SIAM J. MATRIX ANAL. APPL.
Vol. 22, No. 1, pp. 213–229

ACCURACY OF TWO THREE-TERM AND THREE TWO-TERM RECURRENCES FOR KRYLOV SPACE SOLVERS*

MARTIN H. GUTKNECHT† AND ZDENĚK STRAKOŠ‡

Abstract. It has been widely observed that Krylov space solvers based on two three-term recurrences can give significantly less accurate residuals than mathematically equivalent solvers implemented with three two-term recurrences. In this paper we attempt to clarify and justify this difference theoretically by analyzing the gaps between recursively and explicitly computed residuals.

It is shown that, in contrast with the two-term recurrences analyzed by Sleijpen, van der Vorst, and Fokkema [*Numer. Algorithms*, 7 (1994), pp. 75–109] and Greenbaum [*SIAM J. Matrix Anal. Appl.*, 18 (1997), pp. 535–551], in the two three-term recurrences the contributions of the local roundoff errors to the analyzed gaps may be dramatically amplified while propagating through the algorithm. This result explains, for example, the well-known behavior of three-term-based versions of the biconjugate gradient method, where large gaps between recursively and explicitly computed residuals are not uncommon. For the conjugate gradient method, however, such a devastating behavior—although possible—is not observed frequently in practical computations, and the difference between two-term and three-term implementations is usually moderate or small. This can also be explained by our results.

Key words. system of linear algebraic equations, iterative method, Krylov space method, conjugate gradient method, three-term recurrence, accuracy, roundoff

AMS subject classifications. 65F10, 65G05

PII. S0895479897331862

1. Introduction.

Among the Krylov space solvers for linear systems $\mathbf{A}\mathbf{x} = \mathbf{b}$ (with $\mathbf{A}$ an $(N \times N)$-matrix and $\mathbf{b}$ an N-vector) there are quite a few that are based on three-term recurrences for both the *residuals* $\mathbf{r}_n$ and the *iterates* $\mathbf{x}_n$. Given an initial approximation $\mathbf{x}_0$, we let $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\mathbf{r}_{-1} = \mathbf{o}$, $\mathbf{x}_{-1} = \mathbf{o}$, $\beta_{-1} = 0$ and consider for $n \geq 0$, while $\gamma_n \neq 0$,

$$\begin{aligned} \mathbf{r}_{n+1} &= (\mathbf{A}\mathbf{r}_n - \mathbf{r}_n\alpha_n - \mathbf{r}_{n-1}\beta_{n-1})/\gamma_n\,, \\ \mathbf{x}_{n+1} &= -(\mathbf{r}_n + \mathbf{x}_n\alpha_n + \mathbf{x}_{n-1}\beta_{n-1})/\gamma_n\,. \end{aligned} \tag{1.1}$$

In order that the recurrences (1.1) be consistent with the residual definition $\mathbf{r}_n \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_n$, the scaling coefficients γ_n need to be chosen according to

$$\gamma_n = -(\alpha_n + \beta_{n-1}), \tag{1.2}$$

which means that the tridiagonal matrix with coefficients β_{n-1}, α_n, and γ_n in its $(n+1)$st column has column sums zero; see, for example, section 4.3 of [14].

The list of algorithms based on (1.1) and (1.2) includes the Chebyshev iteration [24, 21, 19], the second-order Richardson iteration [21] (which is the stationary form

*Received by the editors December 23, 1997; accepted for publication (in revised form) by G. H. Golub on September 22, 1999; published electronically June 3, 2000.

http://www.siam.org/journals/simax/22-1/33186.html

†Seminar for Applied Mathematics, ETH Zürich, ETH-Zentrum, CH-8092 Zürich, Switzerland (mhg@sam.math.ethz.ch).

‡Department of Mathematics and Computer Science, Emory University, Atlanta, GA 30322 (on leave from Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague) (strakos@mathcs.emory.edu). This author's work was supported by ASCR grant A2030706 and by GA CR grant 205/96/0921. Part of the work was performed while he visited the Swiss Center for Scientific Computing (CSCS/SCSC) in 1997.

of the Chebyshev iteration), the three-term versions (ORes) of the conjugate gradient (CG) and the conjugate residual (CR) methods [24, 15], and the three-term version (BiORes) of the unsymmetric or two-sided Lanczos method [18, 14] (which is a variation of the biconjugate gradient (BiCG) method); see also [2, 15]. On the other hand, for example, neither the version of CG suggested by Rutishauser [21] (based on recurrences for the increments in $\mathbf{x}$ and $\mathbf{r}$) nor the MINRES algorithm of Paige and Saunders [20], which implements the CR method for symmetric indefinite matrices, nor their SYMMLQ algorithm is covered by our assumptions. An interesting contribution to the rounding error analysis of MINRES and SYMMLQ can be found in [23].

The CG, CR, and BiCG methods have better known versions (OMin and BiOMin) that are instead based on three two-term recurrences involving, in addition to the iterates and their residuals, direction vectors $\mathbf{p}_n$: for $n \geq 0$,

$$\begin{aligned} \mathbf{p}_n &= \mathbf{r}_n + \mathbf{p}_{n-1}\psi_{n-1}\,, \\ \mathbf{r}_{n+1} &= \mathbf{r}_n - \mathbf{A}\mathbf{p}_n\omega_n\,, \\ \mathbf{x}_{n+1} &= \mathbf{x}_n + \mathbf{p}_n\omega_n \end{aligned} \tag{1.3}$$

with $\mathbf{p}_0 = \mathbf{r}_0$. Other methods like OrthoMin [28] use the last two of these recurrences, but have a more complex update formula for the direction vectors. In principle, the version (1.3) can be obtained from the three-term version (1.1)–(1.2) by an LU decomposition of the tridiagonal matrix of recurrence coefficients; see [1, 5, 14, 20]. The folklore is that implementations based on the two-term recurrences (1.3) are less affected by roundoff than those based on the three-term recurrences (1.1)–(1.2). It should be pointed out that the meaning of the phrase *less affected by roundoff* should be carefully specified, otherwise the previous statement is imprecise and can be misleading.

Recent work of Greenbaum [10, 11] shows that under the sole assumption that the last two recurrences (1.3) hold, there is a limitation on the accuracy of the iterates computed in finite precision arithmetic; the corresponding residuals $\mathbf{b}-\mathbf{A}\mathbf{x}_n$ cannot be expected to decrease below a certain level. (A similar but somewhat weaker result was given by Sleijpen, van der Vorst, and Fokkema [22].) This level depends primarily on the largest norm of an approximate solution $\mathbf{x}_n$ generated during the iteration, but it does not explicitly depend on how the coefficients ω_n and ψ_n are determined. Since, for example, the BiCG method may produce very large intermediate iterates and residuals, this result is of great importance in practice. In contrast, related work on GMRES showed that the size of intermediate iterates does not play a role [4, 12].

In this paper we investigate and answer the question when and why algorithms based on two three-term recurrences of the form (1.1)–(1.2) usually do not produce as small residuals as mathematically equivalent algorithms based on three two-term recurrences (1.3). Similarly to [10, 11, 22, 4], we investigate the gap $\mathbf{f}_n \equiv (\mathbf{b} - \mathbf{A}\mathbf{x}_n) - \mathbf{r}_n$ between the explicitly computed residuals $\mathbf{b} - \mathbf{A}\mathbf{x}_n$ and the recursively updated residuals $\mathbf{r}_n$. We will refer to the former as *true* residuals and to the latter as *updated* residuals. We show that for computations based on (1.1)–(1.2), the gap $\mathbf{f}_n$ satisfies a nonhomogeneous second-order difference equation. By writing n steps of this difference equation as the superposition of $n+1$ homogeneous difference equations (in a different context, this idea has been used by Grcar [8]), we receive an explicit formula for $\mathbf{f}_n$ in terms of the local roundoff errors. The resulting formula contains, in addition to the sum of local errors (which is the analog of the sum that represents the gap $\mathbf{f}_n$ in the case of two-term recurrences analyzed by Greenbaum), each local error

multiplied by a set of potentially large multipliers. Moreover, the local errors may become for the two three-term recurrences much larger than for two-term recurrences.

