

## PREDICT-AND-RECOMPUTE CONJUGATE GRADIENT VARIANTS\*

TYLER CHEN<sup>†</sup> AND ERIN CARSON<sup>‡</sup>

**Abstract.** The standard implementation of the conjugate gradient algorithm suffers from communication bottlenecks on parallel architectures, due primarily to the two global reductions required every iteration. In this paper, we study conjugate gradient variants which decrease the runtime per iteration by overlapping global synchronizations, and in the case of pipelined variants, matrix-vector products. Through the use of a predict-and-recompute scheme, whereby recursively updated quantities are first used as a predictor for their true values and then recomputed exactly at a later point in the iteration, these variants are observed to have convergence behavior nearly as good as the standard conjugate gradient implementation on a variety of test problems. We provide a rounding error analysis which provides insight into this observation. It is also verified experimentally that the variants studied do indeed reduce the runtime per iteration in practice and that they scale similarly to previously studied communication-hiding variants. Finally, because these variants achieve good convergence without the use of any additional input parameters, they have the potential to be used in place of the standard conjugate gradient implementation in a range of applications.

**Key words.** Krylov subspace methods, conjugate gradient, parallel algorithms, numerical algorithms

**AMS subject classifications.** 65F10, 65N12, 65Y05

**DOI.** 10.1137/19M1276856

**1. Introduction and background.** The conjugate gradient method (CG), due to Hestenes and Stiefel [22], is a widely used method for solving a linear system of equations  $\mathbf{Ax} = \mathbf{b}$ , when  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is a large symmetric positive definite matrix. While the low storage costs and low number of floating point operations per iteration make CG an attractive choice for solving very large sparse systems, the dependency structure of the standard conjugate gradient algorithm given in [22], which we call HS-CG and present in Algorithm 1.1, requires that nearly every computation be done in sequence. In particular, it requires two inner products and one (typically sparse) matrix-vector product per iteration, none of which can occur simultaneously. On distributed memory parallel systems, this results in a communication bottleneck [13, 1, 2, 14]. This is because inner products require a costly global reduction involving communication between all nodes, and the matrix-vector product, even if sparse or structured, requires some level of communication between individual nodes.

To address this bottleneck, many *mathematically equivalent* variants of CG have been introduced; see, for instance, [36, 30, 5, 25, 31, 6, 35, 16, 15, 9, 10]. Broadly speaking, these variants rearrange the HS-CG algorithm such that the communication occurs less frequently or is overlapped with other computations. As a result, the time per iteration of these methods may be reduced on parallel machines in certain

\*Submitted to the journal's Methods and Algorithms for Scientific Computing section July 23, 2019; accepted for publication (in revised form) June 29, 2020; published electronically October 7, 2020.

<https://doi.org/10.1137/19M1276856>

**Funding:** The work of the first author was supported by the NSF through grant DGE-1762114. The work of the second author was supported by Charles University Primus program project PRIMUS/19/SCI/11.

<sup>†</sup>Department of Applied Mathematics, University of Washington, Seattle, WA 98195 USA (chentyl@uw.edu).

<sup>‡</sup>Faculty of Mathematics and Physics, Charles University, Prague, 13000, Czech Republic (carson@karlin.mff.cuni.cz).

settings. The algorithmic variants of CG presented in this paper are most closely related to the communication-hiding variants which we call M-CG, CG-CG, and GV-CG (pipelined CG), introduced by Meurant [25], Chronopoulos and Gear [6], and Ghysels and Vanroose [16], respectively. These communication-hiding variants maintain the iteration structure of HS-CG but add auxiliary vectors such that the computations within an iteration can be arranged so that expensive ones can occur simultaneously, effectively “hiding” the cost of communication. A comparison of these variants is given in Table 1 and full descriptions are given in Appendix D.

