Electronic Transactions on Numerical Analysis.
Volume 40, pp. 338-355, 2013.

ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

ENERGY BACKWARD ERROR: INTERPRETATION IN NUMERICAL SOLUTION OF ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS AND BEHAVIOUR IN THE CONJUGATE GRADIENT METHOD*

SERGE GRATTON†, PAVEL JIRÁNEK‡, AND XAVIER VASSEUR‡

Abstract. Backward error analysis is of great importance in the analysis of the numerical stability of algorithms in finite precision arithmetic, and backward errors are also often employed in stopping criteria of iterative methods for solving systems of linear algebraic equations. The backward error measures how far we must perturb the data of the linear system so that the computed approximation solves it exactly. We assume that the linear systems are algebraic representations of partial differential equations discretised using the Galerkin finite element method. In this context, we try to find reasonable interpretations of the perturbations of the linear systems which are consistent with the problem they represent and consider the optimal backward perturbations with respect to the energy norm, which is naturally present in the underlying variational formulation. We also investigate its behaviour in the conjugate gradient method by constructing approximations in the underlying Krylov subspaces which actually minimise such a backward error.

Key words. symmetric positive definite systems, elliptic problems, finite element method, conjugate gradient method, backward error

AMS subject classifications. 65F10, 65F50

1. Introduction.

Backward error analysis in numerical linear algebra, pioneered by von Neumann and Goldstein [28], Turing [26], Givens [10] and further developed and popularised by Wilkinson (see, e.g., [30, 31]), is a widely used technique employed in the study of effects of rounding errors in numerical algorithms. When solving a given algebraic problem for some data by means of a certain numerical algorithm, we would normally be satisfied with an approximate solution with a small relative error (the forward error) close to the precision of our arithmetic. This is, however, not always possible, so we may ask instead for what data we actually solved our problem. Thus we interpret the computed solution as a solution of the perturbed problem and identify the norm of the data perturbation with the backward error associated with the computed approximate solution. (There might be many such perturbations, so we are interested in the smallest one).

In practical problems, the data are often affected by errors due to, e.g., measurements, truncation, and round-off. We could hence be satisfied with a solution which solves the given problem for some data lying within a certain neighbourhood of the provided data. The backward error provides natural means for quantifying the accuracy of computed solutions with respect to the accuracy of the problem data. In addition, the bounds on forward errors can often be obtained from backward errors using the perturbation theory associated with the problem to be solved, which is independent of the algorithm used to obtain the solution. For more details, see [12, Chapter 1]. See also [17, Section 5.8] for a recent overview of the relations between the concepts of numerical stability and backward error.

Backward error analysis provides an elegant way how to study numerical stability of algorithms, that is, their sensitivity with respect to rounding errors. If an algorithm is guaranteed to provide a solution with a backward error close to the machine precision of the given

*Received December 13, 2011. Accepted June 4, 2013. Published online on September 4, 2013. Recommended by Z. Strakos. The work of the second author was supported by the ADTAO project funded by the foundation STAE, Toulouse, France, within RTRA.

†INPT-IRIT, University of Toulouse and ENSEEIHT, 2 Rue Camichel, BP 7122, 31071 Toulouse Cedex 7, France (serge.gratton@enseeiht.fr).

‡CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France ({jiranek, vasseur}@cerfacs.fr).

ETNA
Kent State University
http://etna.math.kent.edu

finite precision arithmetic for any data (the backward stable algorithm), one could be satisfied with such an algorithm and solution it provides. Indeed the problem data cannot be stored exactly in finite precision arithmetic anyway independently of the means how they were obtained. It is therefore perfectly reasonable to consider the backward error as a meaningful accuracy measure for quantities obtained from algorithms which would (in the absence of the rounding errors) deliver the exact solution of the given problem.

The backward error concept is sometimes used to construct accuracy criteria for computations which are inherently inexact even in exact arithmetic. In particular, we are interested in its use in stopping criteria for iterative solvers for linear algebraic systems

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \qquad \mathbf{A} \in \mathbb{R}^{N\times N}, \tag{1.1}$$

where $\mathbf{A}$ is assumed to be nonsingular. For a given approximation $\hat{\mathbf{u}}$ of the solution of (1.1), the backward error represents a measure by which $\mathbf{A}$ and $\mathbf{f}$ have to be perturbed so that $\hat{\mathbf{u}}$ solves the problem $(\mathbf{A}+\hat{\mathbf{E}})\hat{\mathbf{u}} = \mathbf{f}+\hat{\mathbf{g}}$. The norm-wise relative backward error

$$\min\{\varepsilon : (\mathbf{A}+\hat{\mathbf{E}})\hat{\mathbf{u}} = \mathbf{f}+\hat{\mathbf{g}},\ \|\hat{\mathbf{E}}\| \le \varepsilon\|\mathbf{A}\|,\ \|\hat{\mathbf{g}}\| \le \varepsilon\|\mathbf{f}\|\}$$

was shown by Rigal and Gaches [21] to be given by

$$\frac{\|\mathbf{f}-\mathbf{A}\hat{\mathbf{u}}\|}{\|\mathbf{A}\|\|\hat{\mathbf{u}}\|+\|\mathbf{f}\|}, \tag{1.2}$$

where $\|\cdot\|$ is any vector norm and its associated matrix norm, although in practice one usually chooses the standard Euclidean one. There are reasons why the backward error (1.2) should be preferred over the standard relative residual norm as the guide for stopping the iterative solvers when more relevant and sophisticated measures are not available; see, e.g., [3, 12], and [17, Section 5.8.3]. This might be certainly supported by the fact that some iterative methods, e.g., the methods based on the generalised minimum residual method [23, 29], are backward stable [2, 8, 13, 19] and thus may deliver solutions with an accuracy in terms of the backward error close to the machine precision if required. We also point out the related discussion in [25], in particular in Sections 1 and 2 there.

Iterative methods are in practice chiefly applied for solving linear systems (1.1) arising from discretised partial differential equations (PDE), e.g., by the finite element method (FEM). Here the main source of errors is due to the truncation of the continuous differential operator, which, however, does not need to be reflected simply by the data errors in the coefficients of the resulting linear algebraic system. The basic FEM discretisation of the one-dimensional Poisson equation considered in Section 2 represents this fact; the coefficient matrix can be stored exactly even in finite precision arithmetic. The stopping criteria for iterative solvers based on the norm-wise backward error (in the Euclidean norm) might be at least questionable in this context. More sophisticated criteria balancing the inaccuracy of the solution obtained by the iterative solver and the inaccuracy due to truncation (the discretisation error) should be used; see, e.g., [4] and the references therein.

We believe that when a certain stopping criterion based on data perturbations such as the backward error is considered, the effects of these perturbations in the original problem to be solved should be clarified. Here the system (1.1) is the algebraic representation of a FEM discretisation of an elliptic PDE and solved inaccurately, e.g., by an iterative method. When a stopping criterion based on the backward error is used and hence the computed approximation is interpreted as the solution of a perturbed linear system, we may ask whether such perturbations have meaningful representations in the underlying discrete problem as well.

In Section 2 we consider a general weak formulation of a self-adjoint elliptic PDE which can be characterised by a variational equation involving a continuous, symmetric, and elliptic

ETNA
Kent State University
http://etna.math.kent.edu

bilinear form defined on a real Hilbert space and a general discretisation by the Galerkin finite element method. We also introduce a simple one-dimensional model problem, which we use throughout the paper to illustrate our results. In Section 3 we assume to have an approximate solution $\hat{\mathbf{u}}$ of the algebraic representation (1.1) of the discretised variational problem in a fixed basis of the discrete space, which we associate with perturbed problems

$$\mathbf{A}\hat{\mathbf{u}} = \mathbf{f} + \hat{\mathbf{g}} \quad \text{and} \quad (\mathbf{A} + \hat{\mathbf{E}})\hat{\mathbf{u}} = \mathbf{f}, \tag{1.3}$$

and look for possible interpretations of the data perturbations $\hat{\mathbf{g}}$ and $\hat{\mathbf{E}}$ in the discrete variational equation. Although the role of $\hat{\mathbf{g}}$ in (1.3) is well known (see, e.g., [1]), the interpretation of $\hat{\mathbf{E}}$ is in our opinion worth some clarification. A similar idea of perturbing the operator was considered before by Arioli et al. [5] as the so-called functional backward error. It is, however, not obvious whether such an operator perturbation still may be identified with a (discretised) PDE or how it "physically" affects the original PDE. In Section 3 we try to interpret $\hat{\mathbf{E}}$ as a certain perturbation of the FEM basis for which the second system in (1.3) can be associated with the algebraic form of the original discretised PDE. In addition, we look for the operator $\hat{\mathbf{E}}$ optimal with respect to the norm relevant in our setting, that is, the energy norm, and find a simple characterisation of such a definition of the backward error (called the energy backward error here) in the functional setting. Our approach is related to the work in [20]. There the authors interpret the total error (that is, the difference between the solution of the continuous problem and the approximate discrete solution) as the error of the exact discrete solution on a modified mesh. Here, on the other hand, we keep the discrete space fixed.

Throughout the paper we illustrate our observations at a simple one-dimensional model problem introduced in Section 2 and consider solving the resulting algebraic system by the conjugate gradient method (CG) [11]. It is known that CG minimises the $\mathbf{A}$-norm (the discrete representation of the energy norm) of the error over the Krylov subspace constructed using the initial residual vector and the matrix $\mathbf{A}$. It appears that the energy backward error introduced in Section 3 is closely related to the relative $\mathbf{A}$-norm of the error, that is, the forward error. According to this fact, we look in Section 4 for an approximation in the same Krylov subspace which actually minimises the energy backward error. We show that it is just a scalar multiple of the CG approximation. There is also an interesting "symmetry" with respect to the CG approximations showing that they are in a sense equivalent. We do not consider the effects of rounding errors throughout Section 4, although we are aware of the limits of the presented results in practice.

2. Galerkin FEM and model problem. In this section we recall the abstract weak formulation of a linear partial differential equation and its discretisation using the Galerkin finite element method. For more details, see, e.g., [6, 7]. Although we use a simple one-dimensional Poisson equation as an illustrative model problem, our ideas can be kept in this very general setting.

We consider an abstract variational problem on a real Hilbert space $\mathcal{V}$: find $u \in \mathcal{V}$ such that

$$a(u, v) = \langle f, v \rangle \qquad \forall v \in \mathcal{V}, \tag{2.1}$$

where we assume that a is a continuous, symmetric, and elliptic bilinear form on $\mathcal{V}$, $f \in \mathcal{V}'$, where $\mathcal{V}'$ denotes the space of continuous linear functionals on $\mathcal{V}$, and $\langle \cdot, \cdot \rangle$ is the duality pairing between $\mathcal{V}$ and $\mathcal{V}'$. The bilinear form $a(\cdot, \cdot)$ defines an inner product on $\mathcal{V}$ and its associated norm is $\|\cdot\|_a \equiv [a(\cdot, \cdot)]^{1/2}$ (usually called the energy norm). Due to the Lax-Milgram lemma [16] (see also, e.g., [7, Theorem 1.1.3]), the problem (2.1) is uniquely solvable.

