ON DISTRIBUTION OF THE DISCRETIZATION AND ALGEBRAIC ERROR IN 1D POISSON MODEL PROBLEM [*]

J. PAPEŽ [†], J. LIESEN [‡], AND Z. STRAKOŠ [§]

Abstract. In adaptive numerical solution of partial differential equations, the local mesh refinement is used together with a posteriori error analysis in order to equilibrate the discretization error distribution over the domain. Since the discretized algebraic problems are *not solved exactly*, a natural question is whether the distribution of the algebraic error is analogous to the distribution of the discretization error. This paper illustrates on an example of a simple one-dimensional boundary value model problem that this may not hold. On the contrary, the algebraic error can have large local components and it can therefore significantly dominate the total error in some part of the domain. This can happen even if the globally measured algebraic error is comparable to or smaller than the globally measured discretization error.

This phenomenon is on purpose illustrated on the simplest 1D model problem frequently used in literature; the presented discrepancy between the spatial distribution of the discretization and algebraic errors has not been reported, to our knowledge, in this context before.

Key words. Numerical solution of partial differential equations, finite element method, adaptivity, a posteriori error analysis, discretization error, algebraic error, spatial distribution of the error.

AMS subject classifications. 65F10, 65N15, 65N30, 65N22, 65Y20

1. Introduction.

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

The FEM generates an approximate solution in form of a linear combination of basis functions with *local* supports. Each basis function multiplied by the proper coefficient thus approximates the desired solution only locally. The *global* approximation property of the FEM discrete solution is then ensured by solving the linear algebraic system for the unknown coefficients; the linear algebraic system links the local approximation of the unknown function in different parts of the domain. If the linear algebraic system is solved *exactly*, then all is fine. But in practice we do not solve exactly. In hard problems we even *do not want* to achieve a small algebraic error. That might be too costly or even impossible to set; see, e.g., [2, Sections 1–3],

[*] Received . . . , accepted . . . , published . . .

[†] Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Prague, Czech Republic; Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic, E-mail: papez@cs.cas.cz. The research of this author was supported by the GAUK grant 695612, by the ERC-CZ project LL1202 and by the project IAA100300802 of the Grant Agency of the ASCR.

[‡] Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany, E-mail: liesen@math.tu-berlin.de. The work of this author was supported by the Heisenberg Program of Deutsche Forschungsgemeinschaft (DFG).

[§] Faculty of Mathematics and Physics, Charles University in Prague, Sokolovská 83, 186 75 Prague, Czech Republic, E-mail: strakos@karlin.mff.cuni.cz. The research of this author was supported by the GACR grant 201/09/0917 and by the ERC-CZ project LL1202.

[13, Sections 1 and 6], [19, Section 2.6], the discussion in [20, pp. 36 and 72], and [22, Section 1]. Then, however, one should naturally ask whether the spatial distribution of the algebraic error in the domain can significantly differ from the distribution of the discretization error. There is no a priori evidence that these distributions are to be analogous. On the contrary, from the nature of algebraic solvers, either direct or iterative, there seems to be no reason for equilibrating the algebraic error over the domain. Presented results then indeed demonstrate that the algebraic error can have large local components and it can therefore significantly dominate the total error in some part of the domain.

Following the standard methodology used in the numerical PDE literature for decades (see, e.g., [3, 6, 8]), we consider the simplest one-dimensional boundary value problem. Furthermore, in order to plot illustrative figures, we use a small number of discretization nodes. Like in the standard literature we believe that the simplicity of the model problem does not diminish the message. Since the model problem has appeared in a vast amount of literature, it seems surprising that the presented phenomenon has not been reported elsewhere.

The paper is organised as follows. We describe the model problem and present the experimental observations in Section 2. In Section 3 the total error is interpreted via the modification of the discretization mesh. Section 4 explains the local behavior of the algebraic error using the spectral analysis and the approximation properties of the algebraic solver (here the conjugate gradient (CG) method [12]). The paper ends with concluding remarks.

2. Model problem. We consider the one-dimensional Poisson boundary value problem

$$-u''(x) = f(x)\,, \quad 0 < x < 1\,, \qquad u(0) = u(1) = 0\,, \tag{2.1}$$

where $f(x)$ is a given (continuous) function, $0 \leq x \leq 1$. This model problem is frequently used in mathematical literature for illustrations of various analytical as well as numerical phenomena; see, e.g., [6, Section 6.2.2], [8, Section 5.5], [17], [18, Section 3.2.1].

Denoting by $H_0^1(\Omega)$ the standard Sobolev space of functions having square integrable (weak) derivatives in $\Omega \equiv (0,1)$ and vanishing on the end points (in the sense of traces), the weak formulation of (2.1) looks for $u \in H_0^1(\Omega)$ such that

$$a(u,v) = \ell(v) \quad \text{for all } v \in H_0^1(\Omega)\,, \tag{2.2}$$

where

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

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

$$\|v'\| = a(v,v)^{1/2}\,, \quad v \in H_0^1(\Omega)\,. \tag{2.3}$$

We discretize the problem (2.2) by the FEM on the uniform mesh with n inner nodes, i.e. with the mesh size $h = 1/(n+1)$, using the continuous piecewise linear basis functions $\phi_j,\ j = 1, \dots, n$, satisfying

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

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

$$a(u_h, v_h) = \ell(v_h) \quad \text{for all } v_h \in V_h \,. \tag{2.4}$$

The finite-dimensional problem (2.4) can be equivalently formulated as the system of the linear algebraic equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \,, \tag{2.5}$$

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

$$\mathbf{A} = [A_{ij}] \,, \qquad A_{ij} = a(\phi_j, \phi_i), \tag{2.6}$$

$$\mathbf{b} = [b_1, \dots, b_n]^T \,, \qquad b_i = \ell(\phi_i) \,, \qquad i, j = 1, \dots, n \,. \tag{2.7}$$

The solution $\mathbf{x} = [\xi_1, \dots, \xi_n]^T$ of (2.5) contains the coefficients of the Galerkin FEM solution u_h of (2.4) with the respect to the FEM basis $\phi_1, \dots, \phi_n$, i.e.

$$u_h = \sum_{j=1}^{n} \xi_j \phi_j \,. \tag{2.8}$$

In the one-dimensional problem (2.1), the Galerkin FEM solution u_h is known to coincide with the solution u at the nodes of the mesh; see, e.g., [3, Corollary 4.1.1]. Therefore the coefficients ξ_j are equal to the values of u in the nodes,

$$\xi_j = u(jh) \,, \quad j = 1, \dots, n \,. \tag{2.9}$$

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

$$\mathbf{A} = h^{-1} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} . \tag{2.10}$$

The eigenvalues λ_i and eigenvectors $\mathbf{y}_i = [\eta_{1i}, \dots, \eta_{ni}]^T$ of $\mathbf{A}$, $i = 1, \dots, n$, are known analytically (for details and their relationship to the eigenvalues and eigenfunctions of the continuous Laplace operator see, e.g., [4]),

$$\lambda_i = 4\, h^{-1} \sin^2\left(\frac{i\,\pi}{2(n+1)}\right) , \tag{2.11}$$

$$\eta_{ji} = \sqrt{\frac{2}{n+1}}\, \sin\left(\frac{j\, i\, \pi}{n+1}\right) , \quad j = 1, \dots, n \,. \tag{2.12}$$

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

$$w_i = \sum_{j=1}^{n} \eta_{ji}\, \phi_j \,, \qquad w_i(\ell h) = \eta_{\ell i} \,. \tag{2.13}$$

Remark: Please note, that (unlike in 2D) the stiffness matrix $\mathbf{A}$ (2.10) corresponding to the one-dimensional discretized Laplace operator (and therefore also its eigenvalues) depends on the size h of the mesh through the multiplicative factor h^{-1}. This is often avoided by multiplying the system $\mathbf{Ax} = \mathbf{b}$ by h, which does not affect the conditioning of the matrix. Since the algebraic energy norms $\|z\|_{\mathbf{A}}$ and $\|z\|_{(h\mathbf{A})}$ are different, such scaling would later be inconvenient. We will therefore stay with the algebraic problem $\mathbf{Ax} = \mathbf{b}$ as above with $\mathbf{A}$ and $\mathbf{b}$ given by (2.6) and (2.7) respectively.

