BIT Numerical Mathematics (2005) 45: 789–817
DOI: 10.1007/s10543-005-0032-1

ERROR ESTIMATION IN PRECONDITIONED CONJUGATE GRADIENTS*

ZDENĚK STRAKOŠ[1,**] and PETR TICHÝ[1,***]

[1] *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic.*
email: {strakos,tichy}@cs.cas.cz

Abstract.

In practical problems, iterative methods can hardly be used without some acceleration of convergence, commonly called preconditioning, which is typically achieved by incorporation of some (incomplete or modified) direct algorithm as a part of the iteration. Effectiveness of preconditioned iterative methods increases with possibility of stopping the iteration when the desired accuracy is reached. This requires, however, incorporating a proper measure of achieved accuracy as a part of computation.

The goal of this paper is to describe a simple and numerically reliable estimation of the size of the error in the preconditioned conjugate gradient method. In this way this paper extends results from [Z. Strakoš and P. Tichý, ETNA, 13 (2002), pp. 56–80] and communicates them to practical users of the preconditioned conjugate gradient method.

AMS subject classification (2000): 15A06, 65F10, 65F25, 65G50.

Key words: preconditioned conjugate gradient method, error bounds, stopping criteria, evaluation of convergence, numerical stability, finite precision arithmetic, rounding errors.

1 Introduction.

Discretization of mathematical models of real-world problems often leads to large and sparse (possibly structured) systems of linear algebraic equations. All steps of mathematical modeling (mathematical description of reality in the form of a mathematical model, its discretization and numerical solution of the discretized problem) are subject to errors (errors of the model, discretization errors

* Received September 2004. Accepted in final form August 2005. Communicated by Iain Duff.
** This work was supported by the Program Information Society, project 1ET400300415, and by the Grant Agency of Academy of Sciences of the Czech Republic under grant No. KJB1030306.
*** This work was performed during the academic year 2003/2004 while this coauthor was on leave at the Institute of Mathematics, TU Berlin, Germany, sponsored by the Emmy Noether Program of the Deutsche Forschungsgemeinschaft.

and computational errors, the last being often composed of two parts – truncation errors and errors due to roundoff). An output of the solution process must therefore be confronted with its possible errors through *verification and validation*. While verification addresses the question – whether and how accurately the obtained (approximate) solution conforms to the mathematical model, validation deals with the more general question – to which extent the whole modeling process represents the modeled reality (for a recent discussion of these fundamental topics we refer to [7]). It is desirable that the errors of the model, discretization errors and computational errors are in some balance. They do not need to be of the same order; the discretization and computational errors should not significantly contribute to the total error and affect negatively the validation process [7].

When the linear algebraic systems arising from mathematical modeling are very large (of orders of hundreds of thousands or millions of unknowns), preconditioned iterative methods are taking ground over the purely direct methods. Iterative methods can in very large scale computations exploit a fundamental advantage – they can increase effectiveness of the whole solution process by stopping the iteration when the desired accuracy (as compared to the discretization error) is reached (cf. [1, 4]). This requires, however, a cheap and reliable evaluation of convergence, which is the essential ingredience for choosing proper stopping criteria.

In this paper we consider a system of linear algebraic equations

$$Ax = b \tag{1.1}$$

where A is a symmetric positive definite n by n matrix and b is n-dimensional vector (for simplicity of notation we consider A, b real; all results presented here can trivially be extended to the complex case). For such systems the preconditioned conjugate gradient method [22, 26, 34, 40] represents in most large scale cases a good choice. A goal of this paper is to summarize and discuss evaluation of convergence in the preconditioned conjugate gradient method. In particular, we will focus on estimating the A-norm of the error.

Estimating the A-norm of the error in the conjugate gradient method was subject of many papers, reports and subsections in the books. History and various aspects of estimating the A-norm of the error in the unpreconditioned conjugate gradient method were thoroughly described in [38]. The formulas presented in [38] were published (in some form) previously, e.g. in [22, 12] and [6]. The original contribution of [38] consists, to our opinion, in providing *theoretical justification for practical use of the error estimates* and in putting different estimates in the proper context. Our present paper extends the results from [38] to the preconditioned conjugate gradient method. A need for such paper can be seen from [6, Section 6] or [1, Section 3], which thoroughly and extensively examine estimating error norms in the preconditioned conjugate gradients. Both papers [6, 1] present interesting original results and offer new insight into the error estimation in the preconditioned conjugate gradients. They do not consider, however, an influence of rounding errors. All derivations in [6, Section 6] or [1, Section 3] assume exact arithmetic. Consequently, they unrealistically as-

sume preserving orthogonality, and the results are based on exploiting the finite termination property, i.e., on getting the *exact solution* in a finite number of steps (which does not exceed the dimension of the problem). These assumptions are clearly drastically violated in most practical computations. In order to be widely used, practical error estimators need a proper justification including a thorough analysis of rounding error effects (for a related discussion, see [38] and also [16]).

Section 2 summarizes fundamentals of the conjugate gradient method and briefly recalls several possible ways of convergence evaluation. Section 3 presents a simple estimate for the A-norm of the error in the preconditioned conjugate gradient method. Section 4 deals with numerical stability of the proposed estimate and Section 5 contains numerical experiments which demonstrate its effectivity and possible drawbacks. The paper ends with concluding remarks.

2 Fundamentals of convergence evaluation.

The conjugate gradient method (CG) [22] belongs to the class of the so-called Krylov subspace methods. Starting with an initial approximation x_0, it constructs the subsequent approximations x_j, $j = 1, 2, \ldots$ to the solution x on the linear manifolds

$$x_j \in x_0 + \mathcal{K}_j(A, r_0) \tag{2.1}$$

where

$$\mathcal{K}_j(A, r_0) = \operatorname{span}\left\{r_0, Ar_0, \ldots, A^{j-1}r_0\right\}$$

represents the jth Krylov subspace, $r_0 = b - Ax_0$. CG determines its approximations by orthogonal projections, i.e., the residual $r_j = b - Ax_j$ of the jth approximate solution is orthogonal to the jth Krylov subspace $\mathcal{K}_j(A, r_0)$. This means that $x_j = x_0 + y_j$ can be obtained from the solution y_j of the j-dimensional problem

$$P_j\{r_0 - Ay\} = 0\,, \tag{2.2}$$

where P_j stands for the orthogonal projection onto $\mathcal{K}_j(A, r_0)$, and $y \in \mathcal{K}_j(A, r_0)$ (the operator A is in (2.2) restricted to $\mathcal{K}_j(A, r_0)$). It is well known [22] that, until x_j converges to the exact solution x (which must in the absence of roundoff happen in at most n steps), x_j is uniquely determined by (2.2).

In practical problems we hope that the acceptable approximate solution is attained for j much smaller than the dimension of the problem n. Thus, CG represents a typical model-reduction approach, in which the original problem (represented by the large discretized model) is reduced (here by restriction and orthogonal projection onto the Krylov subspace) to the problem of much smaller dimension. The resulting reduced problem determines the approximate solution. Quality of the approximate solution depends on the amount of significant information about the original problem passed to the reduced problem.

The condition (2.2) is equivalent to the minimization of the A-norm of the error over the manifold (2.1). The jth CG approximation is therefore uniquely

determined by the minimizing condition

$$\|x - x_j\|_A \;=\; \min_{u \in x_0 + \mathcal{K}_j(A, r_0)} \|x - u\|_A \,, \tag{2.3}$$

where

$$\|x - u\|_A \;=\; (x - u, A(x - u))^{\frac{1}{2}} \;. \tag{2.4}$$

The A-norm of the error on the algebraic level (2.4) typically has a counterpart in the original real-world problem. In some applications it can be interpreted as the discretized measure of energy which is to be minimized see, e.g. [1, 4]. Then CG with stopping criterion based on the A-norm of the error consistently reduces large discretized models to small ones. In other applications (such as in image processing) the Euclidean norm of the error $\|x - x_j\|$ plays an important role. In this paper we focus in particular on estimating the A-norm of the error.

Hestenes and Stiefel [22] considered the A-norm of the error a possible candidate for measuring the "goodness" of x_j as an estimate of x. They showed that though it was impossible to compute the A-norm of the jth error without knowing the solution x, it was possible to estimate it. Later, and independently of [22], the idea of estimating errors in CG was promoted by Golub in relation to the problem of moments, Gauss quadrature and its modifications [10, 11]. A comprehensive summary of this approach was given in the papers coauthored with Meurant [14, 15].

In [38] it was shown that the lower bound for the A-norm of the error based on the Gauss quadrature is mathematically equivalent to the lower bound derived from the identity given by Hestenes and Stiefel in [22]. The estimate by Hestenes and Stiefel can be computed at a negligible cost of several floating point operations per iteration. Until the A-norm of the error reaches its ultimate level of accuracy, this estimate is numerically stable.

In [32, 3], backward error perturbation theory (see e.g. [30, 35, 2]) was used to derive a family of stopping criteria for iterative methods. In particular, given x_j, the relative norms $\|\Delta A\|/\|A\| = \|\Delta b\|/\|b\|$ of the smallest perturbations ΔA and Δb such that the *approximate solution* x_j represents the *exact solution* of the perturbed system

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

can be computed by the normwise backward error

$$\frac{\|r_j\|}{\|A\|\|x_j\| + \|b\|} \,. \tag{2.5}$$

This approach can be generalized in order to quantify levels of confidence in A and b, see [32, 3]. Normwise backward error is, as a base for stopping criteria, frequently recommended in the numerical analysis literature, see, e.g. [8, 23], and it is used and popularized by numerical analysts [29, 13]. Despite this effort, evaluating convergence is in most of scientific computations still based on the

relative residual norm

$$\frac{\|r_j\|}{\|r_0\|}. \tag{2.6}$$

With $x_0 = 0$, it measures the relative norm $\|\Delta b\|/\|b\|$ of the smallest perturbation Δb in the right-hand side b only (A is considered unperturbed) such that x_j is the exact solution of the perturbed system $A\,x_j = b + \Delta b$. For $x_0 \neq 0$ (2.6) strongly depends on the initial approximation x_0 and can give a misleading information about convergence, see, e.g. [33]. For some additional information see also [5, 20].

We do not argue that the relative residual norm can not be useful. In some cases it is a proper quantity to be checked. Sometimes it is a part of more complex convergence considerations, e.g. in solving nonlinear systems or in numerical optimization. We do argue, however, that in many other cases, and in particular in numerical solving of partial differential equations, the relative residual norm is often uncritically used as the only measure of convergence.

Mathematically (ignoring effects of rounding errors), extension of the approaches mentioned above to preconditioned methods does not represent a problem, see, e.g., [29, 13]. Extension of the Gauss quadrature-based formulas for estimating the A-norm of the error in CG (algorithm CGQL [15]) to the preconditioned conjugate gradient method (PCG) was published in [27, 28] (algorithm PCGQL). In the following section we deal with the extension of error estimates based on the Hestenes and Stiefel formula [22, 38].