---

**Algorithm 1.1** Hestenes and Stiefel Conjugate Gradient (preconditioned)
 

---

```

1: procedure HS-CG(A, M, b, x0)
2:   INITIALIZE()
3:   for  $k = 1, 2, \dots$  do
4:      $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_{k-1} \mathbf{p}_{k-1}$ 
5:      $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{r}}_k = \mathbf{M}^{-1} \mathbf{r}_k$ 
6:      $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ 
7:      $\beta_k = \nu_k / \nu_{k-1}$ 
8:      $\mathbf{p}_k = \tilde{\mathbf{r}}_k + \beta_k \mathbf{p}_{k-1}$ 
9:      $\mathbf{s}_k = \mathbf{A} \mathbf{p}_k$ 
10:     $\mu_k = \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ 
11:     $\alpha_k = \nu_k / \mu_k$ 
12:   end for
13: end procedure
    
```

---

Notably, however, many communication-hiding variants suffer from numerical problems due to this rearranging of computations within each iteration. Indeed, it is well known that CG is particularly sensitive to rounding errors, and any modification to the HS-CG algorithm can have a significant effect on numerical behavior; see [19, 27] for summaries of CG and Lanczos in finite precision. Specifically, both the rate of convergence (the number of iterations to reach a given level of accuracy) and the maximal attainable accuracy of any algorithmic variant of CG may be *severely* impacted by carrying out computations in finite precision.

These effects are particularly pronounced in pipelined methods, such as GV-CG, because the additional auxiliary recurrences can cause rounding errors to be amplified [4]. For example, as shown in section 4, there are many problems for which the final accuracy attainable by GV-CG is orders of magnitude worse than that attainable by HS-CG. As a result, the practical use of some of the pipelined variants is potentially limited because the algorithms may fail to reach an acceptable level of accuracy, or require so many iterations to do so that there is no benefit to the overall runtime. While there have been many approaches to improving the numerical properties of these variants, such as residual replacement [16, 8] and the use of shifts in auxiliary recurrences [9], these strategies typically require certain parameters to be selected ahead of time based on user intuition or rely on heuristics for when and how to apply corrections. Moreover, as in the case of residual replacement, these strategies sometimes result in delayed convergence.

In this paper, we present a communication-hiding variant similar to M-CG which requires a single global synchronization per iteration. Like the M-CG algorithm, our variants employ a predict-and-recompute scheme which recursively computes quantities as a *predictor* for their true values and then *recomputes* them later in the iteration. We then introduce “pipelined” versions of both this variant and of M-CG which allow the computation and communication from the matrix-vector product and

preconditioning steps to be overlapped with the global reduction associated with the inner products. We demonstrate numerically that the convergence behavior of our pipelined variants is comparable to that of HS-CG despite overlapping all matrix products with global reductions as in other pipelined variants such as GV-CG. This observation is complemented by a rounding error analysis, which provides new insight into the way the predict-and-recompute schemes can improve the numerical behavior. More broadly, this work demonstrates that it is reasonable to add extra computation to improve the numerical properties of a communication-hiding CG variant, provided that the computation is done in a way which does not affect the communication pattern of the algorithm. All of the algorithms introduced in this paper require exactly the same inputs as HS-CG and therefore require no additional tuning by the end user.

Our primary contributions are as follows. First, we provide rounding error analyses of the variants introduced in this paper. In particular, we derive expressions for the residual gap, which is the quantity that dictates the maximum attainable accuracy, as well as for the error in the three-term Lanczos recurrence, which gives some indication of the rate of convergence. Our analysis provides rigorous support for the conjecture made by Meurant in [25] that the predict-and-recompute scheme resolves a potential source of instability. In addition, we provide a range of numerical experiments to further support the claim that our pipelined variants can significantly improve the rate of convergence and ultimately attainable accuracy versus existing pipelined variants. In particular, we note that in the numerical experiments in section 4, our variants appear to converge similarly to HS-CG without any additional input parameters. As such, they have the potential to be used as black box solvers wherever HS-CG is used. Finally, we demonstrate through a strong scaling experiment that the new variants can reduce the time per iteration in a parallel setting.