Let the system $\mathbf{Ax} = \mathbf{b}$ be solved, for the purpose of numerical experiment, via the CG method[1]. We certainly do not advocate using CG for *practical* solving of similar model problems. We only wish to demonstrate on the simplest model problem the possible irregular distribution of the algebraic error. Let

$$u_h^{(k)} = \sum_{j=1}^{n} \xi_j^{(k)} \phi_j \tag{2.14}$$

be the approximation to the Galerkin FEM solution u_h (see (2.8)) given by the coordinate vector $\mathbf{x}_k = [\xi_1^{(k)}, \dots, \xi_n^{(k)}]^T$ computed at the kth step of the CG method. Then the squared energy norm of the error $\|(u - u_h^{(k)})'\|^2$ satisfies as a simple consequence of the Galerkin orthogonality the Pythagorean equality

$$\begin{aligned} \|(u - u_h^{(k)})'\|^2 &= \|(u - u_h)'\|^2 + \|(u_h - u_h^{(k)})'\|^2 \\ &= \|(u - u_h)'\|^2 + \|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{A}}^2 \, ; \end{aligned} \tag{2.15}$$

see, e.g., [5, Theorem 1.3, p. 38]. Given an initial approximation $\mathbf{x}_0$ and the corresponding initial residual $\mathbf{r}_0 \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_0$, the CG method minimizes the $\mathbf{A}$-norm of the algebraic error over the manifold $\mathbf{x}_0 \; + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$, where

$$\mathcal{K}_k(\mathbf{A}, \mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \dots, \mathbf{A}^{k-1}\mathbf{r}_0\}$$

is called the kth Krylov subspace generated by $\mathbf{A}$ and $\mathbf{r}_0$; see, e.g., [12, Theorem 4.3]. Consequently, the error of the approximation $u_h^{(k)}$ determined by the CG approximation $\mathbf{x}_k$ computed using exact arithmetic has the minimal energy norm $\|(u - u_h^{(k)})'\|^2$ over all approximations determined by the coefficient vectors from $\mathbf{x}_0 + \mathcal{K}_k(\mathbf{A}, \mathbf{r}_0)$. The energy norm is relevant in many applications; see, e.g., [9, Section 2.2.1].

Remark: The equality (2.15) holds for any vector $\mathbf{x}_k \in \mathbb{R}^n$ and the corresponding approximation $u_h^{(k)}$. In particular, it holds also for the results of the finite precision CG computations.

Following [6, p. 120], we consider, as an example, the exact solution

$$u = \exp(-5\,(x - 0.5)^2) - \exp(-5/4)\,. \tag{2.16}$$

We consider the FEM discretization using 19 inner nodes[2], i.e. we set $n = 19$. The solution u and the discretization error $u - u_h$ are given in Figure 2.1, with the squared

[1] We will use a general notation considering an initial approximation $\mathbf{x}_0$. All computations below are performed, however, with the zero initial approximation $\mathbf{x}_0 = \mathbf{0}$.

[2] Such small number of nodes allows us to plot illustrative figures. However, similar results can be obtained for any choice of n.

energy and L_2 norms of the discretization error equal (up to the negligible rounding errors in evaluation of the norms) to

$$\|(u-u_h)'\|^2 = 6.8078\text{e-}3 \quad \text{respectively} \quad \|u-u_h\|^2 = 1.7006\text{e-}6. \tag{2.17}$$

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

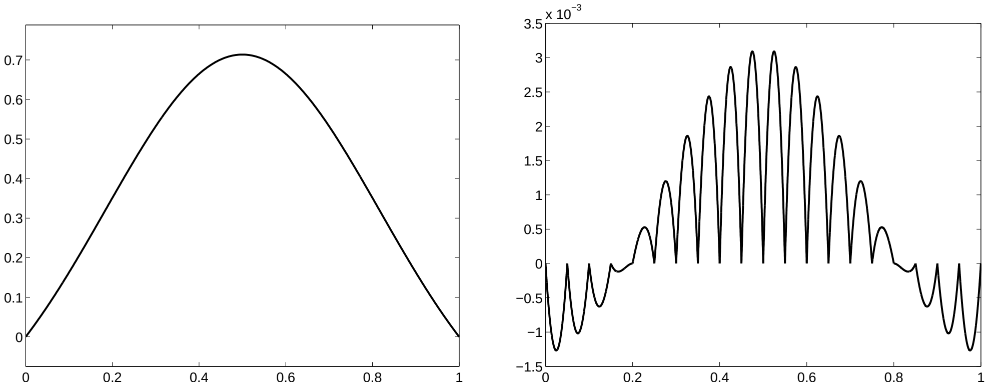


FIG. 2.1. *Left: the exact solution u (see (2.16)). Right: the discretization error $u - u_h$; the vertical axis is scaled by 10^{-3}.*

The squared $\mathbf{A}$-norm of the algebraic error $\|\mathbf{x}-\mathbf{x}_k\|^2_{\mathbf{A}}$ at the iteration steps $k = 7, 8, 9, 10$ of CG is given in the first column of Table 2.1. The second column contains, for comparison, the squared Euclidean norm $\|\mathbf{x}-\mathbf{x}_k\|^2$. For the energy and the L_2 norm of the total error $u - u_h^{(k)}$ see the third and the fourth column, respectively (please recall the corresponding norms of the discretization error $u - u_h$ given by (2.17)).

TABLE 2.1

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

Figure 2.2 shows the relative $\mathbf{A}$-norm of the algebraic error $\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}/\|\mathbf{x}-\mathbf{x}_0\|_{\mathbf{A}}$ together with the loss of orthogonality among the normalized residual vectors (measured in the Frobenius norm) for the standard CG implementation (see [12]) and for the CG implementation with double reorthogonalized residuals (see, e.g., [11]). Since for the given data the loss of orthogonality remains close to the machine precision level, the effect of rounding errors in the standard CG implementation is here negligible. Consequently, the standard finite precision CG behaves very similarly to the double-reorthogonalized CG that simulates the computation in exact arithmetic; see [11]. Taking into account the distribution of the eigenvalues of $\mathbf{A}$ and the choice $\mathbf{x}_0 = \mathbf{0}$, this is to be expected; see [16].