3 PCG error estimates.

In the standard view of preconditioning, the CG method is thought of as being applied to a "preconditioned" system

$$\hat{A}\hat{x} = \hat{b}, \tag{3.1}$$

$$\hat{A} = L^{-1}AL^{-T}, \quad \hat{b} = L^{-1}b, \tag{3.2}$$

where L represents a proper nonsingular (lower triangular) matrix, giving

ALGORITHM 1. *CG for* $\hat{A}\hat{x} = \hat{b}$

given $\hat{x}_0$, $\hat{r}_0 = \hat{b} - \hat{A}\hat{x}_0$,
for $j = 0, 1, \ldots$

$$\gamma_j = \frac{(\hat{r}_j, \hat{r}_j)}{(\hat{p}_j, \hat{A}\hat{p}_j)}$$
$$\hat{x}_{j+1} = \hat{x}_j + \hat{\gamma}_j\,\hat{p}_j$$
$$\hat{r}_{j+1} = \hat{r}_j - \hat{\gamma}_j\,\hat{A}\hat{p}_j$$
$$\hat{\delta}_{j+1} = \frac{(\hat{r}_{j+1}, \hat{r}_{j+1})}{(\hat{r}_j, \hat{r}_j)}$$
$$\hat{p}_{j+1} = \hat{r}_{j+1} + \hat{\delta}_{j+1}\,\hat{p}_j$$

end for.

Defining

$$\gamma_j \equiv \hat{\gamma}_j, \quad \delta_j \equiv \hat{\delta}_j, \tag{3.3}$$
$$x_j \equiv L^{-T}\hat{x}_j, \quad r_j \equiv L\,\hat{r}_j, \quad p_j \equiv L^{-T}\hat{p}_j, \quad s_j \equiv L^{-T}L^{-1}r_j \equiv M^{-1}r_j,$$

(here x_j and r_j represent the approximate solution and residual for the original problem $Ax = b$), we obtain the standard version of the PCG method

ALGORITHM 2. *PCG for* $Ax = b$

given x_0, $r_0 = b - Ax_0$, $s_0 = M^{-1}r_0$, $p_0 = s_0$,
for $j = 0, 1, \ldots$

$$\begin{aligned}
\gamma_j &= \frac{(r_j, s_j)}{(p_j, Ap_j)} \\
x_{j+1} &= x_j + \gamma_j\, p_j \\
r_{j+1} &= r_j - \gamma_j\, Ap_j \\
s_{j+1} &= M^{-1}r_{j+1} \\
\delta_{j+1} &= \frac{(r_{j+1}, s_{j+1})}{(r_j, s_j)} \\
p_{j+1} &= s_{j+1} + \delta_{j+1}\, p_j
\end{aligned}$$

end for.

The preconditioner

$$M = LL^T \tag{3.4}$$

is chosen so that the linear system with the matrix M is easy to solve, while the matrix $L^{-1}AL^{-T}$ should ensure fast convergence of CG. The last goal is fulfilled, e.g., when $L^{-1}AL^{-T}$ is well conditioned (approximates the identity matrix) or has properly clustered eigenvalues. Here we emphasize that *location* as well as *diameter* of the clusters are important; improperly located clusters of very small diameter do not necessarily ensure fast convergence, see [21, 37]. Location of the clusters is sometimes omitted from consideration, and this leads to inaccurate or even false statements, which can be found in widespread literature.

3.1 Estimating the A-norm of the error.

In PCG, the A-norm of the error can be estimated similarly as in ordinary CG. For a given d, the approximate solutions $\hat{x}_j$ of the system (3.1) satisfy

$$\|\hat{x} - \hat{x}_j\|_{\hat{A}}^2 = \sum_{i=j}^{j+d-1} \hat{\gamma}_i \|\hat{r}_i\|^2 + \|\hat{x} - \hat{x}_{j+d}\|_{\hat{A}}^2, \tag{3.5}$$

see [38, (4.4)]. Using (3.3),

$$\|\hat{r}_j\|^2 = r_j^T L^{-T}L^{-1}r_j = r_j^T M^{-1}r_j = (r_j, s_j)\,,$$

and

$$\|\hat{x} - \hat{x}_j\|_{\hat{A}}^2 = (L^Tx - L^Tx_j)^T L^{-1}AL^{-T}(L^Tx - L^Tx_j) = \|x - x_j\|_A^2.$$

The identity (3.5) can therefore be written in the form

$$\|x - x_j\|_A^2 = \sum_{i=j}^{j+d-1} \gamma_i (r_i, s_i) + \|x - x_{j+d}\|_A^2. \tag{3.6}$$

Assuming a reasonable decrease of the A-norm of the error in the steps $j+1$ through $j+d$, the square root of the quantity

$$\nu_{j,d} \equiv \sum_{i=j}^{j+d-1} \gamma_i (r_i, s_i) \tag{3.7}$$

gives a tight lower bound for the A-norm of the jth error of PCG applied to the system $Ax = b$. Please notice that (similarly as in the ordinary CG) the quantities γ_i and (r_i, s_i) are at our disposal during the PCG iterations. For earlier publications of these identities please see [39, 1].

3.2 Estimating the relative A-norm of the error.

Consider PCG applied to linear algebraic systems arising from a finite element discretization of self-adjoint elliptic partial differential equations. Then it is natural to use the stopping criterion that compares the relative A-norm of the error

$$\frac{\|x - x_j\|_A}{\|x\|_A} \tag{3.8}$$

with the discretization error, see [1].

In [1], however, the A-norm of the jth error is estimated using (3.7), while the estimate of the A-norm of the solution $\|x\|_A$ is based on the formula

$$\|x\|_A^2 = r_0^T x_j + b^T x_0 + \|x - x_j\|_A^2 \tag{3.9}$$

which gives the lower bound

$$\|x\|_A^2 \geq \tilde{\xi}_j \equiv r_0^T x_j + b^T x_0 \,. \tag{3.10}$$

Estimating the A-norm of the solution using the value $\tilde{\xi}_j^{1/2}$ has, besides computing an unnecessary scalar product $r_0^T x_j$, a possible disadvantage. Derivation of the identity (3.9) assumes preserving of global orthogonality during the PCG computations, cf. [1, p. 9]. In particular, it can be shown that in finite precision arithmetic it holds (up to some small inaccuracy)

$$\|x\|_A^2 \approx r_0^T x_j + b^T x_0 + r_j^T (x_j - x_0) + \|x - x_j\|_A^2. \tag{3.11}$$

In exact arithmetic, the term $r_j^T (x_j - x_0)$ is equal to zero. In finite precision arithmetic, however, its size can be close to $\|r_j\| \, \|x_j - x_0\|$. Consequently, the estimate $\tilde{\xi}_j^{1/2}$ can for large $r_j^T (x_j - x_0) + \|x - x_j\|_A^2$ (in comparison to $\|x\|_A^2$) provide misleading information about the size of $\|x\|_A$.

A mathematically equivalent identity to (3.9) that overcomes previous difficulties can be obtained in the following way. Subtracting

$$\|x - x_0\|_A^2 = \nu_{0,j+d} + \|x - x_{j+d}\|_A^2,$$

$$\|x - x_0\|_A^2 = \|x\|_A^2 - 2b^T x_0 + \|x_0\|_A^2 = \|x\|_A^2 - b^T x_0 - r_0^T x_0, \tag{3.12}$$

the identity

$$\|x\|_A^2 = \nu_{0,j+d} + b^T x_0 + r_0^T x_0 + \|x - x_{j+d}\|_A^2 \tag{3.13}$$

gives the corresponding lower bound

$$\|x\|_A^2 \;\geq\; \xi_{j+d} \equiv \nu_{0,j+d} + b^T x_0 + r_0^T x_0\,. \tag{3.14}$$

With $d = 0$, the identities (3.13) and (3.9), as well as the estimates ξ_j and $\tilde{\xi}_j$ are *mathematically equivalent*. However, the evaluation of ξ_j is cheaper than the evaluation of $\tilde{\xi}_j$ and, more substantially, (3.13) holds with a small inaccuracy also in finite precision PCG computations independently on the loss of global orthogonality, cf. Section 4.

Replacing the squared A-norm of the solution $\|x\|_A^2$ by the lower bound ξ_{j+d} and the squared jth A-norm of the error $\|x - x_j\|_A^2$ by the lower bound $\nu_{j,d}$, we obtain the estimate $\varrho_{j,d}$ for the squared relative A-norm of the error

$$\varrho_{j,d} \equiv \frac{\nu_{j,d}}{\xi_{j+d}}. \tag{3.15}$$

It should be noted that an improper choice of x_0 can give $\xi_{j+d} \leq 0$ which makes the estimate $\varrho_{j,d}$ in such case useless. We will, however, explain that $\xi_{j+d} \leq 0$ means a meaningless choice of x_0. First, a nonzero x_0 should not be used in an application of the CG method (and of any other Krylov subspace method) unless there is a good reason for using it. In CG, the very natural condition

$$\|x - x_0\|_A^2 \;\leq\; \|x\|_A^2 \tag{3.16}$$

should always be imposed. Though we can not compute the individual values $\|x\|_A^2$, $\|x - x_0\|_A^2$, its difference can easily be checked using (3.12). Second, if $\xi_{j+d} \leq 0$, then from (3.14)

$$b^T x_0 + r_0^T x_0 = \|x\|_A^2 - \|x - x_0\|_A^2 < 0 \tag{3.17}$$

and x_0 violates the condition (3.16). In such case, x_0 should be discarded or properly scaled in order to satisfy (3.16). In particular, x_0 can be scaled such that $\|x - \alpha x_0\|_A^2$ is minimal using

$$\alpha = \frac{b^T x_0}{x_0^T A x_0},$$

(for another application of the same little trick see [33, p. 1903]). With (3.16) $\xi_{j+d} > 0$ and, using (3.13),

$$0 < \varrho_{j,d} \;=\; \frac{\|x - x_j\|_A^2 - \|x - x_{j+d}\|_A^2}{\|x\|_A^2 - \|x - x_{j+d}\|_A^2} \;\leq\; \frac{\|x - x_j\|_A^2}{\|x\|_A^2},$$

i.e. $\varrho_{j,d}^{1/2}$ is a lower bound on the jth relative A-norm of the error. Please note that $\varrho_{j,d}^{1/2}$ can be close to the relative A-norm of the error even when $\nu_{j,d}^{1/2}$ is far from $\|x - x_j\|_A$.

3.3 Estimating the M-norm of the error.

In our paper [38] we described an estimate of the Euclidean norm of the error in CG. For CG applied to $\hat{A}\hat{x} = \hat{b}$, Algorithm 1, the estimate is based on the identity

$$\|\hat{x} - \hat{x}_j\|^2 = \sum_{i=j}^{j+d-1} \frac{\|\hat{p}_i\|^2}{(\hat{p}_i, \hat{A}\hat{p}_i)} \left(\|\hat{x} - \hat{x}_i\|_{\hat{A}}^2 + \|\hat{x} - \hat{x}_{i+1}\|_{\hat{A}}^2\right) + \|\hat{x} - \hat{x}_{j+d}\|^2. \tag{3.18}$$

Using (3.3), (3.18) can be rewritten as

$$\|x - x_j\|_M^2 = \sum_{i=j}^{j+d-1} \frac{\|p_i\|_M^2}{(p_i, Ap_i)} \left(\|x - x_i\|_A^2 + \|x - x_{i+1}\|_A^2\right) + \|x - x_{j+d}\|_M^2 \tag{3.19}$$

where x_j represents the PCG approximate solution for the original problem $Ax = b$. Replacing the unknown $\|x - x_i\|_A^2$ for $i = j, \ldots, j + d$ by the estimates $\nu_{i,2d-i+j}$ (see [38]) we obtain

$$\|x - x_j\|_M^2 \geq \tau_{j,d} + \|x - x_{j+d}\|_M^2 \tag{3.20}$$

where the square root of the quantity

$$\tau_{j,d} \equiv \sum_{i=j}^{j+d-1} \frac{\|p_i\|_M^2}{(p_i, Ap_i)} \left(\gamma_i (r_i, s_i) + 2 \sum_{k=i+1}^{j+2d-1} \gamma_k (r_k, s_k)\right) \tag{3.21}$$

represents a lower bound for the M-norm of the error.

