The Chebyshev iteration revisited

Martin H. Gutknecht

Seminar for Applied Mathematics, ETH-Zentrum HG, CH-8092 Zürich, Switzerland [1]

Stefan Röllin

Integrated Systems Laboratory, ETH-Zentrum ETZ, CH-8092 Zürich, Switzerland [2]

Abstract

Compared to Krylov space methods based on orthogonal or oblique projection, the Chebyshev iteration does not require inner products and is therefore particularly suited for massively parallel computers with high communication cost. Here, six different algorithms that implement this method are presented and compared with respect to roundoff effects, in particular, the ultimately achievable accuracy. Two of these algorithms replace the three-term recurrences by more accurate coupled two-term recurrences and seem to be new. It is also shown that, for real data, the classical three-term Chebyshev iteration is never seriously affected by roundoff, in contrast to the corresponding version of the conjugate gradient method. Even for complex data, strong roundoff effects are seen to be limited to very special situations where convergence is anyway slow.

The Chebyshev iteration is applicable to symmetric definite linear systems and to nonsymmetric matrices whose eigenvalues are known to be confined to an elliptic domain that does not include the origin. Also considered is a corresponding stationary 2-step method, which has the same asymptotic convergence behavior and is additionally suitable for mildly nonlinear problems.

Key words: sparse linear systems, Chebyshev iteration, second order Richarson iteration, coupled two-term recurrences, roundoff error analysis
1991 MSC: 65F10, 65G05, 65H10

[1] Email: mhg@sam.math.ethz.ch, URL: http://www.sam.math.ethz.ch/~mhg
[2] Email: roellin@iis.ee.ethz.ch

1 Introduction

The Chebyshev iteration [2–4] has been one of the favorite Krylov space methods for solving a large sparse linear system of equations in a parallel environment, since, unlike methods based on orthogonalization (such as the conjugate gradient (CG) and biconjugate gradient (BiCG) methods and GMRes—to name a few), it does not require to compute communication-intensive inner products for the determination of the recurrence coefficients. Only the monitoring of the convergence, that is, the determination of the norm of the residuals requires inner products, and even this norm needs to be evaluated only occasionally because its time-dependence, that is, the convergence rate, can be forecast reliably.

The Chebyshev iteration, which in the older literature has often been referred to as Chebyshev semiiterative method, requires some preliminary knowledge about the spectrum $\sigma(\mathbf{A})$ of the coefficient matrix $\mathbf{A}$: an elliptic domain $\mathcal{E} \supset \sigma(\mathbf{A})$ with $0 \notin \mathcal{E}$ is normally assumed to be known in advance. Denote the center of the ellipse by α, its foci by $\alpha \pm c$, and the lengths of the large and the small semi-axes by a and b. When $b = 0$, the elliptic domain turns into a straight line segment (or, interval) $\mathcal{I} :\equiv [\alpha - c, \alpha + c]$. At this point, both α and c may be complex. Manteuffel [1] devised a technique to determine a suitable ellipse from a given nonsymmetric matrix.

Mathematically the method can be defined by translating the Chebyshev polynomials T_n from the interval $[-1, 1]$ to the interval $\mathcal{I}$ and scaling them so that their value at 0 is 1. On $\mathbb{R}$ the Chebyshev polynomials are defined by [3]

$$T_n(\xi) :\equiv \begin{cases} \cos(n \arccos(\xi)) & \text{if } |\xi| \le 1\,, \\ \cosh(n\, \operatorname{arcosh}(\xi)) & \text{if } \xi \ge 1\,, \\ (-1)^n \cosh(n\, \operatorname{arcosh}(-\xi)) & \text{if } \xi \le -1\,. \end{cases}$$

T_n is even or odd if n is even or odd, respectively. All three formulas are valid when we extend the definition to the complex plane, which we will indicate by using the variable ζ. For example, we may define

$$T_n(\zeta) :\equiv \frac{1}{2}\left[\left(\zeta + \sqrt{\zeta^2 - 1}\right)^n + \left(\zeta - \sqrt{\zeta^2 - 1}\right)^n\right] .$$

The translated and scaled residual polynomials p_n that characterize the Cheby-

[3] Definitions are marked by the symbol $:\equiv$, while $:=$ is used for algorithmic assignments; often either one of the symbols could be used.

shev iteration are

$$p_n(\zeta) :\equiv \frac{T_n\left(\frac{\zeta-\alpha}{c}\right)}{T_n\left(-\frac{\alpha}{c}\right)} . \tag{1}$$

If we let $\mathbf{x}_0$ be an initially chosen approximation of the solution of the linear system $\mathbf{Ax} = \mathbf{b}$ that has to be solved, and if $\mathbf{r}_0 :\equiv \mathbf{b} - \mathbf{Ax}_0$ denotes the corresponding residual, then, by definition, the nth approximation and residual satisfy

$$\mathbf{b} - \mathbf{Ax}_n = \mathbf{r}_n = p_n(\mathbf{A})\mathbf{r}_0 .$$

The classical case for applying the method is when $\mathbf{A}$ is symmetric positive definite (spd)—as assumed in CG—and, therefore, the interval $\mathcal{I}$ lies on the positive real axis and contains the spectrum of $\mathbf{A}$. In this case Chebyshev iteration is known to be optimal in the sense that it yields, for every n, the smallest nth maximum residual if the maximum is taken over all normal matrices with spectrum on $\mathcal{I}$; see [2–4].

Due to a wrong claim in [5] it has often been assumed that this optimality also holds for the class of matrices with spectrum inside or on an ellipse whose foci lie on the positive real axis, but Fischer and Freund [6,7] have shown that this is not true in general; the exceptional cases are rather ill-conditioned, however. In any case, for any elliptic compact set not containing 0 the correspondingly chosen Chebyshev iteration is asymptotically optimal, as its recurrence coefficients approach those of a second order Richardson iteration, which is a stationary 2-step method based on conformal mapping [8–11] and can be viewed as the limit of the Chebyshev iteration; see our discussion in Section 6.

Of course, in practice we need an algorithm that generates the approximations $\mathbf{x}_n$ recursively. The usual approach is to derive a three-term recurrence from the standard recursion for the Chebyshev polynomials. However, as has recently been shown by Gutknecht and Strakoš [12], Krylov space methods based on three-term recursions for iterates and residuals may suffer from a large gap between recursively computed residuals $\mathbf{r}_n$ and true residuals $\mathbf{b} - \mathbf{Ax}_n$, and, therefore, may stagnate early with relatively large true residuals. In other words, the ultimately achievable accuracy may be quite low. In particular, this effect may even occur when CG is applied to an spd problem.

We will show here that the Chebyshev iteration, even in this implementation, is not seriously affected by roundoff. Moreover, we will discuss five other implementations that produce even more accurate solutions, that is, stagnate ultimately with smaller true residuals. We also point out that the aforementioned stationary second order Richardson iteration can as well be realized by six analogous different algorithms.

We note that similar analytical techniques have been applied by Golub and Overton [13] for the analysis of the behavior of the Chebyshev iteration when a preconditioner is applied inexactly.

2 Chebyshev iteration with three-term recursion

Recursions for the residuals $\mathbf{r}_n$ and the iterates $\mathbf{x}_n$ of the Chebyshev iteration are easily found from the standard three-term recursions for the classical Chebyshev polynomials T_n,

$$T_{n+1}(\zeta) := 2\zeta T_n(\zeta) - T_{n-1}(\zeta) \quad (n > 1). \tag{2}$$

The following first realization of the method results.

