Analysis of the finite precision s-step biconjugate gradient method

Erin Carson
James Demmel

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-18
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-18.html

March 13, 2014
We analyze the s-step biconjugate gradient algorithm in finite precision arithmetic and derive a bound for the residual norm in terms of a minimum polynomial of a perturbed matrix multiplied by an amplification factor. Our bound enables comparison of s-step and classical biconjugate gradient in terms of amplification factors. Our results show that for s-step biconjugate gradient the amplification factor depends heavily on the quality of s-step polynomial bases generated in each outer loop.
Copyright © 2014, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Acknowledgement

Research partially funded by DARPA Award Number HR0011-12-2-0016, the Center for Future Architecture Research, a member of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA, and ASPIRE Lab industrial sponsors and affiliates Intel, Google, Nokia, NVIDIA, Oracle, and Samsung. Any opinions, findings, conclusions, or recommendations in this paper are solely those of the authors and does not necessarily reflect the position or the policy of the sponsors.
ANALYSIS OF THE FINITE PRECISION S-STEP BICONJUGATE GRADIENT METHOD

ERIN CARSON AND JAMES DEMMEL

Abstract. We analyze the s-step biconjugate gradient algorithm in finite precision arithmetic and derive a bound for the residual norm in terms of a minimum polynomial of a perturbed matrix multiplied by an amplification factor. Our bound enables comparison of s-step and classical biconjugate gradient in terms of amplification factors. Our results show that for s-step biconjugate gradient, the amplification factor depends heavily on the quality of s-step polynomial bases generated in each outer loop.

1. Introduction. Krylov subspace methods (KSMs) are a class of iterative algorithms commonly used to solve the linear system $Ax = b$. In classical KSM implementations, in iteration $n$, the updates to the solution $x_{n+1}$ and residual $r_{n+1}$ consist of one or more sparse matrix-vector multiplications (SpMVs) and vector operations in each iteration. On modern computer architectures, the performance of these operations is communication-bound; the movement of data, rather than the computation, is the limiting factor.

Communication-avoiding KSMs (CA-KSMs), based on s-step formulations ([4, 6, 8, 12, 27, 23, 24]), reduce the total communication cost by a factor of $O(s)$ by performing $O(s)$ computation steps per communication step (see, e.g., [3, 7, 13]). This asymptotic reduction in communication cost yields significant speedups in practice for many problems [19].

Although CA-KSMs are mathematically equivalent to their classical counterparts, their finite precision behavior may differ. It has been empirically observed that the rate of convergence of CA-KSMs deviates further from the convergence of the classical method as $s$ increases, and that the severity of this deviation is heavily influenced by the polynomials used for the s-step Krylov bases (see, e.g., [3, 13, 14, 1]).

In this work, we derive Lanczos-type matrix recurrences governing the s-step biconjugate gradient method (BICG) in finite precision arithmetic, which demonstrates the algorithm’s relationship to classical BICG. Using the recurrence, we extend the results of Tong and Ye for classical BICG [25] to derive an upper bound on the norm of the updated residual in finite precision s-step BICG. Our bound provides an analytical explanation for commonly-observed convergence behavior of s-step BICG.

2. Related work. We briefly outline the available literature on relevant topics, namely the analysis of KSMs in finite precision and s-step KSMs.

2.1. Analysis of finite precision Krylov methods. There are two primary effects of roundoff error in finite precision KSMs: the maximum attainable accuracy of the solution is decreased, and convergence may deteriorate. Much research has been devoted to better understanding this behavior, and to devise more robust and stable algorithms.

An upper bound on the maximum attainable accuracy for finite precision KSMs, limited by the deviation of the Lanczos residual from the true residual, was obtained by Greenbaum [10]. Greenbaum proved that this bound can be given a priori for methods like CG, but cannot be predetermined for methods like BICG, which can have arbitrarily large intermediate iterates. There are also techniques for alleviating this loss of accuracy, namely, residual replacement strategies, where the computed residual is replaced by the finite precision evaluation of the true residual at carefully
chosen iterations (see, e.g., [22, 26]. In this way, agreement between the true and computed residual is maintained to within a factor of $O(\epsilon)$.

In [11], Greenbaum proved backward stability of the finite precision CG algorithm, by showing that the computed Ritz values lie in small intervals around the eigenvalues of $A$. There are many other analyses of the behavior of various KSMs in finite precision arithmetic (see, e.g. [17, 16, 25]). The reader is also directed to the bibliography in [20].

Our analysis is most closely related to the work of Tong and Ye [25]. The authors derived a bound for the residual norm of classical BICG in finite precision, expressed as the product of a minimum polynomial of a perturbed matrix and an amplification factor. Our analysis generalizes the work of Tong and Ye to the $s$-step BICG case.

2.2. $s$-step Krylov subspace methods. The first instance of an $s$-step method in the literature is Van Rosendale’s $s$-step CG [27]. Van Rosendale’s implementation was motivated by exposing more parallelism using the PRAM model. Chronopoulos and Gear later created the $s$-step GMRES method with a similar goal [5]. Walker used $s$-step bases to improve stability in GMRES by replacing the modified Gram-Schmidt orthogonalization process with Householder QR [28]. These authors used monomial bases, and found that convergence often could not be guaranteed for $s > 5$. It was later discovered that this behavior was due to the inherent instability of the monomial basis, which motivated research into the use of other bases for the Krylov subspace.

Hindmarsh and Walker tried a scaled monomial basis to improve convergence [12], but saw only minimal improvement. Joubert and Carey implemented a scaled and shifted Chebyshev basis which led to more accurate results [14]. Bai et al. improved convergence using a Newton basis [1]. Although successively scaling the basis vectors can lower the condition number of the basis, this computation reintroduces communication dependencies. Hoemmen solved this using a novel matrix equilibration and balancing approach as a preprocessing step, which often alleviated the need for scaled basis vectors [13].

Hoemmen et al. [7, 13, 19] derived communication-avoiding variants of Lanczos, Arnoldi, CG and GMRES. We use communication-avoiding to specifically refer to $s$-step variants implemented using the communication-avoiding matrix powers kernel, which applies to well-partitioned sparse matrices (see, e.g., [18]). Derivations of communication-avoiding variants of nonsymmetric Lanczos-based KSMs, such as BICG, CGS, and BICGSTAB can be found in [3].

2.3. $s$-step BICG. We briefly review $s$-step BICG for solving $Ax = b$, where $A \in \mathbb{R}^{N \times N}$ (see Alg.1). Note that this overview is meant for the familiar reader; in the interest of space, we defer to numerous other works on the topic, such as [3, 4, 5, 7, 13, 15, 27, 24]. For simplicity, we assume $A$ is full rank.

Throughout the remainder of the paper, $0_{i,\ell}$ denotes a zero matrix of size $i \times \ell$ and $0$, is a column vector of $i$ zeros. We use $I$ to denote the square identity matrix; dimensions are either given as a single subscript, or are implicit from context. We use $e_i$ to denote the $i$th column of appropriately sized $I$.