Unless otherwise stated, matrices should be assumed to be of size  $n \times n$  and vectors of size  $n \times 1$ . The transpose of a matrix is denoted by the superscript  $\mathsf{T}$ , and the inverse of the transpose denoted by the superscript  $-\mathsf{T}$ . The standard Euclidean inner product and corresponding spectral/operator norm are denoted, respectively, by  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$ , and the inner product and norm induced by a positive definite matrix  $\mathbf{B}$  are denoted by  $\langle \cdot, \cdot \rangle_{\mathbf{B}}$  and  $\| \cdot \|_{\mathbf{B}}$ . While much of the theory about the conjugate gradient algorithm applies to complex systems, we consider real systems for convenience.

The variants studied in this paper are all compatible with a preconditioner  $\mathbf{M}^{-1} = \mathbf{R}^{-\mathsf{T}} \mathbf{R}^{-1}$ , which allows the algorithms to implicitly solve the system  $\mathbf{R}^{-\mathsf{T}} \mathbf{A} \mathbf{R}^{-1} \mathbf{y} = \mathbf{R}^{-\mathsf{T}} \mathbf{b}$  and then set  $\mathbf{x} = \mathbf{R}^{-1} \mathbf{y}$ , using only products with  $\mathbf{M}^{-1}$ ; i.e., the  $\mathbf{R}^{-1}$  factors *need not be known*. As summarized in Table 1, the preconditioned variants require some additional memory to store the preconditioned residual and any related auxiliary vectors, and updating these additional vectors has the potential to introduce additional rounding errors. Throughout this paper, we use a tilde (“ $\sim$ ”) above a vector to indicate that, in exact arithmetic, the tilde vector is equal to the preconditioner applied to the nontilde vector; i.e.,  $\tilde{\mathbf{r}}_k = \mathbf{M}^{-1} \mathbf{r}_k$ ,  $\tilde{\mathbf{s}}_k = \mathbf{M}^{-1} \mathbf{s}_k$ , etc.

**2. Derivation of new variants.** In this section we describe PR-CG, a new communication-hiding variant which requires only one global synchronization point per iteration. This variant is similar to M-CG, introduced by Meurant in [25], and the relationship between the two algorithms is discussed.

Then, in the same way that GV-CG is obtained from CG-CG, we “pipeline” PR-CG to overlap the matrix-vector product with the inner products. The order in which operations are done in the pipelined version of PR-CG allows for a vector quantity to be recomputed using an additional matrix-vector product, giving the pipelined

TABLE 1

Summary of costs for various conjugate gradient variants. Values in parenthesis are the additional costs for the preconditioned variants. The quantity Mem. gives the number of vectors stored, Vec. gives the number of vector updates (AXPYs) per iteration, Scal. gives the number of inner products per iteration, and Time gives the dominant costs (ignoring vector updates and inner products).  $C_{gr}$  is the time spent on communication for a global reduction.  $T_{mv}$  and  $C_{mv}$  are the times spent on computation and communication, respectively, for a matrix-vector product which depend on the method of matrix multiplication (for instance, a dense matrix has  $C_{mv} = C_{gr}$ ).  $T_{2mv}$  is the cost of computing two matrix-vector products simultaneously, which is less than  $2T_{mv}$  if implemented in an efficient way. Note that in this abstraction we assume that the communication time is independent of the size of messages sent.

