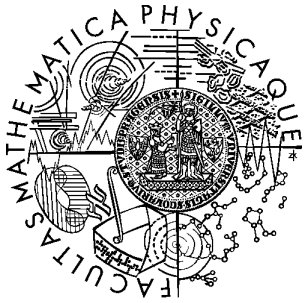
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Jan Papež

Odhady algebraické chyby a zastavovací kritéria v numerickém řešení parciálních diferenciálních rovnic

Katedra numerické matematiky

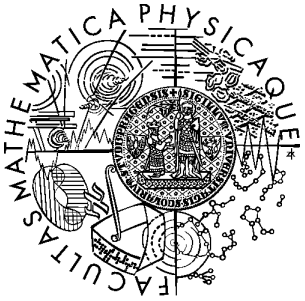
Vedoucí diplomové práce: prof. Ing. Zdeněk Strakoš, DrSc.

Studijní program: Matematika
Studijní obor: Numerická a výpočtová matematika

Praha 2011

Charles University in Prague
Faculty of Mathematics and Physics

MASTER THESIS



Jan Papež

Estimation of the algebraic error and stopping criteria in numerical solution of partial differential equations

Department of Numerical Mathematics

Supervisor of the master thesis: prof. Ing. Zdeněk Strakoš, DrSc.

Study Programme: Mathematics
Specialization: Computational Mathematics

Prague 2011

Rád bych poděkoval své rodině a přátelům za podporu a pomoc, doktoru Tichému za cenné připomínky a především profesoru Strakošovi za trpělivé a odborné vedení.

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 25. 7. 2011 Jan Papež

Název práce: Odhady algebraické chyby a zastavovací kritéria v numerickém řešení parciálních diferenciálních rovnic
Autor: Jan Papež
Katedra: Katedra numerické matematiky
Vedoucí diplomové práce: prof. Ing. Zdeněk Strakoš, DrSc.

Abstrakt: Po uvedení modelového problému a jeho vlastností je v práci popsána metoda sdružených gradientů (Conjugate Gradient Method - CG), jsou uvedeny odhady energetické normy chyby a je navržena heuristika pro adaptivní zpřesňování odhadů ve výpočtech. Na konkrétních příkladech je ukázán rozdíl v lokálním chování algebraické a diskretizační chyby v numerickém řešení modelového problému. Dále jsou uvedeny *a posteriori* odhady diskretizační a celkové chyby, které zahrnují chybu řešení algebraické soustavy. Myšlenka použití více sítí při řešení modelového problému je ukázána na víceúrovňové metodě (multigrid method). Poté je popsána Deuflhardova metoda Cascadic Conjugate Gradient Method (CCG), pro kterou jsou odvozena nová zastavovací kritéria s využitím odhadů algebraické a diskretizační chyby popsaných v předchozích částech předložené práce. Na závěr je metoda CCG s novými zastavovacími kritérii testována.

Klíčová slova: numerické řešení parciálních diferenciálních rovnic, chyba diskretizace, algebraická chyba, odhady chyby, lokální chování chyby, adaptivita

Title: Estimation of the algebraic error and stopping criteria in numerical solution of partial differential equations
Author: Jan Papež
Department: Department of Numerical Mathematics
Supervisor of the master thesis: Zdeněk Strakoš

Abstract: After introduction of the model problem and its properties we describe the Conjugate Gradient Method (CG). We present the estimates of the energy norm of the error and a heuristic for the adaptive refinement of the estimate. The difference in the local behaviour of the discretization and the algebraic error is illustrated by numerical experiments using the given model problem. *A posteriori* estimates for the discretization and the total error that take into account the inexact solution of the algebraic system are then discussed. In order to get a useful perspective, we briefly recall the multigrid method. Then the Cascadic Conjugate Gradient Method of Deuflhard (CCG) is presented. Using the estimates for the error presented in the preceding parts of the thesis, the new stopping criteria for CCG are proposed. The CCG method with the new stopping criteria is then tested.

Keywords: numerical PDE, discretization error, algebraic error, error estimates, locality of the error, adaptivity

Contents

Introduction

The introduction of *a posteriori* error estimates for finite element method in the paper of Babuška and Rheinboldt [3] brought considerable advance in the error analyzis in the solution of partial differential equations. The subject was further developed in vast amount of literature, we reffer, e.g., to the book of Ainsworth and Oden [1]. Apart from few exceptions *a posteriori* error estimates rely on the assumption that the linear algebraic system resulting from discretization is solved exactly. A moderately sized system can be solved by direct methods; for large systems the (preconditioned) iterative methods become competitive and with increasing side they represent the only viable alternative. Moreover, they enable saving the computational work by stopping whenever the algebraic error drops to the level at which it does not significantly affect the whole error. In last years *a posteriori* error estimates which take into account an inexact solving of the linear algebraic system were derived, see, e.g., [4, 23, 35, 36, 47].

The solution of a partial differential equation (PDE) by the finite element method reduces the original mathematical model to the *discretized problem*, where the approximate solution is restricted to some finite-dimensional function subspace, and to the *algebraic problem* that determinates the coefficients for the approximate solution with respect to the given basis of the finite-dimensional subspace (see, e.g., [25, Section 1]). We believe that any reliable and effective PDE solver requires the understanding of the relations between these problems.

The goals of the thesis are to explain the connection between the discretization and the algebraic error, to present an overview of the estimates for the algebraic error and to implement the estimates for the error and the stopping criteria in an adaptive finite element method.

The thesis is organized as follows. In the first chapter we describe the model problem and its discretization and we show the existence and the uniqueness of the solution. In Chapter 2, the Conjugate Gradient Method

(CG, [21]) is described, estimates for the energy norm of the error in CG are derived and a heuristic for the adaptive estimate is proposed and numerically tested. In the numerical experiments we also show the difference in the local distribution of the algebraic and the discretization error. Then the *a posteriori* error estimates including the algebraic error are presented in Chapter 3. In order to get a useful perspective we describe multigrid methods in Chapter 4. In Chapter 5 the Cascadic Conjugate Gradient Method (CCG, [10]) is described. Using the results from Chapter 2 we propose new stopping criteria for the CCG method. The new implementation of the CCG method with the *a posteriori* error estimates described in Chapter 3 is then tested.

Chapter 1

Second-order elliptic PDEs and their discretization

In this chapter we present the model problem considered in the thesis, its basic properties and discretization by the Galerkin Finite Element Method (FEM). The theorems are presented briefly and without proofs (the details and proofs can be found, e.g., in [14, 6, 18]).

1.1 Sobolev spaces

Let D be a domain (open, bounded and connected set) in $\mathbb{R}^d$, $d = 2, 3$. We say that ∂D is a *Lipschitz boundary* and D is called a *Lipschitz domain*, if for each point $s = (s_1, \ldots, s_d) \in \partial D$ there exists a ball $B_r(s)$ of radius $r > 0$ centered at s and a Lipschitz function $\gamma : \mathbb{R}^{d-1} \to \mathbb{R}$ such that – upon relabeling and reorienting the coordinate axes if necessary – we have

$$D \cap B_r(s) = \{x \in B_r(s); x_d > \gamma(x_1, \ldots, x_{d-1})\} \ .$$

For $1 \leq p < \infty$, let

$$\|u\|_{L^p(D)} \equiv \left(\int_D u^p \, ds \right)^{1/p}$$

and for the case $p = \infty$ let

$$\|u\|_{L^\infty(D)} \equiv \operatorname{ess\,sup}_D |u| = \inf \left\{ C \in \mathbb{R};\ \int_{\{\xi; |u(\xi)| > C\}} ds = 0 \right\} .$$

We define the *Lebesgue space* $L^p(D)$, $1 \leq p \leq \infty$

$$L^p(D) \equiv \left\{ u : D \to \mathbb{R}; \|u\|_{L^p(D)} < \infty \right\} .$$

In order to avoid trivial differences between two functions, we identify two functions $u, v \in L^p(D)$ that satisfy $\|u - v\|_{L^p(D)} = 0$.

Let

$$\alpha = (\alpha_1, \ldots, \alpha_d), \quad \alpha_i \in \{0, 1, \ldots\}, \ 1 \leq i \leq d$$

be a multiindex and $|\alpha| = \sum_{i=1}^{d} \alpha_i$. By $C_c^\infty(D)$ we denote the space of infinitely differentiable functions with compact support in D. For $\varphi \in C_c^\infty(D)$ we define the (strong) multiindex derivation

$$D^\alpha \varphi = \frac{\partial^{|\alpha|} \varphi}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} .$$

We define the space of locally integrable functions

$$L^1_{loc} \equiv \{ u : D \to \mathbb{R}; u \in L^1(K) \quad \forall \textit{compact } K \subset D \}.$$

As D is bounded, $L^p(D) \subset L^1_{loc}, 1 \leq p \leq \infty$. For $u \in L^1_{loc}(D)$ we define the *weak derivation* $v = D^\alpha u$ such that

$$\int_D v\varphi \, dx = (-1)^{|\alpha|} \int_D u D^\alpha \varphi \, dx \quad \forall \varphi \in C_c^\infty(D) .$$

The weak derivation is uniquely determined (see [14, Section 5.2.1]). We define the *Sobolev space* $W^{k,p}(D), k \in \mathbb{N}, 1 \leq p \leq \infty$

$$W^{k,p}(D) \equiv \{ u \in L^p(D); D^\alpha u \in L^p(D) \quad \forall \alpha; |\alpha| \leq k \} ,$$

with the norm

$$\|u\|_{k,p,D} \equiv \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p(D)} \right) .$$

For $p = 2$, $W^{k,2}(D)$ is a Hilbert space that we denote by

$$H^k(D) \equiv W^{k,2}(D)$$

and the norm

$$\|u\|_{k,D} \equiv \|u\|_{k,2,D} .$$

According to this notation, we denote by $H^0(D)$ the Hilbert space $L^2(D)$ and its norm

$$\|u\|_{0,D} \equiv \|u\|_{L^2(D)} .$$

For a Lipschitz domain D we define the *trace operator* (see, e.g., [14, Section 5.5])

$$T : W^{1,p}(D) \to L^p(\partial D), \quad 1 \le p < \infty ,$$

such that

$$Tu = u|_{\partial D} \quad \forall u \in C(\bar{D}) \cap W^{1,p}(D) .$$

For the definition of *Sobolev-Slobodetskii spaces* $W^{s,p}(D)$ with $s \notin \mathbb{N}$ see, e.g., [30].

1.2 Model problem, weak solution

Consider a second-order elliptic pure diffusion model problem

$$\begin{aligned} -\nabla \cdot (\mathbf{S}\nabla u) &= f \quad \text{in} \quad \Omega , && (1.1)\\ u &= g_{\mathcal{D}} \quad \text{on} \quad \partial\Omega_{\mathcal{D}} , && (1.2)\\ \frac{\partial u}{\partial n} &= g_{\mathcal{N}} \quad \text{on} \quad \partial\Omega_{\mathcal{N}} , && (1.3) \end{aligned}$$

where

A1: Ω is a Lipschitz domain in $\mathbb{R}^d$, $d = 2, 3$, $\partial\Omega = \partial\Omega_{\mathcal{D}} \cup \partial\Omega_{\mathcal{N}}$ and $\partial\Omega_{\mathcal{D}} \cap \partial\Omega_{\mathcal{N}} = \emptyset$;

A2: $\mathbf{S}$ is a symmetric, bounded and uniformly positive diffusion tensor, i.e.

$$\mathbf{S} = (s_{ij})_{i,j=1}^d, \quad s_{ij} \in L^\infty(\Omega), \quad s_{ij} = s_{ji}, \quad i, j = 1, \dots, d,$$
$$\|\mathbf{S}\| \equiv \sup_{\xi\in\Omega} \sup_{0\neq z\in\mathbb{R}^d} \frac{\|\mathbf{S}(\xi)z\|}{\|z\|} < \infty$$

and there exists constant $c_{\mathbf{S}} > 0$ such that

$$c_{\mathbf{S}}\|z\|^2 \le z^T\mathbf{S}(\xi)z, \quad \forall \xi \in \Omega, \ \forall z \in \mathbb{R}^d.$$

A3: $f \in L^2(\Omega)$ is a source term;

A4: $g_{\mathcal{D}} \in L^2(\partial\Omega_{\mathcal{D}})$ prescribes the Dirichlet boundary condition. We assume that there exists $u_{\mathcal{D}} \in H^1(\Omega)$ such that

$$Tu_{\mathcal{D}} = g_{\mathcal{D}}\,;$$

A5: $g_{\mathcal{N}} \in L^2(\partial\Omega_{\mathcal{N}})$ prescribes the Neumann boundary condition.

Here $\partial u/\partial n$ denotes the derivative in the direction normal to the boundary $\partial\Omega$ (conventionally pointing outwards). Moreover we assume[1] that

$$\int_{\partial\Omega_{\mathcal{D}}} ds \neq 0,$$

so that (1.2)–(1.3) does not represent the pure Neumann condition.

Multiplying (1.1) by an admissible test function v, integrating over Ω and using the Gauss–Green theorem ([14, Appendix C.2]) we get for any test function v from the test space

$$\mathcal{H}^1_0 \equiv \{v \in H^1(\Omega);\, Tv = 0 \text{ on } \partial\Omega_{\mathcal{D}}\}$$

the equation

$$\int_\Omega \mathbf{S}\nabla u \cdot \nabla v = \int_\Omega vf + \int_{\partial\Omega_{\mathcal{N}}} v\, g_{\mathcal{N}}\,. \tag{1.4}$$

The solution u of (1.4) belongs to the solution space

$$\mathcal{H}^1_{\mathcal{D}} \equiv \{u \in H^1(\Omega);\, Tu = g_{\mathcal{D}} \text{ on } \partial\Omega_{\mathcal{D}}\} = \mathcal{H}^1_0 + u_{\mathcal{D}}\,. \tag{1.5}$$

Considering the bilinear form $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$

$$a(u,v) \equiv (\mathbf{S}\nabla u, \nabla v)_\Omega \equiv \int_\Omega \mathbf{S}\nabla u \cdot \nabla v\,,$$

and the linear functional $\ell : H^1(\Omega) \to \mathbb{R}$

$$\ell(v) \equiv (f,v)_\Omega + (g_{\mathcal{N}}, v)_{\partial\Omega_{\mathcal{N}}} \equiv \int_\Omega vf + \int_{\partial\Omega_{\mathcal{N}}} v\, g_{\mathcal{N}}\,,$$

the equation (1.4) can be restated as the *weak formulation* of (1.1)–(1.3):

Find $u \in \mathcal{H}^1_{\mathcal{D}}$ such that

$$a(u,v) = \ell(v) \quad \forall v \in \mathcal{H}^1_0\,. \tag{1.6}$$

[1] this assumption can be removed, see [6, Section 5.2]

1.3 Uniqueness of the solution, regularity

For the study of uniqueness of the weak solution and its properties we rewrite the weak formulation (1.6). Since

$$u \in \mathcal{H}^1_{\mathcal{D}} = \mathcal{H}^1_0 + u_{\mathcal{D}}\,,$$

there exists $w \in \mathcal{H}^1_0$ such that $u = w + u_{\mathcal{D}}$. Substituting into (1.6) we get

$$a(w,v) = \ell(v) - a(u_{\mathcal{D}}, v) \quad \forall v \in \mathcal{H}^1_0\,.$$

Let now consider the problem equivalent to (1.6):

Find $w \in \mathcal{H}^1_0$ such that

$$a(w,v) = \bar{\ell}(v) \equiv \ell(v) - a(u_{\mathcal{D}}, v) \quad \forall v \in \mathcal{H}^1_0\,. \tag{1.7}$$

Theorem 1.1 ([14, Section 5.8.1], Poincaré-Friedrichs inequality)**.** *Assume that Ω is a Lipschitz domain in $\mathbb{R}^d$. Then there exists a constant $C(\Omega) > 0$ depending only on Ω such that*

$$\|v\|_{0,\Omega} \le C(\Omega)\|\nabla v\|_{0,\Omega}$$

holds for any $v \in H^1(\Omega)$ satisfying

$$\int_\Omega v\,dx = 0 \quad \textit{or} \quad Tv = 0 \ \textit{on}\ \Gamma\,,$$

where $\Gamma \subset \partial\Omega$ such that $\int_\Gamma ds \neq 0$.

Using Poincaré-Friedrichs inequality

$$\begin{aligned} a(v,v) &= (\mathbf{S}\nabla v, \nabla v)_\Omega = \|\mathbf{S}^{1/2}\,\nabla v\|_{0,\Omega} \ge \\ &\ge c_{\mathbf{S}}\|\nabla v\|_{0,\Omega} \ge \frac{c_{\mathbf{S}}}{C(\Omega)}\|v\|_{0,\Omega}\,. \end{aligned}$$

With the assumption $\int_{\partial\Omega_{\mathcal{D}}} ds \neq 0$, i.e. the Dirichlet boundary condition is prescribed on a nontrivial part of $\partial\Omega$, and $Tv = 0$ on $\Omega_{\mathcal{D}}$ we get

$$a(v,v) = 0 \iff v = 0\,.$$

Thus the bilinear form $a(\cdot,\cdot)$ represents an inner product over the test space $\mathcal{H}_0^1$ and it induces there the *energy norm*,

$$\|v\|_a^2 \equiv a(v,v)^{1/2} = \|\mathbf{S}^{1/2}\nabla v\|_{0,\Omega}\,, \quad \forall v \in \mathcal{H}_0^1\,. \tag{1.8}$$

Using Theorem 1.1 we get $\forall v \in \mathcal{H}_0^1$ the inequality

$$\|\nabla v\|_{0,\Omega} \geq \frac{1}{1+C(\Omega)}(\|v\|_{0,\Omega} + \|\nabla v\|_{0,\Omega}) = c_{H^1}\|v\|_{1,\Omega}\,. \tag{1.9}$$

The uniqueness of the weak solution w of (1.7) follows from the Lax-Milgram theorem (see, e.g., [14, Section 6.2.1]) presented below. We state it for a general Hilbert space and then we present its application for (1.7).

Theorem 1.2 (Lax-Milgram)**.** *Let V be a Hilbert space and $a(\cdot,\cdot)$ a bilinear form on V, which is*

- *bounded: $a(w,v) \leq C\|w\|_V\|v\|_V\,, \quad \forall w,v \in V$ and*
- *coercive: $a(v,v) \geq c\|v\|_V^2\,, \quad \forall v \in V$*

Then, for any $\bar{\ell} \in V'$, where V' be the dual space to V, there exists a unique solution w to the equation $a(w,v) = \bar{\ell}(v)$ and

$$\|w\|_V \leq \frac{1}{c}\|\bar{\ell}\|_{V'}\,.$$

In our case $V = \mathcal{H}_0^1\,,\ H^1(\Omega) \subset V' = [\mathcal{H}_0^1]'$ with the $H^1(\Omega)$-norm $\|\cdot\|_{1,\Omega}$. Using the definition of $\|\cdot\|_{1,\Omega}$ and (1.9)

- $a(w,v) \leq \|\mathbf{S}\|\|\nabla w\|_{0,\Omega}\|\nabla v\|_{0,\Omega} \leq \|\mathbf{S}\|\|w\|_{1,\Omega}\|v\|_{1,\Omega}\,, \quad \forall w,v \in \mathcal{H}_0^1$
- $a(v,v) \geq c_{\mathbf{S}}\|\nabla v\|_{0,\Omega}^2 \geq c_{\mathbf{S}}c_{H^1}\|v\|_{1,\Omega}^2\,, \quad \forall v \in \mathcal{H}_0^1\,.$

Consequently Theorem 1.2 holds, the solution w of (1.7) is unique and

$$\|w\|_{1,\Omega} \leq c_0\|\bar{\ell}\|_{1,\Omega}\,. \tag{1.10}$$

It can be easily seen that the assumptions of Theorem 1.2 are fulfilled for any finite-dimension subspace $\mathcal{S}$ of the test space $\mathcal{H}_0^1$ so that the problem

Find $w \in \mathcal{S}$ such that

$$a(w,v) = \bar{\ell}(v) \quad \forall v \in \mathcal{S}\,, \tag{1.11}$$

has also the unique solution w. We use this property in the Galerkin discretization of (1.6).

With further assumptions on the diffusion tensor $\mathbf{S}$ we can even improve the estimate (1.10) and prove the higher regularity of the weak solution w of (1.7), see [14, Section 6.3].

Theorem 1.3 ($H^{1+\epsilon}$-regularity of the weak solution)**.** *Let w be the weak solution of (1.7). Let the diffusion tensor $\mathbf{S} = (s_{ij})_{i,j=1}^{d}$ satisfies*

$$s_{ij} \in W^{2,\infty} \quad i,j = 1,\dots,d\,.$$

Then there exists $\epsilon \in (0,1]$ such that $w \in H^{1+\epsilon}(\Omega)$ and the following inequality holds

$$\|w\|_{1+\epsilon,\Omega} \le c_1 \|\bar{\ell}\|_{1-\epsilon,\Omega}\,. \tag{1.12}$$

ϵ depends only on the shape of the Lipschitz domain Ω.

For Ω polygonal/polyhedral, ϵ depends on the greatest interior angle. For a convex Ω, we have $\epsilon = 1$ and

$$\|w\|_{2,\Omega} \le c_1 \|\bar{\ell}\|_{0,\Omega}\,. \tag{1.13}$$

The norm $\|\cdot\|_{k+\epsilon,\Omega}$ is defined as (see, e.g., [30])

$$\|v\|^2_{k+\epsilon,\Omega} = \|v\|^2_{k,\Omega} + \sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x-y|^{d+2\epsilon}}\,dx\,dy$$

1.4 Galerkin discretization

In this section we describe the idea of the Galerkin FEM discretization, its basic properties and, as an example, the P^1-conforming FEM discretization.

The Galerkin approximation to (1.6) can be briefly summarized in the following way. Let $\mathcal{S}_0^h \subset \mathcal{H}_0^1$ be a finite-dimensional subspace of the test space and let the problem on $\mathcal{S}_0^h$ is considered

Find $u_h \in \mathcal{S}_{\mathcal{D}}^h \equiv u_{\mathcal{D}} + \mathcal{S}_0^h$ such that

$$a(u_h, v_h) = \ell(v_h) \quad \forall v_h \in \mathcal{S}_0^h\,. \tag{1.14}$$

Theorem 1.2 assures the existence and the uniqueness of the solution u_h.

Let $\{\phi_1, \ldots, \phi_n\}$ form the basis of $\mathcal{S}_0^h$. Then the finite element approximation $u_h \in \mathcal{S}_{\mathcal{D}}^h$ can be written in the form

$$u_h = \sum_{j=1}^{n} \zeta_j \phi_j + u_{\mathcal{D}} \,, \tag{1.15}$$

By substituting (1.15) into (1.14) and taking $v_h = \phi_i$, $i = 1, 2, \ldots, n$, we get the system of linear algebraic equations

$$Ax = b, \tag{1.16}$$

with

$$\begin{aligned} A &= [a_{ij}] \;, \quad a_{ij} = \int_\Omega \mathbf{S} \nabla \phi_i \cdot \nabla \phi_j \;, \\ b &= \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \;, \quad b_i = \int_\Omega \phi_i f + \int_{\partial\Omega_{\mathcal{N}}} \phi_i g_{\mathcal{N}} - \int_\Omega \nabla \phi_i \cdot \nabla u_{\mathcal{D}} \;, \\ x &= \begin{bmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{bmatrix} . \end{aligned}$$

The system (1.16) is called the Galerkin system, the symmetric matrix A is called the stiffness matrix and the solution (1.15) obtained by solving (1.16) is called the Galerkin solution. In order to have the arisen matrix A as sparse as possible (for effective solving of (1.16)), we usually consider basis functions ϕ_i of small support.

Using the Poincaré-Friedrichs inequality (Theorem 1.1) we showed that the bilinear form $a(\cdot, \cdot)$ represents the inner product over the test space $\mathcal{H}_0^1$ and induces there the energy norm. Consequently the (symmetric) matrix A is positive-definite.

Consider two arbitrary functions $\tilde{w}_h, \tilde{z}_h \in \mathcal{S}_0^h$. Denoting by

$$\Phi = [\phi_1, \phi_2, \ldots, \phi_n]$$

we can write these functions as $\tilde{w}_h = \Phi\, x$, $\tilde{z}_h = \Phi\, y$ for some coefficient vectors $x, y \in \mathbb{R}^n$. Then the inner product of $\tilde{w}_h, \tilde{z}_h$ is given by

$$a(\tilde{w}_h, \tilde{z}_h) = \int_\Omega \mathbf{S} \nabla \tilde{w}_h \cdot \nabla \tilde{z}_h = (x, Ay) \;.$$

Therefore, for any $\tilde{w}_h \in \mathcal{S}_0^h$,

$$\|x\|_A^2 \equiv (x, Ax) = a(\tilde{w}_h, \tilde{w}_h) = \|\tilde{w}_h\|_a^2 \tag{1.17}$$

defines the *algebraic energy norm* on $\mathbb{R}^n$.

We point out that in our considerations we have excluded for clarity of exposition the pure Neumann boundary conditions. Moreover, we will further assume that the Galerkin solution u_h is *conforming*, i.e. the Dirichlet boundary conditions are interpolated by $u_{\mathcal{D}}$ exactly, hence for any $\tilde{u}_h \in \mathcal{S}_{\mathcal{D}}^h$

$$u - \tilde{u}_h \in \mathcal{S}_0^h \,. \tag{1.18}$$

For any Galerkin approximation u_h given by (1.15) and (1.16) and for any $v_h \in \mathcal{S}_0^h$,

$$a(u, v_h) = \ell(v_h) \quad \text{and} \quad a(u_h, v_h) = \ell(v_h)$$

give the *Galerkin orthogonality condition*

$$a(u - u_h, v_h) = 0 \quad \text{for all} \quad v_h \in \mathcal{S}_0^h \,, \tag{1.19}$$

i.e. the discretization error $u - u_h$ is orthogonal to the subspace $\mathcal{S}_0^h$ with respect to the energy inner product. Since

$$u - u_h = (u - u_{\mathcal{D}}) - \sum_{j=1}^{N} \zeta_j \phi_j \,,$$

$u - u_h$ can be seen as the result of the orthogonalization of $u - u_{\mathcal{D}}$ against $\phi_1, \phi_2, \ldots, \phi_n$ in the energy inner product. Subsequently $u - u_h$ must have the minimal energy norm among all possible vectors $u - w_h$ for $w_h \in \mathcal{S}_{\mathcal{D}}^h$. As an immediate consequence of conforming approximations (1.18) and the Galerkin orthogonality (1.19) we therefore get the best approximation property of u_h in the energy norm (1.8), see, e.g., [13, Theorem 8.1]).

$$\|u - u_h\|_a = \min_{w_h \in \mathcal{S}_{\mathcal{D}}^h} \|u - w_h\|_a \,. \tag{1.20}$$

1.4.1 P^1-conforming FEM discretization

Now we present the simplest FEM discretization using piecewise linear continuous functions.

Let T_h be the partition of polygonal/polyhedral domain Ω into closed simplices i.e. triangles (d = 2) or tetrahedra (d = 3) such that $\bar{\Omega} = \bigcup_{K \in T_h} K$. We assume that if $K, L \in T_h$, $K \neq L$, then $K \cap L$ is either an empty set, a common face, edge or vertex of K and L. By h we denote the mesh size

$$h \equiv \max_{K \in T_h} \operatorname{diam}(K).$$

We consider the finite element space

$$\mathcal{S}^{h,P^1} \equiv \{u \in C(\Omega); u|_K \in P^1(K) \quad \forall K \in T_h\}$$

and the appropriate test and solution spaces $\mathcal{S}_0^{h,P^1} \subset \mathcal{S}^{h,P^1}$, $\mathcal{S}_{\mathcal{D}}^{h,P^1} \subset \mathcal{S}^{h,P^1}$ respectively.

We consider the $\mathcal{S}_0^{h,P^1}$-basis functions ϕ_i of small support, function ϕ_i corresponds to a single vertex Z_i of the triangulation T_h and

$$\phi_i(Z_j) = \delta_{ij}\,, \quad i, j = 1, \ldots n.$$

Here δ_{ij} stands for Kronecker delta. The illustration of the basis function ϕ_i is given in Figure 1.1. Consequently $a_{ij} \neq 0$ only for neighboring vertices Z_i, Z_j, i.e. vertices of the same element K of triangulation T_h.

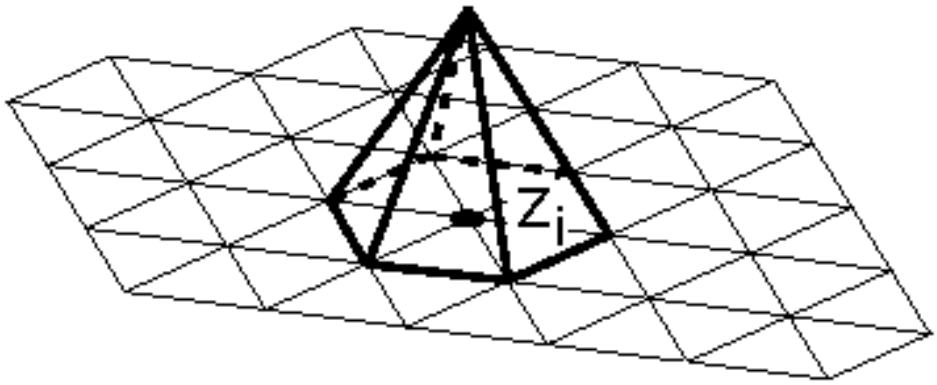


Figure 1.1: *Basis function ϕ_i .*

In the previous section we showed the equality of the energy norm (1.8) and the algebraic energy norm (1.17)

$$\|\tilde{v}_h\|_a = \|x\|_A\,, \quad \tilde{v}_h = \Phi x\,.$$

There exists constants $c_2 > 0, c_3 > 0$ such that the Euclidean norm $\|x\|$ and the L^2-norm $\|\tilde{v}_h\|_{0,\Omega}$ fulfill the inequality with factor h (see [40, relation 5.12])

$$c_2 h\|x\| \leq \|\tilde{v}_h\|_{0,\Omega} \leq c_3 h\|x\|\,, \quad \forall \tilde{v}_h \in \mathcal{S}_0^h\,; \tilde{v}_h = \Phi\, x\,. \tag{1.21}$$

The convergence of the Galerkin approximation u_h to the exact solution u of the (1.6) has been studied in many papers. For the problem (1.6) using the P^1-conforming FEM discretization the following estimates hold (see, e.g., [9, Section 3.2])

Theorem 1.4 (*a priori* estimates)**.** *Let (1.14) be the weak formulation of problem (1.1)–(1.3) with assumptions* ***A1*** *–* ***A5****. Then there exists the unique solution* $u_h \in \mathcal{S}_{\mathcal{D}}^{h,P^1}$ *of (1.14) and obeys the estimates*

$$\|u - u_h\|_a \;\leq\; c_4 h\ \|\ell\|_{0,\Omega}\,, \tag{1.22}$$

$$\|u - u_h\|_{0,\Omega} \;\leq\; c_5 h^2 \|\ell\|_{0,\Omega}\,. \tag{1.23}$$

However, the inequalities (1.21), (1.22) and (1.23) give only theoretical bounds due to unspecified constants c_2, c_3, c_4 and c_5. For computations we need reliable and computable bounds for the errors. Such bounds are presented in Chapter 3.

Chapter 2

Error estimates and stopping criteria in Conjugate Gradient Method (CG)

We start the introduction to CG by its derivation (following [26]) in order to present some important properties and relationship to other mathematical disciplines. We show the consistency of CG and Galerkin discretization and the energy interpretation of the A-norm. Then the estimates of the CG error are presented using the relations of CG with the Lanczos method and Gauss quadrature, and a simple heuristic for an adaptive estimate is proposed (following [33]). Finally, we present stopping criterion for CG using the proposed heuristic. This topic will be further developed in Chapter 5.

2.1 Derivation of CG

Derivation of CG historically proceeded in a different way but the following one very naturally illustrates the properties of CG. It is motivated by the standard optimization approach to CG and employs the well-known equivalence between solving the linear algebraic system

$$Ax = b \ ,$$

where $A \in \mathbb{R}^{n \times n}$ is real symmetric positive-definite (SPD) matrix and $b \in \mathbb{R}^n$ is a right hand side vector, and minimizing the quadratic functional

$$F(x) = \frac{1}{2}(x, Ax) - (x, b) \ .$$

Defining the A-norm

$$\|z\|_A = (z, Az)^{1/2}$$

and considering x_k an approximation to the solution x, the equality

$$\begin{aligned} F(x_k) &= \tfrac{1}{2}\left((x - x_k), A(x - x_k)\right) - \tfrac{1}{2}(x, Ax) = \\ &= \tfrac{1}{2}\|x - x_k\|_A^2 - \tfrac{1}{2}\|x\|_A^2 \ , \end{aligned}$$

shows that the minimization of $F(z)$ over some subspace of $\mathbb{R}^N$ is mathematically equivalent to minimization of $\|x - z\|_A$ over the same subspace. Therefore the A-norm is natural to measure the error of the approximate solution. (In Section 2.2.2 we will further elaborate on the relevance of the A-norm of the error.)

Let x_0 be an initial approximation, and let the sequence of approximations to the solution x be constructed by the simple recurrence

$$x_k = x_{k-1} + \gamma_{k-1} p_{k-1} \ , \quad k = 1, 2, \ldots$$

where p_{k-1} represents the search direction at the step k. The new approximation x_k is determined as the point along the line $x_{k-1} + \gamma_{k-1} p_{k-1}$ where $\|x - z\|_A$ is minimimal (i.e. the minimum of $F(z)$). By simple calculation

$$\|x - x_k\|_A^2 = \|x - x_{k-1}\|_A^2 - 2\gamma_{k-1}(p_{k-1}, r_{k-1}) + \gamma_{k-1}^2 (p_{k-1}, Ap_{k-1}) \ ,$$

so the minimum is attained for

$$\gamma_{k-1} = \frac{(p_{k-1}, r_{k-1})}{(p_{k-1}, Ap_{k-1})} \ . \tag{2.1}$$

As an immediate consequence we get the orthogonality of the residual r_k (corresponding to the new minimum x_k) to the search vector p_{k-1},

$$(p_{k-1}, r_k) = (p_{k-1}, (r_{k-1} - \gamma_{k-1} A p_{k-1})) = 0 \ ,$$

which geometrically means that the gradient $\nabla F(x_k)$ at x_k is orthogonal to the equipotential surface determined by the equation $F(y) = F(x_k)$.