In each outer loop $k$ of $s$-step BICG, we generate Krylov bases with the current search direction and residual vectors, $p_{sk}$ and $r_{sk}$, which we denote as $V^p_k$, having basis length $s + 1$, and $V^r_k$, having basis length $s$. The basis vectors, or columns of $V^p_k = [v^p_{k,0}, \ldots, v^p_{k,s}]$, are generated by the three-term polynomial recurrence

$$v^p_{k,i+1} = \gamma_i (A - \theta_i I) v^p_{k,i} + \sigma_i v^p_{k,i-1}$$

(2.1)
Algorithm 1 s-step BICG

1: $x_0, r_0 = \hat{r}_0 = b - Ax_0, p_0 = \hat{p}_0 = r_0, k = 0$
2: while not converged do
3: Calculate $V^p_k$, $V^k_r$, $V^p_k$, $B_k$
4: $V_k = [V^p_k, V^k_r]$, $V_k = [V^p_k, V^k_r]$, $G_k = \tilde{V}_k TV_k$
5: $p'_{k,0} = [1, 0, 1, 2s]$, $r'_{k,0} = [0, 1, s+1, 1, 0, 1, s-1]$, $x'_{k,0} = [0, 2s+1]$
6: $\tilde{p}'_{k,0} = [1, 0, 1, 2s]$, $\tilde{r}'_{k,0} = [0, 1, s+1, 1, 0, 1, s-1]^T$
7: for $j = 0 : s - 1$ do
8: $\alpha_{sk+j} = ((r'_{k,j})^T G_k r'_{k,j}) / ((\tilde{p}'_{k,j})^T G_k B_k \tilde{p}'_{k,j})$
9: $x'_{k,j+1} = x'_{k,j} + \alpha_{sk+j} \tilde{p}'_{k,j}$
10: $r'_{k,j+1} = r'_{k,j} - B_k (\alpha_{sk+j} \tilde{p}'_{k,j})$
11: $\tilde{r}'_{k,j+1} = \tilde{r}'_{k,j} - B_k (\alpha_{sk+j} \tilde{p}'_{k,j})$
12: $\beta_{sk+j+1} = ((\tilde{r}'_{k,j+1})^T G_k r'_{k,j+1}) / ((\tilde{r}'_{k,j})^T G_k r'_{k,j})$
13: $p'_{k,j+1} = r'_{k,j+1} + \beta_{sk+j+1} \tilde{p}'_{k,j}$
14: $\tilde{p}'_{k,j+1} = \tilde{r}'_{k,j+1} + \beta_{sk+j+1} \tilde{p}'_{k,j}$
15: end for
16: $x_{sk+s} = V_k x_{k,s} + x_{sk}$, $r_{sk+s} = V_k r'_{k,s}$, $p_{sk+s} = V_k p'_{k,s}$
17: $\tilde{r}_{sk+s} = V_k \tilde{r}'_{k,s}$, $\tilde{p}_{sk+s} = V_k \tilde{p}'_{k,s}$
18: $k = k + 1$
19: end while
20: return $x_{sk}$

with starting vector $v^p_{k,0} = p_{sk}$. We assume we use the same recurrence in constructing $v^p_{k,i}$. The choice of parameters $\gamma_j$, $\theta_i$, and $\sigma_i$ play a large role in determining the quality of the resulting basis, which in turn affects both stability and convergence in s-step BICG. We denote $V_k = [V^p_k, V^k_r]$. We also denote $V_k = [V^k_p, 0_N, V^k_r]$ where $V^p_k$ and $V^k_r$ are $V^p_k$ and $V^k_r$, resp., with their last columns omitted.

Within the inner loop, in step $j$ of outer loop $k$, we update the length-$(2s + 1)$ coefficients for the BICG vectors as linear combinations of the columns in $V_k$ (rather than explicitly update the length-$N$ BICG vectors, as in classical BICG). The coefficient vectors are denoted with prime symbols (i.e., $r_{sk+j} = V_k r'_{k,j}$, and similarly for $p_{sk+j}$ and $x_{sk+j}$). The inner iteration updates then become

$$r'_{k,j+1} = r'_{k,j} - \alpha_{sk+j} B_k \tilde{p}'_{k,j} \quad \text{and} \quad (2.2)$$

$$p'_{k,j+1} = r'_{k,j+1} + \beta_{sk+j+1} \tilde{p}'_{k,j} \quad \text{and} \quad (2.3)$$

where

$$B_k = \begin{bmatrix} C_{k,s+1} & 0_{s+1,1} \\ 0_{s,1} & C_{k,s} \end{bmatrix}$$

with

$$C_{k,s} = \begin{bmatrix} \theta_0 & -\sigma_1/\gamma_1 \\ 1/\gamma_0 & \theta_1 \\ & \ddots \\ & & \theta_{j-1} \\ & & 1/\gamma_{j-1} & -\sigma_{j-1}/\gamma_{j-1} \end{bmatrix}$$
We can rearrange (2.2) and (2.3) as

\[ B_k p_{k,j} = \frac{1}{\alpha_{sk+j}} (r'_{k,j} - r'_{k,j+1}) \quad \text{and} \quad (2.4) \]

\[ r'_{k,j} = p'_{k,j} - \beta_{sk+j} p'_{k,j-1}. \quad (2.5) \]

Premultiplying (2.5) by \( V_k \), we obtain

\[ V_k r'_{k,j} = V_k p'_{k,j} - \beta_{sk+j} V_k p'_{k,j-1}. \quad (2.6) \]

This equation is valid for \( 1 \leq j < s \), since \( p'_{k,-1} \) is undefined. When \( j = 0 \), we have

\[ V_k r'_{k,0} = V_{k-1} r'_{k-1,s} = V_{k-1} (p'_{k-1,s} - \beta_{sk} p'_{k-1,s-1}) = V_k p'_{k,0} - \beta_{sk} V_{k-1} p'_{k-1,s-1}, \]

which gives a valid expression for the \( j = 0 \) case.

Now, let

\[ R'_{k,j} = [r'_{k,0}, r'_{k,1}, \ldots, r'_{k,j}] \quad \text{and} \quad P'_{k,j} = [p'_{k,0}, p'_{k,1}, \ldots, p'_{k,j}]. \]

We can write (2.6) in block form as

\[ V_k R'_{k,j} = V_k P'_{k,j} U_{k,j} - \beta_{sk} V_{k-1} p'_{k-1,s-1} e_1^T, \quad (2.7) \]

where

\[ U_{k,j} = \begin{bmatrix} 1 & \beta_{sk+1} & & & \\ & 1 & \ddots & & \\ & & \ddots & -\beta_{sk+j} & \\ & & & 1 & \end{bmatrix}. \]

Premultiplying (2.7) by \( A \), we obtain

\[ A V_k R'_{k,j} = A V_k P'_{k,j} U_{k,j} - \beta_{sk} A V_{k-1} p'_{k-1,s-1} e_1^T. \quad (2.8) \]

We can also write (2.4) in block form as

\[ B_k P'_{k,j} = R'_{k,j} L_{k,j} \Lambda^{-1}_{k,j} - \frac{1}{\alpha_{sk+j}} r'_{k,j+1} e_{j+1}^T, \quad (2.9) \]

where \( \Lambda_{k,j} = \text{diag}(\alpha_{sk}, \ldots, \alpha_{sk+j}) \) and

\[ L_{k,j} = \begin{bmatrix} 1 & 1 & & \\ -1 & 1 & \ddots & \\ & \ddots & \ddots & \ddots \\ & & -1 & 1 \end{bmatrix}. \]

