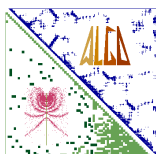


Energy backward error: interpretation in numerical solution of elliptic partial differential equations and convergence of the conjugate gradient method

SERGE GRATTON, PAVEL JIRÁNEK, AND XAVIER VASSEUR

Technical Report TR/PA/12/3



Publications of the Parallel Algorithms Team

<http://www.cerfacs.fr/algor/publications/>

Energy backward error: interpretation in numerical solution of elliptic partial differential equations and convergence of the conjugate gradient method

Serge Gratton* Pavel Jiránek† Xavier Vasseur‡

January 10, 2012

Abstract

We derive backward error formulas for a linear system of equations in norms induced by given symmetric positive definite matrices. We consider a special case of a backward error induced by the energy norm when the system matrix is symmetric positive definite and provide its interpretation in variational approximation of elliptic problems. Next, we study the convergence of the conjugate gradient method (CG) with respect to this energy backward error. For that purpose, we construct approximations to the solution, which minimize the energy backward error over the Krylov subspace generated by CG. We show that these approximations are scalar multiples of the CG approximations, approach the approximations of CG with the increasing iteration number, and start to be very close to each other as soon as CG makes a moderate progress in terms of the energy norm of the error.

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

AMS subject classifications. 65F10, 65F50

1 Introduction

We consider a system of linear algebraic equations

$$Ax = b \tag{1}$$

*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@cerfacs.fr).
The work of this author was supported by the ADTAO project funded by the foundation STAE, Toulouse, France, within RTRA.

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

where $A \in \mathbb{R}^{N \times N}$ is a symmetric and positive definite matrix and $b \in \mathbb{R}^N$ is a nonzero right-hand side vector. When one has to decide whether a given approximation \hat{x} to the solution x of (1) is satisfactory, a common approach is to compute the associated backward error [16, 26, 27] and compare it either with an uncertainty contained in the problem (1) at hand if any, or with a modest multiple of machine precision. Such a strategy is often used and recommended for stopping criteria for iterative methods [2, 3].

Let \hat{x} be an approximation of the solution of (1). The (normwise) backward error of \hat{x} is defined as a size of perturbations E and f of the data A and b , respectively, minimal in a suitable norm such that $(A + E)\hat{x} = b + f$. Rigal and Gaches [27] provided the explicit solution to the problem

$$\min\{\xi; (A + E)\hat{x} = b + f, \|E\| \leq \xi\|A\|, \|f\| \leq \xi\|b\|\} \quad (2)$$

given by the properly scaled Euclidean norm of the residual vector $b - A\hat{x}$ as

$$\frac{\|b - A\hat{x}\|}{\|A\|\|\hat{x}\| + \|b\|}. \quad (3)$$

In (2) and (3), any vector norm and its subordinate matrix norm can be used. The concept of the backward error originates from the rounding error analysis (see, e.g., [35, 17]) and in this context, the Euclidean norm $\|\cdot\|_2$ is widely used. However, when solving a linear system by an iterative method, it is appropriate to measure the backward error in proper spaces and using the proper norms; see [3]. In Section 2, we derive backward error formulas which use generalized norms characterized by given symmetric positive definite matrices. We consider an application in the Galerkin discretization of elliptic partial differential equations in Section 3 and provide two different, however very closely related, interpretations of the backward error induced by the energy norm generated by the symmetric positive matrix A in (1).

Modern iterative solvers are based on minimizing certain quantity over the Krylov subspace $\mathcal{K}_n(A, b) \equiv \text{span}\{b, Ab, \dots, A^{n-1}b\}$ [14, 28, 34]. It might be natural to minimize the backward error over the associated Krylov subspace, when it is used as the stopping criterion as well. This was pointed out by Kasenally and Simoncini in [20, 21], where the algorithms minimizing the Frobenius norm of the backward perturbation $[E, f]$ such that $(A + E)x = b + f$ were proposed to complement the algorithm of GMRES [29], which minimizes the Euclidean norm of the residual. In the next part, we are interested in minimizing the backward error induced by the energy norm which we refer to as the *energy backward error*. In Section 4, we review the Lanczos algorithm [22] and the conjugate gradient method [15] (CG) minimizing the energy norm of the error (or, equivalently, the dual A^{-1} -norm of the residual). The convergence of the CG approximations with respect to the energy backward error is monotonic, but not optimal in the sense of the energy backward error. We ask whether the CG approximation can be improved in the actual Krylov subspace. For this purpose, we derive in Section 5 the approximations, which at a given iteration step minimize the energy backward error in the Krylov subspace. In Section 6, we analyze the relations between the approximations minimizing the energy backward error and the approximations of CG and illustrate the results on a numerical experiment in Section 7.

2 The energy backward error

The matrix $A \in \mathbb{R}^{N \times N}$ can be considered as a mapping from the space $\mathcal{V} = \mathbb{R}^N$ to $\mathcal{W} = \mathbb{R}^N$. We equip \mathcal{V} and \mathcal{W} with the inner products induced by symmetric positive definite matrices K and L and the corresponding norms defined by $\|x\|_K \equiv \sqrt{x^T K x} = \|K^{1/2}x\|_2$ and $\|x\|_L \equiv \sqrt{x^T L x} = \|L^{1/2}x\|_2$, respectively. The matrix norm of A induced by the K - and L -norm is then given by

$$\|A\|_{K,L} = \max_{\substack{z \in \mathbb{R}^N \\ z \neq 0}} \frac{\|Az\|_L}{\|z\|_K}. \quad (4)$$

The norm $\|\cdot\|_{K,L}$ defined by (4) is related to the spectral matrix norm via

$$\begin{aligned} \|A\|_{K,L} &= \max_{\substack{z \in \mathbb{R}^N \\ z \neq 0}} \frac{\|Az\|_L}{\|z\|_K} = \max_{\substack{z \in \mathbb{R}^N \\ z \neq 0}} \frac{\|L^{1/2}Az\|_2}{\|K^{1/2}z\|_2} \\ &= \max_{\substack{y \in \mathbb{R}^N \\ y \neq 0}} \frac{\|L^{1/2}AK^{-1/2}y\|_2}{\|y\|_2} = \|L^{1/2}AK^{-1/2}\|_2. \end{aligned} \quad (5)$$

We define the backward error using the K - and L -norms by

$$\eta_{\alpha,\beta}^{K,L}(\hat{x}) \equiv \min_{E,f} \left\{ \left\| \left(\frac{\|E\|_{K,L}}{\alpha\|A\|_{K,L}}, \frac{\|f\|_L}{\beta\|b\|_L} \right) \right\|_2; (A+E)\hat{x} = b+f \right\}, \quad (6)$$

where α and β are positive weighting parameters. In the following theorem, we provide the solution to the backward error problem (6).

Theorem 2.1. *Let $A \in \mathbb{R}^{N \times N}$, $b \in \mathbb{R}^N$, $\hat{x} \in \mathbb{R}^N$ nonzero, and $r = b - A\hat{x}$. Let α and β be positive and let $K, L \in \mathbb{R}^{N \times N}$ be symmetric positive definite. Then $\eta_{\alpha,\beta}^{K,L}(\hat{x})$ defined by (6) is equal to*

$$\eta_{\alpha,\beta}^{K,L}(\hat{x}) = \frac{\|r\|_L}{\sqrt{\alpha^2\|A\|_{K,L}^2\|\hat{x}\|_K^2 + \beta^2\|b\|_L^2}}. \quad (7)$$

In addition,

$$\eta_{\alpha,0}^{K,L}(\hat{x}) \equiv \min_E \left\{ \frac{\|E\|_{K,L}}{\alpha\|A\|_{K,L}}; (A+E)\hat{x} = b \right\} = \frac{\|r\|_L}{\alpha\|A\|_{K,L}\|\hat{x}\|_K}, \quad \text{for } \alpha > 0,$$

and

$$\eta_{0,\beta}^{K,L}(\hat{x}) \equiv \min_f \left\{ \frac{\|f\|_L}{\beta\|b\|_L}; A\hat{x} = b+f \right\} = \frac{\|r\|_L}{\beta\|b\|_L}, \quad \text{for } \beta > 0.$$