Assume that—in any application for which they are suitable—the methods based on the recurrences (1.1)–(1.2) or (1.3) will eventually produce small updated residuals (whose norm will decrease to the level of roundoff occurring in the finite precision computation of the residual $\mathbf{b} - \mathbf{A}\mathbf{x}$ for the exact solution $\mathbf{x}$). Then the size of the gap $\mathbf{f}_n$ determines the ultimate attainable accuracy measured by the size of the true residual; a large gap will eventually mean a poor residual $\mathbf{b} - \mathbf{A}\mathbf{x}_n$. The methods based on (1.1)–(1.2) are proven to be *in this sense* potentially much less accurate than those based on (1.3). In this sense, the folklore statement mentioned above is correct.

Our theoretical conclusions are well supported by numerical experiments.

It should be mentioned that the question of the ultimate attainable accuracy of iterative methods was studied by several other authors in addition to those mentioned above; see, for example, [3, 17, 25, 26, 27]. For a more detailed discussion we refer to [11]. However, to our knowledge, the problem of numerical differences between the recurrences (1.1)–(1.2) and (1.3) was not analyzed in these papers.

2. Local roundoff and the basic recurrence for the gap. In finite precision arithmetic, recurrences (1.1) have to be replaced by

$$\begin{aligned} \mathbf{r}_{n+1} &= (\mathbf{A}\mathbf{r}_n - \mathbf{r}_n\alpha_n - \mathbf{r}_{n-1}\beta_{n-1} + \mathbf{g}_n)/\gamma_n\,, \\ \mathbf{x}_{n+1} &= -(\mathbf{r}_n + \mathbf{x}_n\alpha_n + \mathbf{x}_{n-1}\beta_{n-1} - \mathbf{h}_n)/\gamma_n\,, \end{aligned} \tag{2.1}$$

where $\mathbf{g}_n$ and $\mathbf{h}_n$ contain all the local rounding errors produced at the step $n+1$, and $\mathbf{r}_n$, $\mathbf{x}_n$, etc., denote the actually computed quantities.

The first step of the analysis consists of estimating these local errors. We make the usual assumption that the floating-point arithmetic with roundoff unit ϵ satisfies

$$\mathrm{fl}(a \pm b) = a(1+\epsilon_1) \pm b(1+\epsilon_2), \qquad |\epsilon_1|, |\epsilon_2| \le \epsilon, \tag{2.2}$$

$$\mathrm{fl}(a \ \mathrm{op}\ b) = (a \ \mathrm{op}\ b)(1+\epsilon_3), \qquad |\epsilon_3| \le \epsilon, \qquad \mathrm{op} = *, /. \tag{2.3}$$

Then the roundoff in the matrix-vector multiplication (computed in a standard way) is bounded according to

$$|\mathrm{fl}(\mathbf{A}\mathbf{p}) - \mathbf{A}\mathbf{p}| \le m\,\epsilon\,|\mathbf{A}|\,|\mathbf{p}| + \mathcal{O}(\epsilon^2), \tag{2.4}$$

where $|\mathbf{A}|$ and $|\mathbf{p}|$ denote the elementwise absolute values of $\mathbf{A}$ and $\mathbf{p}$, and m is the maximal number of nonzeros in any row of $\mathbf{A}$. Assuming that the first and the third terms in (1.1) are summed up first, by applying these rules we get

$$|\mathbf{g}_n| \le ((m+3)\,|\mathbf{A}|\,|\mathbf{r}_n| + 3\,|\mathbf{r}_n\alpha_n| + 4\,|\mathbf{r}_{n-1}\beta_{n-1}|)\,\epsilon + \mathcal{O}(\epsilon^2), \tag{2.5}$$

$$|\mathbf{h}_n| \le (3\,|\mathbf{r}_n| + 3\,|\mathbf{x}_n\alpha_n| + 4\,|\mathbf{x}_{n-1}\beta_{n-1}|)\,\epsilon + \mathcal{O}(\epsilon^2). \tag{2.6}$$

Both $\mathbf{g}_n$ and $\mathbf{h}_n$ are bounded by a quantity proportional to ϵ, but the behavior of their bounds close to convergence is different. While the updated residual will become eventually small in reasonable computations, and the bound for $|\mathbf{g}_n|$ will decrease correspondingly, the bound for $|\mathbf{h}_n|$ will not. Note that we could consider a norm of $\mathbf{g}_n$ and $\mathbf{h}_n$ here, but there is no real need for this.

In the following estimates we assume that the computed coefficients α_n, β_{n-1}, and γ_n satisfy, in analogy to (1.2),

$$\gamma_0 = -\alpha_0\,, \qquad \gamma_n = -(\alpha_n + \beta_{n-1}) + \varepsilon_n \quad (n > 0) \tag{2.7}$$

with error terms ε_n (note that this symbol is distinct from ϵ) that are bounded by

$$|\varepsilon_n| \le (|\alpha_n| + |\beta_{n-1}|)\,\nu\,\epsilon \quad (n > 0)\,, \tag{2.8}$$

where ν is a suitable small constant. Note that $\nu = 1$ when γ_n is computed using (1.2). For later convenience we set $\varepsilon_0 = 0$.

We want to estimate the norm of the difference (or gap) between updated and true residuals, hence, of

$$\mathbf{f}_n \equiv \mathbf{b} - \mathbf{A}\mathbf{x}_n - \mathbf{r}_n\,.$$

For $n = 0$, the gap $\mathbf{f}_0$ is the roundoff in computing $\mathbf{r}_0$ from $\mathbf{A}$, $\mathbf{x}_0$, and $\mathbf{b}$; that is, $\mathbf{f}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0 - \mathrm{fl}(\mathbf{b} - \mathbf{A}\mathbf{x}_0)$, and this is bounded by

$$|\mathbf{f}_0| \le ((m+1)\,|\mathbf{A}|\,|\mathbf{x}_0| + |\mathbf{b}|)\ \epsilon + \mathcal{O}(\epsilon^2). \tag{2.9}$$

Inserting the recursions (2.1) and the equality (2.7) we have

$$\begin{aligned}
\mathbf{f}_{n+1} &= \mathbf{b} + (\mathbf{A}\mathbf{r}_n + \mathbf{A}\mathbf{x}_n\alpha_n + \mathbf{A}\mathbf{x}_{n-1}\beta_{n-1} - \mathbf{A}\mathbf{h}_n)\frac{1}{\gamma_n} \\
&\quad - (\mathbf{A}\mathbf{r}_n - \mathbf{r}_n\alpha_n - \mathbf{r}_{n-1}\beta_{n-1} + \mathbf{g}_n)\frac{1}{\gamma_n} \\
&= -\left[(\mathbf{b} - \mathbf{A}\mathbf{x}_n - \mathbf{r}_n)\,\alpha_n + (\mathbf{b} - \mathbf{A}\mathbf{x}_{n-1} - \mathbf{r}_{n-1})\,\beta_{n-1} - \mathbf{b}\,\varepsilon_n + \mathbf{A}\mathbf{h}_n + \mathbf{g}_n\right]\frac{1}{\gamma_n} \\
&= -\left[\mathbf{f}_n\,\alpha_n + \mathbf{f}_{n-1}\,\beta_{n-1} - \mathbf{b}\,\varepsilon_n + \mathbf{A}\mathbf{h}_n + \mathbf{g}_n\right]\frac{1}{\gamma_n}.
\end{aligned} \tag{2.10}$$

Let us gather the last three terms, the local errors, in

$$\mathbf{l}_n \equiv (-\mathbf{b}\,\varepsilon_n + \mathbf{A}\mathbf{h}_n + \mathbf{g}_n)\,\frac{1}{\gamma_n}\,.$$

By inserting the estimates (2.5), (2.6), and (2.8) we see that

$$\begin{aligned}
|\mathbf{l}_n| \le [\,&|\mathbf{b}|\,(|\alpha_n| + |\beta_{n-1}|)\,\nu + (m+6)\,|\mathbf{A}|\,|\mathbf{r}_n| + 3\,(|\mathbf{A}|\,|\mathbf{x}_n| + |\mathbf{r}_n|)\,|\alpha_n| \\
&+ 4\,(|\mathbf{A}|\,|\mathbf{x}_{n-1}| + |\mathbf{r}_{n-1}|)\,|\beta_{n-1}|\,]\,\frac{\epsilon}{|\gamma_n|} + \mathcal{O}(\epsilon^2).
\end{aligned}$$