It remains to determine the search directions. The simplest choice of the initial direction p_0 is $p_0 \equiv r_0 = b - Ax_0$. If we take $p_k \equiv r_k$, where r_k is the residual at k-th step $r_k = b - Ax_k$, also in the subsequent steps, $k = 1, 2, \ldots$, we get the method of steepest descent (see, e.g, [37, Section 5.3]), which can exhibit a rather poor convergence, because at each step the norm of

the error is minimized only over one-dimensional space determined by r_k. In order to get minimization property over more dimensional subspaces, we must combine in the choice of p_k the information from several iteration steps. The simplest choice generates the new search direction as a combination of the previous search direction and the (new) residual,

$$p_k = r_k + \delta_k p_{k-1}, \text{ for some } \delta_k \in \mathbb{R} \ .$$

In order to motivate the choice of δ_k below, we first notice that the change of the error from the step $k-1$ to k

$$x - x_k = (x - x_{k-1}) - \gamma_{k-1} p_{k-1}$$

with the value

$$\gamma_{k-1} = \frac{(p_{k-1}, r_{k-1})}{(p_{k-1}, Ap_{k-1})} = \frac{(p_{k-1}, A(x - x_{k-1}))}{(p_{k-1}, Ap_{k-1})} \ ,$$

can be regarded as the A-orthogonalization (i.e. the orthogonalization with the respect to the inner product defined by (z, Ay), $z, y \in \mathbb{R}^n$) of the error $x - x_{k-1}$ against p_{k-1}. Then

$$(x - x_{k-1}) = \gamma_{k-1} p_{k-1} + (x - x_k)$$

can be interpreted as the A-orthogonal decomposition of $x - x_{k-1}$. The Pythagorean theorem then gives

$$\|x - x_{k-1}\|_A^2 = |\gamma_{k-1}|^2 \|p_{k-1}\|_A^2 + \|x - x_k\|_A^2 \ .$$

The recursive application of this gives

$$x - x_0 = \sum_{j=1}^{k} \gamma_{j-1} p_{j-1} + (x - x_k) \ .$$

and

$$\|x - x_0\|_A^2 = \sum_{j=1}^{k} |\gamma_{j-1}|^2 \|p_{j-1}\|_A^2 \ + \ \|x - x_k\|_A^2 \, . \tag{2.2}$$

Now *assume* that all search directions $p_0, p_1, \ldots$ are A-orthogonal, i.e.

$$(p_i, Ap_j) = 0, \ i \neq j$$

holds. Then

$$x - x_k = (x - x_0) - \sum_{j=1}^{k} \gamma_{j-1} p_{j-1}$$

represents the A-orthogonal decomposition of $x - x_0$, and, as a consequence, $\|x - x_k\|_A$ is minimal over all possible approximations in the subspace generated by the search directions $p_0, \ldots, p_{k-1}$,

$$\|x - x_k\|_A = \min_{u \in x_0 + \mathrm{span}\{p_0, \ldots, p_{k-1}\}} \|x - u\|_A \ . \tag{2.3}$$

Moreover, the assumed A-orthogonality of p_j, $j = 0, 1, \ldots$ implies $p_n = 0$, i.e., the algorithm would reach the true solution (in exact arithmetic) in at most n steps.

Having only one undetermined scalar coefficient δ_k, we can't get closer to desired A-orthogonality of the search directions, than requiring local A-orthogonality of the two *subsequent* search directions, i.e.,

$$(p_{k-1}, Ap_k) = 0 \, ,$$

which gives

$$\delta_k = -\frac{(p_{k-1}, Ar_k)}{(p_{k-1}, Ap_{k-1})} \, .$$

Now the algorithm is fully determined. In Section 2.2 we will prove that

$$(r_i, r_j) = 0 \quad \text{and} \quad (p_i, Ap_j) = 0, \quad i \neq j \ ,$$

for $r_j \equiv b - Ax_j$. Consequently, the local A-orthogonality of p_k and p_{k-1} guarantees the global orthogonality of residuals (with respect to the standard Euclidean inner product) and global A-orthogonality of all search directions. So the orthogonality assumption mentioned prior to the minimization property (2.3) is satisfied.

Finally, using

$$(p_{k-1}, r_{k-1}) = (r_{k-1}, r_{k-1})$$

and

$$-Ap_{k-1} = \frac{r_k - r_{k-1}}{\gamma_{k-1}} = \frac{(r_k - r_{k-1})(p_{k-1}, Ap_{k-1})}{(p_{k-1}, r_{k-1})} \ .$$

we get

$$\delta_k = \frac{(r_k, r_k)}{(r_{k-1}, r_{k-1})} \ , \quad \gamma_{k-1} = \frac{(r_{k-1}, r_{k-1})}{(p_{k-1}, Ap_{k-1})} \ ,$$

that leads to the standard algorithm implementation of the CG method, see [21].

2.2 CG and its properties

Consider the linear algebraic system

$$Ax = b\,, \tag{2.4}$$

where $A \in \mathbb{R}^{n\times n}$ is a SPD matrix and $b \in \mathbb{R}^n$ a right hand side vector. The CG algorithm was first derived in [21], in our notation

Given $x_0,\ r_0 = b - Ax_0,\ p_0 = r_0$ for $i = 1, 2, \ldots$

$$\begin{aligned}
&\gamma_{i-1} = (r_{i-1}, r_{i-1})/(p_{i-1}, Ap_{i-1})\ ,\\
&x_i = x_{i-1} + \gamma_{i-1} p_{i-1}\ ,\\
&r_i = r_{i-1} - \gamma_{i-1} A p_{i-1}\ ,\\
&\delta_i = (r_i, r_i)/(r_{i-1}, r_{i-1})\ ,\\
&p_i = r_i + \delta_i p_{i-1}\ .
\end{aligned} \tag{2.5}$$

We will recall some important properties.

2.2.1 Orthogonality

Theorem 2.1 ([21, Theorem 5.1], orthogonality in CG)**.** *Residual vectors r_i and search directions p_i determined by the CG algorithm satisfy*

$$\begin{aligned}
(r_i, r_j) &= 0 \quad i \neq j\ ,\\
(p_i, Ap_j) &= 0 \quad i \neq j\ ,\\
(p_i, r_j) &= 0 \quad i < j\ , \quad (p_i, r_j) = \|r_i\|^2 \quad i \geq j\ ,\\
(r_i, Ap_i) &= (p_i, Ap_i)\ , \quad (r_i, Ap_j) = 0 \quad i \neq j,\ i \neq j+1\ .
\end{aligned}$$

From the definition of p_i, r_i it follows that

$$\begin{aligned}
\mathcal{K}_k(A, r_0) &\equiv \operatorname{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}\\
&= \operatorname{span}\{p_0, p_1, \ldots, p_{k-1}\}\\
&= \operatorname{span}\{r_0, r_1, \ldots, r_{k-1}\}, \quad \forall k = 0, 1, \ldots
\end{aligned} \tag{2.6}$$

$\mathcal{K}_k(A, r_0)$ is called k^{th} *Krylov subspace* generated by A and the initial residual $r_0 = b - Ax_0$. From Theorem 2.1 one can see that $\{r_0, r_1, \ldots, r_{k-1}\}$ form orthogonal basis and $\{p_0, p_1, \ldots, p_{k-1}\}$ A-orthogonal basis of $\mathcal{K}_k(A, r_0)$.

2.2.2 CG and the Galerkin discretization

Very important characteristic of CG is the energy error minimizing property,

$$\|x - x_k\|_A = \min_{u \in x_0 + \text{span}\{p_0,\ldots,p_{k-1}\}} \|x - u\|_A \ ,$$

using previous notation

$$\|x - x_k\|_A = \min_{u \in x_0 + \mathcal{K}_k(A,r_0)} \|x - u\|_A \ . \tag{2.7}$$

Following [26] (and resuming Chapter 1) we will show the consistency of CG and Galerkin finite element method. For details and references see [26, 12].

Consider the problem (1.1)–(1.3) and its weak formulation (1.6) derived in Section 1.2. Consider the Galerkin discretization given in Section 1.4, with the system of linear algebraic equations

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n \tag{2.8}$$

and the Galerkin approximation

$$u_h = \Phi\, x + u_{\mathcal{D}}\,,$$

corresponding to the *exact* solution x of the system (2.8).

Let $x_k \in \mathbb{R}^n$ represents an approximation to x and

$$u_h^{(k)} \equiv \Phi\, x_k + u_{\mathcal{D}} \ \in \mathcal{S}_D^h\,. \tag{2.9}$$

By construction and under the conformity assumption

$$u_h - u_h^{(k)} \in \mathcal{S}_0^h\,.$$

(Recall that $\mathcal{S}_{\mathcal{D}}^h, \mathcal{S}_0^h$ are the finite-dimensional subspaces of the solution and test space respectively, $\mathcal{S}_{\mathcal{D}}^h = \mathcal{S}_0^h + u_{\mathcal{D}}$, the columns of Φ form the basis of $\mathcal{S}_0^h$ and $u_{\mathcal{D}}$ interpolates the Dirichlet boundary condition exactly.)

Consider the error of the computed approximate solution $u_h^{(k)}$ in the energy norm $\|u - u_h^{(k)}\|_a$ defined in (1.8). The error $u - u_h^{(k)}$ consists of two parts:

- the discretization error $u - u_h$, with its energy norm defined in (1.8) given by
$$\|u - u_h\|_a\,;$$

- the algebraic error $u_h - u_h^{(k)}$, which can be measured using the induced algebraic energy norm (see (1.17)) as

$$\begin{aligned} \|u_h - u_h^{(k)}\|_a^2 &= ((x - x_k), A(x - x_k)) \\ &= \|x - x_k\|_A^2 . \end{aligned}$$

The relationship between the size of the total error, the Galerkin discretization error and the algebraic error is summarized in the following theorem ([12]).

Theorem 2.2 (Consistency of CG and Galerkin FEM)**.** *Let u_h be the Galerkin approximation of the solution u of the problem (1.1)–(1.3) and let $u_h^{(k)}$ given by (2.9) corresponds to the approximate solution x_k of the linear algebraic system (2.8). Assuming the conformity (1.18),*

$$\|u - u_h^{(k)}\|_a^2 = \|u - u_h\|_a^2 + \|x - x_k\|_A^2 .$$

Proof. A simple manipulation gives

$$\begin{aligned} \|u - u_h^{(k)}\|_a^2 &= a(u - u_h^{(k)}, u - u_h^{(k)}) \\ &= a(u - u_h + u_h - u_h^{(k)}, u - u_h + u_h - u_h^{(k)}) \\ &= a(u - u_h, u - u_h) + a(u_h - u_h^{(k)}, u_h - u_h^{(k)}) , \end{aligned}$$

since $a(u - u_h, u_h - u_h^{(k)}) = 0$ due to the Galerkin orthogonality condition (1.19) as $u_h - u_h^{(k)} \in \mathcal{S}_0^h$. The rest of the proof follows trivially by rewriting the terms in the form of the energy norms. □

Theorem 2.2 has the following interpretation. Consider the approximate solution x_k from a given algebraic subspace $\mathcal{G} \subset \mathbb{R}^n$, that results through (2.9) in $u_h^{(k)}$; an element of the corresponding functional subspace $\mathcal{F} \subset \mathcal{S}_{\mathcal{D}}^h$ determined by $\mathcal{G}$. In order to get the best possible approximation over $\mathcal{F}$ measured in the functional energy norm, the distance of x_k to the exact solution x of the system (1.16) measured by the algebraic energy norm (i.e. $\|x - x_k\|_A$) has to be minimized over $\mathcal{G}$.

If we use Krylov subspace methods (see, e.g., [37]) for solving (1.16), the algebraic subspace $\mathcal{G}$ is the Krylov subspace (2.6). Then the best approximation $u_h^{(k)}$ of the solution of (1.1)–(1.3) in the corresponding subspace $\mathcal{F}_k$

determined by the Krylov subspace $\mathcal{K}_k(A, r_0)$ is given by the CG method. This shows the consistency of CG with the Galerkin FEM discretization of the continuous problem which is unique among all iterative methods seeking the approximate solution of the linear algebraic system in the given Krylov subspace.

Theorem 2.2 has also another important meaning. When solving PDE problems, our goal should seek a balance between the algebraic and the discretization parts of the error. Solving arisen linear systems too accurately (which means often too costly) cannot change the order of the total error (as we see from Theorem 2.2). We discuss this further in Section 3.1 where we look for the way how to estimate the discretization (and total) error. Estimates for the algebraic energy norm are described in Section 2.3.

2.2.3 Lanczos Method and its relationship to CG

In this subsection we present following [43], where the references to the original sources can be found the well-known relationship between CG and the the Lanczos method.

For A and r_0 the Lanczos method generates (ideally, see [28] for the finite precision behaviour) an orthonormal basis $\{v_1, v_2, \ldots, v_k\}$ of the Krylov subspace $\mathcal{K}_k(A, r_0)$ via the recurrence:

Given $v_1 = r_0/\|r_0\|,\ v_0 \equiv 0,\ \beta_1 \equiv 0$, for $i = 1, 2, \ldots$

$$\begin{aligned} \alpha_i &= (Av_i - \beta_i v_{i-1}, v_i) \ , \\ w_i &= Av_i - \alpha_i v_i - \beta_i v_{i-1} \ , \\ \beta_{i+1} &= \|w_i\| \ , \\ v_{i+1} &= w_i/\beta_{i+1} \ . \end{aligned} \tag{2.10}$$

Comparing CG (2.5) with (2.10) gives

$$v_{i+1} = (-1)^i \frac{r_i}{\|r_i\|} \, . \tag{2.11}$$

Denoting by $V_i \equiv [v_1, \ldots, v_i] \in \mathbb{R}^{n \times i}$ the matrix having the Lanczos vectors v_i as its columns and by T_i the symmetric tridiagonal matrix with positive subdiagonals

$$T_i \equiv \begin{pmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_i \\ & & \beta_i & \alpha_i \end{pmatrix} , \tag{2.12}$$

equations (2.10) can be rewritten in the matrix form

$$AV_i = V_i T_i + \beta_{i+1} v_{i+1} e_i^T \,, \tag{2.13}$$

where e_i is i-th column of the identity matrix $I \in \mathbb{R}^{i \times i}$. Using the change of variables

$$x_i = x_0 + V_i \, y_i \,. \tag{2.14}$$

Using the orthogonality relation between r_i and the basis vectors $\{v_1, \ldots, v_i\}$ of $\mathcal{K}_i(A, r_0)$, we get

$$\begin{aligned} 0 = V_i^T r_i = V_i^T (b - A x_i) = V_i^T (r_0 - A V_i y_i) = \\ = e_1 \|r_0\| - V_i^T A V_i y_i = e_1 \|r_0\| - T_i y_i \,. \end{aligned}$$

Consequently, the CG approximation x_i is determined by (2.14) and by solving

$$T_i y_i = \|r_0\| e_1 \,. \tag{2.15}$$

2.2.4 Orthogonal polynomials and Gauss quadrature

The relation of CG to orthogonal polynomials and the Gauss quadrature is the key for understanding both mathematical properties and the highly nonlinear behaviour of the CG method. The connection of CG to the Gauss quadrature was pointed out in the original paper [21], with the orthogonality of CG residuals as a link to orthogonal polynomials. In our exposition we follow [43].

Using (2.5), the i-th error resp. residual can be written as a polynomial in the matrix A applied to the initial error resp. residual,

$$x - x_i = \varphi_i(A) \, (x - x_0), \quad r_i = \varphi_i(A) \, r_0, \quad \varphi_i \in \Pi_i \,, \tag{2.16}$$

where Π_i denotes the class of polynomials of degree at most i having the property $\varphi(0) = 1$ (the constant term is equal to one). Consider the eigen-decomposition of the symmetric matrix A in the form

$$A = U \Lambda U^T, \quad U U^T = U^T U = I, \tag{2.17}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $U = [u_1, \ldots, u_n]$ is the matrix having the normalized eigenvectors of A as its columns. Substituting (2.17) and (2.16)

into the CG minimizing property (2.7) gives

$$\begin{aligned}\|x - x_i\|_A &= \|\varphi_i(A)(x - x_0)\|_A = \min_{\varphi\in\Pi_i} \|\varphi(A)(x - x_0)\|_A = \\ &= \min_{\varphi\in\Pi_i} \|\varphi(A)r_0\|_{A^{-1}} = \min_{\varphi\in\Pi_i} \left\{ \sum_{k=1}^{n} \frac{(r_0, u_k)^2}{\lambda_k} \varphi^2(\lambda_k) \right\}^{1/2} \end{aligned} \tag{2.18}$$

Consequently, for A SPD the rate of convergence of CG is determined by the distribution of its eigenvalues *and* by the size of components of the initial residual r_0 in the direction of the individual eigenvectors.

Similarly to (2.16), Lanczos vector v_{i+1} is linked with a monic polynomial ψ_i,

$$v_{i+1} = \psi_i(A)\, v_1 \cdot \frac{1}{\beta_2\beta_3 \dots \beta_{i+1}}\,. \tag{2.19}$$

Using the orthogonality of v_{i+1} to $v_1, \dots, v_i$, the polynomial ψ_i is determined by the minimizing condition

$$\|\psi_i(A)v_1\| = \min_{\psi\in\mathcal{M}_i} \|\psi(A)v_1\| = \min_{\psi\in\mathcal{M}_i} \left\{ \sum_{k=1}^{n} (v_1, u_k)^2\, \psi^2(\lambda_k) \right\}^{1/2}, \tag{2.20}$$

where $\mathcal{M}_i$ denotes the class of monic polynomials of degree i.

We showed that the CG residuals or the Lanczos basis vectors (defined by (2.5) resp. by (2.10)) can be linked with a sequence $1, \psi_1, \psi_2, \dots$ of the monic orthogonal polynomials determined by (2.20). These polynomials are orthogonal with respect to the discrete inner product

$$(f, g) = \sum_{k=1}^{n} \omega_k f(\lambda_k) g(\lambda_k)\,, \tag{2.21}$$

where the weights ω_k are determined as

$$\omega_k = (v_1, u_k)^2, \quad \sum_{k=1}^{n} \omega_k = 1\,, \quad \left(v_1 = \frac{r_0}{\|r_0\|} \right). \tag{2.22}$$

For simplicity of notation we assume that the eigenvalues of A are distinct and increasingly ordered (an extension to the case of multiple eigenvalues will be obvious). Let ζ, ξ be such that

$$\zeta < \lambda_1 < \lambda_2 < \ldots < \lambda_n < \xi\,.$$

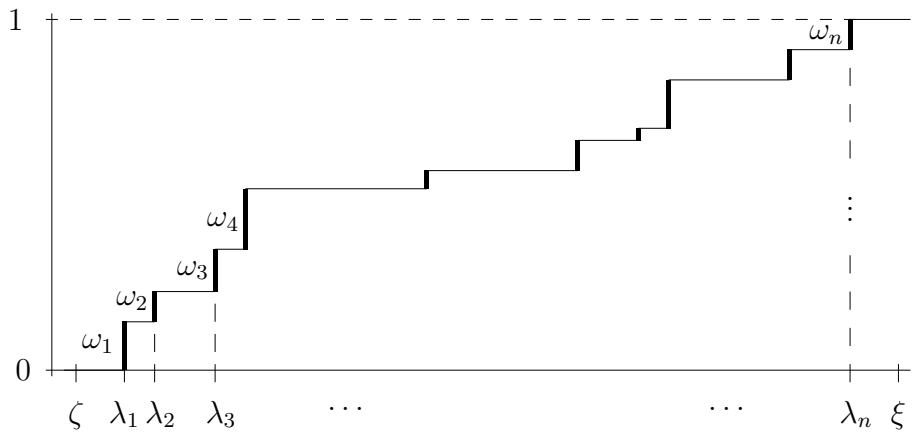


Figure 2.1: *Distribution function* $\omega(\lambda)$

Consider the distribution function $\omega(\lambda)$ with the finite points of increase $\lambda_1, \lambda_2, \ldots, \lambda_n$,

$$\begin{aligned} &\omega(\lambda) = 0 && \text{for} && \zeta \le \lambda < \lambda_1\,, \\ &\omega(\lambda) = \sum_{k=1}^{j} \omega_k && \text{for} && \lambda_j \le \lambda < \lambda_{j+1}\,, \\ &\omega(\lambda) = 1 && \text{for} && \lambda_n \le \lambda \le \xi\,, \end{aligned} \tag{2.23}$$

see Figure 2.1, and the corresponding Riemann-Stieltjes integral

$$\int_{\zeta}^{\xi} f(\lambda)\, d\omega(\lambda) = \sum_{k=1}^{n} \omega_k f(\lambda_k)\,. \tag{2.24}$$

Then (2.20) can be rewritten as

$$\psi_i = \arg \min_{\psi \in \mathcal{M}_i} \left\{ \int_{\zeta}^{\xi} \psi^2(\lambda)\, d\omega(\lambda) \right\}, \quad i = 0, 1, 2, \ldots, n\,. \tag{2.25}$$

Consider, analogously to (2.17), the eigendecomposition of the symmetric tridiagonal matrix T_i in the form

$$T_i = S_i \Theta_i S_i^T, \quad S_i^T S_i = S_i S_i^T = I, \tag{2.26}$$

$\Theta_i = \mathrm{diag}(\theta_1^{(i)}, \ldots, \theta_i^{(i)})$, $S_i = [s_1^{(i)}, \ldots, s_i^{(i)}]$. T_i is determined by the i steps of the CG or Lanczos method for matrix A starting with $\|r_0\| v_1$ resp. v_1. It

can be also regarded as determined by the CG applied to the i-dimensional problem $T_i y_i = e_1\|r_0\|$ with the initial residual $e_1\|r_0\|$ (resp. by the Lanczos method for T_i with the starting vector e_1). The eigenvalues of T_i are called Ritz values and they are distinct (see e.g. [34, Chapter 7]). Obviously, we can construct a Riemann-Stieltjes integral for this i-dimensional problem similarly as above. Let

$$\zeta < \theta_1^{(i)} < \theta_2^{(i)} < \ldots < \theta_i^{(i)} < \xi$$

and

$$\omega_k^{(i)} = (e_1, s_k^{(i)})^2, \quad \sum_{k=1}^{i} \omega_k^{(i)} = 1 \tag{2.27}$$

be the weights determined by the squared size of the components of e_1 in the direction of the eigenvectors of T_i, and

$$\begin{aligned} &\omega^{(i)}(\lambda) = 0 && \text{for} \quad \zeta \le \lambda < \theta_1^{(i)}\,, \\ &\omega^{(i)}(\lambda) = \sum_{k=1}^{j} \omega_k^{(i)} && \text{for} \quad \theta_j^{(i)} \le \lambda < \theta_{j+1}^{(i)}, \\ &\omega^{(i)}(\lambda) = 1 && \text{for} \quad \theta_i^{(i)} \le \lambda \le \xi\,. \end{aligned}$$

Then the first i polynomials from the set $\{1, \psi_1, \ldots, \psi_n\}$ determined by (2.25) are also determined by the condition based on the Riemann-Stieltjes integral with the distribution function $\omega^{(i)}(\lambda)$

$$\psi_l = \arg \min_{\psi \in \mathcal{M}_l} \left\{ \int_\zeta^\xi \psi^2(\lambda)\, d\omega^{(i)}(\lambda) \right\}, \quad l = 0, 1, \ldots, i\,. \tag{2.28}$$

(we can look at the sequence $\{1, \psi_1, \ldots, \psi_i$ as determined by CG or the Lanczos method to the i-dimensional problem described above.) The integral

$$\int_\zeta^\xi f(\lambda)\, d\omega^{(i)}(\lambda) = \sum_{k=1}^{i} \omega_k^{(i)} f(\theta_k^{(i)}) \tag{2.29}$$

is the i-th Gauss quadrature approximation of the integral (2.24), see, e.g., [17]. Thus, the CG and Lanczos methods determine the sequence of distribution functions $\omega^{(1)}(\lambda), \omega^{(2)}(\lambda), \ldots, \omega^{(i)}(\lambda), \ldots$ approximating in the sense of the Gauss quadrature the original distribution function $\omega(\lambda)$ (i.e. the value of the original integral (2.24) is approximated by (2.29) exactly for any polynomial of degree less than of equal to $2i - 1$).

With $f(\lambda) = \lambda^{-1}$ we have from (2.18)

$$\|x - x_0\|_A^2 = \|r_0\|^2 \sum_{k=1}^{n} \frac{\omega_k}{\lambda_k} = \|r_0\|^2 \int_\zeta^\xi \lambda^{-1} \, d\omega(\lambda) \,, \tag{2.30}$$

and, using (2.13) with $i = n$, i.e. $AV_n = V_n T_n$,,

$$\|x - x_0\|_A^2 = (r_0, A^{-1} r_0) = \|r_0\|^2 (e_1, T_n^{-1} e_1) \equiv \|r_0\|^2 \, (T_n^{-1})_{11} \,.$$

Consequently,

$$\int_\zeta^\xi \lambda^{-1} \, d\omega(\lambda) = (T_n^{-1})_{11} \,. \tag{2.31}$$

Repeating the same considerations using the CG method for T_i with the initial residual $\|r_0\| e_1$, (or the Lanczos method for T_i with e_1)

$$\int_\zeta^\xi \lambda^{-1} \, d\omega^{(i)}(\lambda) = (T_i^{-1})_{11} \,. \tag{2.32}$$

Applying the i-point Gauss quadrature to (2.24) gives

$$\int_\zeta^\xi f(\lambda) \, d\omega(\lambda) = \int_\zeta^\xi f(\lambda) \, d\omega^{(i)}(\lambda) + R_i(f), \tag{2.33}$$

where $R_i(f)$ denotes the (approximation) error in the Gauss quadrature. For $f(\lambda) = \lambda^{-1}$ this gives

$$\|x - x_0\|_A^2 = \|r_0\|^2 (T_n^{-1})_{11} = \|r_0\|^2 (T_i^{-1})_{11} + \|r_0\|^2 R_i(\lambda^{-1}) \,. \tag{2.34}$$

In [20] it was proved that

$$R_i(\lambda^{-1}) = \frac{\|x - x_i\|_A^2}{\|r_0\|^2} \,.$$

Then (2.33) can be rewritten

$$\|x - x_0\|_A^2 = \|r_0\|^2 (T_i^{-1})_{11} + \|x - x_i\|_A^2 \,. \tag{2.35}$$

Using (2.15) and $e_1^T = v_1^T V_i$ resulting from the *global orthogonality* of $v_1, \ldots, v_i$,

$$\begin{aligned} \|r_0\|^2 (T_i^{-1})_{11} &= \|r_0\| \, e_1^T T_i^{-1} e_1 \, \|r_0\| \\ &= \|r_0\| \, v_1^T V_i T_i^{-1} e_1 \, \|r_0\| = (\|r_0\| \, v_1)^T (V_i T_i^{-1} e_1 \, \|r_0\|) \\ &= r_0^T (x_i - x_0) \,. \end{aligned} \tag{2.36}$$

Hence using (2.35),

$$\|x - x_0\|_A^2 = r_0^T(x_i - x_0) + \|x - x_i\|_A^2 \,. \tag{2.37}$$

Although (2.35) and (2.37) are mathematically equivalent, (2.37) was derived from (2.35) using the global orthogonality between v_1 and all other basis vectors $v_2, v_3, \ldots, v_i$. Therefore one can expect that in finite precision computations, where the global orthogonality is not well preserved due to rounding errors propagation throughout the Lanczos recurrences, (2.37) will be significantly less accurate than (2.35). This was confirmed in [43].

In the next section we present relations mathematically equivalent to (2.37). We should keep in mind that the quadratures and orthogonal polynomials underlying these algebraic equations represent the highly nonlinear objects with respect to the original data.

2.3 CG error estimates

2.3.1 Lower bounds for the A-norm of the error

In this section we present (following [43]) three expressions for the A-norm of the error at the i-th step of the CG method. Assuming its sufficient rate of decrease we show that the lower bounds for the A-norm of the error are sufficiently close to the actual size of the error and we discuss their properties. Details and the analysis of numerical stability in finite precision arithmetics can be found in [43].

Using the Gauss quadrature and mutual orthogonality between Lanczos vectors $v_1, \ldots, v_i$ we showed in Section 2.2.4 that

$$\|x - x_0\|_A^2 = r_0^T(x_i - x_0) + \|x - x_i\|_A^2 \,. \tag{2.38}$$

By simple algebraic manipulations (without any knowledge about the Gauss quadrature connection), a similar mathematically equivalent relation can be derived (see [43])

$$\begin{aligned}
(x - x_0)^T A(x - x_0) &= (x - x_i + x_i - x_0)^T A(x - x_0) \\
&= (x - x_i)^T A(x - x_0) + (x_i - x_0)^T A(x - x_0) \\
&= (x - x_i)^T A(x - x_i + x_i - x_0) + (x_i - x_0)^T r_0 \\
&= \|x - x_i\|_A^2 + (x - x_i)^T A(x_i - x_0) + r_0^T(x_i - x_0) \\
&= \|x - x_i\|_A^2 + r_i^T(x_i - x_0) + r_0^T(x_i - x_0) \,,
\end{aligned}$$

consequently

$$\|x - x_0\|_A^2 = r_i^T(x_i - x_0) + r_0^T(x_i - x_0) + \|x - x_i\|_A^2 \ . \tag{2.39}$$

The right-hand side of (2.39) contains, in comparison to (2.38), the additional term $r_i^T(x_i - x_0)$. This term is in exact arithmetics equal to zero, but has a correction effect in finite precision computations (see [43, Section 6]). Hestenes and Stiefel in [21, Theorem 6.1] presented the relation

$$\|x - x_{i-1}\|_A^2 - \|x - x_i\|_A^2 = \gamma_{i-1}\|r_{i-1}\|^2 \ . \tag{2.40}$$

Its derivation is simple employing only the local A-orthogonality,

$$\begin{aligned}
\|x - x_{i-1}\|_A^2 - \|x - x_i\|_A^2 &= \|x - x_i + x_i - x_{i-1}\|_A^2 - \|x - x_i\|_A^2 \\
&= \|x_i - x_{i-1}\|_A^2 + 2(x - x_i)^T A(x_i - x_{i-1}) \\
&= \gamma_{i-1}^2 p_{i-1}^T A p_{i-1} + 2r_i^T(x_i - x_{i-1}) \\
&= \gamma_{i-1}\|r_{i-1}\|^2 \ .
\end{aligned}$$

Recurrently

$$\|x - x_0\|_A^2 = \sum_{l=0}^{i-1} \gamma_l \|r_l\|^2 + \|x - x_i\|_A^2 \ . \tag{2.41}$$

Consider (2.38), (2.39) and (2.41) for i and $i+d$, where d is some positive integer. Subtracting the identities for i and $i+d$ results in equivalent relations

$$\|x - x_i\|_A^2 = r_0^T(x_{i+d} - x_j) + \|x - x_{i+d}\|_A^2 \ , \tag{2.42}$$

$$\begin{aligned}
\|x - x_i\|_A^2 &= r_0^T(x_{i+d} - x_i) - r_i^T(x_i - x_0) + r_{i+d}^T(x_{i+d} - x_0) \\
&\qquad + \|x - x_{i+d}\|_A^2 \ ,
\end{aligned} \tag{2.43}$$

$$\|x - x_i\|_A^2 = \sum_{l=i}^{i+d-1} \gamma_l \|r_l\|^2 + \|x - x_{i+d}\|_A^2 \ . \tag{2.44}$$

Neglecting $\|x - x_{i+d}\|_A^2$ on the right-hand side of (2.42), (2.43) and (2.44) we get the lower bounds for $\|x - x_i\|_A^2$. We denote

$$\mu_{i,d} \equiv r_0^T(x_{i+d} - x_i) \ , \tag{2.45}$$

$$\vartheta_{i,d} \equiv r_0^T(x_{i+d} - x_i) - r_i^T(x_i - x_0) + r_{i+d}^T(x_{i+d} - x_0) \ , \tag{2.46}$$

$$\nu_{i,d} \equiv \sum_{l=i}^{i+d-1} \gamma_l \|r_l\|^2 \ . \tag{2.47}$$

Now recall that the A-norm of the error in CG is strictly decreasing. If d is chosen such that

$$\|x - x_i\|_A^2 \gg \|x - x_{i+d}\|_A^2 \,, \tag{2.48}$$

then lower bounds $\mu_{i,d}\,, \vartheta_{i,d}$ and $\nu_{i,d}$ approximate the value of $\|x - x_{i+d}\|_A^2$ with an acceptable inaccuracy $\|x - x_{i+d}\|_A^2$.

Mathematically (in exact arithmetic)

$$\mu_{i,d} = \vartheta_{i,d} = \nu_{i,d} \ .$$

The derivation of $\mu_{i,d}$ is based on the global orthogonality of the Lanczos vectors, that is lost rapidly in finite precision computations (see [43, chapters 5, 6]). Hence, in practice $\mu_{i,d}$ can significantly differ from the other bounds. This happens not because of errors in its evaluation, but due to the fact that the relations used in derivation of (2.37) do not hold.

Bound $\vartheta_{i,d}$ contains, in comparison with $\mu_{i,d}\,$, additional terms $r_i^T(x_i - x_0)$ and $r_{i+d}^T(x_{i+d} - x_0)$. They are equal to zero in exact arithmetic but they have corrective effect in finite precision calculations. On the other hand, they increase the cost of the evaluation and in applications with a variable choice of the parameter d the change of d forces us to recompute whole bound $\vartheta_{i,d}$.