Algorithm 1 (Three-term Chebyshev iteration) *For solving* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *choose* $\mathbf{x}_0$ *and let* $\mathbf{r}_0 := \mathbf{b} - \mathbf{A}\mathbf{x}_0$. *Also set* $\mathbf{r}_{-1} := \mathbf{x}_{-1} := \mathbf{o}$. *Choose the parameters* α *and* c *so that the spectrum of* $\mathbf{A}$ *lies on the straight line segment* $\mathcal{I} :\equiv [\alpha - c, \alpha + c]$ *or on an elliptical domain* $\mathcal{E}$ *with foci* $\alpha \pm c$ *that does not contain* 0. *Then let* $\eta :\equiv -\alpha/c$,

$$\beta_{-1} :\equiv 0\,, \quad \beta_0 :\equiv \frac{c}{2}\,\frac{1}{\eta} = -\frac{c^2}{2\alpha}\,, \quad \gamma_0 :\equiv -\alpha\,, \tag{3}$$

and compute, for $n = 0, 1, \dots$ *until convergence,*

$$\beta_{n-1} :\equiv \frac{c}{2}\,\frac{T_{n-1}(\eta)}{T_n(\eta)} = \left(\frac{c}{2}\right)^2 \frac{1}{\gamma_{n-1}} \quad \textit{if } n \geq 2, \tag{4}$$

$$\gamma_n :\equiv \frac{c}{2}\,\frac{T_{n+1}(\eta)}{T_n(\eta)} = -(\alpha + \beta_{n-1}) \quad \textit{if } n \geq 1, \tag{5}$$

$$\mathbf{x}_{n+1} := -\big(\mathbf{r}_n + \mathbf{x}_n\alpha + \mathbf{x}_{n-1}\beta_{n-1}\big)/\gamma_n\,, \tag{6}$$

$$\mathbf{r}_{n+1} := \big(\mathbf{A}\mathbf{r}_n - \mathbf{r}_n\alpha - \mathbf{r}_{n-1}\beta_{n-1}\big)/\gamma_n\,. \tag{7}$$

We cannot expect that a solver produces ultimately a much smaller residual than what we get when we insert (the machine approximation of) the exact solution $\mathbf{x}_\star :\equiv \mathbf{A}^{-1}\mathbf{b}$ into the definition of the residual: $\|\mathrm{fl}(\mathbf{b}-\mathbf{A}\mathbf{x}_\star)\|$. However, due to the accumulation of rounding errors the achievable accuracy might be perhaps much lower. Actually, the ultimate accuracy of Algorithm 1 (and many others) is determined by the size of the gap $\mathbf{f}_n$ between the updated (or, recursively computed) residual $\mathbf{r}_n$ and the true (or, explicitly computed) residual $\mathbf{b} - \mathbf{A}\mathbf{x}_n$:

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

Here $\mathbf{x}_n$ and $\mathbf{r}_n$ denote the vectors computed in floating-point arithmetic from (6) and (7). In fact, if $\mathbf{A}$ satisfies the spectral assumption, then, normally, $\mathbf{r}_n \to \mathbf{o}$ as $n \to \infty$ even in floating-point arithmetic. Thus,

$$\|\mathrm{fl}(\mathbf{b} - \mathbf{A}\mathbf{x}_n)\| \approx \|\mathbf{f}_n\| \quad \text{for large } n\,.$$

A general result on this gap for methods updating residuals by three-term recurrences was given in [12].

Theorem 1 ([12]) *Assume iterates and residuals are updated according to*

$$\mathbf{x}_{n+1} := -\Big(\mathbf{r}_n + \mathbf{x}_n\alpha_n + \mathbf{x}_{n-1}\beta_{n-1}\Big)/\gamma_n\,, \tag{8}$$

$$\mathbf{r}_{n+1} := \Big(\mathbf{A}\mathbf{r}_n - \mathbf{r}_n\alpha_n - \mathbf{r}_{n-1}\beta_{n-1}\Big)/\gamma_n\,, \tag{9}$$

where $\gamma_n := -\alpha_n - \beta_{n-1}$*. Then the gap* $\mathbf{f}_n :\equiv \mathbf{b} - \mathbf{A}\mathbf{x}_n - \mathbf{r}_n$ *satisfies, up to* $\mathcal{O}(\epsilon^2)$ *(where* ϵ *denotes the machine epsilon),*

$$\begin{aligned}
\mathbf{f}_{n+1} = \mathbf{f}_0 \; & - \mathbf{l}_0 \\
& - \mathbf{l}_0\frac{\beta_0}{\gamma_1} - \mathbf{l}_1 \\
& - \mathbf{l}_0\frac{\beta_0\beta_1}{\gamma_1\gamma_2} - \mathbf{l}_1\frac{\beta_1}{\gamma_2} - \mathbf{l}_2 \\
& \;\vdots \\
& - \mathbf{l}_0\frac{\beta_0\beta_1\cdots\beta_{n-1}}{\gamma_1\gamma_2\cdots\gamma_n} - \cdots - \mathbf{l}_{n-1}\frac{\beta_{n-1}}{\gamma_n} - \mathbf{l}_n\,,
\end{aligned} \tag{10}$$

where

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

is the local error whose components come from

$$\begin{aligned}
\mathbf{r}_{n+1} &= (\mathbf{A}\mathbf{r}_n - \mathbf{r}_n\alpha_n - \mathbf{r}_{n-1}\beta_{n-1} + \mathbf{g}_n)/\gamma_n\,, \\
\mathbf{x}_{n+1} &= -(\mathbf{r}_n + \mathbf{x}_n\alpha_n + \mathbf{x}_{n-1}\beta_{n-1} + \mathbf{h}_n)/\gamma_n\,, \\
\gamma_n &= -(\alpha_n + \beta_{n-1} + \varepsilon_n).
\end{aligned} \tag{11}$$

In (10) and (11) the quantities $\mathbf{x}_k$, $\mathbf{r}_k$, α_k, β_{k-1}, and γ_k are those computed in floating-point arithmetic. If we assume that each row of $\mathbf{A}$ contains at most m nonzero elements and that matrix-vector products with $\mathbf{A}$ are computed in the standard way, then for the local error holds componentwise

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

But more important than the size of the local errors is the size of the potentially large factors β_{k-1}/γ_k and of their products in (10). In Algorithm 1, the factors and their products are (in exact arithmetic)

$$
\frac{\beta_0}{\gamma_1} = \frac{1}{T_2(\eta)}\,, \tag{12a}
$$

$$
\frac{\beta_{n-1}}{\gamma_n} = \frac{T_{n-1}(\eta)}{T_{n+1}(\eta)}\,, \tag{12b}
$$

$$
\frac{\beta_k\beta_{k+1}\cdots\beta_{n-1}}{\gamma_{k+1}\gamma_{k+2}\cdots\gamma_n} = \frac{T_k(\eta)T_{k+1}(\eta)}{T_n(\eta)T_{n+1}(\eta)} \quad (0 \le k < n-1)\,. \tag{12c}
$$

Strictly speaking, we should consider here the values of β_{k-1} and γ_k that are obtained in floating-point arithmetic. Then the three relations (12a)–(12c) are only correct up to a roundoff error of order $\mathcal{O}(\epsilon)$. However, because we are free to compute the recurrence coefficients at little extra cost in multiple-precision arithmetic, and since we are only concerned about quotients that are very large, it seems well justified to neglect these errors here. Otherwise we would have to analyze the roundoff errors in the recursions (4) and (5), or in any other formulas used to calculate β_{n-1} and γ_n.

If $0 < c < \alpha$ (as in the case when $\mathbf{A}$ is spd), so that $\eta < -1$ and $|T_k(\eta)| = \cosh(k\,\mathrm{arcosh}(|\eta|))$, we have, since cosh is monotone increasing on the positive real axis,

$$
|T_k(\eta)| < |T_n(\eta)| \quad \text{if } \; k < n\,, \tag{13}
$$

and therefore all the factors in (10) are less than 1 if the recurrence coefficients are computed accurately enough. Since $\mathbf{l}_k$ appears in $n-k+1$ terms of (10), it may still get amplified by a factor smaller than $n-k+1$, but this is not too serious, in particular since typically most of the factors are rather small.