4 Numerical stability analysis.

In [38] we showed that the Hestenes and Stiefel estimate is numerically stable (i.e. it is in finite precision CG computations not substantially affected by rounding errors) until the A-norm of the error approaches its ultimate level of accuracy. A similar result can be shown for the estimate (3.7) of the A-norm of the error in PCG.

PCG computes at each step an additional vector s_{j+1} as a solution of the linear system

$$Ms_{j+1} = r_{j+1}\,, \tag{4.1}$$

and uses

$$(r_{j+1}, s_{j+1}) \tag{4.2}$$

for computation of the coefficients γ_{j+1} and δ_{j+1} needed for determining of the new direction vector p_{j+1}. This is the difference which must be addressed in extension of the results from CG [38] to PCG.

From now on x_{j+1}, x_j, γ_j, p_j, r_{j+1}, r_j, s_{j+1}, δ_{j+1} and p_{j+1} will represent numerically computed quantities. Numerical stability analysis of the estimate (3.7)

must answer a question to which extent the identity (3.6) holds for quantities computed in finite precision arithmetic. Please note that this question is fundamentally different from its trivial part examining the error in *computing* $\nu_{j,d}$ from γ_i and $\mathrm{fl}[(r_i, s_i)]$, where $\mathrm{fl}[\cdot]$ denotes the result of the operation performed in finite precision arithmetic, using (3.7). In order to justify the estimate (3.7), we have to derive the identity for the computed quantities analogous to (3.6) without using any assumption which does not hold in finite precision computations. In particular, we can not use any assumption about orthogonality or finite termination.

The key step considers the *exact* identity for *numerically computed* quantities

$$\begin{aligned} \|x - x_j\|_A^2 &= \|x - x_{j+1} + x_{j+1} - x_j\|_A^2 \\ &= \|x - x_{j+1}\|_A^2 + 2(x - x_{j+1})^T A(x_{j+1} - x_j) + \|x_j - x_{j+1}\|_A^2 \end{aligned}$$

which gives the desired one-step difference

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 &= \|x_j - x_{j+1}\|_A^2 \\ &\quad + 2(x - x_{j+1})^T A(x_{j+1} - x_j). \end{aligned} \tag{4.3}$$

The technically complicated and quite tedious analysis which must follow can be summarized in several logically simple steps:

- First, the difference $x_{j+1} - x_j$ is equal to $\gamma_j p_j$ perturbed by inaccuracies due to rounding errors. Consequently, $\|x_{j+1} - x_j\|_A^2$ can be expressed as $\gamma_j(r_j, s_j)$ plus some additional terms depending on machine precision ε characterizing the finite precision arithmetic. These additional terms are small (this is not obvious; the proof requires a careful analysis).
- Second, considering the *approximation* of $A(x - x_{j+1})$ by the residual vector r_{j+1} computed in the $(j+1)$th iteration, the term $2(x - x_{j+1})^T A(x_{j+1} - x_j)$ can be seen as $2\gamma_j(r_{j+1}, p_j)$ plus additional small terms depending on ε (again, bounding the size of these terms needs nontrivial work).

The whole problem of justification of the estimate (3.7) in finite precision arithmetic is in this way reduced to proving that local orthogonality between the computed $(j+1)$th residual r_{j+1} and the computed jth direction vector p_j is in PCG maintained proportionally to machine precision. This represents the technically most complicated part of the whole analysis.

In following four subsections we present a detailed rounding error analysis of the identity (3.6). Subsection 4.1 describes the rounding errors arising in PCG iterates due to finite precision arithmetic. In Subsection 4.2 we develop a finite precision analogue of the identity (3.6) for $d = 1$. Subsection 4.3 shows that the local orthogonality between the vectors r_{j+1} and p_j is preserved, up to a term proportional to machine precision, in finite precision PCG computation. We finalize the rounding error analysis in Subsection 4.4.

Readers who wish to skip the details of our rounding error analysis may proceed immediately to Subsection 4.4 or even to numerical experiments in Section 5.

4.1 Finite precision PCG computations.

In the analysis we assume the standard model of floating point arithmetic with machine precision ε, see, e.g. [23, (2.4)],

$$\mathrm{fl}[a \circ b] = (a \circ b)(1+\delta), \quad |\delta| \le \varepsilon, \tag{4.4}$$

where a and b stands for floating-point numbers and the symbol $\circ$ stands for the operations addition, subtraction, multiplication and division. We assume that this model holds also for the square root operation. Under this model, we have for operations involving vectors v, w, a scalar α and the matrix A the following standard results [17], see also [19], [31]

$$\|\alpha\, v - \mathrm{fl}[\alpha\, v]\| \le \varepsilon\, \|\alpha\, v\|, \tag{4.5}$$

$$\|v + w - \mathrm{fl}[v+w]\| \le \varepsilon\, (\|v\| + \|w\|), \tag{4.6}$$

$$|(v,w) - \mathrm{fl}[(v,w)]| \le \varepsilon\, n\, (1+\mathcal{O}(\varepsilon))\, \|v\|\, \|w\|, \tag{4.7}$$

$$\|Av - \mathrm{fl}[Av]\| \le \varepsilon\, c\, \|A\| \|v\|. \tag{4.8}$$

When A is a matrix with at most h nonzeros in any row and if the matrix-vector product is computed in the standard way, $c = hn^{1/2}$. In the following analysis we count only for the terms linear in the machine precision ε and express the higher order terms as $\mathcal{O}(\varepsilon^2)$. By $\mathcal{O}(const)$ where *const* is different from ε^2 we denote *const* multiplied by a bounded positive term of an insignificant size which is independent of the *const* and of any other variables present in the bounds.

Numerically, the PCG iterates satisfy

$$x_{j+1} = x_j + \gamma_j p_j + \varepsilon z_j^x, \tag{4.9}$$

$$r_{j+1} = r_j - \gamma_j A p_j + \varepsilon z_j^r, \tag{4.10}$$

$$p_{j+1} = s_{j+1} + \delta_{j+1} p_j + \varepsilon z_j^p, \tag{4.11}$$

where εz_j^x, εz_j^r and εz_j^p account for the local roundoff ($r_0 = b - Ax_0 - \varepsilon f_0$, $\varepsilon\|f_0\| \le \varepsilon\{\|b\| + \|Ax_0\| + c\|A\|\|x_0\|\} + \mathcal{O}(\varepsilon^2)$). The local roundoff can be bounded according to the standard results (4.5)–(4.8) in the following way

$$\begin{aligned} \varepsilon\, \|z_j^x\| &\le \varepsilon\, \{\|x_j\| + 2\, \|\gamma_j p_j\|\} + \mathcal{O}(\varepsilon^2) \\ &\le \varepsilon\, \{3\|x_j\| + 2\|x_{j+1}\|\} + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{4.12}$$

$$\varepsilon\, \|z_j^r\| \le \varepsilon\, \{\|r_j\| + 2\, \|\gamma_j A p_j\| + c\, \|A\| \|\gamma_j p_j\|\} + \mathcal{O}(\varepsilon^2), \tag{4.13}$$

$$\begin{aligned} \varepsilon\, \|z_j^p\| &\le \varepsilon\, \{\|s_{j+1}\| + 2\, \|\delta_{j+1} p_j\|\} + \mathcal{O}(\varepsilon^2) \\ &\le \varepsilon\, \{3\|s_{j+1}\| + 2\|p_{j+1}\|\} + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{4.14}$$

Similarly, the computed coefficients γ_j and δ_j satisfy

$$\gamma_j = \frac{(r_j, s_j)}{(p_j, Ap_j)} + \varepsilon \zeta_j^{\gamma}, \quad \delta_j = \frac{(r_j, s_j)}{(r_{j-1}, s_{j-1})} + \varepsilon \zeta_j^{\delta}. \tag{4.15}$$

In order to bound the local terms $|\varepsilon\zeta_j^{\gamma}|$ and $|\varepsilon\zeta_j^{\delta}|$ we need following two lemmas.

LEMMA 4.1. *Consider the standard model of floating point arithmetic with machine precision* ε [23, 38], $\varepsilon n \ll 1$. *Let* L *be a nonsingular lower triangular matrix and* $M = LL^T$. *Then the numerically computed vector* s_{j+1} *is the exact solution of the perturbed system*

$$(M + \Delta M)\, s_{j+1} = r_{j+1}, \quad \|\Delta M\| \leq \frac{\varepsilon n^2}{1 - \varepsilon n} \|M\|. \tag{4.16}$$

PROOF. To prove (4.16) we use standard results of backward error analysis [23]. Using the Theorem 9.4 [23, p. 175] and the fact that we have exact Cholesky factorization of the matrix $M = LL^T$ we obtain

$$(M + \Delta M)\, s_{j+1} = r_{j+1}, \quad |\Delta M| \leq \frac{\varepsilon n}{1 - \varepsilon n} |L||L^T|$$

where $|L|$ denotes the matrix L with elements in absolute value. As shown in the proof of the Theorem 10.4 in [23, p. 206],

$$\| \, |L||L^T| \, \| \leq n \, \|M\|.$$

Summarizing,

$$\|\Delta M\| \leq \| \, |\Delta M| \, \| \leq \frac{\varepsilon n}{1 - \varepsilon n} \| \, |L||L^T| \, \| \leq n \, \frac{\varepsilon n}{1 - \varepsilon n} \, \|M\|$$

which completes the proof. □

REMARK. The assumption $M = LL^T$ is not substantial. The result similar to (4.16) and the following analysis, will remain valid also if the Cholesky decomposition of M is computed numerically, see e.g. [17].