The algebraic and total errors are visualized for $k = 8, 9$ in Figure 2.3. At the 9th step, the energy norm of the total error is dominated by the discretization error,

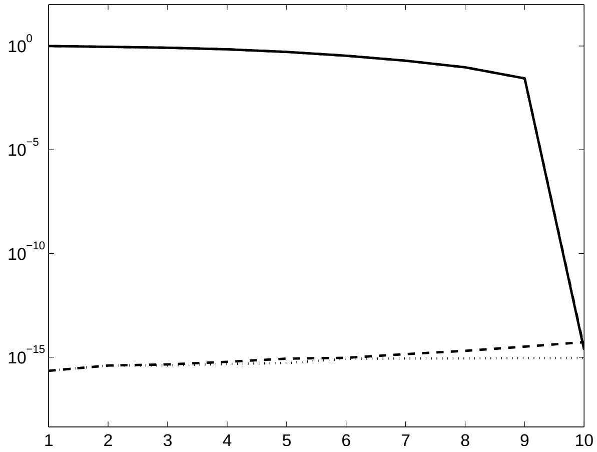


FIG. 2.2. *The relative* $\mathbf{A}$*-norm of the error* $\|\mathbf{x}-\mathbf{x}_k\|_{\mathbf{A}}/\|\mathbf{x}-\mathbf{x}_0\|_{\mathbf{A}}$ *(solid line), the loss of orthogonality in the standard CG implementation (dashed line) and the loss of orthogonality in the CG implementation with double reorthogonalized residuals (dotted line). In our computations, rounding errors do not play a significant role.*

see (2.17) and Table 2.1. Providing that the spatial distributions of the discretization and the algebraic error are similar, the contribution of the algebraic error to the total error would be at any part of the domain Ω marginal. However, quite the opposite is true. As shown in the right part of Figure 2.3, the algebraic error is significantly localized at the 10th component $\xi_{10}^{(9)}$ of the vector $\mathbf{x}_9$ which is much less accurate, in comparison to the exact solution $\mathbf{x}$, than any of its other components. Despite the relatively small energy norm $\|\mathbf{x}-\mathbf{x}_9\|_{\mathbf{A}}$, the total error $u-u_h^{(9)}$ at the node 10 (which is in this 1D model problem nothing but the size of the 10th component of the corresponding algebraic error $\mathbf{x}-\mathbf{x}_9$) is much different than the total error throughout the whole interval. The algebraic error substantially affects the shape of the total error. The left part of Figure 2.3 shows for illustration the same quantities for $k=8$.

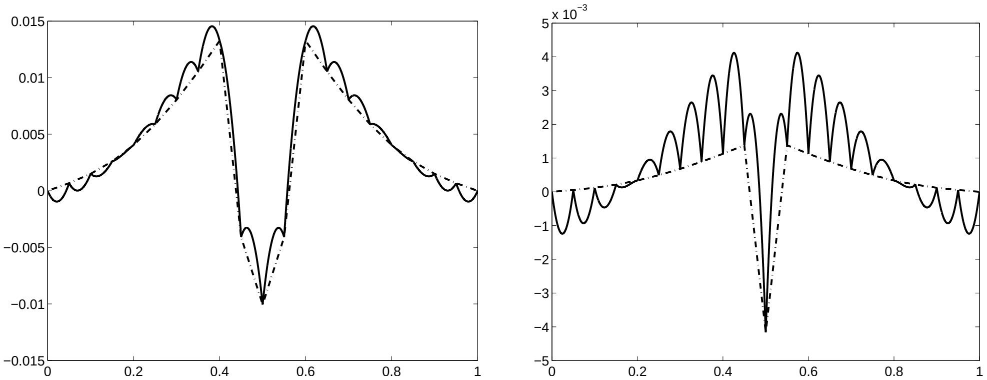


FIG. 2.3. *The algebraic error* $u_h-u_h^{(k)}$ *(dashed-dotted line) and the total error* $u-u_h^{(k)}$ *(solid line) at the 8th iteration (left) and at the 9th iteration (right). The vertical axis in the right part of the figure is scaled by* 10^{-3}.

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

One may object that if the error is measured in the L_2 norm instead of the energy norm, one does not see much discrepancy — both $\|\mathbf{x}-\mathbf{x}_9\|_\mathbf{A}$ and $\|\mathbf{x}-\mathbf{x}_9\|$ are still relatively large in comparison to $\|u-u_h\|$. This objection is, however, not to the points that the global *energy norm is not descriptive* and that the spatial distribution of the discretization and algebraic errors can be very different. Moreover, it can be easily verified that in the 2D Poisson problem an objection concerning the L_2 norm does not substantiate; see [15, Section 5.1].

When the polynomial exact solution

$$u = (x-2)\,(x-1)\,x\,(x+1) \tag{2.18}$$

is used instead of (2.16), we get with the same number of inner discretization nodes $n = 19$ the following results. The exact solution u and the discretization error $u-u_h$ are given in Figure 2.4; the discretization error $u-u_h$ is nonnegative. The squared energy and L_2 norms of the discretization error are equal to

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

Table 2.2 and Figures 2.5 and 2.6 give results analogous to those presented above in Table 2.1 and Figures 2.2 and 2.3 respectively.

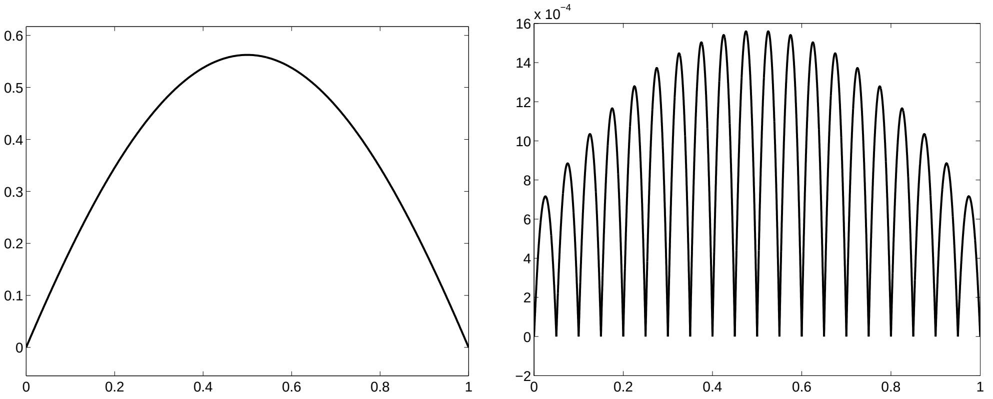


FIG. 2.4. *Left: the exact solution u (see (2.18)). Right: the discretization error $u-u_h$; the vertical axis is scaled by 10^{-4}.*

TABLE 2.2

k	$\|\mathbf{x}-\mathbf{x}_k\|_\mathbf{A}^2$	$\|\mathbf{x}-\mathbf{x}_k\|^2$	$\|(u-u_h^{(k)})'\|^2$	$\|u-u_h^{(k)}\|^2$
7	1.0112e-2	1.3654e-2	1.3612e-2	6.0367e-5
8	2.6905e-3	3.6997e-3	6.1905e-3	9.3021e-6
9	2.5563e-4	3.5534e-4	3.7556e-3	1.1605e-6
10	5.6776e-30	3.8081e-30	3.5000e-3	8.7495e-7