ETNA
Kent State University
http://etna.math.kent.edu

Let $\mathcal{V}_h$ be a subspace of $\mathcal{V}$ of finite dimension N. The Galerkin method for approximating the solution u of (2.1) reads: find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v_h) = \langle f, v_h \rangle \qquad \forall v_h \in \mathcal{V}_h. \tag{2.2}$$

It is well known that the discrete problem (2.2) has a unique solution. The discretisation error $u - u_h$ is orthogonal to $\mathcal{V}_h$ with respect to the inner product $a(\cdot,\cdot)$ and, equivalently, the discrete solution u_h minimises the energy norm of $u - u_h$ over $\mathcal{V}_h$, that is,

$$\|u - u_h\|_a = \min_{v_h \in \mathcal{V}_h} \|u - v_h\|_a.$$

In order to transform the discrete problem (2.2) to a system of linear algebraic equations, we choose a basis of $\mathcal{V}_h$. For simplicity, we use the same notation for the basis and for the matrix representing it. In other words, we do not distinguish between $\Phi = \{\phi_1, \ldots, \phi_N\}$ and the matrix $\Phi = [\phi_1, \ldots, \phi_N]$. Thus we choose a basis $\mathbf{\Phi} \equiv [\phi_1, \ldots, \phi_N]$ of $\mathcal{V}_h$ so that we can express the solution u_h in terms of the basis $\mathbf{\Phi}$ as $u_h = \mathbf{\Phi}\mathbf{u}$ for some vector $\mathbf{u} \in \mathbb{R}^N$ representing the coordinates of u_h in the basis $\mathbf{\Phi}$. Then (2.2) holds if and only if $a(u_h, \phi_i) = \langle f, \phi_i \rangle$ for $i = 1, \ldots, N$, which leads to a system of algebraic equations (1.1) with

$$\mathbf{A} = (A_{ij}), \qquad A_{ij} = a(\phi_j, \phi_i), \quad i, j = 1, \ldots, N, \tag{2.3a}$$

$$\mathbf{f} = (f_i), \qquad f_i = \langle f, \phi_i \rangle. \tag{2.3b}$$

As an illustrative example used in further sections, we consider a simple one-dimensional Poisson problem

$$-u''(x) = f(x), \qquad x \in \Omega \equiv (0,1), \qquad u(0) = u(1) = 0, \tag{2.4}$$

where f is a given continuous function on $[0,1]$. The weak formulation of (2.4) is given by (2.1) with

$$\mathcal{V} \equiv H_0^1(\Omega), \qquad a(u,v) \equiv \int_\Omega u'(x)v'(x)\mathrm{d}x, \qquad \langle f, v \rangle \equiv \int_\Omega f(x)v(x)\mathrm{d}x,$$

where $H_0^1(\Omega) = \{v \in L^2(\Omega) : v' \in L^2(\Omega), v(0) = v(1) = 0\}$ is the Sobolev space of square integrable functions on the interval Ω which have square integrable (weak) first derivatives and vanish at the end points of the interval (in the sense of traces). We use here $f(x) = 2\alpha[1 - 2\alpha(x - 1/2)^2]\exp[-\alpha(x - 1/2)^2]$ for which the solution of (2.4) is given by $u(x) = \exp[-\alpha(x - 1/2)^2] - \exp(-\alpha/4)$ with $\alpha = 5$. For the discretisation of (2.4), we partition Ω into $N+1$ intervals of constant length $h = 1/(N+1)$ and identify $\mathcal{V}_h$ with the space of continuous functions linear on each interval $[ih, (i+1)h]$ $(i = 0, \ldots, N)$ and choose the standard "hat-shaped" basis $\mathbf{\Phi} = [\phi_1, \ldots, \phi_N]$ of piecewise linear functions such that $\phi_i(jh) = 1$ if $i = j$ and $\phi_i(jh) = 0$ if $i \neq j$. The matrix $\mathbf{A}$ and the right-hand side vector $\mathbf{f}$ are respectively given by

$$\mathbf{A} = h^{-1} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{N \times N},$$

$$\mathbf{f} = (f_i), \quad f_i = \int_0^1 f(x)\phi_i(x)\mathrm{d}x, \quad i = 1, \ldots, N.$$

We set $N = 20$ but the actual dimension is not important for the illustrative purpose.

ETNA
Kent State University
http://etna.math.kent.edu

3. Energy backward error and its interpretation in the Galerkin FEM. Let $\hat{\mathbf{u}} \in \mathbb{R}^N$ be an approximation to the solution $\mathbf{u}$ of (1.1). In the backward error analysis, the vector $\hat{\mathbf{u}}$ is interpreted as the solution of a problem (1.1), where the system data $\mathbf{A}$ and $\mathbf{f}$ are perturbed. We restrict ourselves here to the extreme cases where we consider perturbations only in the right-hand side or the system matrix.

In this section, we discuss how such perturbations in the linear algebraic system may be interpreted in the problem it represents, that is, in the discrete problem (2.2). The representation of the residual vector is quite straightforward and well known (see, e.g., [1, 5]) but we include this case for the sake of completeness. We are, however, mainly interested in interpreting the perturbations in the matrix $\mathbf{A}$ itself, where some interesting questions may arise, e.g., whether the symmetry and positive definiteness of the perturbed matrix is preserved and whether the perturbed problem still represents a discrete variational problem.

In order to measure properly the perturbation norms in the algebraic environment, we discuss first the choice of the vector norms relevant to the original variational problem, more precisely its discretisation (2.2), where the energy norm induced by the bilinear form $a(\cdot,\cdot)$ is considered. Let $v_h, w_h \in \mathcal{V}_h$ and let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$ be respectively the coordinates of v_h and w_h in the basis $\mathbf{\Phi}$ so that $v_h = \mathbf{\Phi}\mathbf{v}$ and $w_h = \mathbf{\Phi}\mathbf{w}$. From (2.3a) we have

$$a(v_h, w_h) = a(\mathbf{\Phi}\mathbf{v}, \mathbf{\Phi}\mathbf{w}) = \mathbf{w}^T\mathbf{A}\mathbf{v}, \qquad \|v_h\|_a = \|\mathbf{v}\|_{\mathbf{A}} \equiv \sqrt{\mathbf{v}^T\mathbf{A}\mathbf{v}}. \tag{3.1}$$

The energy norm of v_h is hence equal to the $\mathbf{A}$-norm of the vector of their coordinates with respect to the basis $\mathbf{\Phi}$. Let $g_h \in \mathcal{V}_h'$ be such that $\langle g_h, \phi_i\rangle = g_i$, $i = 1, \ldots, N$, and let the vector $\mathbf{g} = [g_1, \ldots, g_N]^T \in \mathbb{R}^N$ represent the discrete functional g_h with respect to the basis $\mathbf{\Phi}$. For any $v_h = \mathbf{\Phi}\mathbf{v} \in \mathcal{V}_h$ with $\mathbf{v} = [v_1, \ldots, v_N]^T$, we have

$$\langle g_h, v_h\rangle = \sum_{i=1}^{N} v_i \langle g_h, \phi_i\rangle = \sum_{i=1}^{N} g_i v_i = \mathbf{g}^T\mathbf{v}. \tag{3.2}$$

From (3.1) and (3.2), the dual norm of g_h is given by

$$\|g_h\|_{a,\star} \equiv \max_{v_h \in \mathcal{V}_h \setminus \{0\}} \frac{\langle g_h, v_h\rangle}{\|v_h\|_a} = \max_{\mathbf{v} \in \mathbb{R}^N \setminus \{0\}} \frac{\mathbf{g}^T\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{A}}} = \|\mathbf{g}\|_{\mathbf{A}^{-1}}, \tag{3.3}$$

that is, the dual norm of g_h is equal to the $\mathbf{A}^{-1}$-norm of the vector of its coordinates with respect to $\mathbf{\Phi}$. The last equality can be obtained using the Cauchy-Schwarz inequality

$$\frac{\mathbf{g}^T\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{A}}} = \frac{\mathbf{g}^T\mathbf{A}^{-1/2}\mathbf{A}^{1/2}\mathbf{v}}{\|\mathbf{v}\|_{\mathbf{A}}} \leq \frac{\|\mathbf{g}\|_{\mathbf{A}^{-1}}\|\mathbf{v}\|_{\mathbf{A}}}{\|\mathbf{v}\|_{\mathbf{A}}} = \|\mathbf{g}\|_{\mathbf{A}^{-1}} \tag{3.4}$$

and choosing $\mathbf{v} = \mathbf{A}^{-1}\mathbf{g}$, which gives equality in (3.4). We can thus consider the matrix $\mathbf{A}$ as the mapping from $\mathbb{R}^N$ to $\mathbb{R}^N$ equipped with the $\mathbf{A}$-norm and $\mathbf{A}^{-1}$-norm, respectively:

$$\mathbf{A} : (\mathbb{R}^N, \|\cdot\|_{\mathbf{A}}) \to (\mathbb{R}^N, \|\cdot\|_{\mathbf{A}^{-1}}). \tag{3.5}$$

The accuracy of the given approximation $\hat{\mathbf{u}}$ of the solution of (1.1) is characterised by the residual vector $\hat{\mathbf{r}} = [\hat{r}_1, \ldots, \hat{r}_N]^T \equiv \mathbf{f} - \mathbf{A}\hat{\mathbf{u}}$. By definition, the vector $\hat{\mathbf{u}}$ satisfies the perturbed algebraic system

$$\mathbf{A}\hat{\mathbf{u}} = \mathbf{f} - \hat{\mathbf{r}}. \tag{3.6}$$

Let $\hat{u}_h = \mathbf{\Phi}\hat{\mathbf{u}} \in \mathcal{V}_h$ be the approximation to the solution u_h of the discrete problem (2.2) obtained from the inexact solution $\hat{\mathbf{u}}$ of the system (1.1) and let $\hat{r}_h \in \mathcal{V}_h'$ be defined

ETNA
Kent State University
http://etna.math.kent.edu

by $\langle \hat{r}_h, \phi_i \rangle = \hat{r}_i$, $i = 1, \ldots, N$. It is straightforward to verify that the system (3.6) is the algebraic representation of the perturbed discrete problem*

$$a(\hat{u}_h, v_h) = \langle f, v_h \rangle - \langle \hat{r}_h, v_h \rangle \qquad \forall v_h \in \mathcal{V}_h. \tag{3.7}$$

From (3.3), the relation $\mathbf{A}(\mathbf{u} - \hat{\mathbf{u}}) = \hat{\mathbf{r}}$, and (3.1), we have for the dual norm of the residual functional $\hat{r}_h$ the relation

$$\|\hat{r}_h\|_{a,\star} = \|\hat{\mathbf{r}}\|_{\mathbf{A}^{-1}} = \|\mathbf{u} - \hat{\mathbf{u}}\|_{\mathbf{A}} = \|u_h - \hat{u}_h\|_a.$$

Note that (3.7) still represents a discretisation of a PDE. In particular for our model Poisson equation, the functional $\hat{r}_h$ can be identified with a piecewise linear perturbation of the right-hand side f and the approximate discrete solution $\hat{u}_h$ can be considered as the (exact) solution of the discretisation of the original problem with the right-hand side f replaced by $f - \hat{r}_h$.