If we premultiply (2.9) by \( V_k \) and postmultiply by \( U_{k,j} \), we obtain

\[ V_k B_k P'_{k,j} U_{k,j} = V_k R'_{k,j} L_{k,j} \Lambda^{-1}_{k,j} U_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T, \]
which can be written

\[ AV_k p'_k U_{k,j} = V_k R'_{k,j} L_{k,j} \Lambda_{k,j}^{-1} U_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T \]  \tag{2.10}

since \( AV_k = V_k B_k \) and \( V_k R'_k = V_k R'_j \) for \( j \leq s - 1 \). We can then combine (2.8) and (2.10) to obtain

\[ AV_k R'_{k,j} = V_k R'_{k,j} T_{k,j} - \frac{\beta_{sk}}{\alpha_{sk-1}} V_{k-1} r'_{k-1,s-1} e_1^T - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T, \]  \tag{2.11}

where \( T_{k,j} = L_{k,j} \Lambda_{k,j}^{-1} U_{k,j} + e_1 \frac{\beta_{sk}}{\alpha_{sk-1}} e_1^T \). Note when \( k = 0 \), \( \frac{\beta_{sk}}{\alpha_{sk-1}} \) is defined to be 0.

We can now combine outer loop iterations in block form to write the \( s \)-step BICG recurrence for iterations 0 through \( sk + j \). Let \( V_k = [V_0, V_1, \ldots, V_k] \). Let

\[ R'_{k,j} = \begin{bmatrix} R'_{0,s-1} \\ R'_{1,s-1} \\ \vdots \\ R'_{k,j} \end{bmatrix} \]

and

\[ T_{k,j} = \begin{bmatrix} \frac{1}{\alpha_0} & -\frac{\beta_1}{\alpha_1} & & & \\ -\frac{1}{\alpha_0} & \frac{\beta_1}{\alpha_1} & \frac{\beta_2}{\alpha_2} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \frac{\beta_{sk+j}}{\alpha_{sk+j-1}} \\ -\frac{1}{\alpha_{sk+j-1}} & \frac{\alpha_{sk+j-1}}{\alpha_{sk+j-1}} & & & \frac{\beta_{sk+j}}{\alpha_{sk+j}} \end{bmatrix}. \]

Then by (2.11), we can write

\[ A V_k R'_{k,j} = V_k R'_{k,j} T_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T. \]

Since we can write the residual vectors as \( R_n = [r_0, \ldots, r_n] = V_k R'_{k,j} \), where \( n = sk + j \), we can write the above as

\[ A R_n = R_n T_n - \frac{1}{\alpha_n} r_{n+1} e_{n+1}^T, \]

which gives us the same governing equation for iterations 0 through \( sk + j \) as the classical BICG algorithm in exact arithmetic [25]. Note that a similar relation holds for the dual Krylov vectors \( \tilde{r}_{sk+j} \) and \( \tilde{p}_{sk+j} \).

3. \textbf{s-step BICG in finite precision.} The goal of this section is to derive a Lanczos-type recurrence for finite precision \( s \)-step BICG of the form

\[ AV_k R'_{k,j} = V_k R'_{k,j} T_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{sk+j+1} + \epsilon \Delta_{k,j} \]

and upper bound the size of the error term \( \epsilon \Delta_{k,j} \). We assume a standard model of floating point arithmetic, where

\[ \text{fl} (\alpha x + y) = \alpha x + y + \delta_1, \quad \text{where} \quad |\delta_1| \leq \epsilon 2 |\alpha x| + |y| + O(\epsilon^2), \quad \text{and} \]

\[ \text{fl} (Ax) = Ax + \delta_2, \quad \text{where} \quad |\delta_2| \leq \epsilon N |A| |x| + O(\epsilon^2), \]
where \(x, y \in \mathbb{R}^N\), \(\alpha \in \mathbb{R}\). In the remaining analysis we drop higher powers of \(\epsilon\) for simplicity. Let \(\epsilon\) be the machine precision unit. For simplicity of notation, we now let \(r_{k,j}', p_{k,j}', \alpha_{sk+j}, r_{k,s}, p_{k,s}, \beta_{sk+j}, V_k\), and \(B_k\) be the computed quantities in finite precision \(s\)-step BICG.

At the \((sk + j)\)th iteration, to compute \(r_{k,j+1}'\) we first compute \(B_k p_{k,j}'\) and have

\[
\text{fl} (B_k p_{k,j}') = B_k p_{k,j}' + g, \quad \text{where} \quad |g| \leq \epsilon (2s + 1) |B_k| |p_{k,j}'|.
\]

Then

\[
r_{k,j+1}' = \text{fl} \left( r_{k,j}' - \alpha_{sk+j} \cdot \text{fl} \left( B_k p_{k,j}' \right) \right) = r_{k,j}' - \alpha_{sk+j} B_k p_{k,j}' - \alpha_{sk+j} g + g',
\]

where \(|g'| \leq \epsilon \left( |r_{k,j}'| + 2 |\alpha_{sk+j}| |B_k p_{k,j}'| \right)\). Let \(\delta_{r_{k,j}} = (\alpha_{sk+j} g + g') / (\epsilon |\alpha_{sk+j}|)\). Then using (3.1) we obtain

\[
\frac{1}{\alpha_{sk+j}} (r_{k,j+1}' - r_{k,j}') = -B_k p_{k,j}' + \epsilon \delta_{r_{k,j}'},
\]

where

\[
|\delta_{r_{k,j}}| \leq (2s + 1) |B_k| |p_{k,j}'| + \frac{|r_{k,j}'|}{|\alpha_{sk+j}|} + 2 |B_k p_{k,j}'|.
\]

Similarly,

\[
p_{k,j+1}' = \text{fl} \left( r_{k,j+1}' + \beta_{sk+j+1} p_{k,j}' \right) = r_{k,j+1}' + \beta_{sk+j+1} p_{k,j}' + f,
\]

where \(|f| \leq \epsilon \left( |r_{k,j+1}'| + 2 |\beta_{sk+j+1}| |p_{k,j}'| \right)\). Letting \(\delta_{p_{k,j+1}} = f / \epsilon\), we have

\[
p_{k,j+1}' = r_{k,j+1}' + \beta_{sk+j+1} p_{k,j}' + \epsilon \delta_{p_{k,j+1}'},
\]

where

\[
|\delta_{p_{k,j+1}}| \leq |r_{k,j+1}'| + 2 |\beta_{sk+j+1}| |p_{k,j}'|.
\]

Rearranging (3.2) and (3.3), we can write

\[
B_k p_{k,j}' = \frac{1}{\alpha_{sk+j}} (r_{k,j}' - r_{k,j+1}') + \epsilon \delta_{r_{k,j}'}, \quad \text{and}
\]

\[
r_{k,j}' = p_{k,j}' - \beta_{sk+j} p_{k,j-1}' + \epsilon \delta_{p_{k,j}'},
\]

and premultiplying (3.4) by \(V_k\) gives

\[
V_k r_{k,j}' = V_k p_{k,j}' - \beta_{sk+j} V_k p_{k,j-1}' + \epsilon V_k \delta_{p_{k,j}'}.
\]