For $n = 0$, we have $\gamma_0 = -\alpha_0$, $\varepsilon_0 = 0$, and thus

$$\mathbf{l}_0 = (\mathbf{A}\mathbf{h}_0 + \mathbf{g}_0)\frac{1}{\gamma_0}, \qquad \mathbf{f}_1 = \mathbf{f}_0 - \mathbf{l}_0.$$

In summary, (2.10) yields for the gaps $\mathbf{f}_n$ the linear second-order difference equation

$$\mathbf{f}_1 = \mathbf{f}_0 - \mathbf{l}_0\,, \qquad \mathbf{f}_{n+1} = -\left(\mathbf{f}_n\frac{\alpha_n}{\gamma_n} + \mathbf{f}_{n-1}\frac{\beta_{n-1}}{\gamma_n} + \mathbf{l}_n\right) \qquad (n \ge 1), \tag{2.11}$$

or, equivalently, the pair of first-order difference equations

$$\begin{bmatrix} \mathbf{f}_n \\ \mathbf{f}_{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{O} & \mathbf{I} \\ -\frac{\beta_{n-1}}{\gamma_n}\mathbf{I} & -\frac{\alpha_n}{\gamma_n}\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{f}_{n-1} \\ \mathbf{f}_n \end{bmatrix} - \begin{bmatrix} \mathbf{o} \\ \mathbf{l}_n \end{bmatrix} \qquad (n \ge 1) \tag{2.12}$$

with $\mathbf{f}_1 = \mathbf{f}_0 - \mathbf{l}_0$. These recurrences describe the propagation of the local rounding errors $\mathbf{l}_k$, $k = 0, \ldots, n$. We see that the gap $\mathbf{f}_n$ between the updated and the true residuals after n steps is determined by a nonhomogeneous second-order difference equation. This is in sharp contrast to the error behavior of the coupled two-term recurrences, where the gap after n steps is just a simple sum of local errors; see [11]. Consequently, as we will see in the next section, the two three-term recurrences may suffer from a strong amplification of the local errors.

3. Formula for the gap between true and updated residuals. For the moment, assume that the term ε_n in (2.7) vanishes, that is,

$$-\frac{\alpha_n}{\gamma_n} - \frac{\beta_{n-1}}{\gamma_n} = 1 \tag{3.1}$$

holds even in finite precision arithmetic. Denote by $\mathbf{z}_{n+1} = \mathcal{D}(\mathbf{z}_{n-m+1}, \mathbf{z}_{n-m}; m)$ the result of m steps of the recurrence

$$\mathbf{z}_{k+1} = -\mathbf{z}_k \frac{\alpha_k}{\gamma_k} - \mathbf{z}_{k-1} \frac{\beta_{k-1}}{\gamma_k}\,, \qquad k = n-m+1, \ldots, n, \tag{3.2}$$

started at the step $n-m$. Note that due to (3.1), $\mathbf{z}_{n-m+k+1} = \mathcal{D}(\mathbf{z}_{n-m+1}, \mathbf{z}_{n-m}; k) = \mathbf{z}_{n-m}$ for all k whenever $\mathbf{z}_{n-m+1} = \mathbf{z}_{n-m}$. Our discussion will rely heavily on this fact.

First, we derive how the gap $\mathbf{f}_{n+1}$ is affected by $\mathbf{f}_0$. Clearly, the part of this gap that depends on $\mathbf{f}_0$ is given by

$$\mathcal{D}(\mathbf{f}_0, \mathbf{f}_0; n) = \mathbf{f}_0\,,$$

that is, $\mathbf{f}_0$ is not amplified in the process. Next we have to analyze the dependence of $\mathbf{f}_{n+1}$ on the elementary rounding errors $\mathbf{l}_0$ born in the first step of the algorithm. Clearly, considering (2.11) for $n = 1$, subtracting and adding $\mathbf{l}_0 \frac{\beta_0}{\gamma_1}$, the contribution of $\mathbf{l}_0$ to the gap $\mathbf{f}_{n+1}$ can be decomposed into two parts: the part which propagates through the recurrence without any change,

$$\mathcal{D}(-\mathbf{l}_0, -\mathbf{l}_0; n) = -\,\mathbf{l}_0\,,$$

and the part depending on the modified local error of the first step,

$$\widetilde{\mathbf{l}}_1 \equiv\ \mathbf{l}_0 \frac{\beta_0}{\gamma_1} + \mathbf{l}_1\,,$$

which has yet to be analyzed. Repeating the same idea for the steps 2 through n, we can conclude that the gap $\mathbf{f}_{n+1}$ can be written as the following superposition of effects of local errors:

$$\begin{aligned}
\mathbf{f}_{n+1} = \mathbf{f}_0 \quad & -\mathbf{l}_0 \\
& -\mathbf{l}_0 \frac{\beta_0}{\gamma_1} - \mathbf{l}_1 \\
& -\mathbf{l}_0 \frac{\beta_0 \beta_1}{\gamma_1 \gamma_2} - \mathbf{l}_1 \frac{\beta_1}{\gamma_2} - \mathbf{l}_2 \\
& \ \vdots \\
& -\mathbf{l}_0 \frac{\beta_0 \beta_1 \cdots \beta_{n-1}}{\gamma_1 \gamma_2 \cdots \gamma_n} - \ldots - \mathbf{l}_{n-1} \frac{\beta_{n-1}}{\gamma_n} - \mathbf{l}_n.
\end{aligned} \tag{3.3}$$

Let us give another derivation of this fundamental result. From (2.12) we see that, in view of $\mathbf{f}_1 = \mathbf{f}_0 - \mathbf{l}_0$,

$$\left[\begin{array}{c} \mathbf{f}_n \\ \mathbf{f}_{n+1} \end{array}\right] = \prod_{k=1}^{n} \left[\begin{array}{cc} \mathbf{O} & \mathbf{I} \\ -\frac{\beta_{k-1}}{\gamma_k}\mathbf{I} & -\frac{\alpha_k}{\gamma_k}\mathbf{I} \end{array}\right] \left[\begin{array}{c} \mathbf{f}_0 \\ \mathbf{f}_0 \end{array}\right] - \sum_{j=0}^{n} \prod_{k=j+1}^{n} \left[\begin{array}{cc} \mathbf{O} & \mathbf{I} \\ -\frac{\beta_{k-1}}{\gamma_k}\mathbf{I} & -\frac{\alpha_k}{\gamma_k}\mathbf{I} \end{array}\right] \left[\begin{array}{c} \mathbf{o} \\ \mathbf{l}_j \end{array}\right]. \tag{3.4}$$

Here, due to (3.1), the matrices in the first product leave $[\ \mathbf{f}_0^\top \quad \mathbf{f}_0^\top\ ]^\top$ invariant. In the product that appears after the sum, we split off the last matrix (the one where $k = j+1$) and apply it to $[\ \mathbf{o}^\top \quad \mathbf{l}_j^\top\ ]^\top$ to get

$$\left[\begin{array}{c} \mathbf{l}_j \\ -\mathbf{l}_j \frac{\alpha_{j+1}}{\gamma_{j+1}} \end{array}\right] = \left[\begin{array}{c} \mathbf{l}_j \\ \mathbf{l}_j \end{array}\right] + \left[\begin{array}{c} \mathbf{o} \\ \mathbf{l}_j \frac{\beta_j}{\gamma_{j+1}} \end{array}\right].$$

Now we have again a first term that is left invariant by the matrices it is multiplied with and a second term of the form $[\ \mathbf{o}^\top \quad \star\ ]^\top$ that can be treated in the same way that $[\ \mathbf{o}^\top \quad \mathbf{l}_j^\top\ ]^\top$ was treated before. Repeating this trick we finally obtain

$$\left[\begin{array}{c} \mathbf{f}_n \\ \mathbf{f}_{n+1} \end{array}\right] = \left[\begin{array}{c} \mathbf{f}_0 \\ \mathbf{f}_0 \end{array}\right] - \sum_{j=0}^{n-1} \left[\begin{array}{c} \mathbf{l}_j \\ \mathbf{l}_j \end{array}\right] \left(1 + \frac{\beta_j}{\gamma_{j+1}} + \cdots + \frac{\beta_j \cdots \beta_{n-2}}{\gamma_{j+1} \cdots \gamma_{n-1}}\right) - \sum_{j=0}^{n} \left[\begin{array}{c} \mathbf{o} \\ \mathbf{l}_j \end{array}\right] \frac{\beta_j \cdots \beta_{n-1}}{\gamma_{j+1} \cdots \gamma_n}, \tag{3.5}$$

which is the same as formula (3.3), written for both $\mathbf{f}_n$ and $\mathbf{f}_{n+1}$.