Now we make an attempt to find a suitable interpretation of the perturbation of the system matrix $\mathbf{A}$. Let the approximation $\hat{\mathbf{u}}$ be nonzero and let the matrix $\hat{\mathbf{E}} \in \mathbb{R}^{N \times N}$ be such that $\hat{\mathbf{E}}\hat{\mathbf{u}} = \hat{\mathbf{r}}$ so that the vector $\hat{\mathbf{u}}$ satisfies the perturbed system

$$(\mathbf{A} + \hat{\mathbf{E}})\hat{\mathbf{u}} = \mathbf{f}. \tag{3.8}$$

Note that such an $\hat{\mathbf{E}}$ is not unique; we will consider finding certain optimal perturbations later. According to (3.5), we consistently measure the size of the perturbation $\hat{\mathbf{E}}$ by the norm

$$\|\hat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}} \equiv \max_{\mathbf{v} \in \mathbb{R}^N \setminus \{0\}} \frac{\|\hat{\mathbf{E}}\mathbf{v}\|_{\mathbf{A}^{-1}}}{\|\mathbf{v}\|_{\mathbf{A}}} = \|\mathbf{A}^{-1/2}\hat{\mathbf{E}}\mathbf{A}^{-1/2}\|_2, \tag{3.9}$$

where $\|\cdot\|_2$ denotes the spectral matrix norm and $\mathbf{A}^{1/2}$ the unique SPD square root of the matrix $\mathbf{A}$. We will refer to the norm defined by (3.9) as the *energy norm* of the matrix $\hat{\mathbf{E}}$.

We can consider an approach similar to what is called the functional backward error in [5]. The matrix $\hat{\mathbf{E}} = (\hat{E}_{ij})$ can be identified with the bilinear form $\hat{e}_h$ on $\mathcal{V}_h$ defined by $\hat{e}_h(\phi_j, \phi_i) = \hat{E}_{ij}$, $i, j = 1, \ldots, N$. It is then straightforward to show that†

$$a(\hat{u}_h, v_h) + \hat{e}_h(\hat{u}_h, v_h) = \langle f, v_h \rangle \qquad \forall v_h \in \mathcal{V}_h. \tag{3.10}$$

That is, the discrete variational problem (3.10) is represented in the basis $\mathbf{\Phi}$ by the perturbed system (3.8). The norm of $\hat{e}_h$ is given by the energy norm of $\hat{\mathbf{E}}$

$$\max_{v_h, w_h \in \mathcal{V}_h \setminus \{0\}} \frac{\hat{e}_h(v_h, w_h)}{\|v_h\|_a \|w_h\|_a} = \max_{\mathbf{v}, \mathbf{w} \in \mathbb{R}^N \setminus \{0\}} \frac{\mathbf{w}^T \hat{\mathbf{E}} \mathbf{v}}{\|\mathbf{v}\|_{\mathbf{A}} \|\mathbf{w}\|_{\mathbf{A}}} = \|\hat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}}.$$

Note that the matrix $\mathbf{A} + \hat{\mathbf{E}}$ does not need to be sparse nor symmetric (depending on the structure of the perturbation matrix $\hat{\mathbf{E}}$), and in general it does not need to be nonsingular. The form $\hat{e}_h$ therefore does not need to be symmetric either.

It is not easy (if possible) to find a reasonable interpretation of the bilinear form $\hat{e}_h$, e.g., to find out whether the perturbed variational problem (3.10) still represents a discretised PDE. We thus look for a different interpretation of (3.8) which might preserve the character of the

*For the sake of simplicity, we restrict ourselves to the discrete space $\mathcal{V}_h$, although we could interpret (3.7) as the discretisation of a perturbed (continuous) variational problem (2.1) with $\hat{r}_h$ replaced by a proper norm-preserving extension to $\mathcal{V}'$ due to the Hahn-Banach theorem; see, e.g., [22].

†Again, we restrict ourselves to the discrete space and do not consider the extension of $\hat{e}_h$ to $\mathcal{V}$.

ETNA
Kent State University
http://etna.math.kent.edu

original problem. In particular, we will see that the perturbed system (3.8) can be considered as a certain perturbation of the basis $\mathbf{\Phi}$ in which the approximate solution $\hat{\mathbf{u}}$ provides coordinates of the (exact) discrete solution u_h.

Let $\hat{\mathbf{\Phi}} = [\hat{\Phi}_1, \dots, \hat{\Phi}_N]$ be a basis of $\mathcal{V}_h$ obtained from the basis $\mathbf{\Phi}$ by perturbing its individual components by linear combinations of the original basis $\mathbf{\Phi}$. We can write

$$\hat{\mathbf{\Phi}} = \mathbf{\Phi}(\mathbf{I} + \hat{\mathbf{D}}), \quad \text{that is,} \quad \hat{\phi}_j = \phi_j + \sum_{k=1}^{N} \hat{D}_{kj}\phi_k, \qquad j = 1, \dots, N, \tag{3.11}$$

where $\hat{\mathbf{D}} = (\hat{D}_{ij}) \in \mathbb{R}^{N\times N}$ is a matrix of perturbation coefficients and $\mathbf{I}$ denotes the identity matrix. We assume that $\mathbf{I}+\hat{\mathbf{D}}$ is nonsingular so that $\hat{\mathbf{\Phi}}$ is indeed a basis of $\mathcal{V}_h$. We look for the discrete solution u_h given by the linear combination of the modified basis $\hat{\mathbf{\Phi}}$ with coefficients given by the vector $\hat{\mathbf{u}}$. If $u_h = \hat{\mathbf{\Phi}}\hat{\mathbf{u}}$ with $\hat{\mathbf{u}} = [\hat{u}_1, \dots, \hat{u}_N]^T$ and $\hat{\mathbf{\Phi}}$ as in (3.11), we have

$$\begin{aligned} a(u_h, \phi_i) &= \sum_{j=1}^{N} a(\hat{\phi}_j, \phi_i)\hat{u}_j = \sum_{j=1}^{N} \left(a(\phi_j, \phi_i) + \sum_{k=1}^{N} \hat{D}_{kj} a(\phi_k, \phi_i) \right) \hat{u}_j \\ &= \sum_{j=1}^{N} \left(A_{ij} + \sum_{k=1}^{N} A_{ik}\hat{D}_{kj} \right) \hat{u}_j = \left[(\mathbf{A} + \mathbf{A}\hat{\mathbf{D}})\hat{\mathbf{u}} \right]_i , \end{aligned}$$

where $[\cdot]_i$ denotes the i-th component of the vector given in the argument. Hence requiring (2.2) to hold for $v_h = \phi_i$, $i = 1, \dots, N$, leads to

$$(\mathbf{A} + \hat{\mathbf{E}})\hat{\mathbf{u}} = \mathbf{f}, \qquad \hat{\mathbf{E}} = \mathbf{A}\hat{\mathbf{D}},$$

that is, to the perturbed system (3.8) with $\hat{\mathbf{E}} = \mathbf{A}\hat{\mathbf{D}}$. Equivalently, given an approximation $\hat{\mathbf{u}}$ of the solution of the algebraic system (1.1) and the perturbation $\hat{\mathbf{E}}$ such that $\hat{\mathbf{u}}$ satisfies (3.8), there is a basis $\hat{\mathbf{\Phi}}$ given by $\hat{\mathbf{\Phi}} = \mathbf{\Phi}(\mathbf{I} + \hat{\mathbf{D}})$, where $\hat{\mathbf{D}} = \mathbf{A}^{-1}\hat{\mathbf{E}}$ such that the vector $\hat{\mathbf{u}}$ represents the coordinates of the (exact) discrete solution u_h of (2.2) with respect to the modified basis $\hat{\mathbf{\Phi}}$. Note that $\hat{\mathbf{\Phi}}$ is a (linearly independent) basis of $\mathcal{V}_h$ if (and only if) the matrix $\mathbf{A} + \hat{\mathbf{E}}$ (as well as the matrix $\mathbf{I} + \hat{\mathbf{D}}$) is nonsingular.

In order to give the interpretation to the energy norm of $\hat{\mathbf{E}} = \mathbf{A}\hat{\mathbf{D}}$, we define a relative distance between the two bases $\hat{\mathbf{\Phi}}$ and $\mathbf{\Phi}$ by

$$d(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = \max_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{\|\hat{\mathbf{\Phi}}\mathbf{v} - \mathbf{\Phi}\mathbf{v}\|_a}{\|\mathbf{\Phi}\mathbf{v}\|_a}. \tag{3.12}$$

From (3.11) we have

$$\begin{aligned} d(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) &= \max_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{\|\hat{\mathbf{\Phi}}\mathbf{v} - \mathbf{\Phi}\mathbf{v}\|_a}{\|\mathbf{\Phi}\mathbf{v}\|_a} = \max_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{\|\mathbf{\Phi}\hat{\mathbf{D}}\mathbf{v}\|_a}{\|\mathbf{\Phi}\mathbf{v}\|_a} \\ &= \max_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{\|\hat{\mathbf{D}}\mathbf{v}\|_{\mathbf{A}}}{\|\mathbf{v}\|_{\mathbf{A}}} = \max_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{\|\mathbf{A}^{-1}\hat{\mathbf{E}}\mathbf{v}\|_{\mathbf{A}}}{\|\mathbf{v}\|_{\mathbf{A}}} = \max_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{\|\hat{\mathbf{E}}\mathbf{v}\|_{\mathbf{A}^{-1}}}{\|\mathbf{v}\|_{\mathbf{A}}} \\ &= \|\hat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}}, \end{aligned}$$

that is, the relative distance between the bases $\hat{\mathbf{\Phi}}$ and $\mathbf{\Phi}$ related by (3.11) is equal to the energy norm of the matrix $\hat{\mathbf{E}} = \mathbf{A}\hat{\mathbf{D}}$. We summarise the discussion above in the following theorem.

ETNA
Kent State University
http://etna.math.kent.edu

THEOREM 3.1. *Let $\hat{\mathbf{u}}$ be a nonzero approximate solution of the system (1.1) representing algebraically the discretised variational problem (2.2) with respect to the basis $\mathbf{\Phi}$ of $\mathcal{V}_h$. Let $\hat{\mathbf{E}}$ be such that $\hat{\mathbf{u}}$ satisfies the perturbed system (3.8) and let $\mathbf{A}+\hat{\mathbf{E}}$ be nonsingular. Then the vector $\hat{\mathbf{u}}$ contains the coordinates of the solution u_h of (2.2) with respect to the basis $\hat{\mathbf{\Phi}}$ given by (3.11) with $\hat{\mathbf{D}} = \mathbf{A}^{-1}\hat{\mathbf{E}}$. In addition, the perturbed system (3.8) is the algebraic representation of the discrete variational problem (2.2) with respect to the bases $\hat{\mathbf{\Phi}}$ and $\mathbf{\Phi}$. The relative distance (3.12) between $\hat{\mathbf{\Phi}}$ and $\mathbf{\Phi}$ is given by the energy norm of $\hat{\mathbf{E}}$.*

For a given nonzero vector $\hat{\mathbf{u}}$, there are "many" perturbations $\hat{\mathbf{E}}$ so that $\hat{\mathbf{E}}\hat{\mathbf{u}} = \hat{\mathbf{r}}$. Equivalently, there are many bases $\hat{\mathbf{\Phi}}$ which can be (linearly) combined to u_h using the vector of coordinates $\hat{\mathbf{u}}$. We look hence for the perturbation $\hat{\mathbf{E}}$ optimal with respect to the energy norm. For this purpose we define the *energy backward error* by

$$\xi(\hat{\mathbf{u}}) \equiv \min\left\{ \|\hat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}} : \ \hat{\mathbf{E}} \in \mathbb{R}^{N\times N},\ (\mathbf{A}+\hat{\mathbf{E}})\hat{\mathbf{u}} = \mathbf{f} \right\}. \tag{3.13}$$

The following theorem holds for any system (1.1) with a symmetric positive definite matrix $\mathbf{A}$.