Variant	Mem.	Vec.	Scal.	Time
HS-CG	4 (+1)	3 (+0)	2	$2C_{gr} + T_{mv} + C_{mv}$
CG-CG	5 (+1)	4 (+0)	2	$C_{gr} + T_{mv} + C_{mv}$
M-CG	4 (+2)	3 (+1)	3	$C_{gr} + T_{mv} + C_{mv}$
PR-CG	4 (+2)	3 (+1)	4	$C_{gr} + T_{mv} + C_{mv}$
GV-CG	7 (+3)	6 (+2)	2	$\max(C_{gr}, T_{mv} + C_{mv})$
PIPE-M-CG	6 (+4)	5 (+3)	3	$\max(C_{gr}, T_{2mv} + C_{mv})$
PIPE-PR-CG	6 (+4)	5 (+3)	4	$\max(C_{gr}, T_{2mv} + C_{mv})$

predict-and-recompute variant PIPE-PR-CG. Since this matrix-vector product can occur at the same time as the other matrix-vector product and as the inner products, the communication costs per iteration are not increased. For comparison, we also derive a pipelined version of Meurant's M-CG method, which we call PIPE-M-CG.

Table 1 provides a comparison between some commonly used communication-avoiding variants and the newly introduced variants. It should be noted that although the numbers of matrix-vector products and inner products in PIPE-M-CG and PIPE-PR-CG are increased, most of this work can be done locally, and they have the same dominant communication costs as GV-CG.

**2.1. A simple communication-hiding variant.** Like the derivation of M-CG in [25] and the variants introduced in [23, 36, 30, 31], we derive PR-CG by substituting recurrences into the inner product  $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ . This allows us to obtain an equivalent expression for the inner product involving quantities which are known earlier in the iteration.

To this end, we first recall that given  $\tilde{\mathbf{s}}_k = \mathbf{M}^{-1}\mathbf{s}_k$ , we have

$$(2.1) \quad \tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_{k-1} - \alpha_{k-1}\tilde{\mathbf{s}}_{k-1}.$$

Then, by substituting the recurrences for  $\tilde{\mathbf{r}}_k$  and  $\mathbf{r}_k$  from Algorithm 1.1 into  $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ , we can write

$$(2.2) \quad \begin{aligned} \nu_k &= \langle \tilde{\mathbf{r}}_{k-1} - \alpha_{k-1}\tilde{\mathbf{s}}_{k-1}, \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1} \rangle \\ &= \langle \tilde{\mathbf{r}}_{k-1}, \mathbf{r}_{k-1} \rangle - \alpha_{k-1}\langle \tilde{\mathbf{r}}_{k-1}, \mathbf{s}_{k-1} \rangle - \alpha_{k-1}\langle \tilde{\mathbf{s}}_{k-1}, \mathbf{r}_{k-1} \rangle + \alpha_{k-1}^2\langle \tilde{\mathbf{s}}_{k-1}, \mathbf{s}_{k-1} \rangle. \end{aligned}$$

Since  $\mathbf{M}$ , and therefore  $\mathbf{M}^{-1}$ , are symmetric,  $\langle \tilde{\mathbf{s}}_{k-1}, \mathbf{r}_{k-1} \rangle = \langle \mathbf{s}_{k-1}, \tilde{\mathbf{r}}_{k-1} \rangle$ . Thus,

$$(2.3) \quad \nu_k = \nu_{k-1} - 2\alpha_{k-1}\langle \tilde{\mathbf{r}}_{k-1}, \mathbf{s}_{k-1} \rangle + \alpha_{k-1}^2\langle \tilde{\mathbf{s}}_{k-1}, \mathbf{s}_{k-1} \rangle.$$

Once  $\mathbf{s}_k = \mathbf{A}\mathbf{p}_k$  and  $\tilde{\mathbf{s}}_k = \mathbf{M}^{-1}\mathbf{s}_k$  have been computed, we can simultaneously compute the three inner products

$$(2.4) \quad \mu_k = \langle \mathbf{p}_k, \mathbf{s}_k \rangle, \quad \sigma_k = \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle, \quad \gamma_k = \langle \tilde{\mathbf{s}}_k, \mathbf{s}_k \rangle.$$

