Accurate conjugate gradient methods for families of shifted systems*

Jasper van den Eshof† Gerard L. G. Sleijpen*

December 11, 2003

Abstract

We consider the solution of the linear system

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})\mathbf{x}^{\sigma} = \mathbf{A}^{\mathrm{T}}\mathbf{b},$$

for various real values of σ. This family of shifted systems arises, for example, in Tikhonov regularization and computations in lattice quantum chromodynamics. For each single shift σ this system can be solved using the conjugate gradient method for least squares problems (CGLS). In literature various implementations of the, so-called, multishift CGLS methods have been proposed. These methods are mathematically equivalent to applying the CGLS method to each shifted system separately but they solve all systems simultaneously and require only two matrix-vector products (one by $\mathbf{A}$ and one by $\mathbf{A}^{\mathrm{T}}$) and two inner products per iteration step. Unfortunately, numerical experiments show that, due to roundoff errors, in some cases these implementations of the multishift CGLS method can only attain an accuracy that depends on the square of condition number of the matrix $\mathbf{A}$. In this paper we will argue that, in the multishift CGLS method, the impact on the attainable accuracy of rounding errors in the Lanczos part of the method is independent of the effect of roundoff errors made in the construction of the iterates. By making suitable design choices for both parts, we derive a new (and efficient) implementation that tries to remove the limitation of previous proposals. A partial roundoff error analysis and various numerical experiments show promising results.

Key words: Tikhonov regularization , iterative methods , accuracy , finite precision arithmetic , shifted systems

1 Introduction

In various scientific computations the problem arises to compute solutions to

$$(\mathbf{A} + \sigma\mathbf{I})\mathbf{x}^{\sigma} = \mathbf{b}, \tag{1}$$

for various values of σ. The matrix $\mathbf{I}$ denotes the identity matrix. Krylov subspace methods are iterative solution methods for solving linear systems. These methods, with zero initial

*The research of J. van den Eshof was financially supported by the Dutch scientific organization (NWO) through project 613.002.035.

†Mathematical Institute, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands. `www.math.uu.nl/people/[eshof,sleijpen]` Email: `[eshof,sleijpen]@math.uu.nl`.

guess, are characterized by the fact that they construct their approximations in step j from the so-called j dimensional *Krylov subspace* defined as $\mathcal{K}_j(\mathbf{A}, \mathbf{b}) \equiv \text{span}\{\mathbf{b}, \mathbf{A}\mathbf{b} \ldots, \mathbf{A}^{j-1}\mathbf{b}\}$. An important property of Krylov subspaces is that they are shift invariant, that is $\mathcal{K}_j(\mathbf{A}, \mathbf{b}) = \mathcal{K}_j(\mathbf{A} + \sigma\mathbf{I}, \mathbf{b})$. By exploiting this property, Equation (1) can be solved for various values of the shift σ by constructing a basis for the Krylov subspace only once. This observation has led to many efficient implementations of known Krylov subspace methods that can handle multiple shifts simultaneously. We refer the interested reader for further information and applications to [4, 6, 9, 11, 18, 8, 23, 10, 24]. The multishift variants, in general, require the number of matrix-vector products and inner products of the original method applied to a single system and for the solution of each additional shifted system only a few extra vector updates are needed.

In this paper we focus on the numerical solution of the system

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})\mathbf{x}^{\sigma} = \mathbf{A}^{\mathrm{T}}\mathbf{b}, \tag{2}$$

for various real values of $\sigma \geq 0$. The solution of this family of systems plays an important role in Tikhonov regularization [8] and in the computation of the overlap operator in simulations in quantum chromodynamics [20]. Despite the fact that this system is sensitive to the effects of using computer arithmetic, since $\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I}$ can be ill conditioned, there is overwhelming numerical evidence that accurate solutions to this shifted system can be obtained by applying a sufficient number of iterations of the CGLS method, a variant of the CG method designed for solving least squares problems. Unfortunately, previously proposed implementations of multishift-type for solving (2), i.e., *multishift* CGLS methods, can in some cases only achieve a final precision for the shifted systems that depends on the square of the condition number of the matrix due to roundoff errors. In the present paper, we will demonstrate this with a numerical example (cf. Section 6.2). The main goal of this paper is to derive a new implementation that tries to overcome this limitation.

The use of computer arithmetic affects the standard CGLS method in two ways: it alters the convergence properties of the method, and restricts the accuracy of the method that can eventually be achieved. This is also the case for the multishift versions of the CGLS method. In this paper we mainly focus on the second aspect. Our implementation tries to improve previous proposals in this area.The ultimately attainable accuracy of CG-type methods is often investigated, e.g., [14, 3], by considering the *residual gap* which is defined as the difference between the recursively computed approximation to the residual and the unknown true residual. For the CGLS method this is fairly straightforward. For the multishift variants the situation is more complicated and in order to understand the influence of finite precision arithmetic on the attainable accuracy of the multishift CGLS method, we will discriminate between rounding errors made in the construction of the basis for the Krylov subspace (i.e., the Lanczos part) and in the computation of the approximate solution vector from the Krylov subspace (the inversion part). The subdivision into a Lanczos and inversion part is not immediately visible in the standard implementations of the CGLS method and is also not reflected by the analysis of the attainable accuracy through an inspection of the residual gap. We stress that in the multishift context these two parts are necessarily independent and it is *a priori* not clear if it is even possible to develop a multishift version of the CGLS method that is able to obtain approximations with similar precisions as a direct application of the CGLS method to each system separately. In this paper we propose an implementation of the multishift CGLS method by making suitable design choices for the implementation of the Lanczos part and

for the computation of the iterates for the shifted systems. Confidence in the success of our method will be given by a partial rounding error analysis (for the important situations that $\sigma = 0$ and $\sigma \to \infty$) and, partially, by numerical experiments.

This paper has the following structure. In Section 2 we review the CG method and its variant for least squares problems (CGLS) and we discuss some well-known results on their attainable accuracy. An abstract formulation of the multishift CGLS method is given in Section 3. The implementation of the 'Lanczos part' is the subject of Section 4. We review an important result by Paige on the finite precision behavior of this method. As an alternative for computing an orthogonal basis for the Krylov subspace, we propose to use the CGLS recurrences for computing an orthogonal basis for the Krylov subspace. Section 5 deals with the influence of rounding errors made in the 'alternative' Lanczos method on the attainable accuracy of the multishift CGLS method. The topic of Section 6 is the accurate computation of the iterates for the shifted systems. Finally, we show by several numerical experiments that, if both main ingredients are chosen properly, we can achieve high accuracy for the shifted systems.

2 Conjugate gradient methods

In this section we review two variants of the conjugate gradient method from the paper of Hestenes and Stiefel [17] and some of their properties. The conjugate gradient method is an iterative solution method for solving linear systems when the matrix $\mathbf{A}$ is symmetric positive definite. In this method the *iterates*, $\mathbf{x}_j$ and their corresponding *residuals*, $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{x}_j$, are computed for $j = 1, \ldots, k$ using the recurrence relations

$$\mathbf{r}_j = \mathbf{r}_{j-1} - \alpha_{j-1}\mathbf{c}_{j-1}, \quad \mathbf{c}_{j-1} = \mathbf{A}\mathbf{p}_{j-1}, \quad \mathbf{p}_j = \mathbf{r}_j + \beta_{j-1}\mathbf{p}_{j-1}, \tag{3}$$

with the coefficients given by

$$\alpha_{j-1} = \frac{\phi_{j-1}}{\mathbf{p}_{j-1}^{\mathrm{T}}\mathbf{c}_{j-1}}, \quad \beta_{j-1} = \frac{\phi_j}{\phi_{j-1}}, \quad \phi_j = \|\mathbf{r}_j\|^2,$$

and, initially, $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{b}$. (Norms in this paper are Euclidean.) The approximate solution then follows using the recurrence

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \alpha_{j-1}\mathbf{p}_{j-1}, \quad \text{with } \mathbf{x}_0 = \mathbf{0}. \tag{4}$$

In practice nonzero starting vectors are sometimes used but, for future convenience, we will assume that the initial guess, $\mathbf{x}_0$, is zero here. A key characterization of the CG method is that its iterates, $\mathbf{x}_j$, minimize the error in the energy norm (that is an $\mathbf{A}$-weighted norm) over all approximations from the j-th Krylov subspace $\mathcal{K}_j(\mathbf{A}, \mathbf{b})$, see e.g., [17]. As a consequence of this, the residuals $\mathbf{r}_i$ for $i = 0, \ldots, j-1$ form an orthogonal basis for $\mathcal{K}_j(\mathbf{A}, \mathbf{b})$. Another useful property is the fact that $\alpha_i \mathbf{r}_j^{\mathrm{T}}\mathbf{p}_i \geq 0$ for all $i \geq 0$, which follows from Equation (5.2) in [17] and the positivity of the α_i. As a consequence we have, in combination with (4), that

$$\|\mathbf{x}_j\| \leq \|\mathbf{x}_{j+1}\| \quad \text{and} \quad \|\mathbf{x}_i - \mathbf{x}_j\| \leq \|\mathbf{x}_i\| \quad \text{for} \quad 0 \leq j \leq i. \tag{5}$$

In the following proposition we summarize an important result, due to Greenbaum, that helps to provide insight into the attainable accuracy of the standard conjugate gradient method.

Proposition 2.1 (Greenbaum [14]) *The difference between the true residual* $\mathbf{b}-\mathbf{A}\mathbf{x}_k$ *and the computed residual* $\mathbf{r}_k$ *satisfies*

$$\frac{\|(\mathbf{b}-\mathbf{A}\mathbf{x}_k)-\mathbf{r}_k\|}{\|\mathbf{A}\|\|\mathbf{x}\|} \leq \varepsilon\,(k+1+(1+c+k(10+2c))\Theta_k)+\mathcal{O}(\varepsilon^2),$$

with

$$\Theta_k = \max_{j\leq k}\frac{\|\mathbf{x}_j\|}{\|\mathbf{x}\|},$$

ε *is the* unit roundoff*, which for double precision computations is in the order of* 10^{-16} *and* c *is a constant that depends on the error in the matrix-vector product.*

The quantity Θ_k is difficult to bound in computer arithmetic. However, in *exact* arithmetic it immediately follows from (5) that $\Theta_k \leq 1$. The argument to bound this quantity used in [14] depends on the fact that, also in *exact* arithmetic, the errors for the CG method are in Euclidean norm monotonically decreasing [17, Theorem 6.3] and, therefore,

$$\|\mathbf{x}_k-\mathbf{x}\| \leq \|\mathbf{x}_0-\mathbf{x}\| = \|\mathbf{x}\| \quad \Rightarrow \quad \Theta_k \leq 2. \tag{6}$$

Unfortunately, the argument for proving the reduction of the error in Euclidean norm in [17] uses the orthogonality of the residuals which is in general lost in finite precision computations. However, using a relation of the CG method with an *exact* CG method applied to a larger matrix and specific right-hand side, the author argues in [14, Section 3.1] that (6) might also hold approximately in the finite precision context.