This equation is valid for \(1 \leq j < s\), since \(p_{k,j-1}'\) is undefined. When \(j = 0\), we have

\[
V_k r_{k,0}' = \text{fl}(V_{k-1} r_{k-1,0}')
\]

\[
= V_{k-1} r_{k-1,0}' + \epsilon \delta_{r_{k-1,0}'}, \quad \text{and}
\]

\[
V_k p_{k,0}' = \text{fl}(V_{k-1} p_{k-1,0}')
\]

\[
= V_{k-1} p_{k-1,0}' + \epsilon \delta_{p_{k-1,0}'},
\]
where $|\phi'_{k-1}| \leq (2s+1)|V_{k-1}||r'_{k-1,s}|$ and $|\phi''_{k-1}| \leq (2s+1)|V_{k-1}||p'_{k-1,s}|$. Then for $j = 0$, we can write

$$V_k r'_{k,0} = V_{k-1} r'_{k-1,s} + \epsilon \phi''_{k-1}$$

$$= V_{k-1} \left( p'_{k-1,s} - \beta_{sk} p'_{k-1,s-1} + \epsilon \delta'_{k-1,s} + \epsilon \phi''_{k-1} \right) + \epsilon \phi''_{k-1}$$

$$= V_k p'_{k,0} - \epsilon \phi''_{k-1} - \beta_{sk} V_{k-1} p'_{k-1,s-1} + \epsilon V_{k-1} \delta'_{k-1,s} + \epsilon \phi''_{k-1}$$

$$= V_k p'_{k,0} - \beta_{sk} V_{k-1} p'_{k-1,s-1} + \epsilon \left( V_{k-1} \delta'_{k-1,s} + \phi''_{k-1} - \phi''_{k-1} \right).$$

Now, let $\Delta'_{k,j} = \left[ \delta'_{k,0}, \ldots, \delta'_{k,j} \right]$ and $\Delta p'_{k,j} = \left[ 0_{2s+1}, \delta'_{k,i}, \ldots, \delta'_{k,j} \right]$. We can then write

$$V_k R'_{k,j} = V_k P'_{k,j} U_{k,j} - \beta_{sk} V_k V_{k-1} p'_{k-1,s-1} e_1^T + \epsilon V_k \Delta p'_{k,j}$$

$$+ \epsilon \left( V_{k-1} \delta'_{k-1,s} + \phi''_{k-1} - \phi''_{k-1} \right) e_1^T \quad \text{and}$$

(3.5)

$$P_k P'_{k,j} = R'_{k,j} L_{k,j} A^{-1}_{k,j} - \frac{1}{\alpha_{sk+j}} r'_{k,j+1} e_{j+1}^T + \epsilon \Delta R'_{k,j}. \quad \text{(3.6)}$$

Premultiplying (3.5) by $A$ gives

$$AV_k R'_{k,j} = AV_k P'_{k,j} U_{k,j} - \beta_{sk} AV_k V_{k-1} p'_{k-1,s-1} e_1^T + \epsilon AV_k \Delta p'_{k,j}$$

$$+ \epsilon A \left( V_{k-1} \delta'_{k-1,s} + \phi''_{k-1} - \phi''_{k-1} \right) e_1^T,$$

(3.7)

and premultiplying (3.6) by $V_k$ gives

$$V_k B_k P'_{k,j} = V_k R'_{k,j} L_{k,j} A^{-1}_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T + \epsilon V_k \Delta R'_{k,j} \quad \text{(3.8)}$$

for $j = s - 1$.

Now, to write the error in $s$-step BICG in the context of classical BICG, we must account for error in computation of the $s$-step bases. Rearranging the finite precision evaluation of (2.1), we obtain

$$\begin{aligned}
Av_{p}^{i} & = \frac{1}{\gamma_{i}} v_{p,i}^{i+1} + \theta_{i} v_{p}^{i} - \frac{\sigma_{i}}{\gamma_{i}} v_{k,i-1}^{i} + \epsilon \delta_{v,i}^{i}, \\
\end{aligned}$$

where we can write $|\delta_{v,i}^{i}|$ as

$$|\delta_{v,i}^{i}| = (N + 2) |A| |v_{k,i}^{i}| + 3|\theta_{i} v_{p}^{i}| + \frac{3|\sigma_{i} v_{k,i-1}^{i}|}{|\gamma_{i}|}.$$ 

Since we generate $v_{k,i}^{i}$ by the same recurrence, we also have

$$|\delta_{v,i}^{i}| = (N + 2) |A| |v_{k,i}^{i}| + 3|\theta_{i} v_{p}^{i}| + \frac{3|\sigma_{i} v_{k,i-1}^{i}|}{|\gamma_{i}|}.$$ 

Letting $\Delta_{V_k} = \left[ \delta_{V_{k,0}}, \ldots, \delta_{V_{k,s-1}}, 0, \delta_{V_{k,s-1}}, \ldots, \delta_{V_{k,s-2}}, 0 \right]$, we can then write the finite precision basis computation as