LEMMA 4.2. *Consider the standard model of floating point arithmetic with machine precision* ε [23, 38], *let* $\varepsilon n^2 \kappa(M) \ll 1$. *The numerically computed inner product* $\mathrm{fl}[(r_j, s_j)]$ *satisfies*

$$\begin{aligned} \mathrm{fl}[(r_j, s_j)] &= (r_j, s_j) + \varepsilon \zeta_j^{rs}, \\ \varepsilon |\zeta_j^{rs}| &\leq \varepsilon \kappa(M)^{1/2} (r_j, s_j)\, \mathcal{O}(n) + \mathcal{O}(\varepsilon^2)\,, \end{aligned} \tag{4.17}$$

where $\kappa(M)$ *denotes the condition number of the matrix* M. *Moreover,* (r_j, s_j) *is bounded from below by*

$$(r_j, s_j) \geq \frac{\|r_j\| \, \|s_j\|}{\kappa(M)^{1/2}} \, \mathcal{O}(1)\,. \tag{4.18}$$

PROOF. Using (4.7), $\varepsilon |\zeta_j^{rs}|$ can be bounded as

$$\varepsilon |\zeta_j^{rs}| \leq \varepsilon n \, \|r_j\| \, \|s_j\| + \mathcal{O}(\varepsilon^2). \tag{4.19}$$

To prove (4.17), we have to relate $\|r_j\| \, \|s_j\|$ to (r_j, s_j). From (4.16) it follows

$$\begin{aligned} \|r_j\| \, \|s_j\| &\leq \|r_j\| \, \left\|(M + \Delta M)^{-1} r_j\right\| \\ &= \|r_j\| \, \left\|(I + M^{-1}\Delta M)^{-1} M^{-1} r_j\right\| \\ &\leq \|r_j\| \, \left\|M^{-1} r_j\right\| \, \|(I + M^{-1}\Delta M)^{-1}\|\,. \end{aligned} \tag{4.20}$$

Assuming $\varepsilon\, n^2\, \kappa(M) \ll 1$, it holds $\|M^{-1}\Delta M\| \ll 1$ and the matrix inverse $(I+M^{-1}\Delta M)^{-1}$ can be approximated by two terms of the Neumann expansion. Then, (4.20) changes to

$$\|r_j\|\,\|s_j\| \;\le\; \|r_j\|\,\left\|M^{-1}r_j\right\|\,C_M\,(\,1+\mathcal{O}(\|M^{-1}\Delta M\|^2))\,, \tag{4.21}$$

where

$$C_M \;\equiv\; \|\,I - M^{-1}\Delta M\,\|$$

is a constant close to one. It remains to bound the product $\|r_j\|\,\|M^{-1}r_j\|$. A simple manipulation gives

$$\begin{aligned}\|r_j\|\left\|M^{-1}r_j\right\| \;&=\; \frac{\|r_j\|\left\|M^{-1/2}M^{-1/2}r_j\right\|}{\left(M^{-1/2}r_j,\,M^{-1/2}r_j\right)}\left(r_j,M^{-1}r_j\right)\\ &\le\; \|M^{-1/2}\|\frac{\|r_j\|}{\left\|M^{-1/2}r_j\right\|}\,(r_j,M^{-1}r_j).\end{aligned} \tag{4.22}$$

Using $Ms_j + \Delta M s_j = r_j$ we get

$$\begin{aligned}\left(r_j,M^{-1}r_j\right) \;&=\; (r_j,s_j)+\left(r_j,M^{-1}\Delta M s_j\right)\\ &=\; (r_j,s_j)+\left(M^{-1/2}r_j,M^{-1/2}\Delta M s_j\right)\end{aligned}$$

and $\|r_j\|\,\|M^{-1}r_j\|$ can be bounded by

$$\begin{aligned}\|r_j\|\left\|M^{-1}r_j\right\| \;&\le\; \frac{\|M^{-1/2}\|\,\|r_j\|}{\left\|M^{-1/2}r_j\right\|}\,(r_j,s_j)\\ &\quad+\frac{\|M^{-1/2}\|\,\|r_j\|}{\left\|M^{-1/2}r_j\right\|}\left(M^{-1/2}r_j,M^{-1/2}\Delta M s_j\right)\\ &\le\; \kappa(M)^{1/2}(r_j,s_j)+\frac{\varepsilon\,n^2}{1-\varepsilon\,n}\,\kappa(M)\,\|r_j\|\,\|s_j\|\,.\end{aligned} \tag{4.23}$$

From (4.21) and (4.23) it follows

$$\begin{aligned}\|r_j\|\,\|s_j\| \;&\le\; \varepsilon\,\kappa(M)^{1/2}(r_j,s_j)\,C_M\\ &\quad+\frac{\varepsilon\,n^2}{1-\varepsilon\,n}\,\kappa(M)\,\|r_j\|\,\|s_j\|\,C_M+\mathcal{O}(\|M^{-1}\Delta M\|^2)\,.\end{aligned} \tag{4.24}$$

Defining

$$D_M \;\equiv\; C_M\left(1-\frac{\varepsilon\,n^2}{1-\varepsilon\,n}\,\kappa(M)\,C_M\right)^{-1},$$

(4.24) can be written in the form

$$\|r_j\|\,\|s_j\| \le \kappa(M)^{1/2}(r_j,s_j)\,D_M \;+\; \mathcal{O}(\|M^{-1}\Delta M\|^2)\,. \tag{4.25}$$

Since $\varepsilon\,n^2\,\kappa(M) \ll 1$ and C_M is close to one, the definition of D_M implies that D_M is close to one also. The term $\mathcal{O}(\|M^{-1}\Delta M\|^2)$ is under our assumption unimportant and will not be further explicitly considered. Finally, (4.25) gives

$$\|r_j\|\,\|s_j\| \le \kappa(M)^{1/2}(r_j,s_j)\,\mathcal{O}(1)\,, \tag{4.26}$$

where $\mathcal{O}(1)$ stands for a number close to one. (4.17) follows immediately from (4.26) and (4.19). Dividing (4.26) by $\kappa(M)^{1/2}$ gives (4.18), which finishes the proof. □

Assuming $\varepsilon\, n^2\, \kappa(M) \ll 1$, the local term $\varepsilon\zeta_j^\delta$ is bounded, according to (4.4), (4.7) and (4.17), by

$$\varepsilon\big|\zeta_j^\delta\big| \le \varepsilon\, \frac{(r_j, s_j)}{(r_{j-1}, s_{j-1})}\, \kappa(M)^{1/2}\, \mathcal{O}(n) + \mathcal{O}(\varepsilon^2). \tag{4.27}$$

Using (4.5)–(4.8) and $\|A\|\|p_j\|^2/(p_j, Ap_j) \le \kappa(A)$,

$$\begin{aligned} \mathrm{fl}[(p_j, Ap_j)] &= (p_j, Ap_j) + \varepsilon\, \|Ap_j\|\|p_j\|\mathcal{O}(n) + \varepsilon\, \|A\|\|p_j\|^2\mathcal{O}(c) + \mathcal{O}(\varepsilon^2) \\ &= (p_j, Ap_j)\big(1 + \varepsilon\, \kappa(A)\mathcal{O}(n+c)\big) + \mathcal{O}(\varepsilon^2). \end{aligned}$$

Assuming $\varepsilon(n+c)\,\kappa(A) \ll 1$, the local roundoff $\varepsilon\zeta_j^\gamma$ is bounded by

$$\varepsilon\big|\zeta_j^\gamma\big| \le \varepsilon\, (\kappa(A) + \kappa(M)^{1/2})\frac{(r_j, s_j)}{(p_j, Ap_j)}\, \mathcal{O}(n+c) + \mathcal{O}(\varepsilon^2). \tag{4.28}$$

It is well known that in finite precision arithmetic the true residual $b - Ax_j$ differs from the recursively updated residual vector r_j,

$$r_j = b - Ax_j - \varepsilon f_j. \tag{4.29}$$

This topic was studied in [36] and [19]. The results can be written in the following form

$$\|\varepsilon f_j\| \le \varepsilon\, \|A\|\, (\|x\| + \max_{0\le i\le j} \|x_i\|)\, \mathcal{O}(jc), \tag{4.30}$$

$$\|r_j\| = \|b - Ax_j\|\, (1 + \varepsilon F_j), \tag{4.31}$$

where εF_j is bounded by

$$|\varepsilon F_j| = \frac{|\|r_j\| - \|b - Ax_j\||}{\|b - Ax_j\|} \le \frac{\|r_j - (b - Ax_j)\|}{\|b - Ax_j\|} = \frac{\varepsilon\|f_j\|}{\|b - Ax_j\|}. \tag{4.32}$$

Rounding errors affect results of PCG computations in two main ways: they delay convergence and limit the ultimate attainable accuracy. Here we are primarily interested in estimating the convergence rate. We therefore assume that the final accuracy level has not been reached yet and εf_j is, in comparison to the size of the true and iterative residuals, small. In the subsequent text we will relate the numerical inaccuracies to the A-norm of the error $\|x - x_j\|_A$. The following inequalities derived from (4.32) will prove useful,

$$\lambda_1^{1/2}\, \|x - x_j\|_A\, (1 + \varepsilon\, F_j) \le \|r_j\| \le \lambda_n^{1/2}\, \|x - x_j\|_A\, (1 + \varepsilon\, F_j). \tag{4.33}$$

Similarly as in the ordinary CG (see [18], [21]) we can argue that the monotonicity of the A-norm is in PCG preserved (with small additional inaccuracy) also in finite precision computations. Using this fact we get for $j \ge i$

$$\varepsilon\, \frac{\|r_j\|}{\|r_i\|} \;\le\; \varepsilon\, \frac{\lambda_n^{1/2}}{\lambda_1^{1/2}} \cdot \frac{\|x - x_j\|_A}{\|x - x_i\|_A} \cdot \frac{(1 + \varepsilon\, F_j)}{(1 + \varepsilon\, F_i)} \;\le\; \varepsilon\, \kappa(A)^{1/2} + \mathcal{O}(\varepsilon^2). \tag{4.34}$$

This bound will be used later.

4.2 Finite precision analysis – basic identity.

We show that the ideal (exact precision) identity (3.6) changes numerically to

$$\|x - x_j\|_A^2 = \nu_{j,d} + \|x - x_{j+d}\|_A^2 + \widetilde{\nu}_{j,d} \tag{4.35}$$

where $\widetilde{\nu}_{j,d}$ is as small as it can be. We once more emphasize that the difference between (3.6) and (4.35) *is not trivial*. The ideal and numerical counterparts of each individual term in these identities may be orders of magnitude different! Due to the facts that rounding errors in computing $\nu_{j,d}$ numerically from the quantities γ_i and $\mathrm{fl}[(r_i, s_i)]$ are negligible and that $\widetilde{\nu}_{j,d}$ will be related to $\varepsilon \|x - x_j\|_A$, (4.35) will justify the estimate $\nu_{j,d}$ in finite precision computations.

In order to get the desired form leading to (4.35), we will develop the right hand side of (4.3). In this derivation we will rely on local properties (4.9)–(4.11) and (4.15)–(4.16) of the finite precision PCG recurrences.

Using (4.9), the first term on the right hand side of (4.3) can be written as

$$\begin{aligned} \|x_{j+1} - x_j\|_A^2 &= (\gamma_j p_j + \varepsilon\, z_j^x)^T A(\gamma_j p_j + \varepsilon\, z_j^x) \\ &= \gamma_j^2\, (p_j, Ap_j) + 2\varepsilon\, \gamma_j (p_j, Az_j^x) + \mathcal{O}(\varepsilon^2) \\ &= \gamma_j\, (p_j, Ap_j) + 2\varepsilon\, (x_{j+1} - x_j)^T A z_j^x + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{4.36}$$

Similarly, the second term on the right hand side of (4.3) transforms, using (4.29), to the form

$$\begin{aligned} 2\, (x - x_{j+1})^T A(x_{j+1} - x_j) &= 2\, (r_{j+1} + \varepsilon\, f_{j+1})^T (x_{j+1} - x_j) \\ &= 2\, r_{j+1}^T (x_{j+1} - x_j) + 2\varepsilon\, f_{j+1}^T (x_{j+1} - x_j). \end{aligned} \tag{4.37}$$

Combining (4.3), (4.36) and (4.37),

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 = \gamma_j^2\, (p_j, Ap_j) + 2\, r_{j+1}^T (x_{j+1} - x_j) \\ + 2\varepsilon\, (f_{j+1} + Az_j^x)^T (x_{j+1} - x_j) + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{4.38}$$

Substituting for γ_j from (4.15), the first term in (4.38) can be written as

$$\begin{aligned} \gamma_j^2\, (p_j, Ap_j) &= \gamma_j (r_j, s_j) + \varepsilon\, \gamma_j\, (p_j, Ap_j)\, \zeta_j^\gamma \\ &= \gamma_j (r_j, s_j) + \varepsilon\, \gamma_j (r_j, s_j) \left\{ \zeta_j^\gamma \frac{(p_j, Ap_j)}{(r_j, s_j)} \right\}. \end{aligned}$$