Unlike $\vartheta_{i,d}\,$, the evaluation of $\nu_{i,d}$ requires just scalar inputs that are computed during the CG iterations ($\gamma_i, \|r_i\|^2$). Its derivation used only local orthogonality and the change of parameter d leads to simple updates. Moreover, as shown in [43], $\nu_{i,d}$ is numerically stable.

2.3.2 Upper bounds for the A-norm of the error

Using the CG minimizing property (2.7) and the relation (2.18) we can bound the A-norm of the error by the min-max approximation problem on the (discrete) set of eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix A

$$\|x - x_i\|_A = \min_{\varphi \in \Pi_i} \|\varphi(A)(x - x_0)\|_A \leq \min_{\varphi \in \Pi_i} \max_{\lambda_l} |\varphi(\lambda_l)| \, \|x - x_0\|_A \,. \tag{2.49}$$

Using Chebyshev polynomials for φ in (2.49) gives the widely quoted upper bound for the A-norm of the error (the original result can be found in [27], see also, e.g., [37, Section 6.11])

$$\|x - x_i\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^i \|x - x_0\|_A \ . \tag{2.50}$$

Despite the fact that this bound is very often mentioned in the context of conjugate gradient method, it describes the CG convergence very rarely. Indeed, if the bound is tight, i.e.

$$\|x - x_i\|_A \approx 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^i \|x - x_0\|_A \ , \quad i = 1, 2, \dots n$$

then

$$\frac{\|x - x_{i+1}\|_A}{\|x - x_i\|_A} \approx \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right) \ , \quad i = 1, 2, \dots n - 1 \ ,$$

and CG convergence would be linear. The bound (2.50) employs only information about the edges of the spectrum $\lambda_{min}, \lambda_{max}$ of the matrix A and moreover, it represents a worst-case bound, which means that (2.50) holds for *any* right hand side b and matrix $\tilde{A}$, such that $\kappa(\tilde{A}) = \kappa(A)$. As we showed in Section 2.2.4, the particular distribution of eigenvalues of A determines the particular CG convergence.

Other upper bounds are derived in [19] using Gauss-Lobatto and Gauss-Radau quadratures. However, in finite precision computations the boundedness cannot be guaranteed, see [43]. Moreover, these upper bounds expect the tight estimate of the smallest eigenvalue λ_{min} of A that is not usually at the disposal. Hence we cannot use them together with lower bounds in order to set definite stopping criteria.

2.3.3 Other estimates

Although we pointed out the importance of the A-norm of the error, in some applications (for example in image processing) CG is used for problems where the Euclidean norm $\|x - x_i\|$ is to be minimalized. Theoretically, at most in the n-th step CG gives the exact solution for which

$$\|x - x_n\|_A = \|x - x_n\| = 0 \, .$$

For the Euclidean norm of the error no minimizing property similar to (2.7) holds, but it does not mean that CG is useless for the problems mentioned above. In [21, Theorem 6.3] the relation

$$\|x - x_{i-1}\|^2 - \|x - x_i\|^2 = \frac{\|p_{i-1}\|^2}{\|p_{i-1}\|_A^2} \left(\|x - x_{i-1}\|_A^2 + \|x - x_i\|_A^2 \right) \qquad (2.51)$$

was derived. From (2.51) we see that Euclidean norm $\|x - x_i\|$ in CG is also monotonously decreasing. Using similar consideration as for the A-norm we get for positive d

$$\|x-x_i\|^2 = \sum_{l=i}^{i+d-1} \left(\frac{\|p_l\|^2}{\|p_l\|_A^2} (\|x - x_l\|_A^2 + \|x - x_{l+1}\|_A^2) \right) + \|x-x_{i+d}\|^2 , \quad (2.52)$$

assuming

$$\|x - x_i\|^2 \gg \|x - x_{i+d}\|^2 \quad \text{and} \quad \|x - x_{i+d}\|_A^2 \gg \|x - x_{i+2d}\|_A^2 \qquad (2.53)$$

and using $\nu_{i,d}$ as the estimate for $\|x - x_i\|_A$

$$\|x - x_i\|^2 \approx \sum_{l=i}^{i+d-1} \frac{\|p_l\|^2}{\|p_l\|_A^2} \left(\gamma_l \|r_l\|^2 + 2 \sum_{k=l+1}^{i+2d-1} \gamma_k \|r_k\|^2 \right) . \qquad (2.54)$$

This lower bound for $\|x - x_i\|^2$ requires $2d$ additional iterations.

In practical computations CG is used with preconditioning (Preconditioned Conjugate Gradient Method - PCG). Because here no preconditioning is needed, we refer in that respect to [44] where the bounds for the A-norm of the error in PCG are described and studied.

Relation (2.41) can be generalized also for the approximation of the bilinear form $(c, A^{-1}b)$ in nonsymmetric case, see [45]. For the BiConjugate Gradient Method the estimate has the form

$$\xi_i^B \equiv \sum_{l=0}^{i-1} \gamma_l \, (r_l, s_l) \, ,$$

where r_l, s_l stand for the primal, resp. dual residual (see [45] for details).

2.4 Heuristic for the adaptive estimate

In Section 2.3.1 we described the reliable and numerically stable lower bound $\nu_{i,d}$ for $\|x - x_i\|_A^2$ with the inaccuracy equal to the value of $\|x - x_{i+d}\|_A^2$. Assuming the sufficient decrease of the A-norm of the error within d iterations $i + 1, \ldots, i + d$ given by the condition (2.48), $\nu_{i,d}$ gives the tight estimate of the actual error $\|x - x_i\|_A^2$. If the CG method nearly stagnates and the decrease within the steps $i + 1, \ldots, i + d$ is small,

$$\|x - x_i\|_A^2 \approx \|x - x_{i+d}\|_A^2 \, ,$$

then $\nu_{i,d}$ may represent a significant underestimate of $\|x - x_i\|_A^2$. In this section we present the heuristic proposed in [33] that changes the value of d adaptively in order to satisfy the condition (2.48) and thus provide a suitable error estimate.

The algorithm of the CG method including the evaluation of the estimate $\nu_{i,d}$ and the heuristic for the adaptive choice of d is following:

Given $d,\ x_0,\ r_0 = b - Ax_0,\ p_0 = r_0$

for $\ell = 1, 2, \ldots$ **do**

{CG}

$\gamma_{\ell-1} = (r_{\ell-1}, r_{\ell-1})/(p_{\ell-1}, Ap_{\ell-1})$
$x_\ell = x_{\ell-1} + \gamma_{\ell-1}\, p_{\ell-1}$
$r_\ell = r_{\ell-1} - \gamma_{\ell-1}\, Ap_{\ell-1}$
$\delta_\ell = (r_\ell, r_\ell)/(r_{\ell-1}, r_{\ell-1})$
$p_\ell = r_\ell + \delta_\ell\, p_{\ell-1}$

{Evaluation of $\nu_{i,d}$ for $i = \ell - d$ (starting from $\ell = d + 1$)}

$$\nu_{i,d} = \sum_{m=i}^{i+d-1} \gamma_m \|r_m\|^2$$

{Heuristic for the adaptive choice of d (starting from $\ell = d + 1$)}

if $\gamma_{i+d}\|r_{i+d}\|^2 > \sigma\, \nu_{i,d}$ **then**
  $d := d + 1$
else
  while ($d > 1$ and $\gamma_{i+d}\|r_{i+d}\|^2 \le \sigma\, \nu_{i,d}$) **do**
    $d := d - 1$

end

We start the exposition of the heuristic with a simple remark. Figure 2.2 shows the typical behaviour of the lower bound $\nu_{i,d}$ with the fixed value $d = 5$ in the case of near stagnation. We see that the stages where the decrease of the A-norm of the error is slow always end with the increase of the lower bound when the character of the CG convergence changes. The algorithm of the heuristic is derived from the following workflow:

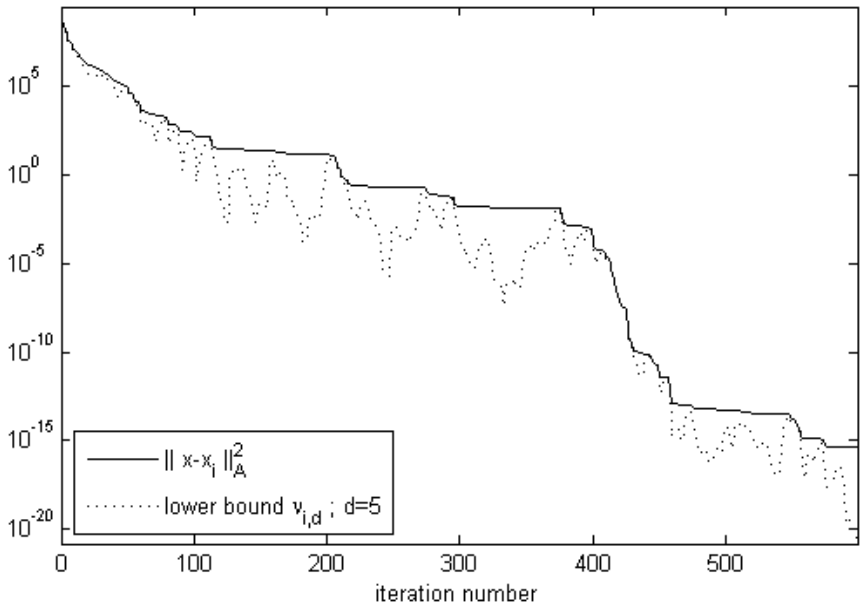


Figure 2.2: *Behaviour of the lower bound in the case of near stagnation;* [33, matrix `maticestred(50,2,2)`]

1) Given d, we test whether the estimate grows:

   if $\gamma_{i+d}\|r_{i+d}\|^2 < \sigma\,\nu_{i,d}$ **then**
   d is too small and has to be increased, we follow to 2)
   else
   we follow to 3) and ask whether a smaller value of d would be sufficient (in order to prevent additional iterations which are not needed)

   Here σ stands for the safety parameter, its choice will be discussed later.

2) We increase $d := d+1$. In order to test, whether this increase is sufficient (whether the condition

$$\|x - x_{i+d}\|_A^2 \ll \|x - x_i\|_A^2 \,.$$

   is fulfilled) we use the heuristic arguments. We replace the unknown values $\|x - x_{i+d}\|_A^2$, $\|x - x_i\|_A^2$ by the estimates $\nu_{i+d,1} = \gamma_{i+d}\|r_{i+d}\|^2$, resp. $\nu_{i,d}$ and demand

$$\gamma_{i+d}\|r_{i+d}\|^2 < \sigma\,\nu_{i,d}\,. \tag{2.55}$$

if (2.55) is satisfied **then**
we proceed out
else
we go back to 2) and further increase d

3) We ask whether d could be decreased. We test (in the same way as in 2)) whether

$$\gamma_{i+d-1}\|r_{i+d-1}\|^2 < \sigma\,\nu_{i,d-1}\,, \tag{2.56}$$

if (2.56) is satisfied **then**
we decrease $d := d - 1$ and go back to 3)
else
we stay with the given value of d and proceed out

The heuristic can be interpreted as the searching for the minimal value of d satisfying

$$\gamma_{i+d}\|r_{i+d}\|^2 < \sigma\,\nu_{i,d}\,,$$

where the safety parameter σ depends on the solved problem. The choice of σ was studied and numerically tested in [33]. In Section 2.6 we present additional numerical experiments and the newly proposed choice of σ for the matrices arising from the discretization of second-order elliptic PDEs.

In the steps 2), 3) we replace $\|x - x_{i+d}\|_A^2$ by the simplest estimate $\nu_{i+d,1}$. In Figure 2.2 we see that even the smallest value of d is sufficient for getting the tight estimate when the A-norm of the error is decreasing. We have numerically tested the estimates $\nu_{i+d,m}$ for $m > 1$ with nearly the same results. Therefore we use the simplest choice $m = 1$.

2.5 Stopping criteria

We start this section with the comment on the choice of the initial vector. Then we present the commonly used stopping criteria for the CG method. The appropriate one has to be always set according to the solved problem. In this section we follow [44, Sections 2 and 3.2]

The proper choice of the initial vector can considerably fasten the convergence of the CG method. However, the initial vector x_0 should be chosen

such that no significant information unrelated with the solution is introduced into the problem. For an unsuitable choice x_0 it may happen that

$$\|x - x_0\|_A \gg \|x\|_A \,.$$

In order to prevent this risk, we can scale the initial vector (i.e. change its size but not the direction) such that

$$\|x - \alpha\, x_0\|_A \leq \|x\|_A \tag{2.57}$$

holds. It can be easily shown that for the choice

$$\alpha \equiv \frac{(b, x_0)}{(x_0, Ax_0)}\,,$$

the error $\|x - \alpha\, x_0\|_A$ is minimal. If we have no relevant information from the problem about the choice of x_0, we use the zero initial vector $x_0 = 0$. For such choice $r_0 = b$.

The relative A-norm of the error

$$\frac{\|x - x_k\|_A}{\|x\|_A} \tag{2.58}$$

is the natural measure of the CG convergence in many cases (e.g. solving PDEs) and can be easily estimated (following [44]). Using the relation

$$\|x\|_A = \|x - x_0\|_A + (b, x_0) + (r_0, x_0)\,,$$

the relations derived in Section 2.3.1 ($d \in \mathbb{N}$)

$$\begin{aligned}
\|x - x_0\|_A^2 &= \sum_{i=0}^{k+d-1} \gamma_i \|r_i\|^2 + \|x - x_{k+d}\|_A^2\,, \\
\|x - x_k\|_A^2 &= \sum_{i=k}^{k+d-1} \gamma_i \|r_i\|^2 + \|x - x_{k+d}\|_A^2\,,
\end{aligned}$$

and assuming $\|x - x_0\|_A \leq \|x\|_A$, the estimate

$$\rho_{k,d}^2 \equiv \frac{\|x - x_k\|_A^2 - \|x - x_{k+d}\|_A^2}{\|x\|_A^2 - \|x - x_{k+d}\|_A^2} \leq \frac{\|x - x_k\|_A^2}{\|x\|_A^2}$$

gives lower bound of the relative A-norm of the error. It can be easily computed

$$\rho_{k,d}^2 \equiv \frac{\sum_{i=k}^{k+d-1} \gamma_i \|r_i\|^2}{\sum_{i=0}^{k+d-1} \gamma_i \|r_i\|^2 + (b, x_0) + (r_0, x_0)} = \frac{\nu_{k,d}}{\nu_{0,k+d} + (b, x_0) + (r_0, x_0)} .$$

and, assuming (2.48), this bound is close to $\|x - x_k\|_A / \|x\|_A$. For the adaptive choice of d the heuristic proposed in the previous section can be used.

Another measure of the CG convergence can be the ratio of the actual and the initial A-norm of the error

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} . \tag{2.59}$$

This value can be bounded (in the same way as the relative A-norm of the error) using the estimates $\nu_{k,d}$

$$\frac{\|x - x_k\|_A}{\|x - x_0\|_A} \geq \frac{\sum_{i=k}^{k+d-1} \gamma_i \|r_i\|^2}{\sum_{i=0}^{k+d-1} \gamma_i \|r_i\|^2} = \frac{\nu_{k,d}}{\nu_{0,k+d}} .$$

Assuming (2.48), this estimate gives suitable results. The heuristic for the choice of d can be applied as above.

The normwise backward error (see, e.g., [31]), measures the smallest relative perturbations $\beta(x_k) = \|\Delta A\| / \|A\| = \|\Delta b\| / \|b\|$ such that

$$(A + \Delta A)\, x_k = b + \Delta b$$

holds. This value can be computed (see [42]) as

$$\beta(x_k) = \frac{\|r_k\|}{\|A\| \|x_k\| + \|b\|} .$$

The spectral norm $\|A\|$ can be estimated using the largest eigenvalue of T_k defined in Section 2.2.3.

The evaluation of convergence is often based on the relative residual norm

$$\frac{\|r_k\|}{\|r_0\|}\,. \tag{2.60}$$

The relative residual norm strongly depends on the initial approximation x_0. For $x_0 = 0$, (2.60) measures the relative norm $\|\Delta b\|/\|b\|$ of the smallest perturbation Δb in the right-hand side b (the matrix A is considered unperturbed), such that x_k solves the perturbed system $Ax_k = b + \Delta b$ exactly. For $x_0 \neq 0$, (2.60) can give a misleading information about the convergence, see [32].

2.6 Numerical experiments

In Section 2.6.1 we demonstrate and explain the different distribution of the discretization and the algebraic error in the FEM discretization. We will see that the local behaviour of the discretization error differs, in general, from the local behaviour of the algebraic error. Moreover, the local behaviour of the total error can be determined by its algebraic part despite the fact, that energy *norm* of the algebraic error is significantly smaller than the energy *norm* of the discretization error. Then in Section 2.6.2 we focus on the local behaviour of the algebraic error. We will see that the algebraic error oscillates with the increasing frequencies with the increasing number of the CG iteration steps. In Section 2.6.3 we test the heuristic proposed in Section 2.4.

2.6.1 Distribution of error

In this section we focus on the local behaviour of the total error, the discretization error and the algebraic error. We demonstrate the influence of the algebraic error on the behaviour of the total error, further elaborating on [25] and [26, Section 5.1].

Problem 1 (Polynomial problem, [10, Example 3])**.**

We consider the problem

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

with right-hand side

$$f = -2\left(x^2 + y^2 - x - y\right).$$

The solution

$$u(x, y) = x\,(x-1)\,y\,(y-1)$$

is shown in the upper left part of Figure 2.3.

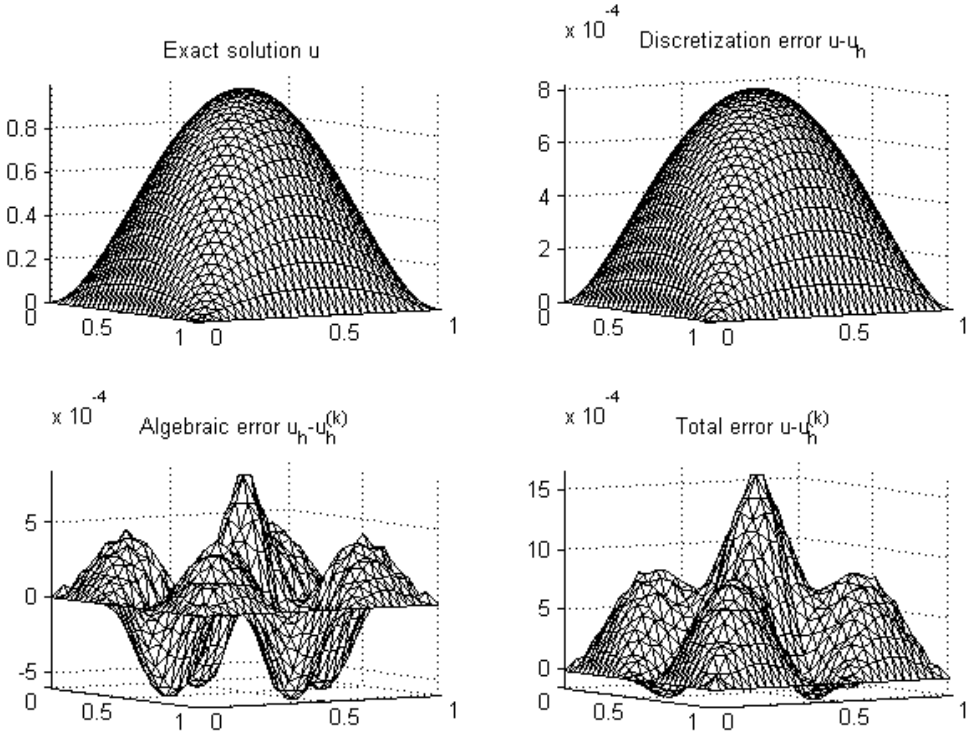


Figure 2.3: *The exact solution u (upper left), the discretization error $u - u_h$ (upper right), the algebraic error $u_h - u_h^{(k)}$ (lower left) and the total error $u - u_h^{(k)}$ (lower right) in Problem 1.*

We discretize the problem using the P^1-conforming FEM discretization on the regular triangular mesh with 30 inner nodes in each direction. The stiffness matrix A has the form

$$A = \text{tridiag}(-I, T, -I) \in \mathbb{R}^{900\times 900}, \quad T = \text{tridiag}(-1, 4, -1) \in \mathbb{R}^{30\times 30}.$$

The right-hand side b is assembled using a two-dimensional Gaussian quadrature formula that is exact for polynomials of degree at most two. The exact solution x of the system $Ax = b$ is approximated (to a sufficient accuracy) using the MATLAB backslash operator. The (closely approximated) squared energy norm of the discretization error is

$$\|u - u_h\|_a^2 = 1.5767\text{e-}2\,.$$

The shape of the discretization error is very similar to the shape of the exact solution, see the upper right part of Figure 2.3.

We apply the CG method with the zero initial vector $x_0 = 0$. We consider the approximation x_k to the exact solution x such that the squared energy norm of the algebraic error is

$$\|x - x_k\|_A^2 = \|u_h - u_h^{(k)}\|_a^2 = 1.6644\text{e-}5\,.$$

Here $u_h^{(k)}$ denotes the approximation to the Galerkin solution u_h given by the approximation x_k, see (2.9).

In Figure 2.3 we can see that the oscillations in the algebraic error $u_h - u_h^{(k)}$ (lower left part) are up to three times greater then the maximum of the discretization error and the local distribution of the algebraic error differs from the local distribution of the discretization error. The behaviour of the resulting total error $u - u_h^{(k)}$, showed in the lower right part of Figure 2.3, is apparently dominated by the algebraic error despite the fact that the energy norm of the algebraic error is significantly lower than the energy norm of the discretization error.

Problem 1 - 1D For simplicity we now consider the one-dimensional boundary value problem

$$\begin{aligned} -u''(x) &= f\,, \qquad x \in [0,1] \\ u(0) &= u(1) = 0 \end{aligned} \tag{2.61}$$

with the right-hand side f chosen differently in the following examples.

We discretize the problem using the piecewise linear basis functions on uniform partition with the step $h = 1/(m+1)$, where m stands for the

number of the inner nodes. The stiffness matrix A has the form

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} .$$

In our numerical experiments we use $m = 49$ (Example 1) and $m = 19$ (Examples 2 and 3) (i.e. 50, respectively 20 elements of the partition). The exact solution x of the stiffness system $Ax = b$ is approximated (to a sufficient accuracy) using the MATLAB backslash operator. We apply the CG method with $x_0 = 0$ and stop the iteration when the normwise backward error $\beta(x_k)$ (see Section 2.5) drops below the prescribed tolerance TOL. The stopping criterion for the CG method is then

$$\beta(x_k) = \frac{\|b - Ax_k\|}{\|b\| + \|A\|\|x_k\|} < TOL\,.$$

In our numerical experiments we evaluate also the componentwise relative backward error (see, e.g., [22])

$$\begin{aligned} \omega(x_k) \equiv \min\{\, \omega \quad | \quad & (A + \Delta A)x_k = (b + \Delta b); \\ & |\Delta A| \le \omega |A|,\ |\Delta b| \le \omega |b| \}\,, \end{aligned}$$

where $|A|$ stands for the matrix and $|b|$ for the vector of the absolute values of the the entries of the matrix A and vector b respectively. Here $\omega(x_k)$ is equal to the absolute value of the smallest relative perturbation in the nonzero entries of A and b such that x_k solves the perturbed system exactly. The componentwise relative backward error $\omega(x_k)$ can be computed using the formula [22, page 130]

$$\omega(x_k) = \max_i \frac{|r_i|}{(|A||x_k| + |b|)_i}\,.$$

We recall the relation (see Theorem 2.2)

$$\begin{aligned} \|u - u_h^{(k)}\|_a^2 &= \|u - u_h\|_a^2 + \|u_h - u_h^{(k)}\|_a^2 \\ &= \|u - u_h\|_a^2 + \|x - x_k\|_A^2 \end{aligned}$$

that holds in finite precision computations up to a small inaccuracy (caused by the computational errors in evaluating the norms and determining the solution) proportional to machine precision. The energy norm $\|\cdot\|_a$ stands in the one-dimensional problem (2.61) for the L^2-norm of the first derivative

$$\|w\|_a = \|w'\|_{(0,1)}\,.$$

Example 1: We consider the problem (2.61) with the constant right-hand side

$$f = 2$$

with the exact solution given by

$$u(x) = -x\,(x-1) \tag{2.62}$$

and the squared energy norm of the discretization error (for $m = 49$)

$$\|u - u_h\|_a^2 = 1.3333\text{e-}4\,.$$

Figure 2.4 shows the solution u given by (2.62) and the corresponding discretization error $u - u_h$. The values of $u - u_h$ at the nodes of the partition are on the machine precision level.

TOL	k	$\beta(x_k)$	$\omega(x_k)$	$\|x - x_k\|_A^2$	$\|u - u_h^{(k)}\|_a^2$
5e-4	23	4.2448e-4	3.2180e-5	1.6000e-4	2.9333e-4
3e-4	24	1.8973e-4	8.0064e-6	1.6000e-5	1.4933e-4
1e-4	25	$\approx$ 1e-17	$\approx$ 1e-17	$\approx$ 1e-30	1.3333e-4

Table 2.1: *The number of CG iterations k, the normwise backward error $\beta(x_k)$, the componentwise relative backward error $\omega(x_k)$, the energy norm of the algebraic error and the energy norm of the total error in Example 1 for the different values of TOL.*

In the Figures 2.5–2.7 we plot the algebraic error $u_h - u_h^{(k)}$ and the total error $u - u_h^{(k)}$ for the values of $TOL =$ 5e-4, 3e-4 and 1e-4 respectively. See Table 2.1 for the number of performed CG iteration k and the values of errors for the particular choices of TOL.

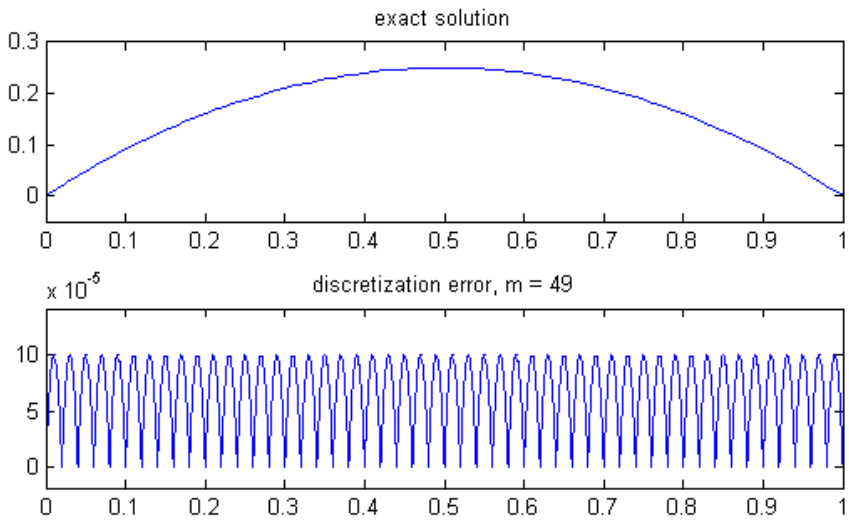


Figure 2.4: *The exact solution u and the discretization error $u - u_h$ in Example 1 for $m = 49$.*

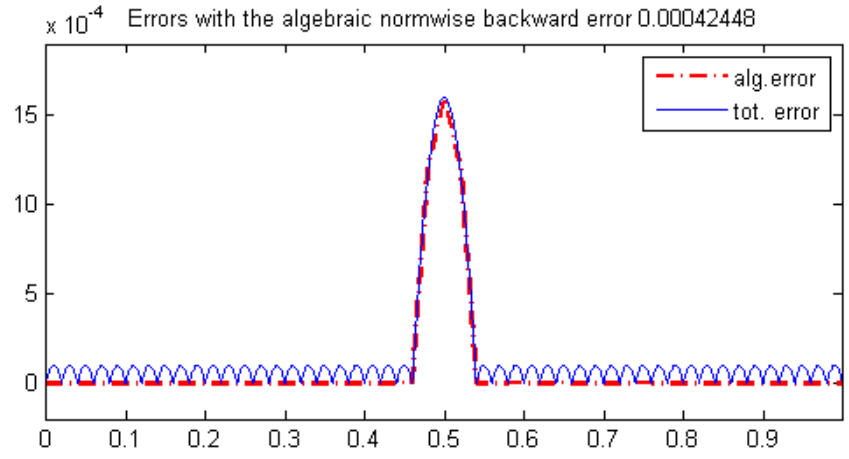


Figure 2.5: *The algebraic error and the total error in Example 1 for the choice TOL = 5e-4. $\beta(x_k)/\|u - u_h\|_a^2 = 3.1837$.*

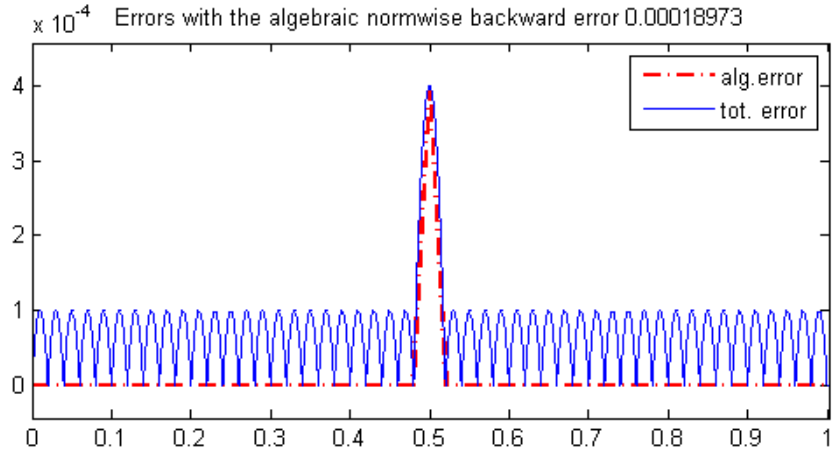


Figure 2.6: *The algebraic error and the total error in Example 1 for the choice* $TOL = 3e\text{-}4$. $\beta(x_k)/\|u - u_h\|_a^2 = 1.4230$.

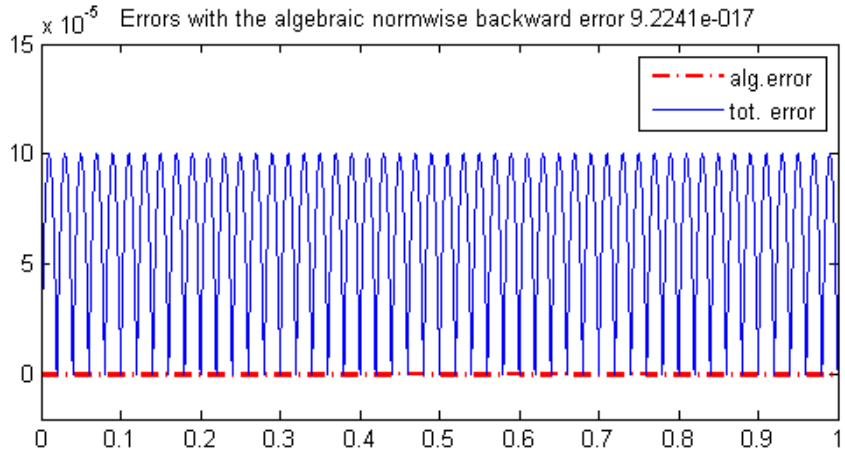


Figure 2.7: *The algebraic error and the total error in Example 1 for the choice* $TOL = 1e\text{-}4$. $\beta(x_k)/\|u - u_h\|_a^2 \approx 1e-13$.

Example 2: In the second example we consider the right-hand side f of (2.61) equal to the polynomial of the second order

$$f = -12x^2 + 12x + 2$$

with the exact solution given by

$$u(x) = (x-2)\,(x-1)\,x\,(x+1) \tag{2.63}$$

and the squared energy norm of the discretization error (for $m = 19$)

$$\|u - u_h\|_a^2 = 3.5000\text{e-}3\,.$$

Figure 2.8 shows the solution u given by (2.63) and the corresponding discretization error $u - u_h$. The values of $u - u_h$ at the nodes of the partition are on the machine precision level.

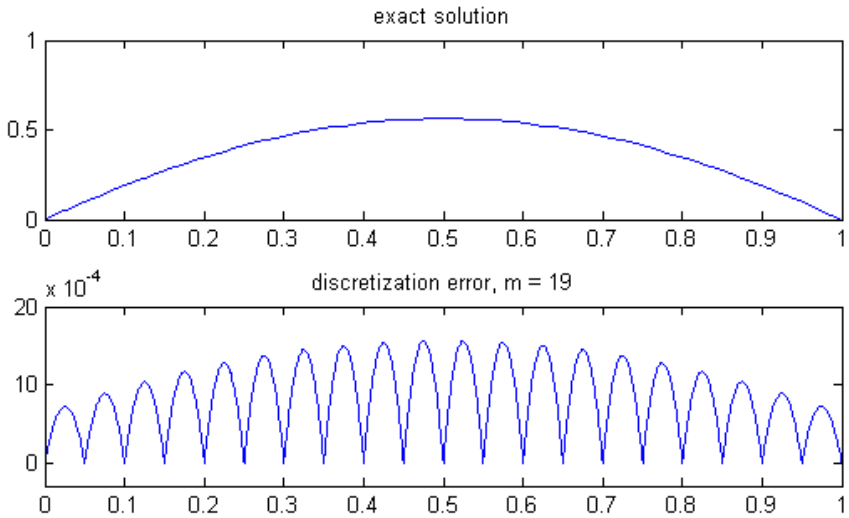


Figure 2.8: *The exact solution u and the discretization error $u - u_h$ in Example 2 for $m = 19$.*

In the Figures 2.9–2.11 we plot the algebraic error $u_h - u_h^{(k)}$ and the total error $u - u_h^{(k)}$ for the values of $TOL =$ 3e-3, 1e-3 and 0.5e-3 respectively. See Table 2.2 for the number of performed CG iteration k and the values of errors for the particular choices of TOL.