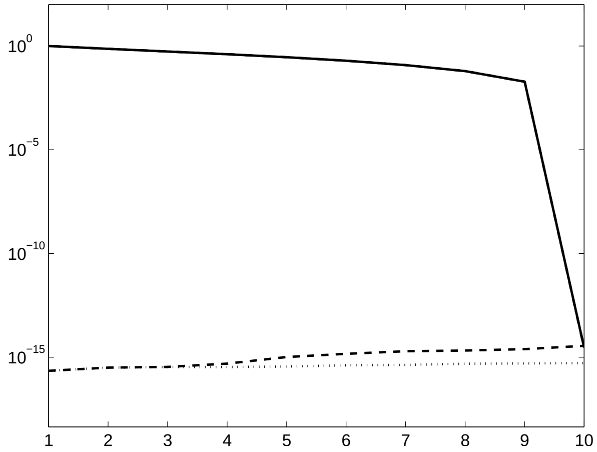


FIG. 2.5. *The relative* $\mathbf{A}$*-norm of the error* $\|\mathbf{x}-\mathbf{x}_k\|_\mathbf{A}/\|\mathbf{x}-\mathbf{x}_0\|_\mathbf{A}$ *(solid line), the loss of orthogonality in the standard CG implementation (dashed line) and the loss of orthogonality in the CG implementation with double reorthogonalized residuals (dotted line).*

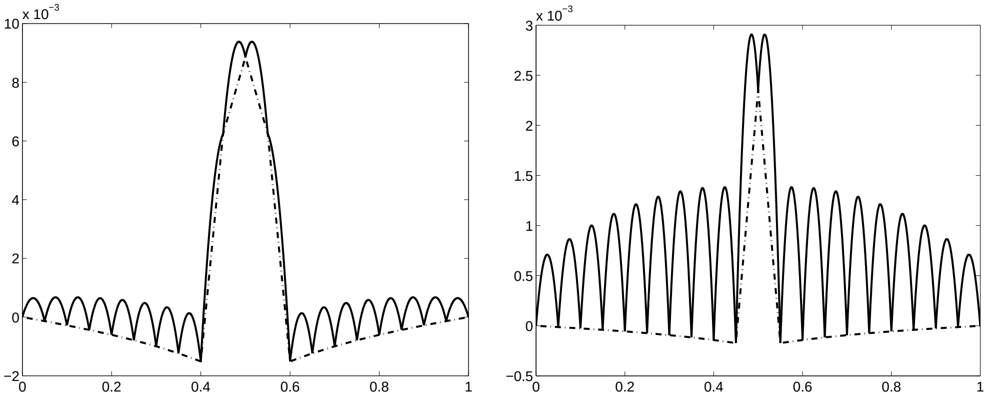


FIG. 2.6. *The algebraic error* $u_h - u_h^{(k)}$ *(dashed-dotted line) and the total error* $u - u_h^{(k)}$ *(solid line) at the 8th iteration (left) and at the 9th iteration (right); the vertical axes are scaled by* 10^{-3}.

3. Interpretation of the total error as a modification of the discretization mesh. As argued in [15, p. 9], it is desirable to interpret the inaccuracies in the solution process (including the algebraic errors) in terms of the meaningful modification of the mathematical model; see also [21, pp. 33–35]. This idea can be related to the so-called functional backward error by Arioli and others (see, e.g., [1]) where the errors are interpreted as (backward) perturbations of the weak formulation (2.2) of the problem. Related to this we observe, however, a serious difficulty. The perturbation of (2.2) should be meaningful in the sense that it preserves the original model. In our case we require that the problem after incorporating the functional backward error would again represent a Poisson problem. Clearly, an introduction of the functional backward error term counting for inaccurate solving of the discretized algebraic problem into the problem (2.2) would not satisfy this natural requirement. Therefore we consider the change of the discretization, i.e. the basis functions or the mesh, a more appealing alternative.

Interpreting the algebraic error as a transformation of the FEM basis has been considered in [10, Section 3]. We will use the idea from [10] but present the result in a slightly different way. Let the transformation of the basis $\Phi = [\phi_1, \dots, \phi_n]$ (in our problem the basis of continuous piecewise linear hat functions) to the basis $\widehat{\Phi} = [\widehat{\phi}_1, \dots, \widehat{\phi}_n]$ be represented by a square matrix $\mathbf{D} = [D_{\ell j}] \in \mathbb{R}^{n \times n}$,

$$\widehat{\phi}_j = \phi_j + \sum_{\ell=1}^{n} D_{\ell j}\, \phi_\ell\,, \quad j = 1, \dots, n\,. \tag{3.1}$$

Please note that unlike the original FEM basis functions ϕ_j, the transformed basis functions $\widehat{\phi}_j\,, j = 1, \dots, n$, need not be of a local support. The relation (3.1) can be written in the compact form as

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

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

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

$$(\mathbf{A} + \mathbf{E})\widehat{\mathbf{x}} = \mathbf{b}\,, \tag{3.2}$$

where

$$\mathbf{E} = \frac{(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}})\widehat{\mathbf{x}}^T}{\|\widehat{\mathbf{x}}\|^2}\,. \tag{3.3}$$

Let the Galerkin FEM solution u_h (see (2.4)–(2.8)) satisfy

$$u_h = \Phi\mathbf{x} = \sum_{j=1}^{n} \xi_j\, \phi_j = \sum_{j=1}^{n} \widehat{\xi}_j\, \widehat{\phi}_j = \widehat{\Phi}\widehat{\mathbf{x}} = \Phi\,(\mathbf{I} + \mathbf{D})\widehat{\mathbf{x}} \tag{3.4}$$

for some (unknown) matrix $\mathbf{D}$. Then, considering the Petrov-Galerkin discretization of (2.2) with $u_h = \widehat{\Phi}\widetilde{\mathbf{x}}$, i.e. the discretization basis $\widehat{\phi}_1, \dots, \widehat{\phi}_n$, and the test functions $\phi_1, \dots, \phi_n$ gives

$$a(u_h, \phi_i) = \ell(\phi_i)\,, \quad i = 1, \dots, n\,, \tag{3.5}$$

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

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

where

$$\begin{aligned} \widehat{A}_{ij} &= a(\widehat{\phi}_j, \phi_i) = a(\phi_j + \sum_{\ell=1}^{n} D_{\ell j}\, \phi_\ell\,, \phi_i) \\ &= A_{ij} + \sum_{\ell=1}^{n} A_{i\ell} D_{\ell j}\,, \end{aligned} \tag{3.6}$$

i.e.

$$\widehat{\mathbf{A}} = \mathbf{A} + \mathbf{A}\mathbf{D}. \tag{3.7}$$

Consequently, knowing the algebraic perturbation matrix $\mathbf{E}$ from (3.2), we can set

$$\mathbf{A}\mathbf{D} = \mathbf{E}\,, \quad \text{giving} \quad \mathbf{D} = \mathbf{A}^{-1}\mathbf{E}\,, \tag{3.8}$$

with $\widehat{\mathbf{x}} = \widetilde{\mathbf{x}}$ the exact algebraic solution of (3.2) representing the Petrov-Galerkin solution u_h of (2.2) in the sense of (3.5).