THEOREM 3.2. *Let $\hat{\mathbf{u}}$ be a nonzero approximation of the solution of (1.1) with a symmetric positive definite matrix $\mathbf{A}$ and let $\hat{\mathbf{r}} = \mathbf{f} - \mathbf{A}\hat{\mathbf{u}}$ be the associated residual vector. Then*

$$\xi(\hat{\mathbf{u}}) = \frac{\|\hat{\mathbf{r}}\|_{\mathbf{A}^{-1}}}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}} = \frac{\|\mathbf{u}-\hat{\mathbf{u}}\|_{\mathbf{A}}}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}}. \tag{3.14}$$

The matrix $\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ for which the minimum in (3.13) is attained is given by

$$\hat{\mathbf{E}}_*(\hat{\mathbf{u}}) \equiv \frac{\hat{\mathbf{r}}\hat{\mathbf{u}}^T\mathbf{A}}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}^2}. \tag{3.15}$$

The matrix $\mathbf{A}+\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ is nonsingular if $\xi(\hat{\mathbf{u}}) < 1$.

Proof. The proof essentially follows that of [12, Theorem 7.1]. Let $\hat{\mathbf{E}}$ be any matrix such that (3.8) holds and hence $\xi(\hat{\mathbf{u}}) \le \|\hat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}}$ due to (3.13). From $\hat{\mathbf{E}}\hat{\mathbf{u}} = \mathbf{f} - \mathbf{A}\hat{\mathbf{u}} = \hat{\mathbf{r}}$ we have that $(\mathbf{A}^{-1/2}\hat{\mathbf{E}}\mathbf{A}^{-1/2})(\mathbf{A}^{1/2}\hat{\mathbf{u}}) = \mathbf{A}^{-1/2}\hat{\mathbf{r}}$. By taking the 2-norm on both sides and using (3.9), we get

$$\|\hat{\mathbf{r}}\|_{\mathbf{A}^{-1}} \le \|\mathbf{A}^{-1/2}\hat{\mathbf{E}}\mathbf{A}^{-1/2}\|_2\|\hat{\mathbf{u}}\|_{\mathbf{A}} = \|\hat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}}\|\hat{\mathbf{u}}\|_{\mathbf{A}}$$

and thus

$$\frac{\|\hat{\mathbf{r}}\|_{\mathbf{A}^{-1}}}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}} \le \|\hat{\mathbf{E}}\|_{\mathbf{A},\mathbf{A}^{-1}}.$$

Hence the ratio $\|\hat{\mathbf{r}}\|_{\mathbf{A}^{-1}}/\|\hat{\mathbf{u}}\|_{\mathbf{A}}$ is a lower bound of $\xi(\hat{\mathbf{u}})$. To prove equality, we consider the matrix $\hat{\mathbf{E}} = \hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ given by (3.15). It is easy to see that $\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\hat{\mathbf{u}} = \hat{\mathbf{r}}$. Indeed,

$$\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\hat{\mathbf{u}} = \frac{\hat{\mathbf{r}}(\hat{\mathbf{u}}^T\mathbf{A}\hat{\mathbf{u}})}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}^2} = \frac{\|\hat{\mathbf{u}}\|_{\mathbf{A}}^2}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}^2}\hat{\mathbf{r}} = \hat{\mathbf{r}}$$

and hence $\hat{\mathbf{E}} = \hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ satisfies (3.8). Its energy norm is given by

$$\|\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\|_{\mathbf{A},\mathbf{A}^{-1}} = \|\mathbf{A}^{-1/2}\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\mathbf{A}^{-1/2}\|_2 = \frac{\|\mathbf{A}^{-1/2}\hat{\mathbf{r}}\hat{\mathbf{u}}^T\hat{\mathbf{A}}^{1/2}\|_2}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}^2} = \frac{\|\hat{\mathbf{r}}\|_{\mathbf{A}^{-1}}}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}},$$

ETNA
Kent State University
http://etna.math.kent.edu

where the last equality follows from the fact that $\|\mathbf{B}\|_2 = \|\mathbf{v}\|_2\|\mathbf{w}\|_2$ holds true for the matrix $\mathbf{B} = \mathbf{v}\mathbf{w}^T \in \mathbb{R}^{N\times N}$ with $\mathbf{v}, \mathbf{w} \in \mathbb{R}^N$; see, e.g., [27, Problem 2.3.9]. Therefore,

$$\|\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\|_{\mathbf{A},\mathbf{A}^{-1}} = \frac{\|\hat{\mathbf{r}}\|_{\mathbf{A}^{-1}}}{\|\hat{\mathbf{u}}\|_{\mathbf{A}}} \le \xi(\hat{\mathbf{u}}) \le \|\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\|_{\mathbf{A},\mathbf{A}^{-1}},$$

which (together with $\mathbf{A}(\mathbf{u} - \hat{\mathbf{u}}) = \hat{\mathbf{r}}$) implies that (3.14) holds. It is well known (see, e.g., [24, Corollary 2.7]) that $\mathbf{A} + \hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ is nonsingular if

$$\frac{\|\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\|_{\mathbf{A},\mathbf{A}^{-1}}}{\|\mathbf{A}\|_{\mathbf{A},\mathbf{A}^{-1}}} < \frac{1}{\kappa_{\mathbf{A},\mathbf{A}^{-1}}(\mathbf{A})},$$

where for a nonsingular matrix $\mathbf{X}$

$$\kappa_{\mathbf{A},\mathbf{A}^{-1}}(\mathbf{X}) = \|\mathbf{X}\|_{\mathbf{A},\mathbf{A}^{-1}}\|\mathbf{X}^{-1}\|_{\mathbf{A}^{-1},\mathbf{A}}.$$

Since $\|\mathbf{A}\|_{\mathbf{A},\mathbf{A}^{-1}} = \|\mathbf{A}^{-1}\|_{\mathbf{A}^{-1},\mathbf{A}} = 1$, we obtain that the matrix $\mathbf{A} + \hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ is nonsingular if $\xi(\hat{\mathbf{u}}) = \|\hat{\mathbf{E}}_*(\hat{\mathbf{u}})\|_{\mathbf{A},\mathbf{A}^{-1}} < 1$. □

The optimal perturbation $\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ defined in Theorem 3.2 is related to certain optimal perturbation of the basis $\mathbf{\Phi}$. In fact, combining Theorems 3.1 and 3.2, we obtain the following result.

THEOREM 3.3. *Let $\hat{\mathbf{u}}$ be a nonzero approximate solution of the system (1.1) representing algebraically the discretised variational problem (2.2) with respect to the basis $\mathbf{\Phi}$ of $\mathcal{V}_h$ and let $\xi(\hat{\mathbf{u}}) < 1$. Then $\hat{\mathbf{u}}$ is the solution of the perturbed problem*

$$\left(\mathbf{A} + \hat{\mathbf{E}}_*(\hat{\mathbf{u}})\right)\hat{\mathbf{u}} = \mathbf{f}$$

with the perturbation matrix $\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ given by (3.15). Furthermore, let $\hat{\mathbf{D}}_*(\hat{\mathbf{u}}) \equiv \mathbf{A}^{-1}\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ and $\hat{\mathbf{\Phi}}_*(\hat{\mathbf{u}}) \equiv \mathbf{\Phi}(\mathbf{I} + \hat{\mathbf{D}}_*(\hat{\mathbf{u}}))$. Then $\hat{\mathbf{\Phi}}_*(\hat{\mathbf{u}})$ is the basis of $\mathcal{V}_h$ closest to the basis $\mathbf{\Phi}$ in terms of the relative distance (3.12) among all bases of $\mathcal{V}_h$ in which the vector $\hat{\mathbf{u}}$ represents the coordinates of the solution u_h of (2.2). Their relative distance is given by the energy backward error $\xi(\hat{\mathbf{u}})$ in (3.13) and (3.14), that is, $d(\hat{\mathbf{\Phi}}_*(\hat{\mathbf{u}}), \mathbf{\Phi}) = \xi(\hat{\mathbf{u}})$.

REMARK 3.4. Backward errors provide bounds on forward errors (relative norms of the error) via the condition number of the matrix $\mathbf{A}$ (with respect to consistently chosen norms). If $\hat{\mathbf{u}}$ satisfies the perturbed system (3.8) and the condition number $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ is such that $\kappa(A)\|\hat{\mathbf{E}}\|/\|\mathbf{A}\| < 1$, the forward error can be bounded by

$$\frac{\|\mathbf{u} - \hat{\mathbf{u}}\|}{\|\mathbf{u}\|} \le \frac{\kappa(\mathbf{A})\|\hat{\mathbf{E}}\|/\|\mathbf{A}\|}{1 - \kappa(\mathbf{A})\|\hat{\mathbf{E}}\|/\|\mathbf{A}\|}, \tag{3.16}$$

see, e.g., [24, Theorem 2.11]. With our choice of norms, both forward and backward errors do coincide since the condition number and the norm of the matrix $\mathbf{A}$ are equal to one. The bound (3.16) then (with $\hat{\mathbf{E}} = \hat{\mathbf{E}}_*(\hat{\mathbf{u}})$) becomes

$$\frac{\|\mathbf{u} - \hat{\mathbf{u}}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} \le \frac{\xi(\hat{\mathbf{u}})}{1 - \xi(\hat{\mathbf{u}})}$$

provided that $\xi(\hat{\mathbf{u}}) < 1$. In addition, from $\|\mathbf{u}\|_{\mathbf{A}} \le \|\hat{\mathbf{u}}\|_{\mathbf{A}}(1 + \xi(\hat{\mathbf{u}}))$, we have

$$\frac{\|\mathbf{u} - \hat{\mathbf{u}}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} \ge \frac{\xi(\hat{\mathbf{u}})}{1 + \xi(\hat{\mathbf{u}})}$$

ETNA
Kent State University
http://etna.math.kent.edu

and hence the forward and backward error in the $\mathbf{A}$-norm are equivalent in the sense that

$$\frac{\xi(\hat{\mathbf{u}})}{1+\xi(\hat{\mathbf{u}})} \leq \frac{\|\mathbf{u}-\hat{\mathbf{u}}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} \leq \frac{\xi(\hat{\mathbf{u}})}{1-\xi(\hat{\mathbf{u}})} \qquad \text{if } \xi(\hat{\mathbf{u}}) < 1.$$

Note that this is simply due to the fact that the condition number of $\mathbf{A}$ is one with respect to the chosen matrix norms.