Proof. Let α and β be positive. From the identity $(A + E)\hat{x} = b + f$ we have that $E\hat{x} - f = b - A\hat{x} = r$ and thus

$$L^{1/2}r = L^{1/2}EK^{-1/2}K^{1/2}\hat{x} - L^{1/2}f,$$

which implies

$$\|r\|_L \leq \|E\|_{K,L}\|\hat{x}\|_K + \|f\|_L.$$

From the inequality $(s+t)^2 \leq (1+p^2)s^2 + (1+p^{-2})t^2$ valid for any real s, t , and nonzero p , we get with $s = \|E\|_{K,L}\|\hat{x}\|_K$, $t = \|f\|_L$, and $p = (\beta\|b\|_L)/(\alpha\|A\|_{K,L}\|\hat{x}\|_K)$ that

$$\begin{aligned} \|r\|_L^2 &\leq (\|E\|_{K,L}\|\hat{x}\|_K + \|f\|_L)^2 \\ &\leq \|E\|_{K,L}^2\|\hat{x}\|_K^2 \left[1 + \left(\frac{\beta\|b\|_L}{\alpha\|A\|_{K,L}\|\hat{x}\|_K} \right)^2 \right] + \|f\|_L^2 \left[1 + \left(\frac{\alpha\|A\|_{K,L}\|\hat{x}\|_K}{\beta\|b\|_L} \right)^2 \right] \\ &= \left(\frac{\|E\|_{K,L}^2}{\alpha^2\|A\|_{K,L}^2} + \frac{\|f\|_L^2}{\beta^2\|b\|_L^2} \right) (\alpha^2\|A\|_{K,L}^2\|\hat{x}\|_K^2 + \beta^2\|b\|_L^2), \end{aligned}$$

which proves

$$\frac{\|r\|_L}{\sqrt{\alpha^2\|A\|_{K,L}^2\|\hat{x}\|_K^2 + \beta^2\|b\|_L^2}} \leq \left\| \left(\frac{\|E\|_{K,L}}{\alpha\|A\|_{K,L}}, \frac{\|f\|_L}{\beta\|b\|_L} \right) \right\|_2. \quad (8)$$

By taking

$$E = \frac{\alpha^2\|A\|_{K,L}^2}{\alpha^2\|A\|_{K,L}^2\|\hat{x}\|_K^2 + \beta^2\|b\|_L^2} r \hat{x}^T K, \quad f = -\frac{\beta^2\|b\|_L^2}{\alpha^2\|A\|_{K,L}^2\|\hat{x}\|_K^2 + \beta^2\|b\|_L^2} r \quad (9)$$

we get $E\hat{x} - f = b - A\hat{x}$ as well as the equality in (8). The formula for $\eta_{\alpha,0}^{K,L}$ can be verified in the similar way while the formula for $\eta_{0,\beta}^{K,L}$ is trivial since there is only one vector f satisfying $A\hat{x} = b + f$. \square

For the sake of simplicity of notation, we introduce a single weighting in the definition of the backward error instead of two. Let $\alpha > 0$ and $\beta \geq 0$ be the weights in (6) and let $\theta \equiv \beta/\alpha$. Then

$$\eta_{\alpha,\beta}^{K,L}(\hat{x}) = \frac{\|r\|_L}{\sqrt{\alpha^2\|A\|_{K,L}^2\|\hat{x}\|_K^2 + \beta^2\|b\|_L^2}} = \frac{1}{\alpha} \frac{\|r\|_L}{\sqrt{\|A\|_{K,L}^2\|\hat{x}\|_K^2 + \theta^2\|b\|_L^2}} \equiv \frac{1}{\alpha} \eta_{\theta}^{K,L}(\hat{x}).$$

In the following text, we work with $\eta_{\theta}^{K,L}$ instead of $\eta_{\alpha,\beta}^{K,L}$.

A particularly important case is when $K = A$ (the A -norm) and $L = A^{-1}$ (the dual norm with respect to the A -norm) for a symmetric positive definite A . We call $\xi_{\theta}(\hat{x}) \equiv \eta_{\theta}^{A,A^{-1}}(\hat{x})$ the *energy backward error* associated with the approximation \hat{x} . Using the relations $\|A\|_{A,A^{-1}} = \|A^{-1/2}AA^{-1/2}\|_2 = \|I\|_2 = 1$, $\|r\|_{A^{-1}} = \|e\|_A$ and $\|b\|_{A^{-1}} = \|x\|_A$, where $e = x - \hat{x}$ denotes the error vector satisfying $Ae = r$, we have the following theorem.

Theorem 2.2. *Let $A \in \mathbb{R}^{N \times N}$ be symmetric positive definite, $b \in \mathbb{R}^N$, $\hat{x} \in \mathbb{R}^N$ nonzero, $r = b - A\hat{x}$, $e = x - \hat{x}$, and $\theta \geq 0$. Then*

$$\xi_\theta(\hat{x}) \equiv \eta_\theta^{A, A^{-1}}(\hat{x}) = \frac{\|r\|_{A^{-1}}}{\sqrt{\|\hat{x}\|_A^2 + \theta^2 \|b\|_{A^{-1}}^2}} = \frac{\|e\|_A}{\sqrt{\|\hat{x}\|_A^2 + \theta^2 \|x\|_A^2}}. \quad (10)$$

3 Energy backward error in numerical solution of elliptic problems

In this section, we provide a simple interpretation of the energy backward error in the context of numerical solution of partial differential equation, in particular the Galerkin approximation of self-adjoint elliptic problems. We consider an abstract variational problem on a real Hilbert space \mathcal{V} : find $u \in \mathcal{V}$ such that

$$a(u, v) = \langle g, v \rangle \quad \forall v \in \mathcal{V}, \quad (11)$$

where a is a continuous, symmetric, and elliptic bilinear form, $g \in \mathcal{V}'$ is a continuous linear functional on \mathcal{V} (their space is denoted by \mathcal{V}') and $\langle \cdot, \cdot \rangle$ denotes the duality pairing between \mathcal{V} and \mathcal{V}' . For details, see, e.g., [10]. The properties of the bilinear form a imply that a is an inner product on \mathcal{V} and thus defines a norm $\|v\|_a \equiv [a(v, v)]^{1/2}$ for $v \in \mathcal{V}$, called usually the energy norm.

Let \mathcal{V}_h be a subspace of \mathcal{V} of finite dimension N . The Galerkin method for approximating the solution u of (11) reads: find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v_h) = \langle g, v_h \rangle \quad \forall v_h \in \mathcal{V}_h. \quad (12)$$

The Galerkin solution u_h minimizes the energy norm of the error $u - u_h$ over all \mathcal{V}_h , i.e.,

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

or, equivalently, the error $u - u_h$ is orthogonal (in the inner product defined by the form a) to the subspace \mathcal{V}_h .

In order to transform the discrete problem (12) to the form of a system of equations, we chose a basis $\Phi \equiv [\phi_1, \dots, \phi_N]$ of \mathcal{V}_h , so that we can express the solution u_h in terms of the basis Φ as $u_h = \Phi x$ for some vector $x \in \mathbb{R}^N$ representing the coordinates of u_h in the basis Φ (we defined the basis Φ as a row vector of ϕ_1, \dots, ϕ_N so that we can use the notation $u_h = \Phi x$ as a shorthand for $\sum_{j=1}^N (x)_j \phi_j$). The equality (12) is satisfied if and only if $a(u_h, \phi_i) = \langle g, \phi_i \rangle$ for $i = 1, \dots, N$, which leads to a system of equations (1) with

$$(A)_{i,j} = a(\phi_j, \phi_i), \quad (b)_i = \langle g, \phi_i \rangle, \quad i, j = 1, \dots, N, \quad (13)$$

and A is symmetric and positive definite since it is the Gram matrix associated with the basis Φ and the inner product $a(\cdot, \cdot)$.

The matrix A represents the bilinear form a on \mathcal{V}_h in the chosen basis Φ . Indeed, for v_h and w_h in \mathcal{V}_h with coordinate vector y and z in \mathbb{R}^N (with respect to the basis Φ), the bilinear form a satisfies

$$a(v_h, w_h) = a(\Phi y, \Phi z) = z^T A y$$

and the energy norm of a $v_h = \Phi y \in \mathcal{V}_h$ can be expressed as $\|v_h\|_a = (y^T A y)^{1/2} = \|y\|_A$.