$$AV_k = V_k B_k + \epsilon \Delta_{V_k}. \quad \text{(3.9)}$$
Using (3.9), we can write (3.8) as
\[
(AV_k - \epsilon \Delta_{V_{k,j}}) P'_{k,j} = V_k R'_{k,j} L_{k,j} A^{-1}_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T + \epsilon V_k \Delta_{R'_{k,j}} + \epsilon \Delta_{V_{k,j}},
\]
which can be rearranged to obtain
\[
AV_k P'_{k,j} = V_k R'_{k,j} L_{k,j} A^{-1}_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T + \epsilon V_k \Delta_{R'_{k,j}} + \epsilon \Delta_{V_{k,j}}, \tag{3.10}
\]
Postmultiplying (3.10) by \( U_{k,j} \) gives
\[
AV_k P'_{k,j} U_{k,j} = V_k R'_{k,j} L_{k,j} A^{-1}_{k,j} U_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T + \epsilon (V_k \Delta_{R'_{k,j}} U_{k,j} + \Delta_{V_{k,j}} P'_{k,j} U_{k,j}), \tag{3.11}
\]
and combining (3.11) and (3.7), we obtain
\[
AV_k R'_{k,j} = V_k R'_{k,j} L_{k,j} A^{-1}_{k,j} U_{k,j} - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T - \beta_{sk} AV_k \epsilon_{k-1,s-1} e_1^T \tag{3.12}
+ \epsilon (AV_k \Delta_{P'_{k,j}} + V_k \Delta_{R'_{k,j}} U_{k,j} + \Delta_{V_{k,j}} P'_{k,j} U_{k,j}) \tag{3.13}
+ \epsilon A \left( V_k \Delta_{\epsilon_{k-1,s}} + \phi_{k-1} - \phi'_{k-1} \right) e_1^T. \tag{3.14}
\]
Since
\[
\beta_{sk} AV_k \epsilon_{k-1,s-1} e_1^T = \beta_{sk} (V_k - \epsilon_{k-1} B_{k-1} + \Delta V_{k-1}) \epsilon_{k-1,s-1} e_1^T
= \beta_{sk} V_k \frac{r'_{k-1,s-1} - r'_{k-1,s}}{\alpha_{sk-1}} \epsilon_{k-1,s-1} e_1^T + \beta_{sk} \Delta V_{k-1} \epsilon_{k-1,s-1} e_1^T
= \beta_{sk} V_k \frac{r'_{k-1,s-1} - r'_{k-1,s}}{\alpha_{sk-1}} \epsilon_{k-1,s-1} e_1^T + \beta_{sk} \Delta V_{k-1} \epsilon_{k-1,s-1} e_1^T,
\]
we can write (3.14) as
\[
AV_k R'_{k,j} = V_k R'_{k,j} T_{k,j} - \frac{\beta_{sk}}{\alpha_{sk-1}} V_k \epsilon_{k-1,s-1} e_1^T - \frac{1}{\alpha_{sk+j}} V_k r'_{k,j+1} e_{j+1}^T + \epsilon \Delta_{k,j},
\]
where
\[
\Delta_{k,j} = AV_k \Delta_{P'_{k,j}} + AV_k \epsilon_{k-1,s} e_1^T + V_k \Delta_{R'_{k,j}} U_{k,j} - \beta_{sk} AV_k \epsilon_{k-1,s-1} e_1^T \tag{3.15}
+ \Delta_{V_{k,j}} \epsilon_{k-1,s} e_1^T + \left( \phi_{k-1} - \phi'_{k-1} \right) e_1^T. \tag{3.16}
\]
Writing \( \Delta_{k,j} = [\delta_{sk}, \ldots, \delta_{sk+j}] \), we have that the \((sk + j + 1)\)th column of \( \Delta_{k,j} \) is
\[
\delta_{sk+j} = AV_k \delta_{p'_{k,j}} + V_k \delta_{r'_{k,j}} - \beta_{sk+j} V_k \delta_{r'_{k,j-1}} + \Delta_{V_{k,j}} r'_{k,j-1}, \tag{3.17}
\]
for \( j > 0 \), and
\[
\delta_{sk} = AV_{k-1}k_{sk-1,s} + V_k\delta_{sk} + V_kV_{k-1}\delta_{sk} + \delta_{sk} + \beta_{sk+j}V_{k-1}k_{sk-1,s-1} + \delta_{sk} + \beta_{sk+j}V_{k-1}k_{sk-1,s-1}
\]
\[
+ \left( A(\phi_{sk}^0 - \phi_{sk}^j) - \frac{\beta_{sk}}{\alpha_{sk}} - \phi_{sk}^j \right)
\]
for \( j = 0 \).

Using the inequalities \( |\beta_{sk+j}p_{k}^j| < |r_{k,j}^j| + |O(\epsilon)| \) and \( |r_{k,j-1}^j| < |r_{k,j}^j| + \alpha_{sk+j}p_{k}^j + O(\epsilon) \), we can bound the norm of the columns as
\[
|\delta_{sk+j}| \leq \left( (N + 6)|A||V_k| + (2s + 8)\frac{1}{N} + \frac{|\beta_{sk+j}|}{|\alpha_{sk+j-1}|} \right) r_{k,j}^j
\]
\[
+ \left( 2|A||V_k| + (4s + 7)\frac{1}{N} p_{k,j}^j \right)
\]
if \( j > 0 \). For the \( j = 0 \) case, we have
\[
|\delta_{sk}| \leq \left( (N + 2s + 7)|A||V_k| + (2s + 8)\frac{1}{N} + \frac{|\beta_{sk}|}{|\alpha_{sk}|} \right) r_{k,j}^j
\]
\[
+ \left( 2|A||V_k| + (4s + 16)\frac{1}{N} p_{k,j}^j \right)
\]
We can thus write the finite precision \( s \)-step BICG recurrence for iterations 0 through \( sk + j \) as
\[
AV_k\hat{R}_{k,j} = V_k\hat{T}_{k,j} - \frac{1}{\alpha_{sk+j}}V_k\hat{T}_{k,j} + \epsilon\Delta_{k,j},
\]
where \( \Delta_{k,j} = [\Delta_{0,s-1}, \Delta_{1,s-1}, \ldots, \Delta_{k,j}] \).

3.1. Comments. Note that we can write \( s \) iterations of finite precision classical BICG as \( n \) iterations of finite precision \( s \)-step BICG with \( s > n \), performed in the standard basis. That is, we have a single outer loop iteration \( k = 0 \) and \( j = n \) inner loop iterations, with \( V_0 = I_n \), \( R_0 = A \), and \( P_{0,n} = P_{0,n} \) defined to be zero. Now, since \( V_0 = I_n \), \( \Delta_{0,0} = 0 \), and since \( k = 0 \), \( \phi_{k-1}^j, \phi_{k}^j \), and \( \delta_{sk}^j \) are defined to be zero. Plugging in to (3.15), we get
\[
\Delta_{s,n} = A\Delta_{p,n} + \Delta_{R_{0,n}} U_{0,n},
\]
which reproduces the error term (modulo a factor of 2) obtained by Tong and Ye for finite precision classical BICG.

Also note that from (3.15), we can see that the first four terms on the right-hand side correspond to the two terms in Tong and Ye’s analysis for classical BICG, and the remaining terms correspond to the error in computing the \( s \)-step Krylov bases and the change of basis operation. We can also see that a bound on the size of the error in each column of the finite precision recurrence depends on both the magnitude of the error in computing the \( s \)-step Krylov bases, i.e., \( \|\Delta V_k\| \), as well as the size of the bases, i.e., \( \|V_k\| \).
3.2. Diagonal scaling. As in [25], it will be more convenient to work with a scaled version of (3.22) in subsequent sections. Let $Z_{k,j} = [z_{0,0}, \ldots, z_{k,j}] = v_k r_{k,j}^t D_{k,j}^{-1}$ where

$$D_{k,j} = \text{diag} (\|v_{0,0}^t\|, \ldots, \|v_{0,0}^t\|; \|v_{1,0}^t\|, \ldots, \|v_{k,j}^t\|).$$

We can then write the scaled version of (3.22) as

$$A Z_{k,j} = Z_{k,j}[\bar{t}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \|r_0\| e_{sk+j+1}^T] e_{sk+j+1} + \epsilon \bar{\Delta}_{k,j}, \quad (3.23)$$

where $\bar{t}_{k,j} = D_{k,j} T_{k,j} D_{k,j}^{-1}$,

$$\bar{\alpha}_{sk+j} = \|v_{k,j}^t\| \alpha_{sk+j} / \|r_0\| = \|v_{k,j}^t\| \alpha_{sk+j} / \|v_{0,0}^t\| = e_{sk+j+1}^T T_{k,j}^{-1} e_1,$$

and

$$\bar{\Delta}_{k,j} = \Delta_{k,j} D_{k,j}^{-1}. \quad (4.4)$$

4. Bounds on $\|r_{sk+j+1}\|$ for finite precision s-step BICG. In this subsection, we upper bound the norm of the updated residual computed in iteration $sk + j$ of s-step BICG. First, we will review a series of Lemmas proved by Tong and Ye [25]. The proofs shown below are nearly identical to those given by Tong and Ye [25], although we have changed the notation and indexing for consistency with our s-step formulation.