TOL	k	$\beta(x_k)$	$\omega(x_k)$	$\|x - x_k\|_A^2$	$\|u - u_h^{(k)}\|_a^2$
3e-3	8	2.0031e-3	1.9944e-4	2.6905e-3	6.1905e-3
1e-3	9	0.8592e-3	5.2429e-5	2.5563e-4	3.7556e-3
0.5e-3	10	$\approx$ 1e-17	$\approx$ 1e-18	$\approx$ 1e-30	3.5000e-3

Table 2.2: *The number of CG iterations k, the normwise backward error $\beta(x_k)$, the componentwise relative backward error $\omega(x_k)$, the energy norm of the algebraic error and the energy norm of the total error in Example 2 for the different values of TOL.*

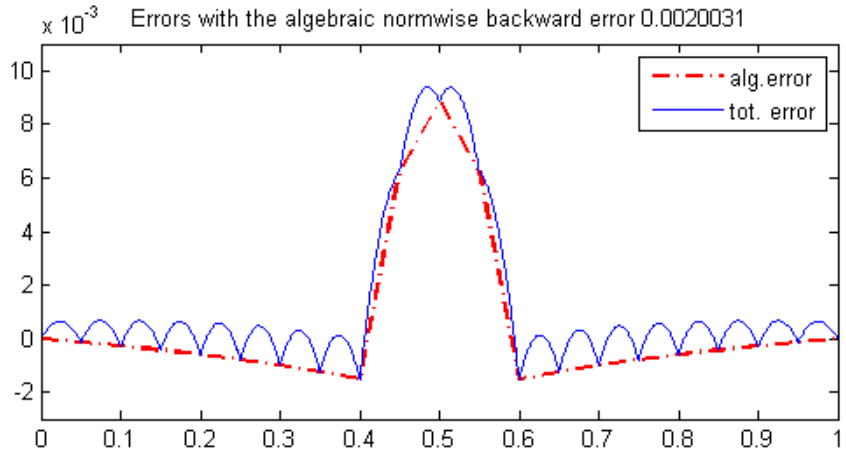


Figure 2.9: *The algebraic error and the total error in Example 2 for the choice $TOL = 3e\text{-}3$. $\beta(x_k)/\|u - u_h\|_a^2 = 0.5723$.*

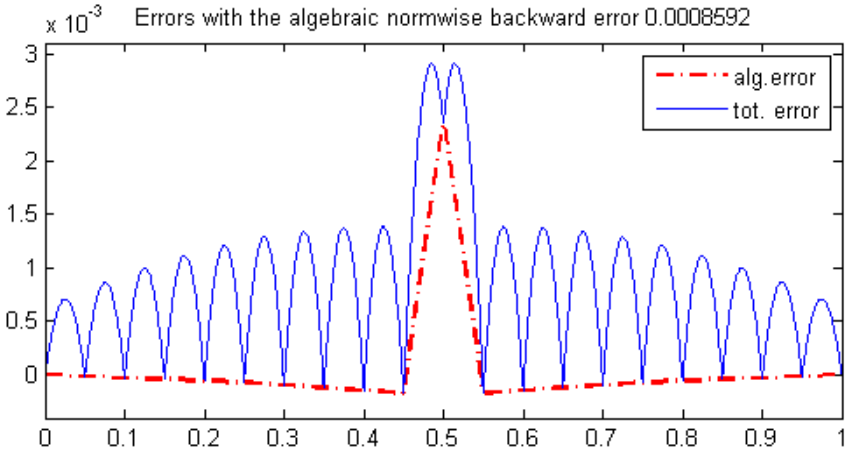


Figure 2.10: *The algebraic error and the total error in Example 2 for the choice* $TOL = 1e\text{-}3$. $\beta(x_k)/\|u - u_h\|_a^2 = 0.2455$.

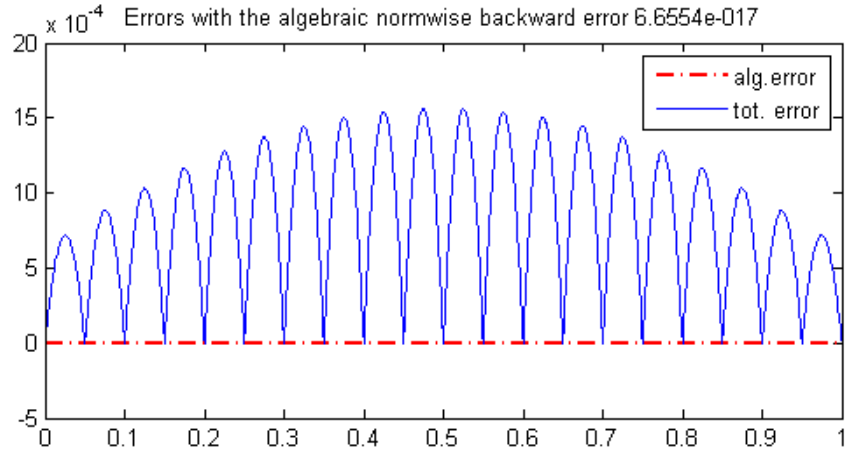


Figure 2.11: *The algebraic error and the total error in Example 2 for the choice* $TOL = 0.5e\text{-}3$. $\beta(x_k)/\|u - u_h\|_a^2 \approx 1e-14$.

Example 3, [13, p. 120]: We consider the problem 2.61 with the right-hand side in (2.61)

$$f = 10\left(1 - 10(x - 0.5)^2\right)\, e^{-5(x-0.5)^2}\,,$$

with the exact solution given by

$$u(x) = e^{-5(x-0.5)^2} - e^{-5/4} \tag{2.64}$$

and the squared energy norm of the discretization error (for $m = 19$)

$$\|u - u_h\|_a^2 = 6.8077\text{e-}3\,.$$

Figure 2.12 shows the solution u given by (2.64) and the corresponding discretization error $u - u_h$. The values of $u - u_h$ at the nodes of the partition are on the machine precision level.

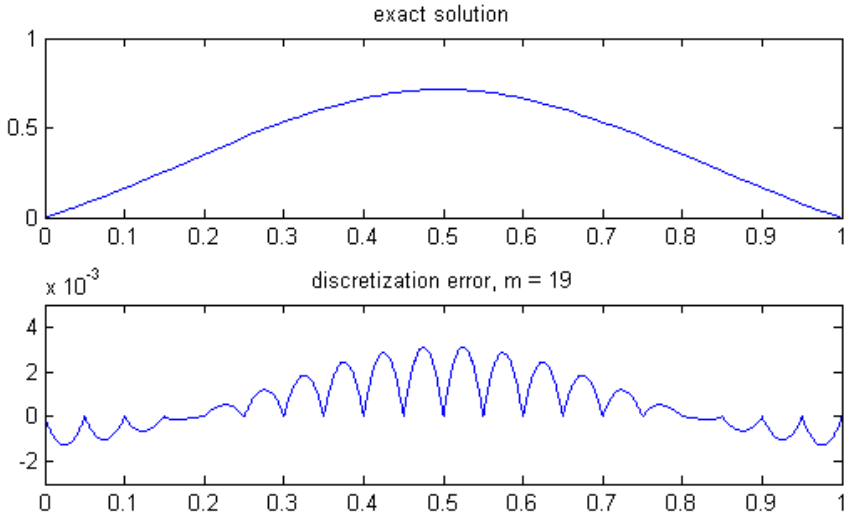


Figure 2.12: *The exact solution u and the discretization error $u - u_h$ in Example 3 for $m = 19$.*

In the Figures 2.13–2.15 we plot the algebraic error $u_h - u_h^{(k)}$ and the total error $u - u_h^{(k)}$ for the values of $TOL =$ 5e-3, 3e-3 and 1e-3 respectively. See Table 2.3 for the number of performed CG iteration n and the values of errors for the particular choices of TOL.

TOL	k	$\beta(x_k)$	$\omega(x_k)$	$\|x - x_k\|_A^2$	$\|u - u_h^{(k)}\|_a^2$
5e-3	8	4.1161e-3	2.5247e-4	1.4504e-2	2.1312e-2
3e-3	9	1.6198e-3	7.2781e-5	1.2381e-3	8.0459e-3
1e-3	10	$\approx$ 1e-17	$\approx$ 1e-17	$\approx$ 1e-30	6.8077e-3

Table 2.3: *The number of CG iterations k, the normwise backward error $\beta(x_k)$, the componentwise relative backward error $\omega(x_k)$, the energy norm of the algebraic error and the energy norm of the total error in Example 3 for the different values of TOL.*

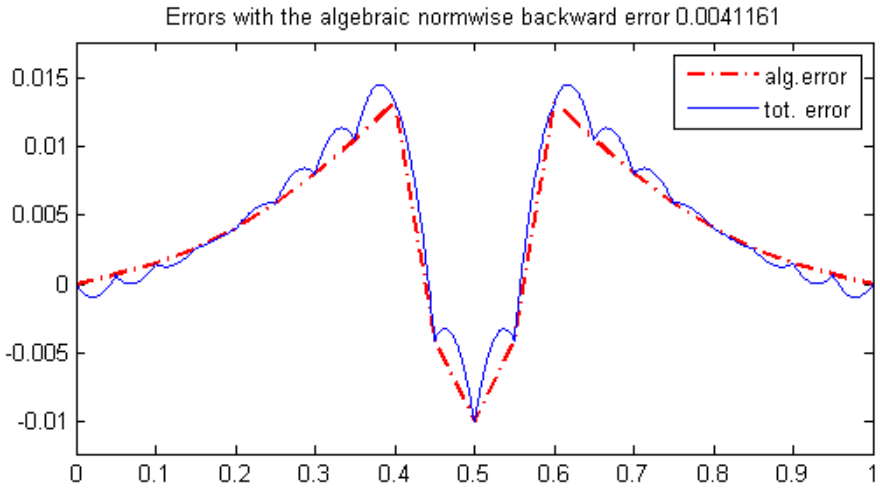


Figure 2.13: *The algebraic error and the total error in Example 3 for the choice $TOL = 5e\text{-}3$. $\beta(x_k)/\|u - u_h\|_a^2 = 0.6046$.*

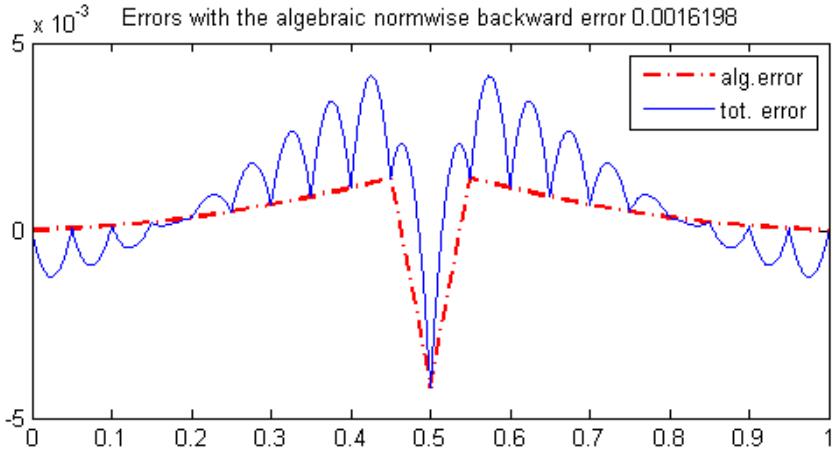


Figure 2.14: *The algebraic error and the total error in Example 3 for the choice TOL = 3e-3.* $\beta(x_k)/\|u - u_h\|_a^2 = 0.2379$.

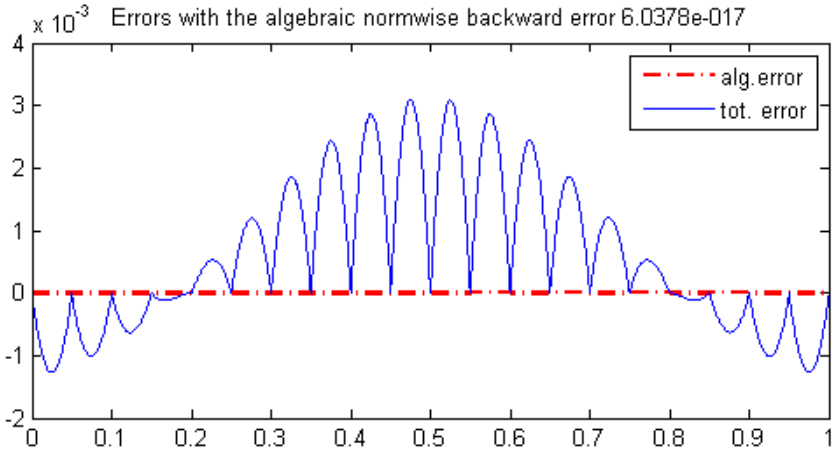


Figure 2.15: *The algebraic error and the total error in Example 3 for the choice TOL = 1e-3.* $\beta(x_k)/\|u - u_h\|_a^2 \approx 1e-14$.

In Figure 2.16 we compare the relative discretization error $(u - u_h)/u$ with the relative total error $(u - u_h^{(k)})/u$ in Example 2 for the choice $TOL = 3\text{e-}3$. The relative errors tend to 1 on the boundaries of the interval $[0, 1]$.

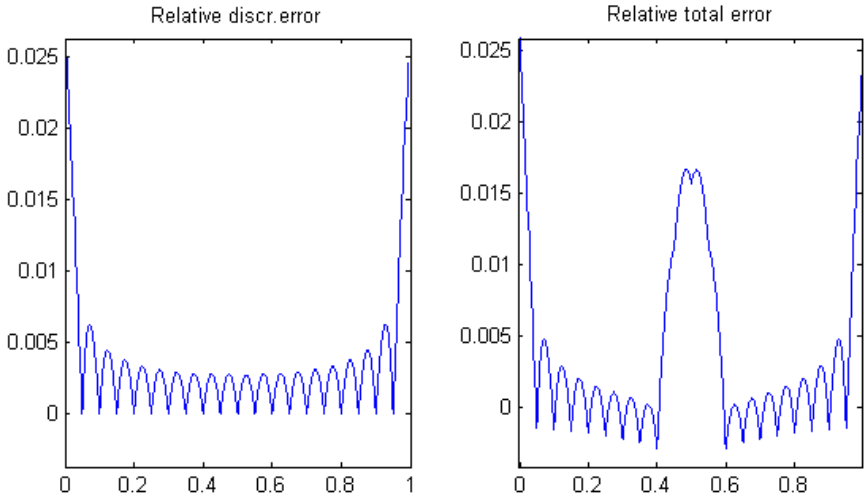


Figure 2.16: *The relative discretization error* $(u - u_h)/u$ *and the relative total error* $(u - u_h^{(k)})/u$ *in Example 2 for the choice* $TOL = 3\text{e-}3$.

Figure 2.17 shows the contributions to the energy norm of the discretization error and the total error

$$\begin{aligned} &\|u - u_h\|^2_{a,[i\,h,(i+1)\,h]} \\ &\|u - u_h^{(k)}\|^2_{a,[i\,h,(i+1)\,h]} \qquad\qquad i = 0, 1, \ldots, 19 \end{aligned}$$

on the elements of the partition $[i\,h, (i+1)\,h] = [i/19, (i+1)/19]$ in Example 2 for the choice $TOL = 3\text{e-}3$. In this case, the energy norm of the algebraic error is about 1.3 times smaller than the energy norm of the discretization error but it is concentrated mostly around the center of interval $[0, 1]$, see Figure 2.9. We recall that the energy norm of the algebraic error is equal to the L^2-norm of the gradient of the algebraic error.

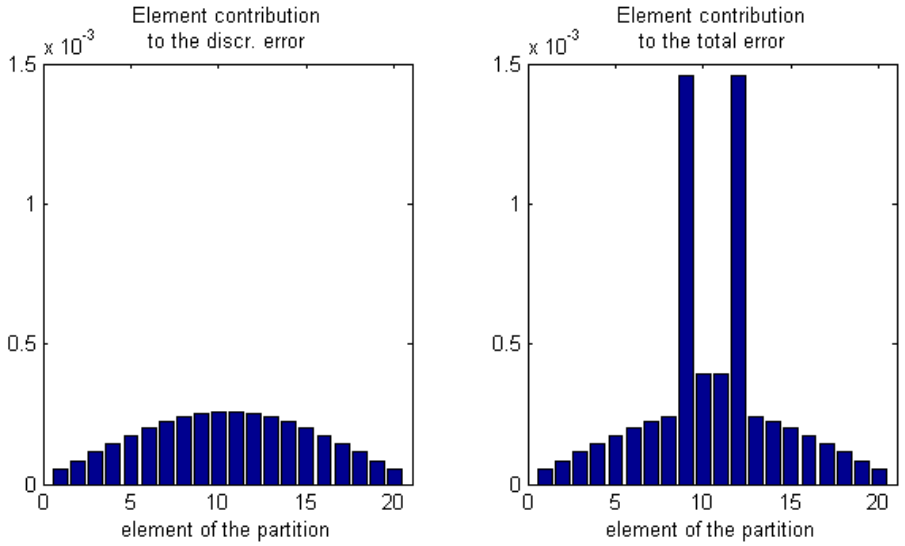


Figure 2.17: *The contributions to the energy norm of the discretization error* $\|u-u_h\|_a^2$ *and to the energy norm of the total error* $\|u-u_h^{(k)}\|_a^2$ *in Example 2 for the choice* $TOL = 3e\text{-}3$.

These simple examples demonstrate:

1. the local distribution of the algebraic error can significantly differ from the local distribution of the discretization error;

2. the energy norm of the algebraic error gives only the *global* information about the behaviour of the algebraic error.

In the FEM discretization we use the basis function of local support and get close to the uniform approximation error per element. This locality is one of the main principles of the finite element method. Since the individual basis functions approximate the solution u only *locally*, the problem of restoring the *global* approximation is transferred to solving the linear algebraic system, where, as the consequence of the local supports, the stiffness matrix A is large and sparse. The fundamental fact pointed out in [26, Section 5.1] is that *a small globally measured algebraic error can exhibit relatively large local components.*

In P^1-FEM discretization either u_h and $u_h^{(k)}$ are piecewise linear functions and the algebraic error $u_h - u_h^{(k)}$ is also piecewise linear function with

the elementwise constant derivatives. The discretization error is, however, generally nonlinear. In Figure 2.3 the discretization error is displayed as the elementwise linear function and this misled our intuition. Since the nonlinear gradient $\nabla(u - u_h)$ is on the individual elements relatively large, the energy norm of the discretization $\|u - u_h\|_a = \|\nabla(u - u_h)\|$ is (in this particular case) of the higher order than the energy norm of the algebraic error $\|u_h - u_h^{(k)}\|_a = \|x - x_k\|_A$. As the consequence, the values of the total error are dominated by the values of the algebraic error, as shown in Figure 2.3. (This phenomenon can be observed also in the one-dimensional examples.) The behaviour of the algebraic error is further studied in the following section.

2.6.2 Smoothness of algebraic error

In the previous section we showed that the local behaviour of the total error may be dominated by the local behaviour of the algebraic error despite the fact that the energy norm of the algebraic error is of the significantly smaller than the energy norm of the discretization error. In this section we therefore focus on the local behaviour of the algebraic error and its smoothness.

In the lower left part of Figure 2.3 one can notice the oscillations in the algebraic error. In Figure 2.18 we show the algebraic errors corresponding to different values of $\|x - x_k\|_A$ in Problem 1 discretized by the P^1-conforming FEM discretization on the regular triangular mesh. Since the CG method tends to approximate well the large eigenvalues of the matrix A (see, e.g. [28]) that corresponds to the low frequencies of the Laplace operator, the algebraic error oscillates with increasing frequencies as the iteration step k increases.

Stationary iterative methods (see, e.g. [37, Chapter 4]) are known to reduce the highly-oscillating parts of the error more effectively than the parts of the error corresponding to low frequencies. This property is called the *smoothing property*. Hence one may suggest to apply a few steps of a stationary method after CG iterations in order to smooth out and further reduce the algebraic error. Such idea is proposed, e.g., in [24, Section 7.13].

In the following experiment we therefore compare the reduction of the error in 3 subsequent iterations of SOR method (see, e.g., [37]) with the reduction in 3 additional CG iterations. The values of the energy norm of

the errors are

k	$\|x - x_k\|_A^2$	$\|x - x_{k+3}\|_A^2$	$\|x - x_{k+SOR(3)}\|_A^2$
12	1.6467e-4	2.8326e-5	5.1591e-5
20	6.5016e-8	1.2616e-8	2.0570e-8
27	6.7929e-10	7.8455e-11	1.1534e-10
34	1.5434e-12	4.1064e-14	7.6223e-14

The comparison of the behaviour of the algebraic errors is given in figures 2.19–2.22. We can see that the SOR method smoothes out the high frequencies but the moderate frequencies that determine the oscillating behaviour of the error stay (almost) unchanged. The additional CG steps reduce either energy norm of the error and the local algebraic errors more effectively. With the increasing number of additional steps the difference is

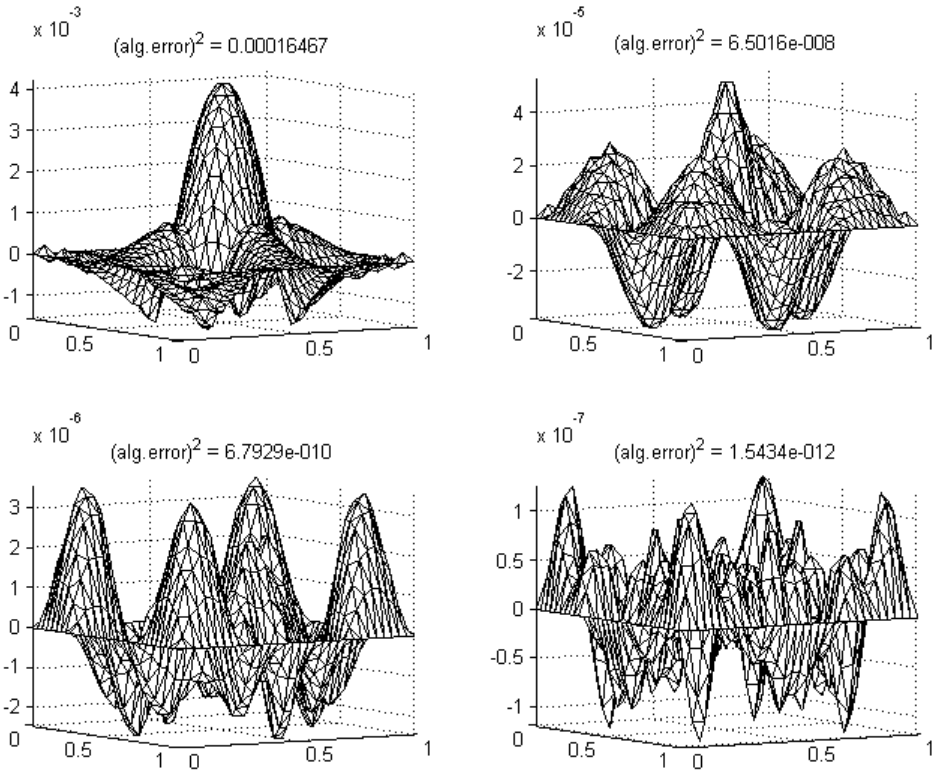


Figure 2.18: *Algebraic error* $u - u_h^{(k)}$ *for different values of the energy norm.*

even more apparent. This property is independent on the chosen stationary method. Jacobi and Gauss-Seidel methods (see [37]) provide even worse results. However, the smoothing property has the application in numerical solving of PDEs as a key feature of multigrid methods, see Chapter 4 of the thesis.

We do not advocate, as pointed out in [25], the CG method for the practical solving of the Poisson problem on regular domains where CG cannot compete with some other methods as the fast Poisson solvers.

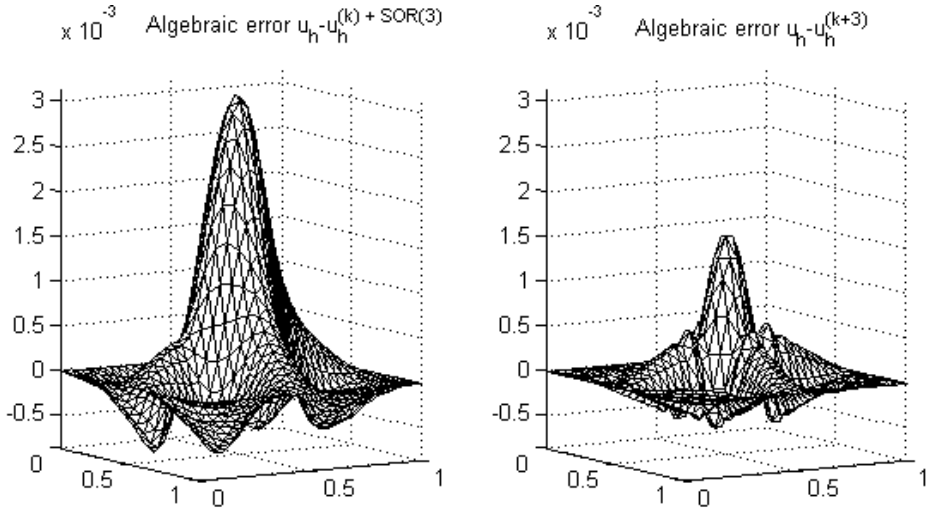


Figure 2.19: *Comparison of the algebraic error* $u_h - u_h^{(k)+SOR(3)}$ *after 3 subsequent steps of the SOR method and the algebraic error* $u_h - u_h^{(k+3)}$ *after 3 additional steps of CG.*

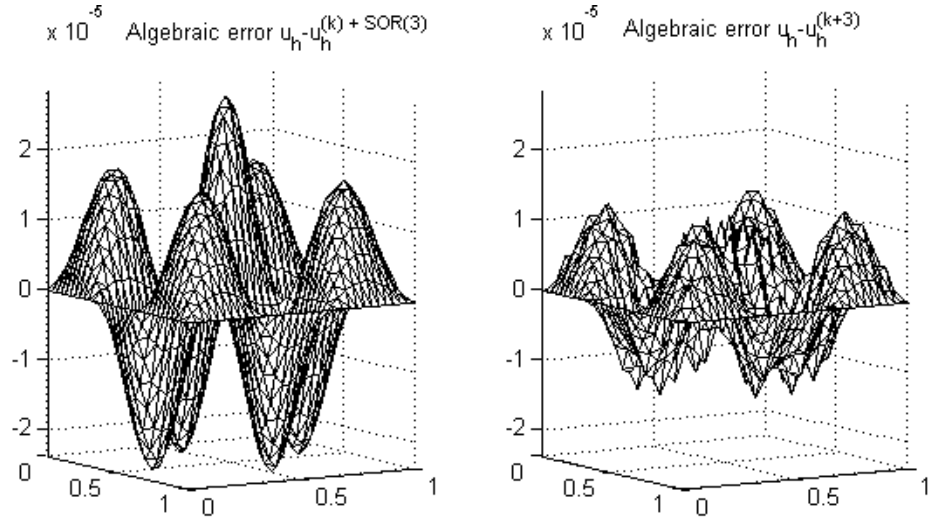


Figure 2.20: *Comparison of the algebraic error* $u_h - u_h^{(k)+SOR(3)}$ *after 3 subsequent steps of the SOR method and the algebraic error* $u_h - u_h^{(k+3)}$ *after 3 additional steps of CG.*

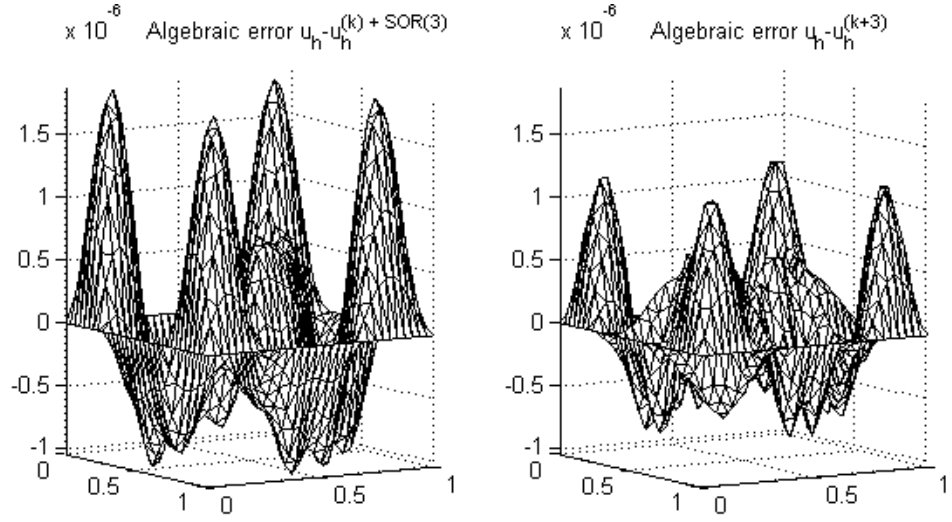


Figure 2.21: *Comparison of the algebraic error* $u_h - u_h^{(k)+SOR(3)}$ *after 3 subsequent steps of the SOR method and the algebraic error* $u_h - u_h^{(k+3)}$ *after 3 additional steps of CG.*

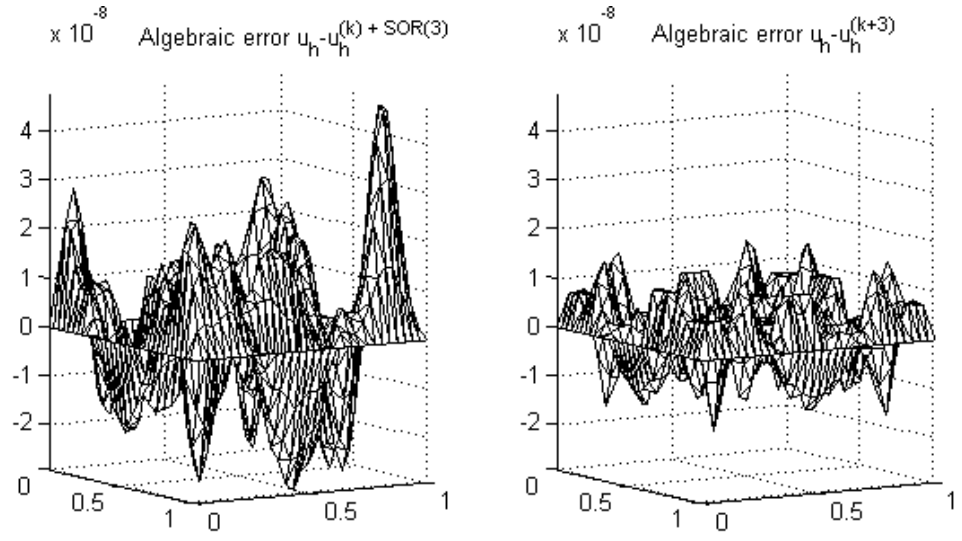


Figure 2.22: *Comparison of the algebraic error* $u_h - u_h^{(k)+SOR(3)}$ *after 3 subsequent steps of the SOR method and the algebraic error* $u_h - u_h^{(k+3)}$ *after 3 additional steps of CG.*

2.6.3 The heuristic

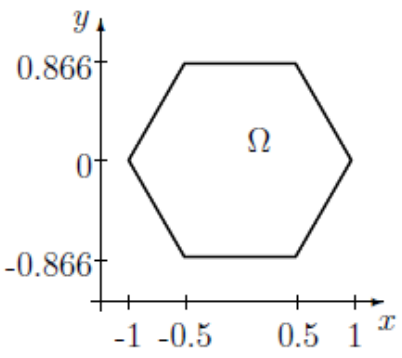
In Section 2.4 we proposed the heuristic that changes adaptively the value of the parameter d in order to satisfy the condition (2.48) and thus provide the suitable estimate for the algebraic error $\|x - x_i\|_A^2$. In this section we propose the value of the safety parameter σ defined in (2.55) and test the adaptively chosen value of d. We apply CG including the heuristic to systems arisen from the discretization of the second-order elliptic problems

Problem 2 (Peak problem, [10, Example 1])**.**

We consider the problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega\,, \\ u &= g_{\mathcal{D}} \quad \text{on } \partial\Omega\,. \end{aligned}$$

on the hexagon domain Ω



with right-hand side f and Dirichlet boundary conditions $g_{\mathcal{D}}$ imposed such that

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

is the solution.