Of course, (13) does not hold in general unless $\eta < -1$ or $\eta > 1$: for example, given $m \in \mathbb{N}$ we have, for η purely imaginary and of sufficiently small absolute value, $|T_{2k}(\eta)| > |T_{2n+1}(\eta)|$ for all $k, n \in \mathbb{N}$ with $k, n \le m$, since $|T_{2k}(0)| \ne 0$ but $|T_{2n+1}(0)| = 0$. But in (12b) the index difference between the numerator and denominator polynomials is 2 and hence this argument is not applicable.

We show next that also in the other case relevant for real-valued problems, namely when $\alpha \in \mathbb{R}$ but c is purely imaginary (so that the ellipse is still centered on and symmetric about the real axis), the quotients (12a)–(12c) are all of absolute value smaller than 1.

For any $\eta \in \mathbb{C}\backslash[-1, 1]$, we let ϑ be the larger solution of the quadratic equation

$$\frac{1}{2}\left(\vartheta + \frac{1}{\vartheta}\right) = \eta\,. \tag{14}$$

Note that the solutions come in pairs ϑ, ϑ^{-1} and that $|\vartheta| = 1$ implies that $\eta = \frac{1}{2}(\vartheta+\vartheta^{-1}) = \frac{1}{2}(\vartheta+\overline{\vartheta}) = \mathrm{Re}\ \vartheta \in [-1, 1]$, which is excluded by assumption. Therefore, we may assume that $|\vartheta| > 1$ here. The mapping $\vartheta \mapsto \eta$ of (14) is the well-known Joukowski transformation, which allows us to express the Chebyshev polynomials simply as

$$T_n(\eta) = \frac{1}{2}\left(\vartheta^n + \frac{1}{\vartheta^n}\right). \tag{15}$$

In fact, if we let

$$\phi :\equiv \mathrm{arcosh}(\eta) \quad \text{with} \quad \mathrm{Re}\ \phi \geq 0\,,$$

so that $e^{\phi} + e^{-\phi} = 2\eta = \vartheta + \vartheta^{-1}$, then, clearly, $e^{\phi} = \vartheta$, and therefore, if $\eta \geq 1$, $T_n(\eta) = \cosh(n\,\mathrm{arcosh}(\eta)) = \frac{1}{2}(\vartheta^n + \vartheta^{-n})$, and this relation can be seen to be valid for any $\eta \in \mathbb{C}$. Consequently, the single factors from (12b) can be written as

$$\frac{\beta_{n-1}}{\gamma_n} = \frac{T_{n-1}(\eta)}{T_{n+1}(\eta)} = \frac{\vartheta^{n-1} + \vartheta^{-(n-1)}}{\vartheta^{n+1} + \vartheta^{-(n+1)}}\,. \tag{16}$$

Obviously, these factors are rational functions of both η and ϑ.

It is well-known that T_{n+1} has $n+1$ simple zeros in $(-1, 1)$. So, $T_{n-1}(\zeta)/T_{n+1}(\zeta)$ (after cancellation of the pole and zero at $\zeta = 0$ if n is even) has at most $n+1$ poles, and they lie all on $(-1, 1)$. If considered as a function of ϑ, the quotient has at most $2(n + 1)$ poles, and they all lie on the unit circle. Clearly, if we choose η close enough to a pole, but not on $[-1, 1]$, the quotient can be made as large as we want. Consequently, as claimed, the factors are in general not all of absolute value less than 1. So, amplification of a local error is possible.

If $0 < c < \alpha$, we have seen already that (13) holds, and, by symmetry, the same is true for $0 > c > \alpha$. If α is still real, say $\alpha > 0$, but $c \in i\,\mathbb{R}^+$, then $\eta :\equiv -\alpha/c \in i\,\mathbb{R}^+$, and since the Joukowski transformation maps the part above i of the imaginary axis on the positive imaginary axis, we have $\vartheta = i\chi$ with $\chi > 1$. Then, from (16) and by setting

$$\tilde{\eta} :\equiv \frac{1}{2}\left(\chi + \frac{1}{\chi}\right),$$

so that $\widetilde{\eta} > 1$ (since the Joukowski transformation maps $[1, \infty)$ onto itself), we obtain

$$\frac{\beta_{n-1}}{\gamma_n} = \frac{T_{n-1}(\eta)}{T_{n+1}(\eta)} = \begin{cases} -\dfrac{\chi^{n-1} + \chi^{-(n-1)}}{\chi^{n+1} + \chi^{-(n+1)}} = -\dfrac{T_{n-1}(\widetilde{\eta})}{T_{n+1}(\widetilde{\eta})} & \text{if } n \text{ odd,} \\ -\dfrac{\chi^{n-1} - \chi^{-(n-1)}}{\chi^{n+1} - \chi^{-(n+1)}} = -\dfrac{U_{n-1}(\widetilde{\eta})}{U_{n+1}(\widetilde{\eta})} & \text{if } n \text{ even}. \end{cases} \tag{17}$$

Here, U_n is the nth Chebyshev polynomial of the second kind. For $\widetilde{\eta} > 1$, U_n can be expressed as

$$U_n(\widetilde{\eta}) = \sinh(n \operatorname{arcosh}(\widetilde{\eta})) .$$

Noting that sinh is monotone increasing, we can conclude that $U_{n+1}(\widetilde{\eta}) > U_{n-1}(\widetilde{\eta}) > 0$ if $\widetilde{\eta} > 1$. As we have seen before, also $T_{n+1}(\widetilde{\eta}) > T_{n-1}(\widetilde{\eta}) > 0$ if $\widetilde{\eta} > 1$. Therefore, also in this situation, the factors β_{n-1}/γ_n have an absolute value smaller than 1. Summarizing, we have proved the following result.

Theorem 2 *For an interval $[\alpha - c, \alpha + c] \subset \mathbb{R}$ or an ellipse with foci $\alpha \pm c$ symmetric about the real axis and not containing the origin, the factors* (12a)–(12c)*, which appear in* (10)*, are of absolute value less than 1 if the recurrence coefficients β_{k-1} and γ_k have been computed with sufficient accuracy.*

In Section 8 we will come back to the question of the size of the factors (12a)–(12c) in the case where the assumptions of this theorem do not hold, that is when the linear system to be solved is complex and does not have a spectrum symmetric about the real axis.

Finally we note that a simple way to avoid the residual gap in the Chebyshev iteration is to replace the recursively computed residuals by explicitly computed residuals:

Algorithm 2 (Three-term recursion, explicitly computed residuals) *Same as Algorithm 1 except that the recursion* (7) *for computing $\mathbf{r}_{n+1}$ is replaced by*

$$\mathbf{r}_{n+1} := \mathbf{b} - \mathbf{A}\mathbf{x}_{n+1} . \tag{18}$$

This remedy could be applied in many Krylov solvers in order to increase the ultimate accuracy. However, explicitly computed residuals are known to slow down the convergence of projection methods like CG and BiCG due to stronger roundoff effects in the Krylov space generation process [14], as they destroy the local (bi)orthogonality of the bases created. But here, unlike in these methods, the recurrence coefficients α, β_{n-1}, and γ_n do not depend on $\mathbf{r}_n$, and therefore the error of $\mathbf{r}_n$ will only have little influence on $\mathbf{x}_m$ $(m > n)$ and the convergence of the method.