The relevance of Proposition 2.1 for explaining the attainable precision of the conjugate gradient method depends on the numerical observation that, in practical computations, the computed vector $\mathbf{r}_k$ converges to zero or, at least, stagnates when its norm is many orders of magnitude smaller than the machine precision ε. With these assumptions it, therefore, follows from Proposition 2.1 that for the iterate, $\mathbf{x}_k$, we essentially have for sufficiently large k that

$$\|\mathbf{b}-\mathbf{A}\mathbf{x}_k\| \leq \varepsilon\,\mathcal{O}(k)\,\|\mathbf{A}\|\|\mathbf{x}\|, \tag{7}$$

which implies the following bound on the relative error:

$$\|\mathbf{x}-\mathbf{x}_k\|/\|\mathbf{x}\| \leq \varepsilon\,\mathcal{O}(k)\,\|\mathbf{A}\|\|\mathbf{A}^{-1}\|.$$

For least squares problems the CG method can be directly applied to the normal equations. Nevertheless, it is common practice to use an alternative, but mathematically equivalent, variation of CG, known as CGLS, in this case [17, Section 10]. For $j=1,\ldots,k$ this method is defined by the following recurrence relations:

$$\mathbf{z}_j = \mathbf{z}_{j-1}-\alpha_{j-1}\mathbf{c}_{j-1}, \quad \mathbf{c}_{j-1}=\mathbf{A}\mathbf{p}_{j-1}, \quad \mathbf{r}_j=\mathbf{A}^{\mathrm{T}}\mathbf{z}_j, \quad \mathbf{p}_j=\mathbf{r}_j+\beta_{j-1}\mathbf{p}_{j-1}, \tag{8}$$

with

$$\alpha_{j-1} \equiv \frac{\phi_{j-1}}{\mathbf{c}_{j-1}^{\mathrm{T}}\mathbf{c}_{j-1}}, \quad \beta_{j-1} \equiv \frac{\phi_j}{\phi_{j-1}}, \quad \phi_j \equiv \|\mathbf{r}_j\|^2, \tag{9}$$

and $\mathbf{z}_0=\mathbf{b}$, $\mathbf{r}_0=\mathbf{p}_0=\mathbf{A}^{\mathrm{T}}\mathbf{z}_0$, and $\mathbf{x}_k$ is computed as in (4).

The advantage of this method, compared to applying CG directly to the normal equations, is that, here the least squares residuals, $\mathbf{z}_j=\mathbf{b}-\mathbf{A}\mathbf{x}_j$, are directly available. Furthermore,

it was shown in [14, Section 3.3] and [3], with similar arguments as used in the proof of Proposition 2.1 for CG, that recurring the residuals for the least squares problem, $\mathbf{z}_j$, improves the attainable accuracy of the method. For future convenience, we state here a result from [3].

Proposition 2.2 (Björck, Elfving and Strakoš [3]) *The difference between the true least squares residual,* $\mathbf{b} - \mathbf{A}\mathbf{x}_k$*, and the computed least squares residual,* $\mathbf{z}_k$*, satisfies*

$$\frac{\|(\mathbf{b} - \mathbf{A}\mathbf{x}_k) - \mathbf{z}_k\|}{\|\mathbf{A}\|\|\mathbf{x}\|} \leq \varepsilon\,(k + 1 + (1 + c + k(10 + 2c))\Theta_k) + \varepsilon(k + 1)\frac{\|\mathbf{z}\|}{\|\mathbf{A}\|\|\mathbf{x}\|} + \mathcal{O}(\varepsilon^2)$$

with

$$\Theta_k = \max_{j \leq k} \frac{\|\mathbf{x}_j\|}{\|\mathbf{x}\|},$$

c is a constant that depends on the error in the matrix-vector product with $\mathbf{A}$ *and* $\mathbf{z} = \mathbf{b} - \mathbf{A}\mathbf{x}$*.*

Again, to apply this result, the authors have to bound Θ_k, which can be accomplished with similar heuristics as used for the CG method. Assuming, furthermore, that the CGLS method is terminated at a point such that $\|\mathbf{z}_k - \mathbf{z}\| \approx d\varepsilon\|\mathbf{A}\|\|\mathbf{x}\|$, the authors conclude that the CGLS method is eventually expected to achieve an approximate solution that is as good as any forward stable method for solving the least squares problem. We return to this in more detail in Section 5.2.

The *damped least squares problem* (2) is equivalent to the least squares problem

$$\min_{\mathbf{x}^\sigma} \left\| \begin{bmatrix} \mathbf{A} \\ \sqrt{\sigma}\mathbf{I} \end{bmatrix} \mathbf{x}^\sigma - \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} \right\|. \tag{10}$$

Hence, the shifted system can be solved, for one shift, by applying the CGLS method to this augmented matrix and right-hand side. However, due to the special structure of this matrix and right-hand side some computational work can be saved in the CGLS method, see, for example, Algorithm 6 in [8]. For convenience of the reader the resulting algorithm is summarized in Alg. 1.

3 An abstract formulation of the multishift CGLS method

In the previous section, we reviewed the celebrated conjugate gradient least squares method from [17]. In exact arithmetic, a mathematical equivalent method can be obtained based on the *Lanczos* method, e.g., [12, Chapter 9], applied to the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ with starting vector $\mathbf{A}^{\mathrm{T}}\mathbf{b}$. This method constructs an orthonormal basis for the Krylov subspace $\mathcal{K}_k(\mathbf{A}^{\mathrm{T}}\mathbf{A}, \mathbf{A}^{\mathrm{T}}\mathbf{b})$. After k iterations, the process can be summarized by the *Lanczos relation*:

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A})\mathbf{V}_k = \mathbf{V}_k T_k + \delta_{k-1}\mathbf{v}_k e_k^{\mathrm{T}} = \mathbf{V}_{k+1}\underline{T}_k, \tag{11}$$

where the columns $\mathbf{v}_0, \ldots, \mathbf{v}_k$ of $\mathbf{V}_{k+1}$, form an orthonormal basis for $\mathcal{K}_{k+1}(\mathbf{A}^{\mathrm{T}}\mathbf{A}, \mathbf{A}^{\mathrm{T}}\mathbf{b})$ and the symmetric $k \times k$ tridiagonal matrix T_k collects the coefficients computed during the execution of the Lanczos algorithm. It is often convenient to include also δ_{k-1} into one $k+1$ by k tridiagonal matrix $\underline{T}_k$ by adding an additional row, $e_k^{\mathrm{T}}\delta_{k-1}$, to T_k. The vector e_k denotes the k-th standard basis vector, i.e., $(e_k)_j = 0$ for all $j \neq k$ and $(e_k)_k = 1$. Furthermore, $\vec{1}$ is the

$\mathbf{z}_0 = \mathbf{b}$, $\mathbf{r}_0 = \mathbf{A}^{\mathrm{T}}\mathbf{z}_0$, $\mathbf{p}_0 = \mathbf{r}_0$, $\mathbf{x}_0 = \mathbf{0}$, $\phi_0 = \|\mathbf{r}_0\|^2$
for $j = 1, \ldots k$

$$\begin{aligned}
&\mathbf{c}_{j-1} = \mathbf{A}\mathbf{p}_{j-1} \\
&\alpha_{j-1} = \phi_{j-1}/(\|\mathbf{c}_{j-1}\|^2 + \sigma\|\mathbf{p}_{j-1}\|^2) \\
&\mathbf{x}_j = \mathbf{x}_{j-1} + \alpha_{j-1}\mathbf{p}_{j-1} \\
&\mathbf{z}_j = \mathbf{z}_{j-1} - \alpha_{j-1}\mathbf{c}_{j-1} \\
&\mathbf{r}_j = \mathbf{A}^{\mathrm{T}}\mathbf{z}_j - \sigma\mathbf{x}_j \\
&\phi_j = \|\mathbf{r}_j\|^2,\ \beta_{j-1} = \phi_j/\phi_{j-1} \\
&\mathbf{p}_j = \mathbf{r}_j + \beta_{j-1}\mathbf{p}_{j-1}
\end{aligned}$$

ALGORITHM 1. *CGLS implementation for solving (2).*

vector with all components one and, similarly, $\vec{0}$ is the vector with all components zero. The dimension of these vectors should be apparent from the context. The approximation in step k for the corresponding CG process is equal to

$$\mathbf{x}_k = \mathbf{V}_k T_k^{-1} e_1 \sqrt{\phi_0}, \quad \text{with } \phi_0 \equiv \mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0, \tag{12}$$

and the vector $\mathbf{r}_0 = \mathbf{A}^{\mathrm{T}}\mathbf{b}$ is the initial residual. The Equations (11) and (12) together give us an abstract formulation of the CGLS method.

It follows from (11) that the Lanczos relation for the shifted system reads

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})\mathbf{V}_k = \mathbf{V}_k(T_k + \sigma I) + \delta_{k-1}\mathbf{v}_k e_k^{\mathrm{T}}. \tag{13}$$

This shows that, if the Lanczos relation is known for one system, the iterates of the CGLS method for the damped least squares problem (10) can be directly computed from this relation and are, in fact, equal to

$$\mathbf{x}_k^{\sigma} = \mathbf{V}_k(T_k + \sigma I)^{-1} e_1 \sqrt{\phi_0}, \quad \text{with } \phi_0 \equiv \mathbf{r}_0^{\mathrm{T}}\mathbf{r}_0. \tag{14}$$

A *multishift* CGLS method constructs the orthonormal basis once and at the same time computes (14) for the required values of σ, of course without storing all Lanczos vectors. Since there are many, mathematically equivalent, ways to compute the Lanczos vectors and tridiagonal matrix T_k and just as many ways to compute the vectors in (14), the number of possible implementations is countless. Two specific implementations are presented in [18] and [8]. However, as discussed in the introduction, the obtainable precision of the computed iterates for these implementations can be limited. We propose a new (and efficient) implementation by choosing a suitable algorithm for the Lanczos part and for computing the iterates (14) and we will discuss the relationship of previously proposed multishift CG-type methods to our implementation in the course of this paper.

4 The implementation for the Lanczos part

As discussed in the previous section, the first key ingredient of a multishift CGLS method is the construction of an orthonormal, or orthogonal, basis for the Krylov subspaces and

the tridiagonal matrix containing the orthogonalization coefficients. The obvious choice is to apply the standard Lanczos method, e.g., [12, Algorithm 9.2.1] to the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ with starting vector $\mathbf{A}^{\mathrm{T}}\mathbf{b}$. It is not difficult to see that this is not an optimal choice in the context of solving least squares problem. We will give detailed arguments in Section 5 when we discuss the impact of rounding errors in the Lanczos part on the attainable accuracy of the multishift CGLS method.