The perturbation matrix $\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ is determined by the errors in solving the system (1.1). Minimising the energy norm of $\hat{\mathbf{E}}$ generally leads to a dense (and nonsymmetric) perturbation matrix $\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ (although structured, in our case of rank one). The corresponding transformation matrix $\hat{\mathbf{D}}_*(\hat{\mathbf{u}}) = \mathbf{A}^{-1}\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ is dense as well, which means that the perturbed matrix $\hat{\mathbf{\Phi}}_*(\hat{\mathbf{u}})$ has global supports even though the supports of $\mathbf{\Phi}$ can be local. This would be the case even if we considered the component-wise perturbations $\hat{\mathbf{E}}$ [18] since the inverse of $\mathbf{A}$ (and hence the transformation matrix $\hat{\mathbf{D}}$) is generally dense. This is, however, not important for the interpretation of the perturbation coefficients itself.

We illustrate our observations at the model problem described in Section 2, which we solve approximately using the conjugate gradient (CG) method [11]. It is well known that, given an initial guess $\mathbf{u}_0$ with the residual $\mathbf{r}_0 \equiv \mathbf{f} - \mathbf{A}\mathbf{u}_0$, CG generates the approximations $\mathbf{u}_n^{\mathrm{CG}} \in \mathbf{u}_0 + \mathcal{K}_n$, where $\mathcal{K}_n$ is the Krylov subspace $\mathcal{K}_n \equiv \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{n-1}\mathbf{r}_0\}$, such that

$$\|\mathbf{u}-\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}} = \min_{\hat{\mathbf{u}} \in \mathbf{u}_0+\mathcal{K}_n} \|\mathbf{u}-\hat{\mathbf{u}}\|_{\mathbf{A}}. \tag{3.17}$$

In Figure 3.1, we display the exact solution of the discrete problem, the relative $\mathbf{A}$-norms

$$\epsilon_n^{\mathrm{CG}} \equiv \frac{\|\mathbf{u}-\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} \tag{3.18}$$

of the errors of the CG approximations $\mathbf{u}_n^{\mathrm{CG}}$ and their associated energy backward errors $\xi(\mathbf{u}_n^{\mathrm{CG}})$ (where we set $\mathbf{u}_0 = 0$). The backward errors of the CG approximations, although monotonically decreasing as we will see in the next section, need not to be necessarily smaller than one as it is the case for the relative error norms ϵ_n^{CG}. For our model problem, we have (note that ξ is not defined for the initial guess $\mathbf{u}_0 = 0$)

$$\xi(\mathbf{u}_1^{\mathrm{CG}}) = 1.2718, \quad \xi(\mathbf{u}_3^{\mathrm{CG}}) = 1.0572, \quad \xi(\mathbf{u}_4^{\mathrm{CG}}) = 0.8658.$$

In order to demonstrate how the perturbation and transformation matrices $\hat{\mathbf{E}}_*(\hat{\mathbf{u}})$ and $\hat{\mathbf{D}}_*(\hat{\mathbf{u}})$ defined in Theorems 3.2 and 3.3, respectively, look like, we consider two approximations $\hat{\mathbf{u}}$ computed by CG at the iterations 1 and 5, that is, we take $\hat{\mathbf{u}} = \mathbf{u}_1^{\mathrm{CG}}$ and $\hat{\mathbf{u}} = \mathbf{u}_5^{\mathrm{CG}}$. In Figure 3.2 we display (together with the exact solution u_h of the discrete problem) the approximations $u_{h,n}^{\mathrm{CG}} = \mathbf{\Phi}\mathbf{u}_n^{\mathrm{CG}}$ of u_h constructed from the CG approximations $\mathbf{u}_n^{\mathrm{CG}}$ (for $n = 1$ and $n = 5$). The entries of the perturbation and transformation matrices $\hat{\mathbf{E}}_*(\mathbf{u}_n^{\mathrm{CG}})$ and $\hat{\mathbf{D}}_*(\mathbf{u}_n^{\mathrm{CG}})$, respectively, corresponding to these approximate solutions are visualised in Figures 3.3 and 3.4 (using the MATLAB command `surf`). Since the standard hat-shaped basis $\mathbf{\Phi}$ is used, the interior nodal values of $u_{h,n}^{\mathrm{CG}}$ are equal to the corresponding components of the vectors $\mathbf{u}_n^{\mathrm{CG}}$. We would get u_h by forming linear combinations of the basis $\mathbf{\Phi}(\mathbf{I} + \mathbf{D}_*(\mathbf{u}_n^{\mathrm{CG}}))$ using the coefficients $\mathbf{u}_n^{\mathrm{CG}}$ obtained by the n-th CG iteration which, at the same time, satisfy the perturbed problems $(\mathbf{A} + \mathbf{E}_*(\mathbf{u}_n^{\mathrm{CG}}))\mathbf{u}_n^{\mathrm{CG}} = \mathbf{f}$.

ETNA
Kent State University
http://etna.math.kent.edu

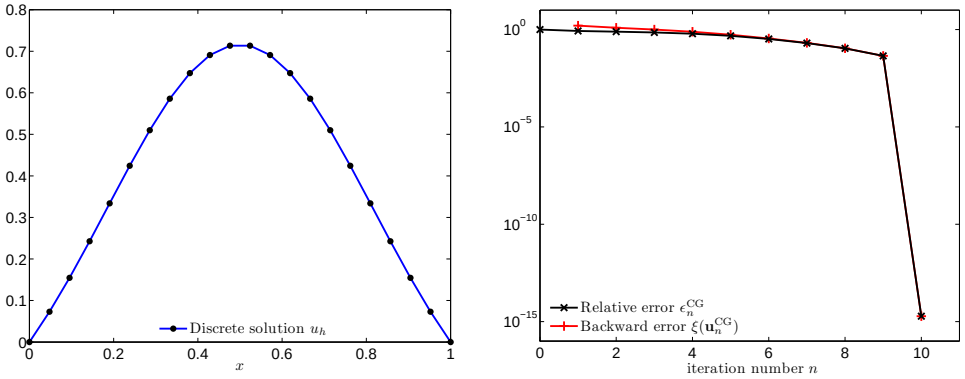

FIG. 3.1. *The discrete solution u_h of the model problem on the left plot and the convergence of CG in terms of the relative $\mathbf{A}$-norm of the error $\epsilon_n^{\mathrm{CG}} = \|\mathbf{u} - \mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}/\|\mathbf{u}\|_{\mathbf{A}}$ and of the energy backward error $\xi(\mathbf{u}_n^{\mathrm{CG}})$ on the right plot.*

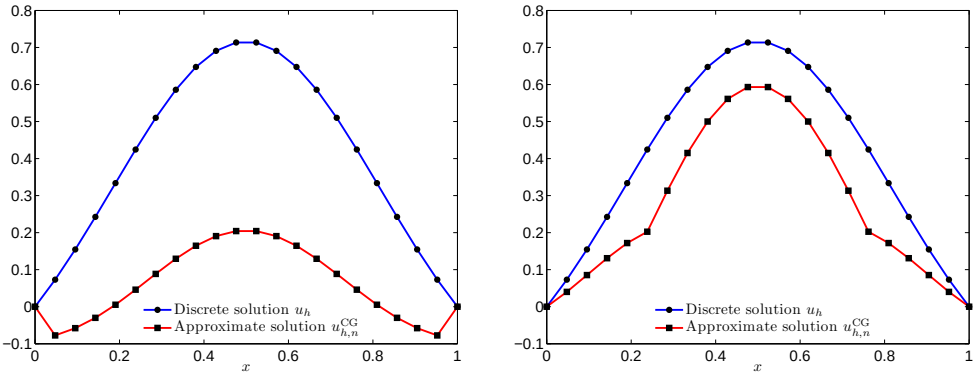

FIG. 3.2. *The discrete solution u_h and the approximate solution $u_{h,n}^{\mathrm{CG}} = \mathbf{\Phi}\mathbf{u}_n^{\mathrm{CG}}$ for $n = 1$ (left plot) and $n = 5$ (right plot).*

4. Conjugate gradient method and energy backward error. The conjugate gradient method constructs, starting from the initial guess $\mathbf{u}_0$, the sequence of approximations $\mathbf{u}_n^{\mathrm{CG}}$ from the (shifted) Krylov subspace $\mathbf{u}_0 + \mathcal{K}_n$. Similarly to the Galerkin method, the approximations $\mathbf{u}_n^{\mathrm{CG}}$ minimise the discrete energy norm ($\mathbf{A}$-norm) of the error $\mathbf{u} - \mathbf{u}_n^{\mathrm{CG}}$ in the sense of (3.17). Equivalently, the error $\mathbf{e}_n^{\mathrm{CG}} \equiv \mathbf{u} - \mathbf{u}_n^{\mathrm{CG}}$ is $\mathbf{A}$-orthogonal to $\mathcal{K}_n$.

REMARK 4.1. In the Galerkin finite element method, there is even more about the optimality of CG than in the iterative method itself. If $u_{h,n}^{\mathrm{CG}} = \mathbf{\Phi}\mathbf{u}_n^{\mathrm{CG}}$ is the associated approximation of the solution of the discrete problem (2.2), we have

$$\|u - u_{h,n}^{\mathrm{CG}}\|_a = \min_{v_h \in \mathbf{\Phi}(\mathbf{u}_0 + \mathcal{K}_n)} \|u - v_h\|_a,$$

where $\mathbf{\Phi}(\mathbf{u}_0 + \mathcal{K}_n) = \{v_h \in \mathcal{V}_h : v_h = \mathbf{\Phi}\mathbf{v},\ \mathbf{v} \in \mathbf{u}_0 + \mathcal{K}_n\}$. It means that CG provides optimal approximations to the solution u of the (continuous) problem (2.1) from the subspaces of $\mathcal{V}_h$ which consist of all linear combinations of the basis $\mathbf{\Phi}$ with coefficients taken from the shifted Krylov subspaces $\mathbf{u}_0 + \mathcal{K}_n$. This follows from the identity

$$\|u - v_h\|_a^2 = \|u - u_h\|_a^2 + \|u_h - v_h\|_a^2 = \|u - u_h\|_a^2 + \|\mathbf{u} - \mathbf{v}\|_{\mathbf{A}}^2,$$