**Lemma 4.1.** Assume

$$A Z_{k,j} = Z_{k,j} \bar{t}_{k,j} - \frac{1}{\alpha_{sk+j}} \|r_0\| e_{sk+j+1}^T e_{sk+j+1}$$

with $r_0 = \|r_0\| z_0$. Then for any polynomial $\rho(x) = \sum_{i=0}^{sk+j+1} \psi_i x^i$ of degree $\leq sk + j + 1$,

$$\rho(A) z_0 = Z_{k,j} \rho(\bar{t}_{k,j}) e_1 + c_{sk+j} v_{k,j}^t e_{sk+j+1},$$

where $c_{sk+j} = (-1)^{sk+j+1} \psi_{sk+j+1}/(\alpha_0 \cdots \alpha_{sk+j} \|r_0\|)$.

**Proof.** First, we will prove by induction that for $1 \leq i \leq sk + j$

$$A^i Z_{k,j} e_1 = Z_{k,j} \tilde{t}_{k,j} e_1. \quad (4.1)$$

For $i = 1$, we have

$$A Z_{k,j} e_1 = (Z_{k,j} \bar{t}_{k,j} - \frac{1}{\alpha_{sk+j}} \|r_0\| e_{sk+j+1}^T e_{sk+j+1}) e_1 = Z_{k,j} \bar{t}_{k,j} e_1.$$

Now, assume (4.1) holds for some $i < sk + j$. Then

$$A^{i+1} Z_{k,j} e_1 = A (A^i Z_{k,j} e_1) = A (Z_{k,j} \bar{t}_{k,j} e_1) = (Z_{k,j} \bar{t}_{k,j} - \frac{1}{\alpha_{sk+j}} \|r_0\| e_{sk+j+1}^T e_{sk+j+1}) \bar{t}_{k,j} e_1 = Z_{k,j} \bar{t}_{k,j} \bar{t}_{k,j} e_1 = Z_{k,j} \bar{t}_{k,j} e_1.$$

---

1 One lemma presented is slightly different than what appears in [25] due to a minor mathematical error that we correct.
where we have used the fact that $e_{sk+j+1}^T T_{k,j} e_1 = 0$ when $i < sk+j$. Therefore the inductive hypothesis holds. Now consider the case $i = sk+j$. We then have

$$A^{sk+j+1} Z_{k,j} e_1 = A(Z_{k,j} T_{k,j}^{sk+j+1}) e_1$$

$$= (Z_{k,j} T_{k,j} - \frac{1}{\alpha_{sk+j}} \frac{V_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T T_{k,j}^{sk+j+1} e_1).$$

Since it can be shown that $e_{sk+j+1}^T T_{k,j} e_1 = (-1)^{sk+j} \|V_k r'_{k,j}\|(A_0 \cdots A_{sk+j} \|r_0\|)^{-1}$ and $\alpha_{sk+j} = \|V_k r'_{k,j}\| \alpha_{sk+j}/\|r_0\|$, we have

$$A^{sk+j+1} Z_{k,j} e_1 = Z_{k,j} T_{k,j}^{sk+j+1} e_1$$

$$= Z_{k,j} T_{k,j}^{sk+j+1} e_1 + \frac{(-1)^{sk+j+1}}{A_0 \cdots A_{sk+j} \|r_0\|} V_k r'_{k,j+1}.$$

The lemma follows. $\Box$

We now use this result in proving the following identity.

**Lemma 4.2.** Assume

$$A Z_{k,j} = Z_{k,j} \bar{T}_{k,j} - \frac{1}{\alpha_{sk+j}} \frac{V_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T$$

with $r_0 = \|r_0\| e_1$ and $\alpha_{sk+j} = e_{sk+j+1}^T \bar{T}_{k,j}^{-1} e_1$. Assume that $W T \in \mathbb{R}^{(sk+j+1) \times N}$ is a matrix such that $W T Z_{k,j} = I$ and $W T V_k r'_{k,j+1} = 0_{sk+j+1}$. Then for any polynomial $\rho(x)$ of degree not exceeding $sk+j$ with $\rho(0) = 1$, we have

$$V_k r'_{k,j+1} = (I - A Z_{k,j} \bar{T}_{k,j}^{-1} W T) \rho(A) r_0.$$

**Proof.** First, we multiply by $\bar{T}_{k,j}^{-1} e_1$ to get

$$A Z_{k,j} \bar{T}_{k,j}^{-1} e_1 = \left( Z_{k,j} \bar{T}_{k,j} - \frac{1}{\alpha_{sk+j}} \frac{V_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T \right) \bar{T}_{k,j}^{-1} e_1,$$

which allows us to write

$$\frac{V_k r'_{k,j+1}}{\|r_0\|} = z_0 - A Z_{k,j} \bar{T}_{k,j}^{-1} e_1.$$

Now, let $\rho(x) = 1 + x \phi(x)$, with $\phi(x) = \sum_{i=0}^{sk+j} \psi_{i+1} x^i$ a polynomial of degree not exceeding $sk+j$. Then

$$\frac{V_k r'_{k,j+1}}{\|r_0\|} = z_0 - A Z_{k,j} \bar{T}_{k,j}^{-1} e_1 + (\rho(A) z_0 - \rho(A) z_0)$$

$$= -A \phi(A) z_0 - A Z_{k,j} \bar{T}_{k,j}^{-1} e_1 + \rho(A) z_0$$

$$= -A Z_{k,j} \phi(\bar{T}_{k,j}) e_1 - A Z_{k,j} \bar{T}_{k,j}^{-1} e_1 + \rho(A) z_0$$

$$= -A Z_{k,j} (\phi(\bar{T}_{k,j}) + \bar{T}_{k,j}^{-1}) e_1 + \rho(A) z_0$$

$$= -A Z_{k,j} \bar{T}_{k,j}^{-1} \rho(\bar{T}_{k,j}) e_1 + \rho(A) z_0. \quad (4.2)$$
By Lemma 4.1, recall that
\[ \rho(A)z_0 = Z_{k,j}\rho(\bar{T}_{k,j})e_1 + c_{sk+j}V_k r'_{k,j+1}, \]
and, multiplying by \( W^T \), we have
\[ W^T \rho(A)z_0 = W^T \left( Z_{k,j}\rho(\bar{T}_{k,j})e_1 + c_{sk+j}V_k r'_{k,j+1} \right) = \rho(\bar{T}_{k,j})e_1, \]
since \( W^T Z_{k,j} = I \) and \( W^T V_k r'_{k,j+1} = 0_{sk+j+1} \). Now, we can write
\[ \frac{V_k r'_{k,j+1}}{\|r_0\|} = -AZ_{k,j}\bar{T}_{k,j}^{-1}W^T \rho(A)z_0 + \rho(A)z_0 = \left( I - AZ_{k,j}\bar{T}_{k,j}^{-1}W^T \right) \rho(A)z_0, \]
and substituting \( z_0 = r_0/\|r_0\| \), we obtain
\[ V_k r'_{k,j+1} = \left( I - AZ_{k,j}\bar{T}_{k,j}^{-1}W^T \right) \rho(A)r_0, \]
which gives the desired result. \( \Box \)

The following lemma describes the construction of the basis \( W \).