It is well known that the coupled two-term recurrences of the CG method can be used as an alternative to the application of the Lanczos method for constructing an orthonormal basis for the Krylov subspace, see [7, 1]. Numerical experiments in these papers showed that alternative recurrences could in some cases considerably improve the robustness of the methods. In a similar spirit, the recurrences in (8) can be used to build an orthonormal basis for the Krylov subspace $\mathcal{K}_j(\mathbf{A}^{\mathrm{T}}\mathbf{A}, \mathbf{A}^{\mathrm{T}}\mathbf{b})$. In this section we consider the use of the CGLS recurrences as alternative Lanczos-type method which we will refer to as the *CGLS-Lanczos* method.

First, a little remark about notational conventions: with $\mathbf{R}_k$ we denote the $n \times k$ matrix with columns $\mathbf{r}_0, \ldots, \mathbf{r}_{k-1}$. Similarly, other capitals will be used to group together the corresponding vectors. Now, the relations in (8) can be summarized by the following matrix formulations

$$\mathbf{Z}_{k+1}\underline{J}_k = \mathbf{C}_k\Delta_k, \quad \mathbf{C}_k = \mathbf{A}\mathbf{P}_k, \quad \mathbf{R}_{k+1} = \mathbf{A}^{\mathrm{T}}\mathbf{Z}_{k+1}, \quad \mathbf{P}_k U_k = \mathbf{R}_k,$$

where

$$U_k \equiv \begin{bmatrix} 1 & -\beta_0 & & & \\ & 1 & -\beta_1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & -\beta_{k-2} \\ & & & & 1 \end{bmatrix}, \quad \Delta_k \equiv \begin{bmatrix} \alpha_0 & & & & \\ & \alpha_1 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \alpha_{k-1} \end{bmatrix},$$

and $\underline{J}_k$ is $k+1 \times k$ lower bidiagonal matrix with 1 and -1 on, respectively, the diagonal and sub-diagonal. Substitution yields the *residual relation*:

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A})\mathbf{R}_k = \mathbf{R}_{k+1}\underline{S}_k, \quad \text{with} \quad \underline{S}_k \equiv \underline{J}_k\Delta_k^{-1}U_k. \tag{15}$$

Apart from some difference in scaling, this is precisely the Lanczos relation given in (11). We can obtain the quantities of the Lanczos method in its standard form by using a simple diagonal scaling Ψ_k with diagonal elements $\psi_0, \ldots, \psi_{k-1}$ where $\psi_j = \phi_j^{1/2}$, which shows that (11) holds with

$$\mathbf{V}_k = \mathbf{R}_k\Psi_k^{-1}, \quad T_k = \Psi_k S_k \Psi_k^{-1} = L_k^{\mathrm{T}}\Delta_k^{-1}L_k \quad \text{with} \quad L_k = \Psi_k U_k \Psi_k^{-1}. \tag{16}$$

Since the ϕ_j are available in the CGLS method, the recurrences in (8) can be used as an alternative to applying the Lanczos method to $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ with starting vector $\mathbf{A}^{\mathrm{T}}\mathbf{b}$ at virtually the same cost.[1]

We have argued that the CGLS recurrences can be used as an alternative to the standard Lanczos method and, therefore, the iterates for the shifted systems can be computed using

[1]Notice that the vectors $\mathbf{v}_j$ (i.e., the columns of $\mathbf{V}_k$) are plus or minus the Lanczos vectors of the standard Lanczos method.

(14) with the quantities given in (16). An, equivalent, alternative that avoids the scaling in (16), is to directly compute

$$\mathbf{x}_k^\sigma = \mathbf{R}_k(S_k + \sigma I)^{-1} e_1,$$

with S_k defined as the upper $k \times k$ block of $\underline{S}_k$ given in (15).

4.1 Perturbed Lanczos relations in computer arithmetic

In finite precision computations, the computed Lanczos vectors $\mathbf{V}_k$ and tridiagonal matrix $\underline{T}_k$ do not satisfy the Lanczos relation exactly. For the standard Lanczos method, Paige [21] proved that, instead, the computed quantities now satisfy a perturbed Lanczos relation. In this section we give an analogous result for the CGLS-Lanczos method which we need in the remainder of this paper.

We assume the standard rules for floating point arithmetic with machine precision ε, see, e.g., [12, Section 2.4.2],

$$\mathrm{fl}[a \circ b] = (a \circ b)(1 + \varepsilon') \quad \text{with} \quad |\varepsilon'| \leq \varepsilon. \tag{17}$$

Here, a and b are floating point numbers and $\circ$ stands for any basic operation like addition, subtraction, multiplication and division. Furthermore, we assume that the errors in the matrix-vector product of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ with some vector $\mathbf{y}$ (of appropriate length) are, respectively, bounded by

$$\varepsilon\, c\, \|\mathbf{A}\|\, \|\mathbf{y}\| \quad \text{and} \quad \varepsilon\, c'\, \|\mathbf{A}^{\mathrm{T}}\|\, \|\mathbf{y}\|.$$

With this notation, the result of Paige, for a certain implementation of the Lanczos method, applied to the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and starting vector $\mathbf{A}^{\mathrm{T}}\mathbf{b}$ is summarized by the following proposition.

Proposition 4.1 (Paige [21]) *The Lanczos method in computer arithmetic results in a perturbed Lanczos relation*

$$(\mathbf{A}^T\mathbf{A})\mathbf{V}_k = \mathbf{V}_{k+1}\underline{T}_k + \mathbf{F}_k, \tag{18}$$

where, ignoring higher order terms,

$$\|\mathbf{F}_k\|_F \leq \varepsilon(6 + c + c')\sqrt{k}\|\mathbf{A}\|^2.$$

We now turn our attention to CGLS-Lanczos, the alternative Lanczos type method discussed in the previous section. Let $\mathbf{f}_j^c$ denote the perturbation in the computation of $\mathbf{c}_j$ caused by the use of computer arithmetic and assume similar notation for the other perturbations. Using the standard model for floating point arithmetic we get (cf. (8) and the initialization computations assuming that all quantities in step minus one have length zero)

$$\|\mathbf{f}_j^z\| \leq \varepsilon\left(\|\mathbf{z}_{j-1}\| + 2\|\alpha_{j-1}\mathbf{c}_{j-1}\|\right) + \mathcal{O}(\varepsilon^2) \leq \varepsilon\left(3\|\mathbf{z}_{j-1}\| + 2\|\mathbf{z}_j\|\right) + \mathcal{O}(\varepsilon^2) \tag{19}$$

$$\|\mathbf{f}_j^c\| \leq \varepsilon\, c\, \|\mathbf{A}\|\, \|\mathbf{p}_j\| \tag{20}$$

$$\|\mathbf{f}_j^r\| \leq \varepsilon\, c'\, \|\mathbf{A}^{\mathrm{T}}\|\, \|\mathbf{z}_j\| \tag{21}$$

$$\|\mathbf{f}_j^p\| \leq \varepsilon\left(\|\mathbf{r}_j\| + 2\|\beta_{j-1}\mathbf{p}_{j-1}\|\right) + \mathcal{O}(\varepsilon^2) \leq \varepsilon\left(3\|\mathbf{r}_j\| + 2\|\mathbf{p}_j\|\right) + \mathcal{O}(\varepsilon^2). \tag{22}$$

Combining this with the notation from the previous section, we find that the quantities computed by the CGLS recurrences in finite precision arithmetic satisfy the perturbed relations given by

$$\mathbf{Z}_{k+1}\underline{J}_k = \mathbf{C}_k\Delta_k + \mathbf{F}_k^z, \quad \mathbf{C}_k = \mathbf{A}\mathbf{P}_k + \mathbf{F}_k^c, \quad \mathbf{R}_{k+1} = \mathbf{A}^{\mathrm{T}}\mathbf{Z}_{k+1} + \mathbf{F}_{k+1}^r, \quad \mathbf{P}_k U_k = \mathbf{R}_k + \mathbf{F}_k^p.$$

Finally, a straightforward sequence of substitutions and multiplications with suitable matrices yields a perturbed relation which is summarized by the following lemma.

Lemma 4.1 *The CGLS recurrences given in* (8) *in finite precision computations lead to a perturbed relation of the form*

$$(\mathbf{A}^T\mathbf{A})\mathbf{R}_k = \mathbf{R}_{k+1}\underline{S}_k + \mathbf{F}_k,$$

where $\underline{S}_k$ *is given by* (15) *and the perturbation is of the form*

$$\mathbf{F}_k \equiv -(\mathbf{A}^T\mathbf{A})\mathbf{F}_k^p - \mathbf{A}^T(\mathbf{F}_k^c + \mathbf{F}_k^z\Delta_k^{-1})U_k - \mathbf{F}_{k+1}^r\underline{S}_k.$$

Upper bounds on the norms of the columns of the perturbations on the right are given by (19)-(22).

This lemma does not explicitly provide a bound on the norm of the perturbation as in Proposition 4.1 for the standard Lanczos method. We note that the norm of this perturbation can be much larger than the norm of the perturbation term for the standard Lanczos method (taking into account difference in scaling). This can be explained from the fact that $\|\mathbf{p}_j\|/\|\mathbf{r}_j\|$ can become very large which typically occurs when the matrix $\mathbf{A}$ has very small isolated eigenvalues. However, the perturbation term in Lemma 4.1 has an interesting structure. In the next section we will exploit this special structure to investigate the attainable accuracy of the multishift CGLS method based on CGLS-Lanczos.

Finally, we conclude this section by giving, for future use, an explicit expression for the iterate in step k of the CGLS method. Assuming that no roundoff errors are made in (4), this approximation is given by

$$\mathbf{x}_k^{\text{CGLS}} = \mathbf{P}_k\Delta_k J_k^{-1}e_1 = (\mathbf{R}_k U_k^{-1} + \mathbf{F}_k^p U_k^{-1})\Delta_k J_k^{-1}e_1 = \mathbf{R}_k S_k^{-1}e_1 + \mathbf{F}_k^p S_k^{-1}e_1. \tag{23}$$

The first expression follows from (4) and the fact that $J_k^{-1}e_1 = \vec{1}$.