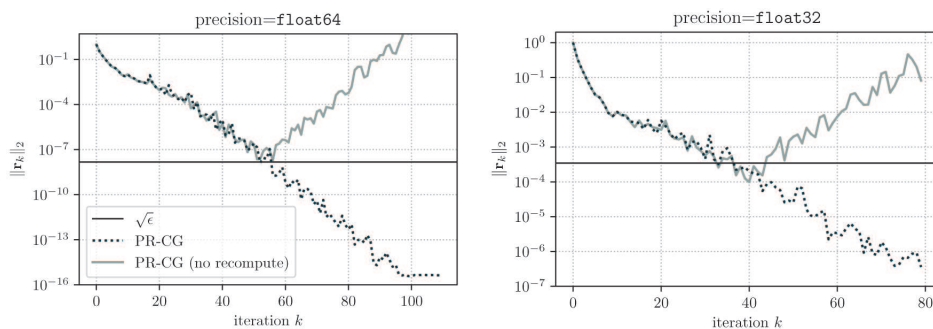


FIG. 1. Demonstration, on *model\_48-8-3* with no preconditioner, of breakdown of PR-CG when  $\nu'_k$  is not recomputed. Breakdown occurs when  $\nu'_k$  becomes negative, which can be avoided using the recompute scheme.

A variant using a similar expression for  $\nu_k$  was suggested in [24] and briefly mentioned in [4]. However, one term of their formula for  $\nu_k$  has a sign difference from (2.3), and no numerical tests or rounding error analysis were provided.

As shown in Figure 1, if (2.3) is used to recursively update  $\nu_k$ , the final accuracy can be severely impacted. This phenomenon was observed previously in [36, 23, 30], all of which study variants which use expressions for  $\nu_k$  similar to that in M-CG. This catastrophic loss of accuracy is caused by the updated value of  $\nu_k$  becoming negative. In [25], Meurant suggests using the recursively updated value of  $\nu_k$  as a *predictor* for the true value in order to update any vectors required for the algorithm to proceed, and then to *recompute*  $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$  at the same time as the other inner products. We observe experimentally that using this strategy effectively brings the ultimately attainable accuracy to a level similar to that of HS-CG. Our rounding error analysis in section 3 explains this improvement. This algorithm, denoted PR-CG, is given in Algorithm 2.1. Note that we use a prime (“’”) to distinguish the recursively updated quantity  $\nu'_k$  from the explicitly computed quantity  $\nu_k$ .

---

**Algorithm 2.1** Predict-and-Recompute Conjugate Gradient

---

```

1: procedure PR-CG(A, M, b, x0)
2:   INITIALIZE()
3:   for  $k = 1, 2, \dots$  do
4:      $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_{k-1} \mathbf{p}_{k-1}$ 
5:      $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_{k-1} - \alpha_{k-1} \tilde{\mathbf{s}}_{k-1}$ 
6:      $\nu'_k = \nu_{k-1} - 2\alpha_{k-1}\sigma_{k-1} + \alpha_{k-1}^2\gamma_{k-1}$ 
7:      $\beta_k = \nu'_k / \nu_{k-1}$ 
8:      $\mathbf{p}_k = \mathbf{r}_k + \beta_k \mathbf{p}_{k-1}$ 
9:      $\mathbf{s}_k = \mathbf{A} \mathbf{p}_k$ ,  $\tilde{\mathbf{s}}_k = \mathbf{M}^{-1} \mathbf{s}_k$ 
10:     $\mu_k = \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ ,  $\sigma_k = \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle$ ,  $\gamma_k = \langle \tilde{\mathbf{s}}_k, \mathbf{s}_k \rangle$ ,  $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ 
11:     $\alpha_k = \nu_k / \mu_k$ 
12:   end for
13: end procedure

