BIT
2001, Vol. 41, No. 1, pp. 086–114
0006-3835/01/4101-0086 $16.00

RESIDUAL SMOOTHING TECHNIQUES: DO THEY IMPROVE THE LIMITING ACCURACY OF ITERATIVE SOLVERS? *

MARTIN H. GUTKNECHT and MIROSLAV ROZLOŽNIK †

Seminar for Applied Mathematics, Swiss Federal Institute of Technology (ETH) Zurich, ETH-Zentrum, CH-8092 Zurich, Switzerland
email: mhg@sam.math.ethz.ch, miro@cs.caz.cz

Abstract.

Many iterative methods for solving linear systems, in particular the biconjugate gradient (BiCG) method and its "squared" version CGS (or BiCGS), produce often residuals whose norms decrease far from monotonously, but fluctuate rather strongly. Large intermediate residuals are known to reduce the ultimately attainable accuracy of the method, unless special measures are taken to counteract this effect. One measure that has been suggested is residual smoothing: by application of simple recurrences, the iterates x_n and the corresponding residuals $r_n := b - Ax_n$ are replaced by smoothed iterates y_n and corresponding residuals $s_n := b - Ay_n$. We address the question whether the smoothed residuals can ultimately become markedly smaller than the primary ones. To investigate this, we present a roundoff error analysis of the smoothing algorithms. It shows that the ultimately attainable accuracy of the smoothed iterates, measured in the norm of the corresponding residuals, is, in general, not higher than that of the primary iterates. Nevertheless, smoothing can be used to produce certain residuals, most notably those of the minimum residual method, with higher attainable accuracy than by other frequently used algorithms.

AMS subject classification: 65F10.

Key words: Iterative methods, residual smoothing, limiting accuracy.

1 Introduction.

Observing the sequence of 2-norms of the residuals, sometimes called the *residual norm history*, is the usual way of monitoring the convergence of an iterative method for solving linear systems $\mathbf{Ax} = \mathbf{b}$. Unfortunately, for many methods and problems, in particular for the biconjugate gradient (BiCG) method and even more so for its "squared" version CGS (or BiCGS), this residual norm history shows quite an erratic convergence behavior, far from the monotonous convergence one might hope for. For wide classes of methods based on either

*Received October 1999. Revised May 2000. Communicated by Lothar Reichel.

†Current address: Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží, CZ-182 07 Prague 8, Czech Republic. Part of the second author's work was supported by the Grant Agency of the Czech Republic under grant No. 201/98/P108.

two-term or three-term recurrences, it has been shown that large intermediate residuals reduce the ultimately attainable accuracy of the method [12, 18, 26], unless special measures are taken to counteract this effect [22, 25, 29]. One such measure that has been suggested is *residual smoothing*: the *primary sequences* of approximate solutions (or, iterates) $\mathbf{x}_n$ and corresponding residuals $\mathbf{r}_n := \mathbf{b} - \mathbf{A}\mathbf{x}_n$ provided by some iterative method are replaced by *smoothed sequences* of iterates $\mathbf{y}_n$ and corresponding residuals $\mathbf{s}_n := \mathbf{b} - \mathbf{A}\mathbf{y}_n$. The smoothed residual $\mathbf{s}_n$ is defined as a linear (and normally convex) combination of $\mathbf{r}_n$ and some of the earlier smoothed residuals $\mathbf{s}_0, \ldots, \mathbf{s}_{n-1}$, and the smoothed iterate $\mathbf{y}_n$ is defined consistently. Alternatively, the smoothed residual $\mathbf{s}_n$ could be chosen as a linear combination of $\mathbf{r}_n$, $\mathbf{s}_{n-1}$ and some of the earlier primary residuals $\mathbf{r}_0, \ldots, \mathbf{r}_{n-1}$ [33, 3]. Consequently, smoothing can be viewed as a weighting process, although not all weights need to be positive. Particular smoothing methods differ by how many old smoothed iterates and residuals get involved and by how the weights are chosen. We will restrict ourselves to schemes which, for computing $\mathbf{y}_n$ and $\mathbf{s}_n$, only involve the last previous smoothed iterate and residual, $\mathbf{y}_{n-1}$ and $\mathbf{s}_{n-1}$, and the new primary iterate and residual, $\mathbf{x}_n$ and $\mathbf{r}_n$. The two best known methods, *minimal residual* (*MR*) *smoothing* and *quasi-minimal residual* (QMR) *smoothing* are discussed below. Both MR and QMR smoothing can effectively eliminate peaks in the residual norm history; for MR smoothing the convergence becomes even monotonous. However, as has been noticed in many numerical experiments and was recently explained by suitably interpreting known connections between the primary and the smoothed residuals [2, 4, 5, 15, 14, 30, 31], in exact arithmetic the smoothed residuals do not converge considerably faster than those of the primary method, unless the primary iterative method converges extremely slowly [17]. We will show here — and this does not come as a surprise — that this is also true in finite precision arithmetic.

Therefore, all we can hope for is that in finite precision arithmetic the ultimate accuracy of the smoothed iterates, measured by the residual norm, is higher. This is the main question addressed in this paper. We will see that the answer is again rather negative. Nevertheless, there is a useful consequence of the fact that the ultimate accuracy of the smoothed residuals is on the same level as that of the primary iterates: there are cases where smoothing can be applied to produce smoothed residuals that are ultimately more accurate than those obtained by other, mathematically equivalent algorithms frequently used in practice. To obtain an accurate implementation of the smoothed method we just need an accurate implementation of the primary method; smoothing will not spoil its ultimate accuracy. The minimum residual solution resulting from smoothing the iterates and residuals of the conjugate gradient (CG) method will be seen to be an example for this effect.

Our results on the ultimate accuracy of the smoothed iterates are based on an analysis of the *gap* between the true and updated residuals. In theory, by definition, the primary and the smoothed residuals are $\mathbf{r}_n := \mathbf{b} - \mathbf{A}\mathbf{x}_n$ and $\mathbf{s}_n := \mathbf{b} - \mathbf{A}\mathbf{y}_n$, respectively, but in finite precision arithmetic we have to distinguish between the *updated residuals* computed by some recurrence formula and the

true residual, obtained by evaluating the above definitions. For distinction, we denote quantities computed in finite precision arithmetic by bars: for the primary iterates and the primary updated residuals we write $\bar{\mathbf{x}}_n$ and $\bar{\mathbf{r}}_n$, respectively, and for those of the smoothing method $\bar{\mathbf{y}}_n$ and $\bar{\mathbf{s}}_n$. Consequently, the *primary gap* is defined as $\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_n - \bar{\mathbf{r}}_n$, and the corresponding gap of the smoothed method, which will be mostly referred to just as the *gap*, is $\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n$. Essentially, our result will be that both gaps are roughly of the same size except when the updated primary residual is considerably larger than the updated smoothed residual and the final accuracy has not yet been reached.

We assume that the usual rules of a well designed floating-point arithmetic hold, and use occasionally the notation $\mathrm{fl}(\cdot)$ for the computed result of an expression. The machine precision is denoted by ϵ. In particular, for a sparse matrix-vector multiplication the error bound

$$\|\mathrm{fl}(\mathbf{Ap}) - \mathbf{Ap}\| \le m\, N^{1/2}\, \epsilon \|\mathbf{A}\|\|\mathbf{p}\| \tag{1.1}$$

is used, where N is the order of $\mathbf{A}$, m refers to the maximum number of nonzeros of $\mathbf{A}$ per row, and $\|\mathbf{p}\|$ is the 2-norm of $\mathbf{p}$, which is generally applied in this paper. For the matrix $\mathbf{A}$, we also make use of the spectral norm $\|\mathbf{A}\|$ and the corresponding condition number $\kappa(\mathbf{A})$. Finally, we apply the $\mathcal{O}$-notation when suitable.

2 Two smoothing methods and their three implementations.

Since roundoff errors depend on the particular formulas used, we do not only investigate various smoothing methods but also several mathematically equivalent implementations for each of these methods.

The *Schönauer–Weiss implementation* [24, 31] assumes only that the primary sequences $\mathbf{x}_n$ and $\mathbf{r}_n$ are provided. While these are generated, we compute the smoothed sequences of iterates $\mathbf{y}_n$ and corresponding residuals $\mathbf{s}_n$ by choosing, for $n = 1, 2, \ldots$, the *smoothing parameter* σ_n and evaluating

$$\mathbf{y}_n = (1-\sigma_n)\mathbf{y}_{n-1} + \sigma_n \mathbf{x}_n\,, \qquad \mathbf{s}_n = (1-\sigma_n)\mathbf{s}_{n-1} + \sigma_n \mathbf{r}_n\,. \tag{2.1}$$

At the start we set $\mathbf{y}_0 = \mathbf{x}_0$ and $\mathbf{s}_0 = \mathbf{r}_0$.

In (2.1) we assume to use the updated primary residual. (We do not yet use bars at this point, since we do not yet analyze the roundoff errors.) But as s a variant we will analyze an *'expensive' Schönauer–Weiss implementation* that uses the true residuals:

$$\mathbf{y}_n = (1-\sigma_n)\mathbf{y}_{n-1} + \sigma_n \mathbf{x}_n\,, \qquad \mathbf{s}_n = (1-\sigma_n)\mathbf{s}_{n-1} + \sigma_n(\mathbf{b} - \mathbf{A}\mathbf{x}_n)\,. \tag{2.2}$$

The *Zhou–Walker implementation* [33] assumes additionally that the primary iterates and residuals are updated by recurrences involving direction vectors $\mathbf{p}_n$:

$$\mathbf{x}_n = \mathbf{x}_{n-1} + \alpha_{n-1}\mathbf{p}_{n-1}\,, \qquad \mathbf{r}_n = \mathbf{r}_{n-1} - \alpha_{n-1}\mathbf{A}\mathbf{p}_{n-1}\,. \tag{2.3}$$

Actually, it does not make use of $\mathbf{x}_n$ and $\mathbf{r}_n$ directly, but only of $\alpha_{n-1}\mathbf{p}_{n-1}$ and $\alpha_{n-1}\mathbf{A}\mathbf{p}_{n-1}$. The smoothed sequences of iterates $\mathbf{y}_n$ and residuals $\mathbf{s}_n$ are

recursively computed as follows: starting from $\mathbf{y}_0 = \mathbf{x}_0$, $\mathbf{s}_0 = \mathbf{r}_0$, $\mathbf{u}_0 = \mathbf{v}_0 = \mathbf{0}$ (where $\mathbf{0}$ denotes the zero vector), and $\sigma_0 = 1$, for $n = 1, 2, \ldots$ we first update the "smoothed" direction vector $\mathbf{v}_n$ and its image $\mathbf{u}_n :\equiv \mathbf{A}\mathbf{v}_n$ according to

$$\mathbf{v}_n = (1-\sigma_{n-1})\mathbf{v}_{n-1}+\alpha_{n-1}\mathbf{p}_{n-1}\,, \quad \mathbf{u}_n = (1-\sigma_{n-1})\mathbf{u}_{n-1}+\alpha_{n-1}\mathbf{A}\mathbf{p}_{n-1}\,, \tag{2.4}$$

then determine the smoothing parameter σ_n (see formulas below), and finally let

$$\mathbf{y}_n = \mathbf{y}_{n-1} + \sigma_n\mathbf{v}_n\,, \qquad \mathbf{s}_n = \mathbf{s}_{n-1} - \sigma_n\mathbf{u}_n\,. \tag{2.5}$$

One can verify that this procedure generates the same sequences of $\{\mathbf{y}_n\}$ and $\{\mathbf{s}_n\}$ as (2.1) if the same smoothing parameters are used. Moreover, comparing (2.1) and (2.5) we can then conclude that

$$\mathbf{r}_n = \mathbf{s}_{n-1} - \mathbf{u}_n\,. \tag{2.6}$$

Hence, if the primary residuals are not updated according to the second equation in (2.3), they can be obtained from (2.6).