Now we describe how the picture changes when the coefficients α_n, β_{n-1}, and γ_n are computed imprecisely, that is, when (3.1) is replaced by (2.7). We can follow the analysis described above with the only difference being that we should add the effect of the quantity $\mathbf{f}_0\varepsilon_1/\gamma_1$ propagating through $n-1$ steps of the recurrence (3.2) with $\mathbf{z}_1 := \mathbf{o}$, the effect of $\widetilde{\mathbf{l}}_1\varepsilon_2/\gamma_2$ propagating through $n-2$ steps of (3.2) with $\mathbf{z}_2 := \mathbf{o}$, and so on. As long as the constant ν is small and ε_n is close to the machine precision ϵ, these modifications will only cause effects proportional to $\mathcal{O}(\epsilon^2)$. In (3.3) we should therefore add terms $\mathcal{O}(\epsilon^2)$ to individual terms of the sum. However, once the size of these terms is considered, the new $\mathcal{O}(\epsilon^2)$ contributions can be thought of as being incorporated in the $\mathcal{O}(\epsilon^2)$ terms already present in the bounds for $\mathbf{f}_0, \mathbf{l}_0, \ldots, \mathbf{l}_n$. Therefore, we can use (3.3) in the further analysis with no change and no limitation.

We summarize our main result in the following theorem.

THEOREM 3.1. *Up to a term $\mathcal{O}(\epsilon^2)$, the gap $\mathbf{f}_{n+1}$ between true and updated*

residuals is given by the formula

$$
\begin{aligned}
\mathbf{f}_{n+1} = \mathbf{f}_0 \quad & - \sum_{j=0}^{n} \mathbf{l}_j \\
& - \mathbf{l}_0 \left(\frac{\beta_0}{\gamma_1} + \frac{\beta_0\beta_1}{\gamma_1\gamma_2} + \cdots + \frac{\beta_0 \cdots \beta_{n-1}}{\gamma_1 \cdots \gamma_n} \right) \\
& - \mathbf{l}_1 \left(\frac{\beta_1}{\gamma_2} + \cdots + \frac{\beta_1 \cdots \beta_{n-1}}{\gamma_2 \cdots \gamma_n} \right) \\
& \vdots \\
& - \mathbf{l}_{n-1} \frac{\beta_{n-1}}{\gamma_n} \, .
\end{aligned} \tag{3.6}
$$

It is tempting to estimate $||\mathbf{f}_n||$ directly on the basis of (3.4), using an appropriate norm for the 2×2 block matrices. However, the resulting estimate is too generous, as it does not take into account the fundamental special properties of these block matrices.

4. Comparison with three coupled two-term recurrences. In our notation, Greenbaum's gap [11] for the coupled two-term recurrences (1.3) is

$$
\mathbf{f}^{\mathrm{G}}_{n+1} = \mathbf{f}_0 - \sum_{j=0}^{n} \mathbf{l}^{\mathrm{G}}_j \, , \qquad \text{where} \quad \mathbf{l}^{\mathrm{G}}_j \equiv \mathbf{A}\mathbf{h}^{\mathrm{G}}_j + \mathbf{g}^{\mathrm{G}}_j \, , \tag{4.1}
$$

with $\mathbf{g}^{\mathrm{G}}_n$ and $\mathbf{h}^{\mathrm{G}}_n$ denoting the local rounding errors in the computation of the first two recurrences of (1.3), analogously to $\mathbf{g}_n$ and $\mathbf{h}_n$ in (2.1). A comparison of (4.1) with (3.6) is instructive.

We point out that the size of the local rounding errors may be considerably larger in the two three-term recurrences than in the three two-term recurrences; the size of the local error $\mathbf{l}^{\mathrm{G}}_j$ in the step n is essentially bounded by $\mathcal{O}(\epsilon)||\mathbf{A}|| \ \max_{1 \le j \le n} ||\mathbf{x}_j||$ (see [11]), where $||\mathbf{A}||$ denotes the spectral norm of $\mathbf{A}$. In our case, a similar term in the bound for $||\mathbf{l}_n||$ would be multiplied by the factor $(3|\alpha_n|+4|\beta_{n-1}|)/|\gamma_n|$, which can be substantially larger than 1; see section 5 for the specific case of the CG method. Nevertheless, as documented by our numerical experiments in section 6, the difference between the implementations based on the two three-term recurrences (1.1)–(1.2) and those using the three two-term recurrences (1.3) cannot be explained by the size of the local rounding error terms only. The amplification of the local errors due to possibly large multipliers plays a substantial if not decisive role: the additional terms in (3.6) can be similar in size to or even dominate the sum of local rounding errors. If the multipliers become very large, then the two three-term recurrences (1.1)–(1.2) are likely to exhibit a dramatically wider gap than the two-term recurrences (1.3).

Assuming, as in [11], that the updated residuals become eventually negligible, the relations (3.6) and (4.1) determine the ultimate attainable accuracy of the methods based on (1.1)–(1.2) and (1.3), respectively, measured by the norm of the true residuals.

5. Example: CG method. For the following discussion of the size of the multiplicative factors

$$
\prod_{j=i}^{k} \frac{\beta_{j-1}}{\gamma_j} \qquad (1 \le i \le k)
$$

we restrict ourselves to symmetric positive definite matrices $\mathbf{A}$ and to the CG method. First, for the simplicity of our exposition, we assume exact arithmetic.

The coefficients in the two-term recurrences (1.3) are for CG given by [16]

$$\omega_n = \frac{\langle \mathbf{r}_n, \mathbf{r}_n \rangle}{\langle \mathbf{p}_n, \mathbf{A}\mathbf{p}_n \rangle}, \qquad \psi_n = \frac{\langle \mathbf{r}_{n+1}, \mathbf{r}_{n+1} \rangle}{\langle \mathbf{r}_n, \mathbf{r}_n \rangle}. \tag{5.1}$$

Both ω_n and ψ_n are positive. Without specific knowledge about $\mathbf{A}$ and $\mathbf{r}_0$ we cannot say anything more about their values. More precisely, given any two sequences of positive numbers, $\omega_0, \ldots, \omega_{N-1}$ and $\psi_0, \ldots, \psi_{N-2}$, there is a symmetric positive definite matrix $\mathbf{A}$ and a vector $\mathbf{r}_0$ such that the classical OMin form (the Hestenes–Stiefel (HS) implementation) of the CG method applied to $\mathbf{A}$ with the initial residual $\mathbf{r}_0$ generates the given coefficients; see Theorem 18:3 of Hestenes and Stiefel [16]. This result allows us to construct examples having any given set of multipliers, and thus to find some with very large gaps. On the other hand, if the matrix $\mathbf{A}$ is reasonably well conditioned and if the CG method converges well, then the bounds derived for the multipliers will show that no substantial amplification of the local rounding errors will occur.

It is well known [20, 5, 1] that by eliminating the direction vectors $\mathbf{p}_n$ in (1.3) we obtain the three-term (ORes) variant of the CG method with recurrences of the form (1.1)–(1.2). From the orthogonality of the residuals we receive

$$\alpha_n = \frac{\langle \mathbf{r}_n, \mathbf{A}\mathbf{r}_n \rangle}{\langle \mathbf{r}_n, \mathbf{r}_n \rangle}, \qquad \beta_{n-1} = \gamma_{n-1} \frac{\langle \mathbf{r}_n, \mathbf{r}_n \rangle}{\langle \mathbf{r}_{n-1}, \mathbf{r}_{n-1} \rangle}. \tag{5.2}$$

Using (5.1) and $\gamma_n = -(\alpha_n + \beta_{n-1})$, we see that the coefficients of the two implementations are related by

$$\gamma_n = -\frac{1}{\omega_n} < 0, \qquad \frac{\beta_{n-1}}{\gamma_n} = \frac{\psi_{n-1}\omega_n}{\omega_{n-1}} \geq 0, \qquad \frac{\alpha_n}{\gamma_n} = -1 - \frac{\psi_{n-1}\omega_n}{\omega_{n-1}} \leq -1, \tag{5.3}$$

where $\psi_{-1} = 0$, $\omega_{-1} = 1$. The equality is attained in the last two formulas only if $\mathbf{x}_n = \mathbf{x}$, that is, if we have reached the solution. We conclude that the multiplicative factors in (3.3) have the form

$$\prod_{j=i}^{k} \frac{\beta_{j-1}}{\gamma_j} = \frac{\omega_k}{\omega_{i-1}} \prod_{j=i}^{k} \psi_{j-1}\,, \tag{5.4}$$

and therefore they may exhibit, in general, an arbitrary behavior.