```

---

It is not hard to see that (2.3) can be simplified further. Indeed, observing that  $\langle \tilde{\mathbf{s}}_k, \mathbf{r}_k \rangle = \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ , which follows from the **A**-orthogonality of  $\mathbf{p}_k$  and  $\mathbf{p}_{k-1}$ , and using the expression  $\alpha_k = \nu_k / \mu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle / \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ , we can write

$$\nu_k = -\langle \tilde{\mathbf{r}}_{k-1}, \mathbf{r}_{k-1} \rangle + \alpha_k^2 \langle \tilde{\mathbf{s}}_{k-1}, \mathbf{s}_{k-1} \rangle = -\nu_{k-1} + \alpha_k^2 \langle \tilde{\mathbf{s}}_{k-1}, \mathbf{s}_{k-1} \rangle.$$

This is identical to the expressions studied in [25] and used in M-CG; i.e., M-CG can be obtained by replacing line 6 of PR-CG with  $\nu'_k = -\nu_k + \alpha_k^2 \gamma_k$ . A similar expression,  $\nu_k = -\alpha_k \mu_k + \alpha_k^2 \gamma_k$ , was studied in [35].

The natural question is then, what is the advantage of the expression for  $\nu_k$  used in PR-CG over the expression used in M-CG? The incomplete response is that experimentally it seems to work a bit better, especially in the pipelined versions of these variants. Intuitively, this is because the less simplification that occurs, the closer the variant is to HS-CG. Of course then it is reasonable to wonder if we should use (2.3) instead of (2.2). The again incomplete response is that this change doesn't seem to significantly influence the numerical behavior. This is perhaps not particularly surprising; for instance, in the case of the unpreconditioned variants, (2.2) and (2.3) are equivalent even in finite precision. An extended discussion on this topic is included in Appendix C; a full theoretical explanation remains for future work.

**2.2. New pipelined variants.** Recall that our goal is to be able to compute the matrix-vector product and inner products simultaneously. To this end, note that in both M-CG and PR-CG we have the recurrence  $\mathbf{p}_k = \tilde{\mathbf{r}}_k + \beta_k \mathbf{p}_{k-1}$ . Thus, defining  $\mathbf{w}_k = \mathbf{A}\tilde{\mathbf{r}}_k$ , we can write

$$(2.5) \quad \mathbf{s}_k = \mathbf{A}\mathbf{p}_k = \mathbf{A}\tilde{\mathbf{r}}_k + \beta_k \mathbf{A}\mathbf{p}_{k-1} = \mathbf{w}_k + \beta_k \mathbf{s}_{k-1}.$$

Similarly, defining  $\mathbf{u}_k = \mathbf{A}\tilde{\mathbf{s}}_k$ ,

$$(2.6) \quad \mathbf{w}_k = \mathbf{A}\tilde{\mathbf{r}}_k = \mathbf{A}\tilde{\mathbf{r}}_{k-1} - \alpha_{k-1} \mathbf{A}\tilde{\mathbf{s}}_{k-1} = \mathbf{w}_{k-1} - \alpha_{k-1} \mathbf{u}_{k-1}.$$

Using these recurrences allows us to compute the product  $\mathbf{u}_k = \mathbf{A}\tilde{\mathbf{s}}_k$  at the same time as all of the inner products.

To move the preconditioning step, we define  $\tilde{\mathbf{w}}_k = \mathbf{M}^{-1} \mathbf{w}_k$  so that

$$(2.7) \quad \tilde{\mathbf{s}}_k = \mathbf{M}^{-1} \mathbf{s}_k = \mathbf{M}^{-1} \mathbf{w}_k + \beta_k \mathbf{M}^{-1} \mathbf{s}_{k-1} = \tilde{\mathbf{w}}_k + \beta_k \tilde{\mathbf{s}}_{k-1},$$

and we define  $\tilde{\mathbf{u}}_k = \mathbf{M}^{-1} \mathbf{u}_k$  so that

$$(2.8) \quad \tilde{\mathbf{w}}_k = \mathbf{M}^{-1} \mathbf{w}_k = \mathbf{M}^{-1} \mathbf{w}_{k-1} - \alpha_{k-1} \mathbf{M}^{-1} \mathbf{u}_{k-1} = \tilde{\mathbf{w}}_{k-1} - \alpha_{k-1} \tilde{\mathbf{u}}_{k-1}.$$