We next recall two ways of choosing the smoothing parameters σ_n. First, the *minimal residual (MR) smoothing* proposed by Schönauer [24] and investigated by Weiss [31, 32] takes σ_n such that $\|\mathbf{s}_n\|$ becomes as small as possible, assuming $\mathbf{s}_n$ is of the form given in (2.1), (2.2), or (2.5). This one-dimensional minimization problem is solved by making $\mathbf{s}_n$ orthogonal to $\mathbf{s}_{n-1} - \mathbf{r}_n$, that is

$$\sigma_n :\equiv \frac{\langle \mathbf{s}_{n-1}, \mathbf{s}_{n-1} - \mathbf{r}_n\rangle}{\|\mathbf{s}_{n-1} - \mathbf{r}_n\|^2} \tag{2.7}$$

in case of the Schönauer–Weiss implementation, and, in view of (2.6),

$$\sigma_n :\equiv \frac{\langle \mathbf{s}_{n-1}, \mathbf{u}_n\rangle}{\|\mathbf{u}_n\|^2} \tag{2.8}$$

for the Zhou–Walker implementation. From Pythagoras' theorem we have then

$$\|\mathbf{s}_n\|^2 = \|\mathbf{s}_{n-1}\|^2 - \|\mathbf{s}_{n-1} - \mathbf{r}_n\|^2|\sigma_n|^2 = \|\mathbf{s}_{n-1}\|^2 - \|\mathbf{u}_n\|^2|\sigma_n|^2. \tag{2.9}$$

Note that by definition of $\mathbf{s}_n$,

$$\|\mathbf{s}_n\| \le \|\mathbf{s}_{n-1}\|, \qquad \|\mathbf{s}_n\| \le \|\mathbf{r}_n\|. \tag{2.10}$$

Under the additional assumption that the residuals of the primary method are mutually orthogonal, we have $\mathbf{r}_n \perp \mathbf{s}_{n-1}$, so that (2.7) and (2.8) simplify to

$$\sigma_n = \frac{\|\mathbf{s}_{n-1}\|^2}{\|\mathbf{s}_{n-1}\|^2 + \|\mathbf{r}_n\|^2} \qquad \in\ (0, 1]. \tag{2.11}$$

Inserting this into (2.9) yields then [31]

$$\frac{1}{||\mathbf{s}_n||^2} = \frac{||\mathbf{s}_{n-1}||^2 + ||\mathbf{r}_n||^2}{||\mathbf{s}_{n-1}||^2||\mathbf{r}_n||^2} = \frac{1}{||\mathbf{s}_{n-1}||^2} + \frac{1}{||\mathbf{r}_n||^2} = \sum_{k=0}^{n} \frac{1}{||\mathbf{r}_k||^2}. \tag{2.12}$$

Second, the *quasi-minimal residual (*QMR*) smoothing* proposed by Zhou and Walker [33] defines the smoothing parameters by

$$\sigma_n :\equiv \frac{\tau_n^2}{\|\mathbf{r}_n\|^2}\,, \qquad \text{where} \quad \frac{1}{\tau_n^2} = \sum_{k=0}^{n} \frac{1}{\|\mathbf{r}_k\|^2}\,, \quad \tau_n > 0\,. \tag{2.13}$$

The scalar τ_n can be seen to be the norm of the so-called *quasi-residual* $\mathbf{q}_n$, the coefficient vector which represents the smoothed residual in the constructed underlying basis of the Krylov space:

$$\mathbf{s}_n = \mathbf{V}_{n+1}\mathbf{q}_n\,, \quad \text{where} \quad \mathbf{V}_{n+1} :\equiv \left[\begin{array}{ccc} \mathbf{v}_0 & \cdots & \mathbf{v}_n \end{array}\right] :\equiv \left[\begin{array}{ccc} \dfrac{\mathbf{r}_0}{\|\mathbf{r}_0\|} & \cdots & \dfrac{\mathbf{r}_n}{\|\mathbf{r}_n\|} \end{array}\right];$$

see, e.g., [15, Section 5.2] or [17]. Clearly, τ_n^{-2} can be computed recursively by

$$\frac{1}{\tau_0^2} = \frac{1}{\|\mathbf{r}_0\|^2}\,, \qquad \frac{1}{\tau_k^2} = \frac{1}{\tau_{k-1}^2} + \frac{1}{\|\mathbf{r}_k\|^2} \qquad (k = 1, \ldots, n). \tag{2.14}$$

If the Zhou–Walker implementation of QMR smoothing is applied to a primary method that does not generate residuals directly, we have to replace $\|\mathbf{r}_k\|$ in the above formulas by $\|\mathbf{s}_{k-1} - \mathbf{u}_k\|$ $(k = 1, \ldots, n)$; see (2.6). From (2.14) we obtain conversely,

$$\|\mathbf{r}_n\| = \sqrt{\frac{\tau_n^2}{1 - (\tau_n/\tau_{n-1})^2}}\,. \tag{2.15}$$

Note that in view of (2.12) the relations (2.14) and (2.15) also hold for MR smoothing if the primary method produces orthogonal residuals (or at least if we have $\mathbf{s}_{n-1} \perp \mathbf{r}_n$ at every step), and if we replace τ_n by $\|\mathbf{s}_n\|$. These two relations are the basis of the so-called peak-plateau connection between primary and smoothed methods, as has been clarified by Cullum and Greenbaum [5], following earlier work of Brown [2], Cullum [4], Walker [30], and others. Actually, this work is about the connections between the residual norms of FOM and GMRES as well as those of CG and CR. In fact, the above relations of MR smoothing hold for these residuals if we denote those of FOM (or CG) by $\mathbf{r}_n$ and those of GMRES (or CR) by $\mathbf{s}_n$: Weiss [31] showed that the GMRES iterates and residuals can be obtained from those of FOM by applying MR smoothing.

QMR smoothing must not be confused with the QMR method due to Freund and Nachtigal [8], which combines the (look-ahead) Lanczos process [7] with ideas from MINRES [23] to generate iterates whose residuals converge typically much smoother than those of the related biconjugate gradient (BICG) method. However, the denomination "QMR smoothing" is justified since, in exact arithmetic, QMR smoothing applied to the BICG iterates produces exactly the QMR iterates.

For our roundoff error analysis we will need some crude estimates of the size of the smoothing parameters and of the norm of the smoothed residuals. First, note that the recurrences (2.1) imply (recall that $\sigma_0 = 1$)

$$\mathbf{y}_n = \sum_{k=0}^{n} \mathbf{x}_k \sigma_k \prod_{j=k+1}^{n} (1 - \sigma_j)\,, \qquad \mathbf{s}_n = \sum_{k=0}^{n} \mathbf{r}_k \sigma_k \prod_{j=k+1}^{n} (1 - \sigma_j)\,. \tag{2.16}$$

From (2.13) follows in particular that for the quasi-residual norm τ_n holds

$$\frac{1}{\sqrt{n+1}} \min_{k=0,\ldots,n} \{\|\mathbf{r}_k\|\} \leq \tau_n \leq \min_{k=0,\ldots,n} \{\|\mathbf{r}_k\|\} . \tag{2.17}$$

We also make use of the simple estimate $\|\mathbf{s}_n\| \leq \sqrt{n+1}\,\tau_n$ from [33], which also results from $\mathbf{s}_n = \mathbf{V}_{n+1}\mathbf{q}_n$ and $\|\mathbf{V}_{n+1}\| \leq \sqrt{n+1}$, see [8], and leads to

$$\|\mathbf{s}_n\| \leq \|\mathbf{V}_{n+1}\|\,\tau_n \leq \sqrt{n+1}\,\tau_n \leq \sqrt{n+1} \min_{k=0,\ldots,n} \{\|\mathbf{r}_k\|\} . \tag{2.18}$$

A lower bound can be given in terms of the smallest singular value of $\mathbf{V}_{n+1}$:

$$\|\mathbf{s}_n\| \geq \sigma_{\min}(\mathbf{V}_{n+1})\,\tau_n .$$

Clearly, the extremal singular values of $\mathbf{V}_{n+1}$ determine how close the norms of the smoothed residuals and quasi-residuals are.

3 How to make the roundoff effects in MR smoothing worse than in QMR smoothing.

The size of the smoothing coefficients σ_n may affect the numerical behavior of the smoothing procedure. For QMR smoothing we have $0 \leq \sigma_n \leq 1$, and the same is true for MR smoothing provided $\mathbf{s}_{n-1} \perp \mathbf{r}_n$ $(\forall n)$, which holds in particular when the residuals of the primary method are mutually orthogonal, such as in CG or FOM. However, as we will see here, in general, when MR smoothing is applied to nonorthogonal residuals, the local rounding errors may get amplified due to the size of the smoothing coefficients; this case seems to be rare in practice, but we can construct examples with such a behavior.

To identify such a case, we assume real vectors and consider σ_n of (2.7) as a function of $\gamma :\equiv \langle \mathbf{s}_{n-1}, \mathbf{r}_n \rangle$:

$$\sigma_n(\gamma) :\equiv \frac{\|\mathbf{s}_{n-1}\|^2 - \gamma}{\|\mathbf{s}_{n-1}\|^2 + \|\mathbf{r}_n\|^2 - 2\gamma} . \tag{3.1}$$

Assuming $\mathbf{s}_{n-1} \neq \mathbf{r}_n$ we see that $|\sigma_n(\gamma)| > 1$ if and only if one of the following two cases hold:

$$\begin{cases} \sigma_n(\gamma) > 1 & \Longleftrightarrow \quad \gamma > \|\mathbf{r}_n\|^2 , \\ \sigma_n(\gamma) < -1 & \Longleftrightarrow \quad \gamma > \frac{2}{3}\|\mathbf{s}_{n-1}\|^2 + \frac{1}{3}\|\mathbf{r}_n\|^2 . \end{cases} \tag{3.2}$$

We are going to choose the first case for constructing examples. For most methods and examples the inner product $\gamma = \langle \mathbf{s}_{n-1}, \mathbf{r}_n \rangle$ is close to zero. In particular, when smoothing is applied to orthogonal residual methods like CG or FOM, this inner product is usually close to the level of machine precision. But in a method with non-orthogonal residuals the smoothing coefficients $\sigma_n = \sigma_n(\gamma)$ may be large when $\mathbf{s}_{n-1}$ and $\mathbf{r}_n$ are nearly parallel. Making use of this observation we can construct examples of MR smoothing where some large σ_n amplify the rounding errors in the recurrences for the smoothed residuals and iterates. In

such cases QMR smoothing is superior to MR smoothing. But we stress again that this behavior is rare and hardly appears in practice when methods exhibit strongly oscillating residual norm histories. In fact, such oscillations indicate that the vectors $\mathbf{r}_n$ and $\mathbf{s}_{n-1}$ differ significantly.

We now describe how we can construct examples with this behavior. First, solving (3.1) for γ we obtain

$$\gamma = \frac{(\sigma_n - 1)\|\mathbf{s}_{n-1}\|^2 + \sigma_n \|\mathbf{r}_n\|^2}{2\sigma_n - 1}. \tag{3.3}$$

Choosing $\sigma :\equiv \sigma_n > 1$ independent of n, we can write the condition $\gamma > \|\mathbf{r}_n\|^2$ of (3.2), complemented by the Schwarz inequality $\gamma \le \|\mathbf{r}_n\|\|\mathbf{s}_{n-1}\|$, as

$$\|\mathbf{r}_n\|^2 < \frac{(\sigma - 1)\|\mathbf{s}_{n-1}\|^2 + \sigma\|\mathbf{r}_n\|^2}{2\sigma - 1} \le \|\mathbf{r}_n\|\|\mathbf{s}_{n-1}\|.$$

The second inequality can be expressed as $(\sigma - 1)\|\mathbf{s}_{n-1}\|^2 + \sigma\|\mathbf{r}_n\|^2 - (2\sigma - 1)\|\mathbf{r}_n\|\|\mathbf{s}_{n-1}\| \le 0$. Here, the quadratic function of $\|\mathbf{r}_n\|$ on the left vanishes at $\|\mathbf{r}_n\| = \|\mathbf{s}_{n-1}\|$ and $\|\mathbf{r}_n\| = (\sigma - 1)/\sigma \, \|\mathbf{s}_{n-1}\|$, and its (negative) minimum is at

$$\|\mathbf{r}_n\| = \frac{2\sigma - 1}{2\sigma}\|\mathbf{s}_{n-1}\| \, . \tag{3.4}$$

We can enforce this relation by starting with $\mathbf{s}_0 = \mathbf{r}_0 = \mathbf{e}_1$, and, at step n, defining the primary residual $\mathbf{r}_n$ by, e.g.,

$$\mathbf{r}_n = \frac{8\sigma^2 - 8\sigma + 1}{8\sigma^2 - 4\sigma}\mathbf{s}_{n-1} + \eta \mathbf{e}_{n+1} \, , \tag{3.5}$$

where $\mathbf{e}_{n+1}$ is the $(n+1)$st unit vector and where the parameter η is chosen such that (3.4) holds. This is possible since, according to (2.1) and (3.5), $\mathbf{s}_{n-1} \in \operatorname{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \perp \mathbf{e}_{n+1}$ and since the fraction in (3.5) can be seen to be smaller than the one in (3.4), so that we can set, by Pythagoras's theorem,

$$\eta = \sqrt{\left(\frac{2\sigma - 1}{2\sigma}\right)^2 - \left(\frac{8\sigma^2 - 8\sigma + 1}{8\sigma^2 - 4\sigma}\right)^2}.$$

Given any nonsingular system matrix $\mathbf{A}$, in theory the primary iterate $\mathbf{x}_n$ that corresponds to the residual $\mathbf{r}_n$ constructed according to (3.5) could be found from $\mathbf{x}_n = \mathbf{A}^{-1}(\mathbf{b} - \mathbf{r}_n)$. Using this construction we also achieve that there is almost no primary gap (that is, difference between the updated primary residuals generated according to (3.5) and the true primary residuals). In any case, the primary gap is of order ϵ. The smoothed approximate solutions $\mathbf{y}_n$ and residuals $\mathbf{s}_n$ are then computed as in (2.1), and once $\mathbf{s}_n$ is known, (3.5) provides $\mathbf{r}_{n+1}$. Generating such a sequence of primary residuals and applying MR smoothing to them we thus obtain, in exact arithmetic, the given smoothing parameters $\sigma = \sigma_n$. However, note that the primary iterates and residuals are not the result of a common iterative method; but they can always be understood

