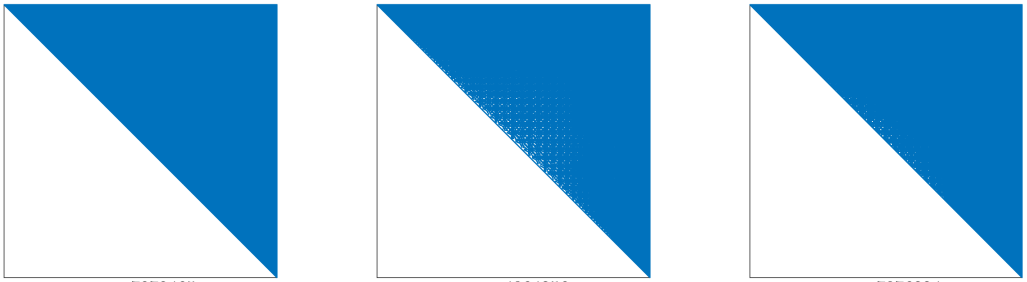


(c) **Symamd** ordering

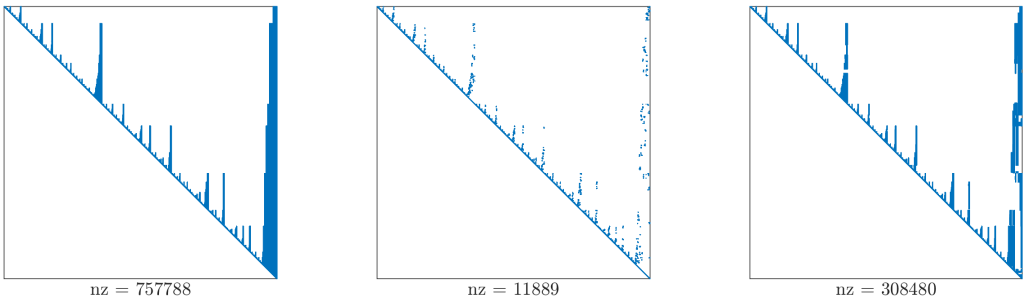
Figure 6.6: Inhomogeneous tensor problem I, uniform mesh: sparsity pattern of the Cholesky factor $\widehat{\mathbf{L}}$ of the preconditioner. Left: **laplace**; middle: **ichol0**; right: **ichol(TOL)**. "nz" stands for the number of nonzero elements.