ETNA
Kent State University
http://etna.math.kent.edu

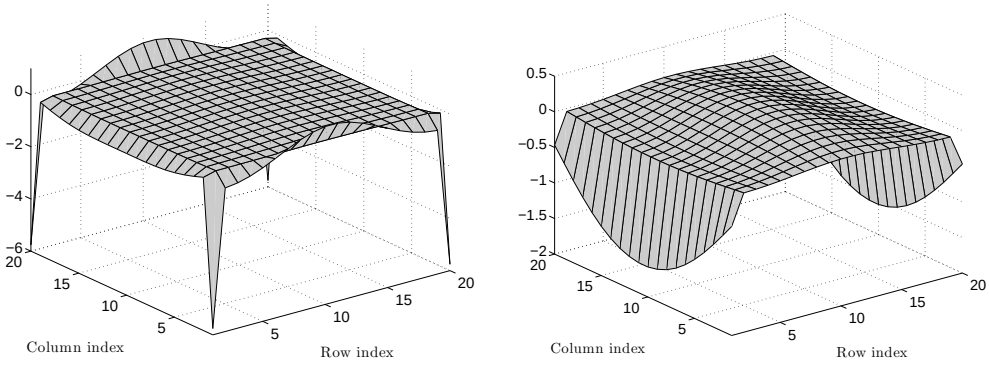

FIG. 3.3. *Surface plots of the perturbation matrix* $\hat{\mathbf{E}}_*(\mathbf{u}_1^{\mathrm{CG}})$ *(left plot) and the transformation matrix* $\hat{\mathbf{D}}_*(\mathbf{u}_1^{\mathrm{CG}})$ *(right plot).*

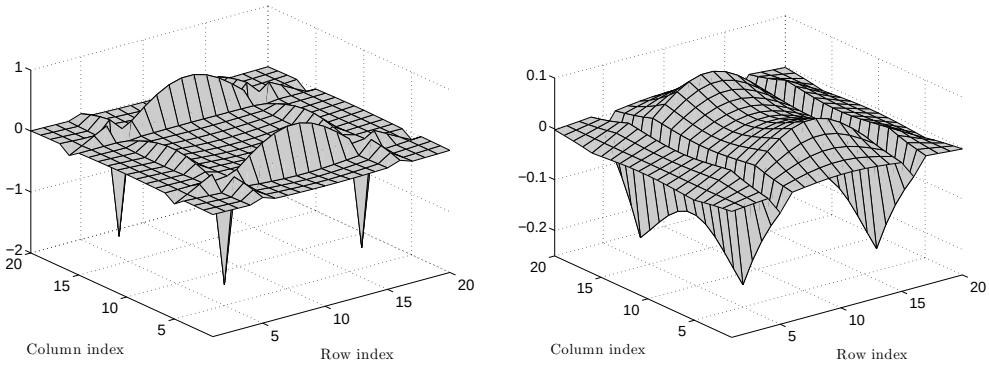

FIG. 3.4. *Surface plots of the perturbation matrix* $\hat{\mathbf{E}}_*(\mathbf{u}_5^{\mathrm{CG}})$ *(left plot) and the transformation matrix* $\hat{\mathbf{D}}_*(\mathbf{u}_5^{\mathrm{CG}})$ *(right plot).*

which holds for any $v_h = \mathbf{\Phi v} \in \mathcal{V}_h$ and is a consequence of the a-orthogonality of $u - u_h$ to $\mathcal{V}_h$; see also [9, Section 2.1], [17, Section 2.5.2], and [20] for more details.

In the following we assume that $\mathbf{u}_0 = 0$. We use a simple relation between the $\mathbf{A}$-norms of the CG error $\mathbf{e}_n^{\mathrm{CG}}$, the solution $\mathbf{u}$, and the CG approximation $\mathbf{u}_n^{\mathrm{CG}}$ of the form

$$\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2 = \|\mathbf{u}\|_{\mathbf{A}}^2 - \|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2, \tag{4.1}$$

which follows from the fact that $\mathbf{u}_n^{\mathrm{CG}} \in \mathcal{K}_n$ and the $\mathbf{A}$-orthogonality of $\mathbf{u} - \mathbf{u}_n^{\mathrm{CG}}$ to $\mathcal{K}_n$:

$$\mathbf{u} = \mathbf{u}_n^{\mathrm{CG}} + (\mathbf{u} - \mathbf{u}_n^{\mathrm{CG}}) \quad \Rightarrow \quad \|\mathbf{u}\|_{\mathbf{A}}^2 = \|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2 + \|\mathbf{u} - \mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2.$$

Using (4.1), the energy backward error of the CG approximation $\mathbf{u}_n^{\mathrm{CG}}$ can be expressed as

$$\xi(\mathbf{u}_n^{\mathrm{CG}}) = \frac{\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}}{\|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}} = \frac{\epsilon_n^{\mathrm{CG}}}{\sqrt{1 - (\epsilon_n^{\mathrm{CG}})^2}}, \tag{4.2}$$

where ϵ_n^{CG} is the relative $\mathbf{A}$-norm of the error $\mathbf{e}_n^{\mathrm{CG}}$; see (3.18). The energy backward error is well defined for every CG iteration except for the zero initial guess. It is due to the fact that the energy norm of the error in CG decreases strictly monotonically at each step. Since ϵ_n^{CG} is decreasing, the energy backward error (4.2) decreases as well in CG. Both $\xi(\mathbf{u}_n^{\mathrm{CG}})$ and ϵ_n^{CG}

ETNA
Kent State University
http://etna.math.kent.edu

are close (as can be observed in Figure 3.1 for our model problem) provided that ϵ_n^{CG} is small enough due to

$$\frac{\epsilon_n^{\mathrm{CG}}}{\xi(\mathbf{u}_n^{\mathrm{CG}})} = \sqrt{1-(\epsilon_n^{\mathrm{CG}})^2}.$$

Note also that $\xi(\mathbf{u}_n^{\mathrm{CG}}) < 1$ if $\epsilon_n^{\mathrm{CG}} < 1/\sqrt{2}$.

One could ask whether it is possible (instead of the $\mathbf{A}$-norm of the error) to minimise the energy backward error ξ over the same Krylov subspace $\mathcal{K}_n$. Let $\mathbf{u}_n$ be an arbitrary vector from $\mathcal{K}_n$ and let $\mathbf{e}_n \equiv \mathbf{u} - \mathbf{u}_n$ be the associated error vector. From $\mathbf{u}_n^{\mathrm{CG}} - \mathbf{u}_n \in \mathcal{K}_n$, the $\mathbf{A}$-orthogonality of $\mathbf{e}_n^{\mathrm{CG}}$ to $\mathcal{K}_n$, and the Pythagorean theorem, we get that

$$\|\mathbf{e}_n\|_{\mathbf{A}}^2 = \|\mathbf{e}_n^{\mathrm{CG}} + (\mathbf{u}_n^{\mathrm{CG}} - \mathbf{u}_n)\|_{\mathbf{A}}^2 = \|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2 + \|\mathbf{u}_n^{\mathrm{CG}} - \mathbf{u}_n\|_{\mathbf{A}}^2. \tag{4.3}$$

From (3.14) and (4.3), we have

$$\xi^2(\mathbf{u}_n) = \frac{\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2 + \|\mathbf{u}_n^{\mathrm{CG}} - \mathbf{u}_n\|_{\mathbf{A}}^2}{\|\mathbf{u}_n\|_{\mathbf{A}}^2}. \tag{4.4}$$

LEMMA 4.2. *Let $\mathbf{v} \in \mathbb{R}^n$ be a given nonzero vector, $\alpha \in \mathbb{R}$, and*

$$\varphi(\mathbf{w}) = \frac{\alpha^2 + \|\mathbf{v} - \mathbf{w}\|_2^2}{\|\mathbf{w}\|_2^2}.$$

Then $\mathbf{w}_* = \gamma\mathbf{v}$ with $\gamma = 1 + (\alpha/\|\mathbf{v}\|_2)^2$ is the unique minimiser of φ over all nonzero vectors $\mathbf{w}$ and it holds that $\varphi(\mathbf{w}_*) = \alpha^2/(\alpha^2 + \|\mathbf{v}\|_2^2)$.

Proof. Let $\mathbf{w} = \eta\mathbf{v} + \mathbf{v}_\perp$ where $\eta \in \mathbb{R}$ and $\mathbf{v}_\perp$ is an arbitrary vector orthogonal to $\mathbf{v}$, that is, $\mathbf{v}_\perp^T\mathbf{v} = 0$. From the Pythagorean theorem we have

$$\varphi(\eta\mathbf{v} + \mathbf{v}_\perp) = \frac{\alpha^2 + (1-\eta)^2\|\mathbf{v}\|_2^2 + \|\mathbf{v}_\perp\|_2^2}{\eta^2\|\mathbf{v}\|_2^2 + \|\mathbf{v}_\perp\|_2^2}. \tag{4.5}$$

Note that φ does not depend on the vector $\mathbf{v}_\perp$ itself but only on its norm. Dividing both the numerator and denominator in (4.5) by the (nonzero) value $\|\mathbf{v}\|_2$, we obtain

$$\varphi(\eta\mathbf{v} + \mathbf{v}_\perp) = \frac{\tilde{\alpha}^2 + (1-\eta)^2 + \zeta^2}{\eta^2 + \zeta^2} \equiv \psi(\eta, \zeta),$$

where $\tilde{\alpha} \equiv \alpha/\|\mathbf{v}\|_2$ and $\zeta \equiv \|\mathbf{v}_\perp\|_2/\|\mathbf{v}\|_2$. Hence the statement is proved by showing that ψ has a global minimum at $(\eta, \zeta) = (\gamma, 0) = (1+\tilde{\alpha}^2, 0)$ and that $\psi(1+\tilde{\alpha}^2, 0) = \tilde{\alpha}^2/(1+\tilde{\alpha}^2)$, which can be shown by standard calculus. The function ψ is smooth everywhere except for $(\eta, \zeta) = 0$. We have

$$\nabla\psi(\eta, \zeta) = -\frac{2}{(\eta^2+\zeta^2)^2}\begin{bmatrix}\eta(\tilde{\alpha}^2+1) - \eta^2 + \zeta^2 \\ \zeta(1+\tilde{\alpha}^2 - 2\eta)\end{bmatrix},$$

and thus we have $\nabla\psi(\eta, \zeta) = 0$ if (and only if) $\eta = 1 + \tilde{\alpha}^2$ and $\zeta = 0$. The minimum can be verified by checking the positive definiteness of the matrix of second derivatives at the stationary point $(\eta, \zeta) = (1 + \tilde{\alpha}^2, 0)$, which holds since

$$\nabla^2\psi(1+\tilde{\alpha}^2, 0) = \frac{2}{(\tilde{\alpha}^2+1)^3}\begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}.$$

ETNA
Kent State University
http://etna.math.kent.edu

Substituting the stationary point into ψ gives

$$\psi(1+\tilde{\alpha}^2,0) = \tilde{\alpha}^2/(1+\tilde{\alpha}^2) = \alpha^2/(\alpha^2+\|\mathbf{v}\|_2^2) < 1.$$

The minimum is also global since $\varphi(t\mathbf{w}) \to 1$ as $t \to \infty$ for any fixed $\mathbf{w}$. ∎

THEOREM 4.3. *Let $\mathbf{u}_n^{\mathrm{CG}}$ be the approximation of CG with the initial guess $\mathbf{u}_0 = 0$ at the step $n > 1$. Then the unique vector $\mathbf{u}_n^*$ minimising the energy backward error ξ over all $\mathbf{v}_n \in \mathcal{K}_n$ is given by*

$$\mathbf{u}_n^* = \gamma_n \mathbf{u}_n^{\mathrm{CG}},$$

where

$$\gamma_n = 1 + \xi^2(\mathbf{u}_n^{\mathrm{CG}}) = \frac{1}{1-(\epsilon_n^{\mathrm{CG}})^2}.$$