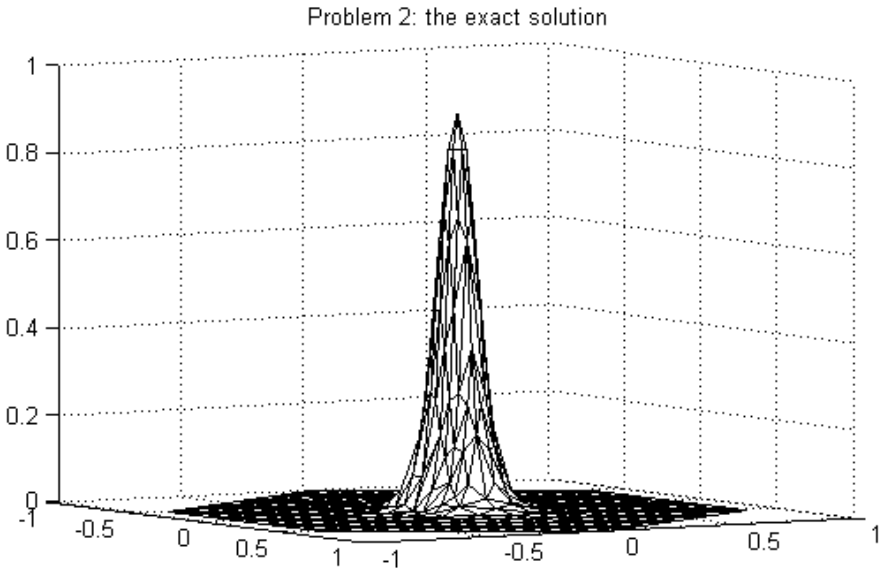
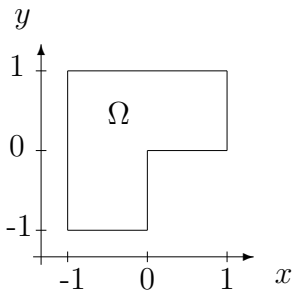


Figure 2.23: *The exact solution of Problem 2.*

Problem 3 (L-shape problem, [2, Test Problem 1])**.**

We consider the L-shape domain Ω



and the problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega\,, \\ u &= g_{\mathcal{D}} \quad \text{on } \partial\Omega\,. \end{aligned}$$

with right-hand side f and Dirichlet boundary conditions $g_{\mathcal{D}}$ imposed such that the solution has in polar coordinates (r, θ) the form

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

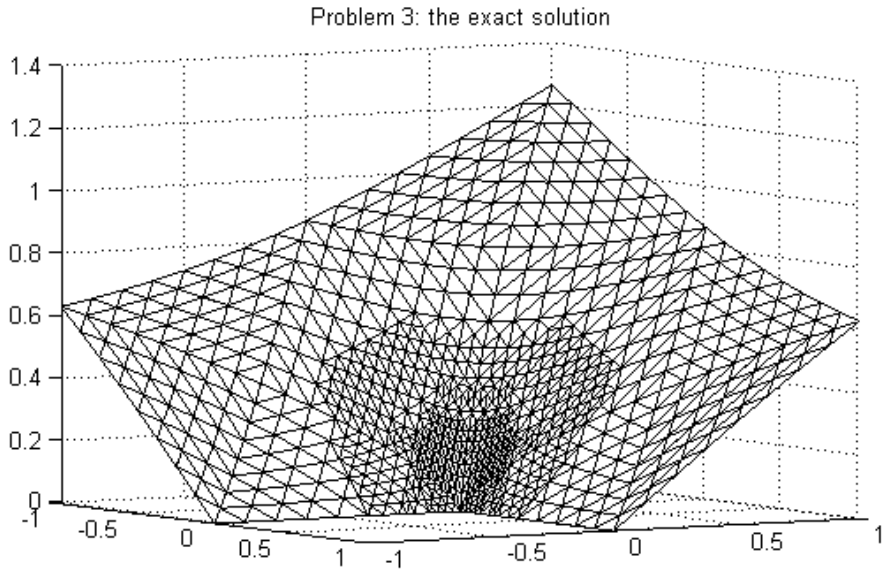


Figure 2.24: *The exact solution of Problem 3.*

Problem 4 (Inhomogenous diffusion tensor, [23, Example 8.2])**.**

We consider the problem

$$\begin{aligned} -\nabla\cdot(\mathbf{S}\nabla u) &= 0 \quad \text{in } \Omega\,, \\ u &= g_{\mathcal{D}} \quad \text{on } \partial\Omega\,. \end{aligned}$$

on square domain $\Omega \equiv [-1,1]^2$ divided into four subdomains Ω_i corresponding to axis quadrants numbered counterclockwise, with $\mathbf{S}$ the piecewise constant tensor equal to $s_i I$ on Ω_i (I stands for the identity matrix). The Dirichlet boundary conditions $g_{\mathcal{D}}$ are imposed such that the solution in each subdomain Ω_i has in polar coordinates (r,θ) the form

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

The parameters are set

$s_1 = s_3 = 5,\ s_2 = s_4 = 1$			
$\alpha =$	0.53544095		
$a_1 =$	0.44721360	$b_1 =$	1.00000000
$a_2 =$	−0.74535599	$b_2 =$	2.33333333
$a_3 =$	−0.94411759	$b_3 =$	0.55555556
$a_4 =$	−2.40170264	$b_4 =$	−0.48148148

. . . . . .

The condition (2.48) on the decrease of the A-norm of the error in d steps is in [33] rewritten using a prescribed parameter $G,\ 0 < G < 1$ as

$$\frac{\|x - x_{i+d}\|_A^2}{\|x - x_i\|_A^2} \le G^2\,. \tag{2.65}$$

The adaptively chosen value of parameter d given by the heuristic is compared to the *ideal value* d_{ideal} defined as the minimal d satisfying (2.65). When the value of d is smaller than d_{ideal} the condition (2.65) is not satisfied and the estimate $\nu_{i,d}$ underestimates the value $\|x - x_i\|_A^2$. On the other hand, a higher value of d (in comparison to d_{ideal}) increases the cost of computation and give no significant improvement to the accuracy of the estimate $\nu_{i,d}$.

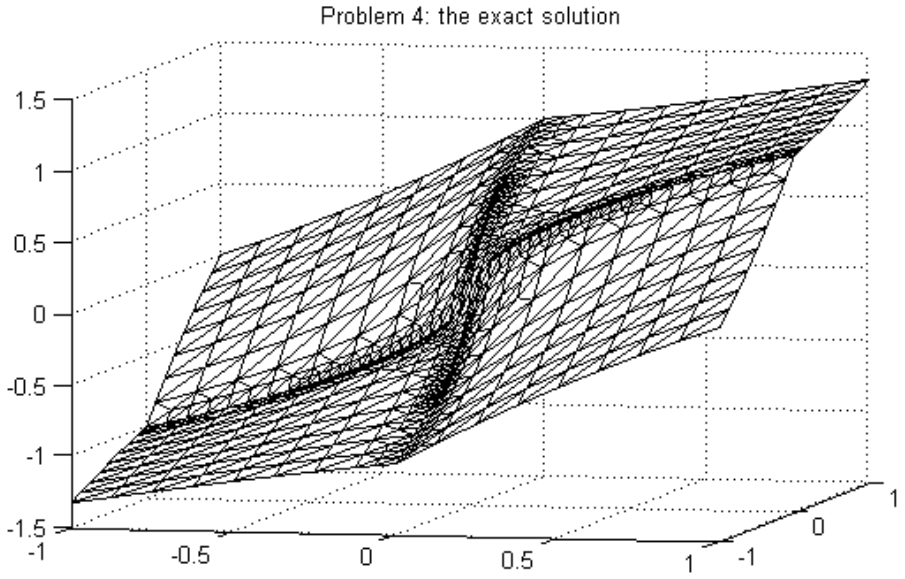


Figure 2.25: *The exact solution of Problem 4.*

	$\dim(A)$	$\|A\|$	$\kappa(A)$
Problem 1, a)	7872	7.9119	3.0716e3
Problem 1, b)	22934	7.9753	1.4110e4
Problem 2, a)	3185	7.9204	1.2990e3
Problem 2, b)	7548	7.9690	2.9711e3
Problem 3, a)	1288	12.3800	3.9377e3
Problem 3, b)	3506	13.3258	1.2913e4
Problem 4, a)	2096	38.4189	3.7698e3
Problem 4, b)	4054	38.4619	1.1045e4

Table 2.4: $\dim(A)$, $\|A\|$ and $\kappa(A)$ for the test matrices

The safety parameter σ from (2.55) was in [33] proposed to be set

$$\sigma = G \cdot \kappa(A)^{-1/4},$$

where $\kappa(A)$ denotes the condition number of the matrix A. After further numerical experiments we do not consider this choice generally recommendable. Moreover, the condition number $\kappa(A)$ cannot be easily estimated.

In this section we consider the system assembled in the discretization of Problems 1–4. For each problem we consider two meshes that are obtained after 5, resp. 6 iterations of the adaptively finite element method. The details of the adaptive mesh refinement and the method will be described in Chapter 5. In Table 2.4 we show the size of the systems, the norm $\|A\|$ and the condition number $\kappa(A)$ of the test matrices.

We propose the safety parameter σ to be set as

$$\sigma = G \cdot \|A\|^{-1/2} . \tag{2.66}$$

In the Figures 2.26–2.33 we compare the ideal value d_{ideal} and the adaptively chosen value of d for the choice $G = 0.4$ and the zero initial vector $x_0 = 0$. In the experiments we estimate the value of $\|A\|$ using the MATLAB command `normest(A)` suitable for sparse matrices.

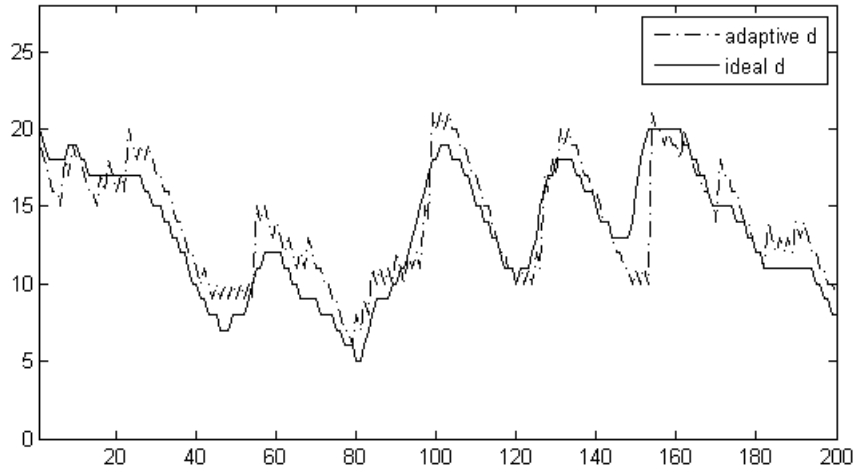


Figure 2.26: *Comparison of the adaptive and ideal value of d, Problem 1, a).*

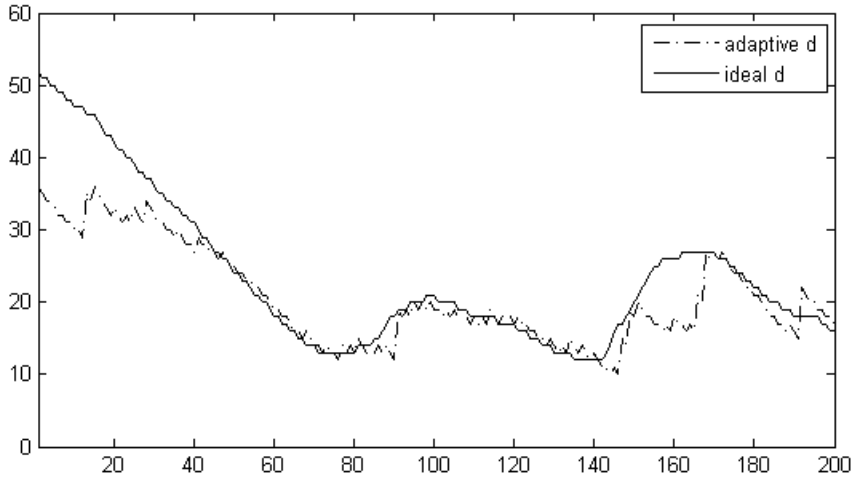


Figure 2.27: *Comparison of the adaptive and ideal value of d, Problem 1, b).*

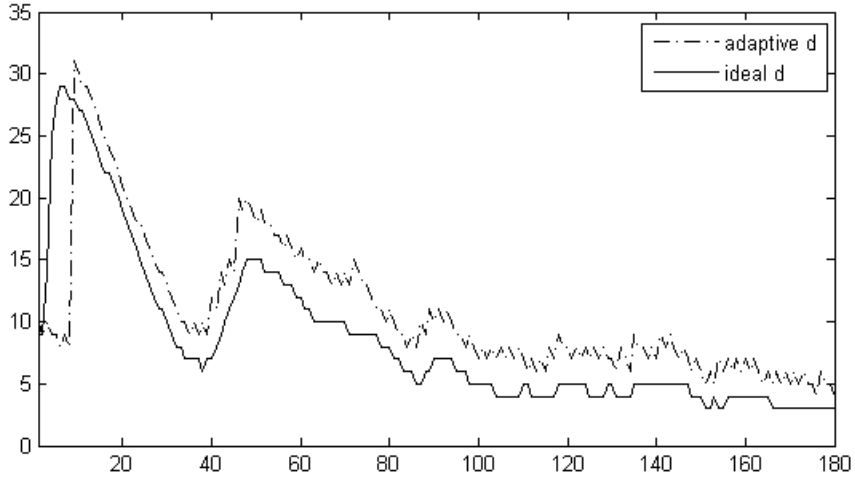


Figure 2.28: *Comparison of the adaptive and ideal value of d, Problem 2, a).*

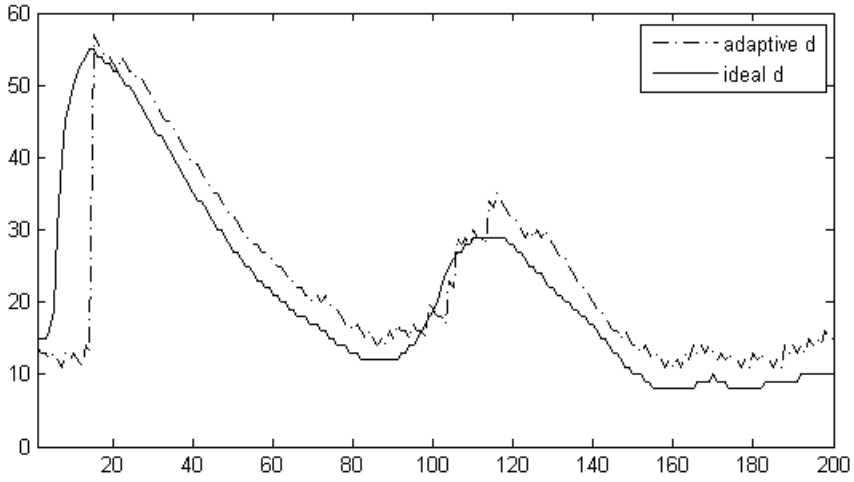


Figure 2.29: *Comparison of the adaptive and ideal value of d, Problem 2, b).*

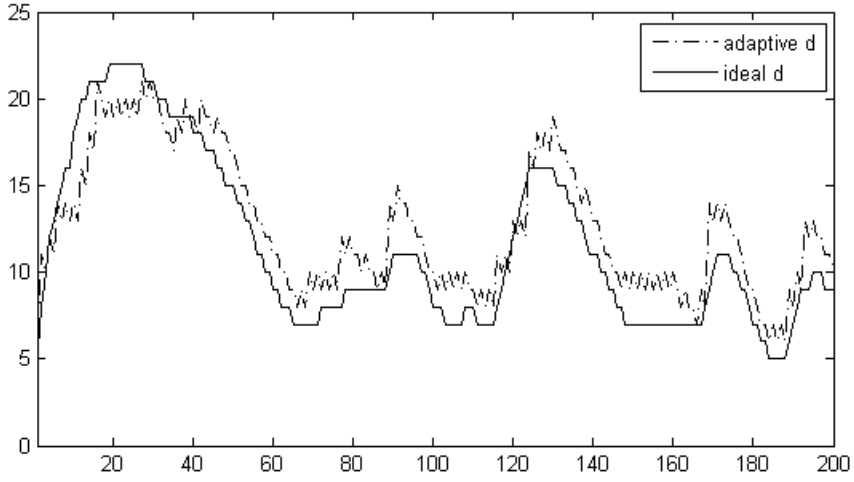


Figure 2.30: *Comparison of the adaptive and ideal value of d, Problem 3, a).*

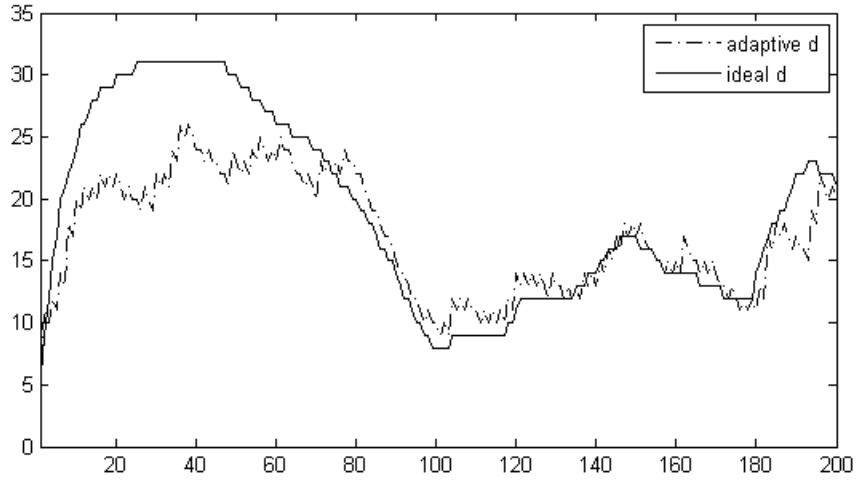


Figure 2.31: *Comparison of the adaptive and ideal value of d, Problem 3, b).*

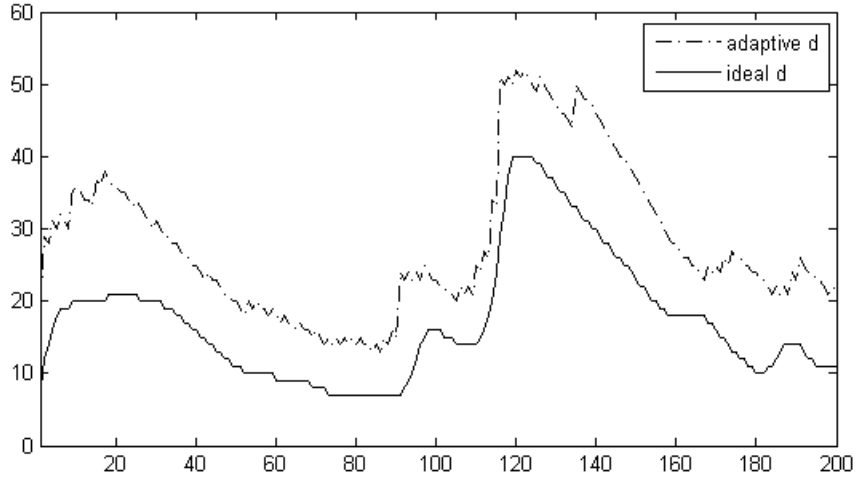


Figure 2.32: *Comparison of the adaptive and ideal value of d, Problem 4, a).*

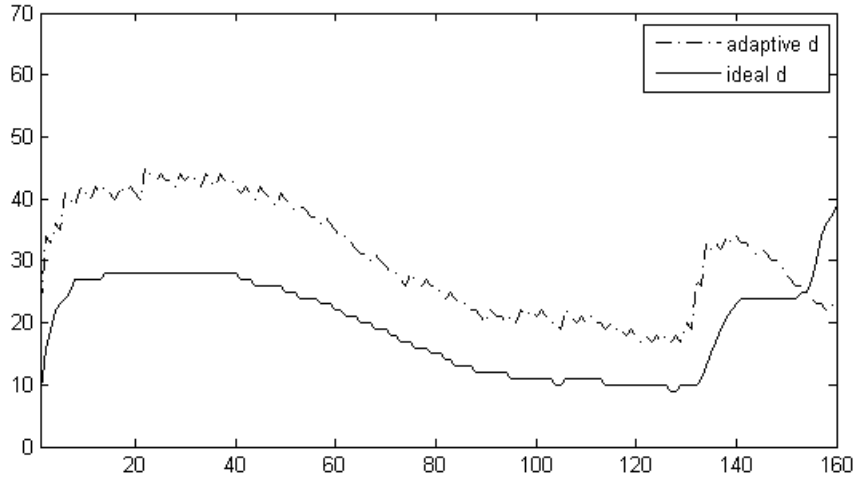


Figure 2.33: *Comparison of the adaptive and ideal value of d, Problem 4, b).*

For Problems 1–3 the adaptively chosen value of d is very close to the ideal value d_{ideal}. For such d the lower bound $\nu_{i,d}$ gives the tight estimate for $\|x - x_i\|_A^2$ with the innacuracy close to $G^2 \|x - x_i\|_A^2$. For Problem 4 the adaptively chosen value of d is higher than d_{ideal} (see Figures 2.32, 2.33). This means that the estimate $\nu_{i,d}$ is more accurate than we demanded in (2.65). However, the evaluation of $\nu_{i,d}$ requires more CG iteration than needed.

The key element of the heuristic is the choice of the safety parameter σ. We do not claim that there exists a value suitable in general for any system. This subject deserves further study and additional experiments. For the systems arisen from the discretization of the model problem (i.e. second order self-adjoint elliptic problem) we propose (based on the experiments above) the value $\sigma = G \cdot \|A\|^{-1/2}$.

Chapter 3

Including algebraic error into the a posteriori error estimates

The energy norm of the total error depends, as shown in Theorem 2.2, on the energy norm of the error of the Galerkin solution and on the error in the solution of the system measured by the algebraic energy norm. A moderately sized system can be solved by a direct method. For large systems iterative methods represents the only possible alternative. These methods — in contrast to direct methods — produce an approximation of the solution in the every iteration step and the considerable amount of computational time (and work) can be saved by stopping the algebraic solver whenever the algebraic error drops to the level at which it does not significantly effect the total error. This approach requires an a posteriori error estimate that takes into account an inexact solving of the linear algebraic system. The analysis of such estimates advanced recently.

In Section 3.1 we present following [23] the estimate for the total error for the cell-centered finite volume discretization. The estimate consists of three parts – the algebraic, the nonconformity and the oscillation one. Then we present the stopping criteria for the algebraic solver proposed in [23] relating the nonconformity and the algebraic parts of the error.

In Section 3.2 we present the bounds for the total error for the piecewise linear FEM discretization presented in [4]. Based on these bounds we propose and test in Section 3.2.1 an estimate for the total error.

Consider the model problem (1.1)

$$-\nabla \cdot (\mathbf{S}\nabla u) = f \quad \text{in} \quad \Omega \tag{3.1}$$

with the pure Dirichlet boundary condition (1.2)

$$u = g_{\mathcal{D}} \quad \text{in} \quad \partial\Omega_{\mathcal{D}} = \partial\Omega\,. \tag{3.2}$$

Let T_h be the partition of polygonal/polyhedral domain Ω into closed simplices i.e. triangles (d = 2) or tetrahedra (d = 3) such that $\bar{\Omega} = \bigcup_{K\in T_h} K$. We assume that if $K, L \in T_h$, $K \neq L$, then $K \cap L$ is either an empty set, a common face, edge or vertex of K and L. We denote the space of elementwise polynomial functions

$$P^m(T_h) \equiv \{w; w|_K \in P^m(K),\ \forall K \in T_h\}\,.$$

In this chapter we assume:

A1': $\Omega \in \mathbb{R}^d$ is a polygonal ($d = 2$) or polyhedral ($d = 3$) domain,

A2: $\mathbf{S}$ is a symmetric, bounded and uniformly positive diffusion tensor, i.e.

$$\mathbf{S} = (s_{ij})_{i,j=1}^d, \quad s_{ij} \in L^\infty(\Omega), \quad s_{ij} = s_{ji}, \quad i,j = 1,\dots,d,$$
$$\|\mathbf{S}\| \equiv \sup_{\xi\in\Omega} \sup_{0\neq z\in\mathbb{R}^d} \frac{\|\mathbf{S}(\xi)z\|}{\|z\|} < \infty$$

and there exists constant $c_{\mathbf{S}} > 0$ such that

$$c_{\mathbf{S}}\|z\|^2 \leq z^T\mathbf{S}(\xi)z, \quad \forall \xi \in \Omega,\ \forall z \in \mathbb{R}^d.$$

A3: $f \in L^2(\Omega)$ is a source term;

A4: $g_{\mathcal{D}} \in L^2(\partial\Omega_{\mathcal{D}})$ prescribes the Dirichlet boundary condition. We assume that there exists $u_{\mathcal{D}} \in H^1(\Omega)$ such that

$$Tu_{\mathcal{D}} = g_{\mathcal{D}}\,;$$

A6: $f \in P^l(T_h)$ is a piecewise l-degree polynomial function[1].

We recall that the assumptions **A1' – A4** ensure the existence and the uniqueness of the weak formulation

Find $u \in \mathcal{H}^1_{\mathcal{D}}$ such that

$$a(u,v) = \ell(v) = (f,v)_\Omega \quad \forall v \in \mathcal{H}^1_0 \tag{3.3}$$

see Section 1.3.

[1] this assumption is usually satisfied in practice. Otherwise, an interpolation can be used.

3.1 A posteriori error estimates for the finite volume discretization

In [23] the cell-centered finite volume discretization of (3.3) is considered and a posteriori error estimate that takes into account an inexact solution of linear algebraic system is derived. It is shown that algebraic error can be bounded and the stopping criterion for iterative algebraic solvers is proposed. We briefly present the results from [23].

Let the space of piecewise constant function

$$\mathcal{S}^h = P^0(T_h)\,,$$

and the appropriate test and the solution spaces $\mathcal{S}_0^h \subset \mathcal{S}^h$, resp. $\mathcal{S}_{\mathcal{D}}^h \subset \mathcal{S}^h$ be considered. Let the diffusive fluxes through the sides ∂K of the elements K depend linearly on the values $v_h|_K$, $v_h \in \mathcal{S}^h$. Then each row of the arisen linear algebraic system $Ax = b$ corresponds to a single element $K \in T_h$ and it represents the principle of mass conservation (the amount of diffusive fluxes through the sides of K is equal to the amount of sources in K).

Vector x represents the cell-values of the finite volume solution $u_h \in \mathcal{S}_{\mathcal{D}}^h$ see, e.g., [15]. Since u_h is piecewise constant, it is not appropriate for an energy error estimates as $\nabla u_h = 0$. Hence the postprocessed approximation $\tilde{u}_h \in P^2(T_h)$ is constructed using the diffusive fluxes prescribed by the values of u_h, see [23] for details. Such postprocessed approximation exists but is not conforming, i.e. $\tilde{u}_h \notin H^1(\Omega)$. So the Oswald interpolation operator $\mathcal{I}_{OS}$ (see [23] and the references given there),

$$\mathcal{I}_{OS} : P^2(T_h) \to P^2(T_h) \cap H^1(\Omega)$$

is used. Then we can state the following a posteriori error estimate ([23, Theorem 5.2]).

Theorem 3.1 (A posteriori error estimate including the algebraic error)**.** *Let assumptions **A1', A2 – A4, A6** be satisfied. Let u be the weak solution of (3.3). Let x_k be the approximation to the exact solution x of the system $Ax = b$, let $u_h^{(k)}$ be the approximation to the finite volume solution u_h given by x_k and let $\tilde{u}_h^{(k)}$ be the postprocessed approximation prescribed by the values of $u_h^{(k)}$. Then*

$$\|u - \tilde{u}_h^{(k)}\|_a \leq \eta_{NC} + \eta_O + \eta_{AE}\,, \tag{3.4}$$

where the global nonconformity *and* oscillation estimators *are given by the local estimators*

$$\eta_{NC} := \left\{ \sum_{K \in T_h} \eta_{NC,K}^2 \right\}^{\frac{1}{2}} \quad \textit{and} \quad \eta_O := \left\{ \sum_{K \in T_h} \eta_{O,K}^2 \right\}^{\frac{1}{2}} ,$$

respectively, and η_{AE} *stands for the* algebraic error estimator.

The algebraic error estimator η_{AE} estimates the algebraic error due to inexact solving of the system $Ax = b$ and can be bounded using estimates for $\|x - x_k\|_A$.

The local nonconformity estimator $\eta_{NC,K}$ measures the distance of $\tilde{u}_h^{(k)}$ from $H^1(\Omega)$ and it is given by

$$\eta_{NC,K} \equiv \|\tilde{u}_h^{(k)} - \mathcal{I}_{OS}(\tilde{u}_h^{(k)})\|_{a,K} ,$$

where $\|\cdot\|_{a,K}$ stands for the energy norm over element K.

The local oscillation estimator $\eta_{O,K}$ estimates the interpolation error in the right-hand side. Whenever $f \in H^1(\Omega)$, η_O is of the higher order (due to Poincaré inequality, see [23, relation (5.1)]) and its value is significant only on coarse meshes or for highly varying diffusion tensor $\mathbf{S}$.

The proposed stopping criterion relates the algebraic error estimator and the nonconformity one via

$$\eta_{AE} \leq \rho\, \eta_{NC} , \quad 0 < \rho \leq 1 , \tag{3.5}$$

where ρ is typically close to 1. The stopping criterion (3.5) gives the bound

$$\|u - \tilde{u}_h^{(k)}\|_a \leq (1 + \rho)\, \eta_{NC} + \eta_O .$$

Supposing that η_{AE} can be constructed using local contributions $\eta_{AE,K}$ as

$$\eta_{AE} := \left\{ \sum_{K \in T_h} \eta_{AE,K}^2 \right\}^{\frac{1}{2}} ,$$

we can consider a local stopping criterion

$$\eta_{AE,K} \leq \rho_K \eta_{NC,K} , \quad 0 < \rho_K \leq 1 , \quad \forall K \in T_h . \tag{3.6}$$

With the stopping criteria (3.5) and (3.6), the global and local efficiency of the estimates can be proved (see [23, Section 6.2]).

As the matrix A is SPD, CG can be applied for solving the system $Ax = b$. In the numerical examples in [23] the estimate $\nu_{i,d}$ described in Section 2.3.1 gives tight lower bound for η_{AE}. The estimator η_{NC} depends also on the approximate solution x_k but it is too expensive to be evaluated in every CG iteration. Having a cheap and reliable estimate for η_{AE}, η_{NC} can be evaluated only when $\nu_{i,d}$ drops to a certain level.

3.2 A posteriori error estimates for FEM discretization

In [4] the convergence and the complexity of an adaptive FEM algorithm are studied. As the algebraic solver authors propose a multigrid method assuming the existence of the upper bound for the error suppression operator (see Section 4.3). The most interesting (related to the content of the thesis) are the lemmas in [4, Section 3].

Consider the space of piecewise linear continuous functions

$$\mathcal{S}^h \equiv P^1(T_h) \cap C(\Omega)$$

and the appropriate solution space

$$\mathcal{S}^h_{\mathcal{D}} \equiv \left\{ v \in \mathcal{S}^h; v = g_{\mathcal{D}} \text{ on } \partial\Omega \right\} .$$

Denote by $\mathcal{E}_h$ the set of the edges of the triangulation and define for $E \in \mathcal{E}_h$ and $v_h \in \mathcal{S}^h$

$$J_E(v_h) \equiv |E|^{1/2} \left[\frac{\partial v_h}{\partial n_E} \right]_E , \quad J_h(v_h) \equiv \left(\sum_{E \in \mathcal{E}_h} J_E^2(v_h) \right)^{1/2}$$

the edge residuals. Here $|E|$ stands for the Lebesgue measure of E, $[\cdot]_E$ for the jump of an elementwise constant function over E and $\partial v_h / \partial n_E$ for the derivative of v_h in the direction normal to E.

Denote by $\mathcal{N}_h$ the set of the vertices of the triangulation T_h, by w_Z the set of elements joining a vertex $Z \in \mathcal{N}_h$ and for $\omega \subset \Omega$ the mean-value operator π_w

$$\pi_w(f) \equiv \frac{1}{|w|} \int_w f \, dx .$$

We define the oscillation terms

$$\begin{aligned} \mathrm{osc}_Z &\equiv |w_Z|^{1/2}\|f - \pi_{w_Z}(f)\|_{0,w_Z}\,, \quad Z \in \mathcal{N}_h, \\ \mathrm{osc}_h &\equiv \left(\sum_{Z\in\mathcal{N}_h} \mathrm{osc}_Z^2\right)^{1/2}. \end{aligned}$$

Then we can state the theorem giving the upper bound for the error.

Theorem 3.2 ([4, Lemma 3.1], upper bound)**.** *There exists a constant $C_1 > 0$ depending only on the maximum angle of the elements of the triangulation T_h such that for the solution u of (3.3), the Galerkin solution u_h and an arbitrary $w_h \in \mathcal{S}^h_{\mathcal{D}}$*

$$\|u - w_h\|_a^2 \le C_1\left(J_h^2(w_h) + \mathrm{osc}_h^2\right) + 2\|u_h - w_h\|_a^2 \tag{3.7}$$

Consider the sequence $T_{h_k}, k = 0, \ldots, K$ of locally refined meshes. With assumptions on the shape regularity and the number of refined elements, see [4, Assumption 2.1], it can be proved (see [4, Lemma 3.1]) that the constant C_1 depends only on the maximum angle of the initial mesh T_{h_0}.

Theorem 3.3 ([4, Lemma 3.2], lower bound)**.** *There exists a constant $C_2 > 0$ depending only on the minimum angle of the elements of the triangulation T_h such that for all $w_h \in \mathcal{S}^h_{\mathcal{D}}$*

$$J_h^2(w_h) \le C_2(\|u - w_h\|_a^2 + \mathrm{osc}_h^2)\,. \tag{3.8}$$

Similarly as in Theorem 3.2 for a sequence of refined meshes, C_2 depends only on the minimum angle of the initial mesh elements, see [4, Lemma 3.2].

3.2.1 Numerical experiments

Theorems 3.2 and 3.3 give the bounds for the total error. In this section we propose an estimate for the total error and test the bounds and the estimate. In the experiments we consider Problems 1–3 described on pages 39, 58, 59.

We discretize the problem using the P^1-conforming FEM discretization (for details see Section 1.4.1) on the given mesh. We approximate the exact solution x of the Galerkin system $Ax = b$ using the MATLAB backslash operator and we apply the CG method with the zero initial approximation

$x_0 = 0$. We denote by $u_h^{(k)}$ the approximation to the Galerkin solution u_h given by the CG approximation x_k.