(a) Original ordering of the degrees of freedom

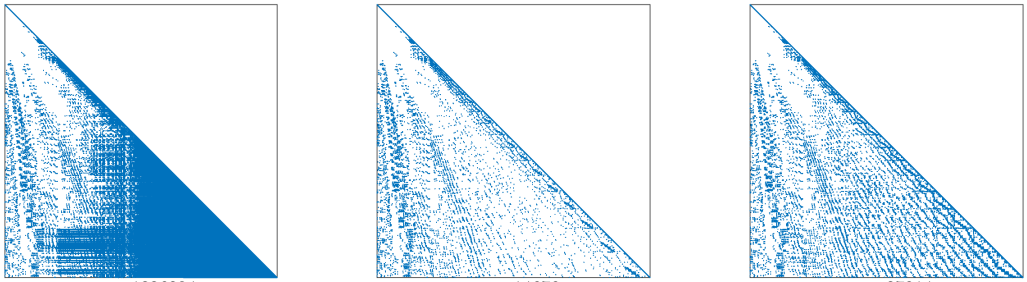


(b) **Zigzag** ordering

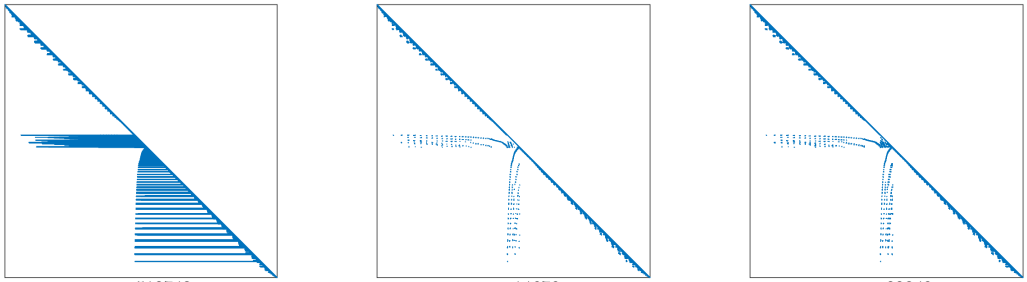


(c) **Symamd** ordering

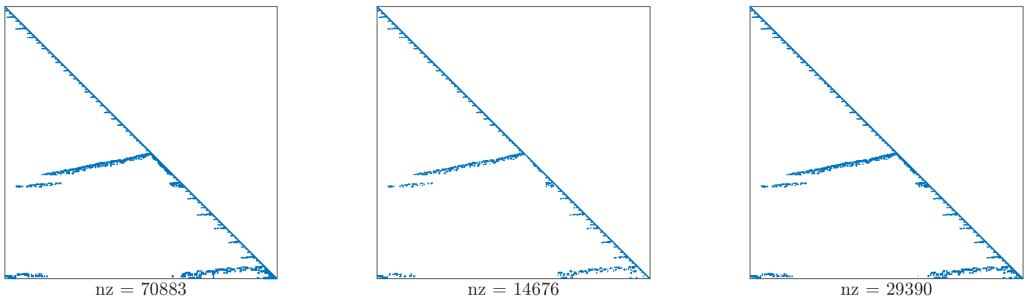
Figure 6.7: Inhomogeneous tensor problem I, uniform mesh: sparsity pattern of the upper triangular matrix $(\widehat{\mathbf{L}}^*)^{-1}$ of the orthogonalization coefficients. Left: **laplace**; middle: **ichol0**; right: **ichol(TOL)**.



(a) Original ordering of the degrees of freedom

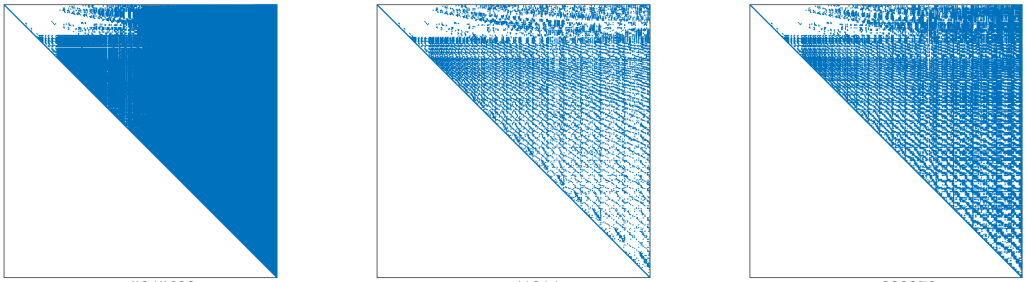


(b) **Zigzag** ordering

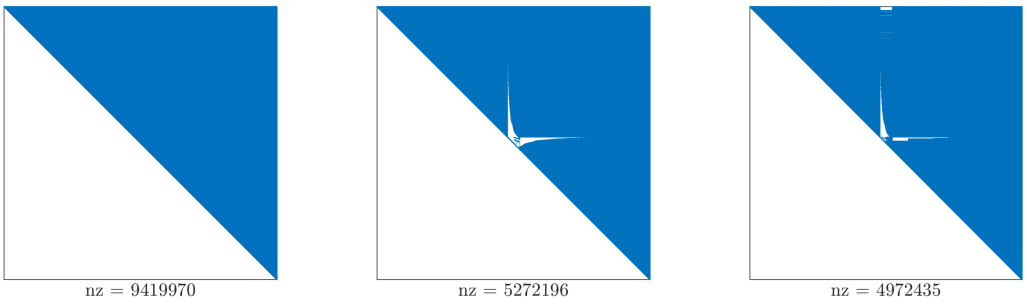


(c) **Symamd** ordering

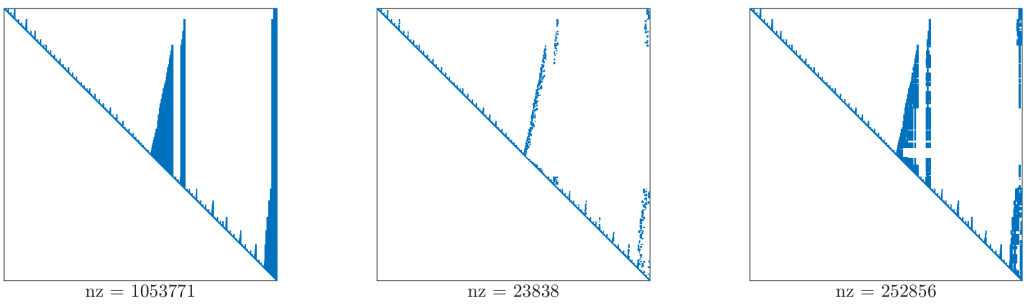
Figure 6.8: Inhomogeneous tensor problem II, adaptively refined mesh: sparsity pattern of the Cholesky factor $\widehat{\mathbf{L}}$ of the preconditioner. Left: **laplace**; middle: **ichol0**; right: **ichol(TOL)**.