Consequently, the difference between the squared A-norms of the error in the consecutive steps can be written in the form convenient for the further analysis

$$\begin{aligned} \|x - x_j\|_A^2 - \|x - x_{j+1}\|_A^2 \;=\; \gamma_j (r_j, s_j) + \varepsilon\, \gamma_j (r_j, s_j) \left\{ \zeta_j^\gamma \frac{(p_j, Ap_j)}{(r_j, s_j)} \right\} \\ + 2\, r_{j+1}^T (x_{j+1} - x_j) + 2\varepsilon\, (f_{j+1} + Az_j^x)^T (x_{j+1} - x_j) + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{4.39}$$

The goal of the following analysis is to show that until $\|x - x_j\|_A$ reaches its ultimate attainable accuracy level, the terms on the right hand side of (4.39)

are, except for $\gamma_j(r_j, s_j)$ insignificant. Bounding the second term will not represent a problem. The norm of the difference $x_{j+1} - x_j = (x - x_j) - (x - x_{j+1})$ is bounded by $2\|x - x_j\|_A/\lambda_1^{1/2}$, and therefore the size of the fourth term is proportional to $\varepsilon\,\|x - x_j\|_A$. The third term is related to the line-search principle. Ideally (in exact arithmetic), the $(j+1)$-th residual $\hat{r}_{j+1}$ is orthogonal to the difference between the $(j+1)$-th and j-th approximation $\hat{x}_{j+1} - \hat{x}_j$ (which is a multiple of the j-th direction vector $\hat{p}_j$). This is equivalent to the line-search: ideally, in terms of the transformed quantities used in Algorithm 2, the $(j+1)$-th PCG approximation minimizes the A-norm of the error along the line determined by the j-th approximation and the j-th direction vector. Here the term $r_{j+1}^T(x_{j+1} - x_j)$, with r_{j+1}, x_j and x_{j+1} computed numerically, examines how closely the line-search holds in finite precision arithmetic. In fact, bounding the local orthogonality $r_{j+1}^T(x_{j+1} - x_j)$ represents the technically most difficult part of the remaining analysis.

4.3 Local orthogonality.

Since the classical work of Paige it is well known that in the three-term Lanczos recurrence local orthogonality is preserved close to the machine epsilon (see [31]). We will derive an analogue of this for the PCG algorithm, and state it as an independent result.

The local orthogonality term $r_{j+1}^T(x_{j+1} - x_j)$ can be written in the form

$$r_{j+1}^T(x_{j+1} - x_j) = r_{j+1}^T(\gamma_j p_j + \varepsilon\, z_j^x) = \gamma_j(r_{j+1}, p_j) + \varepsilon\,(r_{j+1}, z_j^x). \tag{4.40}$$

Using the bound

$$\|r_{j+1}\| \le \lambda_n^{1/2}\|x - x_{j+1}\|_A(1 + \varepsilon\, F_{j+1}) \le \lambda_n^{1/2}\|x - x_j\|_A(1 + \varepsilon\, F_{j+1})\,,$$

see (4.33), the size of the second term in (4.40) is proportional to $\varepsilon\,\|x - x_j\|_A$. The main step consist of showing that the term (r_{j+1}, p_j) is sufficiently small. Scalar multiplying the recurrence (4.10) for r_{j+1} by the vector p_j gives (using (4.11) and (4.15))

$$\begin{aligned}
(p_j, r_{j+1}) &= (p_j, r_j) - \gamma_j(p_j, Ap_j) + \varepsilon\,(p_j, z_j^r) \\
&= (s_j + \delta_j p_{j-1} + \varepsilon\, z_{j-1}^p)^T r_j \\
&\quad -\left(\frac{(r_j, s_j)}{(p_j, Ap_j)} + \varepsilon\,\zeta_j^\gamma\right)(p_j, Ap_j) + \varepsilon\,(p_j, z_j^r) \\
&= \delta_j\,(p_{j-1}, r_j) + \varepsilon\left\{(r_j, z_{j-1}^p) - \zeta_j^\gamma(p_j, Ap_j) + (p_j, z_j^r)\right\}. \qquad (4.41)
\end{aligned}$$

Denoting

$$G_j \equiv (r_j, z_{j-1}^p) - \zeta_j^\gamma(p_j, Ap_j) + (p_j, z_j^r), \tag{4.42}$$

the identity (4.41) is

$$(p_j, r_{j+1}) = \delta_j\,(p_{j-1}, r_j) + \varepsilon\, G_j. \tag{4.43}$$

Recursive application of (4.43) for $(p_{j-1}, r_j), \ldots, (p_1, r_2)$ with $(p_0, r_1) = (p_0, r_0) - \gamma_0 (p_0, Ap_0) + \varepsilon (p_0, z_0^r) = \varepsilon \left\{ - \zeta_0^\gamma (s_0, As_0) + (s_0, z_0^r) \right\} \equiv \varepsilon G_0$, gives

$$(p_j, r_{j+1}) = \varepsilon G_j + \varepsilon \sum_{i=1}^{j} \Big(\prod_{k=i}^{j} \delta_k \Big) G_{i-1}. \tag{4.44}$$

Since

$$\varepsilon \prod_{k=i}^{j} \delta_k = \varepsilon \prod_{k=i}^{j} \frac{(r_k, s_k)}{(r_{k-1}, s_{k-1})} + \mathcal{O}(\varepsilon^2) = \varepsilon \frac{(r_j, s_j)}{(r_{i-1}, s_{i-1})} + \mathcal{O}(\varepsilon^2),$$

we can express (4.44) as

$$(p_j, r_{j+1}) = \varepsilon (r_j, s_j) \sum_{i=0}^{j} \frac{G_i}{(r_i, s_i)} + \mathcal{O}(\varepsilon^2). \tag{4.45}$$

Using (4.42),

$$\frac{|G_i|}{(r_i, s_i)} \leq \frac{\|r_i\| \|z_{i-1}^p\|}{(r_i, s_i)} + |\zeta_i^\gamma| \frac{(p_i, Ap_i)}{(r_i, s_i)} + \frac{\|p_i\| \|z_i^r\|}{(r_i, s_i)}. \tag{4.46}$$

When bounding the first and the last terms on the right hand side of (4.46), we will use the inequality (4.18) proved in Lemma 4.2. From (4.14) it follows

$$\varepsilon \frac{\|r_i\| \|z_{i-1}^p\|}{(r_i, s_i)} \leq \varepsilon \kappa(M)^{1/2} \left\{ 3 + 2 \frac{\|p_i\|}{\|s_i\|} \right\} \mathcal{O}(1) \; + \; \mathcal{O}(\varepsilon^2). \tag{4.47}$$

Using (4.28),

$$\varepsilon |\zeta_i^\gamma| \frac{(p_i, Ap_i)}{(r_i, s_i)} \leq \varepsilon \left(\kappa(A) + \kappa(M)^{1/2} \right) \mathcal{O}(n + c) + \mathcal{O}(\varepsilon^2). \tag{4.48}$$

The last part of (4.46) is bounded using (4.13) and (4.18)

$$\begin{aligned}
\varepsilon \frac{\|p_i\| \|z_i^r\|}{(r_i, s_i)} \; &\leq \; \varepsilon \left\{ \kappa(M)^{1/2} \frac{\|p_i\| \|r_i\|}{\|s_i\| \|r_i\|} \mathcal{O}(1) \right\} \\
&\quad + \varepsilon \left\{ 2 \gamma_i \frac{\|p_i\| \|Ap_i\|}{(r_i, s_i)} + c \gamma_i \frac{\|p_i\| \|A\| \|p_i\|}{(r_i, s_i)} \right\} + \mathcal{O}(\varepsilon^2) \\
&= \varepsilon \left\{ \kappa(M)^{1/2} \frac{\|p_i\|}{\|s_i\|} \mathcal{O}(1) \right\} \\
&\quad + \varepsilon \left\{ 2 \frac{\|p_i\| \|Ap_i\|}{(p_i, Ap_i)} + c \frac{\|A\| \|p_i\|^2}{(p_i, Ap_i)} \right\} + \mathcal{O}(\varepsilon^2) \\
&\leq \varepsilon \left\{ \kappa(M)^{1/2} \frac{\|p_i\|}{\|s_i\|} \mathcal{O}(1) + (2 + c) \kappa(A) \right\} + \mathcal{O}(\varepsilon^2),
\end{aligned} \tag{4.49}$$

where

$$\begin{aligned}
\varepsilon \frac{\|p_i\|}{\|s_i\|} \; &\leq \; \varepsilon \frac{\|s_i\| + \delta_i \|p_{i-1}\|}{\|s_i\|} + \mathcal{O}(\varepsilon^2) \\
&\leq \varepsilon \left\{ 1 + \delta_i \frac{\|s_{i-1}\|}{\|s_i\|} \frac{\|p_{i-1}\|}{\|s_{i-1}\|} \right\} + \mathcal{O}(\varepsilon^2).
\end{aligned} \tag{4.50}$$

Recursive application of (4.50) for $\|p_{i-1}\|/\|s_{i-1}\|$, $\|p_{i-2}\|/\|s_{i-2}\|$, $\ldots$, $\|p_1\|/\|s_1\|$ with $\|p_0\|/\|s_0\| = 1$ gives

$$\begin{aligned}\varepsilon\frac{\|p_i\|}{\|s_i\|} &\leq \varepsilon\left\{1+\frac{(s_i,r_i)}{(s_{i-1},r_{i-1})}\frac{\|s_{i-1}\|}{\|s_i\|}+\cdots+\frac{(s_i,r_i)}{(s_0,r_0)}\frac{\|s_0\|}{\|s_i\|}\right\}+\mathcal{O}(\varepsilon^2)\\ &\leq \varepsilon\left\{1+\frac{\|r_i\|\|s_{i-1}\|}{(s_{i-1},r_{i-1})}+\cdots+\frac{\|r_i\|\|s_0\|}{(s_0,r_0)}\right\}+\mathcal{O}(\varepsilon^2)\\ &\leq \varepsilon\left\{1+\kappa(M)^{1/2}\frac{\|r_i\|}{\|r_{i-1}\|}+\cdots+\kappa(M)^{1/2}\frac{\|r_i\|}{\|r_0\|}\right\}\mathcal{O}(1)+\mathcal{O}(\varepsilon^2).\end{aligned}$$

The size of $\varepsilon\,\|r_i\|/\|r_k\|$, $i \geq k$ is, according to (4.34), less or equal than the value $\varepsilon\,\kappa(A)^{1/2}+\mathcal{O}(\varepsilon^2)$. Consequently,

$$\varepsilon\,\frac{\|p_i\|}{\|s_i\|} \leq \varepsilon\,\{1+i\,\kappa(A)^{1/2}\kappa(M)^{1/2}\}\,\mathcal{O}(1)+\mathcal{O}(\varepsilon^2). \tag{4.51}$$

Denote

$$\kappa(A,M) \equiv \max(\kappa(A),\kappa(M)\kappa(A)^{1/2}).$$

Summarizing (4.47), (4.48), (4.49) and (4.51), the ratio $\varepsilon\,|G_i|/(r_i,s_i)$ is bounded as

$$\varepsilon\,\frac{|G_i|}{(r_i,s_i)} \leq \varepsilon\,\kappa(A,M)\,\mathcal{O}(8+3c+2n+3i)+\mathcal{O}(\varepsilon^2). \tag{4.52}$$

Combining this result with (4.45) proves the following theorem.