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

When we set $\widehat{\mathbf{x}} = \mathbf{x}_8$ for our experimental illustration with the exact solution (2.16), the norms of the perturbation and transformation matrices are

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

Figure 3.1 gives the matrices $\mathbf{E}$ (see (3.3)) and $\mathbf{D}$ (see (3.8)) visualized using the Matlab `surf` command. We can see the effect of the multiplication by $\mathbf{A}^{-1}$: the transformation matrix $\mathbf{D}$ has significantly more entries with the size far from zero than the perturbation matrix $\mathbf{E}$. It should be pointed out that our example is on purpose very simple and the mapping from $\mathbf{E}$ to $\mathbf{D} = \mathbf{A}^{-1}\mathbf{E}$ is for the given $\mathbf{A}$ rather benign (the norm $\|\mathbf{D}\|$ is even smaller than $\|\mathbf{E}\|$). In more practical problems this may not be the case and $\mathbf{D}$ can have large nonzero elements. The left part of Figure 3.2 shows (for the same approximation $\widehat{\mathbf{x}} = \mathbf{x}_8$) the example of the transformed basis function $\widehat{\phi}_j$ (here $\widehat{\phi}_5$; see (3.1)). Since the entries of the matrix $\mathbf{D}$ are of the order 10^{-3}, $\widehat{\phi}_5$ looks visually the same as ϕ_5. The difference $\widehat{\phi}_5 - \phi_5$ is plotted in the right part of Figure 3.2. For other basis functions the situation is analogous. The size of the differences $\widehat{\phi}_j - \phi_j$, $j = 1, \ldots, n$, corresponds to the size of the algebraic error (as well as the discretization error when the algebraic and discretization errors are in balance).

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

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

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

For the second example with the exact solution (2.18) and the approximation $\widehat{\mathbf{x}} = \mathbf{x}_9$ given at the 9th CG iteration step, the norms of the perturbation and transformation matrices are

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

Figure 3.3 gives the matrix $\mathbf{E}$ and the matrix $\mathbf{D}$. For the transformed basis function $\widehat{\phi}_{11}$ and the difference $\widehat{\phi}_{11} - \phi_{11}$ see Figure 3.4.

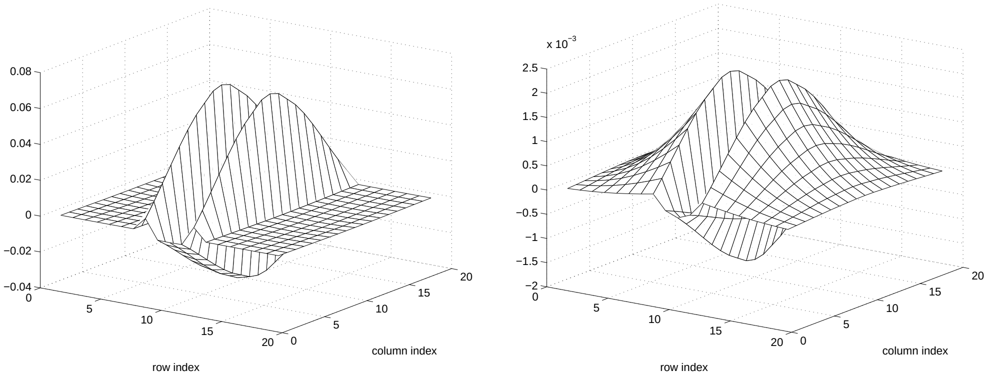


FIG. 3.1. *The perturbation matrix* $\mathbf{E}$ *(left) and the transformation matrix* $\mathbf{D}$ *(right) (with the entries visualized using the Matlab* ***surf*** *command) for the approximation* $\widehat{\mathbf{x}} = \mathbf{x}_8$ *in the example with the exact solution (2.16). The right vertical axis is scaled by* 10^{-3}.

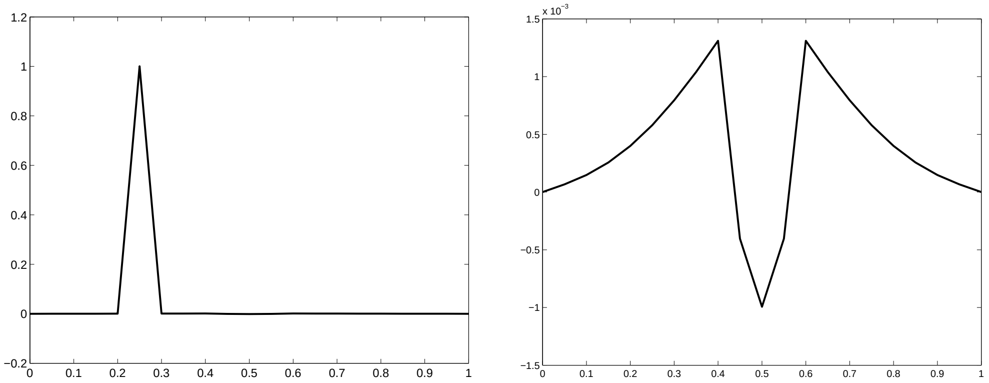


FIG. 3.2. *The transformed basis function* $\widehat{\phi}_5$ *(left) and the difference* $\widehat{\phi}_5 - \phi_5$ *(right) for the approximation* $\widehat{\mathbf{x}} = \mathbf{x}_8$ *in the example with the exact solution (2.16). For the other basis functions the situation is analogous. The right vertical axis is scaled by* 10^{-3}*; see the scale in the right part of Figure 2.1.*

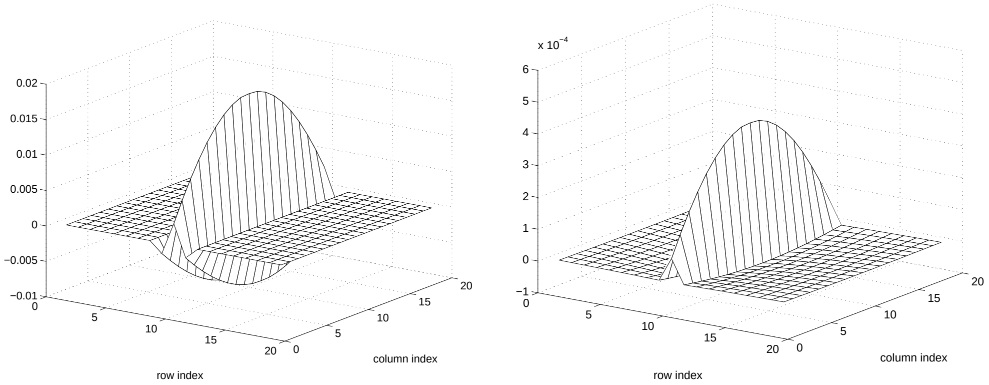


FIG. 3.3. *The perturbation matrix* $\mathbf{E}$ *(left) and the transformation matrix* $\mathbf{D}$ *(right) (with the entries visualized using the Matlab* ***surf*** *command) for the approximation* $\widehat{\mathbf{x}} = \mathbf{x}_9$ *in the example with the exact solution (2.18). The right vertical axis is scaled by* 10^{-4}.