(a) Original ordering of the degrees of freedom



(b) **Zigzag** ordering



(c) **Symamd** ordering

Figure 6.9: Inhomogeneous tensor problem II, adaptively refined mesh: sparsity pattern of the upper triangular matrix $(\widehat{\mathbf{L}}^*)^{-1}$ of the orthogonalization coefficients. Left: **laplace**; middle: **ichol0**; right: **ichol(TOL)**.

Figures 6.6 to 6.9 illustrate that the sparsity pattern of the Cholesky factors $\widehat{\mathbf{L}}$ as well as of the transformation matrices $(\widehat{\mathbf{L}}^*)^{-1}$ can be significantly affected by the chosen ordering of the degrees of freedom. For the **zigzag** ordering, as expected from its construction, the transformation matrix in **laplace** preconditioning has full upper triangle. Also the transformation matrices in **ichol0** and **ichol(TOL)** are denser than for the other orderings; see Figures 6.7b and 6.9b. The **symamd** permutation in our experiments indeed reduces the number of nonzeros in the Cholesky factor of $\widehat{\mathbf{M}}_{\text{laplace}}$; the number of nonzeros in the Cholesky factor of $\widehat{\mathbf{M}}_{\text{ichol0}}$ is the same for any ordering, and the Cholesky factor of $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ is for **symamd** slightly denser than for the original and **zigzag** orderings. However, **symamd** permutation significantly reduces the fill-in in the transformation matrices for all three considered preconditioners; see Figures 6.7c and 6.9c.

Figures 6.7 and 6.9 demonstrate that, independently of the ordering, some columns of the transformation matrices $(\widehat{\mathbf{L}}^*)^{-1}$ for $\widehat{\mathbf{M}}_{\text{laplace}}$ and $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ are dense. In other words, **laplace** and **ichol(TOL)** preconditioners transform some of the original locally supported hat-functions into discretization basis functions with global support. In the following experiment we will focus on the shape of these transformed functions.

6.5.3 Transformed discretization basis functions

We now illustrate how the transformed discretization basis functions $\widehat{\Phi} = \Phi(\widehat{\mathbf{L}}^*)^{-1}$ look like for the given preconditioners. As it is unfeasible to plot the whole basis, the following figures depict only the function $\widehat{\phi}_N$ determined by the coefficients in the last column of $(\widehat{\mathbf{L}}^*)^{-1}$. In the figures, we compare $\widehat{\phi}_N$ with $\widehat{\phi}_{a,N}$, the last of the basis functions $\widehat{\Phi}_a$ orthogonal with respect to the inner product induced by the bilinear form $a(\cdot,\cdot)$. This means that $\widehat{\Phi}_a = \Phi(\mathbf{L}^*)^{-1}$, where $\mathbf{L}$ is the Cholesky factor of the stiffness matrix $\mathbf{A} = (\Phi,\Phi)_a$. The functions $\widehat{\Phi}_a$ are of special interest; when using them for the algebraic representation of the problem (6.10), the discrete operator $\mathcal{A}_h$ is represented by the identity matrix; see (6.11). A natural question arises, to what extent $\widehat{\phi}_N \approx \widehat{\phi}_{a,N}$.

In the figures, it is in some parts of the domain difficult to get an information about the value of the transformed function. We therefore show also the percentage of nonzero entries in the coefficient vector that is denoted as “fill”.