The energy backward error of $\mathbf{u}_n^*$ is equal to the relative $\mathbf{A}$-norm of the CG error

$$\xi(\mathbf{u}_n^*) = \frac{\|\mathbf{u}-\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} = \epsilon_n^{\mathrm{CG}}.$$

Proof. The relation (4.4) can be written as

$$\xi^2(\mathbf{u}_n) = \frac{\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2 + \|\mathbf{A}^{1/2}(\mathbf{u}_n^{\mathrm{CG}}-\mathbf{u}_n)\|_2^2}{\|\mathbf{A}^{1/2}\mathbf{u}_n\|_2^2}.$$

If we set $\mathbf{w} \equiv \mathbf{A}^{1/2}\mathbf{u}_n$, $\mathbf{v} \equiv \mathbf{A}^{1/2}\mathbf{u}_n^{\mathrm{CG}}$, $\alpha \equiv \|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}$, we have from Lemma 4.2 that the minimum of $\xi^2(\mathbf{u}_n)$ is attained at $\mathbf{u}_n^* = \gamma_n \mathbf{u}_n^{\mathrm{CG}}$ with

$$\gamma_n = 1 + \frac{\alpha^2}{\|\mathbf{v}\|_2^2} = 1 + \frac{\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2}{\|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2} = 1 + \xi^2(\mathbf{u}_n^{\mathrm{CG}}) = \frac{1}{1-(\epsilon_n^{\mathrm{CG}})^2},$$

where the last equality follows from (4.2). The minimum is given by

$$\xi(\mathbf{u}_n^*) = \frac{\alpha}{\sqrt{\alpha^2+\|\mathbf{v}\|_2^2}} = \frac{\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}}{\sqrt{\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2 + \|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2}} = \frac{\|\mathbf{e}_n^{\mathrm{CG}}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} = \epsilon_n^{\mathrm{CG}}$$

using (4.1) again. ∎

The approximations $\mathbf{u}_n^*$ minimising the energy backward error ξ over the Krylov subspace $\mathcal{K}_n$ are thus given by a simple scalar multiple of the CG approximations $\mathbf{u}_n^{\mathrm{CG}}$. It is clear that $\mathbf{u}_n^* \approx \mathbf{u}_n^{\mathrm{CG}}$ provided that the relative error ϵ_n^{CG} is small enough and the difference between both approximations gets smaller with the decreasing $\mathbf{A}$-norm of the CG approximations.

REMARK 4.4. There is an interesting "symmetry" between the relative $\mathbf{A}$-norms of the errors and the energy backward errors of the approximations $\mathbf{u}_n^{\mathrm{CG}}$ and $\mathbf{u}_n^*$ illustrated in Table 4.1. The expression for the relative energy norm of the error of $\mathbf{u}_n^*$ follows from (3.14) and Theorem 4.3

$$\xi(\mathbf{u}_n^*) = \epsilon_n^{\mathrm{CG}} = \frac{\|\mathbf{u}-\mathbf{u}_n^*\|_{\mathbf{A}}}{\|\mathbf{u}_n^*\|_{\mathbf{A}}},$$

and hence together with (4.1) we get

$$\frac{\|\mathbf{e}_n^*\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} = \xi(\mathbf{u}_n^*)\frac{\|\mathbf{u}_n^*\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} = \gamma_n \xi(\mathbf{u}_n^*)\frac{\|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}} = \frac{\epsilon_n^{\mathrm{CG}}\sqrt{1-(\epsilon_n^{\mathrm{CG}})^2}}{1-(\epsilon_n^{\mathrm{CG}})^2} = \frac{\epsilon_n^{\mathrm{CG}}}{\sqrt{1-(\epsilon_n^{\mathrm{CG}})^2}}.$$

ETNA
Kent State University
http://etna.math.kent.edu

TABLE 4.1
Symmetry between $\mathbf{u}_n^{\mathrm{CG}}$ *and* $\mathbf{u}_n^*$.

	$\mathbf{u}_n^{\mathrm{CG}}$: minimises $\|\mathbf{e}_n\|_{\mathbf{A}}$	$\mathbf{u}_n^*$: minimises $\xi(\mathbf{u}_n)$
$\frac{\|\mathbf{e}_n\|_{\mathbf{A}}}{\|\mathbf{u}\|_{\mathbf{A}}}$	ϵ_n^{CG}	$\epsilon_n^{\mathrm{CG}}[1-(\epsilon_n^{\mathrm{CG}})^2]^{-1/2}$
$\xi(\mathbf{u}_n)$	$\epsilon_n^{\mathrm{CG}}[1-(\epsilon_n^{\mathrm{CG}})^2]^{-1/2}$	ϵ_n^{CG}

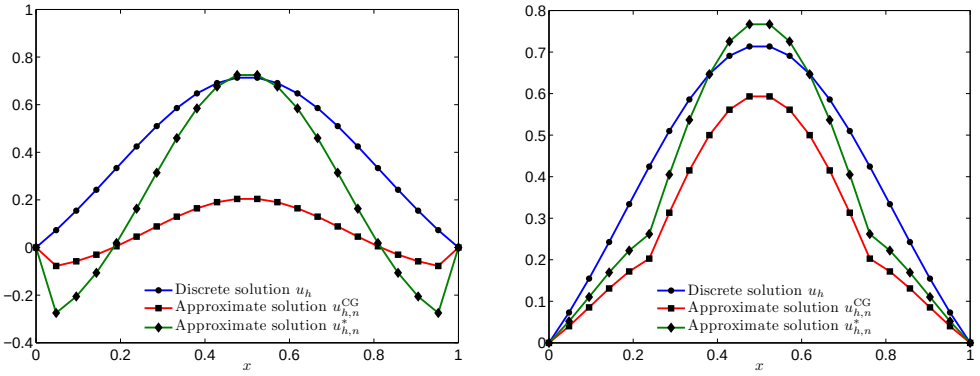

FIG. 4.1. *The discrete solution* u_h *and the approximate solutions* $u_{h,n}^{\mathrm{CG}} = \mathbf{\Phi}\mathbf{u}_n^{\mathrm{CG}}$ *and* $u_{h,n}^* = \mathbf{\Phi}\mathbf{u}_n^*$ *for* $n = 1$ *(left plot) and* $n = 5$ *(right plot).*

In fact, we can also say that the forward error of $\mathbf{u}_n^{\mathrm{CG}}$ is equal to the backward error of $\mathbf{u}_n^*$ and vice versa.

In order to demonstrate the effects of the minimisation of $\xi(\hat{\mathbf{u}})$, we consider as in the previous section the CG approximations obtained at iterations 1 and 5. In Figure 4.1 we show, together with the discrete solution u_h of our model problem, the approximations $u_{h,n}^{\mathrm{CG}} = \mathbf{\Phi}\mathbf{u}_n^{\mathrm{CG}}$ obtained from the CG iterates at steps $n = 1$ and $n = 5$ and the approximations $u_{h,n}^* = \mathbf{\Phi}\mathbf{u}_n^*$ obtained from the CG approximations scaled according to Theorem 4.3. In Figures 4.2 and 4.3, we also show the surface plots of the corresponding perturbations and transformation matrices $\hat{\mathbf{E}}_*(\mathbf{u}_n^*)$ and $\hat{\mathbf{D}}_*(\mathbf{u}_n^*)$ of these scaled CG approximations. It is interesting to observe that although the perturbation matrices $\hat{\mathbf{E}}_*(\mathbf{u}_n^{\mathrm{CG}})$ and $\hat{\mathbf{E}}_*(\mathbf{u}_n^*)$ (left plots of Figures 3.3, 3.4, 4.2, and 4.3) visually look very similar, this is not the case for the transformation matrices $\hat{\mathbf{D}}_*(\mathbf{u}_n^{\mathrm{CG}})$ and $\hat{\mathbf{D}}_*(\mathbf{u}_n^*)$ (right plots of the same figures). This means that (in our example) the scaling of the CG approximations does not change much (at least visually) the coefficients of the perturbation matrices $\hat{\mathbf{E}}(\mathbf{u}_n^{\mathrm{CG}})$, while the changes in the transformation matrices $\hat{\mathbf{D}}(\mathbf{u}_n^{\mathrm{CG}})$ seem to be more prominent.

In order to explain this phenomenon, we evaluate the relative 2-norm of the differences $\hat{\mathbf{E}}_*(\mathbf{u}_n^{\mathrm{CG}}) - \hat{\mathbf{E}}_*(\mathbf{u}_n^*)$ and $\hat{\mathbf{D}}_*(\mathbf{u}_n^{\mathrm{CG}}) - \hat{\mathbf{D}}_*(\mathbf{u}_n^*)$. Let $\mathbf{r}_n^{\mathrm{CG}} \equiv \mathbf{f} - \mathbf{A}\mathbf{u}_n^{\mathrm{CG}}$ be the residual vector of a nonzero CG approximation $\mathbf{u}_n^{\mathrm{CG}}$ different from the exact solution $\mathbf{u}$ of (1.1). Using $\mathbf{u}_n^* = \gamma_n \mathbf{u}_n^{\mathrm{CG}}$ (defined in Theorem 4.3), $\mathbf{f} - \mathbf{A}\mathbf{u}_n^* = \mathbf{f} - \gamma_n \mathbf{A}\mathbf{u}_n^{\mathrm{CG}} = \gamma_n \mathbf{r}_n^{\mathrm{CG}} + (1-\gamma_n)\mathbf{f}$, $(1-\gamma_n)/\gamma_n = -(\epsilon_n^{\mathrm{CG}})^2$, and (3.15), we find

$$\hat{\mathbf{E}}_*(\mathbf{u}_n^{\mathrm{CG}}) = \hat{\mathbf{E}}_*(\mathbf{u}_n^*) + (\epsilon_n^{\mathrm{CG}})^2 \frac{\mathbf{f}(\mathbf{u}_n^{\mathrm{CG}})^T \mathbf{A}}{\|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2}.$$

ETNA
Kent State University
http://etna.math.kent.edu

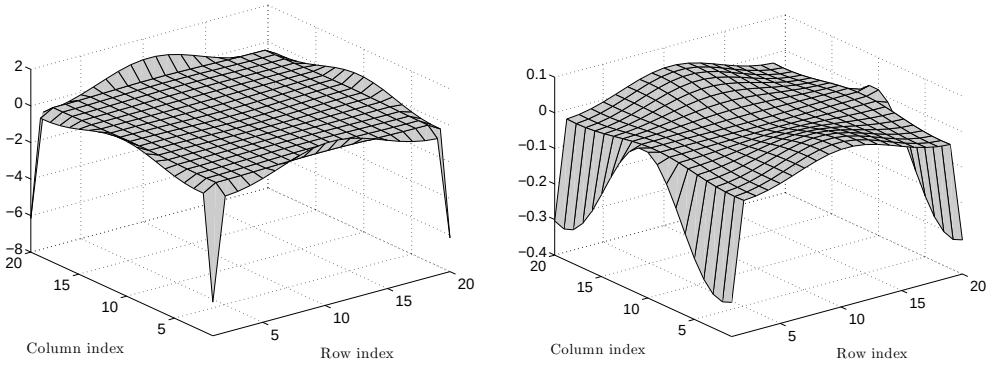

FIG. 4.2. *Surface plots of the perturbation matrix* $\hat{\mathbf{E}}_*(\mathbf{u}_1^*)$ *(left plot) and the transformation matrix* $\hat{\mathbf{D}}_*(\mathbf{u}_1^*)$ *(right plot).*