Let \hat{x} be an approximation of the solution x of (1) with A and b given by (13) and let $r = b - A\hat{x}$ and $e = x - \hat{x} = A^{-1}r$ denote, respectively, the residual and error vectors associated with \hat{x} . In this section, we focus on the interpretation of the quantities

$$\xi_b(\hat{x}) \equiv \min_f \{ \|f\|_{A^{-1}}; A\hat{x} = b + f \} = \|r\|_{A^{-1}} = \|e\|_A, \quad (14a)$$

$$\xi_A(\hat{x}) \equiv \min_E \{ \|E\|_{A, A^{-1}}; (A + E)\hat{x} = b \} = \frac{\|r\|_A^{-1}}{\|\hat{x}\|_A} = \frac{\|e\|_A}{\|\hat{x}\|_A}, \quad (14b)$$

see Theorem 2.1, and on the interpretation of the optimal backward perturbations (see (9) in the proof of Theorem 2.1)

$$f_* \equiv -r \quad \text{and} \quad E_* \equiv \frac{r\hat{x}^T A}{\|\hat{x}\|_A^2}, \quad (15)$$

for which the minima in (14a) and (14b), respectively, are attained.

3.1 Interpretation of $\xi_b(\hat{x})$

Let $\hat{u}_h = \Phi\hat{x}$ denote the approximation of u_h ; i.e., let the approximation \hat{x} of x represent the coordinate vector of \hat{u}_h in the basis Φ . We can define a functional $r_h \in \mathcal{V}'_h$ by $\langle r_h, \phi_i \rangle = (r)_i$ for $i = 1, \dots, N$. It can be easily verified that the perturbed equation $A\hat{x} = b + f_* = b - r$ represents the variational equality

$$a(\hat{u}_h, v_h) = \langle g - r_h, v_h \rangle \quad \forall v_h \in \mathcal{V}_h$$

with $\xi_b(\hat{x})$ being the dual norm of r_h which turns out to be the energy norm of the error $u_h - \hat{u}_h$:

$$\|r_h\|_a = \max_{\substack{v_h \in \mathcal{V}_h \\ v_h \neq 0}} \frac{\langle r_h, v_h \rangle}{\|v_h\|_a} = \max_{\substack{y \in \mathbb{R}^N \\ y \neq 0}} \frac{r^T y}{\|y\|_A} = \|r\|_A^{-1} = \|e\|_A = \|u_h - \hat{u}_h\|_a.$$

Evaluating or estimating the quantity $\xi_b(\hat{x})$ is used in the construction of stopping criteria taking into account the discretization error in the context of numerical solution of partial differential equations; see, e.g., [1, 19]. Indeed, it represents the part of the total error $u - \hat{u}_h$ stemming from the inexact solution of the discrete problem (12):

$$\|u - \hat{u}_h\|_a^2 = \|u - u_h\|_a^2 + \|u_h - \hat{u}_h\|_a^2 = \|u - u_h\|_a^2 + \xi_b^2(\hat{x});$$

see, e.g., [11, Section 2.1].

3.2 Interpretation of $\xi_A(\hat{x})$

The vector \hat{x} satisfying $(A + E)\hat{x} = b$ can be naturally interpreted as a solution of the system with a perturbed matrix (or operator) A . We can define a (generally nonsymmetric) bilinear form e_h on \mathcal{V}_h similarly to the definition of the matrix A in (13) by $e_h(\phi_j, \phi_i) = (E)_{i,j}$, $i, j = 1, \dots, N$, and consider $\hat{u}_h = \Phi\hat{x}$ as the solution of a perturbed Galerkin approximation

$$a(\hat{u}_h, v_h) + e_h(\hat{u}_h, v_h) = \langle g, v_h \rangle \quad \forall v_h \in \mathcal{V}_h \quad (16)$$

with the norm of e_h given by

$$\|e_h\|_a = \max_{v_h, w_h \in \mathcal{V}_h \setminus \{0\}} \frac{e_h(v_h, w_h)}{\|v_h\|_a \|w_h\|_a} = \|E\|_{A, A^{-1}}.$$

The quantity $\xi_A(\hat{x})$ is therefore the solution to the problem

$$\min_{e_h \in \mathcal{B}(\mathcal{V}_h)} \{ \|e_h\|_a; a(\Phi\hat{x}, v_h) + e_h(\Phi\hat{x}, v_h) = \langle g, v_h \rangle \quad \forall v_h \in \mathcal{V}_h \},$$

where $\mathcal{B}(\mathcal{V}_h)$ is the space of bilinear forms defined on \mathcal{V}_h .

Since from $(A + E)\hat{x} = b$ we have $x = (I + A^{-1}E)\hat{x}$, we can also interpret E as a perturbation of a basis in which the vector \hat{x} represents coordinates of the solution x of (1). In order to interpret $\xi_A(\hat{x})$ in the context of the Galerkin approximation, we can proceed in the following way. Let $\hat{\Phi} \equiv [\hat{\phi}_1, \dots, \hat{\phi}_N]$ denote another basis of \mathcal{V}_h and let $\hat{\Phi}$ and Φ be related by

$$\hat{\phi}_j = \phi_j + \sum_{k=1}^N (D)_{k,j} \phi_k, \quad j = 1, \dots, N, \quad (17)$$

for some matrix $D \in \mathbb{R}^{N \times N}$; the relations (17) can be also written in a compact form as $\hat{\Phi} = \Phi(I + D)$. We look for a basis $\hat{\Phi}$ such that u_h can be written as $u_h = \hat{\Phi}\hat{x}$, i.e., the approximate solution \hat{x} of (1) represents the coordinates of the solution of (12) in a different basis.

Let $u_h = \hat{\Phi}\hat{x}$, where $\hat{\Phi}$ is related to Φ by (17). We have

$$a(u_h, \phi_i) = a(\hat{\Phi}\hat{x}, \phi_i) = a(\Phi(I + D)\hat{x}, \phi_i) = (A(I + D)\hat{x})_i$$

and thus the equations $a(u_h, \phi_i) = \langle g, \phi_i \rangle$, $i = 1, \dots, N$, lead to the system

$$(A + E)\hat{x} = b, \quad E = AD. \quad (18)$$

If \hat{x} solves (18) then \hat{x} forms the coordinate vector of u_h in terms of the basis $\hat{\Phi} = \Phi(I + D)$ with $D = A^{-1}E$.

For a given \hat{x} , there are many backward perturbations E for which $(A + E)\hat{x} = b$ holds. If E' is any particular backward perturbation, then $(A + E)\hat{x} = b$ for any $E = E' + ZP$, where $Z \in \mathbb{R}^{N \times N}$ and P is an orthogonal projection such that $Px = 0$; see, e.g., [9].

Equivalently, there are many bases $\hat{\Phi}$ for which \hat{x} represents the coordinate vector of u_h . Let us define

$$d_{\Phi}(\hat{\Phi}) \equiv \max_{\substack{y \in \mathbb{R}^N \\ y \neq 0}} \frac{\|\Phi y - \hat{\Phi} y\|_a}{\|\Phi y\|_a}$$

the relative distance between bases Φ and $\hat{\Phi}$ with respect to Φ measured in the energy norm. Then using (17) and $E = AD$ we get

$$\begin{aligned} d_{\Phi}(\hat{\Phi}) &= \max_{\substack{y \in \mathbb{R}^N \\ y \neq 0}} \frac{\|\Phi y - \hat{\Phi} y\|_a}{\|\Phi y\|_a} = \max_{\substack{y \in \mathbb{R}^N \\ y \neq 0}} \frac{\|\Phi y - \Phi(I + D)y\|_a}{\|\Phi y\|_a} = \max_{\substack{y \in \mathbb{R}^N \\ y \neq 0}} \frac{\|\Phi D y\|_a}{\|\Phi y\|_a} \\ &= \max_{\substack{y \in \mathbb{R}^N \\ y \neq 0}} \frac{\|D y\|_A}{\|y\|_A} = \|A^{1/2} D A^{-1/2}\|_2 = \|A^{-1/2} E A^{-1/2}\|_2 = \|E\|_{A, A^{-1}}. \end{aligned}$$

The quantity $\xi_A(\hat{x})$ thus represents the relative distance to the basis $\hat{\Phi}_*$ closest to Φ (in terms of the relative distance d_{Φ}) such that \hat{x} is the coordinate vector of u_h in $\hat{\Phi}_*$. Using (17), (15), and (14b), we have

$$\hat{\Phi}_* = \Phi(I + D_*), \quad D_* = A^{-1} E_* = \frac{e \hat{x}^T A}{\|\hat{x}\|_A^2}, \quad d_{\Phi}(\hat{\Phi}_*) = \frac{\|e\|_A}{\|\hat{x}\|_A}.$$