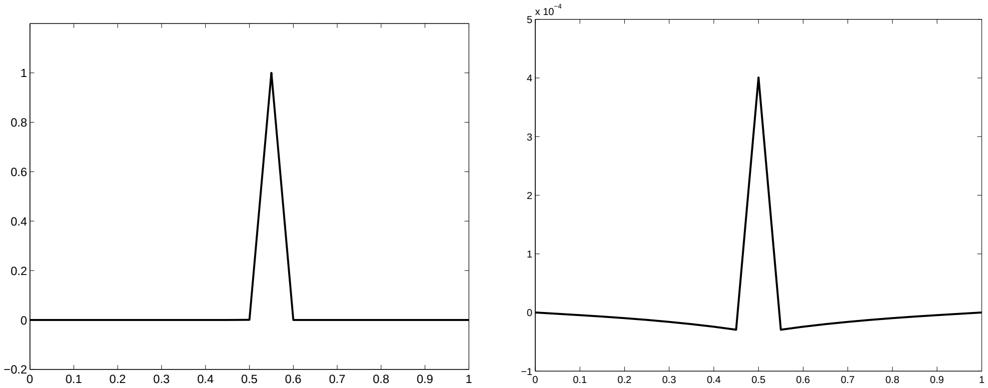


FIG. 3.4. *The transformed basis function $\widehat{\phi}_{11}$ (left) and the difference $\widehat{\phi}_{11} - \phi_{11}$ (right) for the approximation $\widehat{\mathbf{x}} = \mathbf{x}_9$ in the example with the exact solution (2.18). For the other basis functions the situation is analogous. The right vertical axis is scaled by 10^{-4}; see the scale in the right part of Figure 2.4.*

In the rest of this section we interpret (with some unimportant inaccuracy) the total error $u - u_h^{(9)}$ for the last example (the exact solution u is given by (2.18) and $u_h^{(k)}$ is determined using the approximation $\mathbf{x}_9$ computed at the 9th CG step) as the discretization error $u - u_H$, where the Galerkin FEM solution u_H corresponds to a *new mesh* and new basis functions which *preserve the locality of their support*. The Galerkin FEM solution u_H coincides with the solution u at the nodes of the mesh; see [3, Corollary 4.1.1]. Therefore we construct the new mesh in such way that the new nodes τ_i are given as the roots of the total error $u - u_h^{(9)}$ (i.e. the discretization error $u - u_H$) and therefore

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

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

$$\begin{aligned} \tau_i\,,\ i = 1,\dots,18 \quad &= \quad \text{roots of } u - u_h^{(9)} \text{ for } 0 < x < 0.5\,,\\ \tau_{19} \quad &= \quad 0.5 - 0.7h\,,\\ \tau_{20} \quad &= \quad 0.5 + 0.7h\,,\\ \tau_i\,,\ i = 21,\dots,38 \quad &= \quad \text{roots of } u - u_h^{(9)} \text{ for } 0.5 < x < 1\,. \end{aligned} \tag{3.9}$$

The new mesh now consists of $n = 38$ inner nodes, with 36 of them forming 18 close pairs. Please note that the new central element is 1.4 times longer than the elements in the original (uniform) mesh[3] , i.e. $\tau_{20} - \tau_{19} = 1.4\,h$.

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

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

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

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

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

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

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

while

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

The comparison of the discretization error $u - u_H$ with the total error $u - u_h^{(9)}$ is given in the left part of Figure 3.5. With our choice of the nodes (3.9), the positive values of $u - u_h^{(9)}$ coincide, except for $\tau_{18} < x < \tau_{21}$, with the error $u - u_H$; see the detail of the comparison in the right part of Figure 3.5. There is a slight discrepancy between $u - u_H$ and $u - u_h^{(9)}$ for $\tau_{18} < x < \tau_{21}$.

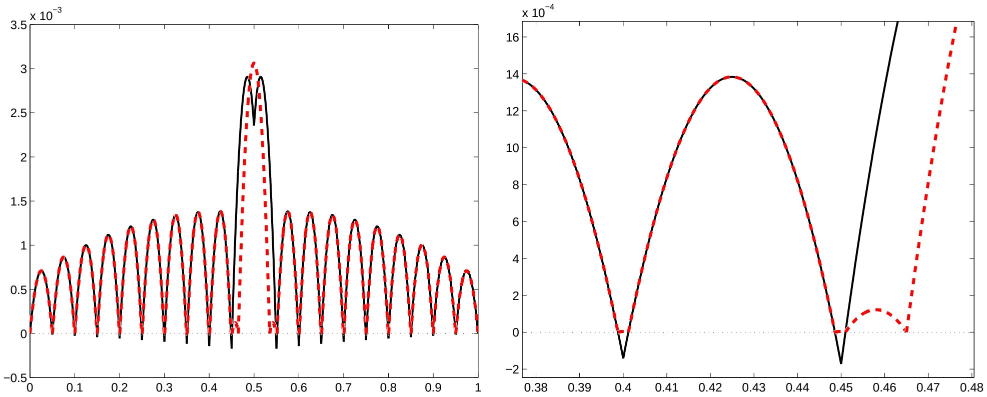


FIG. 3.5. *Left: the total error $u - u_h^{(9)}$ for the original mesh (solid line) and the discretization error $u - u_H$ on the modified mesh (dashed line); the vertical axis is scaled by 10^{-3}. Right: the detail showing the coincidence of the positive values of $u - u_h^{(9)}$ with $u - u_H$ for most of the interval and their slight discrepancy in the middle; the vertical axis is scaled by 10^{-4}.*

Interpretation of the total error as the error of the *exact* discretized solution using a modified discretization mesh can rise, as illustrated above, interesting points. First, the algebraic error can be interpreted, in the sense described above, as the loss of locality of the support of the modified (Petrov-) Galerkin basis functions. Second, the computed approximate solution $u_h^{(k)}$ which includes the error in the solution of the algebraic system can be interpreted (here with a small inaccuracy) as the discrete solution (with the vanishing algebraic error) for a mesh which can possibly have "holes" in the areas where the algebraic error is large (in our example the mesh has a "hole" in the center of the interval).

4. Spatial distribution of the error in CG computations. In this section we explain the behavior of the algebraic error observed above. In the following we present the experimental illustration with the exact solution (2.16); see also Figures 3.1 and 3.2. The exposition uses the close relationship between CG and the Lanczos algorithm; for details see the original papers [12, 14] and also the survey [16].

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

$$\mathbf{x} - \mathbf{x}_k = \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_k, \mathbf{y}_i)\, \mathbf{y}_i \,, \tag{4.1}$$

where, as above, $\mathbf{y}_i$ denotes the ith normalized eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ_i ; see (2.11)-(2.12). We denote by $\theta_j^{(k)}$, $j = 1, \ldots, k$, the approximations of the eigenvalues of the matrix $\mathbf{A}$ (*Ritz values*) given at the kth iteration of the Lanczos algorithm applied to the matrix $\mathbf{A}$ and the starting vector $\mathbf{r}_0/\|\mathbf{r}_0\|$. Assuming exact arithmetic, a close approximation of the eigenvalue λ_i by a Ritz value $\theta_j^{(k)}$ means that the size of the ith component $|(\mathbf{x} - \mathbf{x}_k, \mathbf{y}_i)|$ of the error $\mathbf{x} - \mathbf{x}_k$ of the kth CG approximation in the direction $\mathbf{y}_i$ becomes small; see, e.g., [16, Theorem 3.3]. As mentioned above, the effect of rounding errors is in our example negligible. Consequently, the previous statement holds also for the presented results of finite precision computations.