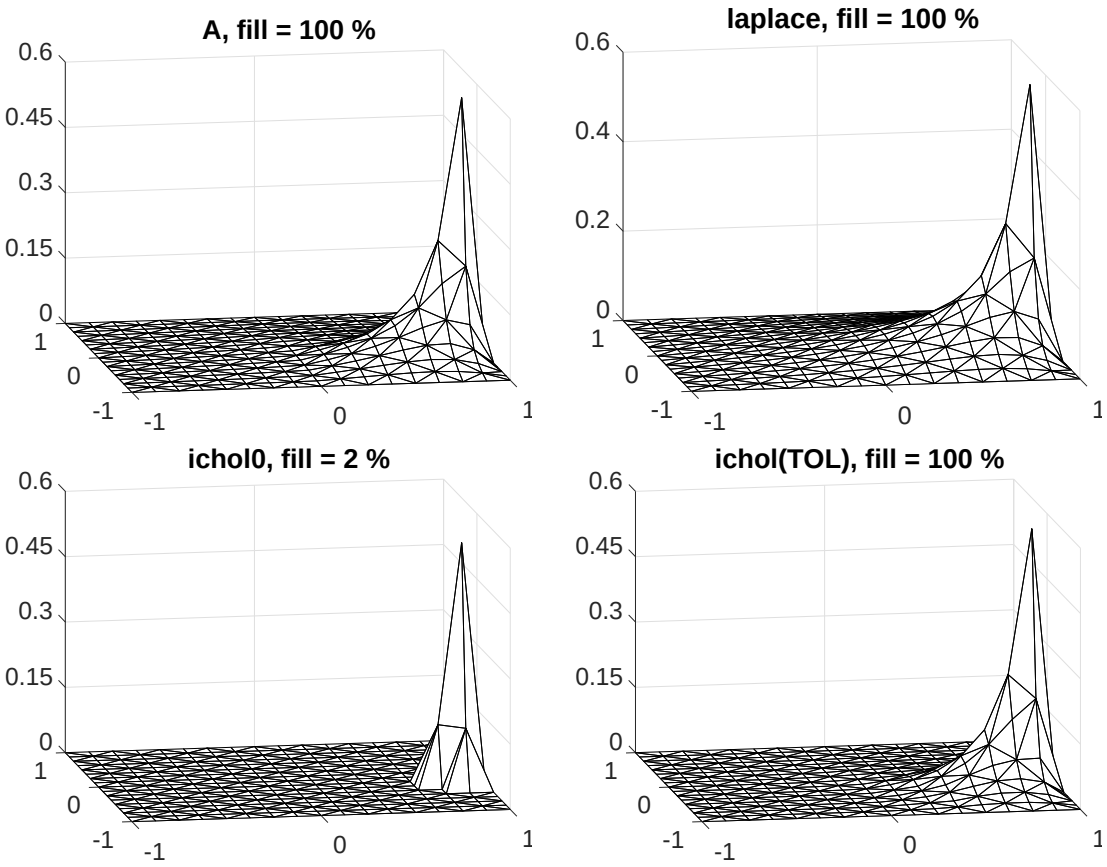


Figure 6.10: Inhomogeneous tensor problem I, uniform mesh, **symamd** ordering: the last of the transformed discretization basis functions $\widehat{\Phi} = \Phi(\widehat{\mathbf{L}}^*)^{-1}$ (see (6.16)) with $\widehat{\mathbf{L}}$ equal to the Cholesky factor of $\mathbf{A}$ (upper left), of $\widehat{\mathbf{M}}_{\text{laplace}}$ (upper right), of $\widehat{\mathbf{M}}_{\text{ichol0}}$ (bottom left), and of $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ (bottom right).