Therefore, $\xi_A(\hat{x})$ is also the solution of the problem

$$\min_{\substack{\hat{\Phi} = \Phi(I + D) \\ D \in \mathbb{R}^{N \times N}}} \left\{ d_{\Phi}(\hat{\Phi}); a(\hat{\Phi} \hat{x}, v_h) = \langle g, v_h \rangle \quad \forall v_h \in \mathcal{V}_h \right\}.$$

Note that not every $\hat{\Phi} = \Phi(I + D)$, where Φ is a basis of \mathcal{V}_h , is a basis of \mathcal{V}_h as well. The elements of $\hat{\Phi}$ is linearly dependent if and only if the matrix $I + D$ is singular. A sufficient condition for the nonsingularity of $I + D$ is that $d_{\Phi}(\hat{\Phi}) = \|A^{1/2} D A^{-1/2}\|_2 < 1$. Assume that $I + D$ is singular. Then there is a nonzero vector z such that $(I + D)z = 0$, which implies that $A^{1/2} z = -(A^{1/2} D A^{-1/2})(A^{1/2} z)$, $\|A^{1/2} z\|_2 \leq \|A^{1/2} D A^{-1/2}\|_2 \|A^{1/2} z\|_2$, and thus $d_{\Phi}(\hat{\Phi}) = \|A^{1/2} D A^{-1/2}\|_2 \geq 1$. It corresponds to the well-known fact that the perturbation E of a nonsingular matrix A leads to a nonsingular matrix $A + E$ if the relative norm of E is smaller than the inverse of the condition number of A (with respect to any norm), i.e.,

$$\frac{\|E\|_{A, A^{-1}}}{\|A\|_{A, A^{-1}}} < \frac{1}{\kappa_{A, A^{-1}}(A)} = \frac{\|A^{-1}\|_{A^{-1}, A}}{\|A\|_{A, A^{-1}}},$$

which is equivalent to $\|E\|_{A, A^{-1}} < 1$ since $\|A\|_{A, A^{-1}} = \|A^{-1}\|_{A^{-1}, A} = \kappa_{A, A^{-1}}(A) = 1$; see, e.g., [30, Corollary 2.7].

4 Lanczos algorithm and the conjugate gradient method

In the next part, starting from this section, we analyze the convergence of the conjugate gradient method in terms of the energy backward error. We start with a short review of the CG method in terms of the Lanczos process [22]. Consider the system (1) with a symmetric positive definite matrix A . The Lanczos algorithm applied to A with the starting vector $v_1 = b/\|b\|_2$ gives at the n th step

$$AV_n = V_n T_n + \delta^{(n+1)} v_{n+1} (e^n)^T, \quad (19)$$

where the columns of the matrix $V_n = [v_1, \dots, v_n]$ form an orthonormal basis of the Krylov subspace $\mathcal{K}_n(A, b)$, e^n denotes the n th vector of the standard basis of \mathbb{R}^n , and the matrix T_n is tridiagonal and symmetric positive definite.

The n th CG [15] approximation x_n^{CG} (starting from the zero initial guess) is defined by

$$x_n^{\text{CG}} = V_n y_n^{\text{CG}}, \quad T_n y_n^{\text{CG}} = \|b\|_2 e^1. \quad (20)$$

It is well-known that the CG approximation minimizes the energy norm (the A -norm) of the error $e_n = x - x_n$, $x_n = V_n y_n \in \mathcal{K}_n(A, b)$, i.e.,

$$\|e_n^{\text{CG}}\|_A = \|x - x_n^{\text{CG}}\|_A = \min_{x_n \in \mathcal{K}_n(A, b)} \|x - x_n\|_A$$

and that $\|e_n^{\text{CG}}\|_A$ decreases at each iteration step. This can be also be observed from the identity

$$\begin{aligned} \|e_n\|_A^2 &= e_n^T A e_n = x^T A x - 2x^T A V_n y_n + y_n^T T_n y_n \\ &= \|x\|_A^2 - \|b\|_2^2 (e^1)^T T_n^{-1} e^1 + \|\|b\|_2 e^1 - T_n y_n\|_{T_n^{-1}}^2. \end{aligned} \quad (21)$$

The A -norm of $e_n = e_n^{\text{CG}}$ is minimized if and only if the last term in (21) is zero, i.e., $y_n = y_n^{\text{CG}}$ defined in (20) giving then, using

$$\|b\|_2^2 (e^1)^T T_n^{-1} e^1 = \|y_n^{\text{CG}}\|_{T_n}^2 = \|x_n^{\text{CG}}\|_A^2, \quad (22)$$

the relation between of the A -norm of the error e_n^{CG} and the A -norms of the solution x and the iteration x_n of the form

$$\|e_n^{\text{CG}}\|_A^2 = \|e_0^{\text{CG}}\|_A^2 - \|b\|_2^2 (e^1)^T T_n^{-1} e^1 = \|x\|_A^2 - \|x_n^{\text{CG}}\|_A^2. \quad (23)$$

The A -norms of the error e_n^{CG} decrease monotonically towards zero with increasing n , the A -norms of the CG iterate x_n^{CG} strictly increase towards the A -norm of the solution x of (1) and we have the following theorem.

Theorem 4.1. *In the conjugate gradient method starting with $x_0^{\text{CG}} = 0$, the energy backward error $\xi_\theta(x_n^{\text{CG}})$ is strictly decreasing with n .*

5 Minimizing the energy backward error

In this section, we construct the approximations to the solution of (1), which minimize the energy backward error (10) over the Krylov subspace $\mathcal{K}_n(A, b)$. We look for $x_n^\theta \in \mathcal{K}_n(A, b)$ such that

$$\xi_\theta(x_n^\theta) = \min_{x_n \in \mathcal{K}_n(A, b)} \xi_\theta(x_n).$$

Let $x_n = V_n y_n$ be an arbitrary vector from $\mathcal{K}_n(A, b)$. For the A -norm of the error $e_n = x - x_n$ we get from (21), (22), and (23) the relation

$$\begin{aligned} \|e_n\|_A^2 &= \|x\|_A^2 - \|x_n^{\text{CG}}\|_A^2 + \|\|b\|_2 e^1 - T_n y_n\|_{T_n^{-1}}^2 \\ &= \|e_n^{\text{CG}}\|_A^2 + \|\|b\|_2 e^1 - T_n y_n\|_{T_n^{-1}}^2. \end{aligned} \quad (24)$$

For the A -norm of x_n we have

$$\|x_n\|_A^2 = y_n^T V_n^T A V_n y_n = \|y_n\|_{T_n}^2 \quad (25)$$

using (19). Therefore from (10), (24), (25), and (20) we obtain

$$\begin{aligned} \xi_\theta(x_n) &= \frac{\|e_n\|_A}{\sqrt{\|x_n\|_A^2 + \theta^2 \|x\|_A^2}} = \left(\frac{\|e_n^{\text{CG}}\|_A^2 + \|\|b\|_2 e^1 - T_n y_n\|_{T_n^{-1}}^2}{\|y_n\|_{T_n}^2 + \theta^2 \|x\|_A^2} \right)^{1/2} \\ &= \left(\frac{\|e_n^{\text{CG}}\|_A^2 + \|T_n(y_n^{\text{CG}} - y_n)\|_{T_n^{-1}}^2}{\|y_n\|_{T_n}^2 + \theta^2 \|x\|_A^2} \right)^{1/2} = \frac{\|L_n z_n\|_2}{\|G_n^\theta z_n\|_2}, \end{aligned} \quad (26)$$

where the matrices L_n , G_n^θ , and the vector z_n are given by

$$L_n \equiv \begin{bmatrix} \|e_n^{\text{CG}}\|_A & 0 \\ T_n^{1/2} y_n^{\text{CG}} & T_n^{1/2} \end{bmatrix}, \quad G_n^\theta \equiv \begin{bmatrix} \theta \|x\|_A & 0 \\ 0 & T_n^{1/2} \end{bmatrix}, \quad z_n \equiv \begin{bmatrix} 1 \\ -y_n \end{bmatrix}. \quad (27)$$

The approximation $x_n^\theta = V_n y_n^\theta$ can be found by minimizing the Rayleigh quotient in (26) over all z_n of the form (27). The matrix L_n is nonsingular provided that $x_n^{\text{CG}} \neq x$. Otherwise, we set $x_n^\theta = x_n^{\text{CG}} = x$ to obtain $\xi_\theta(x_n^\theta) = 0$.