3 Rutishauser's Chebyshev iteration by updating corrections

The recursions (6) and (7) are of the form (8)–(9) with the consistency condition $\alpha_n + \beta_{n-1} + \gamma_n = 0$, which implies that $\mathbf{r}_n = \mathbf{b} - \mathbf{A}\mathbf{x}_n$ for all n if it holds for $n = 0$. Subtracting $\mathbf{x}_n$ and $\mathbf{r}_n$, respectively, on both sides of (8) and (9), using the consistency condition, and setting

$$\Delta\mathbf{x}_n :\equiv \mathbf{x}_{n+1} - \mathbf{x}_n \,, \qquad \Delta\mathbf{r}_n :\equiv \mathbf{r}_{n+1} - \mathbf{r}_n \,,$$

yields

$$\Delta\mathbf{x}_n := \Big(-\mathbf{r}_n + \Delta\mathbf{x}_{n-1}\beta_{n-1}\Big)/\gamma_n \,, \tag{19}$$

$$\Delta\mathbf{r}_n := \Big(\mathbf{A}\mathbf{r}_n + \Delta\mathbf{r}_{n-1}\beta_{n-1}\Big)/\gamma_n \,. \tag{20}$$

This leads to the following reformulation of Algorithm 1.

Algorithm 3 (Chebyshev iteration by updating $\Delta\mathbf{x}_n$ and $\Delta\mathbf{r}_n$) *Same as Algorithm 1 except that the recursions* (6) *and* (7) *for computing* $\mathbf{x}_{n+1}$ *and* $\mathbf{r}_{n+1}$ *are replaced by* (19)–(20) *and*

$$\mathbf{x}_{n+1} := \mathbf{x}_n + \Delta\mathbf{x}_n \,, \tag{21}$$

$$\mathbf{r}_{n+1} := \mathbf{r}_n + \Delta\mathbf{r}_n \,. \tag{22}$$

This is how Rutishauser [4] formulated the Chebyshev iteration and other Krylov space solvers (which he called "gradient methods"). It is easy to also modify this scheme so that the residuals are computed explicitly:

Algorithm 4 (Updating $\Delta\mathbf{x}_n$ and explicitly computing $\mathbf{r}_n$) *Same as Algorithm 1 except that the recursions* (6) *and* (7) *for computing* $\mathbf{x}_{n+1}$ *and* $\mathbf{r}_{n+1}$ *are replaced by* (19), (21), *and* (18), *that is,*

$$\begin{aligned} \Delta\mathbf{x}_n &:= \Big(-\mathbf{r}_n + \Delta\mathbf{x}_{n-1}\beta_{n-1}\Big)/\gamma_n \,, \\ \mathbf{x}_{n+1} &:= \mathbf{x}_n + \Delta\mathbf{x}_n \,, \\ \mathbf{r}_{n+1} &:= \mathbf{b} - \mathbf{A}\mathbf{x}_{n+1} \,. \end{aligned}$$

4 Algorithms based on coupled two-term recurrences

For Krylov space solvers based on two-term updates for $\mathbf{x}_n$ and $\mathbf{r}_n$ (involving additionally *direction vectors* $\mathbf{v}_n$) [15,16],

$$\begin{aligned} \mathbf{v}_n &:= \mathbf{r}_n - \mathbf{v}_{n-1}\psi_{n-1,n} - \mathbf{v}_{n-2}\psi_{n-2,n} - \cdots - \mathbf{v}_0\psi_{0,n} \,, && (23)\\ \mathbf{x}_{n+1} &:= \mathbf{x}_n + \mathbf{v}_n\omega_n \,, && (24)\\ \mathbf{r}_{n+1} &:= \mathbf{r}_n - \mathbf{A}\mathbf{v}_n\omega_n \,, && (25) \end{aligned}$$

the gap between updated and true residuals is known to be often much smaller than for those that update the residuals with three-term recurrences of the form (8)–(9) or even longer ones. It does not matter whether the recursion (23) for $\mathbf{v}_n$ is long or just two-term as in

$$\mathbf{v}_n := \mathbf{r}_n - \mathbf{v}_{n-1}\psi_n \,, \tag{26}$$

because the same possibly inaccurate $\mathbf{v}_n$ is used in (24) and (25). Examples for algorithms of the form (24)–(25) with (26) are the standard Hestenes-Stiefel or OMin version of CG and the standard BiOMin version of BiCG.

The above claim about the higher ultimate accuracy of algorithms with two-term updates (24)–(25) is based on a comparison between Theorem 1 and the following result of Greenbaum [17], which improves on previous similar statements in [19] and [18]. It explains why the gap between updated and true residuals is relatively small: here, the gap is just a sum of local errors; these are not multiplied by any potentially large factors.

Theorem 3 ([17]) *Assume iterates and residuals are updated according* (24)–(25)*. Then the gap* $\mathbf{f}_n :\equiv \mathbf{b} - \mathbf{A}\mathbf{x}_n - \mathbf{r}_n$ *between the true and the updated residual satisfies*

$$\mathbf{f}_n = \mathbf{f}_0 - \mathbf{l}_0 - \cdots - \mathbf{l}_n \,,$$

where

$$\mathbf{l}_n :\equiv \mathbf{A}\mathbf{h}_n + \mathbf{g}_n$$

is the local error whose components come from

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{v}_n\omega_n + \mathbf{h}_n \,, \quad \mathbf{r}_{n+1} = \mathbf{r}_n - \mathbf{A}\mathbf{v}_n\omega_n + \mathbf{g}_n \,.$$

In particular,

$$\frac{\|\mathbf{f}_n\|}{\|\mathbf{A}\|\,\|\mathbf{x}\|} \le (\epsilon + \mathcal{O}(\epsilon^2))\Big[n + 2 \ + \ (1 + \mu + (n+1)(10 + 2\mu))\Theta_n\Big] \,,$$

where ϵ denotes the machine epsilon, $\mu :\equiv m\sqrt{N}$ with m the maximum number of nonzeros in a row of $\mathbf{A}$, N the order of $\mathbf{A}$, and

$$\Theta_n :\equiv \max_{k\leq n} \frac{||\mathbf{x}_k||}{||\mathbf{x}_\star||}.$$

5 Chebyshev iteration based on coupled two-term recurrences

Theorem 3 suggest to search for a coupled-two term recursion as an alternative realization of the Chebyshev method. Recursions (24)–(25) call for the following “Ansatz” in a polynomial formulation:

$$\widehat{p}_n(\zeta) := p_n(\zeta) - \psi_{n-1}\widehat{p}_{n-1}(\zeta), \tag{27}$$

$$p_{n+1}(\zeta) := p_n(\zeta) - \zeta\omega_n\widehat{p}_n(\zeta), \tag{28}$$

with $p_0(\zeta) := \widehat{p}_0(\zeta) := 1$, $\psi_{-1} :\equiv 0$. To determine ω_n and ψ_n we insert (27) into (28), make use of (28) with n replaced by $n-1$, and then compare the result with the polynomial reformulation of (7): if $n \geq 1$,

$$\begin{aligned} p_{n+1}(\zeta) &= p_n(\zeta) - \zeta\omega_n p_n(\zeta) + \zeta\omega_n\psi_{n-1}\widehat{p}_{n-1}(\zeta) \\ &= p_n(\zeta) - \zeta\omega_n p_n(\zeta) + \psi_{n-1}\frac{\omega_n}{\omega_{n-1}}\Big(p_{n-1}(\zeta) - p_n(\zeta)\Big) \\ &= \underbrace{\left(1 - \psi_{n-1}\frac{\omega_n}{\omega_{n-1}}\right)}_{=-\alpha/\gamma_n} p_n(\zeta) \underbrace{-\,\omega_n}_{=1/\gamma_n} \zeta p_n(\zeta) + \underbrace{\psi_{n-1}\frac{\omega_n}{\omega_{n-1}}}_{=-\beta_{n-1}/\gamma_n} p_{n-1}(\zeta)\,. \end{aligned}$$