We denote the bounds

$$\overline{\eta} \equiv C_1 \left(J_h^2(u_h^{(k)}) + \mathrm{osc}_h^2 \right) + 2 \, \|x - x_k\|_A^2 \,, \tag{3.9}$$

$$\underline{\eta} \equiv \frac{1}{C_2} \, J_h^2(u_h^{(k)}) - \mathrm{osc}_h^2 \,. \tag{3.10}$$

In the experiments we set the constants C_1, C_2 such that

$$C_1 = \frac{\|u - u_h\|_a^2}{J_h^2(u_h) + \mathrm{osc}_h^2} \,, \tag{3.11}$$

$$C_2 = \frac{J_h^2(u_h)}{\|u - u_h\|_a^2 + \mathrm{osc}_h^2} \,. \tag{3.12}$$

For x_k such that

$$\|x - x_k\|_A^2 \gg \|u - u_h\|_a^2$$

the total error $\|u - u_h^{(k)}\|_a^2$ satisfies

$$\|u - u_h^{(k)}\|_a^2 \approx \|x - x_k\|_A^2 \,,$$

see Theorem 2.2, and thus we consider the estimate for the total error

$$\eta_h^{(k)} \equiv C_1 \left(J_h^2(u_h^{(k)}) + \mathrm{osc}_h^2 \right) + \|x - x_k\|_A^2 \,. \tag{3.13}$$

We evaluate the term $J_h^2(\cdot)$ following [16]. In the evaluation of osc_h^2 we approximate the right-hand side f using a two-dimensional Gaussian quadrature formula that is exact for polynomials of degree at most two. In the figures we show the total, the discretization and the algebraic errors $\|u - u_h^{(k)}\|_a^2$, $\|u - u_h\|_a^2$, $\|x - x_k\|_A^2$, the bounds $\overline{\eta}, \underline{\eta}$ and the estimate $\eta_h^{(k)}$.

Example 1: We consider Problem 1, see page 39, discretized on the uniform triangular mesh with 16384 right-angled isosceles elements, i.e. the minimal and the maximal angle of the elements of the triangulations are $\alpha_{min} = 45°$, $\alpha_{max} = 90°$. In this example

$\|u - u_h\|_a^2 =$	4.1803e-6	$\mathrm{osc}_h^2 =$	1.9647e-8
$C_1 =$	0.0397	$C_2 =$	25.2866
		$1/C_2 =$	0.0395

In Figure 3.1 we show the errors, the bounds and the estimate $\eta_h^{(k)}$ in Example 1. The solid line and the bold solid line are almost identical, i.e. $\eta_h^{(k)}$ gives a very tight estimate for the total error. When the energy norm of the algebraic error drops below the value of the discretization error ($k > 75$), the values of the bounds, the estimate $\eta_h^{(k)}$ and the values of the discretization and the total error $\|u - u_h\|_a^2$, $\|u - u_h^{(k)}\|_a^2$ are almost equal.

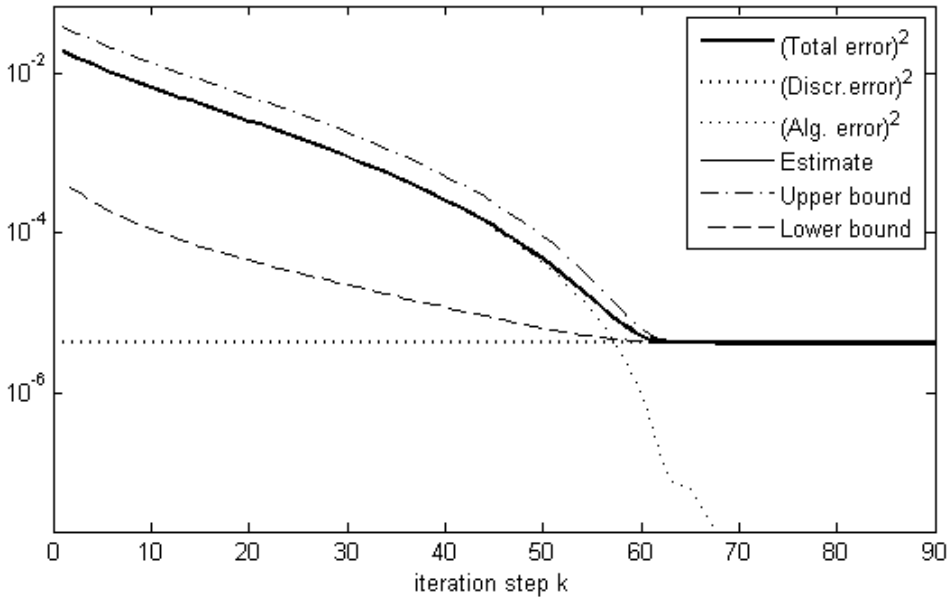


Figure 3.1: *The errors, bounds and the estimate $\eta_h^{(k)}$ in the discretization of Problem 1 on the uniform mesh, Example 1.*

Example 2: In the second example we consider Problem 1 discretized on the mesh shown in Figure 3.2 that consists of 15276 elements. The mesh was obtained by the local refinement from the initial mesh consisting of the right-angled isosceles elements. The values of the discretization error, the oscillation term and constants C_1, C_2 are

$$\begin{aligned} \|u - u_h\|_a^2 &= \text{5.1071e-6} & \text{osc}_h^2 &= \text{3.7856e-8} \\ C_1 &= 0.0396 & C_2 &= 25.4615 \\ & & 1/C_2 &= 0.0393 \end{aligned}$$

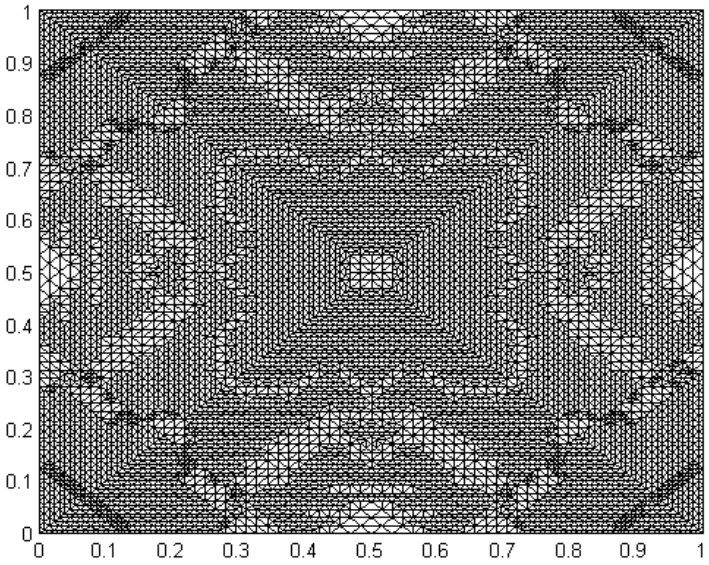


Figure 3.2: *The locally refined mesh in Example 2.*

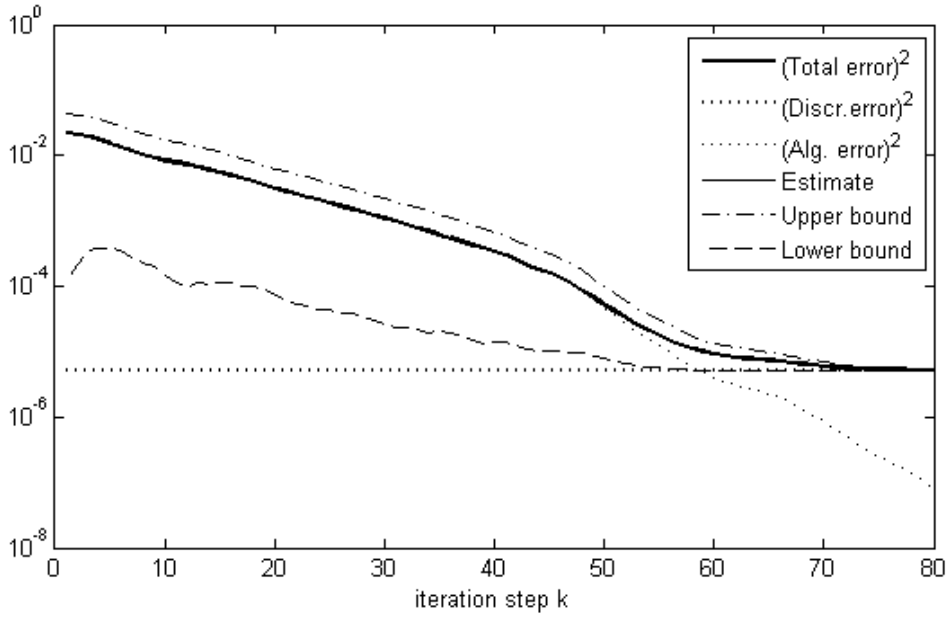


Figure 3.3: *The errors, bounds and the estimate* $\eta_h^{(k)}$ *in the discretization of Problem 1 on the locally refined mesh, Example 2.*

Example 3: We consider Problem 2 described on page 58 discretized on the uniform triangular mesh with 6144 equilateral elements. Due to large values of the right-hand side f around the point [0; 0] the oscillation term osc_h^2 is large (in comparison to discretization error $\|u-u_h\|_a^2$). The values of the discretization error, the oscillation term and the constants C_1, C_2 are

$$\begin{aligned} \|u-u_h\|_a^2 &= 0.0766 \qquad & \mathrm{osc}_h^2 &= 0.2854 \\ C_1 &= 0.0333 \qquad & C_2 &= 5.5567 \\ & & 1/C_2 &= 0.18 \end{aligned}$$

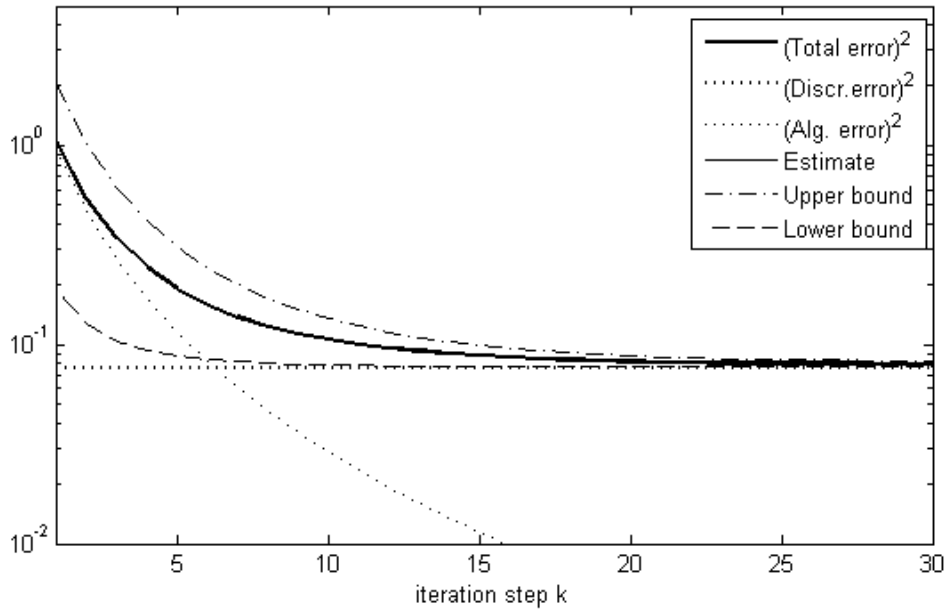


Figure 3.4: *The errors, bounds and the estimate $\eta_h^{(k)}$ in the discretization of Problem 2 on the uniform mesh, Example 3.*

Example 4: We consider Problem 2 discretized on the locally refined mesh that consists of 7514 elements, see Figure 3.5. The mesh is refined mostly around [0; 0] where the exact solution has a steep gradient. Note, that osc_h^2 is significantly smaller than the oscillation term in previous Example 3.

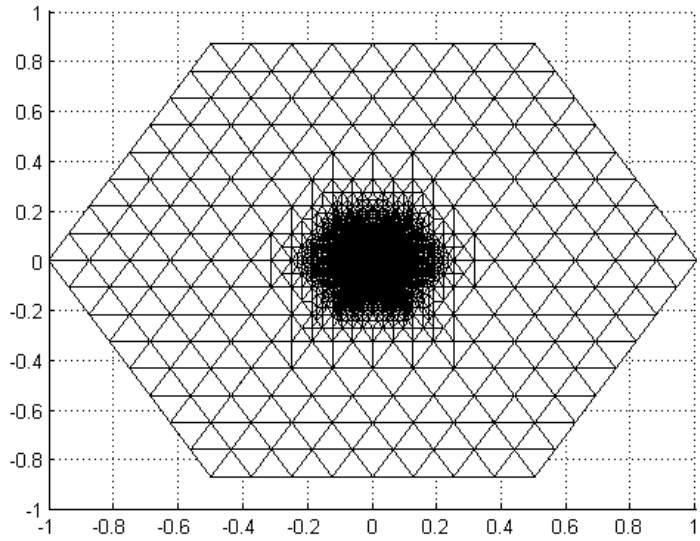


Figure 3.5: *The locally refined mesh in Example 4.*

$$\begin{aligned} \|u-u_h\|_a^2 &= \quad 0.0062 \qquad & \mathrm{osc}_h^2 &= \quad 0.0042 \\ C_1 &= \quad 0.0325 \qquad & C_2 &= \quad 17.8763 \\ & & 1/C_2 &= \quad 0.0559 \end{aligned}$$

In Figure 3.6 we see that $\eta_h^{(k)}$ gives a tight estimate for the total error. Due to the large differences in the element sizes the CG method converges more slowly than in Example 3.

Example 5: We consider Problem 3, see page 59, discretized on the uniform triangular mesh with 12288 right-angled isosceles elements, $\alpha_{min} = 45°$, $\alpha_{max} = 90°$. Since the right-hand side f is constant the oscillation term osc_h^2 is equal to zero and $C_1 = 1/C_2$.

$$\begin{aligned} \|u-u_h\|_a^2 &= \quad 0.0014 \qquad & \mathrm{osc}_h^2 &= \quad 0 \\ C_1 &= \quad 0.0464 \qquad & C_2 &= \quad 21.5426 \\ & & 1/C_2 &= \quad 0.0464 \end{aligned}$$

In Figure 3.7 we see the similar behaviour of the bounds and the estimate $\eta_h^{(k)}$ as in Example 1.

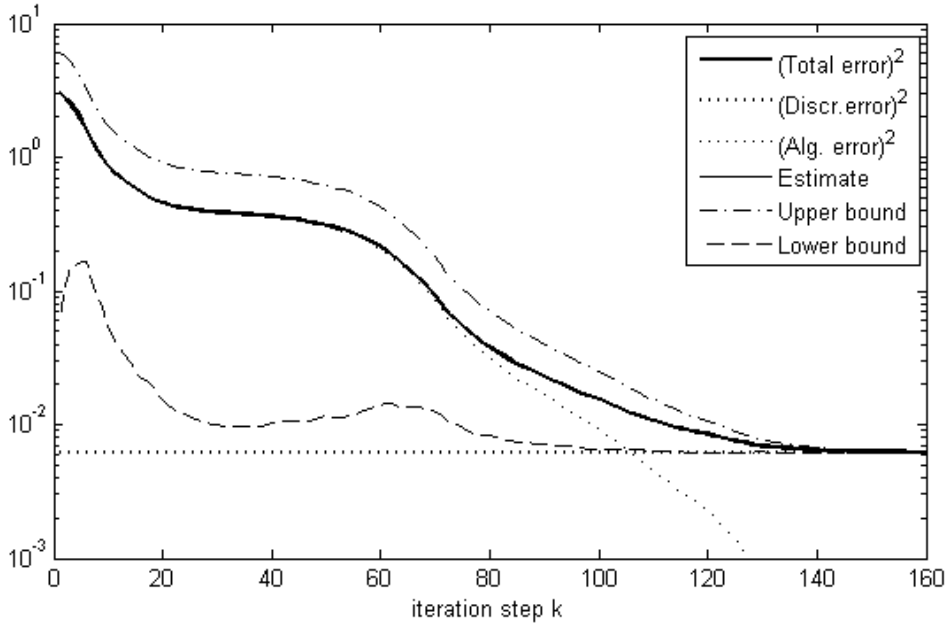


Figure 3.6: *The errors, bounds and the estimate* $\eta_h^{(k)}$ *in the discretization of Problem 2 on the locally refined mesh, Example 4.*

Example 6: In the last example we consider Problem 3 discretized on the mesh shown in Figure 3.8 that consists of 19831 elements. The mesh was obtained by the local refinement from the initial mesh consisting of the right-angled isosceles elements with $\alpha_{min} = 45°$, $\alpha_{max} = 90°$.

$$\begin{aligned} \|u - u_h\|_a^2 = \quad & 1.5661\text{e-}4 \qquad & \mathrm{osc}_h^2 = \quad & 0 \\ C_1 = \quad & 0.0404 \qquad & C_2 = \quad & 24.7376 \\ & & 1/C_2 = \quad & 0.0404 \end{aligned}$$

In Figure 3.9 we see very similar behaviour of the bounds $\underline{\eta}, \overline{\eta}$ and the estimate $\eta_h^{(k)}$ as in the previous Example 5.

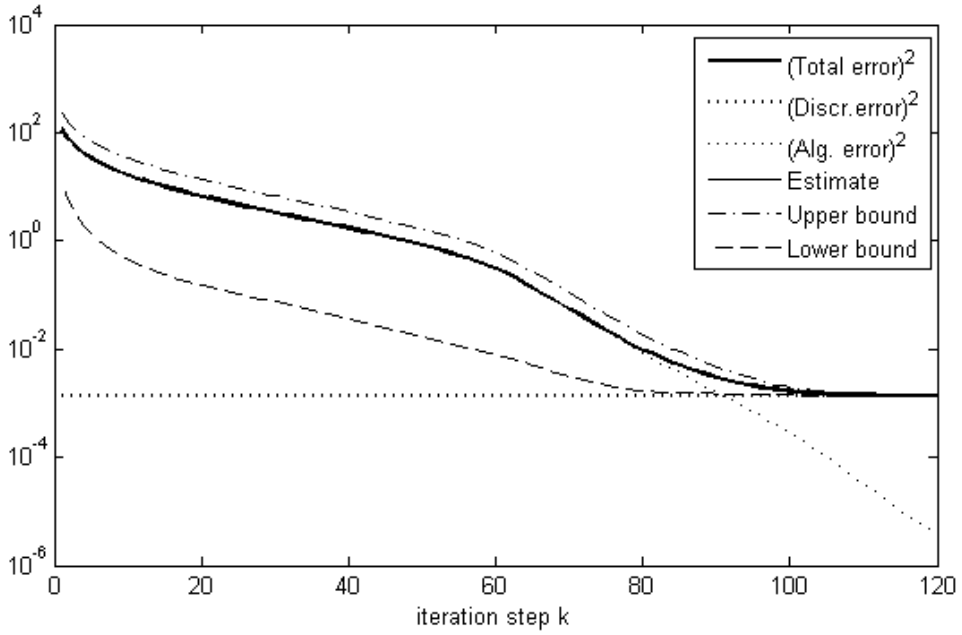


Figure 3.7: *The errors, bounds and the estimate* $\eta_h^{(k)}$ *in the discretization of Problem 3 on the uniform mesh, Example 5.*

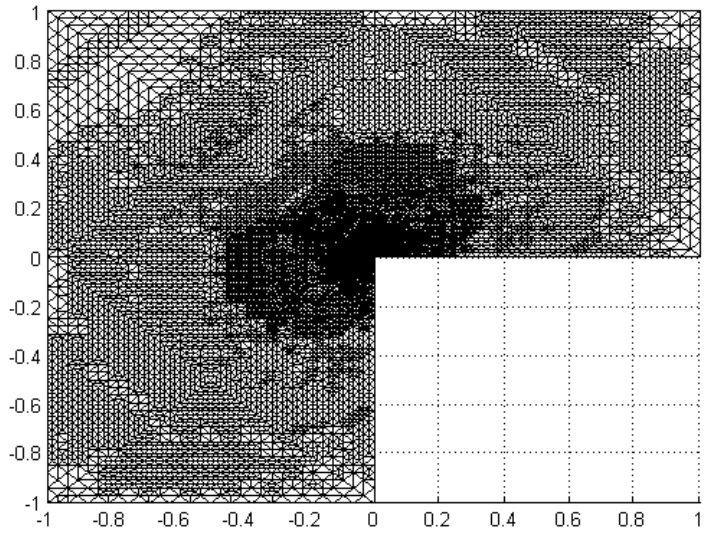


Figure 3.8: *The locally refined mesh in Example 6.*

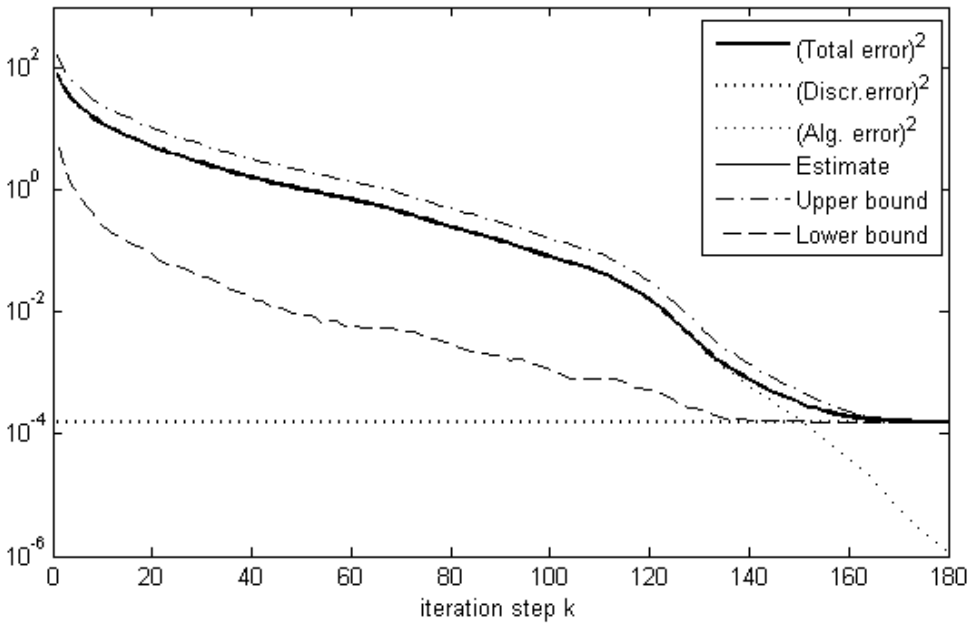


Figure 3.9: *The errors, bounds and the estimate* $\eta_h^{(k)}$ *in the discretization of Problem 3 on the locally refined mesh, Example 6.*

In Figures 3.1–3.9 we see that $\overline{\eta}, \underline{\eta}$ give reliable bounds for the total error $\|u - u_h^{(k)}\|_a^2$ and $\eta_h^{(k)}$ (the bold line) gives the tight estimate for total error. For CG iteration steps k such that

$$\|x - x_k\|_A^2 \ll \|u - u_h\|_a^2 \tag{3.14}$$

the bounds $\overline{\eta}, \underline{\eta}$ are very tight and the discretization error $\|u - u_h\|_a^2$ can be estimated using

$$\eta_h \equiv C_1 \left(J_h^2(u_h^{(k)}) + \mathrm{osc}_h^2 \right) . \tag{3.15}$$

For a smooth right-hand side f or an appropriately refined mesh the value of the oscillation term osc_h^2 is significantly smaller than the value of $\|u - u_h\|_a^2$ (as we can see in Examples 1, 2, 4 – 6) and it can be neglected.

In real computations the values $\|u - u_h\|_a^2$ and $J_h^2(u_h)$ are not known and the constant C_1 cannot be set according to (3.11). For a sequence of locally refined meshes $T_j\,,\ j = 0, 1, \ldots$ (with assumptions on the shape regularity, see [4, Assumption 2.1]) the constant C_1 depends on the maximum angle of

the initial mesh T_0, see [4, Lemma 3.1]. Based on the experiments above we use for $\alpha_{max} = 90^\circ$ the value $C_1 \equiv 0.04$ and for $\alpha_{max} = 60^\circ$ $C_1 \equiv 0.033$.

Chapter 4

Multigrid methods

In order to provide a useful perspective we present in this chapter the idea of using more levels (grids) in multigrid methods. We state the principle, algebraic formulation and some of the basic convergence results.

Starting as the solver for boundary value problems in [5], multigrid methods have grown to be popular in many other domains (e.g. nonlinear problems, see [40, Chapter 6]). The principle of multigrid has been applied also to problems that are not associated to any grid or the grid is too irregular. Such extension is called *algebraic multigrid*, see, e.g., [7, Chapter 8] or [46, Appendix A]. Rather than a single method (or even a family of methods) multigrid now denotes an entire approach, a collection of ideas.

4.1 Principle of multigrid

We describe the principle of multigrid on the simplest multigrid scheme. Consider the (fine) grid Ω_h. Let $A^h x^h = b^h$ be the system arisen from the discretization of the model problem (1.1)–(1.3) on Ω_h (see Section 1.4). Let Ω_H be the coarse grid and let A^H be the matrix assembled in the discretization of the problem on Ω_H. The two-grid correction scheme solves the system $A^h x^h = b^h$ using a stationary iterative method (Jacobi, Gauss-Seidel, SOR, see, e.g., [37, Chapter 4]) and the error correction on the coarse grid. It consists of five phases (see Figure 4.1 for illustration):

- (pre-smoothing)
  Use m_1 iteration steps of a stationary method applied to $A^h x^h = b^h$ with initial guess w^h giving the approximation y^h;

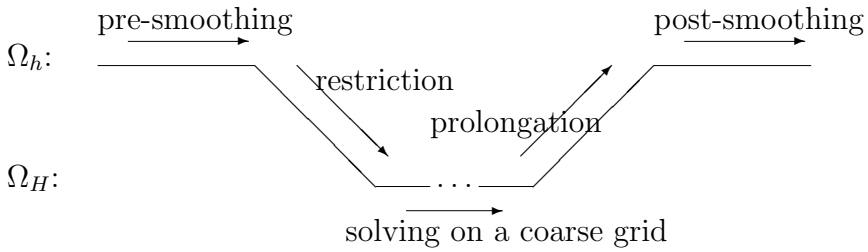


Figure 4.1: *Two-grid correction scheme*

- (restriction)
  Compute the fine-grid residual $r^h = b^h - A^h y^h$ and restrict it to the coarse grid as r^H;

- (solving on the coarse grid)
  Solve $A^H e^H = r^H$;

- (prolongation)
  Prolongate (i.e. interpolate) e^H to the fine grid as e^h and correct the approximation y^h by $y^h := y^h + e^h$;

- (post-smoothing)
  Use m_2 iteration steps of a stationary method applied to $A^h x^h = b^h$ with initial guess y^h giving the final approximation z^h to the exact solution x^h;

The pre-smoothing phase has, due to the smoothing property of stationary iterative methods (see, e.g., [46, Section 2.1]), the effect of damping out the oscillatory part of the error. After few steps of the stationary method the error $x^h - y^h$ is dominated by the smooth components and the additional iterations are not effective. Hence we restrict the error $x^h - y^h$ that satisfies the *residual equation*

$$A^h(x^h - y^h) = b^h - A^h y^h = r^h$$

to the coarse grid where the system $A^H e^H = r^H$ is of smaller dimension and thus it is less costly to be solved. The solution e^H of the restricted system is then prolongated to the fine grid as the approximation e^h of the error $x^h - y^h$. The post-smoothing phase smoothes out the oscillations that may occur in $y^h + e^h$ due to the prolongation.

The system $A^H e^H = r^H$ is smaller than the original system but it can be still too large to be solved effectively by a direct method. For the solution on the coarse grid we can use recursively the correction scheme until the restricted system is small enough to be solved by a direct method.

Now we present two one-dimensional examples to illustrate the smoothing property of stationary iterative methods and to demonstrate the importance of the pre-smoothing phase. In the first example we consider (following the exposition in [7, Chapter 2]) the SPD matrix $A \in \mathbb{R}^{64\times 64}$ arisen from the discretization of the one-dimensional Poisson equation, see [7], in the form

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

the zero right-hand side vector $b = 0$ and the initial vectors $v^{(1)}, v^{(3)}, v^{(8)}$ equal to the (discrete) Fourier basis functions

$$v^{(k)}_{(i)} \equiv \sin\left(ik\pi/64\right), \quad 0 \le i \le 64$$

with the frequencies $k = 1, 3, 8$ respectively. Here $v_{(i)}$ stands for the i-th component of the vector v. The exact solution x of the system $Ax = b$ is equal to the zero vector. For an approximation x_j to the solution x we denote the error vector $e_j \equiv x - x_j$. Figure 4.2 shows on the left the vectors $v^{(1)}, v^{(3)}, v^{(8)}$. On the right we plot the maximum norm of the errors $\|e_j\|_\infty = \max_i |e_{j,(i)}|$ in 100 Jacobi iterations with the initial vectors $v^{(1)}, v^{(3)}$ and $v^{(8)}$. Obviously the maximum norm of the error corresponding to higher frequency is reduced faster, i.e. such error is smoothed out more effectively.

In the second example we illustrate the restriction to the coarse grid and demonstrate the importance of the pre-smoothing phase. We consider the coarse grid as the grid consisting of even-numbered points of the fine grid. The grid size of the coarse grid H is then equal to $2h$. We restrict a vector

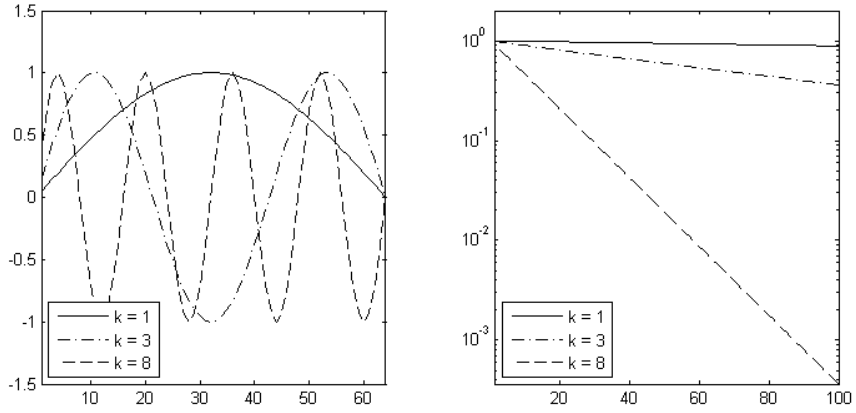


Figure 4.2: $v^{(1)}, v^{(3)}, v^{(8)}$; $\|e_j\|_{infty}$ *for the initial vectors* $v^{(1)}, v^{(3)}, v^{(8)}$

to the coarse grid by using the even indexed components. Figure 4.3 shows the typical restriction of a smooth vector (on the left) and the restriction of an oscillating vector (on the right). We see that the restriction keeps the essential behaviour of the smooth vector while an oscillating vector is misrepresented as a smooth vector. We may expect that the restriction to the coarse grid has the same property also in the d-dimensional case. In the two-grid correction scheme the highly-oscillating parts of the error are smoothed out in the pre-smoothing phase and the error is then restricted to the coarse grid with acceptable accuracy.

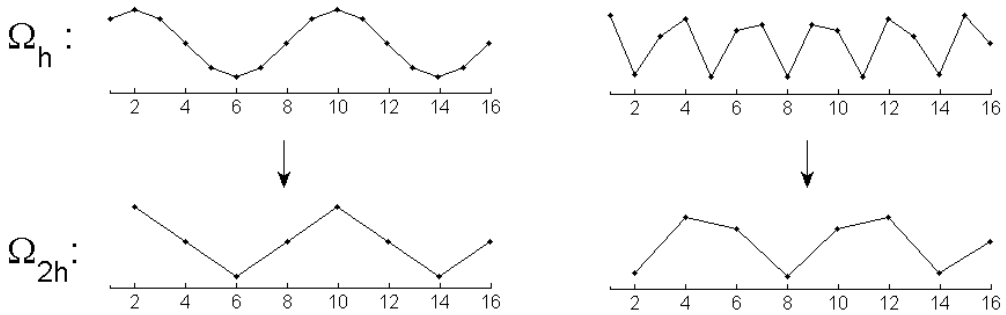


Figure 4.3: *Typical restriction of a smooth vector and an oscillating vector to the coarse grid*

4.2 Algebraic formulation

Let a sequence of finite-dimensional spaces with increasing dimensions

$$M_0, M_1, \ldots, M_k$$

with inner products denoted by $(\cdot,\cdot)_j$ correspondingly in M_j be given. Assume that we have the *prolongation* operators

$$I_j : M_j \to M_{j+1}\,, \quad j = 0, \ldots, k-1\,,$$

the *restriction* operators

$$R_j : M_j \to M_{j-1}\,, \quad j = 1, \ldots, k\,,$$

the invertible operators

$$L_j : M_j \to M_j\,, \quad j = 0, \ldots, k$$

and the sequence of the right-hand sides

$$f_j \in M_j\,, \quad j = 0, \ldots, k\,.$$

Let $\mathcal{D}_j$ be the linear operator corresponding to a stationary iterative method (Jacobi, Gauss-Seidel, SOR)

$$\mathcal{D}_j : M_j \to M_j\,, \quad j = 1, \ldots, k\,.$$

By $MG_j(z, f_j)$ we denote the approximation to the problem

$$L_j u_j = f_j\ , \quad f_j \in M_j \tag{4.1}$$

obtained from the j-th level iteration with the initial guess z. The j-th level iteration of multigrid is defined recursively starting from the lowest level that corresponds to the coarsest grid:

j-th level iteration

The problem on the coarsest level $j = 0$ is solved directly, i.e.

$$MG_0(z, f_0) := L_0^{-1} f_0\ ,$$

For $j > 0$, $MG_j(z, f_j)$ is obtained in 5 steps:

1) (pre-smoothing)
Set $v_0 = z$ and define v_{m_1} by

$$v_{l+1} = v_l - \mathcal{D}_{j,l+1}(L_j v_l - f_j)\,, \quad l = 0, 1, \ldots, m_1 - 1\,;$$

2) (restriction)
Let

$$g_{j-1} = R_j(L_j v_{m_1} - f_j)\,;$$

3) (solving on the coarse grid)
Let $\tilde{w}_0 = 0 \in M_{j-1}$, repeat $(j-1)$-th iteration γ-times:

$$\tilde{w}_s = MG_{j-1}(\tilde{w}_{s-1}, g_{j-1})\,, \quad s = 1, \ldots, \gamma\,;$$

4) (prolongation)
Let

$$y_0 = v_{m_1} - I_{j-1}\tilde{w}_\gamma\,;$$

5) (post-smoothing)
Define y_{m_2} by

$$y_{l+1} = y_l - \mathcal{D}_j(L_j y_l - f_j)\,, \quad l = 0, 1, \ldots, m_2 - 1\,.$$

Finally set

$$MG_j(z, f_j) := y_{m_2}\,.$$

The choice of integers m_1, m_2 depends on the problem (for illustration see [40]). For $\gamma = 1$, this scheme is called *V-cycle multigrid*, for $\gamma = 2$ *W-cycle multigrid*, see Figure 4.4 for illustration.