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

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

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

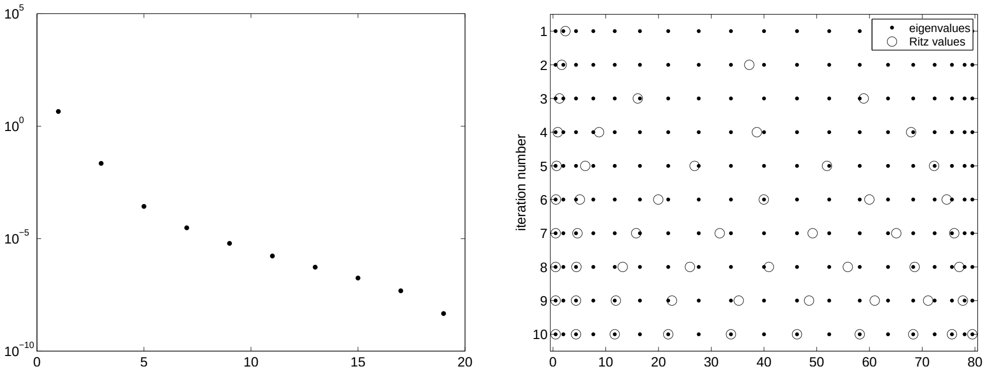


FIG. 4.1. *Left: the squared size of the spectral components* $|(\mathbf{x} - \mathbf{x}_0, \mathbf{y}_i)|^2$*,* $i = 1, \ldots, n$*, of the initial error* $\mathbf{x} - \mathbf{x}_0$*. Right: convergence of the Ritz values (circles) to the eigenvalues of* $\mathbf{A}$ *(dots) in iterations 1 through 10.*

The squared size of the spectral components $|(\mathbf{x} - \mathbf{x}_0, \mathbf{y}_i)|^2$, $i = 1, \ldots, n$, of the initial error $\mathbf{x} - \mathbf{x}_0$ are given in the left part of Figure 4.1. Recall that $\mathbf{x}_0 = \mathbf{0}$ and therefore the initial error is equal to the solution $\mathbf{x}$. Since the solution is symmetric with respect to the center 0.5 of the given interval, the spectral components with even indices vanish (the corresponding projections computed in finite precision arithmetic are on the machine precision level). Since the initial error $\mathbf{x}-\mathbf{x}_0$ is smooth (i.e. nonoscillating), the components of the error with higher indices, which correspond to more oscillating eigenvectors (see (2.12)), significantly decrease with increasing index i. The Ritz values $\theta_j^{(k)}$, $j = 1, \ldots, k$, are for $k = 1, \ldots, 10$ given in the right part of Figure 4.1. The dots represent the eigenvalues of matrix $\mathbf{A}$. As expected, the Ritz values approximate the eigenvalues with odd indices. At the 10th iteration, all such eigenvalues are approximated, all components of the error $\mathbf{x}-\mathbf{x}_{10}$ become very small and the norm of the algebraic error drops to the machine precision level; see Figure 2.2 and Table 2.1. We can observe that the eigenvalues λ_1, λ_2 and partially also λ_3 are approximated much faster (for smaller iteration number) than the others.

In Figure 4.2 the development of the squared size of the spectral components of the algebraic error $\mathbf{x}-\mathbf{x}_k$ is shown for $k = 0, 7, 8, 9$ (only the values with odd indices are plotted; the rest remain at the level 10^{-30}). We can see that the CG method reduces quickly the dominating spectral components of the error which corresponds to the fast approximation of the eigenvalues λ_1 and λ_2 by the Ritz values illustrated above. With increasing k the spectral components of $\mathbf{x}-\mathbf{x}_k$ almost equilibrate. As a consequence, the spatial distribution of the error $\mathbf{x}-\mathbf{x}_k$ changes as k increases and it eventually becomes highly nonuniform in the way substantially different than the spatial distribution of the initial error $\mathbf{x}-\mathbf{x}_0$.

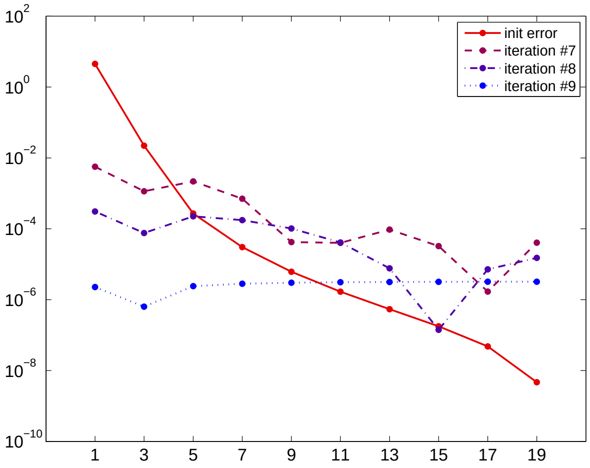


FIG. 4.2. *The development of the squared size of the spectral components of the algebraic error* $|(\mathbf{x}-\mathbf{x}_k, \mathbf{y}_i)|^2$*,* $i = 1, 3, \ldots, 19$*, for the iteration steps* $k = 0, 7, 8, 9$ *(solid, dashed, dashed-dotted and dotted lines respectively). We can observe equilibrating of the size of the spectral components as* k *increases.*

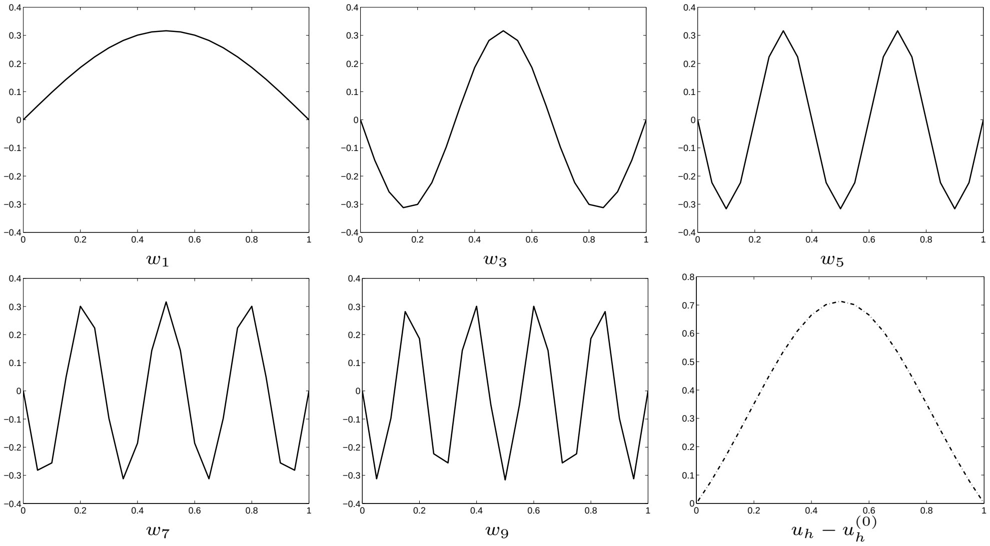
This situation is illustrated in Figures 4.3 and 4.4, where we plot the most dominating approximations w_i to the eigenfunctions of the continuous operator (see (2.13) and (4.1)), corresponding to the initial error $\mathbf{x}-\mathbf{x}_0$ and to the error $\mathbf{x}-\mathbf{x}_9$ respectively. The right bottom part of Figure 4.3 shows the algebraic part of the initial error in the function space, which is given as the linear combination of the eigenfunction approximations with odd indices

$$u_h - u_h^{(0)} = \Phi\left(\mathbf{x}-\mathbf{x}_0\right) = \sum_{i=1}^{10} \left(\mathbf{x}-\mathbf{x}_0, \mathbf{y}_{2i-1}\right) w_{2i-1}\,. \tag{4.2}$$