We obtain

$$\beta_{n-1} = \frac{\psi_{n-1}}{\omega_{n-1}} \quad (n \geq 1), \qquad \gamma_n = -\frac{1}{\omega_n} \quad (n \geq 0), \tag{29}$$

$$\alpha = -\beta_{n-1} - \gamma_n = \frac{1}{\omega_n} - \frac{\psi_{n-1}}{\omega_{n-1}} \quad (n \geq 1), \tag{30}$$

and, conversely,

$$\psi_{n-1} = -\frac{\beta_{n-1}}{\gamma_{n-1}} \quad (n \geq 1), \qquad \omega_n = -\frac{1}{\gamma_n} \quad (n \geq 0). \tag{31}$$

If $n = 0$, we have $\psi_{-1} := 0$ and hence just $p_1(\zeta) = p_0(\zeta) - \zeta\omega_0 p_0(\zeta)$, so

$$\omega_0 = -\frac{1}{\gamma_0} = \frac{1}{\alpha}\,.$$

Like β_{n-1} and γ_n we can express ω_n and ψ_{n-1} in terms of the Chebyshev polynomials and derive recursions for them. First, inserting the left-hand side equations from (4) and (5) into (31) we see that

$$\omega_n = -\frac{2}{c}\frac{T_n(\eta)}{T_{n+1}(\eta)}\,, \qquad \psi_{n-1} = -\left(\frac{T_{n-1}(\eta)}{T_n(\eta)}\right)^2 . \tag{32}$$

Then, inserting the right-hand side equations from (4) and (5) we get

$$\begin{aligned}\omega_n &= -\frac{1}{\gamma_n} = (\alpha+\beta_{n-1})^{-1} = \left(\alpha + \left(\frac{c}{2}\right)^2 \frac{1}{\gamma_{n-1}}\right)^{-1} \\ &= \begin{cases} \left(\alpha - \left(\frac{c}{2}\right)^2 \omega_{n-1}\right)^{-1} & \text{(if } n \geq 2), \\ \left(\alpha - \dfrac{c^2}{2\alpha}\right)^{-1} & \text{(if } n = 1) \end{cases}\end{aligned}$$

and

$$\begin{aligned}\psi_{n-1} &= -\frac{\beta_{n-1}}{\gamma_{n-1}} \\ &= \begin{cases} -\left(\frac{c}{2}\right)^2 \dfrac{1}{\gamma_{n-1}^2} = -\left(\frac{c}{2}\right)^2 \omega_{n-1}^2 & (n \geq 2)\,, \\ -\dfrac{c^2}{2\alpha^2} & (n = 1)\,. \end{cases}\end{aligned}$$

Summarizing we obtain the following *coupled two-term Chebyshev iteration* [20].

Algorithm 5 (Coupled two-term Chebyshev iteration) *For solving* $\mathbf{Ax} = \mathbf{b}$ *choose* $\mathbf{x}_0$ *and let* $\mathbf{r}_0 := \mathbf{b} - \mathbf{Ax}_0$. *Choose the parameters* α *and* c *so that the spectrum of* $\mathbf{A}$ *lies on the straight line segment* $\mathcal{I} :\equiv [\alpha - c, \alpha + c]$ *or on an elliptical domain* $\mathcal{E}$ *with foci* $\alpha \pm c$ *that does not contain* 0. *Then let* $\eta :\equiv -\alpha/c$,

$$\psi_{-1} :\equiv 0, \qquad \psi_0 :\equiv -\frac{1}{2}\left(\frac{c}{\alpha}\right)^2, \tag{33}$$

$$\omega_0 :\equiv \frac{1}{\alpha}, \qquad \omega_1 :\equiv \left(\alpha - \frac{c^2}{2\alpha}\right)^{-1}. \tag{34}$$

and compute, for $n = 0, 1, \ldots$ until convergence,

$$\psi_{n-1} :\equiv -\left(\frac{T_{n-1}(\eta)}{T_n(\eta)}\right)^2 := -\left(\frac{c}{2}\right)^2 \omega_{n-1}^2 \quad (n \geq 2), \tag{35}$$

$$\omega_n :\equiv -\frac{2}{c}\frac{T_n(\eta)}{T_{n+1}(\eta)} := \left(\alpha - \left(\frac{c}{2}\right)^2 \omega_{n-1}\right)^{-1} \quad (n \geq 2), \tag{36}$$

$$\mathbf{v}_n := \mathbf{r}_n - \mathbf{v}_{n-1}\psi_{n-1}\,, \tag{37}$$

$$\mathbf{x}_{n+1} := \mathbf{x}_n + \mathbf{v}_n\omega_n\,, \tag{38}$$

$$\mathbf{r}_{n+1} := \mathbf{r}_n - \mathbf{A}\mathbf{v}_n\omega_n\,. \tag{39}$$

Also in Algorithm 5 we can avoid the residual gap by replacing the recursively computed residuals by explicitly computed residuals.

Algorithm 6 (Two-term recursions and explicitly computed residuals) *Same as Algorithm 5 except that the recursion* (39) *for computing* $\mathbf{r}_{n+1}$ *is replaced by* $\mathbf{r}_{n+1} := \mathbf{b} - \mathbf{A}\mathbf{x}_{n+1}$.

6 The second order Richardson iteration as limiting case

For any $\eta \in \mathbb{C}\backslash[-1, 1]$ we have according to (15) in terms of ϑ defined by (14) and $|\vartheta| > 1$

$$\frac{T_n(\eta)}{T_{n+1}(\eta)} = \frac{\vartheta^n + \vartheta^{-n}}{\vartheta^{n+1} + \vartheta^{-(n+1)}} = \vartheta^{-1}\frac{1 + \vartheta^{-2n}}{1 + \vartheta^{-2(n+1)}} \;\to\; \vartheta^{-1} \tag{40}$$

as $n \to \infty$. We can conclude that for any admissible value of η the coefficients of both the three-term and the two-term Chebyshev iterations converge:

$$\beta_{n-1} \to \frac{c}{2\vartheta} \equiv: \beta \qquad \gamma_n \to \frac{c\,\vartheta}{2} \equiv: \gamma \qquad \text{as } n \to \infty\,, \tag{41}$$

$$\psi_{n-1} \to -\,\vartheta^2 \equiv: \psi \quad \omega_n \to -\,\frac{2}{c\,\vartheta} \equiv: \omega \quad \text{as } n \to \infty\,. \tag{42}$$

(The dependence on the center α of the ellipse or interval is hidden in ϑ.) This gives rise to six additional related algorithms that are analogous to Algorithms 1–6 but use the limit values of the coefficients. For example, for the iterates hold the three-term recurrences

$$\mathbf{x}_{n+1} := -\Big(\mathbf{r}_n + \mathbf{x}_n\alpha + \mathbf{x}_{n-1}\beta\Big)/\gamma\,, \tag{43}$$

and the coupled two-term recurrences

$$\mathbf{v}_n := \mathbf{r}_n - \mathbf{v}_{n-1}\psi\,, \tag{44}$$

$$\mathbf{x}_{n+1} := \mathbf{x}_n + \mathbf{v}_n\omega\,. \tag{45}$$

These additional six algorithms are different implementations of the second-order Euler method that can be associated with the ellipse $\mathcal{E}$. This method belongs to the class of iterative methods based on conformal mappings, introduced by Kublanovskaya in 1959; see [8–11]. It is, at least in the case where the ellipse $\mathcal{E}$ collapses to an interval $\mathcal{I}$, better known as stationary second-order Richardson iteration; see [3]. It can easily be generalized for mildly non-linear systems of equations, and for those it seems more suitable than the nonlinear Chebyshev iteration; see [21,22]. Note that

$$\left|\frac{\beta}{\gamma}\right| = \left|\vartheta^{-2}\right| = e^{-2\ \mathrm{Re}\ \phi} < 1\,. \tag{46}$$

Therefore, for the three-term version of the second-order Richardson iteration, all the multiplicative factors in (10) of Theorem 1 are actually smaller than 1 if β and γ are computed with sufficient accuracy.