The recurrences above could be used to implement a pipelined variant which, in each iteration, overlaps the matrix product and preconditioning steps with the global reduction. However, like GV-CG, this variant suffers from delayed convergence and reduced final accuracy compared to HS-CG, M-CG, and PR-CG. To address this, we observe that  $\mathbf{w}_k = \mathbf{A}\tilde{\mathbf{r}}_k$  and  $\tilde{\mathbf{w}}_k = \mathbf{M}^{-1} \mathbf{w}_k$  can be *recomputed* at the same time as the other matrix-vector product and all inner products are being computed. Thus, in the same way we use the recursively updated value of  $\nu_k$  as a predictor for the true value, we can use the recursively updated value of  $\mathbf{w}_k$  as a predictor for the true value in order to update other vector quantities, and then update the value of  $\mathbf{w}_k$  later in the iteration. Using this predict-and-recompute approach gives PIPE-M-CG and PIPE-PR-CG.

Algorithm 2.2 shows PIPE-PR-CG, from which PIPE-M-CG can be obtained by using the alternate expression  $\nu'_k = -\nu_{k-1} + \alpha_{k-1}^2 \gamma_{k-1}$  in line 7. Using this expression means that  $\sigma_k$  need not be computed. As before, we use a prime to denote predicted quantities.

**Algorithm 2.2** Pipelined Predict-and-Recompute Conjugate Gradient

---

```

1: procedure PIPE-PR-CG( $\mathbf{A}$ ,  $\mathbf{M}$ ,  $\mathbf{b}$ ,  $\mathbf{x}_0$ )
2:   INITIALIZE()
3:   for  $k = 1, 2, \dots$  do
4:      $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1}$ 
5:      $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_{k-1} - \alpha_{k-1}\tilde{\mathbf{s}}_{k-1}$ 
6:      $\mathbf{w}'_k = \mathbf{w}_{k-1} - \alpha_{k-1}\mathbf{u}_{k-1}$ ,  $\tilde{\mathbf{w}}'_k = \tilde{\mathbf{w}}_{k-1} - \alpha_{k-1}\tilde{\mathbf{u}}_{k-1}$ 
7:      $\nu'_k = \nu_{k-1} - 2\alpha_{k-1}\sigma_{k-1} + \alpha_{k-1}^2\gamma_{k-1}$ 
8:      $\beta_k = \nu'_k/\nu_{k-1}$ 
9:      $\mathbf{p}_k = \tilde{\mathbf{r}}_k + \beta_k\mathbf{p}_{k-1}$ 
10:     $\mathbf{s}_k = \mathbf{w}'_k + \beta_k\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{s}}_k = \tilde{\mathbf{w}}'_k + \beta_k\tilde{\mathbf{s}}_{k-1}$ 
11:     $\mathbf{u}_k = \mathbf{A}\tilde{\mathbf{s}}_k$ ,  $\tilde{\mathbf{u}}_k = \mathbf{M}^{-1}\mathbf{u}_k$ 
12:     $\mathbf{w}_k = \mathbf{A}\tilde{\mathbf{r}}_k$ ,  $\tilde{\mathbf{w}}_k = \mathbf{M}^{-1}\mathbf{w}_k$ 
13:     $\mu_k = \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ ,  $\sigma_k = \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle$ ,  $\gamma_k = \langle \tilde{\mathbf{s}}_k, \mathbf{s}_k \rangle$ ,  $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ 
14:     $\alpha_k = \nu_k/\mu_k$ 
15:  end for
16: end procedure

```

---

**2.2.1. Implementation.** The presentation of PIPE-PR-CG in Algorithm 2.2 is intended to match the derivation from HS-CG and to emphasize the mathematical equivalence of the two algorithms. However, as with any parallel algorithm, some care must be taken at implementation time as an inefficient implementation may actually increase the runtime per iteration.