For a given matrix $\mathbf{A}$ and an initial residual $\mathbf{r}_0$, it is possible to relate the size of the multipliers to the condition number of $\mathbf{A}$ and the convergence of the CG process measured by the norm of the residuals. First, according to Theorem 5:5 in [16],

$$\frac{\langle \mathbf{p}_n, \mathbf{A}\mathbf{p}_n \rangle}{\langle \mathbf{p}_n, \mathbf{p}_n \rangle} < \frac{1}{\omega_n} = |\gamma_n| < \frac{\langle \mathbf{r}_n, \mathbf{A}\mathbf{r}_n \rangle}{\langle \mathbf{r}_n, \mathbf{r}_n \rangle},$$

which yields, with the spectral norm,

$$\frac{1}{\|\mathbf{A}^{-1}\|} = \frac{1}{\sigma_{\min}(\mathbf{A})} < \frac{1}{\omega_n} = |\gamma_n| < \|\mathbf{A}\|. \tag{5.5}$$

Rewriting the multipliers in the form

$$\prod_{j=i}^{k} \frac{\beta_{j-1}}{\gamma_j} = \frac{\omega_k}{\omega_{i-1}} \frac{||\mathbf{r}_k||^2}{||\mathbf{r}_{i-1}||^2},$$

we receive the following bounds:

$$\frac{1}{\kappa(\mathbf{A})}\,\frac{||\mathbf{r}_k||^2}{||\mathbf{r}_{i-1}||^2} \le \prod_{j=i}^{k}\frac{\beta_{j-1}}{\gamma_j} \le \kappa(\mathbf{A})\,\frac{||\mathbf{r}_k||^2}{||\mathbf{r}_{i-1}||^2}, \tag{5.6}$$

where $\kappa(\mathbf{A})$ is the spectral condition number of the matrix $\mathbf{A}$. Note that

$$\frac{||\mathbf{r}_k||^2}{||\mathbf{r}_{i-1}||^2} = \frac{||\mathbf{A}^{1/2}\,\mathbf{A}^{1/2}\,(\mathbf{x}-\mathbf{x}_k)||^2}{||\mathbf{A}^{1/2}\,\mathbf{A}^{1/2}\,(\mathbf{x}-\mathbf{x}_{i-1})||^2} \le \frac{||\mathbf{A}||}{\sigma_{\min}(\mathbf{A})}\,\frac{||\mathbf{x}-\mathbf{x}_k||_{\mathbf{A}}^2}{||\mathbf{x}-\mathbf{x}_{i-1}||_{\mathbf{A}}^2} \le \kappa(\mathbf{A})$$

due to the monotonicity of the $\mathbf{A}$-norm of the error. Consequently,

$$\prod_{j=i}^{k}\frac{\beta_{j-1}}{\gamma_j} \le \kappa^2(\mathbf{A}).$$

As mentioned in section 2, the bound for the size of the local rounding errors $\mathbf{l}_n$ in the two three-term recurrences (1.1)–(1.2) contains the factors $|\alpha_n/\gamma_n|$ and $|\beta_{n-1}/\gamma_n|$. In view of (5.2) and (5.5) we have $0 \le \alpha_n \le \|\mathbf{A}\|$ and $|\gamma_n|^{-1} \le \|\mathbf{A}^{-1}\|$. Using (5.3), we obtain the estimate

$$0 \le \frac{\beta_{n-1}}{\gamma_n} \le \left|\frac{\alpha_n}{\gamma_n}\right| \le \kappa(\mathbf{A}). \tag{5.7}$$

Surprisingly, to establish that the developed bounds remain relevant in the case of finite precision computation we do not need any extra work: the results of [9] and [13] imply that in finite precision arithmetic the following slightly relaxed bounds hold:

$$(1-\vartheta)\,\frac{1}{\kappa(\mathbf{A})}\,\frac{||\mathbf{r}_k||^2}{||\mathbf{r}_{i-1}||^2} \le \prod_{j=i}^{k}\frac{\beta_{j-1}}{\gamma_j} \le (1+\vartheta)\,\kappa(\mathbf{A})\,\frac{||\mathbf{r}_k||^2}{||\mathbf{r}_{i-1}||^2}\,, \tag{5.8}$$

$$\prod_{j=i}^{k}\frac{\beta_{j-1}}{\gamma_j} \le (1+\vartheta)\,\kappa^2(\mathbf{A})\,, \tag{5.9}$$

$$\frac{\beta_{n-1}}{\gamma_n} \le \left|\frac{\alpha_n}{\gamma_n}\right| \le (1+\vartheta)\,\kappa(\mathbf{A})\,, \tag{5.10}$$

where $0 \le \vartheta \ll 1$. (Here, we make the usual assumption about the numerical non-singularity of the matrix $\mathbf{A}$; for details see the references mentioned above.) Note, however, that the conclusion we just made is far from trivial. The values of the actually computed recurrence coefficients and of the residual norms may be completely different from their theoretical counterparts. But still, essentially the same bounds hold!

The large size of the upper bounds for ill-conditioned $\mathbf{A}$ suggest that though the size of the local errors may contribute to a possibly large gap between true and updated residuals, the further amplification of the local errors due to large multipliers may have a much stronger effect.

6. Numerical experiments with the CG method. The construction of our numerical experiments follows ideas from [16].

Example 1. We consider $N = 48$ and aim at the following values of the coefficients (5.1) for the classical HS form of the CG method:

$$\begin{aligned} &\omega_0 = \omega_1 = \cdots = \omega_{47} = 1, \\ &\psi_0 = 10, \qquad \psi_1 = \psi_3 = \cdots = \psi_{43} = 0.01, \qquad \psi_2 = \cdots = \psi_{42} = 100, \\ &\psi_{44} = 10^{-2}, \qquad \psi_{45} = 10^{-3}, \qquad \psi_{46} = 10^{-4}. \end{aligned} \tag{6.1}$$

Using the well-known formulas [9]

$$\begin{aligned} \mathbf{T}_{0,0} &= \frac{1}{\omega_0}\,, \\ \mathbf{T}_{i,i} &= \frac{1}{\omega_i} + \frac{\psi_{i-1}}{\omega_{i-1}}\,, \\ \mathbf{T}_{i,i-1} = \mathbf{T}_{i-1,i} &= \frac{\sqrt{\psi_{i-1}}}{\omega_{i-1}}\,, \qquad i = 1, \ldots, N-1, \end{aligned} \tag{6.2}$$

we construct an $N \times N$ symmetric positive definite tridiagonal matrix $\mathbf{T}$ with spectral norm $\|\mathbf{T}\| = 102$ and condition number $\kappa(\mathbf{T}) \approx 2 \times 10^6$ (for $N = 48$). For any unitary $N \times N$ matrix $\mathbf{V}$, the CG method (1.3), (5.1) applied to the system $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} = \mathbf{VTV}^\star$ and $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0 = \mathbf{Ve}_1$ then generates in steps 1 to N the prescribed coefficients ω_j, ψ_j, $j = 0, \ldots, N-1$, and the residual norms

$$\begin{aligned} \|r_j\| &= 10^{1/2} \qquad \text{for } \; j = 1, 3, \ldots, 43, \\ \|r_j\| &= 10^{-1/2} \qquad \text{for } \; j = 2, 4, \ldots, 44, \end{aligned}$$

with $\|r_j\|$ sharply decreasing in the steps 45 through 48. For an initial residual different from $\mathbf{Ve}_1$ the behavior of the residual norms will be different, but we may still expect some oscillations and, consequently, some large multipliers.