The conformal map associated with the recursion (43) is [4]

$$f(\zeta) :\equiv \frac{\gamma}{\zeta} + \alpha + \beta\,\zeta\,. \tag{47}$$

In view of (46) f maps a neighborhood of the unit disk one-to-one onto the exterior of an ellipse with the foci $\alpha \pm c$. In particular, the disk $D_{\widehat{\rho}}$ around 0 with radius $\widehat{\rho} :\equiv |\gamma/\beta| = |\vartheta^2|$ is mapped onto the exterior of the interval or line segment $[\alpha - c, \alpha + c]$. If $1 < \rho < \widehat{\rho}$, the disk D_ρ is mapped onto the exterior of a confocal ellipse, and if all the eigenvalues of $\mathbf{A}$ lie in this ellipse, the iteration converges asymptotically at least with the rate $1/\rho$. If all the eigenvalues lie on $[\alpha - c, \alpha + c]$, the asymptotic rate is $1/\widehat{\rho}$. These asymptotic rates are the same for the Chebyshev iteration.

7 Numerical results

We consider first real matrices of order 500 whose eigenvalues are randomly chosen as complex conjugate pairs in an ellipse with foci $\alpha \pm c$ and longer semi-axis a. These matrices have been constructed by unitarily transforming a block-diagonal matrix (with 2×2 blocks) with these randomly chosen

[4] In [9–11,21,22] the mapping p related to f by $f(\zeta) = 1 - 1/p(\zeta)$ is considered instead.

eigenvalues. Note that these matrices are not very ill-conditioned as long as the ellipse does not come very close to the origin: they are normal and their condition number is bounded by the quotient of the distances from the origin of the farthest point and the closest point. However, if we considered very ill-conditioned matrices instead, the rate of convergence would be very slow. We report the number n_{12} of iterations needed to reduce the residual norm by a factor of 10^{12} and the ultimate relative accuracy where the residual norm stagnates. Table 1 summarizes the results for four such matrices for the three-term, two-term, and Rutishauser versions of the Chebyshev iteration using recursively computed residuals. Table 2 contains the corresponding results if explicitly computed residuals are used instead. We see that in these examples the number of iterations needed to reach relative accuracy 10^{-12} is not affected by the choice of the version. The ultimate accuracy is worst for the three-term version with updated residuals, and by replacing them by explicitly computed residuals we gain nearly up to two orders of magnitude. In other words, for the three-term version with updated residuals the loss of accuracy is notable, but not really serious. This reflects what we can expect from Theorem 2. For all the other versions, the ultimate accuracy is higher than 10^{-14}.

Table 1
Comparison of the three-term, two-term, and Rutishauser versions of the Chebyshev iteration using recursively computed residuals. Normal matrices with eigenvalues in the ellipse with foci $\alpha \pm c$ and semi-axis a.

matrix			**3-term**		**2-term**		**Rutishauser**	
α	c	a	ult.acc.	n_{12}	ult.acc.	n_{12}	ult.acc.	n_{12}
100	50	90	1.6e-14	195	1.6e-15	195	2.1e-15	195
100	70	90	5.9e-15	159	1.7e-15	159	2.3e-15	159
100	70	99	4.9e-14	1663	3.2e-15	1663	3.9e-15	1663
100	90	99	1.1e-13	1040	3.1e-15	1040	5.7e-15	1040

Table 2
Comparison of the three-term, two-term, and Rutishauser versions of the Chebyshev iteration using explicitly computed residuals. Normal matrices with eigenvalues in the ellipse with foci $\alpha \pm c$ and semi-axis a.

matrix			**3-term**		**2-term**		**Rutishauser**	
α	c	a	ult.acc.	n_{12}	ult.acc.	n_{12}	ult.acc.	n_{12}
100	50	90	9.2e-16	195	1.0e-15	195	9.1e-16	195
100	70	90	9.1e-16	159	9.5e-16	159	9.3e-16	159
100	70	99	2.1e-15	1663	1.7e-15	1663	1.7e-15	1663
100	90	99	1.8e-15	1040	1.9e-15	1040	1.7e-15	1040

In Figures 1–3 we show for the first example with $\alpha = 100$, $c = 50$, and $a = 90$ the residual histories for the two three-term versions, the two two-term versions, and the two Rutishauser versions, respectively. For the algorithms with residual recursions, both the true residuals and the recursively updated residuals are plotted. Needless to say that for the algorithms using explicitly computed residuals there is no difference between those and the true residuals and thus only one curve is shown.