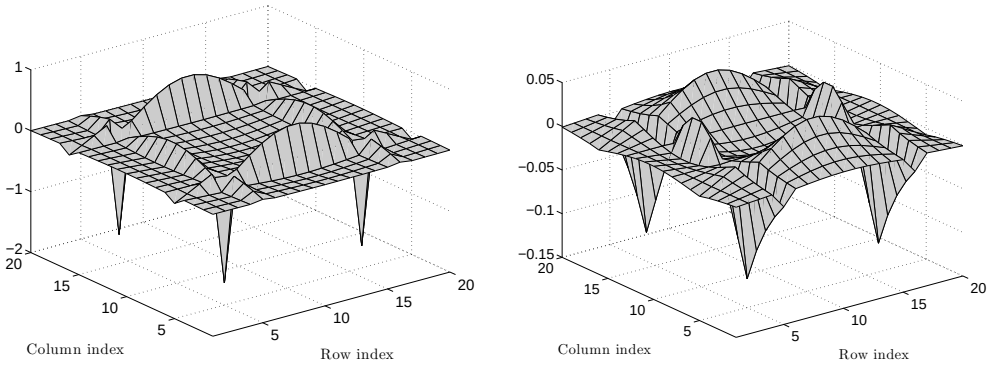

FIG. 4.3. *Surface plots of the perturbation matrix* $\hat{\mathbf{E}}_*(\mathbf{u}_5^*)$ *(left plot) and the transformation matrix* $\hat{\mathbf{D}}_*(\mathbf{u}_5^*)$ *(right plot).*

Due to the relation $\hat{\mathbf{E}}_*(\cdot) = \mathbf{A}\hat{\mathbf{D}}_*(\cdot)$, we have similarly that

$$\hat{\mathbf{D}}_*(\mathbf{u}_n^{\mathrm{CG}}) = \hat{\mathbf{D}}_*(\mathbf{u}_n^*) + (\epsilon_n^{\mathrm{CG}})^2 \frac{\mathbf{u}(\mathbf{u}_n^{\mathrm{CG}})^T\mathbf{A}}{\|\mathbf{u}_n^{\mathrm{CG}}\|_{\mathbf{A}}^2}.$$

Although the approximation $\mathbf{u}_n^*$ is a scalar multiple of $\mathbf{u}_n^{\mathrm{CG}}$, this is not the case for their corresponding perturbation/transformation matrices. On the other hand, their differences are rank one matrices depending only on the CG approximation $\mathbf{u}_n^{\mathrm{CG}}$ and its associated relative $\mathbf{A}$-norm of the error ϵ_n^{CG}. For the 2-norm, we obtain

$$\frac{\|\hat{\mathbf{E}}_*(\mathbf{u}_n^*) - \hat{\mathbf{E}}_*(\mathbf{u}_n^{\mathrm{CG}})\|_2}{\|\hat{\mathbf{E}}_*(\mathbf{u}_n^{\mathrm{CG}})\|_2} = (\epsilon_n^{\mathrm{CG}})^2 \frac{\|\mathbf{f}\|_2}{\|\mathbf{f} - \mathbf{A}\mathbf{u}_n^{\mathrm{CG}}\|_2}$$

and

$$\frac{\|\hat{\mathbf{D}}_*(\mathbf{u}_n^*) - \hat{\mathbf{D}}_*(\mathbf{u}_n^{\mathrm{CG}})\|_2}{\|\hat{\mathbf{D}}_*(\mathbf{u}_n^{\mathrm{CG}})\|_2} = (\epsilon_n^{\mathrm{CG}})^2 \frac{\|\mathbf{u}\|_2}{\|\mathbf{u} - \mathbf{u}_n^{\mathrm{CG}}\|_2}.$$

Thus we can expect more prominent differences in the coefficients of the transformation matrices when the relative norm of the error of the CG approximation $\mathbf{u}_n^{\mathrm{CG}}$ is considerably

ETNA
Kent State University
http://etna.math.kent.edu

smaller than its relative residual norm. In our example, we have $\|\mathbf{u} - \mathbf{u}_1^{\mathrm{CG}}\|_2/\|\mathbf{u}\|_2 = 0.7982$ and $\|\mathbf{f} - \mathbf{A}\mathbf{u}_1^{\mathrm{CG}}\|_2/\|\mathbf{f}\|_2 = 2.195$ for $n = 1$, and for $n = 5$, $\|\mathbf{u} - \mathbf{u}_5^{\mathrm{CG}}\|_2/\|\mathbf{u}\|_2 = 0.2998$ and $\|\mathbf{f} - \mathbf{A}\mathbf{u}_5^{\mathrm{CG}}\|_2/\|\mathbf{f}\|_2 = 2.102$. We can hence expect more prominent differences between the transformation matrices $\hat{\mathbf{D}}_*(\mathbf{u}_n^*)$ and $\hat{\mathbf{D}}_*(\mathbf{u}_n^{\mathrm{CG}})$ than between the perturbation matrices $\hat{\mathbf{E}}_*(\mathbf{u}_n^*)$ and $\hat{\mathbf{E}}_*(\mathbf{u}_n^{\mathrm{CG}})$, in particular at the iteration 5, where the relative residual norm is about 10 times larger than the relative error norm.

5. Conclusions. Motivated by the use of backward errors in stopping criteria for iterative solvers, we made an attempt to find an "easy-to-touch" interpretation of the data perturbations in linear algebraic systems arising from discretisations of elliptic partial differential equations. In particular, we were interested in finding a possible meaning of the perturbations of the system matrix $\mathbf{A}$ and related them to certain perturbations of the basis of the approximation space where the discrete solution of the underlying variational problem is sought. Although we are aware of the limited usability of our results in practice while bearing in mind recent results on dealing with discretisation and algebraic errors in numerical solution of PDEs, we believe that they might be of certain interest and motivate designers of stopping criteria for iterative processes to justify their relevance to the problem to be solved.

We showed that minimising the backward error induced by the $\mathbf{A}$-norm over the Krylov subspace $\mathcal{K}_n$ leads to approximations which are closely related to the approximations computed by CG, which minimise the $\mathbf{A}$-norm of the error over $\mathcal{K}_n$. This is similar to the idea behind the methods called GMBACK and MINPERT introduced in [14, 15] for general nonsymmetric problems. In contrast to the iterates computed by these methods, we showed that the optimal approximations minimising the backward error are just scalar multiples of the CG approximations and they are closer to each other as the $\mathbf{A}$-norm of the CG approximations decreases. Nevertheless, we do not claim that approximations constructed in this way have any superiority with respect to CG which is optimal itself with respect to the closely related measure.

Acknowledgments. We would like to thank to Zdeněk Strakoš and the anonymous referee for their comments and suggestions which considerably helped to improve the presentation of our results.

REFERENCES

[1] M. ARIOLI, *A stopping criterion for the conjugate gradient algorithm in a finite element method framework*, Numer. Math., 97 (2004), pp. 1–24.
[2] M. ARIOLI AND I. S. DUFF, *Using FGMRES to obtain backward stability in mixed precision*, Electron. Trans. Numer. Anal., 33 (2008/09), pp. 31–44. http://etna.math.kent.edu/vol.33.2008-2009/pp31-44.dir
[3] M. ARIOLI, I. S. DUFF, AND D. RUIZ, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 138–144.
[4] M. ARIOLI, J. LIESEN, A. MIEDLAR, AND Z. STRAKOŠ, *Interplay between discretization and algebraic computation in adaptive numerical solution of elliptic PDE problems*, to appear in *GAMM-Mitt.*, 2013.
[5] M. ARIOLI, E. NOULARD, AND A. RUSSO, *Stopping criteria for iterative methods: applications to PDE's*, Calcolo, 38 (2001), pp. 97–112.
[6] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, 3rd ed., Springer, New York, 2008.
[7] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[8] J. DRKOŠOVÁ, A. GREENBAUM, M. ROZLOŽNÍK, AND Z. STRAKOŠ, *Numerical stability of GMRES*, BIT, 35 (1995), pp. 309–330.
[9] H. C. ELMAN, D. J. SILVESTER, AND A. J. WATHEN, *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*, Oxford University Press, New York, 2005.
[10] W. GIVENS, *Numerical computation of the characteristic values of a real symmetric matrix*, Technical Report, ORNL 1574, Oak Ridge National Laboratory, Oak Ridge, 1957.

[11] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[12] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.
[13] P. JIRÁNEK AND M. ROZLOŽNÍK, *Adaptive version of simpler GMRES*, Numer. Algorithms, 53 (2010), pp. 93–112.
[14] E. M. KASENALLY, *GMBACK: a generalized minimum backward error algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 16 (1995), pp. 698–719.
[15] E. M. KASENALLY AND V. SIMONCINI, *Analysis of a minimum perturbation algorithm for nonsymmetric linear systems*, SIAM J. Numer. Anal., 34 (1997), pp. 48–66.
[16] P. D. LAX AND A. N. MILGRAM, *Parabolic equations*, in Contributions to the Theory of Partial Differential Equations, L. Bers, S. Bochner, and F. John, eds., Annals of Mathematics Studies, 33, Princeton University Press, Princeton, 1954, pp. 167–190.
[17] J. LIESEN AND Z. STRAKOŠ, *Krylov Subspace Methods: Principles and Analysis*, Oxford University Press, Oxford, 2012.
[18] W. OETTLI AND W. PRAGER, *Compatibility of approximate solution of linear equation with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409.
[19] C. C. PAIGE, M. ROZLOŽNÍK, AND Z. STRAKOŠ, *Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 264–284.
[20] J. PAPEŽ, J. LIESEN, AND Z. STRAKOŠ, *On distribution of the discretization and algebraic error in 1D Poisson model problem*, Preprint, MORE Report 2012/03, April 2013.
[21] J. L. RIGAL AND J. GACHES, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.
[22] W. RUDIN, *Functional Analysis*, 2nd. ed., McGraw-Hill, New York, 1991.
[23] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[24] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[25] Z. STRAKOŠ AND J. LIESEN, *On numerical stability in large scale linear algebraic computations*, Z. Angew. Math. Mech., 85 (2005), pp. 307–325.
[26] A. M. TURING, *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math., 1 (1948), pp. 287–308.
[27] C. VAN LOAN AND G. H. GOLUB, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.
[28] J. VON NEUMANN AND H. H. GOLDSTEIN, *Numerical inverting of matrices of high order*, Bull. Amer. Math. Soc., 53 (1947), pp. 1021–1099.
[29] H. F. WALKER, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 152–163.
[30] J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, 1963.
[31] ———, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.