Lemma 5.1. *Let the matrices L_n and G_n for $n > 1$ in (26) be given by (27), let $\theta \geq 0$, and $e_n^{\text{CG}} \neq 0$. Then there is a unique vector z_n^θ which minimizes the Rayleigh quotient $\|L_n z_n\|_2 / \|G_n^\theta z_n\|_2$ over nonzero z_n and has the first component equal to one and*

$$\frac{\|L_n z_n^\theta\|_2}{\|G_n^\theta z_n^\theta\|_2} = \min_{z_n \neq 0} \frac{\|L_n z_n\|_2}{\|G_n^\theta z_n\|_2} = \left[\frac{2}{1 + \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2 \epsilon_n^2}} \right]^{1/2} \epsilon_n, \quad (28)$$

where

$$z_n^\theta = \begin{bmatrix} 1 \\ -y_n^\theta \end{bmatrix}, \quad y_n^\theta = \delta_n^\theta y_n^{\text{CG}}, \quad \delta_n^\theta \equiv \frac{1 - \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2 \epsilon_n^2}}{2(1 - \epsilon_n^2)},$$

and $\epsilon_n \equiv \|e_n^{\text{CG}}\|_A / \|x\|_A$ is the relative error of the CG approximation x_n^{CG} .

Proof. Since $e_n^{\text{CG}} \neq 0$, the matrix L_n is nonsingular and we have

$$\begin{aligned} \frac{\|L_n z_n^\theta\|_2}{\|G_n^\theta z_n^\theta\|_2} &= \min_{z_n \neq 0} \frac{\|L_n z_n\|_2}{\|G_n^\theta z_n\|_2} \\ &= \left(\max_{w_n \neq 0} \frac{\|G_n^\theta L_n^{-1} w_n\|_2}{\|w_n\|_2} \right)^{-1} = \left(\frac{\|G_n^\theta L_n^{-1} w_n^\theta\|_2}{\|w_n^\theta\|_2} \right)^{-1}. \end{aligned} \quad (29)$$

If $w_n^\theta = [1, (\bar{w}_n^\theta)^T]^T$ in (29) then

$$z_n^\theta = \begin{bmatrix} 1 \\ -y_n^\theta \end{bmatrix} = \|e_n^{\text{CG}}\|_A L_n^{-1} w_n^\theta = \begin{bmatrix} 1 \\ -y_n^{\text{CG}} + \|e_n^{\text{CG}}\|_A T_n^{-1/2} \bar{w}_n^\theta \end{bmatrix} \quad (30)$$

(note that the vectors z_n and w_n in (29) are defined up to a scalar multiple; we choose to relate them by $L_n z_n = \|e_n^{\text{CG}}\|_A w_n$ in order to normalize the first component of z_n^θ to 1). Using the Courant-Fisher theorem [18, Theorem 4.2.11, p. 179] the maximum of $\|G_n^\theta L_n^{-1} w_n\|_2 / \|w_n\|_2$ in (29) is given by the largest singular value of $G_n^\theta L_n^{-1}$, or equivalently, by the square root of the largest eigenvalue of $M_n^\theta \equiv (G_n^\theta L_n^{-1})^T G_n^\theta L_n^{-1}$, where the vector w_n^θ , for which the maximum is attained, is equal to the corresponding (properly scaled) eigenvector. We have also to show that this eigenvector has nonzero first component. The matrix M_n^θ can be written as

$$\begin{aligned} M_n^\theta &= \begin{bmatrix} \|e_n^{\text{CG}}\|_A^{-2} (\theta^2 \|x\|_A^2 + \|x_n^{\text{CG}}\|_A^2) & -\|e_n^{\text{CG}}\|_A^{-1} (T_n^{-1/2} y_n^{\text{CG}})^T \\ -\|e_n^{\text{CG}}\|_A^{-1} T_n^{-1/2} y_n^{\text{CG}} & I_n \end{bmatrix} \\ &= \begin{bmatrix} \frac{1+\theta^2-\epsilon_n^2}{\epsilon_n^2} & -h_n^T \\ -h_n & I_n \end{bmatrix}. \end{aligned} \quad (31)$$

using (27), (22), and (23), where $h_n = \|e_n^{\text{CG}}\|_A^{-1} T_n^{1/2} y_n^{\text{CG}}$, $\|h_n\|_2^2 = (1 - \epsilon_n^2) / \epsilon_n^2$. The matrix M_n^θ has the eigenvalue 1 of the multiplicity $n - 1$ because there are $n - 1$ linearly independent eigenvectors w of the form $[0, \bar{w}^T]^T$ such that $\bar{w}^T h_n = 0$. Let $w = [1, \gamma h_n^T]^T$ where γ is a (real) scalar. Then

$$M_n^\theta \begin{bmatrix} 1 \\ \gamma h_n \end{bmatrix} = \mu \begin{bmatrix} 1 \\ \gamma h_n \end{bmatrix}$$

holds if and only if

$$1 + \theta^2 - \epsilon_n^2 - \gamma(1 - \epsilon_n^2) = \mu \epsilon_n^2, \quad \gamma(1 - \mu) = 1. \quad (32)$$

Substituting the first equation into the second one leads to the quadratic equation for μ in the form

$$\epsilon_n^2 \mu^2 - (1 + \theta^2) \mu + \theta^2 = 0. \quad (33)$$

The (real) roots of (33) are

$$\mu_{\pm} = \frac{1}{2\epsilon_n^2} \left[1 + \theta^2 \pm \sqrt{(1 + \theta^2)^2 - 4\theta^2 \epsilon_n^2} \right].$$

It can be verified that the root μ_+ of (33) is always greater than 1. Indeed, $\mu_+ > 1$ if and only if

$$\sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2} > 2\epsilon_n^2 - (1 + \theta^2). \quad (34)$$

This inequality is trivially satisfied if the right-hand side is negative. On the other hand, if the right-hand side in (34) is nonnegative, we obtain by squaring both sides of (34) that $\mu_+ > 1$ if $\epsilon_n^2 < 1 + \theta^2 + \theta^4$. This is clearly satisfied since $\epsilon_n^2 < 1$. Hence the largest eigenvalue of the matrix M_n in (31) is

$$\frac{1}{2\epsilon_n^2} \left[1 + \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2} \right] \quad (35)$$

and the minimum in (29) (as well as in (28)) is equal to the square root of the reciprocal value of (35). The scalar γ is from (32) given by

$$\gamma = 1 - \frac{1 - \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2}}{2(1 - \epsilon_n^2)}. \quad (36)$$

By comparing the eigenvector $[1, \gamma h_n^T]^T$ with w_n^θ maximizing (29), we get

$$\bar{w}_n^\theta = \gamma h_n = \gamma \|e_n^{\text{CG}}\|_A^{-1} T_n^{1/2} y_n^{\text{CG}}$$

and from (30) and (36), we obtain

$$\begin{aligned} y_n^\theta &= y_n^{\text{CG}} - \|e_n^{\text{CG}}\|_A T_n^{-1/2} \bar{w}_n^\theta = (1 - \gamma) y_n^{\text{CG}} \\ &= \frac{1 - \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2}}{2(1 - \epsilon_n^2)} y_n^{\text{CG}} = \delta_n^\theta y_n^{\text{CG}}. \quad \square \end{aligned}$$

Theorem 5.2. *Let $x_n^{\text{CG}} \neq 0$ denote the approximation of CG starting with the initial guess $x_0^{\text{CG}} = 0$ at the step $n > 1$ and let $\theta \geq 0$. Then the unique x_n^θ minimizing the backward error $\xi_\theta(x_n)$ over all $x_n \in \mathcal{K}_n(A, b)$ is given by*

$$x_n^\theta = \delta_n^\theta x_n^{\text{CG}}, \quad (37)$$

where

$$\delta_n^\theta \equiv \frac{1 - \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2}}{2(1 - \epsilon_n^2)}, \quad \epsilon_n = \frac{\|e_n^{\text{CG}}\|_A}{\|x\|_A}. \quad (38)$$

The energy backward error of the approximation x_n^θ is

$$\xi_\theta(x_n^\theta) = \left[\frac{2}{1 + \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2}} \right]^{1/2} \epsilon_n. \quad (39)$$

The approximation minimizing the energy backward error (10) over the Krylov subspace $\mathcal{K}_n(A, b)$ is thus given by a scalar multiple of the CG approximation x_n^{CG} with the coefficient given by (38). Note that even though we derived Lemma 5.1 with the assumption that $e_n^{\text{CG}} \neq 0$, the formulas for the δ_n^θ and $\xi_\theta(x_n^\theta)$ in Theorem 5.2 are still

valid for $e_n^{\text{CG}} = 0$ (i.e., $\epsilon_n = 0$). We obtain in this case from (38) and (39) that $\delta_n^\theta = 1$ leading to $x_n^\theta = x_n^{\text{CG}} = x$ and $\xi_\theta(x_n^\theta) = 0$, respectively.