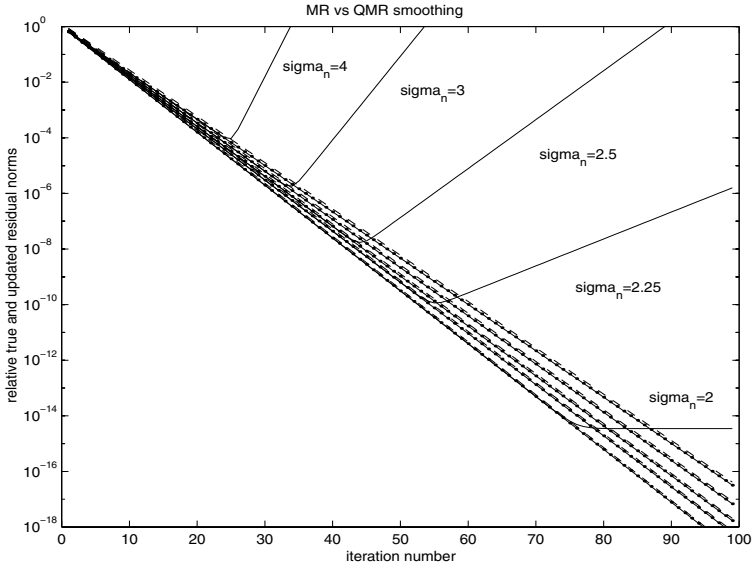

Figure 3.1: Relative norms of updated (solid lines with dots) and true (solid lines) smoothed residuals obtained by MR smoothing of the sequence of residuals constructed according to (3.5) (relative norms shown as dashed lines) for various choices of the smoothing coefficients $\sigma = \sigma_n$ (independent of n).

as the result of a Krylov space solver, since here, the residuals are clearly linearly independent. In fact, such a solver can be viewed as determined by an upper Hessenberg matrix whose column sums are equal to zero; see Section 4.3 of [15]. Given a prescribed sequence of linearly independent residuals, one can construct such a matrix, so that the corresponding Krylov space method generates these residuals. But in the present context the exact specification of the primary iterative method is not important.

Figure 3.1 shows the results of a numerical experiment. We see that despite the fact that the primary gap is close to the level of the machine precision ϵ, for the smoothed residuals there is a strong divergence for various values of σ that exceed 2. If instead, we applied QMR smoothing to the same sequence of primary residuals, then the gap between the true and updated smoothed residuals would become almost invisible.

As we will verify later, it is important to have smoothing coefficients smaller than 2 so that $|1 - \sigma_n| < 1$. In the following we will analyze QMR smoothing, where $0 \leq \sigma_n \leq 1$. In MR smoothing the critical situation where $|1 - \sigma_n| \geq 1$ can be avoided by redefining $\sigma_n :\equiv 1$ whenever the value of (2.7) is bigger than 1, and $\sigma_n :\equiv 0$ whenever this value is negative. We will refer to this as *stabilized MR smoothing*.

4 Relations between computed primary and smoothed residuals.

In this section we study the effects of rounding errors on the relationship between the norms of the recursively computed primary and smoothed residuals. We consider QMR smoothing, but the results remain valid for stabilized MR smoothing. First, we compare the norm of the primary residuals, computed by updating, with the norm of the computed (smoothed) quasi-residual. A relationship similar to (2.15) is given. From this relationship we can conclude that the peak-plateau connection still holds to a close approximation in finite precision arithmetic. Then we show that, up to terms proportional to the machine precision, the inequality (2.18) between the norms of primary and smoothed residuals holds also in finite precision arithmetic.

Our approach is based on introducing an exact smoothing procedure applied to the set of computed primary residuals. We show that the smoothed residuals obtained from this exact procedure are close to the smoothed residuals computed in finite precision arithmetic. Using bounds for the small differences and the fact that the smoothed residuals and the quasi-residuals associated with this exact smoothing procedure satisfy the arithmetic relations (2.15) and (2.18) exactly, we can derive the finite precision analogs of these relations. All three implementations mentioned in Section 1 are considered, and it is shown that there is no substantial difference between them with respect to their local behavior.

4.1 Schönauer–Weiss implementation.

In the *Schönauer–Weiss implementation* the recurrences for the actually computed quantities have the following form:

$$\bar{\mathbf{y}}_0 = \bar{\mathbf{x}}_0\,, \qquad \bar{\mathbf{y}}_n = (1-\bar{\sigma}_n)\bar{\mathbf{y}}_{n-1} + \bar{\sigma}_n\bar{\mathbf{x}}_n + \delta\mathbf{y}_n\,, \tag{4.1}$$

$$\bar{\mathbf{s}}_0 = \bar{\mathbf{r}}_0\,, \qquad \bar{\mathbf{s}}_n = (1-\bar{\sigma}_n)\bar{\mathbf{s}}_{n-1} + \bar{\sigma}_n\bar{\mathbf{r}}_n + \delta\mathbf{s}_n\,, \tag{4.2}$$

where $\delta\mathbf{y}_n$ and $\delta\mathbf{s}_n$ represent the local errors produced at the step n. Applying the standard rules for finite precision arithmetic we find

$$\|\delta\mathbf{y}_n\| \le 3\epsilon|1-\bar{\sigma}_n|\|\bar{\mathbf{y}}_{n-1}\| + 2\epsilon|\bar{\sigma}_n|\|\bar{\mathbf{x}}_n\| + \mathcal{O}(\epsilon^2), \tag{4.3}$$

$$\|\delta\mathbf{s}_n\| \le 3\epsilon|1-\bar{\sigma}_n|\|\bar{\mathbf{s}}_{n-1}\| + 2\epsilon|\bar{\sigma}_n|\|\bar{\mathbf{r}}_n\| + \mathcal{O}(\epsilon^2). \tag{4.4}$$

Here, $\bar{\sigma}_n$ denotes the floating-point result of (2.13) with τ_n^2 obtained from the recurrence (2.14). The square of the quasi-residual norm that is computed as a byproduct of recurrence (2.14) will be denoted by $\bar{\tau}_n^2$, and the value of $1/\tau_n^2$ computed before will be referred to as $\bar{\chi}_n$. We will not need to compute $\bar{\tau}_n$ itself.

THEOREM 4.1. *In the Schönauer–Weiss implementation the computed square $\bar{\tau}_n^2$ of the (smoothed) quasi-residual norm satisfies*

$$\bar{\tau}_n^2 = \hat{\tau}_n^2 + \delta\tau_n, \qquad |\delta\tau_n| \le (N+n+3)\epsilon\hat{\tau}_n^2 + \mathcal{O}(\epsilon^2), \tag{4.5}$$

where $\hat{\tau}_n$ is the exact quasi-residual norm defined by the recursion

$$\hat{\tau}_0 = \|\bar{\mathbf{r}}_0\|, \qquad \frac{1}{\hat{\tau}_k^2} = \frac{1}{\hat{\tau}_{k-1}^2} + \frac{1}{\|\bar{\mathbf{r}}_k\|^2} \qquad (k=1,\ldots,n). \tag{4.6}$$

PROOF. For the computed value $\bar{\chi}_0$ of $1/\tau_0^2 = 1/\|\mathbf{r}_0\|^2$, the standard rules for floating-point error estimates yield $|\bar{\chi}_0 - 1/\|\bar{\mathbf{r}}_0\|^2| \le (N+1)\epsilon/\|\bar{\mathbf{r}}_0\|^2 + \mathcal{O}(\epsilon^2)$. For the later values $\bar{\chi}_n$ we can show that

$$\left|\bar{\chi}_n - \left(\bar{\chi}_{n-1} + \mathrm{fl}\left(\frac{1}{\|\bar{\mathbf{r}}_n\|^2}\right)\right)\right| \le \epsilon|\bar{\chi}_{n-1}| + (N+2)\epsilon\frac{1}{\|\bar{\mathbf{r}}_n\|^2} + \mathcal{O}(\epsilon^2). \tag{4.7}$$

After some manipulation, we obtain

$$\left|\bar{\chi}_n - \sum_{k=0}^{n}\frac{1}{\|\bar{\mathbf{r}}_k\|^2}\right| \le (N+n+3)\epsilon\sum_{k=0}^{n}\frac{1}{\|\bar{\mathbf{r}}_k\|^2} + \mathcal{O}(\epsilon^2). \tag{4.8}$$

Finally, when taking into account the rounding error in computing the reciprocal of $\bar{\chi}_n$, we find the bound given in (4.5). □

Next, using a similar approach as in [5] we can establish a relation between the norms of the updated primary residuals and the computed smoothed quasi-residuals. The following theorem holds.

THEOREM 4.2. *In the Schönauer–Weiss implementation the norm of the computed (smoothed) quasi-residuals and the norm of the primary residual computed by updating are related by*

$$\|\bar{\mathbf{r}}_n\| = \sqrt{\frac{\bar{\tau}_n^2}{1-\frac{\bar{\tau}_n^2}{\bar{\tau}_{n-1}^2}}} + \frac{\frac{\bar{\tau}_n^2}{\bar{\tau}_{n-1}^2}\delta\tau_{n-1} - \delta\tau_n}{2\sqrt{\left(1-\frac{\bar{\tau}_n^2}{\bar{\tau}_{n-1}^2}\right)^3}} + \mathcal{O}(\delta^2), \tag{4.9}$$

where $\mathcal{O}(\delta^2)$ stands for the higher-order terms in $\delta\tau_n$ and $\delta\tau_{n-1}$.

PROOF. The proof is analogous to the one of Theorem 4 in [5]. The exact quasi-residual norms $\hat{\tau}_n$ satisfy

$$\|\bar{\mathbf{r}}_n\| = \sqrt{\frac{\hat{\tau}_n^2}{1-(\hat{\tau}_n^2/\hat{\tau}_{n-1}^2)}}\,.$$

Substituting (4.5) for $\hat{\tau}_n^2$ and $\hat{\tau}_{n-1}^2$ in (4.6) we get the desired result after some manipulation. □

Using (4.5) and standard rounding error analysis we can write the computed smoothing parameters $\bar{\sigma}_n$ in the form

$$\bar{\sigma}_n = \hat{\sigma}_n + \delta\sigma_n, \qquad |\delta\sigma_n| \le (N+n+5)\epsilon\hat{\sigma}_n + \mathcal{O}(\epsilon^2), \tag{4.10}$$

where the coefficient $\hat{\sigma}_n$ is defined as $\hat{\sigma}_n :\equiv \hat{\tau}_n^2/\|\bar{\mathbf{r}}_n\|^2$. We use this coefficient to define an exact smoothing procedure applied to a sequence of computed primary residuals. The corresponding smoothed residuals are then given by the recurrence

$$\hat{\mathbf{s}}_0 = \bar{\mathbf{r}}_0\,, \qquad \hat{\mathbf{s}}_n = (1-\hat{\sigma}_n)\hat{\mathbf{s}}_{n-1} + \hat{\sigma}_n\bar{\mathbf{r}}_n\,. \tag{4.11}$$

THEOREM 4.3. *In the Schönauer–Weiss implementation the norm of the computed smoothed residual is related to the norm of the computed primary residual by*

$$\|\bar{\mathbf{s}}_n\| \le \sqrt{n+1}\,\min_{k=0,\ldots,n}\{\|\bar{\mathbf{r}}_k\|\} + n(N+n+8)(\sqrt{n}+1)\epsilon\|\bar{\mathbf{r}}_0\| + \mathcal{O}(\epsilon^2). \tag{4.12}$$

PROOF. The difference between the computed residual $\bar{\mathbf{s}}_n$ and the exactly smoothed residual $\hat{\mathbf{s}}_n$ can be written as

$$\bar{\mathbf{s}}_n - \hat{\mathbf{s}}_n = (1-\bar{\sigma}_n)(\bar{\mathbf{s}}_{n-1} - \hat{\mathbf{s}}_{n-1}) + \delta\sigma_n(\bar{\mathbf{r}}_n - \hat{\mathbf{s}}_{n-1}) + \delta\mathbf{s}_n. \tag{4.13}$$

Taking norms on both sides, using the bound (4.4), and applying this inequality recursively we get

$$\begin{aligned}\|\bar{\mathbf{s}}_n - \hat{\mathbf{s}}_n\| \le \sum_{k=1}^{n}\prod_{j=k+1}^{n}|1-\bar{\sigma}_j|(1+3\epsilon)\big[&|\delta\sigma_k|(\|\bar{\mathbf{r}}_k\| + \|\hat{\mathbf{s}}_{k-1}\|)\\ &+2\epsilon|\bar{\sigma}_n|\|\bar{\mathbf{r}}_k\| + 3\epsilon|1-\bar{\sigma}_k|\|\hat{\mathbf{s}}_{k-1}\|\big].\end{aligned} \tag{4.14}$$