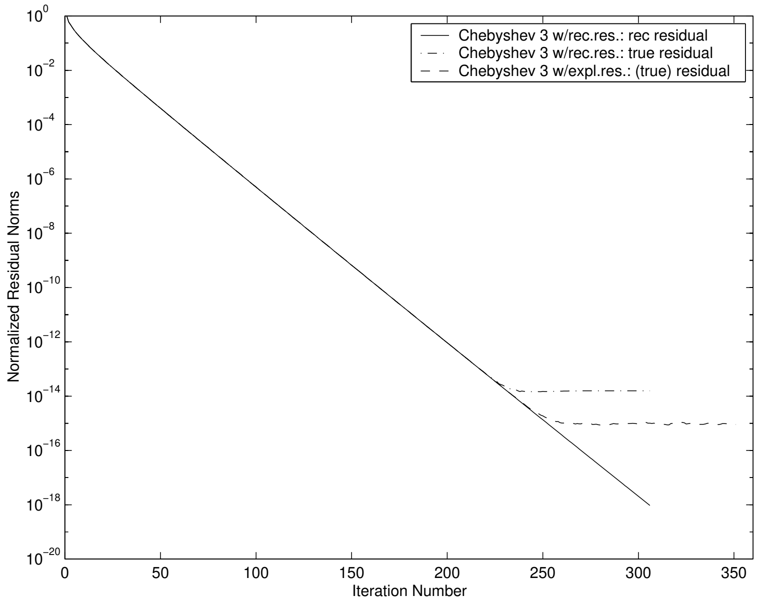

Fig. 1. Chebyshev iteration with three-term recursions

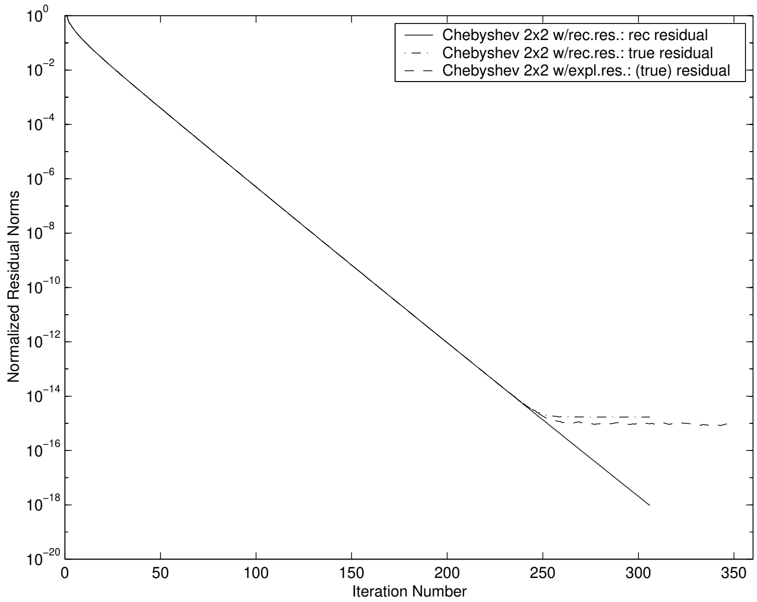

Fig. 2. Chebyshev iteration with coupled two-term recursions

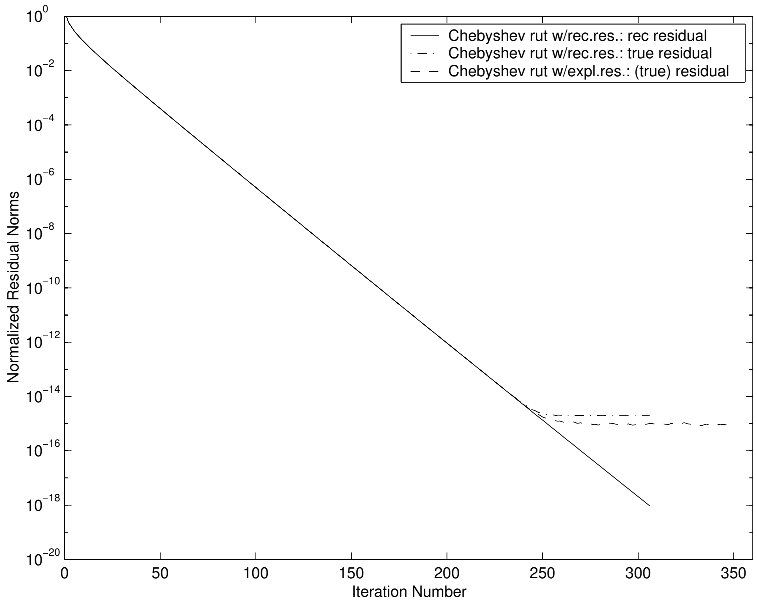

Fig. 3. Chebyshev iteration with Rutishauser's recursions for updating corrections

8 Discussion of the potential roundoff amplification in the three-term Chebyshev algorithm in the case of complex data

Now we want to try to construct an example with much stronger degradation of the ultimate accuracy in case of the three-term version with updated residuals. We know that the influence of the roundoff hinges in this case mainly on the factors (12a)–(12c) in (10). Clearly, the absolute value of the factor (12c),

$$\left|\frac{\beta_k \beta_{k+1} \cdots \beta_{n-1}}{\gamma_{k+1}\gamma_{k+2} \cdots \gamma_n}\right| = \left|\frac{T_k(\eta) T_{k+1}(\eta)}{T_n(\eta) T_{n+1}(\eta)}\right| \quad (0 \le k \le n-1) \tag{48}$$

(which simplifies to the absolute value of (12b) if $k = n-1$) can become large, if the absolute value of the denominator is very small, that is if η is close to a zero of T_n or T_{n+1}. These zeros all lie in the interval $(-1, 1)$, while $|T_n(\eta)| > 1$ if $\eta > 1$ or $\eta < -1$. Hence we need a complex η to get a small denominator. In Figure 4 we display this factor for $k = 1$ and $n = 3$ as a function of η in the domain $0 < \mathrm{Re}\,\eta \le 2$, $0 \le \mathrm{Im}\,\eta \le 0.5$. The poles of the function at the three positive zeros of T_3 and T_4 are well visible, although the values of the function on $\mathrm{Re}\,\eta = 0$ (where the poles are) are not plotted; the zero of T_3 at $\eta = 0$ cancels with the one of T_1. Clearly, we can make the fraction as large as we want by choosing η close enough to a pole. Then at least one term in (10) will be large.

However, if η is close to such a pole (and, hence, to a point in the interval

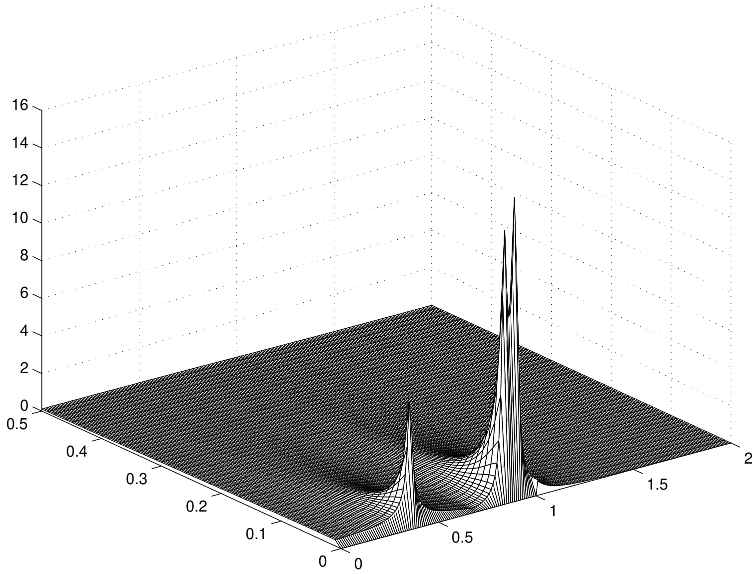

Fig. 4. The factor $(\beta_1\beta_2)/(\gamma_2\gamma_3)$ [that is, (48) for $k = 1$ and $n = 3$] in (10) as a function of $\eta = -\alpha/c$.

$(-1, 1)$), say to a zero of T_n, then the residual polynomial p_n of (1) is large at some points of the prescribed, necessarily very flat elliptic domain. (Recall that the straight line segment determined by the foci of the ellipse must come very close to the origin, but the ellipse must not contain the origin. If the ellipse were not flat, the quotient $\eta = -\alpha/c$ would not be close to a point in the interval (-1,1) unless the ellipse would contain the origin.) Therefore, the residual $\mathbf{r}_n$ of a system with a matrix whose spectrum is spread in this ellipse or on the straight line segment will most likely have some eigensystem components that have not been damped or have even been amplified. It is not of importance whether the matrix is diagonalizable or not.

There is the question what happens with the other quotients in (10). To explore that, we show in Figures 5 and 6 the factors (48) for $0 \le k \le n-1 < 100$ when $\eta = 0.85 + 0.05i$ and when $\eta = 0.865 + 0.001i$, respectively. In the first case, the plot shows a clear ridge where $k = n-1$, but except for small values of n, the quotient $|\beta_{n-1}/\gamma_n|$ remains smaller than one.

In fact, since $\beta_{n-1} \to \beta$ and $\gamma_n \to \gamma$ (see (41)), and since the asymptotic convergence rate $|\beta/\gamma|$ is bounded by 1 (see (46)), this is what we must expect. Moreover, this asymptotic rate is also the asymptotic convergence factor of both the Chebyshev iteration and the second order Richardson iteration if the eigenvalues of $\mathbf{A}$ lie on the straight line segment $[\alpha - c, \alpha + c]$. A rate close to 1 means that, in general (that is, when the eigenvalues of $\mathbf{A}$ can be anywhere on the line segment), the iteration will converge very slowly. In Figure 5 this rate is around 0.83. Away from the ridge, the factors (48) quickly decay.

The second plot shows a few very high peaks and a regular pattern of many

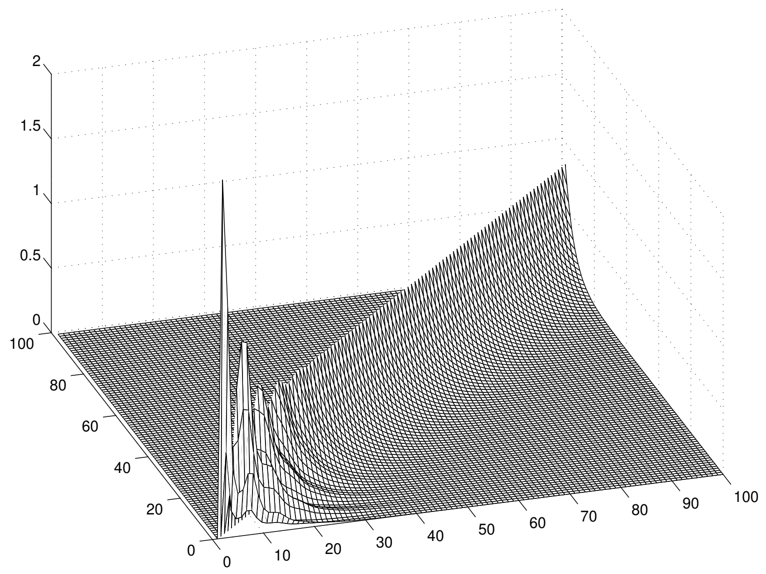

Fig. 5. The factors (48) in (10) for $\eta = 0.85 + 0.05i$ as a function of k and n.