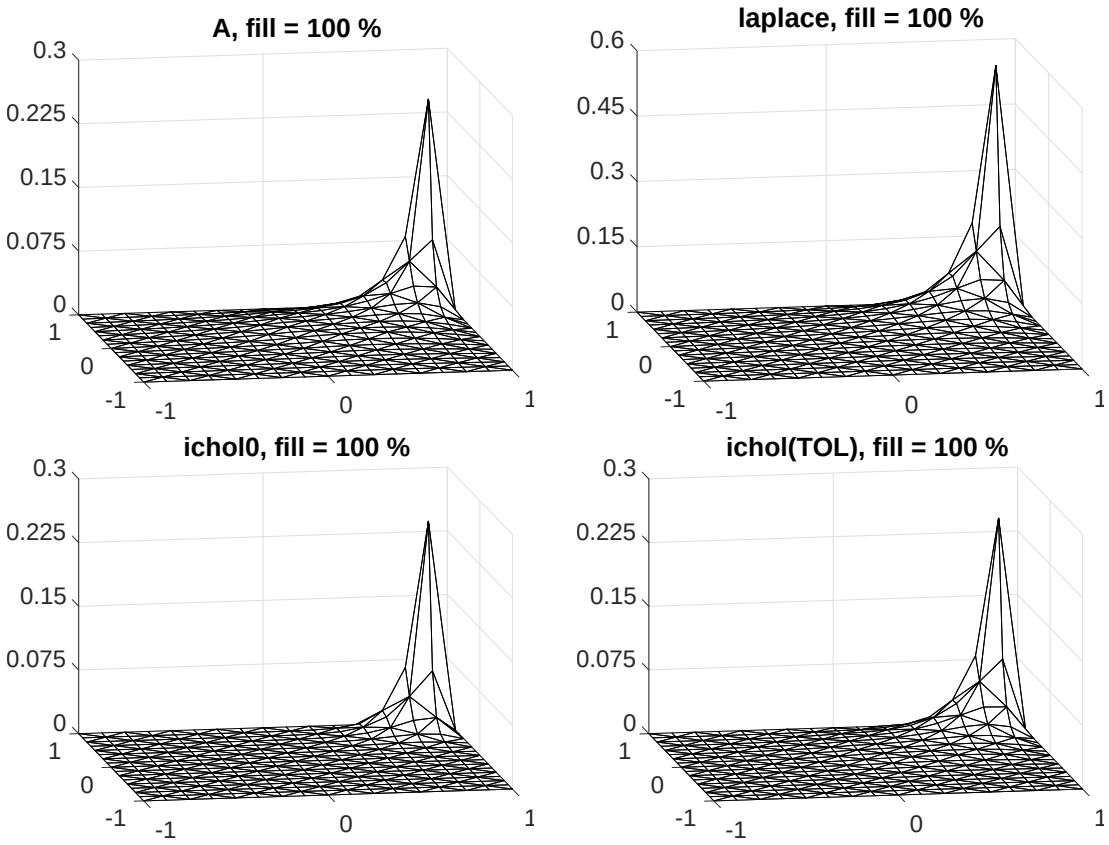


Figure 6.11: Inhomogeneous tensor problem I, uniform mesh, **zigzag** ordering: the last of the transformed discretization basis functions $\widehat{\Phi} = \Phi(\widehat{\mathbf{L}}^*)^{-1}$ (see (6.16)) with $\widehat{\mathbf{L}}$ equal to the Cholesky factor of $\mathbf{A}$ (upper left), of $\widehat{\mathbf{M}}_{\text{laplace}}$ (upper right), of $\widehat{\mathbf{M}}_{\text{ichol0}}$ (bottom left), and of $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ (bottom right).

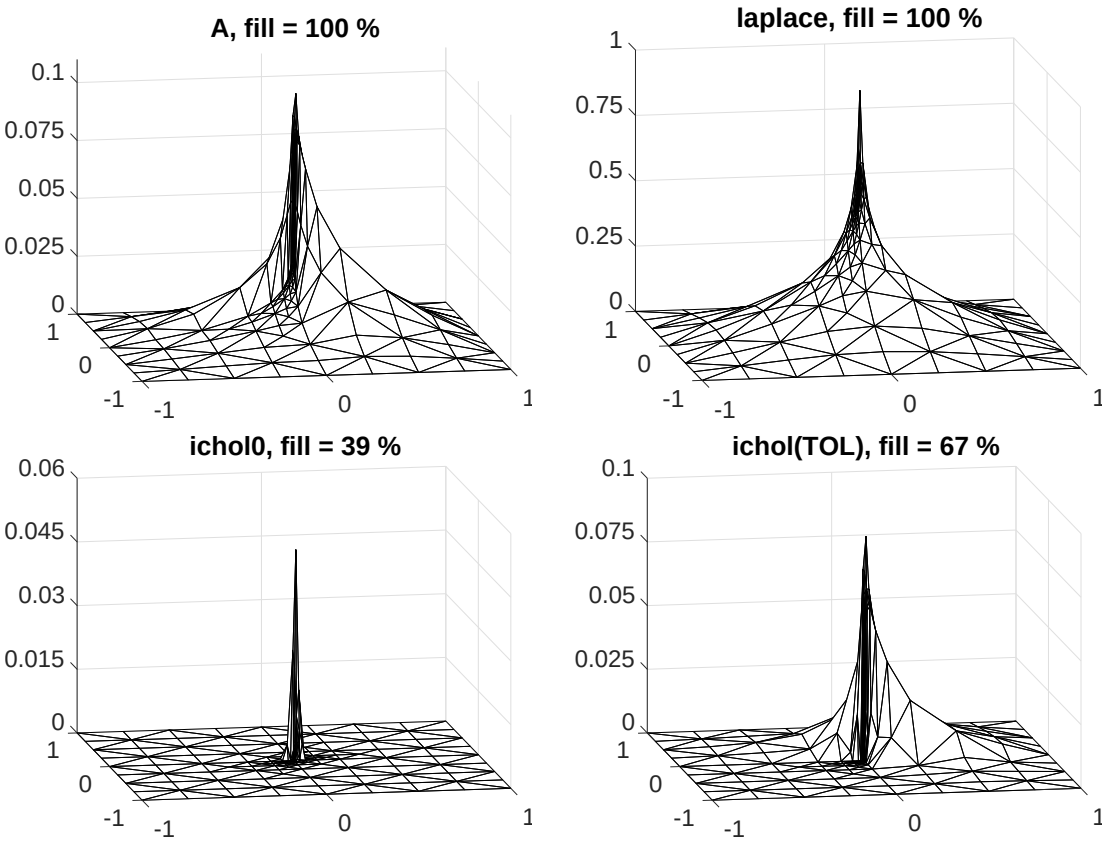


Figure 6.12: Inhomogeneous tensor problem II, adaptively refined mesh, **symamd** ordering: the last of the transformed discretization basis functions $\widehat{\Phi} = \Phi(\widehat{\mathbf{L}}^*)^{-1}$ (see (6.16)) with $\widehat{\mathbf{L}}$ equal to the Cholesky factor of $\mathbf{A}$ (upper left), of $\widehat{\mathbf{M}}_{\text{laplace}}$ (upper right), of $\widehat{\mathbf{M}}_{\text{ichol0}}$ (bottom left), and of $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ (bottom right).