Using the relation (4.10) we can bound the terms $|1-\bar{\sigma}_j|$ by 1. After some manipulation the inequality (4.14) leads us then to

$$\|\bar{\mathbf{s}}_n - \hat{\mathbf{s}}_n\| \le \sum_{k=1}^{n}\big[|\delta\sigma_k|(\|\bar{\mathbf{r}}_k\| + \|\hat{\mathbf{s}}_{k-1}\|) + 3\epsilon\|\hat{\mathbf{s}}_{k-1}\| + 2\epsilon\hat{\sigma}_k\|\bar{\mathbf{r}}_k\|\big] + \mathcal{O}(\epsilon^2). \tag{4.15}$$

Noting that for the exactly smoothed residual we have by (2.18) and (2.13)

$$\|\hat{\mathbf{s}}_n\| \le \sqrt{n+1}\,\hat{\tau}_n = \sqrt{\frac{1}{\frac{1}{n+1}\sum_{k=0}^{n}\frac{1}{\|\bar{\mathbf{r}}_k\|^2}}} \le \sqrt{n+1}\,\min_{k=0,\ldots,n}\{\|\bar{\mathbf{r}}_k\|\} \tag{4.16}$$

and using

$$\hat{\sigma}_k\|\bar{\mathbf{r}}_k\| = \frac{\tau_k^2}{\|\bar{\mathbf{r}}_k\|} \le \frac{1}{\|\bar{\mathbf{r}}_k\|}\left(\min_{j=0,\ldots,k}\{\|\bar{\mathbf{r}}_j\|\}\right)^2 \le \min_{j=0,\ldots,k}\{\|\bar{\mathbf{r}}_j\|\} \le \|\bar{\mathbf{r}}_0\| \tag{4.17}$$

(obtained from (2.13) and (2.17)) and $\|\hat{\mathbf{s}}_{k-1}\| \le \sqrt{k}\,\min_{j=0,\ldots,k-1}\{\|\bar{\mathbf{r}}_j\|\}$ ($k = 1,\ldots,n$) (obtained from (2.18)) we finally find the statement of the theorem. □

The statement of Theorem 4.3 shows that the relationship (2.18) between the norms of smoothed and primary residuals holds to a close approximation also in finite precision arithmetic provided the roundoff term in (4.12) is much smaller than the norms of the computed primary residuals, that is

$$\min_{k=0,\ldots,n}\{\|\bar{\mathbf{r}}_k\|\} \gg n(N+n+8)(\sqrt{n}+1)\epsilon\|\bar{\mathbf{r}}_0\| \qquad (n \ge 0). \tag{4.18}$$

Since we are in this section only interested in the local behavior of the smoothing procedure above the level of the limiting accuracy, this requirement does not

mean a restriction. As we will see later, the level of $n(N+n+8)(\sqrt{n}+1)\epsilon\|\bar{\mathbf{r}}_0\|$ is usually below or equal to the maximum attainable accuracy of any iterative Krylov subspace method. We will return to this point later.

If we consider instead the *expensive Schönauer–Weiss implementation*, the recurrence in (4.2) must be replaced by

$$\bar{\mathbf{s}}_0 = \bar{\mathbf{t}}_0\,, \qquad \bar{\mathbf{s}}_n = (1-\bar{\sigma}_n)\bar{\mathbf{s}}_{n-1} + \bar{\sigma}_n\bar{\mathbf{t}}_n + \delta\mathbf{s}_n\,, \tag{4.19}$$

where $\bar{\mathbf{t}}_n$ is the residual computed directly from the approximate solution $\bar{\mathbf{x}}_n$ and satisfying the relation

$$\|\bar{\mathbf{t}}_n - (\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_n)\| \le \left[(mN^{1/2}+1)\|\mathbf{A}\|\|\bar{\mathbf{x}}_n\| + \|\mathbf{b}\|\right]\epsilon + \mathcal{O}(\epsilon^2). \tag{4.20}$$

It is easy to see that by replacing the vectors $\bar{\mathbf{r}}_k$ by $\bar{\mathbf{t}}_k$, $k = 0, \ldots, n$, we obtain for this implementation the same results as for the Schönauer–Weiss implementation except that the norms $\|\bar{\mathbf{r}}_k\|$ have to be replaced by $\|\bar{\mathbf{t}}_k\|$ in the statements of Theorems 4.1–4.3.

4.2 Zhou–Walker implementation.

Similar results hold for the *Zhou–Walker implementation*. In finite precision arithmetic the recurrences (2.4)–(2.5) must be written as

$$\bar{\mathbf{v}}_0 = \mathbf{0}\,, \qquad \bar{\mathbf{v}}_n = (1-\bar{\sigma}_{n-1})\bar{\mathbf{v}}_{n-1} + \bar{\alpha}_{n-1}\bar{\mathbf{p}}_{n-1} + \delta\mathbf{v}_n\,, \tag{4.21}$$

$$\bar{\mathbf{u}}_0 = \mathbf{0}\,, \qquad \bar{\mathbf{u}}_n = (1-\bar{\sigma}_{n-1})\bar{\mathbf{u}}_{n-1} + \bar{\alpha}_{n-1}\mathbf{A}\bar{\mathbf{p}}_{n-1} + \delta\mathbf{u}_n\,, \tag{4.22}$$

$$\bar{\mathbf{y}}_0 = \mathbf{x}_0\,, \qquad \bar{\mathbf{y}}_n = \bar{\mathbf{y}}_{n-1} + \bar{\sigma}_n\bar{\mathbf{v}}_n + \delta\mathbf{y}_n\,, \tag{4.23}$$

$$\bar{\mathbf{s}}_0 = \bar{\mathbf{r}}_0\,, \qquad \bar{\mathbf{s}}_n = \bar{\mathbf{s}}_{n-1} - \bar{\sigma}_n\bar{\mathbf{u}}_n + \delta\mathbf{s}_n\,, \tag{4.24}$$

where the local errors $\delta\mathbf{v}_n$, $\delta\mathbf{u}_n$, $\delta\mathbf{y}_n$, and $\delta\mathbf{s}_n$ can be bounded according to

$$\|\delta\mathbf{v}_n\| \le 3\epsilon|1-\bar{\sigma}_{n-1}|\|\bar{\mathbf{v}}_{n-1}\| + 2\epsilon\|\bar{\alpha}_{n-1}\bar{\mathbf{p}}_{n-1}\| + \mathcal{O}(\epsilon^2), \tag{4.25}$$

$$\|\delta\mathbf{u}_n\| \le 3\epsilon|1-\bar{\sigma}_{n-1}|\|\bar{\mathbf{u}}_{n-1}\| + (2+mN^{1/2})\epsilon\|\mathbf{A}\|\|\bar{\alpha}_{n-1}\bar{\mathbf{p}}_{n-1}\| + \mathcal{O}(\epsilon^2), \tag{4.26}$$

$$\|\delta\mathbf{y}_n\| \le \epsilon\|\bar{\mathbf{y}}_{n-1}\| + 2\epsilon|\bar{\sigma}_n|\|\bar{\mathbf{v}}_n\| + \mathcal{O}(\epsilon^2), \tag{4.27}$$

$$\|\delta\mathbf{s}_n\| \le \epsilon\|\bar{\mathbf{s}}_{n-1}\| + 2\epsilon|\bar{\sigma}_n|\|\bar{\mathbf{u}}_n\| + \mathcal{O}(\epsilon^2). \tag{4.28}$$

For QMR smoothing, the smoothing coefficient $\bar{\sigma}_n$ is the floating-point result of the computation (2.13), where, however, the primary residuals may have come from (2.6), and thus their computed values $\bar{\mathbf{r}}_k$, $k = 1, \ldots, n$, satisfy

$$\|\bar{\mathbf{r}}_k - (\bar{\mathbf{s}}_{k-1} - \bar{\mathbf{u}}_k)\| \le \epsilon\|\bar{\mathbf{s}}_{k-1} - \bar{\mathbf{u}}_k\|. \tag{4.29}$$

Consequently, the smoothing coefficient $\bar{\sigma}_n$ can be written as

$$\bar{\sigma}_n = \hat{\sigma}_n + \delta\sigma_n, \qquad |\delta\sigma_n| \le (N+n+5)\epsilon\hat{\sigma}_n + \mathcal{O}(\epsilon^2), \tag{4.30}$$

where the 'exact' smoothing coefficient $\hat{\sigma}_n$ is now defined by

$$\hat{\sigma}_n :\equiv \frac{\hat{\tau}_n^2}{\|\bar{\mathbf{r}}_n\|^2} \qquad \text{with} \quad \frac{1}{\hat{\tau}_n^2} = \sum_{k=0}^{n}\frac{1}{\|\bar{\mathbf{r}}_k\|^2}\,, \quad \hat{\tau}_n > 0\,. \tag{4.31}$$

We introduce again an exact smoothing procedure with residual recurrence

$$\hat{\mathbf{s}}_0 = \bar{\mathbf{r}}_0\,, \qquad \hat{\mathbf{s}}_n = (1-\hat{\sigma}_n)\hat{\mathbf{s}}_{n-1} + \hat{\sigma}_n \bar{\mathbf{r}}_n\,. \tag{4.32}$$

Using a similar construction as for the Schönauer–Weiss implementation the following analog of Theorem 4.3 can be derived.

THEOREM 4.4. *In the Zhou–Walker implementation the norm of the computed smoothed residuals are related to the norms of the computed primary residuals by*

$$\|\bar{\mathbf{s}}_n\| \le \sqrt{n+1}\min_{k=0,\ldots,n}\{\|\bar{\mathbf{r}}_k\|\} + n(N+n+8)(\sqrt{n}+1)\epsilon\|\bar{\mathbf{r}}_0\| + \mathcal{O}(\epsilon^2). \tag{4.33}$$

PROOF. See [16]. □

For MR smoothing we can show analogous results assuming that $0 < \sigma_n < 2$ for all n. This can be achieved by stabilizing the MR smoothing with a small modification of the code, as we mentioned at the end of Section 3.

5 Maximum attainable accuracy of the smoothed method.

For many methods the norm of the updated residuals $\bar{\mathbf{r}}_n$ decreases far below the level of machine precision ϵ, and, indeed, there is a heuristic argument to explain this: often, an iterative algorithm scales with respect to the size of the residuals, that is, if we restart it from an initial residual that is smaller by a factor 2^{-M}, then, even in finite precision arithmetic all the following residuals will be as much smaller; in particular, in such an algorithm there is no feedback from the iterates to the residuals. This argument does not prove that updated residuals decrease till they reach underflow — since continuing an iteration is not the same as restarting it — but it makes such a behavior plausible.

On the other hand, we must expect that there is a limitation to the accuracy of the norm of the true residuals $\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_n$. In fact, typically this norm stagnates from a certain point on. We say then that the level of the maximum attainable accuracy has been reached. For certain classes of methods it was shown in [26], [12], and [18] that the gap $\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_n - \bar{\mathbf{r}}_n$ can increase during the iteration, and may even do so drastically in certain methods and examples. Assuming that the norm of the updated residual converges to 0, we obtain from estimates for this gap an estimate for the level of the maximum attainable accuracy of a particular method. Note that in the stagnation phase the true residual may even increase again, an effect that is often observed when local errors are amplified and the updated residuals oscillate.

The analysis of Greenbaum [12] shows that for algorithms using recursions of the form (2.3) the norm of the gap can be bounded by

$$\begin{aligned}\|\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_n - \bar{\mathbf{r}}_n\| \le{}& 2n(3+mN^{1/2})\epsilon\|\mathbf{A}\|\max_{k=0,\ldots,n}\{\|\mathbf{x}-\bar{\mathbf{x}}_k\|\}\\ &+ (1+mN^{1/2})\epsilon\|\mathbf{A}\|\|\mathbf{x}_0\|\\ &+ (n+1)\epsilon\|\mathbf{A}\|\|\mathbf{x}\| + \mathcal{O}(\epsilon^2)\end{aligned} \tag{5.1}$$

$$\begin{aligned} &\le 2n(3+mN^{1/2})\epsilon\kappa(\mathbf{A}) \max_{k=0,\dots,n}\{\|\bar{\mathbf{r}}_k\|\} \\ &\quad + (1+mN^{1/2})\epsilon\kappa(\mathbf{A})\|\mathbf{r}_0\| \\ &\quad + (n+1)\epsilon\kappa(\mathbf{A})\|\mathbf{b}\| + \mathcal{O}(\epsilon^2). \end{aligned} \tag{5.2}$$

Consequently, the maximum attainable accuracy of these algorithms, measured by Greenbaum [12] in terms of the quantity

$$\frac{\|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n-\bar{\mathbf{r}}_n\|}{\|\mathbf{A}\|\,\|\mathbf{x}\|}, \tag{5.3}$$

depends on the largest norm of the error (of the approximate solutions) during the full iteration. It can be bounded further in terms of the largest residual norm, as in (5.2). For the conjugate gradient (CG) and the conjugate residual (CR) methods, where the error norm or the residual norm, respectively, are known to converge monotonously, these bounds depend actually on the initial error or residual, respectively. In contrast, there are other Krylov subspace methods (such as BiCG or CGS) that are well-known to often produce very large intermediate residuals and approximate solutions. This will then affect their maximum attainable accuracy. Since the smoothing techniques avoid such large intermediate residuals, one may wonder whether the maximum attainable accuracy of smoothed residuals can be much higher than that of a primary method affected by large oscillations in the residual norm. Our aim is to answer this question.