The approximation x_n^θ and its associated backward error ξ_θ depend on the choice of the weighting parameter θ . We obtain the CG approximation x_n^{CG} by taking the limit $\theta \rightarrow \infty$, which corresponds to restricting the backward perturbation in (10) only to the right-hand side; indeed,

$$x_n^\infty = \lim_{\theta \rightarrow \infty} x_n^\theta = \lim_{\theta \rightarrow \infty} \delta_n^\theta x_n^{\text{CG}} = x_n^{\text{CG}}.$$

Perturbing only the matrix A in (10) is made by the choice $\theta = 0$. In this case,

$$\delta_n^0 = \frac{1}{1 - \epsilon_n^2}, \quad x_n^0 = \frac{x_n^{\text{CG}}}{1 - \epsilon_n^2}$$

and the energy backward error of the approximation x_n^0 is

$$\xi_0(x_n^0) = \epsilon_n = \frac{\|e_n^{\text{CG}}\|_A}{\|x\|_A},$$

i.e., it is equal to the relative energy norm of the error of the CG approximation.

6 Relations between CG and approximations minimizing the energy backward error

In this section, we study the relations between the approximations x_n^θ minimizing the energy backward error and the CG approximations x_n^{CG} minimizing the energy norm of the error over the Krylov subspace $\mathcal{K}_n(A, b)$. Theorem 5.2 shows that the approximation x_n^θ is a scalar multiple of the CG approximation x_n^{CG} with the coefficient given by (38). For the ratio of (any) norms of x_n^θ and x_n^{CG} , we have the following result.

Theorem 6.1. *Norms of the approximations x_n^θ and of the CG approximations x_n^{CG} satisfy*

$$1 \leq \frac{\|x_n^\theta\|}{\|x_n^{\text{CG}}\|} = \delta_n^\theta \leq \frac{1}{1 - \epsilon_n^2}. \quad (40)$$

The coefficient δ_n^θ (for a fixed θ) decreases monotonically with increasing n to 1.

Proof. For any vector norm $\|\cdot\|$, we have from Theorem 5.2 and the positive homogeneity of vector norms that

$$\|x_n^\theta\| = \|\delta_n^\theta x_n^{\text{CG}}\| = |\delta_n^\theta| \|x_n^{\text{CG}}\|$$

and thus $\|x_n^\theta\|/\|x_n^{\text{CG}}\| = |\delta_n^\theta|$. The coefficient δ_n^θ given by (38) can be bounded from below by 1 and thus we can omit the absolute value. Indeed,

$$\delta_n^\theta \geq 1 \quad \Leftrightarrow \quad \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2} \geq 1 + \theta^2 - 2\epsilon_n^2. \quad (41)$$

If the right-hand side of the latter inequality is negative, the inequality (41) holds trivially, while on the other hand assuming $1 + \theta^2 - 2\epsilon_n^2 \geq 0$ and squaring both sides reduces to $\epsilon_n \geq 0$, which proves the first part of the inequality (40). Observing that $\delta_n^\theta \leq \delta_n^0 = 1/(1 - \epsilon_n^2)$ implies the second part. The monotonicity of δ_n^θ with respect to the iteration number n can be shown by taking the derivative of δ_n^θ with respect to ϵ_n (since ϵ_n strictly decreases from 1 to 0). \square

The theorem implies that when the relative energy norm ϵ_n of the error in CG is small enough, the approximations x_n^θ start to be very close to the CG approximations x_n^{CG} . For example, a decrease of ϵ_n by an order of magnitude implies $\|x_n^\theta\|/\|x_n^{\text{CG}}\| \leq 1/0.99 \approx 1.01$.

Next, we compare the energy backward errors associated with both x_n^θ and x_n^{CG} . The energy backward error of the CG approximation x_n^{CG} is equal to

$$\xi_\theta(x_n^{\text{CG}}) = \frac{\|e_n^{\text{CG}}\|_A}{\sqrt{\|x_n^{\text{CG}}\|_A^2 + \theta^2\|x\|_A^2}} = \frac{\|e_n^{\text{CG}}\|_A}{\sqrt{(1 + \theta^2)\|x\|_A^2 - \|e_n^{\text{CG}}\|_A^2}} = \frac{\epsilon_n}{\sqrt{1 + \theta^2 - \epsilon_n^2}}.$$

For the ratio of the energy backward errors of the approximations x_n^θ and x_n^{CG} , we have hence

$$\frac{\xi_\theta(x_n^\theta)}{\xi_\theta(x_n^{\text{CG}})} = \left[\frac{2(1 + \theta^2 - \epsilon_n^2)}{1 + \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2}} \right]^{1/2}. \quad (42)$$

In the next theorem, we give the bounds for this ratio in terms of the relative energy norm ϵ_n of the error in CG. Its statement can be verified directly from (42).

Theorem 6.2. *The energy backward errors associated with approximations x_n^θ and x_n^{CG} satisfy*

$$\sqrt{1 - \epsilon_n^2} \leq \frac{\xi_\theta(x_n^\theta)}{\xi_\theta(x_n^{\text{CG}})} \leq 1.$$

The approximations x_n^θ minimize the energy backward error ξ_θ , while CG minimizes the energy norm of the error, i.e., the perturbation of the right-hand side measured in the A^{-1} -norm, over the Krylov subspace $\mathcal{K}_n(A, b)$. We compare the relative decrease of these error measures associated with approximations x_n^θ and x_n^{CG} at the given step n .

Let $\epsilon_n = \|e_n^{\text{CG}}\|_A/\|x\|_A$ denote as before the relative energy norm of the error associated with the CG approximation x_n^{CG} . We define the relative decrease of the energy backward error of the approximation x_n^θ by

$$\epsilon_n^\theta \equiv \lim_{\hat{x} \rightarrow 0} \frac{\xi_\theta(x_n^\theta)}{\xi_\theta(\hat{x})}.$$

For $\hat{x} = 0$, the expression (10) cannot be evaluated directly for $\theta = 0$, but the equation (39) can be evaluated for $\epsilon_n = 1$ (i.e., $\hat{x} = 0$). This gives

$$\begin{aligned} \lim_{\hat{x} \rightarrow 0} \xi_\theta(\hat{x}) &= \lim_{\epsilon \rightarrow 1} \left[\frac{2}{1 + \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon^2}} \right]^{1/2} \epsilon \\ &= \left[\frac{2}{1 + \theta^2 + |1 - \theta^2|} \right]^{1/2} = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1, \\ \frac{1}{\theta} & \text{if } 1 < \theta \end{cases} \end{aligned}$$

and hence

$$\frac{\epsilon_n^\theta}{\epsilon_n} = \begin{cases} \left[\frac{2}{1 + \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2}} \right]^{1/2} & \text{if } 0 \leq \theta \leq 1, \\ \left[\frac{2\theta^2}{1 + \theta^2 + \sqrt{(1 + \theta^2)^2 - 4\theta^2\epsilon_n^2}} \right]^{1/2} & \text{if } 1 < \theta. \end{cases} \quad (43)$$

For a fixed θ , the ratio $\epsilon_n^\theta/\epsilon_n$ as a function of $\epsilon_n \in [0, 1]$ is decreasing from the value 1 (for $\epsilon_n = 0$) to the value $1/\sqrt{1 + \theta^2}$ if $0 \leq \theta \leq 1$ and $\theta/\sqrt{1 + \theta^2}$ if $1 < \theta$ (for $\epsilon_n = 1$). The minimum of this final value of $\epsilon_n^\theta/\epsilon_n$ is $1/\sqrt{2}$ and we have the following theorem.