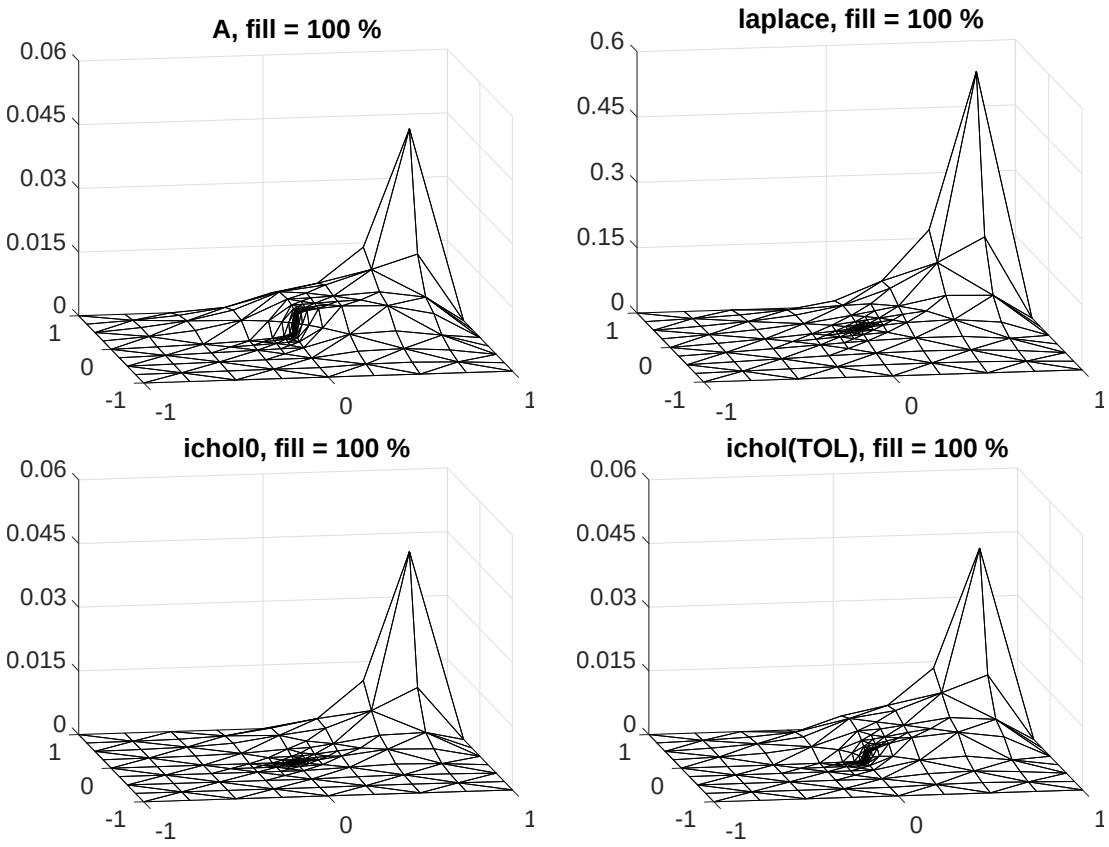


Figure 6.13: Inhomogeneous tensor problem II, adaptively refined mesh, **zigzag** ordering: the last of the transformed discretization basis functions $\widehat{\Phi} = \Phi(\widehat{\mathbf{L}}^*)^{-1}$ (see (6.16)) with $\widehat{\mathbf{L}}$ equal to the Cholesky factor of $\mathbf{A}$ (upper left), of $\widehat{\mathbf{M}}_{\text{laplace}}$ (upper right), of $\widehat{\mathbf{M}}_{\text{ichol0}}$ (bottom left), and of $\widehat{\mathbf{M}}_{\text{ichol(TOL)}}$ (bottom right).

The transformed function given by the **laplace** preconditioner can differ from $\widehat{\phi}_{a,N}$ in the shape (this is in particular visible in the second test problem; see Figures 6.12 and 6.13) and also in the scaling; see Figures 6.11 to 6.13. For the preconditioners based on the incomplete Cholesky decomposition of the stiffness matrix we (mostly) observed a similar scaling of $\widehat{\phi}_N$ and $\widehat{\phi}_{a,N}$. The effect of dropping some coefficients in the construction of the **ichol(TOL)** preconditioner on the shape of the associated transformed function is visible in the second test problem; see in particular Figure 6.12. For the **ichol0** preconditioner and **symamd** ordering of the degrees of freedom the transformed function $\widehat{\phi}_N$ has relatively small support in both test problems; see Figures 6.10 and 6.12. In the previous numerical examples we observed that PCG with **ichol0** converges for **symamd** ordering more slowly than for the other two orderings; cf. Figure 6.5. This is in line with the small support of the transformed functions.