Note that Greenbaum's measure (5.3) differs from ours by a factor $(\|\mathbf{A}\|\,\|\mathbf{x}\|)^{-1}$. Its rationale is that when (5.3) is a small multiple of ϵ, the computed approximate solution $\bar{\mathbf{x}}_n$ is known to be backward stable.

In this section we give bounds for local errors that appear in the recurrences for the quantities computed in the smoothing method. It is shown that although we deal with potentially large primary residuals and iterates, the smoothing techniques keep the local errors small. Then we analyze the gap between the smoothed true and updated residuals and show that for both the Schönauer–Weiss implementation (2.1) and the Zhou–Walker implementation (2.3), this gap remains on the same level as that of the primary method. Only for the expensive implementation (2.2) it is close to the level of machine precision, so that the gap between the true and updated residuals is almost invisible. Nevertheless, for all three variants of smoothing discussed, the maximum attainable accuracy remains at least on the same level as for the primary method.

In accordance with the special case (5.3) we make the general assumption that the primary gap is of order ϵ:

$$\|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n-\bar{\mathbf{r}}_n\| \le \mathcal{O}(\epsilon). \tag{5.4}$$

The constant in front of ϵ need not be small; when the gap is multiplied by another factor $\mathcal{O}(\epsilon)$, the effect will be covered by the $\mathcal{O}(\epsilon^2)$ term that appears in our bounds. In practice, we need $\|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n-\bar{\mathbf{r}}_n\| \ll \|\mathbf{b}\|+\|\mathbf{A}\|\|\bar{\mathbf{x}}_n\|$.

5.1 Schönauer–Weiss implementation.

For the *Schönauer–Weiss implementation* the recurrences (4.1)–(4.2) are valid in finite precision arithmetic. Starting from them we can establish the following theorem for the norm of the gap $\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n$.

THEOREM 5.1. *In the Schönauer–Weiss implementation of QMR residual smoothing the gap between true and updated smoothed residuals satisfies the recurrence*

$$\begin{aligned}(5.5)\qquad \mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n &= (1-\bar{\sigma}_n)(\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_{n-1} - \bar{\mathbf{s}}_{n-1}) \\ &\quad + \bar{\sigma}_n(\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_n - \bar{\mathbf{r}}_n) + \mathbf{A}\delta\mathbf{y}_n + \delta\mathbf{s}_n,\end{aligned}$$

and thus it is given by

$$\begin{aligned}(5.6)\qquad \mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n &= \sum_{k=0}^{n} \bar{\sigma}_k \Big[\prod_{j=k+1}^{n} (1-\bar{\sigma}_j) \Big] (\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k - \bar{\mathbf{r}}_k) \\ &\quad + \sum_{k=1}^{n} \Big[\prod_{j=k+1}^{n} (1-\bar{\sigma}_j) \Big] (\mathbf{A}\delta\mathbf{y}_k + \delta\mathbf{s}_k)\end{aligned}$$

and bounded according to

$$\begin{aligned}(5.7)\quad \|\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n\| &\le \sum_{k=0}^{n} \frac{\|\mathbf{b} - \mathbf{A}\bar{\mathbf{x}}_k - \bar{\mathbf{r}}_k\|}{\|\bar{\mathbf{r}}_k\|} \min_{j=0,\ldots,k} \{\|\bar{\mathbf{r}}_j\|\} \\ &\quad + 3n(\sqrt{n}+1)\epsilon(\kappa(\mathbf{A})+1)\|\bar{\mathbf{r}}_0\| + 6n\epsilon\kappa(\mathbf{A})\|\mathbf{b}\| + \mathcal{O}(\epsilon^2).\end{aligned}$$

PROOF. The recurrence (5.5) follows from (4.1)–(4.2). Applying it inductively and taking into account $\bar{\mathbf{y}}_0 = \bar{\mathbf{x}}_0 = \mathbf{x}_0$, $\bar{\mathbf{s}}_0 = \bar{\mathbf{r}}_0$, and $\bar{\sigma}_0 = 1$ we obtain (5.6). Taking there norms on both sides and using (4.10), (5.4), and the fact that $|1-\bar{\sigma}_j| < 1$ holds under the assumption of the theorem, we conclude that

$$(5.8)\quad \|\mathbf{b}-\mathbf{A}\bar{\mathbf{y}}_n-\bar{\mathbf{s}}_n\| \le \sum_{k=0}^{n} |\hat{\sigma}_k| \|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_k-\bar{\mathbf{r}}_k\| + \sum_{k=1}^{n} \big(\|\mathbf{A}\|\|\delta\mathbf{y}_k\| + \|\delta\mathbf{s}_k\|\big) + \mathcal{O}(\epsilon^2).$$

For bounding the local error in the computation of the updated smoothed residual we note that, by proceeding similarly as in (4.17), half of the second term on the right-hand side of (4.4) can be bounded by

$$\epsilon|\bar{\sigma}_k|\|\bar{\mathbf{r}}_k\| \le \epsilon\hat{\sigma}_k\|\bar{\mathbf{r}}_k\| + \mathcal{O}(\epsilon^2) \le \epsilon \min_{j=0,\ldots,k} \{\|\bar{\mathbf{r}}_j\|\} + \mathcal{O}(\epsilon^2).$$

Using Theorem 4.3 we can write for one third of the first term

$$\epsilon|1-\bar{\sigma}_k|\|\bar{\mathbf{s}}_{k-1}\| \le \epsilon\|\hat{\mathbf{s}}_{k-1}\| + \mathcal{O}(\epsilon^2) \le \epsilon\sqrt{k} \min_{j=0,\ldots,k} \{\|\bar{\mathbf{r}}_j\|\} + \mathcal{O}(\epsilon^2).$$

In summary, the local error $\delta\mathbf{s}_k$ in the computation of the updated smoothed residual satisfies

$$(5.9)\quad \|\delta\mathbf{s}_k\| \le 3(\sqrt{k}+1)\epsilon \min_{j=0,\ldots,k} \{\|\bar{\mathbf{r}}_j\|\} + \mathcal{O}(\epsilon^2) \le 3(\sqrt{k}+1)\epsilon\|\bar{\mathbf{r}}_0\| + \mathcal{O}(\epsilon^2).$$

Similarly for bounding the error $\delta\mathbf{y}_k$ in the computation of the smoothed approximate solution we note that

$$\epsilon|\bar{\sigma}_k|\|\bar{\mathbf{x}}_k\| \leq \epsilon\hat{\sigma}_k\|\mathbf{x}-\bar{\mathbf{x}}_k\|+\epsilon\|\mathbf{x}\|+\mathcal{O}(\epsilon^2) \leq \epsilon\frac{\|\mathbf{x}-\bar{\mathbf{x}}_k\|}{\|\bar{\mathbf{r}}_k\|}\min_{j=0,\ldots,k}\{\|\bar{\mathbf{r}}_j\|\}+\epsilon\|\mathbf{x}\|+\mathcal{O}(\epsilon^2)$$

and

$$\epsilon|1-\bar{\sigma}_k|\|\bar{\mathbf{y}}_{k-1}\| \leq \epsilon\frac{\|\mathbf{x}-\bar{\mathbf{y}}_{k-1}\|}{\|\bar{\mathbf{s}}_{k-1}\|}\sqrt{k}\min_{j=0,\ldots,k-1}\{\|\bar{\mathbf{r}}_j\|\}+\epsilon\|\mathbf{x}\|+\mathcal{O}(\epsilon^2).$$

Recalling again assumption (5.4) and making the inductive assumption that at step $k-1$ the gap between recursive and true smoothed residuals is also $\mathcal{O}(\epsilon)$, that is, $\|\mathbf{b}-\mathbf{A}\bar{\mathbf{y}}_{k-1}-\bar{\mathbf{s}}_{k-1}\| \leq \mathcal{O}(\epsilon)$, we see that $\|\mathbf{x}-\bar{\mathbf{y}}_{k-1}\|/\|\bar{\mathbf{s}}_{k-1}\| \leq \epsilon\,\|\mathbf{A}^{-1}\|+\mathcal{O}(\epsilon^2)$, so the image $\mathbf{A}\delta\mathbf{y}_k$ of the local error $\delta\mathbf{y}_k$ can be bounded by

$$\|\mathbf{A}\|\|\delta\mathbf{y}_k\| \leq 3(\sqrt{k}+1)\epsilon\kappa(\mathbf{A})\|\bar{\mathbf{r}}_0\|+6\epsilon\|\mathbf{A}\|\|\mathbf{x}\|+\mathcal{O}(\epsilon^2). \tag{5.10}$$

Substituting (5.9) and (5.10) into (5.8) we finally obtain (5.7). □

Theorem 5.1 establishes an explicit formula, (5.6), and a bound, (5.7), for the gap. Making the earlier discussed assumption that the updated primary residuals decay far beyond the level of machine precision ϵ, we can conclude that the same is true for the updated smoothed residuals and in this way estimate the maximum attainable accuracy of the smoothed method. *Theorem* 5.1 *shows that the maximum attainable accuracy is not improved by smoothing.* In fact, on the right-hand side of (5.6) we find as the last term of the sum just $\bar{\sigma}_n(\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n-\bar{\mathbf{r}}_n)$ since the corresponding product $\prod_{j=n+1}^{n}(1-\bar{\sigma}_j)$ is empty (and thus 1). The other terms in the first sum cannot be expected to compensate this one as the gap vectors are unlikely to be collinear, and the products are smaller than 1. The second sum contains just local errors, which are typically rather small compared to the gap. When $\|\bar{\mathbf{r}}_n\| \gg \min_{k=0,\ldots,n}\{\|\bar{\mathbf{r}}_k\|\}$ (that is, when we are at a peak in the residual norm history), it follows from (2.13) and (2.17) that $\bar{\sigma}_n$ is small, so that in fact $\|\mathbf{b}-\mathbf{A}\bar{\mathbf{y}}_n-\bar{\mathbf{s}}_n\| \ll \|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n-\bar{\mathbf{r}}_n\|$ is possible. However, under the assumptions made in [12, 18, 26] (which reflect the behavior of Krylov space methods that use recurrences to update the residuals), the primary gap will not shrink later and stagnates on a certain level, although $\|\bar{\mathbf{r}}_n\| \to 0$ in the sense that $\|\bar{\mathbf{r}}_n\|$ will become much smaller than ϵ. Thus, once $\|\bar{\mathbf{r}}_n\| \approx \min_{k=0,\ldots,n}\{\|\bar{\mathbf{r}}_k\|\}$ (that is, once the peak is left behind and the method converges ultimately), (2.13) and (2.17) yield

$$\frac{1}{n+1} \lesssim \bar{\sigma}_n \lesssim 1 \tag{5.11}$$

and we find that the size of the gap of the smoothed residual has caught up with the primary one: roughly,

$$\frac{1}{n+1}\,\|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n-\bar{\mathbf{r}}_n\| \lesssim \|\mathbf{b}-\mathbf{A}\bar{\mathbf{y}}_n-\bar{\mathbf{s}}_n\| \lesssim \|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n-\bar{\mathbf{r}}_n\|, \tag{5.12}$$

which shows that the sizes of the primary and smoothed gaps are very close, up to terms proportional to ϵ.

The above argument has to be modified if in the primary method measures are taken to reduce the gap after peaks of the residual norm. In [22, 25, 29] it has been proposed to replace updated by true residuals (and also to shift the origin in $\mathbf{x}$-space) at certain steps. Then the ultimate accuracy does no longer depend on the largest $\mathbf{x}_n$ or $\mathbf{r}_n$, but the level $p(n,m,N)\,\epsilon\,\|A\|\,\|x\|$ can be achieved, where $p(n,m,N)$ is a low degree polynomial in the number n of iteration steps, the dimension N of the problem, and the maximum number m of nonzeros per row in the matrix $\mathbf{A}$. Since the primary method is then so accurate, the smoothed method cannot be considerably more accurate, so again, *smoothing will not improve the ultimately attained accuracy.*

When the primary method is based on the recursions (2.3), we can use the results (5.1)–(5.2) of Greenbaum [12] to prove the following corollary.

COROLLARY 5.2. *If the Schönauer–Weiss implementation of QMR smoothing is applied to primary iterates* $\mathbf{x}_n$ *and residuals* $\mathbf{r}_n$ *that are computed by the recursion* (2.3)*, the gap between the true and updated smoothed residuals is bounded according to*

$$\begin{aligned}\|\mathbf{b}-\mathbf{A}\bar{\mathbf{y}}_n-\bar{\mathbf{s}}_n\| \le 2n(3+mN^{1/2})\epsilon\kappa(\mathbf{A})\sum_{k=0}^{n}\frac{\max_{j=0,\ldots,k}\{\|\bar{\mathbf{r}}_j\|\}}{\|\bar{\mathbf{r}}_k\|}\min_{j=0,\ldots,k}\{\|\bar{\mathbf{r}}_j\|\}\\ +(n+1)(mN^{1/2}+3n)\epsilon(\kappa(\mathbf{A})+1)\|\bar{\mathbf{r}}_0\|\\ +(7n+1)\epsilon\kappa(\mathbf{A})\|\mathbf{b}\|.\end{aligned}$$

In accordance with Theorem 5.1 this result shows that when the primary method uses the two-term recursions (2.3), then, unfortunately, the maximum attainable accuracy of the smoothed method again depends on the largest norm of the primary residual. Like in [12] we could rewrite this result in terms of the norms of computed approximate solutions.