Theorem 6.3. *The relative decrease of the energy backward error ϵ_n^θ associated with the approximations x_n^θ and the relative decrease of the energy norm of the error associated with the CG approximations x_n^{CG} satisfy*

$$\frac{1}{\sqrt{2}} \leq \frac{\epsilon_n^\theta}{\epsilon_n} \leq 1.$$

Therefore, the relative decrease of the energy backward error of the approximations x_n^θ is very closely related to the relative decrease of the energy norm of the error in CG.

7 Numerical experiments

We illustrate the results of Section 6 on a system with the symmetric positive matrix BCSSTK04 ($N = 132$, $\kappa_2(A) \approx 2.293 \cdot 10^6$) from the Matrix Market collection [6] and with the right-hand side $b = [1, \dots, 1]^T$. The CG approximations are computed using the standard implementation of the conjugate gradient method; see [15]. In order to compute the approximations x_n^θ and the coefficients δ_n^θ in (37) and (38), respectively, we need to evaluate the relative energy norm of the error of the CG approximations ϵ_n using the MATLAB's backslash operator.

In Figures 1-3 we plot, for $\theta = 0$, $\theta = 1/2$, and $\theta = 1$, respectively, the energy backward errors of the CG approximations $\xi_\theta(x_n^{\text{CG}})$ and of the optimal approximations x_n^θ together with their ratio with respect to the iteration number n of the CG method. We also include the plot of the relative energy norm of the error $\epsilon_n = \|e_n^{\text{CG}}\|_A/\|x\|_A$ associated with CG iterates. In the initial stage the CG method nearly stagnates in terms of the relative energy norm of the error and in such a case, we can expect the largest differences between the CG approximations x_n^{CG} and the approximations x_n^θ in terms of their energy backward errors. For $\theta = 0$, the energy backward errors of the approximations x_n^θ are equal to the relative energy norm of the error in CG as expected (see the last paragraph of Section 5), while the difference becomes smaller when the parameter θ increases.

Please note that the approximations x_n^θ are not computable in practice. For the coefficient δ_n^θ , we have to evaluate or estimate the relative A -norm of the error in CG: the A -norm of the error as well as the A -norm of the solution x to (1). Several authors

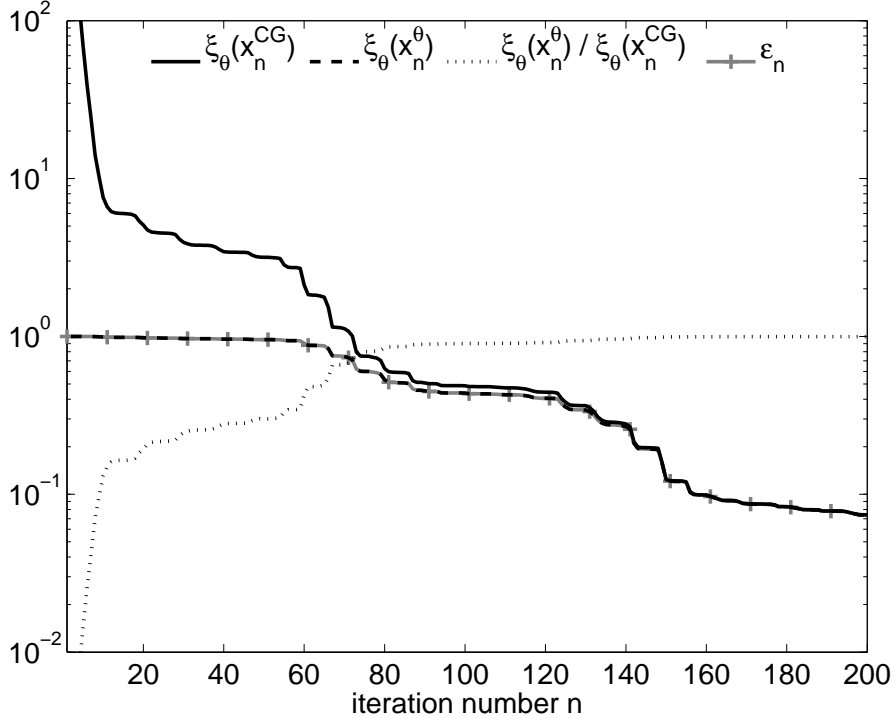


Figure 1: $\theta = 0$: The energy backward errors $\xi_\theta(x_n^{\text{CG}})$ (solid lines) and $\xi_\theta(x_n^\theta)$ (dashed lines), the ratios of the energy backward errors $\xi_\theta(x_n^\theta)/\xi_\theta(x_n^{\text{CG}})$ (dotted lines), and the relative energy norm of the CG error $\epsilon_n = \|e_n^{\text{CG}}\|_A/\|x\|_A$ (gray solid lines with markers) with respect to the iteration number n .

proposed recently bounds for the error norms in CG [4, 5, 7, 8, 12, 13, 23, 24, 31, 32]. One can for instance estimate $\|e_n^{\text{CG}}\|_A$ using

$$\|e_n^{\text{CG}}\|_A^2 = \|e_{n+d}^{\text{CG}}\|_A^2 + \|x_n - x_{n+d}\|_A^2 \approx \|x_n - x_{n+d}\|_A^2$$

and use $\|x\|_A^2 = \|x_{n+d}^{\text{CG}}\|_A^2 + \|e_{n+d}^{\text{CG}}\|_A^2 \approx \|x_{n+d}^{\text{CG}}\|_A^2$ to obtain an approximation of $\|x\|_A$, which leads to the estimate

$$\epsilon_n \approx \frac{\|x_n - x_{n+d}\|_A}{\|x_{n+d}\|_A}. \quad (44)$$

Note that $\|x_n - x_{n+d}\|_A$ and $\|x_{n+d}\|_A$ do not need to be computed directly but can be evaluated cheaply using the coefficients computed during the CG iterations; see, e.g., [33, 31, 25]. In order to obtain an accurate estimate the requirement is that $\|e_{n+d}^{\text{CG}}\|_A \ll \|e_n^{\text{CG}}\|_A < \|e_0^{\text{CG}}\|_A = \|x\|_A$ and thus the relative A -norm of the error e_{n+d}^{CG} must be sufficiently smaller than 1. However, as indicated in Section 6 the small relative error in CG implies that the energy backward error of the CG approximation is already very close to the optimal one. As a consequence a possible implementation of a variant of the CG method minimizing the energy backward error using the estimate (44)

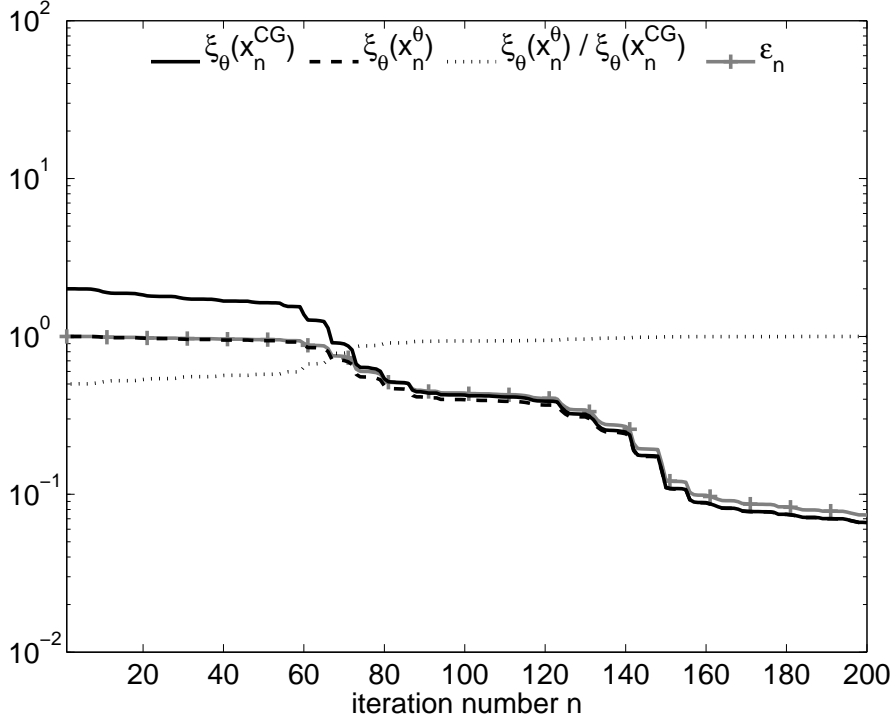


Figure 2: $\theta = 1/2$: The energy backward errors $\xi_\theta(x_n^{\text{CG}})$ (solid lines) and $\xi_\theta(x_n^\theta)$ (dashed lines), the ratios of the energy backward errors $\xi_\theta(x_n^\theta)/\xi_\theta(x_n^{\text{CG}})$ (dotted lines), and the relative energy norm of the CG error $\epsilon_n = \|e_n^{\text{CG}}\|_A/\|x\|_A$ (gray solid lines with markers) with respect to the iteration number n .

could work properly only when the approximations x_n^θ and x_n^{CG} are already practically indistinguishable.