We have used the construction described above, choosing $\mathbf{V}$ as the unitary matrix resulting from the QR decomposition of a randomly generated $N \times N$ matrix; in MATLAB notation $[\mathbf{V}, \mathbf{R}] = qr(\text{randn}(N, N))$. Furthermore, we have chosen $\mathbf{x} = (1, \ldots, 1)^\top$, $\mathbf{b} = \mathbf{Ax}$, $\mathbf{x}_0 = \mathbf{o}$, $\mathbf{r}_0 = \mathbf{b}$. Hence, $\mathbf{r}_0 \neq \mathbf{Ve}_1$. Experiments were performed on an Sun Ultra 10 workstation with $\epsilon \approx 1.11 \times 10^{-16}$ using MATLAB 5.0.

Three implementations of the CG method have been compared: except for Figure 9, solid lines always represent results of the classical OMIN or Hestenes–Stiefel (HS) version given by (1.3) and (5.1), dots those of the Rutishauser (R) variant described in [21], and dashed lines those of the ORES implementation of the form (1.1)–(1.2) presented, for example, in [15, p. 143], and denoted here as HY. In the R variant the recurrences are, for $n \geq 0$, of the form

$$\begin{aligned} \Delta\mathbf{r}_n &= (-\mathbf{Ar}_n + \Delta\mathbf{r}_{n-1}\,\eta_{n-1})\,\tau_n^{-1}\,, & \mathbf{r}_{n+1} &= \mathbf{r}_n + \Delta\mathbf{r}_n\,, \\ \Delta\mathbf{x}_n &= (\mathbf{r}_n + \Delta\mathbf{x}_{n-1}\,\eta_{n-1})\,\tau_n^{-1}\,, & \mathbf{x}_{n+1} &= \mathbf{x}_n + \Delta\mathbf{x}_n\,, \end{aligned} \tag{6.3}$$

and they are started with $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$, $\Delta\mathbf{r}_{-1} = \mathbf{o}$, $\Delta\mathbf{x}_{-1} = \mathbf{o}$, and $\eta_{-1} = 0$. The coefficients are computed according to

$$\tau_n = \frac{\langle \mathbf{r}_n, \mathbf{Ar}_n \rangle}{\langle \mathbf{r}_n, \mathbf{r}_n \rangle} - \eta_{n-1}\,, \qquad \eta_n = \tau_n \frac{\langle \mathbf{r}_{n+1}, \mathbf{r}_{n+1} \rangle}{\langle \mathbf{r}_n, \mathbf{r}_n \rangle}. \tag{6.4}$$

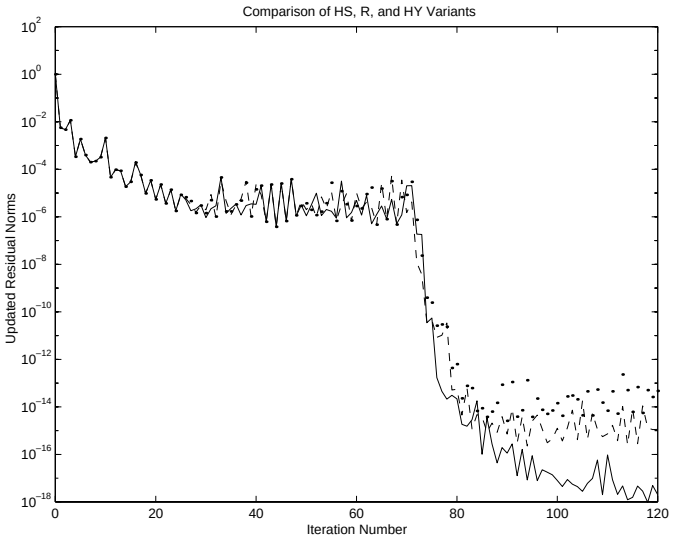

FIG. 1. *Example* 1: *Norms of the updated residuals for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

In the HY variant, the following recurrences are used for $n \geq 0$:

$$\begin{aligned} \mathbf{r}_{n+1} &= \theta_{n+1}(-\mu_{n+1}\mathbf{A}\mathbf{r}_n + \mathbf{r}_n) + (1-\theta_{n+1})\mathbf{r}_{n-1}\,, \\ \mathbf{x}_{n+1} &= \theta_{n+1}(\mu_{n+1}\mathbf{r}_n + \mathbf{x}_n) + (1-\theta_{n+1})\mathbf{x}_{n-1}\,. \end{aligned} \tag{6.5}$$

They are started with $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$, $\theta_1 = 1$, $\mathbf{x}_{-1} = \mathbf{o}$, and $\mathbf{r}_{-1} = \mathbf{o}$, and the coefficients are computed according to

$$\mu_n = \frac{\langle \mathbf{r}_n, \mathbf{r}_n\rangle}{\langle \mathbf{r}_n, \mathbf{A}\mathbf{r}_n\rangle}\,, \qquad \theta_{n+1} = \left(1 - \frac{\mu_{n+1}}{\mu_n}\,\frac{\langle \mathbf{r}_n, \mathbf{r}_n\rangle}{\langle \mathbf{r}_{n-1}, \mathbf{r}_{n-1}\rangle}\,\frac{1}{\theta_n}\right)^{-1}. \tag{6.6}$$

Clearly, the finite precision equivalent of (6.5) can be written in the form (2.1). Consequently, Theorem 3.1 applies, although the bounds for the size of the local errors derived in section 2 have to be modified slightly.

Norms of the updated residuals are compared in Figure 1. We can see the oscillations followed by the fast convergence for n around 70. Of course, theoretically the method should converge in 48 steps, but, as can be explained by the analysis in [9, 13], the convergence is delayed due to roundoff effects. Norms of the true residuals $\|\mathbf{b} - \mathbf{A}\mathbf{x}_n\|$ are shown in Figure 2. Clearly, residual norms of the HY variant stagnate at a significantly worse level than those of the HS variant, as predicted by our analysis.

In Figure 3 the norms of the gaps $\mathbf{f}_n$ we investigated, that is, of the differences between true and updated residuals, are displayed. Note that for the HY variant the gap starts to grow soon, much earlier than one can detect from the two previous figures. Figure 4 shows the behavior of the error norms $\|\mathbf{x} - \mathbf{x}_n\|$. Surprisingly, the differences in the error norms are much less pronounced than those in the true residuals.

Example 2. The second example makes use of the same construction, but now, again for $N = 48$, we aim at

$$\begin{aligned} \omega_0 &= \omega_1 = \cdots = \omega_{47} = 1, \\ \psi_0 &= \psi_1 = \cdots = \psi_{39} = \sqrt{2}, \qquad \psi_{40} = \cdots = \psi_{46} = 2^{-7}, \end{aligned} \tag{6.7}$$

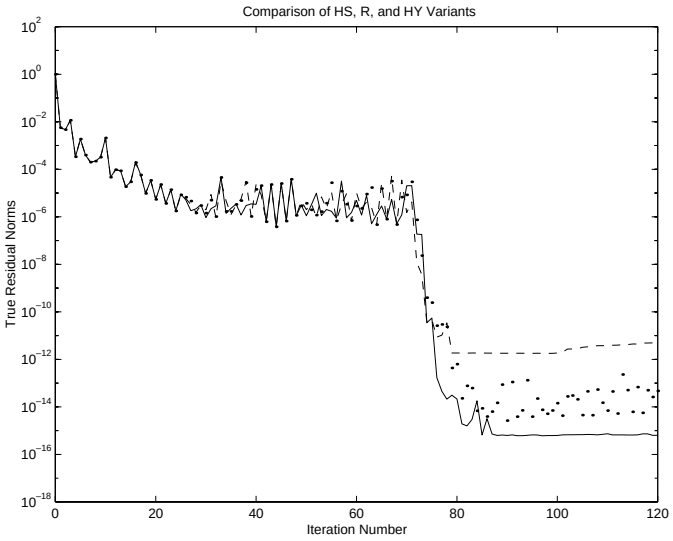

FIG. 2. *Example* 1: *Norms of the true residuals computed as* $\|\mathbf{b}-\mathbf{A}\mathbf{x}_n\|$ *for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