(As mentioned above, we use $\mathbf{x}_0 = \mathbf{0}$ and therefore $u_h - u_h^{(0)} = u_h$.) The right bottom part of Figure 4.4 shows the algebraic part of the error

$$u_h - u_h^{(9)} = \Phi\,(\mathbf{x} - \mathbf{x}_9) \approx \sum_{i=1}^{10} (\mathbf{x} - \mathbf{x}_9, \mathbf{y}_{2i-1})\, w_{2i-1}\,; \tag{4.3}$$

please compare with the algebraic error given in the right part of Figure 2.3. Here we neglect the spectral components of $\mathbf{x} - \mathbf{x}_9$ in the direction of even eigenvectors of $\mathbf{A}$ which remain at the machine precision level (and therefore we use the approximation instead of the equality).



w_1 w_3 w_5

w_7 w_9 $u_h - u_h^{(0)}$

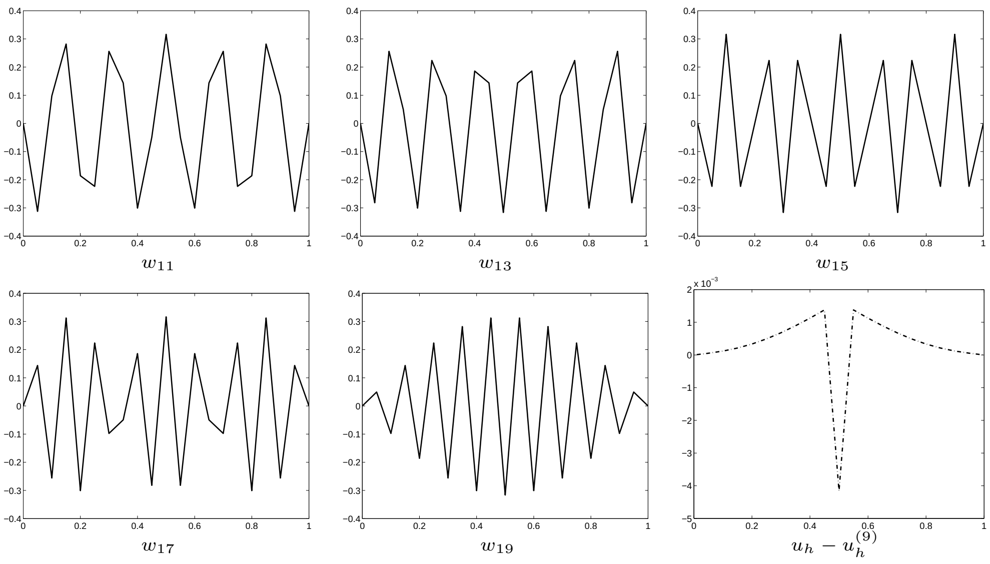
FIG. 4.3. *The approximate eigenfunctions w_i corresponding to the largest components of the initial algebraic error $\mathbf{x}-\mathbf{x}_0$ in the eigenvector basis of the matrix $\mathbf{A}$ and the algebraic part $u_h - u_h^{(0)}$ of the initial error $u - u_h^{(0)}$ (see (4.2)) (the dashed-dotted line in the right bottom part).*

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

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

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

5. Conclusions. Using a simple 1D model problem, it is illustrated that the spatial distribution of the algebraic error can significantly differ from the spatial distribution of the discretization error. Because of its possibly large local components



w_{11} w_{13} w_{15}

w_{17} w_{19} $u_h - u_h^{(9)}$

Fig. 4.4. *The approximate eigenfunctions w_i corresponding to the largest components of the algebraic error $\mathbf{x} - \mathbf{x}_9$ in the eigenvector basis of the matrix $\mathbf{A}$ and the algebraic part $u_h - u_h^{(9)}$ of the error $u - u_h^{(9)}$ (see (4.3)) (the dashed-dotted line in the right bottom part). The vertical axis in the right bottom part of the figure is scaled by 10^{-3}.*

in some parts of the domain, the algebraic error can determine the spatial distribution of the total error $u - u_h^{(k)}$ even when its globally measured size (e.g. the energy norm $\|u - u_h^{(k)}\|$ is small). It can be expected that an analogous phenomenon can be observed for practical problems.

The demonstrated difference between the spatial distributions of the algebraic and the discretization error across the domain (here obtained for the CG method) underlines importance of constructing reliable stopping criteria in iterative algebraic solvers. In particular, such criteria should be related to the spatial distribution of the total error in the function space. A work in this direction has been recently done, e.g., in [13, Section 6] and in a more general nonlinear setting in [7]. One should also recall the goal oriented adaptivity approach of Rannacher, Becker and their collaborators in the context of duality-based error control, which allows balancing discretization and iteration error in the problem-related areas of interest; see, e.g., the survey paper [19] and the references given there. We believe that further developments focusing on the spatial distribution of the algebraic and total errors will be reported in a near future.

REFERENCES

[1] M. Arioli, E. Noulard, and A. Russo, *Stopping criteria for iterative methods: applications to PDE's*, Calcolo, 38 (2001), pp. 97–112.

[2] I. Babuška, *Numerical stability in problems of linear algebra*, SIAM J. Numer. Anal., 9 (1972), pp. 53–77.

[3] I. Babuška and T. Strouboulis, *The finite element method and its reliability*, Numerical Mathematics and Scientific Computation, The Clarendon Press Oxford University Press, New York, 2001.

[4] D. Boffi, *Finite element approximation of eigenvalue problems*, Acta Numer., 19 (2010), pp. 1–120.

[5] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

[6] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational differential equations*, Cambridge University Press, Cambridge, 1996.

[7] A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs.* HAL Preprint 00681422 v2, submitted for publication, 2012.

[8] M. S. Gockenbach, *Partial differential equations: analytical and numerical methods*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[9] ———, *Understanding and implementing the finite element method*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[10] S. Gratton, P. Jiránek, and X. Vasseur, *Energy backward error: interpretation in numerical solution of elliptic partial differential equations and convergence of the conjugate gradient method*, Tech. Report TR/PA/10/95, CERFACS, Toulouse, 2010.

[11] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.

[12] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436 (1953).

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

[14] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Research Nat. Bur. Standards, 49 (1952), pp. 33–53.

[15] J. Liesen and Z. Strakoš, *Krylov subspace methods: principles and analysis*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, 2012.

[16] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542.

[17] A. E. Naiman, I. Babuška, and H. C. Elman, *A note on conjugate gradient convergence*, Numer. Math., 76 (1997), pp. 209–230.

[18] A. Quarteroni, *Numerical models for differential problems*, vol. 2 of MS&A. Modeling, Simulation and Applications, Springer-Verlag Italia, Milan, 2009. Translated from the 4th (2008) Italian edition by Silvia Quarteroni.

[19] R. Rannacher, *A short course on numerical simulation of viscous flow: discretization, optimization and stability analysis*, Discrete Contin. Dyn. Syst. Ser. S, 5 (2012), pp. 1147–1194.

[20] P. J. Roache, *Verification and validation in computational science and engineering*, Hermosa Publishers, Albuquerque, New Mexico, 1998.

[21] ———, *Building PDE codes to be verifiable and validatable*, Comput. Sci. Eng., 6 (2004), pp. 30–38.

[22] Z. Strakoš and J. Liesen, *On numerical stability in large scale linear algebraic computations*, ZAMM Z. Angew. Math. Mech., 85 (2005), pp. 307–325.