**Lemma 4.3.** Assume that \( z_0, \ldots, z_{sk+j+1} \in \mathbb{R}^N \) are linearly independent and write \( Z_{k,j} = [z_0, \ldots, z_{sk+j}] \), \( Z^+_{k,j} = [z_0, z_{sk+j+1}] \). Then \( W_0^T = [I_{sk+j+1}, 0_{sk+j+1}]Z^+_{k,j} \) has the property \( W_0^T Z_{k,j} = I \) and \( W_0^T z_{sk+j+1} = 0_{sk+j+1} \). Furthermore, its spectral norm is minimal among all matrices having this property.

**Proof.** By the definition of \( W_0 \), \( Z^+_{k,j} = [W_0, w]^T \) for some \( w \). Since we assume \( z_0, \ldots, z_{sk+j+1} \) are linearly independent,
\[ [W_0, w]^T[Z_{k,j}, z_{sk+j+1}] = Z^+_{k,j}Z_{k,j} = I. \]

Then \( W_0^T Z_{k,j} = I_{sk+j+1} \) and \( W_0^T z_{sk+j+1} = 0_{sk+j+1} \).

Now, assume \( W \) is some other matrix such that \( W^T Z_{k,j} = I \) and \( W^T z_{sk+j+1} = 0_{sk+j+1} \) hold. Then \( W^T Z_{k,j} = [I_{sk+j+1}, 0_{sk+j+1}]Z^+_{k,j} \). Thus, \( W^T Z_{k,j} Z^+_{k,j} = [I_{sk+j+1}, 0_{sk+j+1}]Z^+_{k,j} = W_0^T \). Hence \( \|W_0\| \leq \|W\| \cdot \|Z^+_{k,j}Z^+_{k,j}\| \leq \|W\| \). \( \Box \)

We can now present the main result.

**Theorem 4.4.** Assume (3.23) holds and let \( W_0^T = [I_{sk+j+1}, 0_{sk+j+1}]Z^+_{k,j} \in \mathbb{R}^{(sk+j+1)\times N}. \) If \( z_0, \ldots, z_{sk+j+1} \) are linearly independent, then
\[ \|V_k r'_{k,j+1}\| \leq (1 + K_{k,j}) \min_{\rho \in \mathbb{R}^{sk+j+1}, \rho(0)=1} \|\rho(A + \delta A_{k,j})r_0\|, \quad (4.3) \]
where \( K_{k,j} = \|\left( AZ_{k,j} - \epsilon \Delta_{k,j}\right)\bar{T}_{k,j}^{-1}W_0^T\| \) and \( \delta A_{k,j} = -\epsilon \Delta_{k,j} Z^+_{k,j} \).

**Proof.** Since \( z_0, \ldots, z_{sk+j+1} \) are linearly independent, \( Z^+_{k,j}Z_{k,j} = I \). Then \( \delta A_{k,j} = -\epsilon \Delta_{k,j} Z^+_{k,j} \in \mathbb{R}^{N\times N} \) satisfies \( \delta A_{k,j} Z_{k,j} = -\epsilon \Delta_{k,j} \). Thus (3.23) can be written as
\[ (A + \delta A_{k,j})Z_{k,j} = Z_{k,j}\bar{T}_{k,j} - \frac{1}{\epsilon \Delta_{sk+j} Z_{k,j}} \left( V_k r'_{k,j+1} / \|r_0\| \right) c_{sk+j+1}. \quad (4.4) \]
Then, by Lemma 4.2, for any $\rho \in P_{sk+j+1}$ with $\rho(0) = 1$, we obtain

$$
\bar{v}_{k+1} = (I - (A + \delta A_{k,j})Z_{k,j} \tilde{T}_{k,j}^{-1} W_0^T) \cdot \rho(A + \delta A_{k,j})r_0
$$

$$
= (I - (AZ_{k,j} - \epsilon \Delta_{k,j}) \tilde{T}_{k,j}^{-1} W_0^T) \cdot \rho(A + \delta A_{k,j})r_0.
$$

Thus, we can bound the norm of the left hand side by

$$
\|\bar{v}_{k+1}\| \leq (1 + \|AZ_{k,j} - \epsilon \Delta_{k,j}\| \tilde{T}_{k,j}^{-1} W_0^T) \cdot \|\rho(A + \delta A_{k,j})r_0\|.
$$

Since this holds for any $\rho(x)$ with $\rho(0) = 1$, the inequality is true for the minimizing polynomial, which leads to the bound. \(\square\)

Note that $\tau_{k,j} = \min_{\rho \in P_{sk+j+1}, \rho(0) = 1} \|\rho(A + \delta A_{k,j})r_0\|$ is the $(sk + j)$th residual norm of exact GMRES applied to the perturbed matrix $A + \delta A_{k,j}$, which decreases monotonically with increasing $(sk + j)$.

Since we have $K_{k,j} = \|(AZ_{k,j} - \epsilon \Delta_{k,j}) \tilde{T}_{k,j}^{-1} W_0^T\|$, we can bound $K_{k,j}$ as

$$
K_{k,j} \leq (\sqrt{sk + j + 1} \|A\| + \epsilon \|\Delta_{k,j}\|) \|\tilde{T}_{k,j}^{-1}\| \cdot \|W_0\|.
$$

Then, assuming $\|\tilde{T}_{k,j}^{-1}\|$ and $\|W_0\|$ are bounded, $\|\bar{v}_{k+1}\|$ is on the order $O(\tau_{k,j})$. We therefore expect convergence of the $s$-step BICG residual when $K_{k,j}$ increases at a slower rate than $\tau_{k,j}$ decreases, for all values of $k$.

Unfortunately, as in the BICG case, we can not determine $K_{k,j}$ a priori, although we can make some meaningful observations based on the bound in (4.3). Clearly, the terms $\epsilon \Delta_{k,j}$ in $K_{k,j}$ and $\|\Delta_{k,j}\|$ are determined by the residual vector $(sk + j)$ of the GMRES step on the perturbed matrix $A + \delta A_{k,j}$.

Note that in the case of CG (SPD $A$), we have $\|\bar{v}_{k+1}\| = \|r_{sk+j+1}\| = \|e_{sk+j+1}\|_A$, where $e_{sk+j+1}$ denotes the solution error $e_{sk+j+1} = x^* - x_{sk+j}$ for the true solution $x^*$. Thus in this case Theorem 4.4 gives a bound on the error of finite precision $s$-step CG. It remains future work to determine under what conditions $\|e_{sk+j+1}\|_A < \|e_{sk+j}\|_A$ for $s$-step CG.

5. The linearly dependent case. In the analysis above, we assumed linear independence among the residual vectors (which are scalar multiples of the Lanczos vectors). For many linear systems, however, convergence of classical BICG in finite precision is still observed despite numerical rank deficiency of the basis. In [25] it is shown how the residual norm can be bounded absent the assumption of linear independence, which gives insight into why convergence still occurs in such cases. We will now prove similar bounds, relaxing the constraint that $x_0, \ldots, x_{sk+j+1} \in \mathbb{R}^N$ be linearly independent. Again, our analysis extends that of Tong and Ye [25] for classical BICG.

We note that in the $s$-step case, there are two potential causes of a rank-deficient basis. Since we have $R_{k,j} = V_{k,j} R_{k,j}^*$, linear dependence can occur as a result of the finite precision Lanczos process, as in the classical method, as well as from numerical rank deficiencies in the generated $s$-step polynomial bases $V_k$.