5 The impact of errors in the Lanczos part on multishift CGLS

We discuss the effects that rounding errors in the Lanczos part have on the attainable accuracy of the multishift CGLS method. The first choice that we mentioned for the Lanczos part was an application of the standard Lanczos method applied to the matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and starting vector $\mathbf{A}^{\mathrm{T}}\mathbf{b}$. This is not a very good choice as will be argued using the same line of arguments as given in [3, Section 4.2] to explain the failure of the LSCG method. The main observation is that we have to startup the Lanczos method by computing $\mathbf{A}^{\mathrm{T}}\mathbf{b}$ to obtain the first Lanczos vector. Since the vector $\mathbf{b}$ does not appear in the Lanczos part, it follows that the multishift CGLS method based on the standard Lanczos method at best computes a solution to

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})\widetilde{\mathbf{x}}^\sigma = \mathbf{A}^{\mathrm{T}}\mathbf{b} + \mathbf{f}, \quad \text{where} \quad \|\mathbf{f}\| \le \varepsilon c'\|\mathbf{A}\|\|\mathbf{b}\|.$$

This shows that, in the nearly consistent case and $\sigma = 0$, the forward error is, at best, much less than optimal. For more details on this argument consult [3, Section 4.2].

We now focus on the alternative Lanczos procedure, CGLS-Lanczos. We assume that the multishift iterates are computed as

$$\mathbf{x}_k^\sigma = \mathbf{R}_k(S_k + \sigma I)^{-1}e_1,$$

where S_k is defined as in (15) and, moreover, for the moment no rounding errors are considered in the computation of this vector, i.e., $\mathbf{x}_k^\sigma$ is assumed to be computed exactly. In the multishift CGLS method the iterates are computed independently of the Lanczos part. As a consequence, this will generate additional sources of errors that are not present in the CGLS method while having impact on the attainable accuracy.

At this point, we warn the reader for some unconventional notation. If we apply a matrix with k columns to an ℓ-vector with $\ell \leq k$, then we assume the vector to be expanded with zeros if necessary (we do the same with other operations and equalities). In the case $\ell > k$, the matrix is assumed to be applied to the first k elements of the vector and the remaining elements are assumed to be unchanged by the operation.

For future convenience we define

$$(\gamma_k^\sigma)^{-1} \equiv -(e_{k+1}^{\mathrm{T}} \underline{S}_k e_k)\, e_k^{\mathrm{T}} (S_k + \sigma I)^{-1} e_1,$$

and we note that

$$e_{k+1}/\gamma_k^\sigma - (e_1 - \underline{S}_k (S_k + \sigma I)^{-1} e_1) = -\sigma (S_k + \sigma I)^{-1} e_1.$$

Using this relation and the relations given in Section 4.1, the true residual, corresponding to the normal equations, can be written as

$$\begin{aligned}
\mathbf{A}^{\mathrm{T}}\mathbf{b} - (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})\mathbf{x}_k^\sigma &= (\mathbf{A}^{\mathrm{T}}\mathbf{b} - \mathbf{r}_0) + \mathbf{r}_0 - (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})\mathbf{R}_k (S_k + \sigma I)^{-1} e_1 \\
&= -\mathbf{F}_{k+1}^r e_1 + \mathbf{r}_k/\gamma_k^\sigma - \mathbf{F}_k (S_k + \sigma I)^{-1} e_1 \\
&= -\mathbf{F}_{k+1}^r e_1 + \mathbf{F}_{k+1}^r e_{k+1}/\gamma_k^\sigma + \mathbf{A}^{\mathrm{T}}\mathbf{z}_k/\gamma_k^\sigma - \mathbf{F}_k (S_k + \sigma I)^{-1} e_1.
\end{aligned}$$

We plug in the expression for the perturbation term given by Lemma 4.1 and we arrive at our main relation:

$$\mathbf{A}^{\mathrm{T}}\mathbf{b} - (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})\mathbf{x}_k^\sigma = \mathbf{A}^{\mathrm{T}}\mathbf{z}_k/\gamma_k^\sigma + \mathbf{A}^{\mathrm{T}}\mathbf{w}_k^{(1)} + \mathbf{w}_k^{(2)} + \mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{w}_k^{(3)} \tag{24}$$

with

$$\mathbf{w}_k^{(1)} = (\mathbf{F}_k^c + \mathbf{F}_k^z \Delta_k^{-1}) U_k (S_k + \sigma I)^{-1} e_1, \quad \mathbf{w}_k^{(2)} = -\sigma \mathbf{F}_{k+1}^r (S_k + \sigma I)^{-1} e_1,$$
$$\mathbf{w}_k^{(3)} = \mathbf{F}_k^p (S_k + \sigma I)^{-1} e_1.$$

This expression plays a similar role in this section as Proposition 2.2 plays for the analysis of the attainable accuracy of the CGLS method. It shows that, if the vector $\mathbf{z}_k$ eventually approaches $\mathbf{z}$, and, therefore, the first term becomes very small, then the true residual corresponding to (2) stagnates at a level determined by the three $\mathbf{w}$-vectors in (24). To be more precise, we assume that we terminate, at an iteration step k, such that

$$\frac{\|\mathbf{z}_k - \mathbf{z}\|}{\|\mathbf{A}\|\|\mathbf{x}\|} \leq d\varepsilon, \tag{25}$$

for some constant d. Although there is no rigorous proof that this condition, eventually, always can be achieved, there is according to [3] "overwhelming experimental evidence" to justify this assumption, see also the numerical experiments in [3].

The attained forward error is given by multiplying the expression for the true residual (24) from the left with $(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}$. This gives

$$\begin{aligned}
\|\mathbf{x}^\sigma - \mathbf{x}_k^\sigma\| \;\leq\; & \|(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{z}_k\|/|\gamma_k^\sigma| + \|(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\|\,\|\mathbf{w}_k^{(1)}\| + \\
& \|(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}\|\,\|\mathbf{w}_k^{(2)}\| + \|(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{A}\|\,\|\mathbf{w}_k^{(3)}\|.
\end{aligned} \tag{26}$$

We stress that the exploitation of the special structure of the perturbation in Lemma 4.1 has played an important role in the derivation of this expression. We can get a similar expression for the least squares residual by multiplying from the left by $\mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\sigma\mathbf{I})^{-1}$. In the next section we bound the size of these $\mathbf{w}$-vectors for the important case that $\sigma = 0$ (notice that already for this simple case multishift CGLS based on the standard Lanczos method fails). In Section 5.2 we combine these bounds with (26) and compare the attainable accuracy of the CGLS method and the multishift CGLS method based on the CGLS-Lanczos method. We will, furthermore, discuss the situation of more general σ.

5.1 Bounding the size of the w-vectors for $\sigma = 0$

We bound the size of the vectors $\mathbf{w}_k^{(1)}$, $\mathbf{w}_k^{(2)}$ and $\mathbf{w}_k^{(3)}$ for the case of σ equal to zero. Since $S_k = J_k\Delta^{-1}U_k$ and $J_k^{-1}e_1 = \vec{1}$, the expression for the first $\mathbf{w}$-vector can be rewritten to

$$\mathbf{w}_k^{(1)} = (\mathbf{F}_k^c + \mathbf{F}_k^z\Delta_k^{-1})U_kS_k^{-1}e_1 = (\mathbf{F}_k^c\Delta_k + \mathbf{F}_k^z)\vec{1} = \sum_{j=0}^{k-1}(\alpha_j\mathbf{f}_j^c + \mathbf{f}_j^z).$$

Therefore, the estimates in (19)-(22) lead to the following bounds on the $\mathbf{w}$-vectors for $\sigma = 0$:

$$\begin{aligned}
\mathbf{w}_k^{(1)} = \sum_{j=0}^{k-1}(\alpha_j\mathbf{f}_j^c + \mathbf{f}_j^z) \quad &\Rightarrow \quad \|\mathbf{w}_k^{(1)}\| \le \varepsilon\sum_{j=0}^{k-1}\big(c\|\mathbf{A}\|\|\alpha_j\mathbf{p}_j\| + 5\|\mathbf{z}_j\|\big) + \mathcal{O}(\varepsilon^2)\\
\mathbf{w}_k^{(2)} = \mathbf{0} \quad &\Rightarrow \quad \|\mathbf{w}_k^{(2)}\| = 0\\
\mathbf{w}_k^{(3)} = \mathbf{F}_k^pS_k^{-1}e_1 \quad &\Rightarrow \quad \|\mathbf{w}_k^{(3)}\| \le \varepsilon\sum_{j=0}^{k-1}\big(3\|\mathbf{r}_j\| + 2\|\mathbf{p}_j\|\big)\,|e_{j+1}^{\mathrm{T}}S_k^{-1}e_1| + \mathcal{O}(\varepsilon^2).
\end{aligned}$$

To further bound these quantities, it turns out to be convenient to use a relation with an *exact* conjugate gradient process applied to the $(k+1)\times(k+1)$-matrix A and right-hand side b that generates the same matrix S_k (and therefore the same coefficients as in the CGLS process). We define

$$\underline{T}_k \equiv \Psi_{k+1}\underline{S}_k\Psi_k^{-1} \quad \text{with } \Psi_k \equiv \mathrm{diag}(\psi_0,\ldots,\psi_{k-1}) \text{ and } \psi_j \equiv \left(\psi_0\prod_{i=0}^{j-1}\beta_i\right)^{1/2}. \tag{27}$$

The elements of the diagonal scaling Ψ_k are chosen such that the $k\times k$ upper block of $\underline{T}_k$ is symmetric. The symmetric matrix A is now defined by taking its left part equal to $\underline{T}_k$. The $(k+1,k+1)$-element is arbitrary but we will assume that the matrix A is positive definite which requires that the computed α-coefficients are all positive. The vector b equals ψ_0e_1. It is interesting to notice that for the matrix A more advanced completions could have been chosen (A is in this case of higher dimension). For example, Greenbaum [13] shows that the matrix A can be completed such that the eigenvalues of A are in tiny clusters around the eigenvalues of $\mathbf{A}$. This gives important insight into the finite precision convergence behavior of the conjugate gradient method. For our purposes the precise completion is not of importance and we work with a simple $k+1$-dimensional matrix.

It is not difficult to see that the *exact* conjugate gradient method applied to the matrix A with right-hand side b, indeed, yields our computed matrix S_k. The vectors of this hypothetical conjugate gradient process will be denoted with non-bold characters, in contrast to

the bold characters denoting the computed vectors in the CGLS-Lanczos method. So, for example, $r_j = e_{j+1}\psi_j$ and x_j equals $T_j^{-1}e_1\psi_0$ appended with $k+1-j$ additional zeros, where T_j is the upper $j \times j$ block of A. The following lemma gives some useful relations.