6.6 Comments and outlook

Efficient numerical solution of difficult problems requires applying various preconditioning techniques. The point made in Málek and Strakoš [2015] and recalled in this chapter is that preconditioning should not be considered separated from discretization.

The ideas linking discretization and preconditioning are certainly not new; see [Málek and Strakoš, 2015, Section 8] for a thorough discussion and the list of references. For example, construction of an algebraic preconditioner based on an appropriate transformation of the original basis was used already a long time ago, e.g., in the so-called hierarchical basis preconditioning (see, e.g., Yserentant [1985, 1986]). However, the presented approach of Málek and Strakoš [2015] is more general, with preconditioning considered *not* only as a remedy for improving the properties of the (unpreconditioned) algebraic problem but also as an inherent issue closely linked with discretization.

Preconditioners incorporating coarse space information (such as, e.g., multilevel preconditioners or domain decomposition techniques with coarse space components) often prove particularly efficient. Algebraically constructed preconditioners are based on approximate solution of (a part of) the problem, often without a direct geometrical analogy. One may ask in which way these preconditioners provide a global exchange of information in function spaces associated with the underlying mathematical model. The interpretation of algebraic preconditioning as transformation of the discretization basis seems suitable for answering this question. The above numerical experiments illustrate that some of the original basis functions with local support are in this way transformed into functions supported over the whole discretization domain. Presented results also allow to investigate algebraic techniques such as reordering of the degrees of freedom that can be used in efficient algebraic preconditioning.

Further research can start from interpreting efficient algebraically constructed preconditioners as the corresponding transformations of the discretization basis (and of the associated inner product), and from investigating whether a similar basis can be used directly for the discretization of the analogous infinite-dimensional problems. Such questions might be relevant, in particular, in massively parallel computer environment.

Bibliography

P. R. Amestoy, T. A. Davis, and I. S. Duff. An approximate minimum degree ordering algorithm. *SIAM J. Matrix Anal. Appl.*, 17(4):886–905, 1996. ISSN 0895-4798.

M. Benzi and M. Tůma. Orderings for factorized sparse approximate inverse preconditioners. *SIAM J. Sci. Comput.*, 21(5):1851–1868, 2000. ISSN 1064-8275. Iterative methods for solving systems of algebraic equations (Copper Mountain, CO, 1998).

P. G. Ciarlet. *Linear and nonlinear functional analysis with applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013. ISBN 978-1-611972-58-0.

I. S. Duff and G. A. Meurant. The effect of ordering on preconditioned conjugate gradients. *BIT*, 29(4):635–657, 1989. ISSN 0006-3835.

T. Gergelits and Z. Strakoš. Composite convergence bounds based on Chebyshev

polynomials and finite precision conjugate gradient computations. *Numer. Algorithms*, 65(4):759–782, 2014. ISSN 1017-1398.

A. Greenbaum. Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra Appl.*, 113:7–63, 1989. ISSN 0024-3795.

R. Hiptmair. Operator preconditioning. *Comput. Math. Appl.*, 52(5):699–706, 2006. ISSN 0898-1221.

J. Liesen and Z. Strakoš. *Krylov subspace methods: principles and analysis*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013. ISBN 978-0-19-965541-0.

J. W. H. Liu. Modification of the minimum-degree algorithm by multiple elimination. *ACM Trans. Math. Software*, 11(2):141–153, 1985. ISSN 0098-3500.

J. Málek and Z. Strakoš. *Preconditioning and the conjugate gradient method in the context of solving PDEs*, volume 1 of *SIAM Spotlights*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015. ISBN 978-1-611973-83-9.

G. Meurant and Z. Strakoš. The Lanczos and conjugate gradient algorithms in finite precision arithmetic. *Acta Numer.*, 15:471–542, 2006. ISSN 0962-4929.

Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, second edition, 2003. ISBN 0-89871-534-2.

J. Štuller. Ordered modified Gram-Schmidt orthogonalization revised. *J. Comput. Appl. Math.*, 63(1-3):221–227, 1995. ISSN 0377-0427. International Symposium on Mathematical Modelling and Computational Methods Modelling 94 (Prague, 1994).