THEOREM 4.3. *Let* $\varepsilon\,(n+c)\,\kappa(A) \ll 1$, $\varepsilon\,n^2\,\kappa(M) \ll 1$. *Then the local orthogonality between the direction vectors and the iteratively computed residuals is in the finite precision implementation of the preconditioned conjugate gradient method* (4.9)–(4.11) *and* (4.15)–(4.16) *bounded by*

$$|(p_j,r_{j+1})| \leq \varepsilon\,(r_j,s_j)\,\kappa(A,M)\,\mathcal{O}((j+1)(8+3c+2n+3j))+\mathcal{O}(\varepsilon^2) \tag{4.53}$$

where

$$\kappa(A,M) \equiv \max\left(\kappa(A),\kappa(M)\kappa(A)^{1/2}\right).$$

4.4 Finite precision analysis – conclusions.

We now return to (4.39) and finalize our discussion. Using (4.40) and (4.45),

$$\begin{aligned}\|x-x_j\|_A^2-\|x-x_{j+1}\|_A^2 =\ & \gamma_j(r_j,s_j)\\ &+\varepsilon\,\gamma_j(r_j,s_j)\left\{\zeta_j^{\gamma}\frac{(p_j,Ap_j)}{(r_j,s_j)}+2\sum_{i=0}^{j}\frac{G_i}{(r_j,s_j)}\right\}\\ &+2\varepsilon\left\{(f_{j+1}+Az_j^x)^T(x_{j+1}-x_j)+(r_{j+1},z_j^x)\right\}+\mathcal{O}(\varepsilon^2).\end{aligned} \tag{4.54}$$

The term

$$E_j^{(1)} \equiv \varepsilon\left\{\zeta_j^{\gamma}\frac{(p_j,Ap_j)}{(r_j,s_j)}+2\sum_{i=0}^{j}\frac{G_i}{(r_j,s_j)}\right\}$$

is bounded using (4.28) and (4.52),

$$|E_j^{(1)}| \leq \varepsilon\kappa(A,M)\,\mathcal{O}\big(2n+2c+2(j+1)(8+3c+2n+3j)\big) + \mathcal{O}(\varepsilon^2). \tag{4.55}$$

We write the remaining term on the right hand side of (4.54) proportional to ε

$$2\varepsilon\left\{(f_{j+1}+Az_j^x)^T(x_{j+1}-x_j)+(r_{j+1},z_j^x)\right\} \equiv \|x-x_j\|_A\, E_j^{(2)} \tag{4.56}$$

where

$$|E_j^{(2)}| = 2\varepsilon\left|(f_{j+1}+Az_j^x)^T\left(\frac{x_{j+1}-x+x-x_j}{\|x-x_j\|_A}\right)+\frac{(r_{j+1},z_j^x)}{\|x-x_j\|_A}\right|$$

$$\leq 2\varepsilon\left\{2\left(\|f_{j+1}\|\lambda_1^{-1/2}+\|A\|^{1/2}\|z_j^x\|\right)+\|A\|^{1/2}\|z_j^x\|\right\}. \tag{4.57}$$

With (4.30) and (4.12),

$$\begin{aligned}|E_j^{(2)}| &\leq 4\varepsilon\|A\|^{1/2}\kappa(A)^{1/2}(\|x\|+\max_{0\leq i\leq j+1}\|x_i\|)\,\mathcal{O}(jc)\\ &\quad+5\|A\|^{1/2}\varepsilon(3\|x_j\|+2\|x_{j+1}\|)+\mathcal{O}(\varepsilon^2)\\ &\leq \varepsilon\|A\|^{1/2}\kappa(A)^{1/2}(\|x\|+\max_{0\leq i\leq j+1}\|x_i\|)\,\mathcal{O}(4jc+25)+\mathcal{O}(\varepsilon^2).\end{aligned} \tag{4.58}$$

Finally, using the fact that the monotonicity of the A-norm is with small additional inaccuracy preserved also in finite precision PCG computations (see also the discussion following (4.33)), we obtain the finite precision analogue of (3.6), which is formulated as a theorem.

THEOREM 4.4. *Let $\varepsilon\,(n+c)\,\kappa(A) \ll 1$, $\varepsilon\,n^2\,\kappa(M) \ll 1$. Then the PCG approximate solutions computed in finite precision arithmetic satisfy*

$$\|x-x_j\|_A^2 - \|x-x_{j+d}\|_A^2 = \nu_{j,d} + \nu_{j,d}\,E_{j,d}^{(1)} + \|x-x_j\|_A\,E_{j,d}^{(2)} + O(\varepsilon^2), \tag{4.59}$$

where

$$\nu_{j,d} = \sum_{i=j}^{j+d-1}\gamma_i\,(r_i,s_i). \tag{4.60}$$

The terms due to rounding errors are bounded by

$$\begin{aligned}|E_{j,d}^{(1)}| &\leq \varepsilon\,\kappa(A,M)\,p^{(1)}(n,d)+O(\varepsilon^2),\\ |E_{j,d}^{(2)}| &\leq \varepsilon\,\|A\|^{1/2}\kappa(A)^{1/2}\,(\|x\|+\max_{0\leq i\leq j+1}\|x_i\|)\,p^{(2)}(n,d)+O(\varepsilon^2),\end{aligned} \tag{4.61}$$

where

$$\kappa(A,M) \equiv \max\left(\kappa(A),\kappa(M)\kappa(A)^{1/2}\right),$$

$p^{(1)}(n,d)$ and $p^{(2)}(n,d)$ represent small degree polynomials in n and d independent of any other variables.

Based on the assumptions we consider $|E_{j,d}^{(1)}| \ll 1$. Then, assuming that the A-norm of the error reasonably decreases, the numerically computed value $\nu_{j,d}$

gives a good estimate for the A-norm of the error $\|x - x_j\|_A^2$ until

$$\|x - x_j\|_A \, |E_{j,d}^{(2)}| \ll \|x - x_j\|_A^2,$$

which is equivalent to

$$\|x - x_j\|_A \gg |E_{j,d}^{(2)}|. \tag{4.62}$$

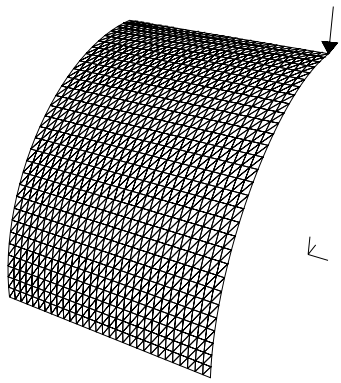
The quantity $E_{j,d}^{(2)}$ represents various terms. Its upper bound is, apart from $\kappa(A)^{1/2}$, which comes into play as an effect of the worst-case rounding error analysis, linearly dependent on an upper bound for $\|x - x_0\|_A$. The value of $E_{j,d}^{(2)}$ is (similar to terms or constants in any other rounding error analysis) not important. What is important is the following possible interpretation of (4.62): until $\|x - x_j\|_A$ reaches a level close to $\varepsilon\|x - x_0\|_A$, the computed estimate $\nu_{j,d}^{1/2}$ must work.

Please note that $\nu_{j,d}$ represents here the exact value determined from the computed inputs γ_i, r_i and s_i. In fact, we should consider the computed value $\mathrm{fl}[\nu_{j,d}]$. Additional rounding errors in evaluating the formula (4.60) are, however, negligible in comparison to the other rounding error terms in (4.59), and need not be considered here.

5 Numerical experiments.

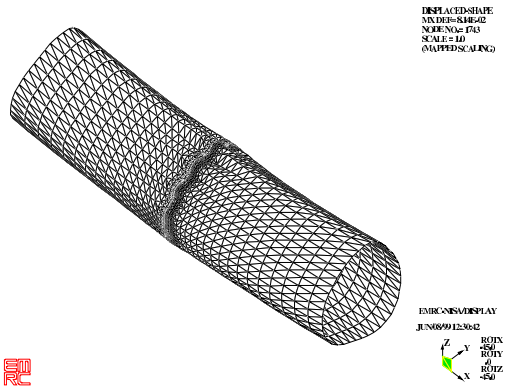
We test our theoretical results on three linear systems with a symmetric positive definite matrix A. The first two systems (by R. Kouhia) arise from cylindrical shell modeling. The matrices are large and sparse, and PCG represents a natural choice for solving the systems in practical computations. The third system (by P. Benner) appears in large-scale control problems. PCG is not used for practical solution of the last (rather small) system. We use it here for illustration of how the estimate of the A-norm of the error works for this type of problem. We describe the problems in more detail.

The system `s3dkt3m2`. The collection Cylshell (by R. Kouhia) from the electronic library Matrix Market [25] contains matrices that represent low order finite element discretization of a shell element test, the pinched cylinder. An illustration of the mesh for this problem provided by R. Kouhia is given below.



In our experiments we use the matrix `s3dkt3m2` of the order $n = 90449$. The matrix has $\mathrm{nnz}(A) = 1921955$ nonzero elements, and the condition number $\kappa(A) = 3.62\mathrm{e}{+11}$. Only the last element of the right-hand side vector b is nonzero, which corresponds to the given physical problem (for more details see [24] and the references in [24]). The preconditioner was determined by incomplete Cholesky decomposition with no fill-in.

The system `tube`. The second system is given at the R. Kouhia's homepage `http://www.hut.fi/~kouhia/` (the system `tube1-2`). The tube is a cylindrical shell with the constant wall thickness, loaded with an axial stress distribution at both ends. The mesh is refined at the center, and it is almost uniform towards the ends.



The order of the matrix A is $n = 21498$, $\mathrm{nnz}(A) = 894490$. The factor L of the preconditioner M is determined by the incomplete Cholesky decomposition with the drop tolerance 1e–5, $\mathrm{nnz}(L) = 4384369$.

The system `stahl`. We consider the problem of optimal cooling of steel profile, that arises, e.g. in a rolling mill when different steps in the production process require different temperatures of the raw material. The problem is modeled using a boundary control (given by the temperature of the cooling fluid) for a heat-diffusion process described by the linearized heat equations. This leads to the Lyapunov equations that are solved by the ADI iterations. For more detail about this problem see [9]. We test the proposed estimates on the system from the initial step of the ADI iteration. The matrix is of the order $n = 5177$, $\kappa(A) = 1.56\mathrm{e}{+05}$, $\mathrm{nnz}(A) = 35241$. The system is preconditioned by incomplete Cholesky decomposition with no fill-in.

In all experiments we use the initial approximation $x_0 = 0$. We do not tune the preconditioner for the best performance; our aim is to demonstrate the behaviour of the estimate of the A-norm of the error in practical computations. The substitutes for the exact solutions x used in the figures are for each system computed in two steps: 1. We apply PCG to the system and iterate until ultimate level of accuracy is reached (the norm of true and recursive residuals start to differ). 2. We apply PCG to the system for the second time, with the initial approximation given by the approximate solution computed in the first step. In this

way, we obtain approximate solutions that represent for our purpose sufficiently accurate approximations to the exact solutions x. Our numerical experiments showed that even for the first step the obtained residual norms were comparable with that ones obtained by the direct Cholesky decomposition solver. After the second step the residual norms further decreased, but less than by a factor of 10.

In experiments with the system `s3dkt3m2` we use a Fortran program CG6 provided us by M. Tůma. The other two systems are solved using our implementation of PCG in Matlab 6.5; we use the Matlab-function `cholinc` to determine the incomplete Cholesky decomposition of the matrix A. All experiments were performed on a AMD Athlon XP 2100+ personal computer with machine precision $\varepsilon \sim 10^{-16}$.