Lemma 5.1 *We have that*

$$\|r_j\|^2 = \psi_j^2 = \|\mathbf{r}_j\|^2(1 + (2n + j)\varepsilon') \quad \textit{with} \quad |\varepsilon'| \leq \varepsilon + \mathcal{O}(\varepsilon^2), \tag{28}$$

and

$$\|\mathbf{p}_j\| \leq \sqrt{j+1}\|p_j\| + \mathcal{O}(\varepsilon). \tag{29}$$

Proof. Let ε_i and ε' denote variables such that $|\varepsilon_i|, |\varepsilon'| \leq \varepsilon + \mathcal{O}(\varepsilon^2)$ (where it precise value varies). We have, using the standard model of floating point arithmetic (17),

$$\phi_i = \|\mathbf{r}_i\|^2(1 + 2n\varepsilon') \quad \text{and} \quad \beta_i = \frac{\phi_{i+1}}{\phi_i}(1 + \varepsilon').$$

From this it follows that

$$\|r_j\|^2 = \psi_j^2 = \phi_0 \prod_{i=0}^{j-1} \beta_i = \phi_0 \prod_{i=0}^{j-1} \frac{\phi_{i+1}}{\phi_i}(1 + \varepsilon_i) = \phi_j(1 + j\varepsilon') = \|\mathbf{r}_j\|^2(1 + (2n + j)\varepsilon').$$

For the second inequality we use that

$$\|\mathbf{p}_j\| \leq \|\mathbf{r}_j\| + \beta_j\|\mathbf{p}_{j-1}\| + \mathcal{O}(\varepsilon) \leq \left(1 + \frac{\|\mathbf{r}_j\|}{\|\mathbf{r}_{j-1}\|}\frac{\|\mathbf{p}_{j-1}\|}{\|\mathbf{r}_{j-1}\|}\right)\|\mathbf{r}_j\| + \mathcal{O}(\varepsilon).$$

Recursive application of this estimate leads to

$$\|\mathbf{p}_j\| \leq \|\mathbf{r}_j\|^2 \sum_{i=0}^{j} \frac{1}{\|\mathbf{r}_i\|} + \mathcal{O}(\varepsilon) \leq \sqrt{j+1}\|\mathbf{r}_j\|^2 \left(\sum_{i=0}^{j} \frac{1}{\|\mathbf{r}_i\|^2}\right)^{1/2} + \mathcal{O}(\varepsilon).$$

Plugging the expression (28) into this expression we find (29). □

Now we continue with our original goal, which is to bound $\|\mathbf{w}_k^{(1)}\|$. We have that

$$\|\alpha_j\mathbf{p}_j\| \leq \sqrt{j+1}\|\alpha_j p_j\| + \mathcal{O}(\varepsilon) = \sqrt{j+1}\|x_{j+1} - x_j\| + \mathcal{O}(\varepsilon) \leq \sqrt{j+1}\|x_{j+1}\| + \mathcal{O}(\varepsilon).$$

where in the last inequality we have used (5). For the second quantity that appears in $\mathbf{w}_k^{(1)}$, Proposition 2.2 shows that

$$\|\mathbf{z}_j\| = \|\mathbf{b} - \mathbf{A}\mathbf{x}_j^{\text{CGLS}}\| + \|(\mathbf{b} - \mathbf{A}\mathbf{x}_j^{\text{CGLS}}) - \mathbf{z}_j\| \leq \|\mathbf{b}\| + \|\mathbf{A}\|\|\mathbf{x}_j^{\text{CGLS}}\| + \mathcal{O}(\varepsilon). \tag{30}$$

Furthermore, by combining Lemma 5.1 and (23), it is not difficult to see that

$$\|\mathbf{x}_j^{\text{CGLS}}\| = \|\mathbf{R}_j S_j^{-1} e_1\| + \mathcal{O}(\varepsilon) \leq \sqrt{j+1}\|x_j\| + \mathcal{O}(\varepsilon).$$

All together we get the following upper bound:

$$\|\mathbf{w}_k^{(1)}\| \leq \varepsilon\left(k^{3/2}(c+5)\|\mathbf{A}\|\|x_k\| + 5k\|\mathbf{b}\|\right) + \mathcal{O}(\varepsilon^2). \tag{31}$$

In order to estimate the size of the vector $\mathbf{w}_k^{(3)}$, the problem is reduced to bounding:

$$\left(3\|\mathbf{r}_j\| + 2\|\mathbf{p}_j\|\right)|e_{j+1}^{\mathrm{T}} S_k^{-1} e_1| + \mathcal{O}(\varepsilon) \leq \sqrt{j+1}\left(3\|r_j\| + 2\|p_j\|\right)|e_{j+1}^{\mathrm{T}} S_k^{-1} e_1| + \mathcal{O}(\varepsilon). \tag{32}$$

For this purpose, the following lemma is of use.

Lemma 5.2 *Let $j < k$. We have that*

$$\|r_j\|\,|e_{j+1}^T S_k^{-1} e_1| \le \|x_k\| \quad \textit{and} \quad \|p_j\|\,|e_{j+1}^T S_k^{-1} e_1| \le \|x_k\|. \tag{33}$$

Proof. First observe that $\|r_j\||e_{j+1}^{\mathrm{T}} S_k^{-1} e_1| = |e_{j+1}^{\mathrm{T}} x_k|$ and the first inequality follows.

From [17, Theorem 5.3] we know that $\|p_j\|/\|r_j\| = \|r_j\|/\rho_j$ with $\rho_j \equiv (\sum_{i=0}^{j} \|r_i\|^{-2})^{-1/2}$. The value ρ_j is essentially the norm of the *minimal residual* approximation corresponding to the approximation x_j^{MR} with only components in the first j elements of its vector, thus $\rho_j = \|e_1\psi_0 - Ax_j^{\mathrm{MR}}\|$. Now, write

$$\begin{aligned} \|r_j\|^2 |e_{j+1}^{\mathrm{T}} S_k^{-1} e_1| &= \|r_j\|\,|e_{j+1}^{\mathrm{T}} x_k| = |r_j^{\mathrm{T}} x_k| = |r_j^{\mathrm{T}}(x_k - x_j^{\mathrm{MR}})| \\ &= |(x_k - x_j)^{\mathrm{T}} T_k (x_k - x_j^{\mathrm{MR}})| \le \|x_k - x_j\|\rho_j \le \|x_k\|\rho_j \end{aligned}$$

and the second inequality in (33) follows. Here, we used that $r_j \perp x_j^{\mathrm{MR}}$ and the second inequality in (5). □

Notice that the obvious estimates $|e_{j+1}^{\mathrm{T}} x_k| \le \|x_k\|$ and $\|r_j\|^2 \le \|T_k\|\|T_k^{-1}\|\,\rho_j^2$ could have been used in the proof of the lemma. This would have given the crude bound

$$(\|r_j\|/\rho_j)\,|e_{j+1}^{\mathrm{T}} x_k| \le (\|T_k\|\|T_k^{-1}\|)^{1/2}\|x_k\|,$$

which is very similar to the estimate (9.13) in [25] for the CG method in finite precision computations. Since it contains the term $\|T_k^{-1}\|^{1/2}$, it would not have been sufficient for our purposes. In the proof we used the fact that a large value of $\|r_j\|$ is canceled against a smaller $j+1$-th elements in the vector x_k.

Combining these expressions for all $j < k$, we finally find

$$\|\mathbf{w}_k^{(3)}\| \le \varepsilon k^{3/2} 5\|x_k\| + \mathcal{O}(\varepsilon^2). \tag{34}$$

5.2 The attainable accuracy of the multishift CGLS method

Using our bounds for the $\mathbf{w}$-vectors and the expression for the true residual (24), we are now in a position to give some discussion on the ultimately attainable accuracy of the multishift CGLS method based on the CGLS-Lanczos method. Recall that we neglect, for the moment, rounding errors made in the computation of the $\mathbf{x}_k^\sigma$.

The condition number for the least squares problem, e.g., [2, Section 1.4.3], can be written as

$$\kappa_{LS}(\mathbf{A}, \mathbf{b}) \equiv \kappa(\mathbf{A}) \left(1 + \kappa(\mathbf{A}) \frac{\|\mathbf{z}\|}{\|\mathbf{A}\|\|\mathbf{x}\|}\right),$$

where $\kappa(\mathbf{A})$ is defined as $\|\mathbf{A}\|\|\mathbf{A}^\dagger\|$ and $\mathbf{A}^\dagger$ denotes the pseudo-inverse of the matrix $\mathbf{A}$ given by $\mathbf{A}^\dagger \equiv (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$. This means that if $\widetilde{\mathbf{x}}$ is computed using a (normwise) backward stable method, the forward error is bounded by

$$\|\widetilde{\mathbf{x}} - \mathbf{x}\| \lesssim \varepsilon\kappa(\mathbf{A})\|\mathbf{x}\| + \varepsilon\kappa(\mathbf{A})^2 \frac{\|\mathbf{z}\|}{\|\mathbf{A}\|}. \tag{35}$$

This provides us with some idea what we can expect.

In case of $\sigma = 0$, we have for the first term in (26) that

$$\|\mathbf{A}^\dagger \mathbf{z}_k\| = \|\mathbf{A}^\dagger(\mathbf{z}_k - \mathbf{z})\| \le \varepsilon d\kappa(\mathbf{A})\|\mathbf{x}\|.$$

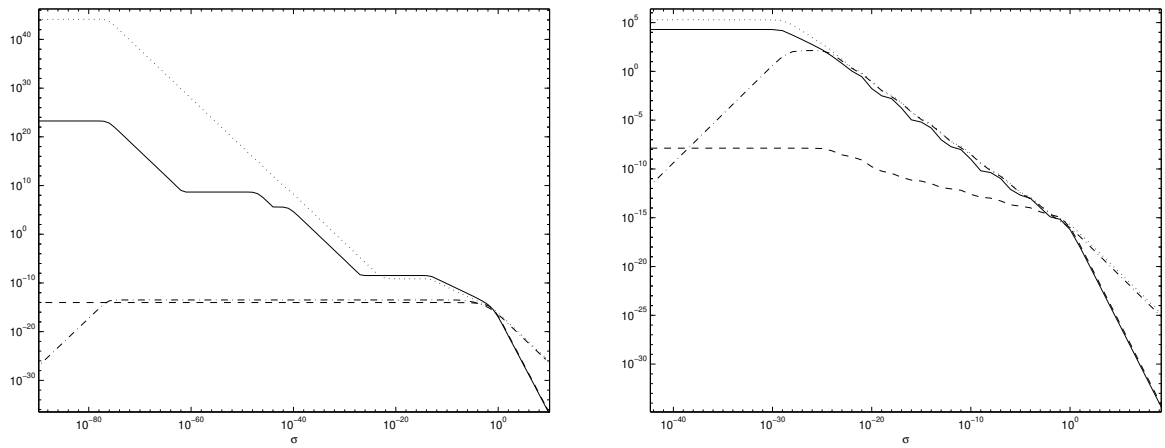

FIGURE 1. *Upper bounds, as a function of* σ*, on the quantities* $\|(\mathbf{A}^T\mathbf{A}+\sigma\mathbf{I})^{-1}\mathbf{A}^T\|\,\|\mathbf{w}_k^{(1)}\|$ *(solid),* $\|(\mathbf{A}^T\mathbf{A}+\sigma\mathbf{I})^{-1}\|\,\|\mathbf{w}_k^{(2)}\|$ *(dash-dot),* $\|(\mathbf{A}^T\mathbf{A}+\sigma\mathbf{I})^{-1}\mathbf{A}^T\mathbf{A}\|\,\|\mathbf{w}_k^{(3)}\|$ *(dashed), and the expression on the right in (36) (dotted). Left picture: HEAT(100). Right picture: URSELL(100).*

To bound the contribution of $\mathbf{w}_k^{(1)}$ and $\mathbf{w}_k^{(3)}$ to the forward error we define

$$\Theta_k \equiv \|x_k\|/\|\mathbf{x}\|.$$

Although in *exact* arithmetic this expression does not exceed one, its finite precision value is unknown. In the following we assume that this quantity is modest. We stress, however, that the quantity $\|\mathbf{x}_j\|/\|\mathbf{x}\|$, as in Proposition 2.1, is argued to be bounded in finite precision computations using similar assumptions.