The full multigrid algorithm for solving the problem on the finest level k

$$L_k u_k = f_k\,, \quad f_k \in M_k \tag{4.2}$$

consists of nested iterations:

Full multigrid algorithm

1. Set $\tilde{v}_0 = L_0^{-1} f_0\,.$
2. For $j = 1, 2, \ldots, k$ do:
   (a) set $\tilde{v}_j = I_{j-1}\tilde{v}_{j-1}\,,$
   (b) repeat t-times:
   $$\tilde{v}_j := MG_j(\tilde{v}_j, f_j)$$

The result after k-th step $\tilde{v}_k$ is an approximation to the solution u_k of (4.2).

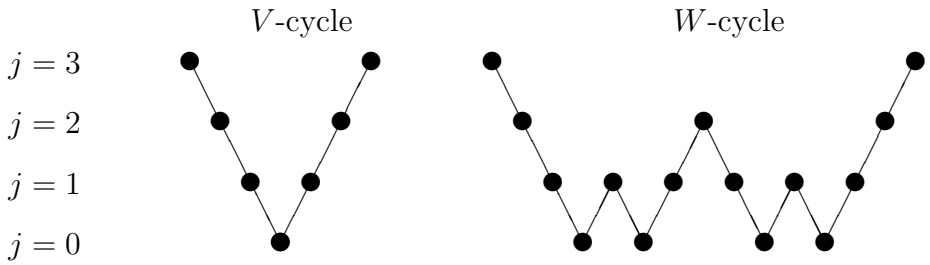


Figure 4.4: *V-cycle, W-cycle multigrid*

4.3 Convergence of multigrid methods

The convergence theory of multigrid methods usually consists of a bounding of the so-called error suppression operator. In this section we present (following [40]) two theorems, for symmetric and for general case.

Let u_j be the solution of (4.1) and let w_0 be an arbitrary initial vector. We denote

$$e_0 \equiv w_0 - u_j$$

the error of the initial vector w_0 and

$$e_1 \equiv w_1 - u_j$$

the error of approximation w_1 to u_j given by the j-th multigrid iteration. We define the operator $B_j : M_j \to M_j\,,\ j = 1, \ldots, k$

$$B_j : e_0 \mapsto e_1\,. \tag{4.3}$$

B_j is called the *operator of error suppression* and has following properties (see [40, Lemma 4.2.1]):

Theorem 4.1. *The operator of error suppression B_j defined in (4.3) is linear for any e_0, independent of f_j, w_0 and has the form*

$$B_j = J_j^{(m_2)}(I - I_{j-1}(I - B_{j-1}^{\gamma})L_{j-1}^{-1}R_jL_j)J_j^{(m_1)}\,,$$

where I stands for the identity operator and

$$J_j^{(m_1)} = (I - \mathcal{D}_jL_j)^{m_1} \tag{4.4}$$

$$J_j^{(m_2)} = (I - \mathcal{D}_jL_j)^{m_2}\,. \tag{4.5}$$

In the symmetric case we assume

SC 1: the operators L_j are self-adjoint and positive-definite;

SC 2: The restriction operator R_j is the transpose to the prolongation operator I_{j-1} with respect to the inner products $(\cdot,\cdot)_{j-1}$ and $(\cdot,\cdot)_j$, i.e.

$$(R_j\, w, v)_{j-1} = (w, I_{j-1}\, v)_j\,, \quad \forall v \in M_{j-1}\,,\ \forall w \in M_j\,;$$

SC 3: Operators $J_j^{(m_1)}$ and $J_j^{(m_2)}$ are adjoint with respect to the inner product $(\cdot,\cdot)_j$, i.e.

$$(J_j^{(m_1)}\, v, w)_j = (v, J_j^{(m_2)}\, w)_j\,, \quad \forall v, w \in M_j\,;$$

SC 4: The operators L_{j-1} , L_j satisfy

$$L_{j-1} = R_j L_j I_{j-1}\,;$$

SC 5: Denoting $J_j = J_j^{(m_1)}$ we assume that $J_j L_j = L_j J_j$.

Consider the L_j-inner product

$$(u, v)_{L_j} = (L_j u, v)_j \quad \forall u, v \in M_j$$

and the corresponding norm

$$\|u\|_{L_j} = (u, u)_{L_j}^{1/2}\,.$$

The convergence criterion is given by the *weak approximation condition*: there exists constant $c^* > 0$ such that $\forall j = 1, \ldots, k$

$$\|v\|_{L_j}^2 - \|J_j v_j\|_{L_j}^2 \geq c^* \|Q_j J_j v\|_{L_j}^2 \quad \forall v \in M_j\,, \tag{4.6}$$

where

$$Q_j = I - I_{j-1} L_{j-1}^{-1} R_j L_j$$

is the operator representing the correction on the coarse grid. The following theorem proves the convergence of the symmetric V-cycle multigrid.

Theorem 4.2 ([40, Theorem 4.5], convergence of the symmetric V-cycle)**.** *Let the assumptions* ***SC 1*** *–* ***SC 5*** *are satisfied and let the weak approximation condition (4.6) holds with constant c^*, let $\gamma = 1$. Then*

$$\|B_j\|_{L_j} \equiv \sup_{0 \neq v \in M_j} \frac{\|B_j v\|_{L_j}}{\|v\|_{L_j}} \leq \frac{1}{1+c^*}\,.$$

In the general case we assume only

GC 1: L_j are invertible operators, $\quad j = 0, 1, \ldots, k\,,$

GC 2: The restriction operator R_j is the transpose to the prolongation operator I_{j-1} with respect to the inner products $(\cdot,\cdot)_{j-1}$ and $(\cdot,\cdot)_j$, i.e.

$$(R_j\, w, v)_{j-1} = (w, I_{j-1}\, v)_j\,, \quad \forall v \in M_{j-1}\,,\ \forall w \in M_j\,;$$

The operators L_j do not induce any norm and hence we consider the norm $\|\cdot\|_j$ induced by the inner product $(\cdot,\cdot)_j$ on M_j. We assume

GC 3: $\|J_j\|_j \leq c_J\,, \quad \forall j = 1, \ldots, k$

GC 4: There exist constants $0 < \underline{c}_I \leq \overline{c}_I$ such that

$$\underline{c}_I \|I_{j-1} v\|_j \leq \|v_j\|_{j-1} \leq \overline{c}_I \|I_{j-1} v\|_j\,, \quad \forall v \in M_{j-1}\,, j = 1, \ldots, k\,.$$

The weak approximation condition (4.6) is replaced by the condition

$$\|L_j^{-1} - I_{j-1} L_{j-1}^{-1} R_j\|_j \cdot \|L_j J_j\|_j \leq \mu(m) \quad \forall j = 1, \ldots, k \tag{4.7}$$

with a function $\mu(m)$ independent of j and tending to zero as $m \to \infty$. Here $m = m_1$ stands for the number of pre-smoothing iterations. In the following theorem ([40, Theorem 4.11]) we consider asymmetric variant of the multigrid algorithm with $m_2 = 0$ and $\gamma = 2$.

Theorem 4.3 ([40, Theorem 4.11], convergence of asymmetric W-cycle)**.** *Let the assumptions* ***GC 1*****–*****GC 4****, are satisfied. Let $m_1 = m$, $m_2 = 0$ and $\gamma = 2$. Then for any $\xi \in (0,1)$ there exists index m_0 such that $\forall m \geq m_0$*

$$\|B_j\|_j \leq \xi \quad \forall j = 1, \ldots, k\,.$$

As the consequence of the convergence of the j-th multigrid iteration, the convergence of the full multigrid method can be shown, see, e.g. [4, Theorem 6.7.1]

The multigrid method provides an optimal order algorithm for solving elliptic boundary value problems, see, e.g., [46, Section 3.2]. The error bounds of the approximate solution obtained from the full multigrid algorithm are comparable to the theoretical bounds for the error in the finite element solution, while the amount of computational work is proportional to the number of unknowns in the discretized equations, see, e.g., [4, Proposition (6.7.4)].

Chapter 5

Cascadic Conjugate Gradient Method (CCG)

Cascadic Conjugate Gradient Method (CCG, [10]) is a method for solving self-adjoint positive-definite problems on a sequence of grids (uniformly as well as adaptively constructed). It combinates the discretization by the Galerkin finite element method and the conjugate gradient method[1] for solving the linear algebraic system arising on each level of the discretization.

In papers [38, 39] the optimal complexity is proved for elliptic second-order Dirichlet problem in 2D for convex and non-convex polygonal domain. The optimal complexity means that CCG converges with the rate that is independent of the number of unknowns and the number of grids.

In CCG (as well as in other PDE solvers) the balancing of the algebraic error and the discretization error on a particular level (grid) is essential. In the original paper [10] an unreliable stopping criterion for CG was used. In the following papers [38, 39] the number of CG iterations was set in order to ensure the algebraic error to drop to a certain level which may lead to many iterations that are not needed. We therefore use the actual algebraic error estimator with the heuristic proposed in Chapter 2.

The CCG method is described in Section 5.1. Using the estimate for the algebraic error and the heuristic for the adaptive choice of the parameter of the estimate proposed in Section 2.4 we propose in Section 5.2 the new stopping criteria for CCG. Then we briefly present the adaptive refinement

[1] In the original paper [10] the preconditioned conjugate gradient method was considered. Nevertheless, the version with no preconditioning was in numerical examples superior (see [10])

technique in Section 5.3. In numerical experiments (Section 5.4) the CCG method with the newly proposed stopping criteria is tested.

5.1 Description of the CCG method

Consider the problem given in the weak formulation

Find $u \in H$ such that

$$a(u, v) = \ell(v), \quad \forall v \in H\,, \tag{5.1}$$

where H is the appropriate Hilbert space, $u \in H$ is the solution and $a(\cdot, \cdot)$ is the symmetric positive-definite bilinear form. In CCG the Hilbert space H is approximated by a sequence of *nested* finite element spaces with increasing dimension

$$S_0 \subset S_1 \subset \ldots S_k \subset H\,. \tag{5.2}$$

The Galerkin FEM discretization of (5.1) on $S_0, S_1, \ldots, S_k$ generates the sequence of problems

Find $u_j \in S_j$ such that

$$a(u_j, v_j) = \ell(v_j), \quad \forall v_j \in S_j\,. \tag{5.3}$$

The Galerkin solution $u_j \in S_j$ minimizes the energy norm of the error over the finite dimensional space S_j

$$\|u - u_j\|_a = \min_{w \in S_j} \|u - w\|_a\,, \tag{5.4}$$

see (1.20). In Section 1.4 we showed that the problem (5.3) can be represented as a linear algebraic system

$$A_j x = b_j\,, \tag{5.5}$$

where the exact solution x corresponds to the Galerkin solution u_j. Hereafter we denote by x the solution of the system (5.5) for any $j = 0, 1, \ldots, k$. From the context it will be clear which discretization level j is considered. We

reserve g, j, k for denoting the levels of the discretization and i, ℓ, n, m for denoting the iteration steps of the CG method.

All matrices A_j are SPD (as the bilinear form $a(\cdot,\cdot)$ is symmetric and positive-definite) so that the conjugate gradient method can be applied. As showed in Section 2.2.2, the approximation x_i given by the CG method applied to system $A_j x = b_j$ with the initial vector x_0 minimizes the algebraic energy norm of the error over the Krylov subspace generated by A_j and the initial residual $r_0 = b_j - A_j x_0$

$$\|x - x_i\|_{A_j} = \min_{u \in x_0 + \mathcal{K}_i(A_j, r_0)} \|x - u\|_{A_j} . \tag{5.6}$$

The CCG method combinates the Galerkin FEM discretization to minimize the discretization error (measured in the energy norm) and the CG method to minimize the algebraic energy norm of the error on each discretization level j.

The CCG method is based on *cascading principle* that we describe on a single level $j > 0$:

Suppose that we have (from the previous level) the Galerkin solution u_{j-1} and marked elements of the previous mesh to refine. Then we proceed in the following way:

1) refine marked elements (and the neighboring ones in order to keep the conformity of the mesh)

2) assembly the matrix A_j and the right-hand side vector b_j

3) interpolate the solution u_{j-1} from the previous level to the new mesh as $u_j^{(0)} \in S_j$

4) denote by x_0 the vector of the coefficients of $u_j^{(0)}$ with respect to the given basis of S_j, see (1.15) [2]

5) solve the system $A_j x = b_j$ by CG with the initial vector x_0

6) compute the Galerkin solution u_j corresponding to the exact solution x

7) compute the global error estimate and test the convergence; we assume that the global estimator is given as the sum of the local error estimates

[2] for P^1 FEM discretization described in Section 1.4.1 the vector x_0 is given by the values of $u_j^{(0)}$ in the vertices of the triangulation

8) use the local estimates for marking the elements with the highest contribution to the global error (or mark all elements when using the uniform refinement)

On the coarsest grid $j = 0$, we replace 1) – 5) by the direct solution of the (small) system $A_0 x = b_0$ and we continue from 6).

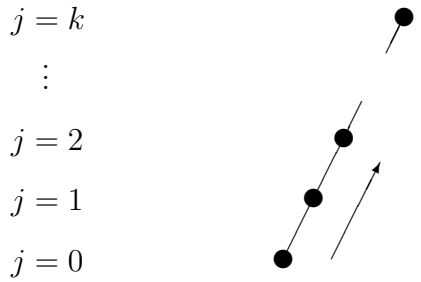


Figure 5.1: *Scheme of the CCG method*

The scheme of the CCG method is shown in Figure 5.1. The CCG method is, e.g. in [38, 39, 41], considered as the simpler version of a multigrid method without coarse-grid correction. However, the CCG method is in principle different from multigrid methods (see Section 4.1 and compare Figure 5.1 with Figure 4.4).

Providing that S_j are nested, the Galerkin orthogonality, see (1.19)

$$a(u - u_{j+1}, v_{j+1}) = 0\,, \quad \forall v_{j+1} \in S_{j+1}$$

gives

$$a(u - u_{j+1}, u_{j+1} - u_j) = 0\,.$$

Then for the Galerkin solutions u_j and u_{j+1}

$$\begin{aligned}
\|u - u_j\|_a^2 &= a(u - u_j, u - u_j) = \\
&= a(u - u_{j+1} + u_{j+1} - u_j, u - u_{j+1} + u_{j+1} - u_j) = \\
&= a(u - u_{j+1}, u - u_{j+1}) + 2\,a(u - u_{j+1}, u_{j+1} - u_j) + \\
&\quad + a(u_{j+1} - u_j, u_{j+1} - u_j) = \\
&= \|u - u_{j+1}\|_a^2 + \|u_{j+1} - u_j\|_a^2\,,
\end{aligned}$$

which gives the relation

$$\|u - u_{j+1}\|_a^2 = \|u - u_j\|_a^2 - \|u_{j+1} - u_j\|_a^2 \,. \tag{5.7}$$

From the cascading principle we have $u_j = u_{j+1}^{(0)}$ and using the equality of the energy norms (1.17) we get the relation

$$\|u_{j+1} - u_j\|_a^2 = \|u_{j+1} - u_{j+1}^{(0)}\|_a^2 = \|x - x_0\|_{A_{j+1}}^2 \,. \tag{5.8}$$

The relations (5.7) and (5.8) give

$$\|u - u_j\|_a^2 - \|u - u_{j+1}\|_a^2 = \|x - x_0\|_{A_{j+1}}^2 \,, \tag{5.9}$$

which means that the reduction of the discretization error from the discretization level j to $j+1$ (measured in the enrgy norm) is equal to to the error of the initial CG approximation x_0 on the discretization level $j+1$ (measured in the algebraic energy norm).

In finite precision computations we are not able to compute the exact solution x of the system $A_j x = b_j$. Following [10], we therefore consider the cascading principle changed, solving the system $A_j x = b_j$ on finer meshes ($j > 0$) approximately by the CG method. For $j > 0$:

1) refine marked elements (and the neighboring ones in order to keep the conformity of the mesh);

2) assembly the matrix A_j and the right-hand side vector b_j

3) interpolate the approximation $\tilde{u}_{j-1}$ from the previous level to the new mesh as $\tilde{u}_j^{(0)}$

4) denote by $\tilde{x}_0$ the vector of the coefficients of $\tilde{u}_j^{(0)}$ with respect to the given basis of S_j

5) apply CG for the system $A_j x = b_j$ with the initial vector $\tilde{x}_0$ giving the approximation $\tilde{x}_{m_j}$ to the exact solution x

6) compute the Galerkin approximation $\tilde{u}_j \equiv \tilde{u}_j^{(m_j)}$ given by the CG approximation $\tilde{x}_{m_j}$

7) compute the global error estimate and test the convergence

8) use local estimates for marking the elements with the highest contribution to the global error (when using the uniform refinement we mark all elements)

Please note, that since $\tilde{u}_{j-1} \neq u_{j-1}$, $\tilde{x}_0 \neq x_0$ and the CG convergence (that is affected by the choice of the initial vector, see Chapter 2) is different from the "exact" case, i.e. $\tilde{x}_i \neq x_i\,, i = 1, 2, \ldots, m_j\,.$

Using (5.4), the error $\|u - \tilde{u}_j\|_a$ satisfy

$$\|u - \tilde{u}_j\|_a \geq \|u - u_j\|_a\,.$$

As a consequence of the cascading principle, Theorem 2.2 and the local orthogonality of CG residuals (2.41)

$$\begin{aligned}
\|u - \tilde{u}_j\|_a^2 &= \|u - \tilde{u}_{j+1}^{(0)}\|_a^2 = \\
&= \|u - u_{j+1}\|_a^2 + \|x - \tilde{x}_0\|_{A_{j+1}}^2 = \\
&= \|u - u_{j+1}\|_a^2 + \|x - \tilde{x}_{m_{j+1}}\|_{A_{j+1}}^2 + \|\tilde{x}_{m_{j+1}} - \tilde{x}_0\|_{A_{j+1}}^2 = \\
&= \|u - \tilde{u}_{j+1}^{(m_{j+1})}\|_a^2 + \|\tilde{u}_{j+1}^{(m_{j+1})} - \tilde{u}_{j+1}^{(0)}\|_a = \\
&= \|u - \tilde{u}_{j+1}\|_a^2 + \|\tilde{u}_{j+1} - \tilde{u}_j\|_a^2\,,
\end{aligned}$$

which give the relation

$$\|u - \tilde{u}_{j+1}\|_a^2 = \|u - \tilde{u}_j\|_a^2 - \|\tilde{u}_{j+1} - \tilde{u}_j\|_a^2\,. \tag{5.10}$$

Analogously to (5.8)

$$\|\tilde{u}_{j+1} - \tilde{u}_j\|_a^2 = \|\tilde{x}_{m_{j+1}} - \tilde{x}_0\|_{A_{j+1}}^2\,. \tag{5.11}$$

and

$$\|u - \tilde{u}_j\|_a^2 - \|u - \tilde{u}_{j+1}\|_a^2 = \|\tilde{x}_{m_{j+1}} - \tilde{x}_0\|_{A_{j+1}}^2\,. \tag{5.12}$$

The illustration of errors in the CCG method is given in Figure 5.2.

5.2 Stopping criteria for the CCG method

In Section 5.2.1 we recall the original stopping criteria for the CCG method presented in [10] and briefly comment their reliability. In Section 5.2.2 we propose new stopping criterion for the CG method using the estimate for the algebraic error presented in Section 2.3.1 and the heuristic proposed in Section 2.4. Then in Section 5.2.3 we propose new stopping criterion for the Galerkin discretization.

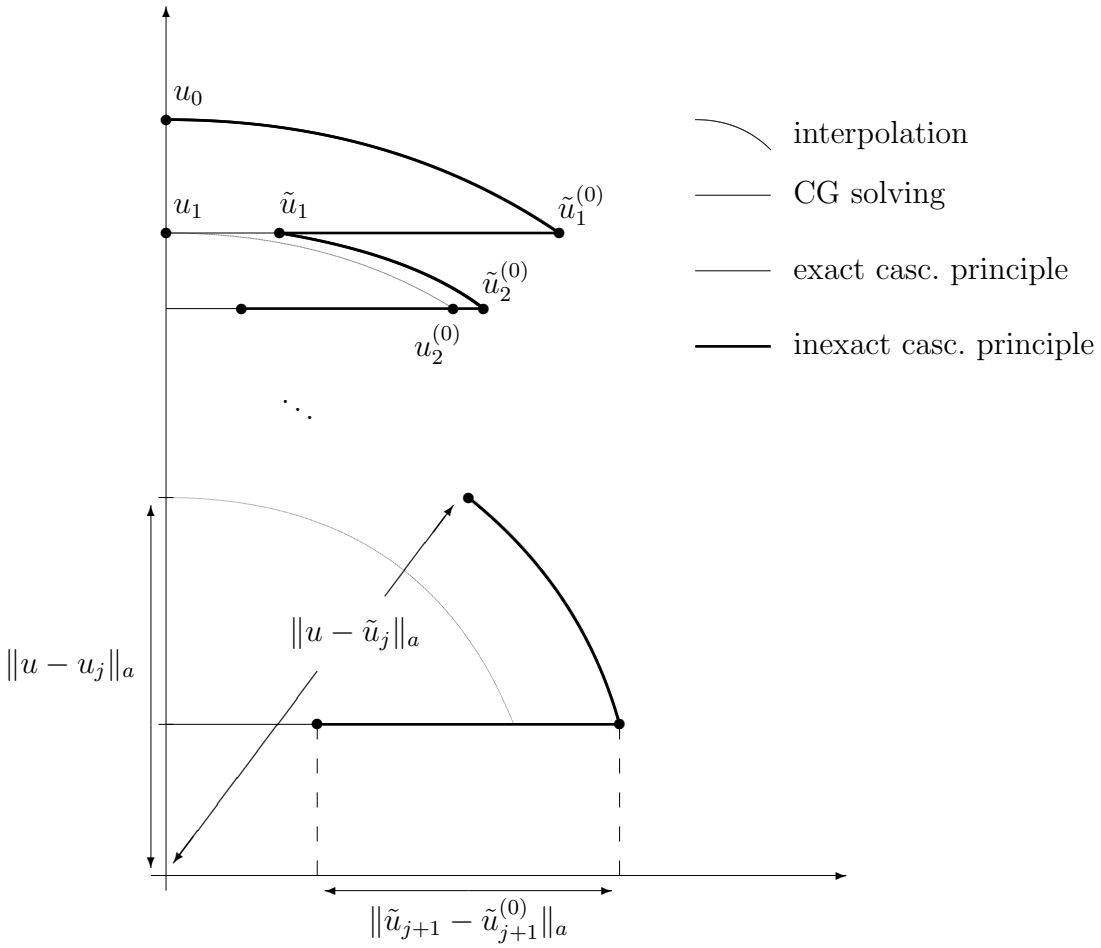


Figure 5.2: *CCG error scheme*

5.2.1 Original stopping criteria

Demanding the reduction of the discretization error given by the user prescribed parameter ϵ_{red}^2

$$\|u - u_j\|_a^2 \leq \epsilon_{red}^2 \|u - u_0\|_a^2 \,, \tag{5.13}$$

the stopping criterion for the Galerkin FEM discretization is in [10] proposed as

$$\frac{\tilde{\Theta}_j \, \|\tilde{u}_j - \tilde{u}_{j-1}\|_a^2}{1 - \Theta_h} \leq \epsilon_{red}^2 \left(\sum_{g=1}^{j} \|\tilde{u}_g - \tilde{u}_{g-1}\|_a^2 \right) , \tag{5.14}$$

where

$$\begin{aligned} \tilde{\Theta}_j &\equiv \left(\frac{n_{j-1}}{n_j} \right)^{2/d} , \quad (n_j = \dim(A_j)) , \\ \Theta_h &\equiv \frac{1}{4} . \end{aligned}$$

The term $\|\tilde{u}_g - \tilde{u}_{g-1}\|_a^2$ can be evaluated using (5.11) as

$$\|\tilde{u}_g - \tilde{u}_{g-1}\|_a^2 = \|\tilde{x}_{m_g} - \tilde{x}_0\|_{A_g}^2 .$$

However, as we will see in the numerical examples, the fulfillment of (5.14) need not assure the fulfillment of the condition (5.13).

The stopping criterion for the CG method on the discretization level j is in [10] proposed as

$$\frac{\|\tilde{x}_i - \tilde{x}_{i-1}\|_{A_j}^2}{1 - \frac{\|\tilde{x}_i - \tilde{x}_{i-1}\|_{A_j}^2}{\|\tilde{x}_{i-1} - \tilde{x}_{i-2}\|_{A_j}^2}} \leq \hat{\rho}^2 \epsilon_{red}^2 \left(\sum_{g=0}^{j} \|\tilde{u}_g - \tilde{u}_{g-1}\|_a^2 \right) , \tag{5.15}$$

where $\hat{\rho}$ is a parameter empirically chosen as $1/4$. The criterion (5.15) is derived assuming the existence of the contraction factor $\Theta < 1$ such that there exist index i_0 satisfying

$$\|x - \tilde{x}_i\|_{A_j}^2 \leq \Theta \, \|x - \tilde{x}_{i-1}\|_{A_j}^2 , \quad \forall i > i_0 , \tag{5.16}$$

In Figure 2.2 on page 35 we can see the stages of near stagnation where $\Theta \approx 1$. Moreover, the unknown factor Θ is in (5.15) estimated using the ratio

$$\frac{\|\tilde{x}_i - \tilde{x}_{i-1}\|_{A_j}^2}{\|\tilde{x}_{i-1} - \tilde{x}_{i-2}\|_{A_j}^2} ,$$

that may give even more misleading information about the CG convergence in finite precision computations (see, e.g., [28]).

5.2.2 Stopping criterion for the CG method

We regard the setting of the proper stopping criterion for CG essential for the effective implementation of CCG. Too stringent criterion leads to CG iterations that are not needed whereas too loose one causes mesh refinement to be ineffective (since the total error is dominated by the algebraic error).

According to Theorem 2.2 we demand on level j the balance of discretization and algebraic error

$$\|x - \tilde{x}_i\|_{A_j}^2 = \|u_j - \tilde{u}_j^{(i)}\|_a^2 \leq \rho_j \|u - u_j\|_a^2 \,, \quad \rho_j \in [0,1] \,. \tag{5.17}$$

As we showed in numerical examples in Section 2.6.1, this criterion controls only the *global* balance of the errors and the *local* distribution of the algebraic and the discretization error may differ. However, we do not have any local estimator for the algebraic error to control the balance of the local contributions to the total error.

The choice $\rho_j = 0$ in (5.17) assumes the exact solution of the Galerkin system. Using the cascading principle of CCG

$$\|u - u_j\|_a^2 = \|u - \tilde{u}_{j-1}\|_a^2 - \|x - \tilde{x}_0\|_{A_j}^2 \,, \quad j = 1, \dots \,, \tag{5.18}$$

where $\tilde{x}_0$ corresponds to the interpolation of the approximation $\tilde{u}_{j-1}$ from the previous level to the new mesh. The relations (5.17) and (5.18) give on the level $j = 1, 2, \dots$

$$\|x - \tilde{x}_i\|_{A_j}^2 \leq \rho_j \left(\|u - \tilde{u}_{j-1}\|_a^2 - \|x - \tilde{x}_0\|_{A_j}^2 \right) . \tag{5.19}$$

In numerical experiments in Section 5.4 we will see that for fixed value $\rho_j = \rho$, $j = 1, 2, \dots$, the number of CG iterations required for the fulfillment of (5.19) is increasing with the size of the matrix A_j where the CG iterations are more costly. In order to prevent this increase we propose to set ρ_j such that the criterion (5.19) is more stringent on the coarser levels.

Fixing the values $\rho_0 = 0$ and $\rho_k = 1$ (k stands for the finest level), we consider for a parameter $\xi \geq 1$ the choice

$$\rho_j = \xi^{j-k} \,, \quad j = 1, \dots, k \,. \tag{5.20}$$

Note that $\xi = 1$ gives $\rho_j = 1$, $j = 1, \dots, k$.

Proposition 1. *Let ρ_j be given by (5.20) with ξ satisfying*

$$\xi > \frac{\|u-u_j\|_a^2}{\|u-u_{j+1}\|_a^2}\,, \quad \forall j = 0,\ldots,k-1\,. \tag{5.21}$$

Then for $j = 1,\ldots,k-1$

$$\rho_{j+1}\left(\|u-\tilde{u}_j\|_a^2 - \|x-\tilde{x}_0\|_{A_{j+1}}^2\right) > \rho_j\left(\|u-\tilde{u}_{j-1}\|_a^2 - \|x-\tilde{x}_0\|_{A_j}^2\right)\,,$$

i.e. the criterion (5.19) is more stringent on the coarser level.

Proof. Using (5.18),

$$\rho_{j+1}\left(\|u-\tilde{u}_j\|_a^2 - \|x-\tilde{x}_0\|_{A_{j+1}}^2\right) = \rho_{j+1}\|u-u_{j+1}\|_a^2 = \xi\rho_j\,\|u-u_{j+1}\|_a^2 >$$
$$> \frac{\|u-u_j\|_a^2}{\|u-u_{j+1}\|_a^2}\,\rho_j\,\|u-u_{j+1}\|_a^2 = \rho_j\left(\|u-\tilde{u}_{j-1}\|_a^2 - \|x-\tilde{x}_0\|_{A_j}^2\right)\,.$$

□

Combinating (5.19) and (5.20) we get the (theoretical) stopping criterion for the CG method on the discretization level $j = 1,2,\ldots,k$ (we recall that k stands for the finest level)

$$\|x-\tilde{x}_i\|_{A_j}^2 \le \xi^{j-k}\left(\|u-\tilde{u}_{j-1}\|_a^2 - \|x-\tilde{x}_0\|_{A_j}^2\right)\,. \tag{5.22}$$

For estimating the error $\|x-\tilde{x}_0\|_{A_j}$ of the initial CG approximation we consider the lower bound

$$\nu_{0,i+d} = \sum_{\ell=0}^{i+d}\gamma_\ell\|r_\ell\|^2\,,$$

see Section 2.3.1.

We estimate the algebraic error $\|x-\tilde{x}_i\|_{A_j}$ using the bound $\nu_{i,d}$ with the adaptively chosen value of d given by the heuristic proposed in Section 2.4. In [2] the estimates for $\|x-\tilde{x}_i\|_{A_j}$ are compared in three numerical examples. The estimate $\nu_{i,d}$ for fixed value $d = 5$ gives worse results in comparison to upper bounds. However, we believe that using the adaptive choice of the parameter d significantly improves the accuracy of the lower bound $\nu_{i,d}$.

Let η_j be an estimator for the discretization error $\|u-u_j\|_a^2$. As

$$\|u-\tilde{u}_j^{(i)}\|_a^2 = \|u-u_j\|_a^2 + \|x-\tilde{x}_i\|_{A_j}\,,$$

see Theorem 2.2, we consider the estimator for the total error $\|u - \tilde{u}_j^{(i)}\|_a^2$

$$\eta_j^{(i)} \equiv \eta_j + \nu_{i,d}\,.$$

Using these estimates in (5.22), we get the criterion

$$\nu_{i,d} \leq \xi^{j-k} \left(\eta_{j-1}^{(m_{j-1})} - \nu_{0,i+d} \right) , \quad j = 1, \ldots, k\,.$$

Since the positivity of the term $\eta_{j-1}^{(m_{j-1})} - \nu_{0,i+d}$ is assured only theoretically, it need not hold in finite precision computations. Hence we include also the other estimate for $\|u - u_j\|_a^2$. Assuming (5.21),

$$\|u - u_j\|_a^2 > \frac{1}{\xi} \|u - u_{j-1}\|_a^2\,,$$

we get

$$\|u - u_j\|_a^2 \;\gtrapprox\; \frac{\eta_{j-1}}{\xi}\,.$$

Then we set the stopping criterion for the CG method as

$$\nu_{i,d} \leq \xi^{j-k} \;\max \left((\eta_{j-1}^{(m_{j-1})} - \nu_{0,i+d}), \frac{\eta_{j-1}}{\xi} \right) , \quad j = 1, \ldots, k\,. \tag{5.23}$$

The choice of parameter ξ is to be further studied and examined. See the numerical experiments in Section 5.4 for the comparison of various choices.

5.2.3 Stopping criterion for the Galerkin FEM discretization

For the Galerkin FEM discretization we consider following [10] the standard stopping criterion

$$\|u - u_j\|_a^2 \leq \epsilon_{red}^2 \|u - u_0\|_a^2 \tag{5.24}$$

demanding the reduction of discretization error prescribed by the parameter ϵ_{red}. Assuming the fulfillment of (5.17), the criterion (5.24) controls also the reduction of the total error

$$\|u - \tilde{u}_j\|_a^2 \leq (1 + \rho_j)\, \epsilon_{red}^2\, \|u - \tilde{u}_0\|_a^2\,. \tag{5.25}$$

Using the estimate η_j the stopping criterion for the Galerkin FEM discretization is set as

$$\eta_j \leq \epsilon_{red}^2\, \eta_0\,. \tag{5.26}$$

The choice of the estimator η_j depends on the solved problem and the chosen discretization. For the second-order elliptic pure diffusion problem with the pure Dirichlet boundary condition (see Section 1.2) discretized by the P^1-conforming FEM discretization described in Section 1.4.1 we propose the estimate η_j defined in (3.15) in Section 3.2.1.