5.1 Estimates for the A-norm of the error.

In the first numerical experiment we test the estimate $\nu_{j,d}^{1/2}$ of the A-norm of the error and the estimate $\varrho_{j,d}^{1/2}$ of the relative A-norm of the error in PCG applied to the three systems described above. The results are presented in the figures Figure 5.1 (`s3dkt3m2`), Figure 5.2 (`tube`) and Figure 5.3 (`stahl`). All three figures consist of two parts. The left part includes various convergence characteristics: the A-norm of the error $\|x - x_j\|_A$ (dashed line), its estimate $\nu_{j,d}^{1/2}$ for some particular value of the parameter d (bold solid line), the residual norm $\|b - Ax_j\|$ (dash-dotted line) and the normwise backward error $\|b - Ax_j\|/(\|A\|\,\|x_j\| + \|b\|)$ (dotted line). In the right part of the figure we plot the relative A-norm of the error $\|x - x_j\|_A/\|x\|_A$ (dashed line) and its estimates $\varrho_{j,d}^{1/2}$ for different values of d (solid lines). The bold line corresponds to the same value of d as the bold line in the left part of the figure.

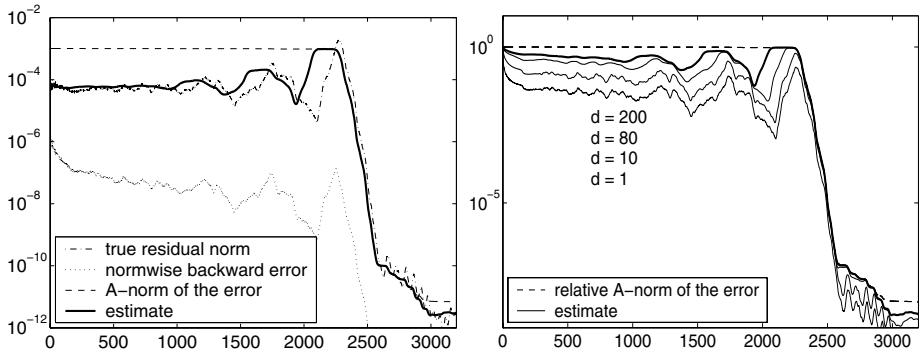


Figure 5.1: The system `s3dkt3m2`. In an extremal case of very slow PCG convergence the estimate $\nu_{j,d}^{1/2}$ can significantly underestimate the actual A-norm of the error (left part). The estimate $\varrho_{j,d}^{1/2}$ of the relative A-norm of the error (right part) is in general much tighter than the estimate of the A-norm of the error.

Figure 5.1 (`s3dkt3m2`), *left part.* We start with the most difficult situation when the A-norm of the error (dashed line) almost stagnates for many steps

(here up to the iteration ~ 2400). Then the estimate $\nu_{j,d}^{1/2}$ (bold solid line) can give a poor information about the actual A-norm of the error. The values of $\|x - x_j\|_A$ and $\nu_{j,d}^{1/2}$, can significantly differ even for a considerably large value of the parameter d (here $d = 200$). Please notice that the situation just described is not frequent in practical computations. It corresponds to an extremely slow convergence of PCG, i.e. to the case of very difficult problem which is hard to precondition. We have chosen such problem on purpose to show the possible drawback of the proposed error estimator. We emphasize that this situation represents an extremal case. Typical situation is demonstrated below on Figure 5.2 (`tube`) and Figure 5.3 (`stahl`). As soon as the convergence takes place (around the iteration 2400), we get a tight lower bound for the A-norm of the error.

In CG, we often observe a close correlation between the behaviour of the residual norm and the estimate $\nu_{j,d}^{1/2}$ for small values of d. This is a consequence of the fact that in ordinary CG the coefficients γ_j usually oscillate around some value and, apart from this oscillations, the behaviour of $\|r_j\|$ determines the behaviour of $\nu_{j,d}^{1/2}$. Similar phenomenon appears also in the PCG iterations. Here $\nu_{j,d}$ and $(r_j, M^{-1}r_j)$ (the squared M^{-1}-norm of the residual r_j) are correlated for small values of d. The M^{-1}-norm of the residual r_j frequently behaves in practical computation similarly as a constant multiple of the Euclidean norm of the residual. Then the correlation between $\|r_j\|$ and $\nu_{j,d}^{1/2}$ is observed also in the PCG iterations. For larger values of d, however, there is, in general, no correlation between the behaviour of $\|r_j\|$ and $\nu_{j,d}^{1/2}$. In the left part of Figure 5.1 (where $d = 200$) we clearly see periods of decrease of $\|r_j\|$ with simultaneous increase of $\nu_{j,d}^{1/2}$, and vice versa.

By the dotted line we plot the normwise backward error. After the convergence becomes steady, the values of $\|x_j\|$ typically stabilize. The residual norm and the normwise backward error are then in a strong correlation. Until then, however, both characteristics can behave differently. This fact is demonstrated by the convergence curves in the first 500 iterations; the backward error decreases while the residual norm stagnates.

Figure 5.1 (`s3dkt3m2`), *right part.* In the right part of the Figure 5.1 we plot the relative A-norm of the error (3.8) (dashed line) and its estimate $\varrho_{j,d}^{1/2}$ for $d = 1$, $d = 10$, $d = 80$ (solid lines) and $d = 200$ (bold solid line). The estimate $\varrho_{j,1}^{1/2}$, and sometimes even $\varrho_{j,10}^{1/2}$, $\varrho_{j,80}^{1/2}$ and $\varrho_{j,200}^{1/2}$, are not tight when the A-norm of the error almost stagnates. In the other cases $\varrho_{j,1}^{1/2}$ as well as the bounds for the larger d are close to the considered convergence curve. By the bold solid line we plot the estimate for $d = 200$. In comparison to the left part of the Figure 5.1, the estimate of the relative A-norm of the error gives better results (it is closer to the approximated curve) than the estimate of the absolute A-norm of the error.

Figure 5.2 (`tube`), *left part.* When the A-norm of the error (dashed line) decreases rapidly (iterations $350 - 400$), we can not visually distinguish this quantity from its estimate $\nu_{j,d}^{1/2}$ (bold solid line). On the other hand, when the convergence is slow (iterations $1 - 350$), the difference between the actual A-norm of the error and its estimate is observable but insignificant. The normwise backward

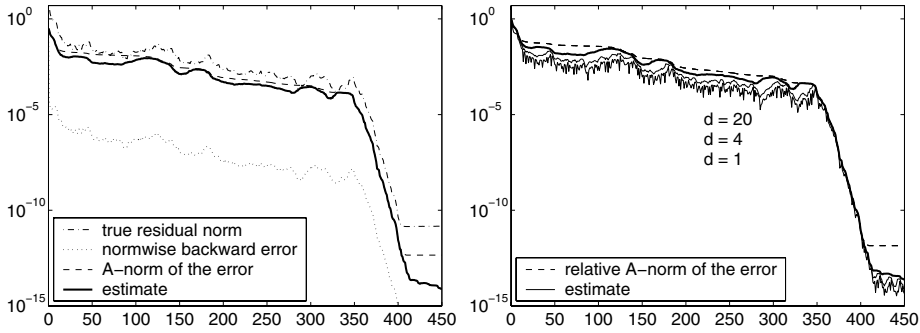


Figure 5.2: The system `tube`. Even a slow decrease of the A-norm of the error is sufficient for obtaining a satisfactory value of the estimate $\nu_{j,d}^{1/2}$ of the A-norm of the error. The erratic behaviour for $d = 1$ is caused by the oscillations of the coefficients γ_j (right part). By increasing the value of d, the curves are more smooth and closer to the relative A-norm of the error.

error (dotted line) behaves similarly, apart from the difference in magnitude, as the residual norm (dash dotted line).

Figure 5.2 (`tube`), *right part.* The right part of the Figure 5.2 contains the curve of the relative A-norm of the error (dashed line) and its estimates for $d = 1$, $d = 4$ (solid lines) and $d = 20$ (bold solid line). For $d = 1$, the curve of the estimate is erratic. The irregularity of the curve is due to the oscillations of the coefficients γ_j. The estimate $\varrho_{j,1}^{1/2}$ does not differ from the actual relative A-norm of the error for more than a single order of magnitude, although the convergence is in iterations 1–350 slow. Increasing d provides a very good estimate throughout the whole computation.

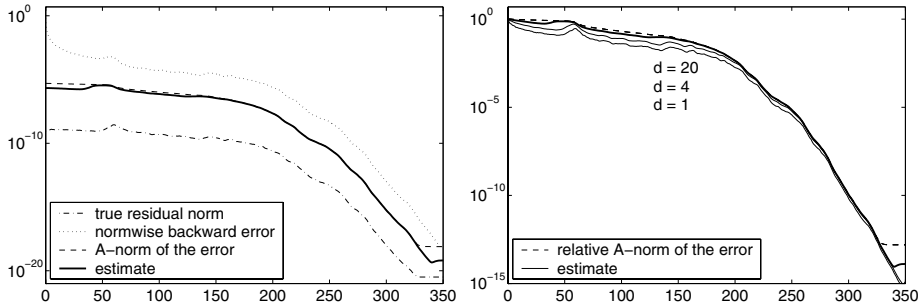


Figure 5.3: The system `stahl`. The estimates for the absolute and relative A-norm of the error are tight throughout the whole computation.

Figure 5.3 (`stahl`), *left part.* The preconditioning by incomplete Cholesky decomposition represents here a very good choice; the convergence of the A-norm of the error (dashed line) is fast during the whole computation and the estimate (bold solid line) for the parameter $d = 20$ describes very well the convergence curve.

Figure 5.3 (`stahl`), *right part.* The estimates of the relative A-norm of the error give a satisfactory information about the convergence also for small values of d ($d = 1,\ d = 4$).

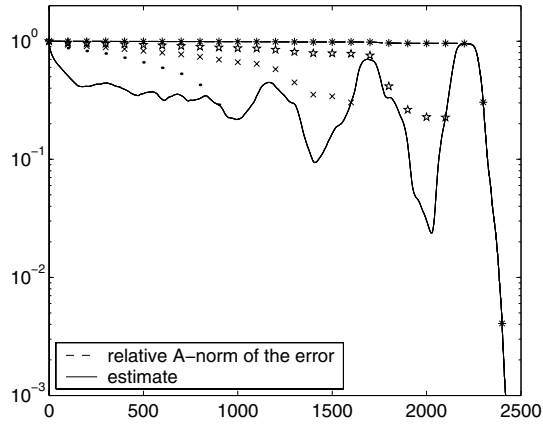


Figure 5.4: The system `s3dkt3m2`. The relative A-norm of the error (dashed line), the estimate of the relative A-norm of the error with $d = 100$ (solid line) and the curves reconstructed at iterations 1000 (dots), 1700 (x-marks), 2200 (pentagrams) and 2500 (stars).