Using (31), (34) and the fact that $\|\mathbf{b}\| \leq \|\mathbf{z}\| + \|\mathbf{A}\|\,\|\mathbf{x}\|$, we finally arrive at the following result.

Lemma 5.3 *Assume that (25) holds and let* $\sigma = 0$*. For the accuracy of the multishift CGLS approximation, we have*

$$\|\mathbf{x}-\mathbf{x}_k\| \leq \varepsilon\left(\kappa(\mathbf{A})(d+5k+k^{3/2}(c+5)\Theta_k)+5k^{3/2}\Theta_k\right)\|\mathbf{x}\| + \varepsilon 5k\kappa(\mathbf{A})\frac{\|\mathbf{z}\|}{\|\mathbf{A}\|} + \mathcal{O}(\varepsilon^2),$$

and for the least squares residual:

$$\|(\mathbf{b}-\mathbf{A}\mathbf{x}_k)-\mathbf{z}\| \leq \varepsilon(d+5k+k^{3/2}(c+10)\Theta_k)\|\mathbf{A}\|\|\mathbf{x}\| + 5\varepsilon k\|\mathbf{z}\| + \mathcal{O}(\varepsilon^2).$$

Notice that the forward error appears to be smaller than the upper bound on the error of a normwise backward stable method given in (35). However, as remarked in [3], the number of iterations k to reach this state might also depend on the conditioning of the matrix $\mathbf{A}$.

With a direct application of the CGLS method, it follows from (23) that $\mathbf{x}_k^{\text{CGLS}} = \mathbf{x}_k + \mathbf{w}_k^{(3)}$. Therefore, the difference in attainable accuracy between the multishift CGLS method and a direct application of the CGLS method is determined by the vector $\mathbf{w}_k^{(3)}$. Our analysis clearly shows that this term is not expected to form an essential difference. In addition, it shows that the residual gap for the CGLS method is not influenced by the third $\mathbf{w}$-vector.

In the preceding, we restricted our attention to the zero shift case. The generalization to the case of $\sigma > 0$ is not straightforward. We want to provide some, preliminary, insight

into using the CGLS-Lanczos method for the Lanczos part of the multishift CGLS method. Perturbation analysis for the Tikhonov regularized problem can be found in [19, 15]. For example, a straightforward argument in [19, Section 3] shows that, if $\widetilde{\mathbf{x}}$ is computed with the matrix $\mathbf{A}$ and right-hand side $\mathbf{b}$, with perturbations in the order of $\varepsilon\|\mathbf{A}\|$ and $\varepsilon\|\mathbf{b}\|$, respectively, then the forward error of this approximate solution is bounded by

$$\|\widetilde{\mathbf{x}} - \mathbf{x}^\sigma\| \lesssim \varepsilon\|(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\|(\|\mathbf{b}\| + \|\mathbf{A}\|\|\mathbf{x}^\sigma\|) + \varepsilon\|(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}\|\|\mathbf{b} - \mathbf{A}\mathbf{x}^\sigma\|. \quad (36)$$

Although, sharper estimates are possible, this estimate already provides us with some idea of the upper bound on the error that can be expected.

If $\sigma \to \infty$, then we have:

$$(\gamma_k^\sigma)^{-1} \to 0, \quad \mathbf{w}_k^{(1)} \to 0, \quad \mathbf{w}_k^{(2)} \to -\mathbf{F}_k^r e_1, \quad \mathbf{w}_k^{(3)} \to 0 \qquad (\sigma \to \infty),$$

which yields

$$\|\mathbf{x}_k^\sigma - \mathbf{x}^\sigma\| \sim \sigma^{-1}\varepsilon\|(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \sigma\mathbf{I})^{-1}\mathbf{F}_k^r e_1\| \lesssim \sigma^{-1}\varepsilon c\|\mathbf{A}\|\|\mathbf{b}\| \qquad (\sigma \to \infty).$$

Notice that only a perturbation of the right-hand side $\mathbf{b}$ of size $\varepsilon\|\mathbf{b}\|$ leads to a forward error of this size and, therefore, the error of the multishift CGLS method is as small as might be expected.

In Figure 1 we have plotted upper bounds on the quantities in (26) for two very ill-conditioned test problems from [16] for various values of σ. For this purpose, we ran the CGLS-Lanczos method for 100 iterations and computed upper bounds on the perturbation terms in Lemma 4.1 using the estimates given in (19)-(22) evaluated using the computed quantities in the CGLS-Lanczos method. Furthermore, $c = c' = 1$ and $\varepsilon \approx 10^{-16}$. Moreover, as a reference we have included the expression on the right in (36). This picture shows the behavior that we observed for all our test problems: for small σ the error is dominated by the contribution of $\mathbf{w}_k^{(1)}$ and for very large values of σ, the contribution of $\mathbf{w}_k^{(2)}$ becomes dominant. The upper bound on the contribution of $\mathbf{w}_k^{(2)}$ increases for increasing σ until in reaches it maximum value after which it appears to decay with a speed proportional to σ^{-1}. Unfortunately, we do not have precise expressions that bound these quantities for general σ from above.

6 The solution of the shifted systems

The second main component of the multishift CGLS method is the computation of the approximate solutions $\mathbf{x}_k^\sigma$ as

$$\mathbf{x}_k^\sigma = \mathbf{R}_k(S_k + \sigma I)^{-1}e_1, \quad (37)$$

or using the equivalent formulation given by (14).

We start this section by summarizing an approach that is used at several places in literature in multishift versions of the (Bi-)CG method based on coupled two-term recurrences, which also requires the computation (37), see [18, 10]. For more details consult these references.

An important observation is that the residuals for shifted systems are *colinear* for the CG method, that is, there exist constants γ_j^σ such that $\mathbf{r}_j^\sigma = \mathbf{r}_j/\gamma_j^\sigma$. Writing out the three-term recurrence of the residuals $\mathbf{r}_k^\sigma$ (similar to (15)) and comparing terms reveals, with $\gamma_{-1}^\sigma = \gamma_0^\sigma = 1$, the three-term relation

$$\gamma_j^\sigma = (1 + \alpha_{j-1}\sigma)\gamma_{j-1}^\sigma + \frac{\alpha_{j-1}}{\alpha_{j-2}}\beta_{j-2}(\gamma_{j-1}^\sigma - \gamma_{j-2}^\sigma), \quad (38)$$

and a recurrences for the iterates and search directions

$$\mathbf{x}_j^\sigma = \mathbf{x}_{j-1}^\sigma + \alpha_{j-1}\left(\frac{\gamma_{j-1}^\sigma}{\gamma_j^\sigma}\right)\mathbf{p}_{j-1}^\sigma, \quad \mathbf{p}_j^\sigma = \mathbf{r}_j/\gamma_j^\sigma + \beta_{j-1}\left(\frac{\gamma_{j-1}^\sigma}{\gamma_j^\sigma}\right)^2\mathbf{p}_{j-1}^\sigma,$$

with initially $\mathbf{p}_0^\sigma = \mathbf{r}_0$ and $\mathbf{x}_0^\sigma = \mathbf{0}$. Scaling $\mathbf{p}_j^\sigma$ by γ_j^σ leads to a version with the same stability properties:

$$\mathbf{x}_j^\sigma = \mathbf{x}_{j-1}^\sigma + \frac{\alpha_{j-1}}{\gamma_j^\sigma}\widetilde{\mathbf{p}}_{j-1}^\sigma, \quad \widetilde{\mathbf{p}}_j^\sigma = \mathbf{r}_j + \beta_{j-1}\frac{\gamma_{j-1}^\sigma}{\gamma_j^\sigma}\widetilde{\mathbf{p}}_{j-1}^\sigma. \tag{39}$$

If a stable method is applied for computing the inverse in (37), then this leads to an approximate solution that is usually sufficiently accurate when dealing with ordinary linear systems. However, this might not be the case for normal equations, since then a dependence of the attainable precision on the square of the condition number of $\mathbf{A}$ may have been introduced. Therefore, a point of concern of the approach (38)-(39) is that it implicitly forms the ill-conditioned tridiagonal matrix S_k (cf., (15)) in the computation of the γ_j^σ in (38). An example of the failure of this algorithm in the multishift CGLS context is given at the end of this section (§6.2).