H. Yserentant. Hierarchical bases of finite-element spaces in the discretizaton of nonsymmetric elliptic boundary value problems. *Computing*, 35(1):39–49, 1985. ISSN 0010-485X.

H. Yserentant. Hierarchical bases give conjugate gradient type methods a multigrid speed of convergence. *Appl. Math. Comput.*, 19(1-4):347–358, 1986. ISSN 0096-3003. Second Copper Mountain conference on multigrid methods (Copper Mountain, Colorado, 1985).

7. Conclusions

Algebraic errors in numerical solution of partial differential equations presents a very broad topic of study. The fact that the numerical computation is, in general, not accurate, and it is even principally desirable in many cases not to perform it to a high accuracy, has consequences that have to be taken into account in numerical analysis and in practical computations. In this thesis we have investigated and illustrated some of them. We have focused on (preconditioned) iterative algebraic solvers.

The first point of the thesis is that the algebraic error can be highly unevenly distributed over the solution domain. It can have large local components, which can significantly dominate the total error in some parts of the domain. This can happen despite the fact that the energy or the L^2 norm of the algebraic error is small in comparison to the norm of the discretization error. It motivates developing a posteriori error estimators that can provide a reliable information not only about the global error norms but also about the local distribution of the various types of the error (Chapter 2).

The second point concerns the backward error interpretation of the algebraic error in the context of function approximations. Using the algebraic backward error, a standard methodology that interprets the inaccuracies in the algebraic solution as a modification (perturbation) of the data defining the problem, we have linked the algebraic inaccuracies with modifications of the original model and its discretization. This underlines importance of understanding the interconnections between the phases of the solution process (such as discretization and algebraic computation). There is much to be done (Chapter 3).

The third point concerns adaptivity and the price to be paid for increasing the reliability and accuracy of the a posteriori error estimates. The key feature of an efficient numerical PDE solver is adaptivity based on a posteriori error estimation. Historically, most a posteriori analysis in numerical PDEs focuses on estimating the discretization error. A posteriori analysis is often based on the assumption of the exact solution of the discretized problem. This assumption is principally restrictive. Using the so-called residual-based error estimator as an example, we have studied the impact of abandoning the assumption of the exact algebraic solution. In order to justify the evaluation of the estimator at the presence of the algebraic error, the construction of the estimator has to be carefully revisited. Then we have numerically illustrated the effect of the algebraic error on the adaptive finite element discretizations based on the local residual-based error indicators. The results of the experiments suggest that in practical computations the effect of the algebraic error to adaptivity is worth to investigate. When using inappropriate stopping criteria, the efficiency of the whole adaptive procedure and also reaching the prescribed accuracy can be endangered (Chapter 4).

Despite the fact that there is a growing body of very substantial work avoiding the unrealistic assumption on the exact solution of the algebraic problem, a mathematically justified, inexpensive and tight estimation of the discretization and algebraic errors that would allow for their comparison in practical computations is not, in our opinion, a fully solved problem. To provide a step towards resolving this problem, we have shown a methodology for computing upper and

lower bounds on the algebraic and total error norms based on the flux reconstruction. The derived bounds allow for estimating the local distribution of the errors over the computational domain. We have discussed bounds on the discretization error, application of the results for constructing rigorously justified stopping criteria for iterative algebraic solvers, and the relationship to the previously published estimates on the algebraic error. The presented results indicate the difficulties one has to cope with in rigorous approach for including algebraic errors into a posteriori error estimates (Chapter 5).

Finally, the last point investigates the link between preconditioning and transformation of the discretization bases. Efficient numerical solution of difficult problems requires applying various preconditioning techniques. We have followed the idea that the discretization of a given mathematical model and preconditioning of the associated algebraic system are tightly coupled. The links between the algebraic preconditioning and the transformation of the discretization basis allow to interpret, e.g., efficient algebraically constructed preconditioners in terms of transformations enlarging the support of the discretization basis functions (Chapter 6).

Many questions regarding the algebraic error in numerical PDEs remain widely open. We list some of them related to the topics in this thesis that are, in our opinion, of particular importance:

- Deriving tight bounds on the errors of different origin (such as discretization and algebraic) that allow for accurate estimating the spatial distribution of the errors across the computational domain and that are inexpensive to evaluate. Derivations should contain no assumptions that are impossible to fulfill in practical computations and clearly declare all assumptions that can restrict applicability of the results.

- Deriving mathematically justified stopping criteria that balance (in the appropriate problem-dependent sense) the errors of different origin and that avoid stopping the algebraic iterations prematurely. Heuristics that must be used in many practical computations should admit possible gaps in rigorous justification in order to open the door for further investigation.

- Investigating procedures that would allow to efficiently reduce the algebraic error in some parts of the computational domain where it is indicated to be large.