5.2 Reconstruction of the convergence curve.

Up to now we estimated the A-norm of the error at the iteration step j at the price of running d extra steps, and we considered d to be fixed. The simple form of the estimate $\nu_{j,d}$, see (3.6), (3.7) enables at the given iteration step j updating of the estimates of the A-norm of the error at the steps $j-d, j-2d, \ldots$ at a negligible cost. Indeed, assuming, for simplicity of exposition, that j is a multiple of the chosen d ($j \bmod d = 0$), the identity (3.6) gives

$$\|x - x_{j-id}\|_A^2 \;=\; \sum_{l=0}^{i} \nu_{j-ld,d} + \|x - x_{j+d}\|_A^2, \quad i = 0, 1, \ldots . \tag{5.1}$$

In this way,

$$\nu_{j-id,(i+1)d}^{1/2} \;=\; \left(\sum_{l=0}^{i} \nu_{j-ld,d} \right)^{1/2} \tag{5.2}$$

approximates $\|x-x_{j-id}\|_A$ with the inaccuracy at most $\|x-x_{j+d}\|_A$. In practical computations we can simply store the values of $\nu_{0,d}$, $\nu_{d,d}$, $\nu_{2d,d}, \ldots, \nu_{j-d,d}$, and with the additional d steps update the estimates for the A-norm of the error in the steps $0, d, 2d, \ldots, j-d$ to

$$\nu_{0,j+d}^{1/2},\ \nu_{d,j}^{1/2},\ \nu_{2d,j-d}^{1/2}, \ldots,\ \nu_{j-d,2d}^{1/2}\,.$$

Dividing by $\nu_{0,j+d}^{1/2}$ we get the corresponding values of the estimates $\varrho_{d,j}^{1/2}$, $\varrho_{2d,j}^{1/2}$, $\ldots$, $\varrho_{j-d,2d}^{1/2}$ for the relative A-norm of the error. We illustrate this "reconstruction" of the convergence curve in Figure 5.4, computed for the problem `s3dkt3m2` with $d = 100$, where we plot the relative A-norm of the error (dashed line), its estimate $\varrho_{j,d}^{1/2}$ (solid line) and the updated estimates of the relative A-norm of the error computed for $j = 1000$ (dots), $j = 1700$ (x-marks), $j = 2200$ (pentagrams) and $j = 2500$ (stars). Please notice that when $\|x - x_j\|_A$ almost stagnates, the updated estimates can significantly differ from the original ones represented by the solid line.

We point out that in this paper we deal with evaluation of convergence, and we left heuristics for proper stopping criteria to further investigation. The problem `s3dkt3m2` illustrates that the last question is not trivial. Though, e.g., the computed estimates (even those updated at the iteration $j = 2200$) significantly decrease in the iterations 1800–2000, the actual value of the A-norm of the error still almost stagnates. We emphasize that neither the residual norm nor the normwise backward error reliably indicate the convergence of the A-norm of the error (cf. Figure 5.1, iterations 1800–2000).

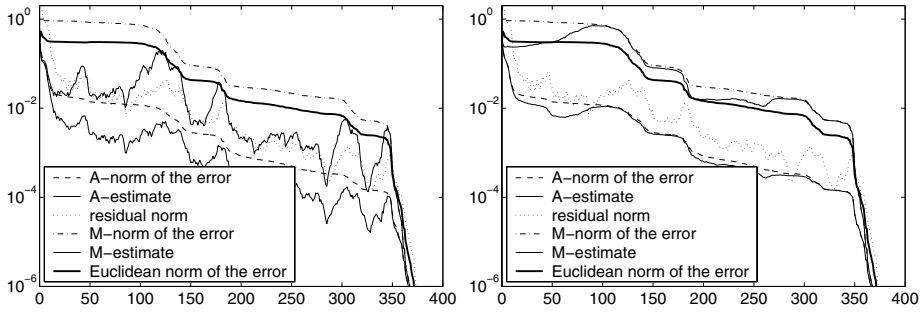


Figure 5.5: The system `tube`. The norms of the errors show similar behaviour. For $d = 4$, the estimates of $\|x - x_j\|_M$ and of $\|x - x_j\|_A$ behave erratically, similarly to the residual norm (left part). For $d = 40$, the estimates are smoother and closer to the approximated curves (right part).

5.3 Comparison of the convergence characteristics.

In Figure 5.5 we plot various convergence characteristics and error estimates for the system `tube`. We have used $d = 4$ (left part) and $d = 40$ (right part). The M-norm of the error $\|x - x_j\|_M$ (dash-dotted line), the Euclidean norm of the error $\|x - x_j\|$ (bold solid line) and the A-norm of the error $\|x - x_j\|_A$ (dashed line) show, except for a few initial iterations, similar behaviour. The estimates both of $\|x - x_j\|_M$ and $\|x - x_j\|_A$ are plotted by the solid lines (no confusion is possible; the line that is always under the dashed curve is the estimate of the A-norm of the error). The A-norm of the error is estimated more accurately than the M-norm of the error; while the estimate $\nu_{j,d}^{1/2}$ differs for no more that one order of magnitude from $\|x - x_j\|_A$, $\tau_{j,d}^{1/2}$ differs often for about two orders

of magnitude. The behaviour of both estimates is similar, but the peaks on the line representing $\tau_{j,d}^{1/2}$ are higher than the peaks on the line representing $\nu_{j,d}^{1/2}$. For $d = 4$ both estimates behave erratically, similarly to the residual norm (dotted line). By increasing the value of d, the estimates are smoother and closer to the approximated curves (see right part). The estimate of the M-norm of the error is in our example more sensitive to a slow decrease of error norms.

6 Conclusions.

We propose to incorporate the estimate for the A-norm of the error $\nu_{j,d}^{1/2}$ (see (3.7)) and the estimate for the relative A-norm of the error $\varrho_{j,d}^{1/2}$ (see (3.15)) into software realizations of the PCG method. They are simple and numerically stable, and can complement with a great benefit the quantities commonly used for evaluating convergence. The estimates are tight if the A-norm of the error reasonably decreases. With a good preconditioner ensuring fast convergence we get an authentic information about convergence in terms of the A-norm of the error. Similarly, the estimate $\tau_{j,d}^{1/2}$ (see (3.21)) for the M-norm of the error should be used whenever appropriate.

The proposed estimates can be combined with the standard quantities, such as residual norm or normwise backward error, for constructing a proper stopping criteria. The last topic as well as the (variable) choice of the parameter d in the estimates still needs further work.

Acknowledgment.

The authors wish to thank to R. Kouhia, M. Tůma and P. Benner for their invaluable help, advice and sharing results of their work, which made possible to prepare the experimental part of our paper, and to J. Liesen for his comments which improved the presentation. The comments of M. Arioli were very helpful in revising Section 3.2 of the original manuscript. The authors are also indebted to an anonymous referee for careful reading of the original manuscript and for valuable suggestions which clarified numerous formulations.

REFERENCES

1. M. Arioli, *A stopping criterion for the conjugate gradient algorithms in a finite element method framework*, Numer. Math., 97 (2004), pp. 1–24.
2. M. Arioli, J. W. Demmel and I. S. Duff, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.
3. M. Arioli, I. Duff and D. Ruiz, *Stopping criteria for iterative solvers*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 138–144.
4. M. Arioli, E. Noulard and A. Russo, *Stopping criteria for iterative methods: applications to PDE's*, Calcolo, 38 (2001), pp. 97–112.
5. O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, 1994.
6. O. Axelsson and I. Kaporin, *Error norm estimation and stopping criteria in preconditioned conjugate gradient iterations*, Numer. Linear Algebra Appl., 8 (2001), pp. 265–286.

7. I. Babuška, *Mathematics of the verification and validation in computational engineering*, in Mathematical and Computer Modelling in Science and Engineering, M. Kočandrlová and V. Kelar, eds., pp. 5–12, Union of Czech Mathematicians and Physicists, Prague, 2003.

8. R. Barrett, M. Berry, T. F. Chan et al., *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, PA, 1994.

9. P. Benner, *Solving large-scale control problems*, to appear in IEEE Control Syst. Magazine, 24 (2004), pp. 44–59.

10. G. Dahlquist, S. Eisenstat and G. H. Golub, *Bounds for the error of linear systems of equations using the theory of moments*, J. Math. Anal. Appl., 37 (1972), pp. 151–166.

11. G. Dahlquist, G. H. Golub and S. G. Nash, *Bounds for the error in linear systems*, in Proc. Workshop on Semi-Infinite Programming, R. Hettich, ed., pp. 154–172, Springer, Berlin, 1978.

12. P. Deuflhard, *Cascadic conjugate gradient methods for elliptic partial differential equations: algorithm and numerical results*, in Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993), Contemp. Math., vol. 180, pp. 29–42, Am. Math. Soc., Providence, RI, 1994.

13. V. Frayssé, L. Giraud, S. Gratton and J. Langou, *A set of GMRES routines for real and complex arithmetics on on high performance computers*, TR/PA/03/3, CERFACS, Toulouse Cedex, France, 2003.

14. G. H. Golub and G. Meurant, *Matrices, moments and quadrature*, in Proc. 15-th Dundee Conf., June 1993, D. Sciffeths and G. Watson, eds., pp. 105–156, Longman Sci. Tech. Publ., 1994.

15. 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. G. H. Golub and Z. Strakoš, *Estimates in quadratic formulas*, Numer. Algorithms, 8 (1994), pp. 241–268.

17. G. H. Golub and C. van Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore MD, third edn., 1996.

18. A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

19. A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.

20. A. Greenbaum, *Iterative methods for solving linear systems*, Frontiers in Applied Mathematics, vol. 17, SIAM, Philadelphia, PA., 1997.

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

22. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bureau Stand., 49 (1952), pp. 409–435.

23. N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

24. R. Kouhia, *Description of the CYLSHELL set*, Laboratory of Structural Mechanics, Finland, May 1998. Matrix Market.

25. Matrix Market, `http://math.nist.gov/MatrixMarket/`. The Matrix Market is a service of the Mathematical and Computational Sciences Division of the Information Technology Laboratory of the National Institute of Standards and Technology.

26. G. Meurant, *Computer solution of large linear systems*, Studies in Mathematics and its Applications, vol. 28, North-Holland Publishing Co., Amsterdam, 1999.

27. G. Meurant, *Numerical experiments in computing bounds for the norm of the error in the preconditioned conjugate gradient algorithm*, Numer. Algorithms 22, 3–4 (1999), pp. 353–365.

28. G. Meurant, *Towards a reliable implementation of the conjugate gradient method*, Invited plenary lecture at the Latsis Symposium: Iterative Solvers for Large Linear Systems, Zurich, February 2002.

29. E. Noulard and M. Arioli, *Vector stopping criteria for iterative methods: Theoretical tools*, pubblicazioni n. 956, Instituto di Analisi Numerica, Pavia, Italy, 1995.

30. 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.

31. C. C. Paige, *Error analysis of the lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl, 18 (1976), pp. 341–349.

32. C. C. Paige and M. A. Saunders, *LSQR: an algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Softw., 8 (1982), pp. 43–71.

33. C. C. Paige and Z. Strakoš, *Residual and backward error bounds in minimum residual Krylov subspace methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1898–1923 (electronic).

34. Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, second edn., 2003.

35. R. D. Skeel, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.

36. G. L. G. Sleijpen, H. A. van der Vorst and D. R. Fokkema, BiCGstab(l) *and other hybrid Bi-CG methods*, Numer. Algorithms, 7 (1994), pp. 75–109.

37. Z. Strakoš, *Theory of Convergence and Effects of Finite Precision Arithmetic in Krylov Subspace Methods*, thesis for the degree doctor of science, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, February 2001.

38. 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 (electronic).

39. Z. Strakoš and P. Tichý, *On estimation of the A-norm of the error in CG and PCG*, PAMM, 3 (2003), pp. 553–554 (published online).

40. H. A. van der Vorst, *Iterative Krylov methods for large linear systems*, Cambridge Monographs on Applied and Computational Mathematics, vol. 13, Cambridge University Press, Cambridge, 2003.