8 Conclusions

We introduce the energy backward error for problems with a symmetric positive definite matrix and provide its interpretations in the Galerkin approximation of elliptic problems. It appears that the energy backward error (with only the system matrix perturbed) can be interpreted as a relative distance to a basis in which the approximate solution of the system represents the coordinates of the Galerkin approximation.

Furthermore, we analyze the convergence of the conjugate gradient method in terms of the energy backward error. For this purpose, we introduce approximations which at each iteration minimize the energy backward error over the given Krylov subspaces. Such approximations are given by scalar multiples of the actual CG approximations; no additional information is extracted from the computed Krylov subspace in order to minimize the energy backward error. When the A -norm of the error in CG mildly

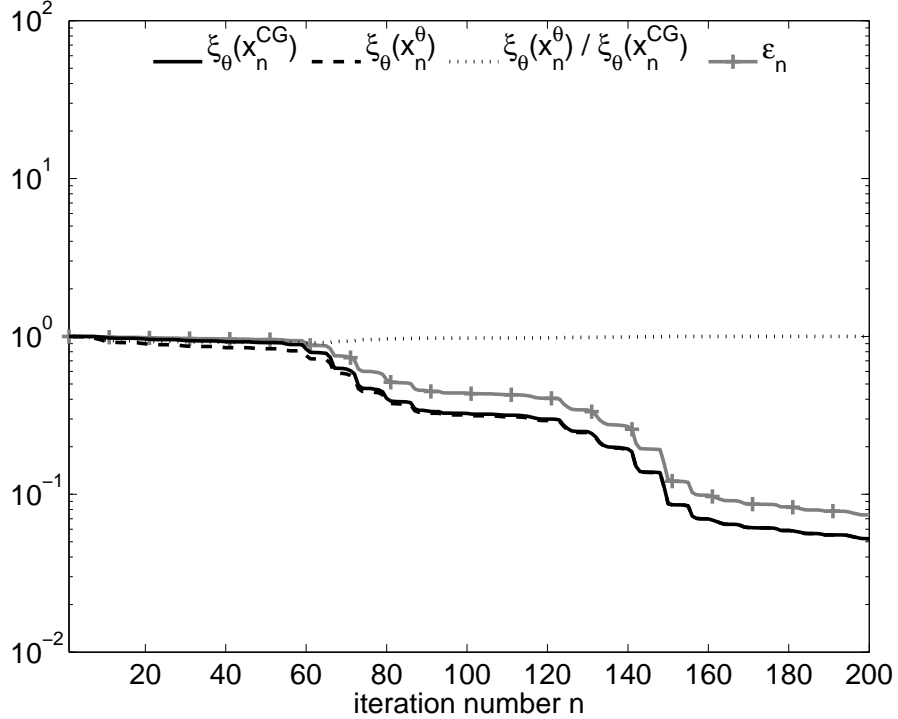


Figure 3: $\theta = 1$: The energy backward errors $\xi_\theta(x_n^{\text{CG}})$ (solid lines) and $\xi_\theta(x_n^\theta)$ (dashed lines), the ratios of the energy backward errors $\xi_\theta(x_n^\theta)/\xi_\theta(x_n^{\text{CG}})$ (dotted lines), and the relative energy norm of the CG error $\epsilon_n = \|e_n^{\text{CG}}\|_A/\|x\|_A$ (gray solid lines with markers) with respect to the iteration number n .

decreases, the energy backward errors of the CG approximations start to be very close to the optimal ones and in this way such a scaling can be regarded only as a sort of smoothing of the energy backward error in CG in the early convergence phase when the energy backward error of the approximations can be large.

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. [6](#)
- [2] M. ARIOLI, I. S. DUFF, AND D. RUIZ, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 138–144. [2](#)
- [3] M. ARIOLI, E. NOULARD, AND A. RUSSO, *Stopping criteria for iterative methods: applications to PDE's*, Calcolo, 38 (2001), pp. 97–112. [2](#)

- [4] S. F. ASHBY, M. J. HOLST, T. A. MANTEUFFEL, AND P. E. SAYLOR, *The role of the inner product in stopping criteria for conjugate gradient iterations*, BIT, 41 (2001), pp. 26–52. 16
- [5] O. AXELSSON AND I. KAPORIN, *Error norm estimation and stopping criteria in preconditioned conjugate gradient iteration*, Numer. Linear Algebra Appl., 8 (2001), pp. 265–286. 16
- [6] R. F. BOISVERT, R. POZO, K. REMINGTON, R. BARRET, AND J. J. DONGARRA, *The Matrix Market: A web resource for test matrix collections*, in Quality of Numerical Software, Assessment and Enhancement, R. F. Boisvert, ed., Chapman & Hall, London, UK, 1997. 15
- [7] D. CALVETTI, S. MORIGI, L. REICHEL, AND F. SGALLARI, *Computable error bounds and estimates for the conjugate gradient method*, Numer. Algorithms, 25 (2000), pp. 75–88. 16
- [8] ———, *An iterative method with error estimators*, J. Comput. Appl. Math., 127 (2001), pp. 93–119. 16
- [9] X.-W. CHANG, C. C. PAIGE, AND D. TITLEY-PELOQUIN, *Characterizing matrices that are consistent with given solutions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1406–1420. 7
- [10] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, vol. 4 of Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1978. 5
- [11] 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, Oxford, UK, 2005. 6
- [12] G. H. GOLUB AND G. MEURANT, *Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods*, BIT, 37 (1997), pp. 687–705. 16
- [13] G. H. GOLUB AND Z. STRAKOŠ, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268. 16
- [14] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997. 2
- [15] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stand., 49 (1952), pp. 409–435. 2, 9, 15
- [16] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996. 2

- [17] ———, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 2nd ed., 2002. 2
- [18] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, NY, 1985. 11
- [19] 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. 6
- [20] E. M. KASENALLY, *GMBACK: a generalised minimum backward error algorithm for nonsymmetric linear systems*, SIAM J. Sci. Comput., 16 (1995), pp. 698–719. 2
- [21] E. M. KASENALLY AND V. SIMONCINI, *Analysis of minimum perturbation algorithm for nonsymmetric linear systems*, SIAM J. Numer. Anal., 34 (1997), pp. 48–66. 2
- [22] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Stand., 45 (1950), pp. 255–282. 2, 9
- [23] G. MEURANT, *The computation of bounds for the norm of the error in the conjugate gradient algorithm*, Numer. Algorithms, 16 (1997), pp. 77–87. 16
- [24] ———, *Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm*, Numer. Algorithms, 22 (1999), pp. 353–365. 16
- [25] G. MEURANT AND Z. STRAKOŠ, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542. 16
- [26] W. OETTLI AND W. PRAGER, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409. 2
- [27] 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. 2
- [28] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, 2003. 2
- [29] Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869. 2
- [30] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Computer Science and Scientific Computing, Academic Press, San Diego, 1990. 8

- [31] Z. STRAKOŠ AND P. TICHÝ, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electron. Trans. Numer. Anal., 13 (2002), pp. 56–80. 16
- [32] ———, *Error estimation in preconditioned conjugate gradients*, BIT, 45 (2005), pp. 789–817. 16
- [33] ———, *On efficient numerical approximation of the bilinear form $c^*A^{-1}b$* . SIAM J. Sci. Comput., 33 (2011), pp. 565–587. 16
- [34] H. A. VAN DER VORST, *Iterative Krylov Methods for Large Linear Systems*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2009. 2
- [35] J. H. WILKINSON, *Algebraic Eigenvalue Problem*, Oxford University Press, New York, NY, 1965. 2