The best we can hope for in finite precision arithmetic is that the backward error associated with the computed smoothed approximate solution $\bar{\mathbf{y}}_n$ is on the level of $p(n,m,N)\,\epsilon$, that is, $\|\mathbf{b}-\mathbf{A}\bar{\mathbf{x}}_n\| \le \|\mathbf{A}\|\,\|\bar{\mathbf{x}}_n\|\,p(n,m,N)\,\epsilon$. Then the true smoothed residual stagnates at most on the level of $p(n,m,N)\,\epsilon\,\kappa(\mathbf{A})\|\mathbf{b}\|$. (The opposite need not be true.) If we assume a properly chosen initial approximation $\mathbf{x}_0$, this level is close to the level of $p(n,m,N)\,\epsilon\,\kappa(\mathbf{A})\|\mathbf{r}_0\|$, which appears in our bounds. In practice, $p(n,m,N)\,\kappa(\mathbf{A})$ is often a rough overestimate of the actual factor, which is normally a small constant. Nevertheless, when $\|\bar{\mathbf{x}}_n\| \approx \|\mathbf{x}\| \approx \|\mathbf{A}^{-1}\|\,\|\mathbf{b}\|$, the true smoothed residual may indeed stagnate on the level $p(n,m,N)\,\epsilon\,\kappa(\mathbf{A})\|\mathbf{b}\|$ for a backward stable method; so the bounds that contain the condition number $\kappa(\mathbf{A})$ explicitly or implicitly through $\|\bar{\mathbf{x}}_n\|$ cannot be essentially improved in the general case.

In contrast, for the *expensive Schönauer–Weiss implementation* based on (2.2) the following result can be established.

THEOREM 5.3. *In the expensive Schönauer–Weiss implementation* (2.2) *of QMR residual smoothing the gap between the true and updated smoothed residuals*

is bounded according to

$$\|\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n\| \le (n+1)(mN^{1/2}+1)\epsilon(\kappa(\mathbf{A})+1)\|\bar{\mathbf{r}}_0\| + (n+1)(mN^{1/2}+6)\epsilon(\kappa(\mathbf{A})+1)\|\mathbf{b}\|.$$

This shows that the gap is small for the expensive implementation. This is due to the fact that in this implementation the updated residuals do not appear in the recursions, and we are smoothing the residuals $\bar{\mathbf{t}}_n$ that are computed directly from the iterates. Therefore the gap becomes almost invisible. However, these updated smoothed residuals do not decay to zero, but, like the true smoothed residuals, remain on the level of the true primary residuals. Roughly speaking, while in the original Schönauer–Weiss implementation the updated smoothed residuals are smoothing the sequence of updated primary residuals, which converge usually far beyond machine precision, in the 'expensive implementation' the updated smoothed residuals behave like the primary residuals that are directly computed from the (inaccurate) approximate solutions. The maximum attainable accuracy consequently remains on the same level.

5.2 Zhou–Walker implementation.

For the *Zhou–Walker implementation* the situation is very similar to the one of Schönauer–Weiss. In this subsection we give a bound for the corresponding gap. It implies that ultimately the norm of the gap must be expected to be of the same order as the limiting accuracy of the primary method. In practice, we frequently see here that also the updated smoothed residual stagnates on a certain level. This, of course, does not mean that the Zhou–Walker implementation is better than the Schönauer–Weiss implementation; its maximum attainable accuracy remains on the same level.

THEOREM 5.4. *In the Zhou–Walker implementation of QMR residual smoothing the gap between the true and updated smoothed residuals satisfies the coupled recurrences*

$$\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n = \mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_{n-1} - \bar{\mathbf{s}}_{n-1} + \bar{\sigma}_n(\bar{\mathbf{u}}_n - \mathbf{A}\bar{\mathbf{v}}_n) + \mathbf{A}\delta\mathbf{y}_n + \delta\mathbf{s}_n, \tag{5.13}$$

$$\bar{\mathbf{u}}_n - \mathbf{A}\bar{\mathbf{v}}_n = (1-\bar{\sigma}_n)(\bar{\mathbf{u}}_{n-1} - \mathbf{A}\bar{\mathbf{v}}_{n-1}) + \delta\mathbf{u}_n - \mathbf{A}\delta\mathbf{v}_n, \tag{5.14}$$

which yield the explicit representation

$$\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n = \mathbf{b} - \mathbf{A}\mathbf{x}_0 - \bar{\mathbf{r}}_0 + \sum_{k=1}^{n}\left\{\mathbf{A}\delta\mathbf{y}_k + \delta\mathbf{s}_k + \bar{\sigma}_k \sum_{j=1}^{k}\left[\prod_{l=j+1}^{k}(1-\bar{\sigma}_l)\right](\delta\mathbf{u}_j - \mathbf{A}\delta\mathbf{v}_j)\right\} \tag{5.15}$$

and the bound

$$\|\mathbf{b} - \mathbf{A}\bar{\mathbf{y}}_n - \bar{\mathbf{s}}_n\| \le 2n^2(6+mN^{1/2})\epsilon\kappa(\mathbf{A}) \times \sum_{k=1}^{n}\left(\frac{\max_{j=0,\dots,k}\{\|\bar{\mathbf{r}}_j\|\}}{\|\bar{\mathbf{r}}_k\|}\min_{j=0,\dots,k}\{\|\bar{\mathbf{r}}_j\|\}\right) \tag{5.16}$$

$$+ \left[3n(\sqrt{n}+1)+1+mN^{1/2}\right]\epsilon(\kappa(\mathbf{A})+1)\|\bar{\mathbf{r}}_0\|$$
$$+ \left[n+1+mN^{1/2}\right]\epsilon(\kappa(\mathbf{A})+1)\|\mathbf{b}\|.$$

PROOF. See [16]. □

Theorem 5.4 again establishes an explicit formula, (5.15), and a bound, (5.16), for the gap, and under the previously made assumption of ultimately very small primary residuals, these formulas then also describe the maximum attainable accuracy. From *Theorem* 5.4 *and its proof we can conclude that also for the Zhou–Walker implementation this maximum attainable accuracy is not essentially improved by smoothing*. Indeed, on the right-hand side of (5.15) the biggest term is typically $\bar{\sigma}_k(\delta\mathbf{u}_k - \mathbf{A}\delta\mathbf{v}_k)$ if $\bar{\sigma}_k$ lies in the range $(k+1)^{-1} \lesssim \bar{\sigma}_k \lesssim 1$ (see (5.11)) as it happens in the phase of ultimate convergence where $\|\bar{\mathbf{r}}_k\| \approx \min_{j=0,\ldots,k}\{\|\bar{\mathbf{r}}_j\|\}$. In fact, when $j < k$, the terms $\delta\mathbf{u}_j - \mathbf{A}\delta\mathbf{v}_j$ are damped by the product $\prod_{l=j+1}^{k}|1-\bar{\sigma}_l| < 1$. For $j = k$ it can be shown using (4.21)–(4.22) and (4.25)–(4.26) that the norms of $\delta\mathbf{u}_k$ and $\mathbf{A}\delta\mathbf{v}_k$ are of the order $p(k,m,N)\,\epsilon\|\mathbf{A}\|\max_{i=1,\ldots,k}\{\|\bar{\alpha}_{i-1}\bar{\mathbf{p}}_{i-1}\|\}$, where $p(k,m,N)$ is again a low degree polynomial in k, m, and N; so the bound for the gap is dominated by a term of this form. Recall now that in the Zhou–Walker setting the primary iterates and residuals are computed according to (2.3), so by recurrences of the form analyzed by Greenbaum [12], which in finite precision arithmetic lead to bounds of the same form (5.1)–(5.2) for the gap. Consequently, the primary gap is of the same order as the one between true and updated smoothed residuals. For details we refer to [12].

6 Examples and numerical experiments.

We report on two sets of numerical experiments, one with a real, symmetric positive definite (spd) matrix, the other with a real nonsymmetric matrix. Both matrices are from the Harwell–Boeing collection. The spd matrix NOS6 has dimension 675 and condition number 7.6505E+6. It originates from a simple 5-point stencil finite difference approximation of Poisson's equation on an L-shape; the bandwidth is 61, and there are at most 5 nonzeros per row. The nonsymmetric matrix ORSREG1 has dimension 2205 and condition number 1.0E+2. It describes a 3d oil reservoir simulation based on a 7-point stencil finite difference approximation on a regular $21 \times 21 \times 5$ grid. Thus, the bandwidth is 881, and there are at most 7 nonzeros per row. In both cases we choose the right-hand side $\mathbf{b}$ as the vector $\mathbf{e}$ with all components 1, and the initial approximation $\mathbf{x}_0 = \mathbf{0}$.

To the spd system we apply various versions of the conjugate gradient (CG) method of Hestenes and Stiefel [19], followed by QMR smoothing. Both the Schönauer–Weiss and the Zhou–Walker implementations, and also the 'expensive' Schönauer–Weiss implementation are tested. In exact arithmetic we could obtain the same results from CG followed by MR smoothing, but also the related conjugate residual (CR) method [28] and the MINRES algorithm due to

Paige and Saunders [23] would produce the same results. To illustrate the influence of finite precision arithmetic on these algorithms, we show residual norm histories obtained with each of them. Considerable differences in the ultimate accuracy will be noticed. Though the local roundoff behavior differs for the various schemes, this has only in few cases an effect on the speed of convergence. The versions of the CG method investigated are (compare [1]):

(i) the classical Hestenes–Stiefel or CG-OMIN implementation based on three coupled two-term recurrences for the iterates, residuals, and direction vectors,

(ii) the CG-ORES implementation based on two three-term recurrences for the iterates and the residuals, and

(iii) the CG-ODIR implementation based on a three-term recurrence for the direction vectors and two coupled two-term recurrences for the the iterates and the residuals.

The algorithms resulting from piping the CG approximates and residuals into a smoothing process will be denoted by CG-OMIN|MR, CG-OMIN|QMR, and CG-ORES|MR, etc. The type of the implementation of the smoother could be displayed additionally in brackets: [SW], [ZW], or [EXP]. For the conjugate residual method there exist analogous versions CR-OMIN, CR-ORES, and CR-ODIR. Actually one even has to differentiate their implementations further. For versions of CR-OMIN we did not notice relevant differences in the error behavior. But for CR-ODIR it is important that we used the version that requires two matrix-vector multiplications (MVs) for computing $\mathbf{A}\mathbf{p}_n$ and $\mathbf{A}^2\mathbf{p}_n$ in each step (in contrast to applying an extra three-term recurrence for the vectors $\mathbf{A}\mathbf{p}_n$, which produces additional large local errors); we will return to this question elsewhere. The term 'accurate' will refer to the ultimate accuracy of an algorithm, that is the level of stagnation of the norm of the true residual.

To nonsymmetric systems we apply the biconjugate gradient (BICG) method due to Lanczos [21] followed by either MR or QMR smoothing implemented according to Schönauer–Weiss. Specifically, we use the following versions:

(i) the classical BIOMIN implementation of BICG due to Lanczos [21] and Fletcher [6], called Lanczos/ORTHOMIN by Jea and Young [20]; it generalizes CG-OMIN and is also based on coupled two-term recurrences;

(ii) the BIORES implementation [13, 15], called Lanczos/ORTHORES in [20], which analogously to CG-ORES is based on three-term recurrences;

(iii) the BIODIR implementation [13, 15] (similar to Lanczos/ORTHODIR in [20]), which is fully analogous to CG-ODIR.

The resulting combined algorithms will be called BIOMIN|MR, BIOMIN|QMR, and BIORES|MR, etc.