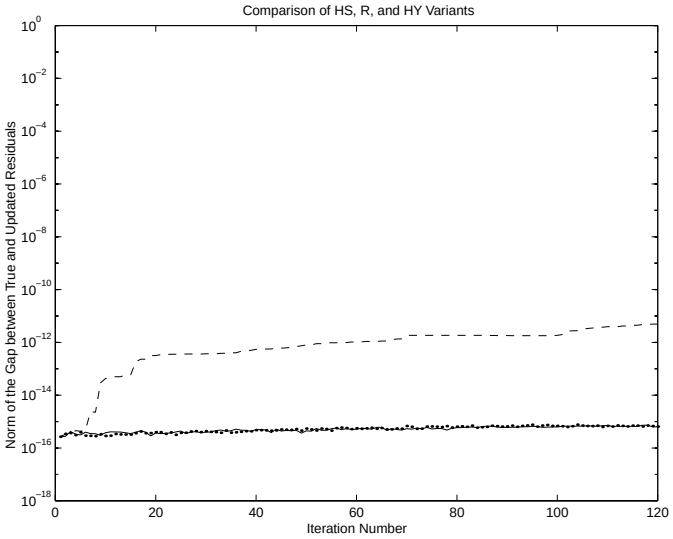

FIG. 3. *Example* 1: *Norms of the differences (gaps)* $\mathbf{f}_n$ *between the true and updated residuals for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

which gives $\|\mathbf{T}\| \approx 4.8$ and $\kappa(\mathbf{T}) \approx 6 \times 10^7$. Again, we consider the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} = \mathbf{V}\mathbf{T}\mathbf{V}^\star$, where $\mathbf{V}$ is determined as in Example 1, $\mathbf{x} = (1, \ldots, 1)^\top$, $\mathbf{b} = \mathbf{A}\mathbf{x}$. If we chose $\mathbf{x}_0$ so that $\mathbf{r}_0 = \mathbf{V}\mathbf{e}_1$, we would find residuals with

$$\|r_n\| = (\sqrt{2})^n \qquad \text{for} \qquad n = 1, 2, \ldots, 40$$

and a sharply decreasing norm in the subsequent steps. However, we have again chosen $\mathbf{x}_0$ differently, namely $\mathbf{x}_0 = \mathbf{o}$, so that $\mathbf{r}_0 = \mathbf{b}$. Then we do not find an initially

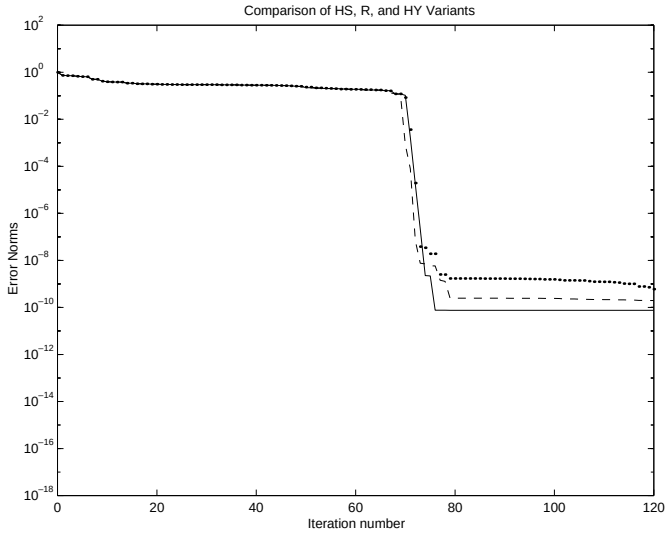

FIG. 4. *Example* 1: *Norms of the errors* $\|\mathbf{x}-\mathbf{x}_n\|$ *for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

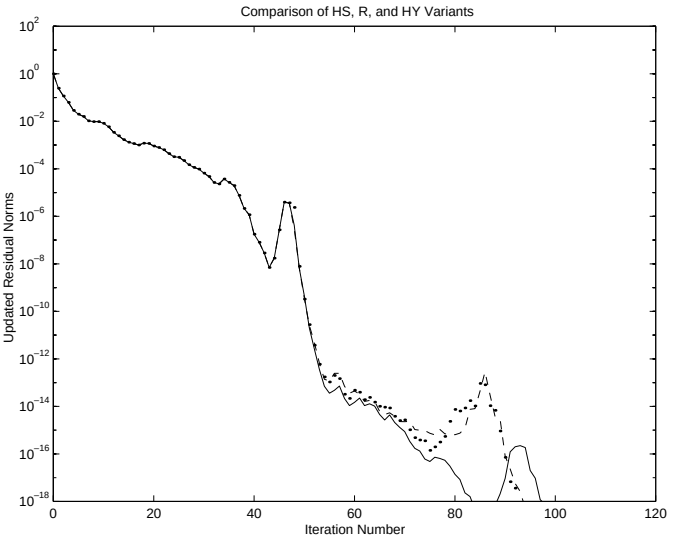

FIG. 5. *Example* 2: *Norms of the updated residuals for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

increasing but rather a quickly decreasing residual norm, both for the updated (see Figure 5) and the true residual (see Figure 6); note the significant oscillation around $n = 45$. The norm of the true residuals of the HY variant stagnates again at a significantly worse level than in the HS variant. Figure 7 shows the norm of the gaps $\mathbf{f}_n$. The differences in the norms of the errors, displayed in Figure 8, are again less pronounced.

To illustrate the contribution of the size of local rounding errors to the gap $\mathbf{f}_n$, we plotted in Figure 9 the size of the coefficients $|\alpha_n/\gamma_n|$, β_n/γ_n and $|1/\gamma_n|$. Clearly, while

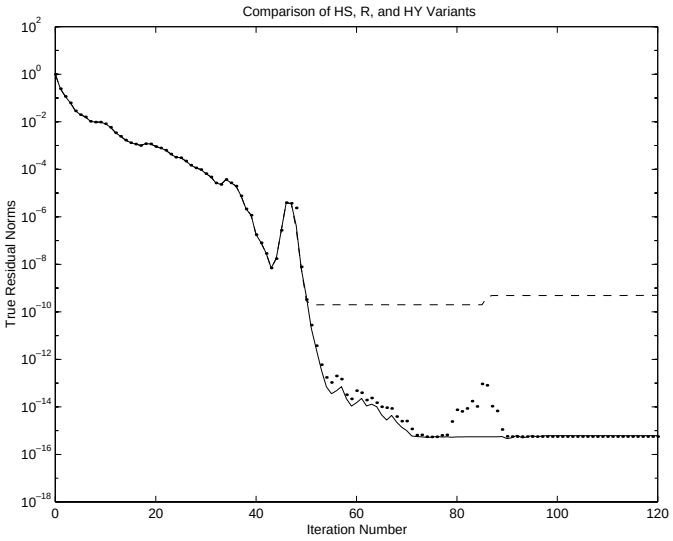

FIG. 6. *Example* 2: *Norms of the true residuals computed as* $\|\mathbf{b}-\mathbf{A}\mathbf{x}_n\|$ *for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

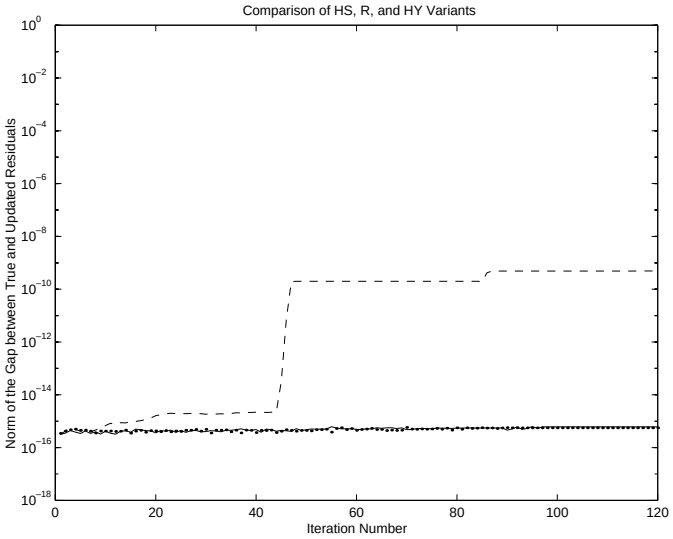

FIG. 7. *Example* 2: *Norms of the differences (gaps)* $\mathbf{f}_n$ *between the true and updated residuals for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

the gap exhibits a loss of accuracy of about six orders of magnitude, the anticipated contribution of the local errors to this gap is not greater than about two orders of magnitude. The disastrous difference between updated and true residuals must therefore be caused by an amplification of the local rounding errors due to large multipliers. In the analogous figure (not shown) for Example 1 the same behavior is slightly less pronounced.

A detailed explanation of the performance of the R variant and of the behavior of the error in all variants requires further work.