Given $A \in \mathbb{R}^{N \times N}$ and $B \in \mathbb{R}^{N' \times N'}$, $AE - EB = Z$ corresponds to the linear system with coefficient matrix $A \otimes I_{N'} - I_N \otimes B$. This system has a unique solution if and only if $\lambda(A) \cap \lambda(B) = \emptyset$, or, equivalently, if $\text{sep}(A, B) := \|(A \otimes I_{N'} - I_N \otimes B)^{-1}\|^{-1} > 0$, which depends on the spectral gap of $A$ and $B$ (see [9]).

THEOREM 5.1. Assume (3.23) holds, and let $\mu$ be a complex number such that $\text{sep}(A - \mu I, \tilde{T}_{k,j}) > 0$. Then

$$
\|\bar{v}_{k+1}\| \leq K_{k,j} \min_{\rho \in P_{sk+j+1}, \rho(0) = 1} \left( \|\rho(\tilde{T}_{k,j})\| + \|\rho(A - \mu I)\| \right) \|r_0\|,
$$
where

\[
K_{k,j} = \frac{\sqrt{sk + j + 1} \text{sep}(A - \mu I, \bar{T}_{k,j}) + |\mu| + \epsilon \|\Delta_{k,j}\|_F \cdot \max(1, \|\rho(A - \mu I)\| \cdot \|\rho(\bar{T}_{k,j})\|)}{\text{sep}(A - \mu I, \bar{T}_{k,j})}.
\]

Proof. By (3.23),

\[
(A - \mu I)Z_{k,j} = Z_{k,j} \bar{T}_{k,j} - \frac{1}{\alpha_{sk+j}} \frac{V_k r_{k,j+1}'}{\|r_0\|} e_{sk+j+1} + \epsilon \Delta_{k,j} - \mu Z_{k,j}.
\]

Then since \text{sep}(A - \mu I, \bar{T}_{k,j}) > 0, the equation

\[
(A - \mu I)E_{k,j} = E_{k,j} \bar{T}_{k,j} - \epsilon \Delta_{k,j} + \mu Z_{k,j}
\]

has a unique solution \(E_{k,j}\) with

\[
\|E_{k,j}\|_F \leq \frac{\|\epsilon \Delta_{k,j} + \mu Z_{k,j}\|_F}{\text{sep}(A - \mu I, \bar{T}_{k,j})} \leq \frac{\|\Delta_{k,j}\|_F + |\mu| \sqrt{sk + j + 1}}{\text{sep}(A - \mu I, \bar{T}_{k,j})}.
\]

Combining (5.1) and (5.2), we can write

\[
(A - \mu I)(Z_{k,j} + E_{k,j}) = (Z_{k,j} + E_{k,j}) \bar{T}_{k,j} - \frac{1}{\alpha_{sk+j}} \frac{V_k r_{k,j+1}'}{\|r_0\|} e_{sk+j+1}.
\]

Thus, for any \(\rho \in \mathbb{P}_{sk+j+1}\), \(\rho(0) = 1\), we have, by (4.2),

\[
\frac{V_k r_{k,j+1}'}{\|r_0\|} = \rho(A - \mu I)(Z_{k,j} + E_{k,j}) e_1 - (A - \mu I)(Z_{k,j} + E_{k,j}) \bar{T}_{k,j}^{-1} \rho(\bar{T}_{k,j}) e_1,
\]

and thus

\[
\frac{\|V_k r_{k,j+1}'\|}{\|r_0\|} \leq (\|Z_{k,j}\| + \|E_{k,j}\|) \|\rho(A - \mu I)\| + \|A - \mu I\| (\|Z_{k,j}\| + \|E_{k,j}\|) \|T_{k,j}^{-1}\| \|\rho(\bar{T}_{k,j})\|.
\]

Since

\[
\|Z_{k,j}\| + \|E_{k,j}\| \leq \sqrt{sk + j + 1} + \epsilon \|\Delta_{k,j}\|_F + |\mu| \sqrt{sk + j + 1} \text{sep}(A - \mu I, \bar{T}_{k,j}),
\]

we obtain the desired result. \(\Box\)

Note that in this case, if \(\mu\) is such that \text{sep}(A - \mu I, \bar{T}_{k,j}) is large, the quantity \(K_{k,j}\) depends heavily on \(\|T_{k,j}^{-1}\|\). The minimizing polynomial part of the bound now depends on both \(\rho(\bar{T}_{k,j})\) and \(\rho(A - \mu I)\).

6. Extensions: perturbation theory. We can think of (4.4) as an exact subspace relation for a perturbed \(A\), i.e., the quantities \(V_k, R'_{k,j}\), and \(T_{k,j}\) produced by the finite precision \(s\)-step BICG algorithm satisfy an exact subspace recurrence (4.4) for the perturbed system \(A + \delta A_{k,j}\). This means that the eigenvalues of the computed matrix \(T_{k,j}\) generated by the \(s\)-step algorithm are among the eigenvalues of the
perturbed matrix $A - \epsilon \Delta_{k,j} R_{k,j}^+ V_k^+$.

In the next theorem, we bound the distance of these eigenvalues to eigenvalues of unperturbed matrix $A$.

**Theorem 6.1.** Let $A$ be a normal $n \times n$ matrix of full rank. For each eigenvalue $\mu$ of the matrix $T_{k,j}$ computed by the finite precision $s$-step (BI)CG method, there exists an eigenvalue $\lambda$ of $A$ such that

$$|\gamma - \mu| \leq \epsilon \|\Delta_{k,j}\|_2 \|R_{k,j}^+\|_2 \|V_k^+\|_2$$

(6.1)

**Proof.** Note that $\bar{T}_{k,j} = D_{k,j} T_{k,j} D_{k,j}^{-1}$ has the same eigenvalues as $T_{k,j}$. By application of the Bauer-Fike theorem [2] to (4.4), there exists an eigenvalue of $\gamma$ of $A$ such that

$$|\gamma - \mu| \leq \epsilon \|\bar{\Delta}_{k,j} Z_{k,j}\|_2.$$  (6.2)

We can then write

$$\|\bar{\Delta}_{k,j} Z_{k,j}\|_2 = \|\Delta_{k,j} D_{k,j}^{-1} D_{k,j} R_{k,j}^+ V_k^+\|_2 \leq \|\Delta_{k,j}\|_2 \|R_{k,j}^+\|_2 \|V_k^+\|_2$$

(6.3)

(6.4)

The right hand side above can be shown to depend on $\kappa(V_k)$ and $\kappa(R_{k,j}^+)$. The above theorem means that the Lanczos vectors computed by the $s$-step (BI)CG algorithm, $V_{k} R'_{k,j}$, span Krylov spaces of a matrix within $\epsilon \|\Delta_{k,j} R_{k,j}^+ V_k^+\|$ of $A$. Similar observations have been made for classical finite precision Krylov methods [21, 29].

In [21], Paige shows that for classical Lanczos without reorthogonalization, the perturbed matrix is very close to $A$ until a Ritz value has stabilized. It is an open question whether a similar result (perhaps with additional restrictions on $V_k$) applies to the $s$-step case.

**REFERENCES**