6.1 An alternative implementation

The goal is to give an alternative algorithm which prevents the formation of the ill-conditioned matrix S_k. In case the CGLS-Lanczos method is used for the Lanczos part, then it is clear from (16), that also the $L^{\mathrm{T}}DL$ factorization of T_k is directly available. The *quotient-difference* algorithms introduced by Rutishauser [22], provide a means to construct an $L^{\mathrm{T}}DL$ factorization of the shifted matrix $T_k + \sigma I$ directly from the factors of T_k. These algorithms construct the factors, D_k^σ and L_k^σ, in a step-by-step fashion such that

$$L_k^{\mathrm{T}}\Delta_k^{-1}L_k + \sigma I = (L_k^\sigma)^{\mathrm{T}}D_k^\sigma L_k^\sigma,$$

where D_k^σ is diagonal with diagonal elements $d_0^\sigma, \ldots, d_{k-1}^\sigma$ and L_k^σ is upper bidiagonal with diagonal elements one and upper diagonal elements $l_0^\sigma, \ldots, l_{k-2}^\sigma$. If we take the *differential form* of the *stationary qd* transformation (dstqds) as presented in [5, Algorithm 4.2], then we have, with $t_0^\sigma = \sigma$, the following recurrence relations for computing the elements of the factors D_k^σ and L_k^σ:

$$d_{j-1}^\sigma = t_{j-1}^\sigma + \alpha_{j-1}^{-1}, \quad l_{j-1}^\sigma = -\frac{\sqrt{\beta_{j-1}}}{\alpha_{j-1}d_{j-1}^\sigma}, \quad t_j^\sigma = \sigma - l_{j-1}^\sigma\sqrt{\beta_{j-1}}t_{j-1}^\sigma.$$

Just as for the conjugate gradient method, the construction of the vector $\mathbf{x}_k^\sigma$ can be accomplished efficiently by introducing the auxiliary vectors $\mathbf{p}_j^\sigma$ defined by the relation $\mathbf{P}_k^\sigma = \mathbf{V}_k(L_k^\sigma)^{-1}$. Starting with $\xi_0 = 1$, $\mathbf{p}_0^\sigma = \mathbf{r}_0/\sqrt{\phi_0}$, $\mathbf{x}_0^\sigma = \mathbf{0}$, this gives

$$\mathbf{x}_j^\sigma = \mathbf{x}_{j-1}^\sigma + \frac{\xi_{j-1}}{d_{j-1}^\sigma}\mathbf{p}_{j-1}^\sigma, \quad \mathbf{p}_j^\sigma = \mathbf{r}_j/\sqrt{\phi_j} - l_{j-1}^\sigma\mathbf{p}_{j-1}^\sigma, \quad \xi_j = -\xi_{j-1}l_{j-1}^\sigma.$$

A simple scaling of the $\mathbf{p}_j^\sigma$ by $\sqrt{\phi_j}$ and denoting $\ell_j^\sigma \equiv \alpha_j d_j^\sigma$ and $\gamma_j^\sigma \equiv \sqrt{\phi_j}/\xi_j$ leads to a slightly more efficient, but equally stable, variant

$$\ell_{j-1}^\sigma = 1 + \alpha_{j-1}t_{j-1}^\sigma, \quad t_j^\sigma = \sigma + \frac{\beta_{j-1}}{\ell_{j-1}^\sigma}t_{j-1}^\sigma, \quad \gamma_j^\sigma = \gamma_{j-1}^\sigma\ell_{j-1}^\sigma. \tag{40}$$

$\mathbf{z}_0 = \mathbf{b}$, $\mathbf{r}_0 = \mathbf{A}^{\mathrm{T}}\mathbf{z}_0$, $\mathbf{p}_0 = \mathbf{r}_0$, $\phi_0 = \|\mathbf{r}_0\|^2$
$\mathbf{x}_0^\sigma = \mathbf{0}$, $t_0^\sigma = \sigma$
for $j = 1, \ldots k$

The Lanczos part
$\mathbf{c}_{j-1} = \mathbf{A}\mathbf{p}_{j-1}$
$\alpha_{j-1} = \phi_{j-1}/\|\mathbf{c}_{j-1}\|^2$
$\mathbf{z}_j = \mathbf{z}_{j-1} - \alpha_{j-1}\mathbf{c}_{j-1}$
$\mathbf{r}_j = \mathbf{A}^{\mathrm{T}}\mathbf{z}_j$
$\phi_j = \|\mathbf{r}_j\|^2$, $\beta_{j-1} = \phi_j/\phi_{j-1}$
$\mathbf{p}_j = \mathbf{r}_j + \beta_{j-1}\mathbf{p}_{j-1}$

The inversion part
$\ell_{j-1}^\sigma = 1 + \alpha_{j-1}t_{j-1}^\sigma$, $t_j^\sigma = \sigma + (\beta_{j-1}/\ell_{j-1}^\sigma)t_{j-1}^\sigma$, $\gamma_j^\sigma = \gamma_{j-1}^\sigma \ell_{j-1}^\sigma$
$\mathbf{x}_j^\sigma = \mathbf{x}_{j-1}^\sigma + (\alpha_{j-1}/\gamma_j^\sigma)\widetilde{\mathbf{p}}_{j-1}^\sigma$
$\widetilde{\mathbf{p}}_j^\sigma = \mathbf{r}_j + (\beta_{j-1}/\ell_{j-1}^\sigma)\widetilde{\mathbf{p}}_{j-1}^\sigma$

Algorithm 2. *Multishift CGLS implementation for solving families of the form (2).*

$$\mathbf{x}_j^\sigma = \mathbf{x}_{j-1}^\sigma + \frac{\alpha_{j-1}}{\gamma_j^\sigma}\,\widetilde{\mathbf{p}}_{j-1}^\sigma, \quad \widetilde{\mathbf{p}}_j^\sigma = \mathbf{r}_j + \beta_{j-1}\frac{1}{\ell_{j-1}^\sigma}\widetilde{\mathbf{p}}_{j-1}^\sigma. \tag{41}$$

We have discussed the two components of our implementation of the multishift CGLS method which consists of combining (40) and (41) with the CGLS recurrences in (8). The resulting algorithm is summarized in Algorithm 2. Comparing (41) and (39) reveals that the search directions $\widetilde{\mathbf{p}}_j^\sigma$ and the scalars γ_j^σ coincide (in exact arithmetic). So, the recursions (40) can be seen as a reformulation of the equivalent recursion given in (38) which is hopefully more stable. Therefore, the norm of the residual of the shifted system (for the normal equations) is equal to $\|\mathbf{r}_j\|/|\gamma_j^\sigma| = \sqrt{\phi_j}/|\gamma_j^\sigma|$.

In Section 5.2, we addressed the influence of rounding errors in the Lanczos part on the attainable accuracy of the multishift CGLS method based on the CGLS recurrences. We now discuss the second main source of errors in the multishift method which is the difference between $\mathbf{x}_k^\sigma$, computed using (40) and (41), and the exact expression $\mathbf{R}_k(S_k + \sigma I)^{-1}e_1$. In [5, Section 4.3], Dhillon and Parlett present a roundoff error analysis of the dstqds algorithm which shows that the outcome of this algorithm is relatively close to the outcome of an exact transformation applied to factors relatively close to the original input. This is also expected to hold for our scaled recursions (40). Alternatively, this stability can also be understood by a close inspection of the recursions in (40): only elements are added with the same sign and therefore the computed value of γ_j^σ is relatively close to the exact value (since there is no cancellation). For this reason, we expect that the influence of rounding errors in our recursion for the γ_j^σ is small. On the other hand, the three-term recurrence in (38) requires the computation of $(\gamma_{j-1}^\sigma - \gamma_{j-2}^\sigma)$ which may be sensitive to roundoff errors due to cancellation in case the two quantities are relatively close.

Another issue is the impact of roundoff errors in the vector updates (41). The impact of

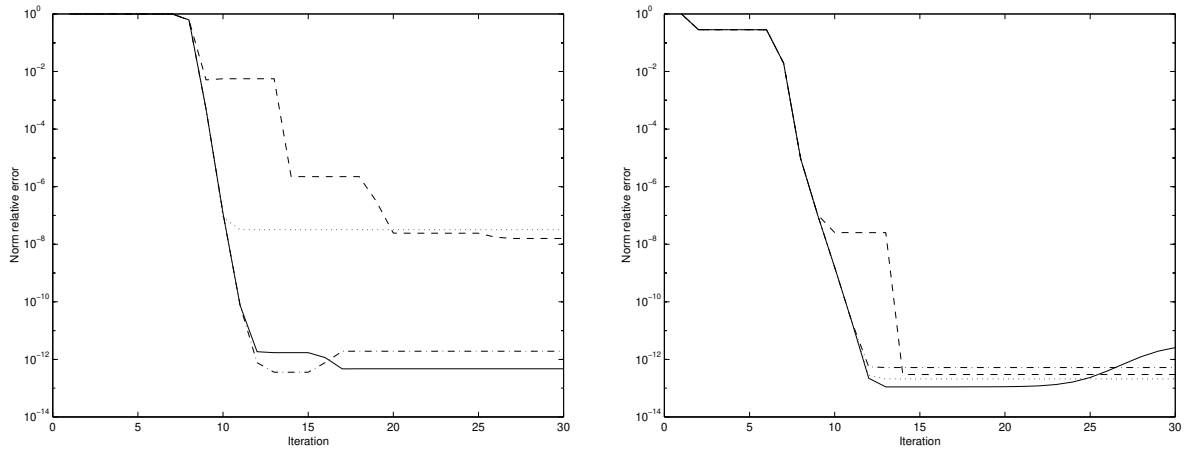

FIGURE 2. *Relative error as function of k for Alg. 1 (solid), Algorithm 6 from [8] (dashed), multishift CGLS based on CGLS-Lanczos with (41) (dash-dot) and (39) (dotted) for two different shifts: $\sigma = 10^{-8}$ (left) and $\sigma = 1$ (right).*

these errors can be analyzed in detail using the techniques in Section 5.1. If we define x_k^σ as the iterate computed in an *exact* conjugate gradient method applied to the matrix $A + \sigma I$ with starting vector b, as defined in Section 5.1, then we can show, assuming that no roundoff errors are made in (40), that

$$\|\mathbf{x}_k^\sigma - \mathbf{R}_k(S_k + \sigma I)^{-1} e_1\| \leq \varepsilon 10 k^{3/2} \|x_k^\sigma\|.$$

Here, we have used, essentially, the same technique as used for bounding the vector $\mathbf{w}_k^{(3)}$ in the case that $\sigma = 0$. Hence, we do not expect that rounding errors in the computation of the vector $\mathbf{x}_k^\sigma$ have a significant influence on the attainable accuracy of the multishift CGLS method.

6.2 A numerical comparison

Numerical experiments suggest that a multishift CGLS method (based on the CGLS-Lanczos method) combined with (38)-(39) for the inversion part, is often remarkably accurate and, in most situations, as accurate as with our alternative given in Alg. 2. Nevertheless, there are examples where differences are clear. We show this for two simple systems. The matrix $\mathbf{A}$ has eigenvalues $\{1/250, 240, 241, \ldots, 250\}$ ($n = 12$), the orthonormal eigenvector basis is random and the right-hand side has equal components in all eigenvector directions. The results for solving the system (2) are presented in Figure 2 for $\sigma = 10^{-8}$ and $\sigma = 1$. In this picture we have also presented the results for Alg. 1 as a reference to the other methods. In this case coupled recurrences (40) clearly give more accurate results than the three-term recurrence (38) for $\sigma = 10^{-8}$. For larger values of σ, and very small values, the differences become small.

The implementation of the multishift CGLS method presented in [8, Alg. 6] uses a variant of the standard Lanczos method for the Lanczos part. This implementation is based on a three-term recurrence for the least squares residuals and results in a tridiagonal matrix in standard form. So, even if these changes in the Lanczos part are improving the attainable accuracy of the method, the accuracy is expected to be limited by the inversion of the tridiagonal since the ill-conditioned tridiagonal is not given in factorized form. The dashed lines in Figure 2 confirm this.