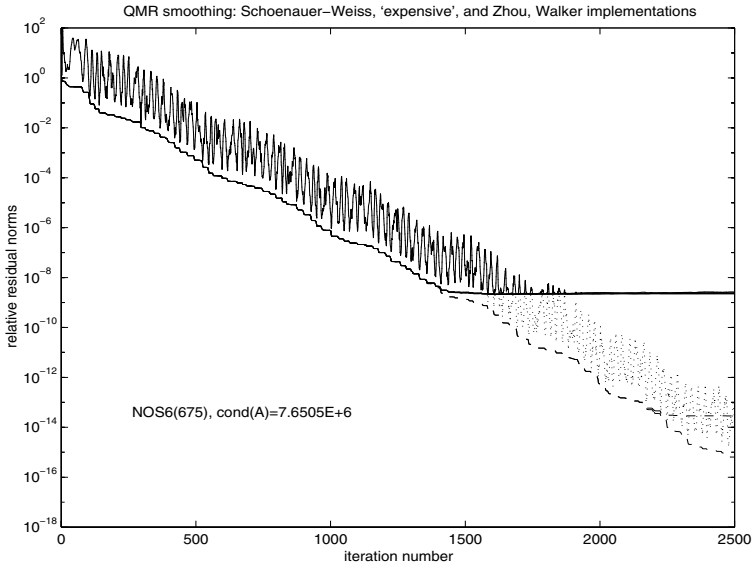

Figure 6.1: QMR smoothing of Hestenes–Stiefel CG (CG-OMin) applied to an spd system with the matrix NOS6. Relative 2-norms of updated (dots) and true (initially oscillating solid line) residuals of CG-OMin. Comparison with three implementations of QMR smoothing applied to the CG results: updated (lower dashed line) and true (solid line) residuals of the Schönauer–Weiss implementation, updated (dashed line ending in lower plateau) and true (solid line) residuals of the Zhou–Walker implementation, and updated (dashed line covered by solid line) and true (solid line) residuals of 'expensive' implementation. On the plateau, the four solid lines cannot be distinguished.

Recall that in exact arithmetic, BiCG|QMR would produce the same results as the QMR method of Freund and Nachtigal [8] if in both cases either no look-ahead or the same look-ahead procedure were applied. Our results published here do not include look-ahead, but we will consider both the original QMR algorithm based on three-term Lanczos recurrences [8] and the one based on coupled two-term recurrences [9, 10, 11]. Unlike in the spd case, in general, BiCG|QMR is not equivalent to BiCG|MR.

We start by solving the spd system with the matrix NOS6 with the Hestenes–Stiefel (OMin) version of CG and applying the three discussed implementations of QMR smoothing ([SW], [ZW], [EXP]) to the CG residuals and iterates. The results are shown in Figure 6.1. Clearly, as expected and predicted by the now well-known analysis based on (2.15) [5], the strongly fluctuating residual norm history of CG is effectively smoothed. However, the most relevant conclusion is that the true residuals (solid lines) of all four algorithms stagnate on the same level. The same would be true for MR-smoothing (not shown). In other words, *smoothing does not increase the ultimate accuracy of* CG. We also note that

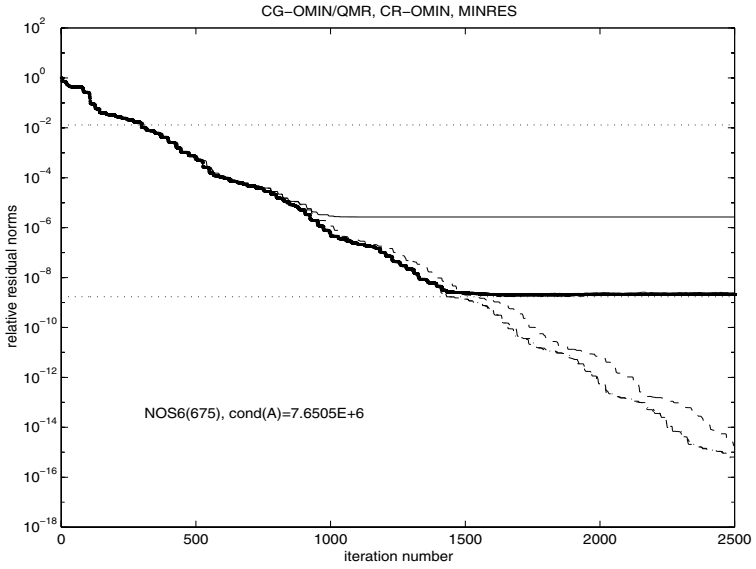

Figure 6.2: Comparison of QMR-smoothed CG-OMin with CR-OMin and MinRes applied to the spd system with the matrix NOS6. Relative 2-norms of true residuals (lower solid line) obtained by QMR smoothing from CG-OMin. Comparison with the norms of the true residuals of CR-OMin (overlapping lower solid line) and MinRes (upper solid line). Also shown are the norms of the updated residuals of CG-OMin|QMR (lower dashed line), and CR-OMin (mostly overlapping dash-dotted line), and the MinRes quantities $\widetilde{\rho}_n$ (upper dashed line), as well as the levels of $\kappa(\mathbf{A})\epsilon$ (lower dotted straight line) and of $\kappa^2(\mathbf{A})\epsilon$ (upper dotted straight line).

the updated smoothed residual norm from the Zhou–Walker implementation stagnates ultimately, but on a much lower level than the true residual. For the expensive implementation, true and updated residuals stagnate on the same level and almost coincide, while the recursive residuals of the Schönauer–Weiss implementation decrease further; however, this is of no relevance for solving the linear system.

In Figure 6.2 we compare various mathematically equivalent residual minimizing algorithms for the solution of the same spd system: MinRes, CR-OMin, and QMR smoothing applied to CG-OMin. Note that CR-OMin and smoothed CG-OMin are equally accurate (overlapping lower solid lines), while MinRes (upper solid line), as has to be expected from the recent analysis of Sleijpen, van der Vorst, and Modersitzki [27], stagnates much earlier. The accuracy of CR-OMin and CG-OMin|QMR is well captured by the estimate $\kappa(\mathbf{A})\epsilon$ (lower dotted straight line), while the level of $\kappa^2(\mathbf{A})\epsilon$ (upper dotted straight line) that one might expect for MinRes from the bound in [27] is here too pessimistic. We also display the norms of the updated residuals of CR-OMin (dash-dotted line) and the quantities $\widetilde{\rho}_n$ of MinRes (dashed line), which is, in exact arith-

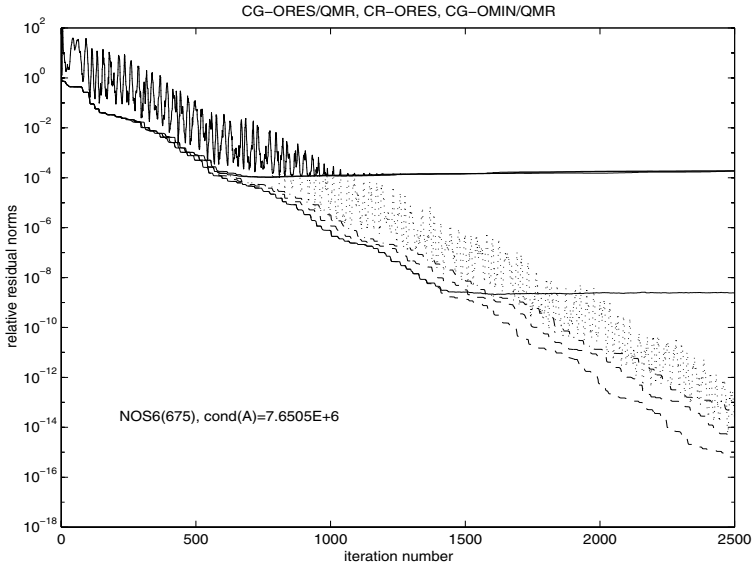

Figure 6.3: QMR smoothing of coupled two-term CG (CG-OMin) vs. three-term CG (CG-ORes), applied to the spd system with the matrix NOS6. Relative 2-norms of updated (dots) and true (initially oscillating solid line) residuals of CG-ORes. Comparison with norms of updated (upper dashed line) and true (upper solid line) residuals of CG-ORes|QMR, and with updated (dash-dotted line) and true (overlapping upper solid line) residual norms of CR-ORes. For reference, the norms of the updated (lower dashed line) and true (lower solid line) residuals of CG-OMin|QMR residuals are shown again.

metic, equal to the residual norm, and comes for free, as a byproduct of the LQ decomposition that is computed. Compared to the updated residuals of CR-OMin and CG-OMin|QMR (which can be understood as differences between $\mathbf{r}_0$ and its computed projection onto a Krylov subspace of growing dimension) the quantities $\widetilde{\rho}_n$ of MinRes seem to be affected by roundoff causing a slight delay of convergence.

Next, we consider CG and CR algorithms based on three-term recurrences. Figure 6.3 shows primarily the relative norms of the updated (dots) and the true (oscillating solid line, stagnating early) residuals of CG-ORes, as well as the corresponding norms obtained by QMR smoothing (upper dashed line and upper solid line, respectively). For comparison, we also show the updated (dash-dotted line) and the true (again mostly the same upper solid line) residual norms of CR-ORes, and the updated (lower dashed line) and true (lower solid line) residual norms of CG-OMin|QMR from Figures 6.1 and 6.2. Clearly, the algorithms based on three-term recurrences are less accurate than CG-OMin of Hestenes and Stiefel, and, again, smoothing does not help to increase the accuracy. There is also a very minor delay of convergence of the three-term versions compared

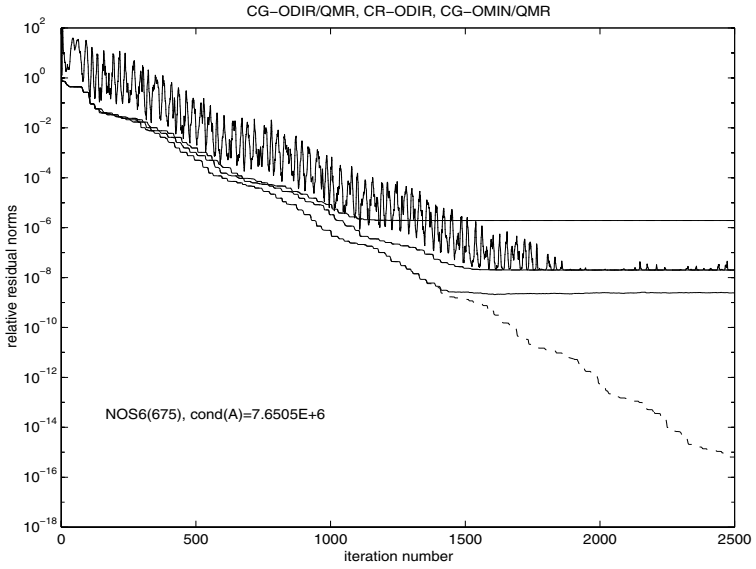

Figure 6.4: QMR smoothing of coupled two-term CG (CG-OMin) vs. coupled two/three-term CG (CG-ODir), applied to the spd system with the matrix NOS6. Relative 2-norms of updated (invisible dots, covered by oscillating solid line) and true (initially oscillating solid line covering the dots) residuals of CG-ODir. Comparison with the norms of updated (covered upper dashed line) and true (middle solid line covering the dashed line) residuals of CG-ODir|QMR, and with updated (dash-dotted line covered by upper solid line) and true (top solid line covering the dash-dotted line) residual norms of CR-ODir. For reference, the norms of the updated (lower, visible dashed line) and the true (bottom solid line) residuals of CG-OMin|QMR are shown once more.

to the two-term one, both for the true and updated residual norms.

In Figure 6.4 we show the results obtained with the ODir version of CG. Remarkably, here norms of the true and updated residuals of the primary method, CG-ODir, overlap (initially oscillating solid line), as do those of CR-ODir (top solid line), and those of smoothed CG-ODir|QMR (middle solid line). Again, QMR smoothing yields no higher ultimate accuracy than the primary method provides, and this accuracy is less than that of CG-OMin and CG-OMin|QMR. Once more, the norms of the true and the updated residuals of the latter algorithm are shown for comparison.

Now we turn to the nonsymmetric system with the matrix ORSREG1. As already anticipated in Schönauer's short note in [24], the natural application of smoothing is to BiCG, and as shown by Zhou and Walker, BiCG|QMR is mathematically equivalent to the QMR method without look-ahead. In Figure 6.5 we first exhibit the effect of smoothing and note how little persists from this equivalence when we turn to finite precision arithmetic. We show the rela-

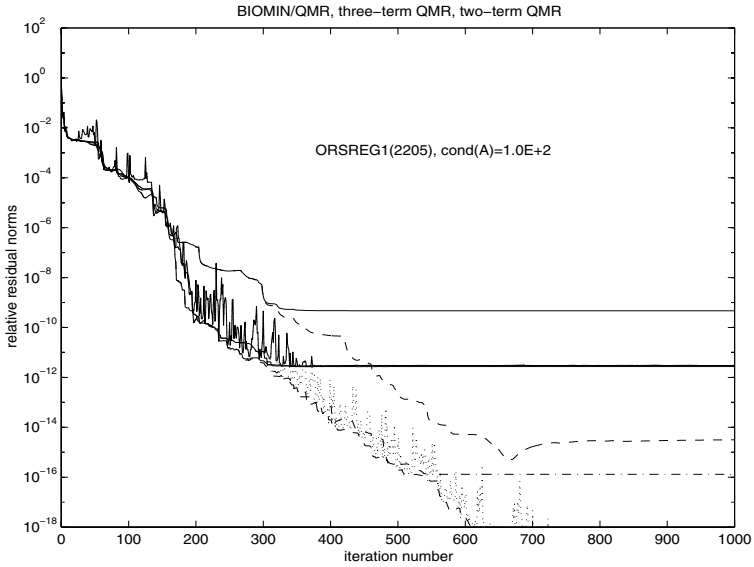

Figure 6.5: QMR smoothing of standard BiCG (BiOMin) compared with the QMR method, when applied to a linear system with the matrix ORSREG1. Relative 2-norms of updated (dots) and true (initially oscillating solid line) residuals of BiOMin. Comparison with the QMR method and QMR smoothing: updated (lower dashed line) and true (lower solid line) residuals of BiOMin|QMR, true residuals of standard three-term QMR (upper solid line) and two-term QMR (lower solid line overlapping the one of BiOMin|QMR), and the quantities $\widetilde{\rho}_n$ of three-term QMR (upper dashed line) and two-term QMR (dash-dotted line ending in lowest plateau).

tive 2-norms of the updated (dots) and the true (initially oscillating solid line) residuals of the standard BiCG algorithm (BiOMin), as well as those of the updated (lower dashed line) and true (lower solid line ending on same plateau as BiOMin) residuals of the smoothed BiOMin|QMR. For comparison we further display the true residual norms of standard three-term QMR (upper solid line) and coupled two-term QMR (lower solid line overlapping with that of BiOMin|QMR), both without look-ahead, and the corresponding quantities $\widetilde{\rho}_n$ (upper dashed and dash-dotted lines, respectively), which are a byproduct of the QR decomposition and are, in exact arithmetic, equal to the quasi-residual norms. They are often used as indicators for the size of the actual residuals. Clearly, three-term QMR is less accurate than the other methods, which is no surprise in view of the mentioned error analysis of Sleijpen et al. [27] for MinRes, from which we must expect *a fortiori* a similar (or even worse) behavior for three-term QMR. In contrast, two-term QMR is as accurate as BiOMin or BiOMin|QMR, but not better.