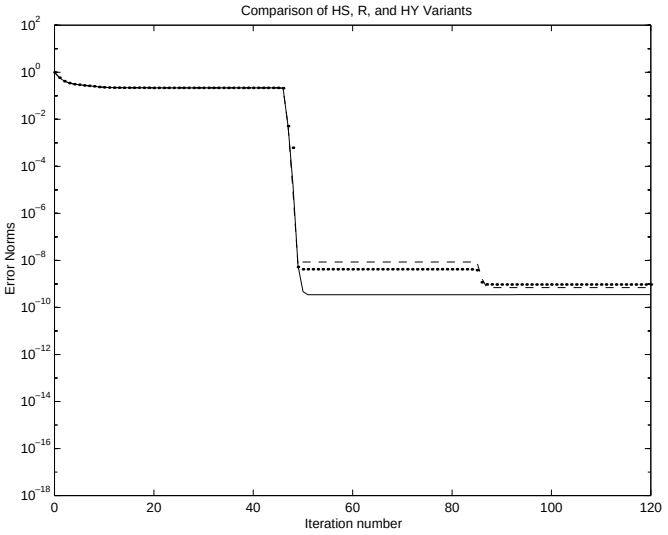

FIG. 8. *Example* 2: *Norms of the errors* $\|\mathbf{x}-\mathbf{x}_n\|$ *for the two-term (HS, solid line), three-term (HY, dashed line), and Rutishauser (R, dots) variants of the CG method.*

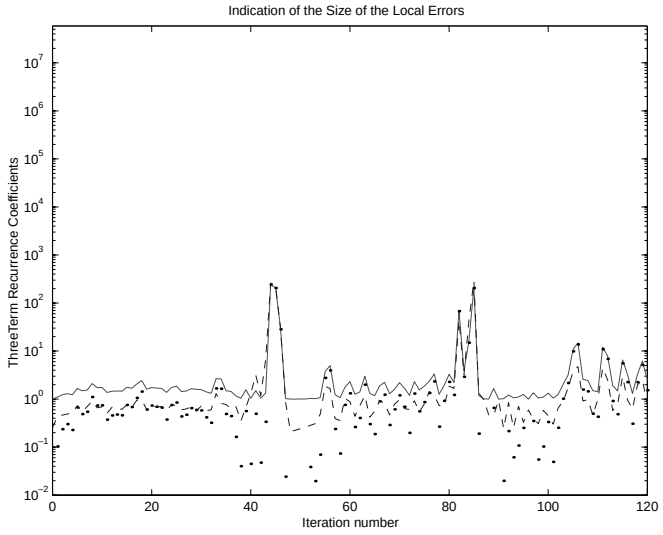

FIG. 9. *Example* 2: *Size of the three-term recurrence coefficients* $|\alpha_n/\gamma_n|$ *(solid line),* β_n/γ_n *(dots) and* $|1/\gamma_n|$ *(dashed line) for the HY variant of the CG method*

.

7. Conclusions. We have explained why the ultimate attainable accuracy measured by the norm of the true residual $\mathbf{b}-\mathbf{A}\mathbf{x}_n$ can be much worse for implementations of Krylov space methods based on the two three-term recurrences (1.1)–(1.2) than for the corresponding implementations based on two-term recurrences of the form (1.3). For example, in the three-term (ORES) version of the CG method, the gap between true and updated residuals is affected not only by the maximum size of the intermediate iterates $||\mathbf{x}_k||$ as in the coupled two-term (OMIN) version, but also by oscillations of the squared norms of the residuals, that is, the quantities $||\mathbf{r}_k||^2/||\mathbf{r}_{i-1}||^2,\ 1 \le i \le k$.

Many well-known algorithms like MINRES and SYMMLQ [20], or the three-term and the coupled two-term versions of the quasi-minimal residual (QMR) method [6, 7], as well as the Rutishauser variant of the CG method are not of the form (1.1)–(1.2) or (1.3). Hence, the results presented in this paper do not apply to them.

Chris Paige suggested another derivation of the results presented in this paper, based entirely on matrix formulations of the algorithms. His approach brings some additional insight into the problem and has potential for further generalization of the results. We hope to report about the results of the joint subsequent work in the near future.

Acknowledgments. The authors would like to thank Anne Greenbaum, Gerard Meurant, Chris Paige, Lisa Perrone, and Miro Rozložník for their helpful comments.

REFERENCES

[1] S. F. Ashby and M. H. Gutknecht, *A matrix analysis of conjugate gradient algorithms*, in Advances in Numerical Methods for Large Sparse Sets of Linear Systems, M. Natori and T. Nodera, eds., Parallel Processing for Scientific Computing 9, Keio University, Yokahama, Japan, 1993, pp. 32–47.

[2] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.

[3] J. A. M. Bollen, *Numerical stability of descent methods for solving linear equations*, Numer. Math., 43 (1984), pp. 361–377.

[4] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical stability of the GMRES method*, BIT, 35 (1995), pp. 308–330.

[5] R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis, G. A. Watson, ed., Lecture Notes in Math. 506, Springer, Berlin, 1976, pp. 73–89.

[6] R. W. Freund and N. M. Nachtigal, *QMR: A quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339.

[7] R. W. Freund and N. M. Nachtigal, *An implementation of the QMR method based on coupled two-term recurrences*, SIAM J. Sci. Comput., 15 (1994), pp. 313–337.

[8] J. F. Grcar, *Analyses of the Lanczos Algorithm and of the Approximation Problem in Richardson's Method*, Ph.D. thesis, Report UIUCDCS-R-81-1074, University of Illinois at Urbana-Champaign, 1981.

[9] A. Greenbaum, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, Linear Algebra Appl., 113 (1989), pp. 7–63.

[10] A. Greenbaum, *Accuracy of computed solutions from conjugate-gradient-like methods*, in Advances in Numerical Methods for Large Sparse Sets of Linear Systems, M. Natori and T. Nodera, eds., Parallel Processing for Scientific Computing 10, Keio University, Yokahama, Japan, 1994, pp. 126–138.

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

[12] A. Greenbaum, M. Rozložník, and Z. Strakoš, *Numerical behaviour of the modified Gram-Schmidt GMRES implementation*, BIT, 37 (1997), pp. 706–719.

[13] 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.

[14] M. H. Gutknecht, *Lanczos-type solvers for nonsymmetric linear systems of equations*, Acta Numerica, 6 (1997), pp. 271–397.

[15] L. Hageman and D. Young, *Applied Iterative Methods*, Academic Press, Orlando, FL, 1981.

[16] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–435.

[17] N. J. Higham and P. A. Knight, *Componentwise error analysis for stationary iterative methods*, in Linear Algebra, Markov Chains, and Queueing Models, C. D. Meyer and R. J. Plemmons, eds., IMA Vol. Math. Appl. 48, Springer-Verlag, New York, 1993, pp. 29–46.

[18] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Research Nat. Bur. Standards, 45 (1950), pp. 255–281.

[19] T. A. Manteuffel, *The Tchebyshev iteration for nonsymmetric linear systems*, Numer. Math., 28 (1977), pp. 307–327.

[20] C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[21] H. Rutishauser, *Theory of gradient methods*, in Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems, Mitt. Inst. angew. Math. ETH Zürich, Birkhäuser-Verlag, Basel, Switzerland, 1959, pp. 24–49.
[22] 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.
[23] G. L. G. Sleijpen, H. A. van der Vorst, and J. Modersitzki, *The main effects of rounding errors in Krylov solvers for symmetric linear systems*, SIAM J. Matrix Anal. Appl., submitted.
[24] E. Stiefel, *Relaxationsmethoden bester Strategie zur Lösung linearer Gleichungssysteme*, Comm. Math. Helv., 29 (1955), pp. 157–179.
[25] H. Woźniakowski, *Numerical stability of the Chebyshev method for the solution of large linear systems*, Numer. Math., 28 (1977), pp. 191–209.
[26] H. Woźniakowski, *Round-off error analysis of iterations for large linear systems*, Numer. Math., 30 (1978), pp. 301–314.
[27] H. Woźniakowski, *Round-off error analysis of a new class of conjugate-gradient algorithms*, Linear Algebra Appl., 29 (1980), pp. 507–529.
[28] D. M. Young and K. C. Jea, *Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods*, Linear Algebra Appl., 34 (1980), pp. 159–194.