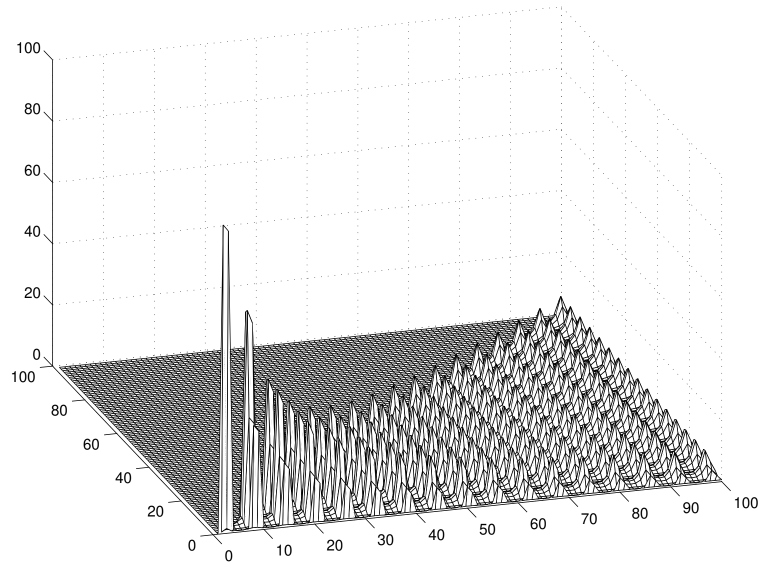

Fig. 6. The factors (48) in (10) for $\eta = 0.865 + 0.001i$ as a function of k and n.

smaller peaks that are still higher than 1. (Note the new scale of the vertical axis!) But, in view of what we just said, this can only mean that on the line $k = n - 1$ the quotients are still far away from their asymptotic value, which is around 0.996 here. So, in an example with a matrix with this kind of spectrum we might notice a serious influence of roundoff propagation on the ultimate accuracy, but the method would converge so slowly that we rather do not want to apply it. In the initial phase the residuals may strongly increase in this situation, because some of the residual polynomials are large on the line segment.

9 Conclusions

We have compared six different implementations of Chebyshev iteration with respect to convergence speed and ultimate accuracy attained. Several conclusions can be drawn from both theoretical and experimental investigations. The same theoretical conclusions also hold, and the same experimental ones can be expected to hold, for the related stationary method, the second order Richardson iteration.

- In our fairly well-conditioned examples, the number of iterations needed to reduce the residual norm by 10^{12} did not depend on which of the six versions is applied.
- The ultimate accuracy turned out worst for the classical 3-term recursion with recursively computed residuals, as had to be expected from theoretical results.
- Explicitly computed residuals yield the higher ultimate accuracy, and, for all three types of iterations, roughly the same.
- In contrast to CG, BiCG, and related methods, explicitly computed residuals do not cause a slowdown of convergence. They also do not have higher computational cost. Therefore they should be preferred.

If the (standard) three-term recursion for the residuals is applied nevertheless, the ultimate accuracy is still likely to be quite high, and this for the following reasons:

- If the Chebyshev iteration is applied to a matrix with spectrum on an interval $[\alpha - c, \alpha + c] \subset \mathbb{R}$ or an ellipse with foci $\alpha \pm c$ symmetric about the real axis, then, in contrast to CG and BiCG, the loss of ultimate accuracy is never very pronounced, because the multiplicative factors in (10) in front of the local errors in the expression for the residual gap are all of absolute value smaller than one if the recurrence coefficients are computed with sufficient accuracy.
- If the Chebyshev iteration is applied to an ellipse whose foci $\alpha \pm c$ do not lie on the real axis, but for which the line segment $[\alpha - c, \alpha + c]$ passes close to the origin (which implies that the ellipse must be very flat or must collapse to an interval), then some local errors may amplify dramatically and might cause a large residual gap, so that the ultimate accuracy deteriorates. However, this can only happen when the Chebyshev iteration converges very slowly.

Acknowledgment. The authors are grateful to Zdeněk Strakoš for pointing out several misprints and suggesting a number of improvements.

References

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

[2] D. A. Flanders, G. Shortly, Numerical determination of fundamental modes, J. Appl. Phys. 21 (1950) 1326–1332.

[3] G. H. Golub, R. S. Varga, Chebyshev semiiterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods, Numer. Math. 3 (1961) 147–168.

[4] 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, Nr. 8, Birkhäuser, Basel, 1959, pp. 24–49.

[5] A. J. Clayton, Further results on polynomials having least maximum modulus over an ellipse in the complex plane, U.K.A.E.A. Memorandum AEEW – M348 (1963).

[6] B. Fischer, R. W. Freund, On the constrained Chebyshev approximation problem on ellipses, Journal of Approximation Theory 62 (1990) 297–315.

[7] B. Fischer, R. W. Freund, Chebyshev polynomials are not always optimal, Journal of Approximation Theory 65 (1991) 261–272.

[8] D. K. Faddeev, V. N. Faddeeva, Computational Methods of Linear Algebra, Freeman, San Francisco, 1963.

[9] W. Niethammer, Iterationsverfahren und allgemeine Euler-Verfahren, Math. Z. 102 (1967) 288–317.

[10] W. Niethammer, R. S. Varga, The analysis of k-step iterative methods for linear systems from summability theory, Numer. Math. 41 (1983) 177–206.

[11] M. Eiermann, W. Niethammer, R. S. Varga, A study of semiiterative methods for nonsymmetric systems of linear equations, Numer. Math. 47 (1985) 505–533.

[12] M. H. Gutknecht, Z. Strakoš, Accuracy of two three-term and three two-term recurrences for Krylov space solvers, SIAM J. Matrix Anal. Appl. 22 (1) (2000) 213–229.

[13] G. H. Golub, M. Overton, The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems, Numer. Math. 53 (1988) 571–591.

[14] A. Greenbaum, Z. Strakoš, Predicting the behavior of finite precision Lanczos and conjugate gradient computations, SIAM J. Matrix Anal. Appl. 13 (1) (1992) 121–137.

[15] D. M. Young, K. C. Jea, Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods, Linear Algebra Appl. 34 (1980) 159–194.

[16] M. H. Gutknecht, Changing the norm in conjugate gradient type algorithms, SIAM J. Numer. Anal. 30 (1993) 40–56.

[17] A. Greenbaum, Estimating the attainable accuracy of recursively computed residual methods, SIAM J. Matrix Anal. Appl. 18 (3) (1997) 535–551.

[18] G. L. G. Sleijpen, H. A. van der Vorst, D. R. Fokkema, BiCGstab(l) and other hybrid Bi-CG methods, Numerical Algorithms 7 (1994) 75–109.

[19] A. Greenbaum, Accuracy of computed solutions from conjugate-gradient-like methods, in: M. Natori, T. Nodera (Eds.), Advances in Numerical Methods for Large Sparse Sets of Linear Systems, no. 10 in Parallel Processing for Scientific Computing, Keio University, Yokahama, Japan, 1994, pp. 126–138.

[20] S. Röllin, Auf 2-Term-Rekursionen beruhende Produktmethoden vom Lanczos-Typ, diploma thesis, Seminar of Applied Mathematics, ETH Zurich (2000).

[21] M. H. Gutknecht, W. Niethammer, R. S. Varga, k-step iterative methods for solving nonlinear systems of equations, Numer. Math. 48 (1986) 699–712.

[22] M. H. Gutknecht, Stationary and almost stationary iterative (k, l)-step methods for linear and nonlinear systems of equations, Numer. Math. 56 (1989) 179–213.