7 Numerical experiments

In this section, we compare the attainable accuracy of the CGLS method for damped least squares problems, Alg. 1, to the attainable accuracy of our version of the multishift CGLS method, Alg. 2. The 'exact' solution, $\mathbf{x}^\sigma$, was computed using a singular value decomposition of the matrix $\mathbf{A}$ and we report the relative error given by

$$\|\mathbf{x}_k^\sigma - \mathbf{x}^\sigma\| / \|\mathbf{x}^\sigma\|,$$

where $\mathbf{x}_k^\sigma$ is the computed approximation with either method. The number of iterations, k, was chosen such that the error for the particular method was minimal. The results for various test problems from [16] are given in Table 1.

σ	10^{-8}	10^{-4}	1	10^4
		HEAT(100)		
Alg. 1	4.9(-13)	5.1(-15)	6.6(-16)	6.5(-16)
Alg. 2	4.8(-13)	5.2(-15)	6.1(-16)	7.0(-16)
		URSELL(100)		
Alg. 1	6.7(-14)	2.9(-15)	2.9(-16)	2.6(-16)
Alg. 2	8.7(-14)	3.3(-15)	2.5(-16)	2.7(-16)
		FOXGOOD(100)		
Alg. 1	2.2(-13)	3.0(-15)	3.7(-16)	6.7(-16)
Alg. 2	2.7(-13)	3.0(-15)	3.7(-16)	7.3(-16)
		ILAPLACE(100)		
Alg. 1	9.2(-13)	1.8(-14)	1.2(-15)	6.7(-16)
Alg. 2	8.4(-13)	1.8(-14)	1.3(-15)	6.0(-16)

TABLE 1. *Attained relative errors for various problems and various choices for σ.*

The results in this table confirm that the proposed implementation of the multishift CGLS method achieves a comparable accuracy to applying the CGLS method directly to the regularized system. However, there are a few interesting differences between both methods that occur now and then and which are not apparent from this table. One property of Alg. 1 is that, for large shifts, the method appears to have the tendency to diverge after reaching its maximal precision. An interesting observation is that the multishift version of CGLS does not have this behavior. This is illustrated in the left picture in Figure 3. The computational costs per step are much lower for the multishift version of CGLS (no matrix-vector multiplication, no inner products, less vector updates for solving the shifted problems) than the direct application of the CGLS method. However, it is remarkable that, in addition, the multishift version sometimes needs less iteration steps. An example of this is given in the right picture in Figure 3.

8 Summary and outlook

In this paper we have proposed a new implementation of the multishift CGLS method for solving families of damped least squares problems. This work was motivated by the observation that, in some cases, previous proposals can only attain a precision that depends on the

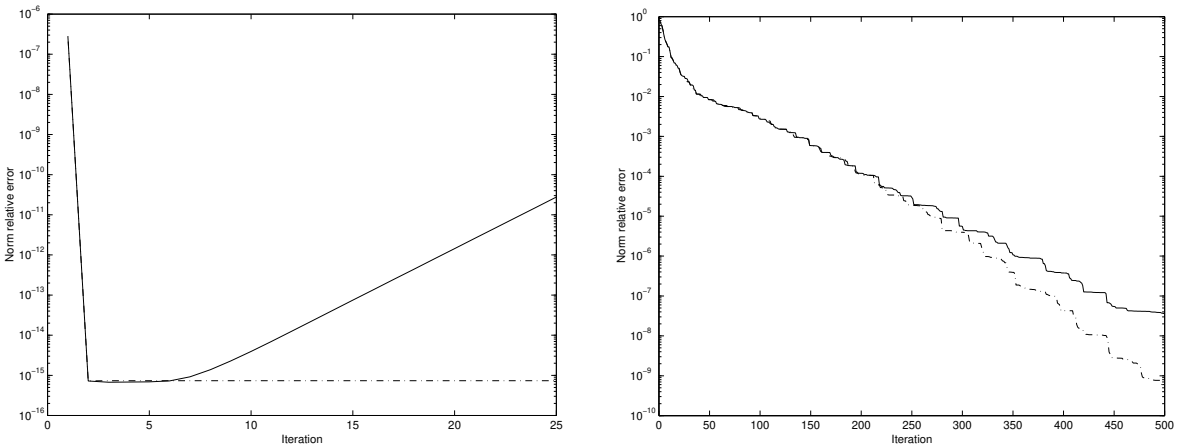

FIGURE 3. *The relative error (as a function of k) of Alg. 1 (solid) and Alg. 2 (dash-dot). Left: problem FOXGOOD(100) with* $\sigma = 10^4$. *Right: HEAT(100) with* $\sigma = 10^{-8}$.

square of the condition number. The first key ingredient of our implementation is the use of coupled recurrences for the construction of an orthogonal basis for the Krylov subspace. We showed that this leads to a perturbed Lanczos-type relation with a perturbation that has a desirable structure. The size of this perturbation can be relatively large compared to the perturbation term of the standard Lanczos process. Nevertheless, using an extensive rounding error analysis for $\sigma = 0$, which exploits the special structure of the perturbation term, we showed that this alternative Lanczos method is desirable in multishift CGLS methods. A second advantage of the use of CGLS recurrences for the Lanczos part, is the fact that the tridiagonal matrix is available in a factorized form. This is important information for the second key ingredient: the construction of the iterates for the shifted systems. We showed, by exploiting the factorized form of the tridiagonal using the stationary qd transformation, that the iterates can be accurately and efficiently constructed.

In future work, we plan to extend our analysis for the Lanczos part to the situation of more general σ. This should theoretically confirm also for more general σ the suitability of the CGLS-Lanczos method for the Lanczos part. Another interesting extension of this paper is to consider the implementation of the multishift CGLS method based on *Lanczos bidiagonalization* e.g., [12, Section 9.3.3].

Furthermore, we believe that our analysis provides the key ingredients for analyzing the advantages of coupled two-term recurrence Lanczos in the QMR method as experimentally shown in [7]. In the QMR method there is also a clear separation of the Lanczos and inversion part, as for the multishift CGLS method. It could also help to analyze the advantages of alternative Lanczos methods in other applications, e.g., [1].

Acknowledgments

We are thankful to the two referees. Their remarks helped us improve the presentation of this paper.

References

[1] Zhaojun Bai and Roland W. Freund, *A symmetric band Lanczos process based on coupled recurrences and some applications*, SIAM J. Sci. Comput. **23** (2001), no. 2, 542–562 (electronic), Copper Mountain Conference (2000). MR 1 861 264

[2] Åke Björck, *Numerical methods for least squares problems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. MR 97g:65004

[3] Åke Björck, Tommy Elfving, and Zdeněk Strakoš, *Stability of conjugate gradient and Lanczos methods for linear least squares problems*, SIAM J. Matrix Anal. Appl. **19** (1998), no. 3, 720–736 (electronic). MR 99a:65051

[4] Biswa Nath Datta and Youcef Saad, *Arnoldi methods for large Sylvester-like observer matrix equations, and an associated algorithm for partial spectrum assignment*, Linear Algebra Appl. **154/156** (1991), 225–244. MR 92b:65032

[5] Inderjit Dhillon and Beresford Parlett, *Orthogonal eigenvectors and relative gaps*, Submitted.

[6] Roland W. Freund, *Solution of shifted linear systems by quasi-minimal residual iterations*, Numerical linear algebra (Kent, OH, 1992), de Gruyter, Berlin, 1993, pp. 101–121. MR 1 244 155

[7] Roland W. Freund and Noël M. Nachtigal, *An implementation of the QMR method based on coupled two-term recurrences*, SIAM J. Sci. Comput. **15** (1994), no. 2, 313–337, Iterative methods in numerical linear algebra (Copper Mountain Resort, CO, 1992). MR 95f:65067

[8] A. Frommer and P. Maass, *Fast CG-based methods for Tikhonov-Phillips regularization*, SIAM J. Sc. Comput. **20** (1999), 1831–1850.

[9] A. Frommer, B. Nöckel, S. Güsken, Th. Lippert, and K. Schilling, *Many masses on one stroke: Economic computation of quark*, Int. J. Modern Physics **C 6** (1995), 627–638.

[10] Andreas Frommer, *BICGSTAB(l) for families of shifted linear systems*, Preprint BUGHW-SC 02/04, Bergische Universität GH Wuppertal, Wuppertal, Germany, November 2002.

[11] Andreas Frommer and Uwe Glässner, *Restarted GMRES for shifted linear systems*, SIAM J. Sci. Comput. **19** (1998), no. 1, 15–26 (electronic), Special issue on iterative methods (Copper Mountain, CO, 1996). MR 99b:65033

[12] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., The John Hopkins University Press, Baltimore, London, 1996.

[13] Anne Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl. **113** (1989), 7–63. MR 90e:65044

[14] ———, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl. **18** (1997), no. 3, 535–551. MR 98c:65048

[15] M. E. Gulliksson, P.-Å. Wedin, and Yimin Wei, *Perturbation identities for regularized Tikhonov inverses and weighted pseudoinverses*, BIT **40** (2000), no. 3, 513–523. MR MR2001g:65047

[16] Per Christian Hansen, *Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms **6** (1994), no. 1-2, 1–35. MR 94k:65062

[17] Magnus R. Hestenes and Eduard Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards **49** (1952), 409–436 (1953). MR 15:651a

[18] B. Jegerlehner, *Krylov space solvers for shifted linear systems*, HEP-LAT hep-lat/9612014, 1996.

[19] A. N. Malyshev, *A unified theory of conditioning for linear least squares and Tikhonov regularization solutions*, SIAM J. Matrix Anal. Appl. **24** (2003), no. 4, 1186–1196.

[20] H. Neuberger, *Overlap Dirac operator*, Numerical challenges in Lattice Quantum Chromodynamics (Berlin) (A. Frommer, Th. Lippert, B. Medeke, and K. Schilling, eds.), Springer-Verlag, 2000.

[21] C. C. Paige, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl. **18** (1976), no. 3, 341–349. MR 58:19082

[22] Heinz Rutishauser, *Der Quotienten-Differenzen-Algorithmus*, Mitt. Inst. Angew. Math. Zürich **1957** (1957), no. 7, 74. MR 19,686b

[23] V. Simoncini and F. Perotti, *On the numerical solution of* $(\lambda^2 A + \lambda B + C)x = b$ *and application to structural dynamics*, SIAM J. Sci. Comput. **23** (2002), no. 6, 1875–1897 (electronic). MR 1 923 717

[24] Valeria Simoncini, *Restarted full orthogonalization method for shifted linear systems*, Tech. report, 2002, To appear in BIT.

[25] Zdeněk Strakoš and Petr Tichý, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, ETNA **13** (2002), 56–80.