5.3 Adaptive refinement

Each iteration j of adaptive finite element method (AFEM) (i.e. the FEM method using the adaptive mesh refinement) can be summarized in the following workflow (see, e.g., [2]):

$$\text{SOLVE} \rightarrow \text{ESTIMATE} \rightarrow \text{MARK} \rightarrow \text{REFINE}$$

The first step includes the computing of the approximation $\tilde{u}_j$ to the Galerkin solution u_j. In the second step we evaluate the *local* estimators for each edge $E \in \mathcal{E}_j$ or for each element of the triangulation $K \in T_j$. Then we mark elements for refinement forming the set $\mathcal{M}_j \subset T_j$. In the fourth step the marked elements are refined using a proper refinement technique and the conformity of mesh is provided by a sufficient additional refinement.

In Section 2.6.1 we showed that the local distribution of the algebraic and the discretization error can significantly differ. However, to the best of our knowledge, there is not yet described any estimator for the local algebraic error. In the implementation of the CCG method we therefore consider the adaptive refinement technique according to the local estimators for the element contributions to the discretization error.

Let the estimator η_j for the discretization error is given as the sum of element contributions $\eta_j^K, K \in T_j$

$$\eta_j = \sum_{K \in T_j} \eta_j^K \,.$$

We mark elements for refinement employing the Dörfler-type marking strategy (see [11]). For a user-defined parameter $0 < \theta \leq 1$ we find a set $\mathcal{M}_j$ of minimal cardinality such that

$$\sum_{K \in \mathcal{M}_j} \eta_j^K \geq \theta \sum_{K \in T_j} \eta_j^K \,. \tag{5.27}$$

The choice $\theta = 1$ stands for the uniform mesh refinement. In numerical experiments we use $\theta = 0.75$. We refine the elements using the red-green-blue (RGB) refinement strategy that avoids degeneracies, i.e. maintain the minimum angle condition, see, e.g., [8]. We believe that the newest-vertex-bisection (NVB, [29]), the other commonly used technique, would give nearly the same results.

5.4 Numerical experiments

In this section we test the CCG method with the newly proposed stopping criteria on Problems 1–4, described on pages 39, 58, 59 and 61. We start with the comparison of the choice of parameter ξ defined in (5.20). Then we test the stopping criteria proposed in Section 5.2 and compare them with the original stopping criteria proposed in [10]. Finally we show the convergence rate of the CCG method.

We compare the choices of the parameter ξ defined in (5.20) according to the number of CG iterations required for the fulfillment of the criterion (5.19) on the levels $j = 1, 2, \ldots, 6$. In this experiment we evaluate the values of $\|x - \tilde{x}_0\|_{A_j}$, $\|x - \tilde{x}_i\|_{A_j}$, $\|u - u_j^{(i)}\|_a$ and $\|u - u_j\|_a$ using the exact solution u and the Galerkin approximation u_j given by the exact solution x approximated (to a sufficient accuracy) using the MATLAB backslash solver `Aj \ b`.

From the Tables 5.1– 5.4 it is evident that the proper choice of ξ can significantly improve the efficiency of CCG. For $\xi = 1$ (i.e. constant ρ_j) the number of CG iterations required for the fulfillment of the criterion (5.19) is increasing with the size of the matrices A_j. We can see that higher values of ξ lead to more CG iterations on lower and middle levels and give no major improvement (in comparison to $\xi = 3$ or $\xi = 4$) on finer levels.

The optimal value ξ_{opt} depends on the problem and on Dörfler marking parameter θ, see (5.27). For the choice $\theta = 0.75$ we expect $\xi_{opt} \in [3, 4]$.

Problem 1

$$2.3929 \leq \frac{\|u-u_{j-1}\|_a^2}{\|u-u_j\|_a^2} \leq 3.2667$$

$\xi \setminus j$	1	2	3	4	5	6
1	2	2	8	15	27	49
2	6	10	8	12	17	29
3	8	12	8	5	8	3
4	9	14	9	4	3	2
8	13	18	19	5	3	2
16	16	23	27	8	3	2

Table 5.1: *The number of CG iterations on the levels* $j = 1, 2, \ldots, 6$ *for different* ξ *in Problem 1.*

Problem 2

$$\frac{\|u-u_0\|_a^2}{\|u-u_1\|_a^2} = 1.3487\,, \quad 2.6597 \leq \frac{\|u-u_{j-1}\|_a^2}{\|u-u_j\|_a^2} \leq 2.9086\,, \quad 2 \leq j \leq 6$$

$\xi \setminus j$	1	2	3	4	5	6
1	1	2	5	11	16	31
2	3	6	8	11	12	23
3	4	9	11	10	13	4
4	6	12	14	12	5	3
8	16	22	18	17	4	2
16	19	26	34	21	5	2

Table 5.2: *The number of CG iterations on the levels* $j = 1, 2, \ldots, 6$ *for different* ξ *in Problem 2.*

Problem 3

$1.9706 \leq \frac{\|u-u_{j-1}\|_a^2}{\|u-u_j\|_a^2} \leq 2.2539$

$\xi \setminus j$	1	2	3	4	5	6
1	2	3	10	16	26	38
2	8	12	9	12	14	12
3	15	16	13	7	5	3
4	20	20	17	8	4	2
8	25	29	29	12	4	2
16	30	34	41	21	5	2

Table 5.3: *The number of CG iterations on the levels* $j = 1, 2, \ldots, 6$ *for different* ξ *in Problem 3.*

Problem 4

$1.7811 \leq \frac{\|u-u_{j-1}\|_a^2}{\|u-u_j\|_a^2} \leq 1.8524$

$\xi \setminus j$	1	2	3	4	5	6
1	3	8	12	17	23	32
2	17	19	20	22	20	18
3	23	27	25	29	18	9
4	27	31	31	34	22	9
8	33	38	41	42	40	9
16	37	43	48	50	53	9

Table 5.4: *The number of CG iterations on the levels* $j = 1, 2, \ldots, 6$ *for different* ξ *in Problem 4.*

Now we present the results of the CCG method with the new stopping criteria proposed in Section 5.2 applied to the Problems 1–4. We consider the estimate for the discretization error η_j defined in (3.15) in Section 3.2.1. The estimate η_j measures the jumps of the gradient of the approximation $u_j^{(i)}$ over the edges of the triangulation, see Section 3.2. We refine the mesh adaptively according to the element contributions to the estimator η_j as described in Section 5.3. We set the parameters

- $\sigma = 0.4 \, \|A\|^{-1/2}$, the safety parameter for the heuristic for the adaptive choice of d, see Section 2.6.3;
- $\xi = 3.5$, see the experiment above;
- $C_1 = 0.04$ for $\alpha_{max} = 90°$ and $C_1 = 0.033$ for $\alpha_{max} = 60°$, the constant in the estimate η_j, see Section 3.2.1;
- $\epsilon_{red}^2 = 0.01$, the demanded reduction of the discretization error in (5.26);

We recall that the squared energy norms of the error satisfy (up to a small innacuracy proportional to machine precision) the Galerkin orthogonality relation

$$\|u - u_j^{(i)}\|_a^2 = \|u - u_j\|_a^2 + \|u_j - u_j^{(i)}\|_a^2 = \|u - u_j\|_a^2 + \|x - \tilde{x}_i\|_{A_j}^2 \,,$$

see Theorem 2.2.

For the particular problem we plot the energy norm of the total, the discretization and the algebraic error and the square roots of the estimates $\eta_j, \nu_{i,d}$.

In Figures 5.3–5.6 we see that η_j gives a tight estimate for the discretization error. The estimate $\nu_{i,d}$ lost its accuracy for higher j where the number of CG iteration decreases. However, the balancing of the energy norms of the algebraic and the discretization error on the finest level k prescribed by the criterion (5.17) is fulfilled, i.e.

$$\|x - \tilde{x}_{m_k}\|_{A_k}^2 = \|u_k - \tilde{u}_k^{(m_k)}\|_a^2 \leq \|u - u_k\|_a^2 \,.$$

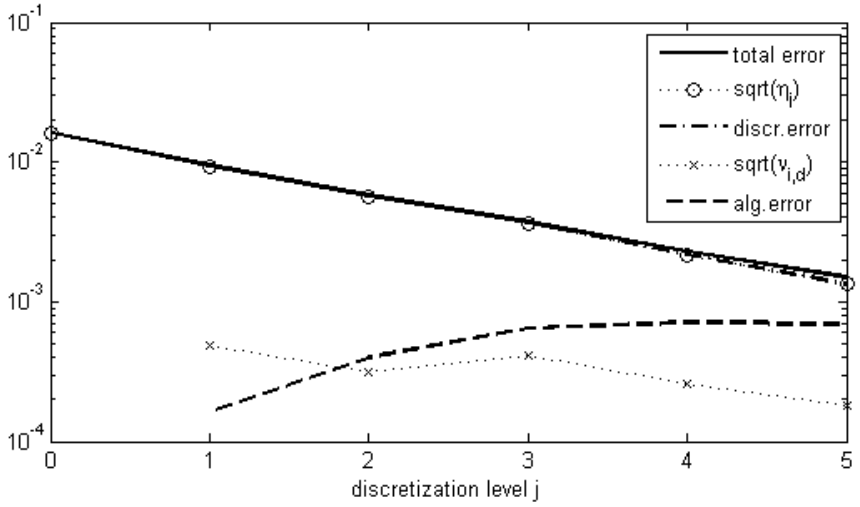


Figure 5.3: *The errors and the estimates in CCG applied to Problem 1.*

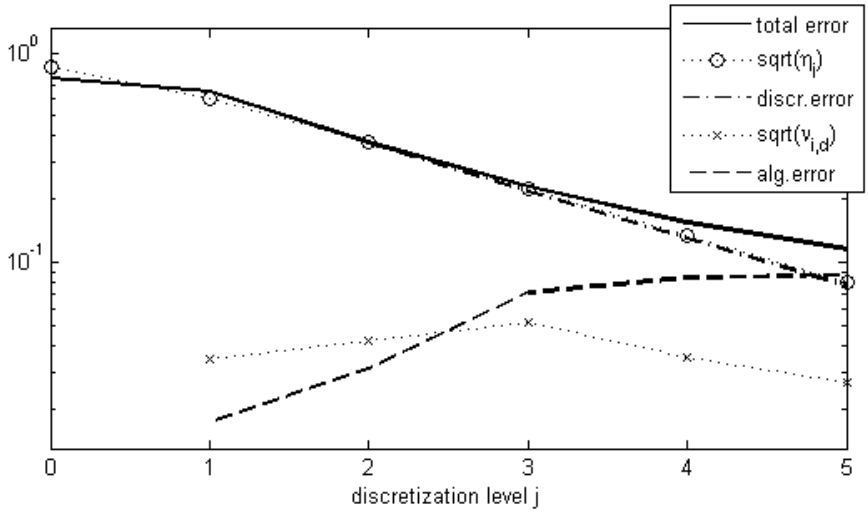


Figure 5.4: *The errors and the estimates in CCG applied to Problem 2.*

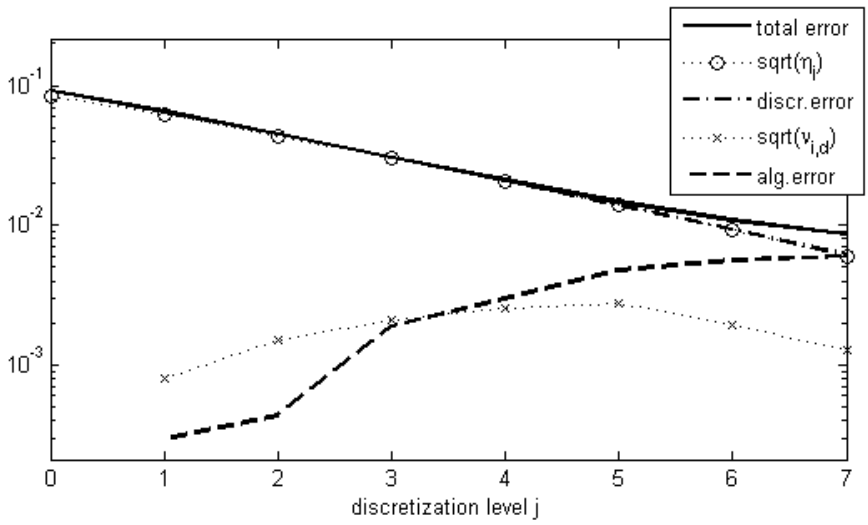


Figure 5.5: *The errors and the estimates in CCG applied to Problem 3.*

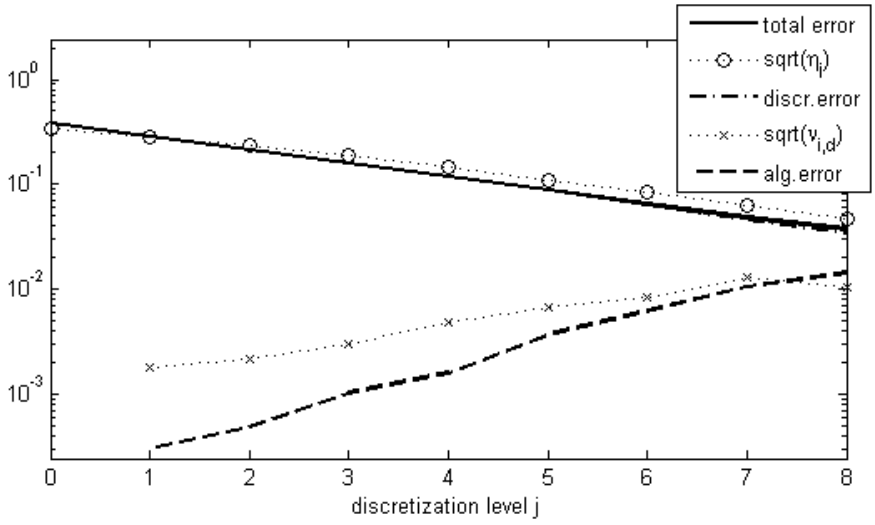


Figure 5.6: *The errors and the estimates in CCG applied to Problem 4.*

In the following set of figures 5.7–5.10 we compare the total and the algebraic errors in the CCG method with the newly proposed stopping criteria and the errors in the original implementation proposed in [10] (using the adaptive mesh refinement described in Section 5.3).

We can see that for Problems 1, 2 the original implementation of CCG gives the similar results as the new one. In Problem 3 and 4 the original implementation stopped on the discretization level $j = 6$, (resp. $j = 7$) when the condition (5.24) is not fulfilled, i.e. the ratio of the initial and the actual discretization error is larger than the prescribed tolerance $\epsilon^2_{red} = 0.01$

$$\frac{\|u - u_j\|^2_a}{\|u - u_0\|^2_a} > \epsilon^2_{red}\,.$$

In Problem 3 this ratio is equal to 0.0104, in Problem 4

$$\|u - u_7\|^2_a / \|u - u_0\|^2_a = 0.0149\,.$$

Moreover, in Problem 4 the algebraic error grows over the value of the discretization error, see Figure 5.10.

The number of performed CG iterations is compared in Table 5.5. We present also the size n_j of matrices A_j for the newly proposed implementation of the CCG method. The sizes of matrices for the original implementation are slightly different as the estimate η_j (and its local contributions) depends on the CG approximation $\tilde{x}_{m_j}$.

Problem 1

discr. level j	1	2	3	4	5
n_j	437	1247	3005	8337	23233
m_j^{new}	15	8	4	4	4
m_j^{orig}	8	8	5	3	3

Problem 2

discr. level j	1	2	3	4	5
n_j	235	313	593	1358	3684
m_j^{new}	6	10	4	4	5
m_j^{orig}	13	14	6	6	4

Problem 3

discr. level j	1	2	3	4	5	6	7
n_j	467	717	1504	3075	6561	13859	31642
m_j^{new}	25	30	21	12	4	4	4
m_j^{orig}	46	24	15	8	5	3	—

Problem 4

discr. level j	1	2	3	4	5	6	7	8
n_j	403	625	1094	2089	4073	7819	14704	27376
m_j^{new}	27	32	37	39	34	18	8	4
m_j^{orig}	10	10	8	7	8	7	5	—

Table 5.5: *Comparison of the number of performed CG iterations m_j^{new} for the newly proposed CCG implementation and m_j^{orig} for the original implementation due to [10].*

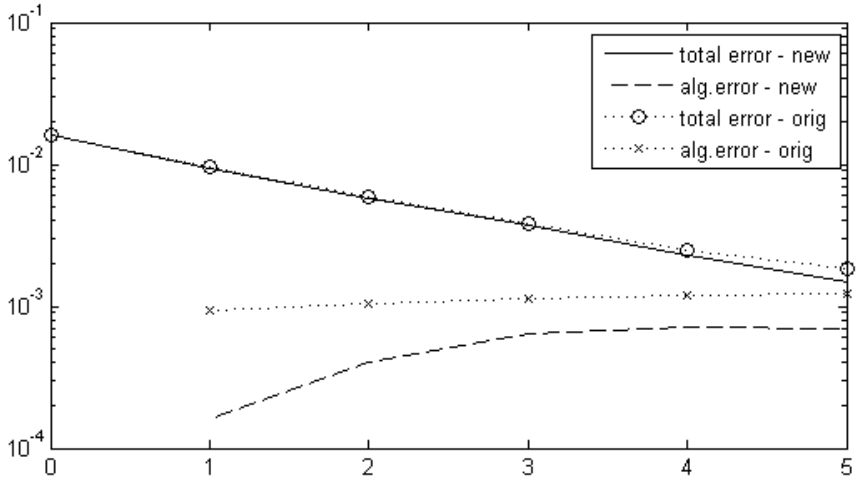


Figure 5.7: *Total and algebraic errors for the new and the original implementation of the CCG method in Problem 1.*

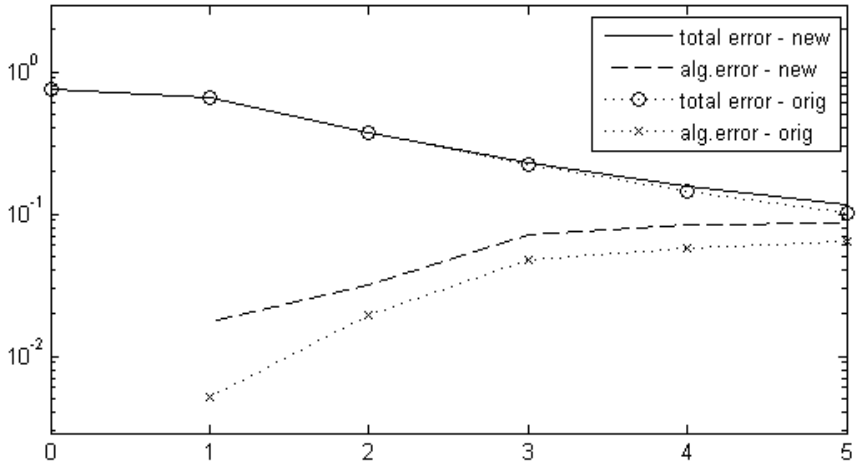


Figure 5.8: *Total and algebraic errors for the new and the original implementation of the CCG method in Problem 2.*

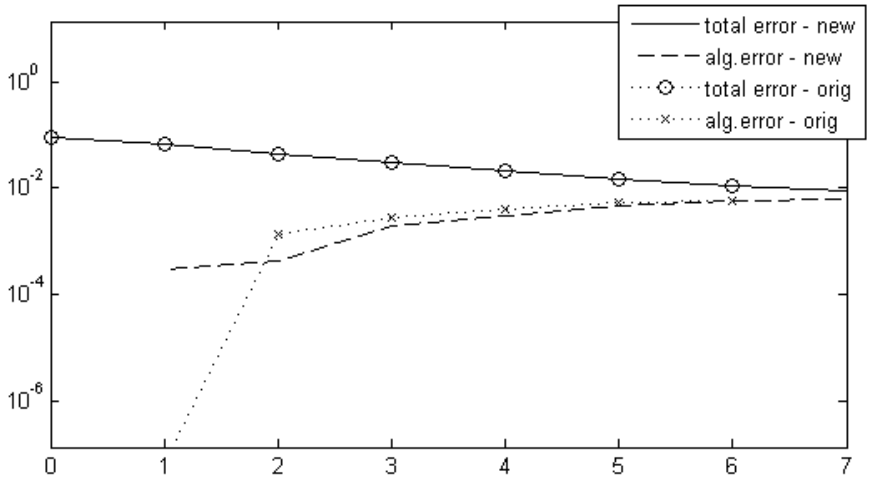


Figure 5.9: *Total and algebraic errors for the new and the original implementation of the CCG method in Problem 3.*

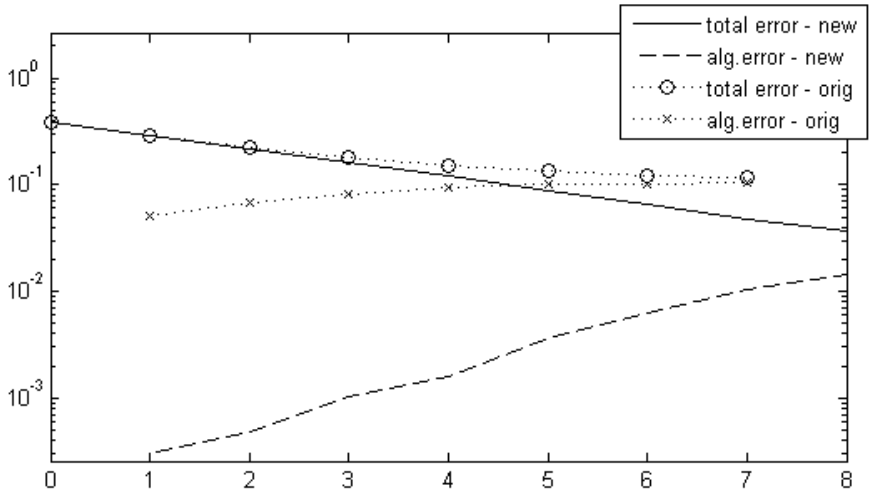


Figure 5.10: *Total and algebraic errors for the new and the original implementation of the CCG method in Problem 4.*

Although the CCG method with the new stopping criteria does not give in the numerical experiments significantly better results than the original implementation, the new stopping criteria have the following advantages:

- they give the estimate for the actual discretization and algebraic error at every iteration step,
- the estimate for the discretization error is locally-based and it enables the adaptive mesh refinement according to local indicators,
- the new stopping criteria are more reliable.

Finally we show in Figure 5.11 the dependance of the total error on the number of degrees of freedom in the new implementation of the CCG method for Problems 1–4. We can see that the total error behaves like $O(n_j^{-1/2})$, where n_j stands for the number of degrees of freedom (i.e. the size of the linear algebraic system $A_j x = b_j$).

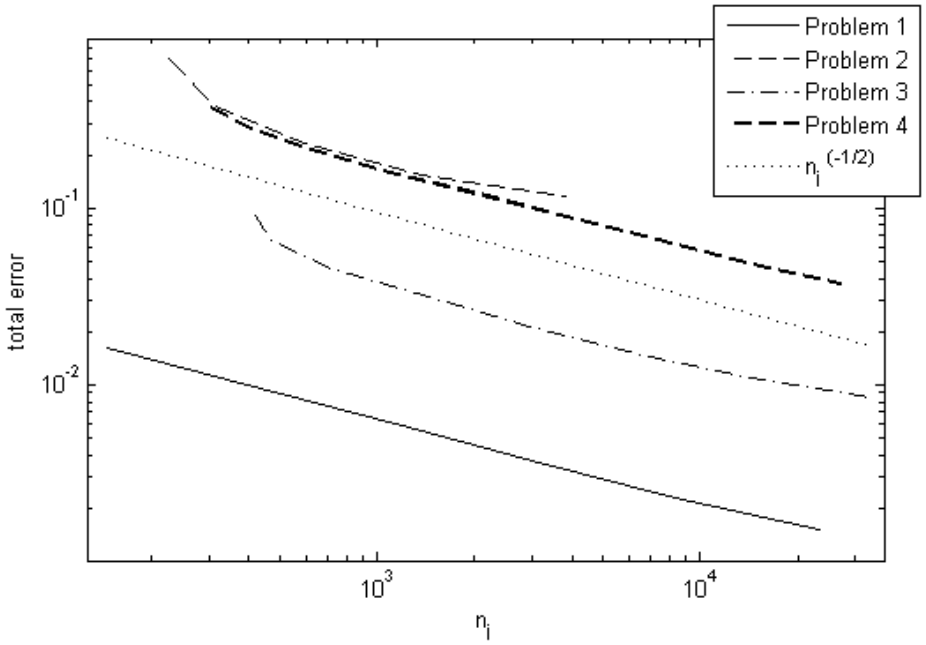


Figure 5.11: *Dependance of the total error on the number of degrees of freedom in the new implementation of the CCG method in Problems 1–4.*

Conclusion

In Chapter 1 of the thesis we have described the model problem and presented some of its basic properties. The discretization of the problem leads to a symmetric positive-definite linear algebraic system. Effective PDE solvers using iterative methods should stop the iteration whenever the algebraic error drops to the level at which it does not significantly affect the total error. Linear system with a SPD matrix can be solved using the conjugate gradient method. The CG method has been described in Chapter 2 including the estimates for the energy norm of the error and the heuristic for the adaptive choice of the parameter of the estimate have been presented. As we have shown on a simple example in Section 2.6.1, the local distribution of the algebraic error can significantly differ from the local distribution of the discretization error. When using the *global* stopping criteria for an algebraic solver we are (generally) not able to find the level at which the algebraic error does not significantly affect the total error *locally*. In Chapter 3 we recall the locally-based estimators for the discretization error for the cell-centered finite volume and for the piecewise linear finite element discretization that take into account the inexact solving of the algebraic system, presented in the literature. After recalling the multigrid method in Chapter 4 we have described in Chapter 5 the CCG method and proposed new stopping criteria relating the algebraic and the discretization error. The new implementation is then tested.

Many questions remain opened. The balance between the algebraic and the discretization error requires further study. This should include the derivation of a locally-based estimator for the algebraic error and setting the local stopping criteria for the algebraic solver. We also intend to compare the CCG method (with the newly proposed stopping criteria) with multigrid methods.

Bibliography

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

[2] **Arioli, M.**, **Georgoulis, E.** and **Loghin, D.** *Convergence of inexact adaptive finite element solvers for elliptic problems.* Technical report, RAL Technical Reports, RAL-TR-2009-021, 2009

[3] **Babuška, I.** and **Rheinboldt, W.** *Error estimates for adaptive finite element computations.* SIAM J. Numer Anal., **15**; pages 736–754, 1978

[4] **Becker, R.** and **Mao, S.** *Convergence and quasi-optimality of a simple adaptive finite element method.* ESAIM: M2AN, **43**; pages 1203 – 1219, 2009

[5] **Brandt, A.** *Multi-level adaptive solutions to boundary value problems.* Math. Comput., **31**; pages 333 – 390, 1977

[6] **Brenner, S.** and **Scott, L.** *The Methamatical Theory of Finite Element Methods.* Springer, 1996

[7] **Briggs, W. L.**, **Henson, V. E.** and **McCormick, S. F.** *A Multigrid Tutorial.* SIAM, second edition, 2000

[8] **Carstensen, C.** *An Adaptive Mesh-Refining Algorithm Allowing for an H^1-Stable L^2-Projection onto Courant Finite Element Spaces.* Constructive Approximation, **20**; pages 549–564, 2004

[9] **Ciarlet, P.** *The Finite Element Methods for Elliptic Problems.* North-Holland, Amsterdam, 1978

[10] **Deuflhard, P.** *Cascading conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results.* Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993), **180**; pages 29 – 42, 1994

[11] **Dörfler, W.** *A convergent adaptive algorithm for Poisson's equation.* SIAM J. Numer Anal., **33** (3); pages 1106–1124, 1996

[12] **Elman, H. C.**, **Silvester, D. J.** and **Wathen, A. J.** *Finite elements and fast iterative solvers; with application in incompressible fluid dynamics.* Numerical Mathematics and Scientific Computation, Oxford University Press, 2005. (Chapter 2)

[13] **Eriksson, K.**, **Estep, D.**, **Hansbo, P.** and **Johnson, C.** *Computational Differential Equations.* Cambridge University Press, 1996

[14] **Evans, L.** *Partial Differential Equations.* Graduate studies in mathematics. American Mathematical Society, 1998

[15] **Eymar, R.**, **Gallouët, T.** and **Herbin, R.** *Finite volume methods.* Handbook of Numerical Analysis, **VII**; page 713–1020, 2000

[16] **Funken, S.**, **Praetorius, D.** and **Wissgott, P.** *Efficient implementation of adaptive P1-FEM in MATLAB.* Technical report, Institute for Analysis and Scientific Computing, Vienna University of Technology, Wien, 2008

[17] **Gautschi, W.** *A survey of Gauss-Christoffel quadrature formulae.* in E.B.Christoffel. The influence of His Work on Mathematics and the Physical Sciences, P.Bultzer and F.Fehér, eds. Birkhausen, 1981

[18] **Gockenbach, M.** *Partial Differential Equations, Analytical and Numerical Methods.* SIAM, 2002

[19] **Golub, G. H.** and **Meurant, G.** *Matrices, moments and quadratures II: How to compute the norm of the error in iterative methods.* BIT, **37**; pages 687 – 705, 1997

[20] **Golub, G. H.** and **Strakoš, Z.** *Estimates in quadratic formulas.* Numer. Algorithm, **8**; pages 253 – 254, 1994

[21] **Hestenes, M. R.** and **Stiefel, E.** *Methods of conjugate gradient for solving linear systems.* J. Research Nat. Bur. Standarts, **49**; pages 409 – 435, 1952

[22] **Higham, N. J.** *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, second edition, 2002

[23] **Jiránek, P.**, **Strakoš, Z.** and **Vohralík, M.** *A posteriori error estimates including algebraic error: computable upper bounds and stopping criteria for iterative solvers.* SIAM Journal on Scientific Computing (SISC), **32**; pages 1567 – 1590, 2010

[24] **Křížek, M.** and **Neittaanmaki, P.** *Finite Element Approximation of Variational Problems and Applications.* Longman Scientific & Technical, Wiley, 1990

[25] **Liesen, J.** and **Strakoš, Z.** *On the interplay between the PDE discretization and numerical solution of the resulting algebraic problems.* (to be published in Foundations of Computational Mathematics)

[26] —. *Principles of Krylov subspace methods.* (to be published by Oxford University Press)

[27] **Meinardus, G.** *Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch.* Numer. Math., **5**; pages 14 – 23, 1963

[28] **Meurant, G.** and **Strakoš, Z.** *The Lanczos and conjugate gradient algorithms in finite precision arithmetic.* Acta Numerica, **15**; pages 471 – 542, 2006

[29] **Morin, P.**, **Nochetto, R.** and **Siebert, K.** *Data Oscillation and Convergence of Adaptive FEM.* SIAM J. Numer. Anal., **38**; pages 466–488, 2000

[30] **Novotný, A.** and **Straškraba, I.** *Introduction to the Mathematical Theory of Compressible Flow.* Oxford University Press, 2004

[31] **Paige, C.** and **Sounders, M.** *LSQR: an algorithm for sparse linear equations and sparse least squares.* ACM Trans. Math. Software, **8**; pages 43 – 71, 1982

[32] **Paige, C.** and **Strakoš, Z.** *Residual and backward bounds in minimum residual Krylov subspace methods.* SIAM J. Sci. Comput, **23**; pages 1898–1923, 2002

[33] **Papež, J.** *Estimation of the energy and Euclidean norms of the error in the conjugate gradient method.* bachelor thesis, Charles University of Prague, 2009. (in Czech)

[34] **Parlett, B. N.** *The symmetric Eigenvalue problem.* Prentice Hall, Englewood Cliffs, 1980

[35] **Repin, S.** *A posteriori error estimation for nonlinear variational problems by duality theory.* Journal of Mathematical Sciences, **99**; pages 927–935, 2000

[36] **Rüde, U.** *Error estimators based on stable splittings.* in Proceedings of the 7th International Conference on Domain decomposition in Science and Engineering Computing, Pennsylvania State University, D. Keyes, ed., **180**; pages 111–118, 1994

[37] **Saad, Y.** *Iterative methods for sparse linear systems.* SIAM, 2003

[38] **Shaidurov, V. V.** *Some estimates of the rate of convergence for the Cascadic Conjugate-Gradient Method.* Technical Report. Otto-von-Guericke-Universitat Magdeburg, 1994

[39] —. *The convergence of the Cascadic Conjugate-Gradient Method under a deficient regularity.* Problems and Methods in Mathematical Physics (L. Jentsch and F. Trőltzsch, eds.), pages 185 – 194, 1994

[40] —. *Multigrid Methods for Finite Elements.* Mathematics and Its Applications, Kluwer Academic Publishers, 1995

[41] **Shaidurov, V. V.** and **Tobiska, L.** *The convergence of the Cascadic Conjugate-Gradient Method applied to elliptic problems in domains with re-entrant corners.* Mathematics of Computation, **69** (230); pages 501 – 520, 1999

[42] **Skeel, R.** *Iterative refinementimplies numerical stability for Gaussian elimination.* Math. Comp., **35**; pages 817 – 832, 1980

[43] **Strakoš, Z.** and **Tichý, P.** *On error estimation in the conjugate gradient method and why it works in finite precision computations.* Electron. Trans. Numer. Anal., **13**; pages 56 – 80, 2002

[44] —. *Error estimation in peconditioned conjugate gradients.* BIT Numerical Mathematics, **45**; pages 789 – 817, 2005

[45] —. *On efficient numerical approximation of the bilinear form $c^*A^{-1}b$.* SIAM Journal on Scientific Computing (SISC), 2011. (to appear)

[46] **Trottenberg, U.**, **Oosterlee, C.** and **Schüller, A.** *Multigrid.* Academic Press, 2001

[47] **Wohlmuth, B.** and **Hoppe, R.** *A comparison of a posteriori error estimators for mixed finite element discretization by Raviart-Thomas elements.* Math. Comp., **68**; pages 1347–1378, 1999