Next we perform also for this nonsymmetric system a comparison of three-term recurrences with coupled two-term recurrences, both of the primary and

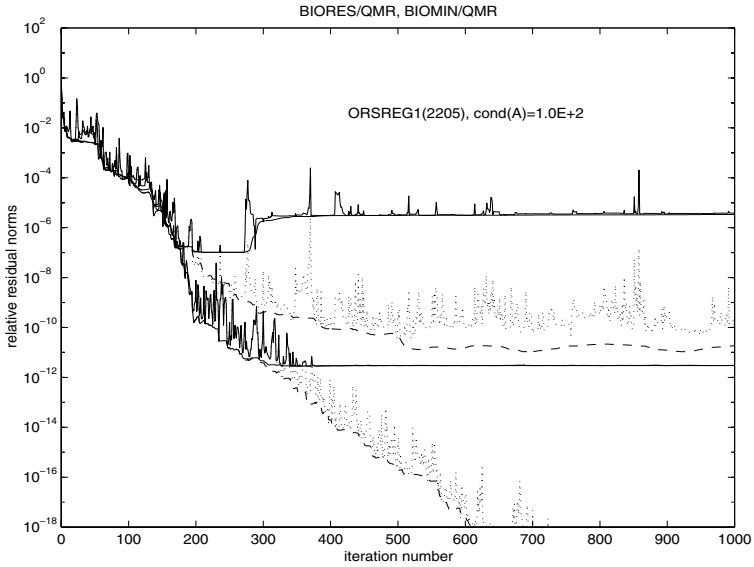

Figure 6.6: QMR smoothing of coupled two-term BiCG (BiOMin) vs. coupled three-term BiCG (BiORes), applied to the nonsymmetric system with the matrix ORSREG1. Relative 2-norms of updated (upper set of dots) and true (initially oscillating top dash-dotted line) residuals of BiORes, and of updated (lower set of dots) and true (initially oscillating solid line) residuals of BiOMin. QMR-smoothed updated (upper dashed line) and true (top solid line) residuals of BiORes|QMR, and of updated (bottom dashed line) and true (lower solid line, overlapping with the one of BiOMin) residuals of BiOMin|QMR.

the QMR-smoothed algorithms. From Gutknecht and Strakoš [18] we have to expect that the primary three-term algorithm BiORes (top dash-dotted line) is much less accurate than standard BiCG (BiOMin) (lower initially oscillating solid line), and this is fully confirmed in Figure 6.6. Moreover, from our results we have to expect that this discrepancy persists if we apply smoothing, and this is indeed confirmed too. Note that here not even the updated BiORes (upper set of dots) and BiORes|QMR (upper dashed line) residual norms decay to zero, which is a counterexample to the heuristic argument that we have given, but, at least, they are several orders of magnitude smaller than the norms of the true residuals. Moreover, the true BiORes residual history has some strong spikes even after stagnation, and due to such a spike, more than an order of magnitude in accuracy is lost after the residuals first stagnates on a 10^{-7} level.

Finally, in Figure 6.7 we compare QMR smoothing with MR smoothing, both for the three-term and the coupled two-term version of BiCG. Compared to the difference between the primary algorithms, those between MR smoothing and QMR smoothing are rather small, both for the updated (dot-dashed and dashed lines, respectively) and for the true residuals (solid lines).

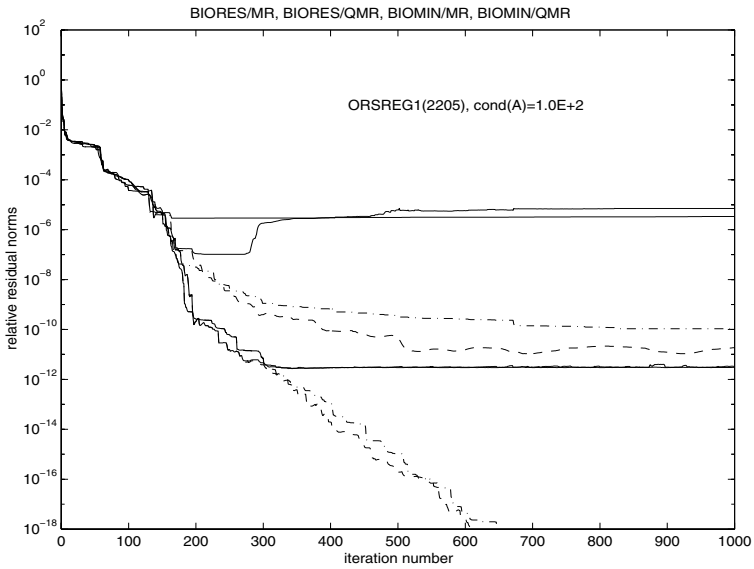

Figure 6.7: QMR smoothing vs. MR smoothing, both for coupled two-term and three-term BiCG, applied to the nonsymmetric system with the matrix ORSREG1. Relative 2-norms of updated BiOMin|QMR residuals (lower dashed line) vs. updated BiOMin|MR residuals (lower dot-dashed line), as well as norms of true BiOMin|QMR vs. BiOMin|MR residuals (nearly overlapping lower solid lines). Comparison with norms of updated BiORes|QMR residuals (upper dashed line) and updated BiORes|MR residuals (upper dot-dashed line), as well as norms of true BiORes|QMR residuals (middle solid line) and BiORes|MR residuals (top solid line).

7 Conclusions.

We have first shown by a contrived example that the ultimate accuracy, that is the level of stagnation of the true residual, can be much worse for MR smoothing than for QMR smoothing. However, such a difference seems to appear rarely in practice. For QMR smoothing and for all three implementations of both CG (OMin, ORes, and ODir) and BiCG (BiOMin, BiORes, and BiODir) considered, the smoothing relations are locally satisfied up to quantities proportional to machine precision. The convergence of the true smoothed residuals is not faster than that of the original residuals, but at least it also does not deteriorate due to rounding errors. In other words, smoothing — despite earlier hopes — does not improve the attainable accuracy. But by examples we have shown that by smoothing residuals of algorithms based on two-term recursions one can construct the minimum residual iterates for an spd system or the quasi-minimal residual iterates for a nonsymmetric system with higher accuracy than with MinRes and three-term QMR, respectively, and with comparable accuracy as with coupled two-term CR or QMR. (Using a result of Sleijpen, van der

Vorst, and Modersitzki [27], this observation can also be justified theoretically.) Generally, algorithms based on coupled two-term recurrences are again seen to be sometimes much more accurate than the corresponding ones using three-term recurrences for residuals and iterates.

Acknowledgment.

The authors like to thank Stefan Röllin for providing equivalent MATLAB code for most of the algorithms.

REFERENCES

1. 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.
2. P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 58–78.
3. L. Chunguang and X. Chengxian, *A generalization of residual smoothing technique for iterative methods*, Manuscript, 1998.
4. J. Cullum, *Peaks, plateaus, numerical instabilities in a Galerkin/minimal residual pair of methods for solving* $Ax = b$, Appl. Numer. Math., 19 (1995), pp. 255–278.
5. J. Cullum and A. Greenbaum, *Relations between Galerkin and norm-minimizing iterative methods for solving linear systems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 223–247.
6. R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Numerical Analysis, Proceedings of the Dundee Conference on Numerical Analysis 1975, G. A. Watson, ed., Vol. 506 of Lecture Notes in Mathematics, Springer, Berlin, 1976, pp. 73–89.
7. R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, *An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 137–158.
8. 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.
9. R. W. Freund and N. M. Nachtigal, *Implementation details of the coupled QMR algorithm*, in Numerical Linear Algebra, L. Reichel, A. Ruttan, and R. S. Varga, eds., W. de Gruyter, 1993, pp. 123–140.
10. R. W. Freund and N. M. Nachtigal, *An implementation of the QMR method based on coupled two-term recurrences*, in Linear Algebra for Large Scale and Real-Time Applications, M. S. Moonen, G. H. Golub, and B. L. R. de Moor, eds., Kluwer Academic Publishers, Dordrecht, 1993, pp. 381–384.
11. 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.
12. A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.
13. M. H. Gutknecht, *The unsymmetric Lanczos algorithms and their relations to Padé approximation, continued fractions, and the qd algorithm*, in Preliminary Proceedings of the Copper Mountain Conference on Iterative Methods, April 1990. `http://www.sam.math.ethz.ch/~mhg/pub/CopperMtn90.ps.gz` and `CopperMtn90-7.ps.gz`.

14. M. H. Gutknecht, *Changing the norm in conjugate gradient type algorithms*, SIAM J. Numer. Anal., 30 (1993), pp. 40–56.
15. M. H. Gutknecht, *Lanczos-type solvers for nonsymmetric linear systems of equations*, Acta Numerica, 6 (1997), pp. 271–397.
16. M. H. Gutknecht and M. Rozložník, *Residual smoothing techniques: do they improve the limiting accuracy of iterative solvers?*, Research Rep. 99-22, Seminar for Applied Mathematics, ETH Zurich, October 1999.
17. M. H. Gutknecht and M. Rozložník, *By how much can residual minimization accelerate the convergence of orthogonal residual methods?*, in Research Rep. 2000-09, Seminar for Applied Mathematics, ETH Zurich, July 2000.
18. M. H. Gutknecht and Z. Strakoš, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., to appear.
19. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bureau Standards, 49 (1952), pp. 409–435.
20. K. C. Jea and D. M. Young, *On the simplification of generalized conjugate-gradient methods for nonsymmetrizable linear systems*, Linear Algebra Appl., 52 (1983), pp. 399–417.
21. C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. Nat. Bureau Standards, 49 (1952), pp. 33–53.
22. A. Neumaier, *Iterative regularization for large-scale ill-conditioned linear systems*, Talk at Oberwolfach, April 1994.
23. C. C. Paige and M. A. Saunders, *Solution of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
24. W. Schönauer, *Scientific Computing on Vector Computers*, Elsevier, Amsterdam, 1987.
25. G. L. G. Sleijpen and H. A. van der Vorst, *Reliable updated residuals in hybrid Bi-CG methods*, Computing, 56 (1996), pp. 141–163.
26. G. L. G. Sleijpen, H. A. van der Vorst, and D. R. Fokkema, *BiCGstab(l) and other hybrid Bi-CG methods*, Numerical Algorithms, 7 (1994), pp. 75–109.
27. G. L. G. Sleijpen, H. A. van der Vorst, and J. Modersitzki, *Differences in the effects of rounding errors in Krylov solvers for symmetric indefinite linear systems*, to appear in SIAM J. Matrix Anal. Appl..
28. E. Stiefel, *Relaxationsmethoden bester Strategie zur Lösung linearer Gleichungssysteme*, Comm. Math. Helv., 29 (1955), pp. 157–179.
29. H. A. van der Vorst and Q. Ye, *Residual replacement strategies for Krylov subspace iterative methods for the convergence of true residuals*, preprint, 1999.
30. H. F. Walker, *Residual smoothing and peak/plateau behavior in Krylov subspace methods*, Appl. Numer. Math., 19 (1995), pp. 279–286.
31. R. Weiss, *Convergence Behavior of Generalized Conjugate Gradient Methods*, PhD thesis, University of Karlsruhe, 1990.
32. R. Weiss, *Properties of generalized conjugate gradient methods*, J. Numer. Linear Algebra Appl., 1 (1994), pp. 45–63.
33. L. Zhou and H. F. Walker, *Residual smoothing techniques for iterative methods*, SIAM J. Sci. Comput., 15 (1994), pp. 297–312.