We suggest computing the scalars  $\alpha_{k-1}$ ,  $\nu'_k$ , and  $\beta_k$  (lines 14, 7, 8) at the beginning of each iteration. This will allow all vector updates (lines 4, 5, 6, 9, 10) to occur simultaneously. The vector updates require only local on-node communication and are therefore assumed to be very fast. Finally, the matrix-vector products/preconditioning (lines 11, 12) and inner products (line 13) can all be computed simultaneously. As a result, the dominant cost per iteration will be the time for the global reduction associated either with the inner products or with the matrix-vector products, thus giving the runtime  $\max(C_{\text{gr}}, T_{2\text{mv}} + C_{\text{mv}})$  as listed in Table 1.

The matrix-vector products (and preconditioning) in lines 11 and 12 can be computed together using efficient kernels. In particular, this means that within a single node, PIPE-PR-CG still requires only one pass over  $\mathbf{A}$  (and  $\mathbf{M}^{-1}$ ) in each iteration. This is an especially important consideration if  $\mathbf{A}$  is too large to store in fast memory. Similarly, three of the inner products involve  $\mathbf{s}_k$ , so the number of passes over  $\mathbf{s}_k$  can be reduced from three to one. However, this is likely not to have a noticeable effect until the cost of reading  $\mathbf{s}_k$  from memory becomes large compared to the reduction time. Finally, there is no need to store  $\mathbf{w}_k$  and  $\mathbf{w}'_k$  as separate vectors.

**3. Rounding error analysis.** We give a rounding error analysis which provides some insight into how predict-and-recompute schemes may lead to the improved maximal accuracy and convergence behavior observed on test problems.

The maximal attainable accuracy of a CG algorithm in finite precision is typically analyzed in terms of the residual gap  $\Delta_{\mathbf{r}_k} := (\mathbf{b} - \mathbf{A}\mathbf{x}_k) - \mathbf{r}_k$  [18, 32], an expression introduced by Greenbaum in [17, Theorem 2]. For many variants, such as HS-CG and those introduced in this paper, it is observed experimentally that the norm of the updated residual  $\mathbf{r}_k$  decreases to much lower than the machine precision. As a result, the size of the residual gap  $\Delta_{\mathbf{r}_k}$  can be used to estimate the size of the smallest true residual which can be attained in finite precision, thereby giving an estimate of the accuracy of the iterate  $\mathbf{x}_k$ . Similar analyses have been done for a three-term CG

variant [21], as well as for CG-CG, GV-CG, and other pipelined CG variants [8, 4]. However, for some variants such as GV-CG, the updated residual  $\mathbf{r}_k$  may not decrease to well below machine precision, so some care must be taken when interpreting such results.

There is also existing theory about the rate of convergence of a CG implementation in finite precision due to Greenbaum [17], which extends the work of Paige [28, 29]. This analysis applies to a perturbed Lanczos recurrence, which, in effect, shows that the error norms of a CG algorithm run in finite precision for  $k$  steps correspond to the error norms of exact arithmetic CG applied to a larger matrix whose eigenvalues lie in small intervals about the eigenvalues of  $\mathbf{A}$ , provided that the updated residuals  $\mathbf{r}_k$  satisfy certain conditions. This provides a method for applying the well-understood theory about the convergence of CG in exact arithmetic to the finite precision setting.

The conditions required for the analysis in [17] are that (i) successive updated residuals are approximately orthogonal, i.e.,  $\langle \mathbf{r}_k, \mathbf{r}_{k-1} \rangle \approx 0$ , and (ii) that they approximately satisfy the three-term Lanczos recurrence, i.e.,

$$(3.1) \quad \mathbf{A}\mathbf{q}_k \approx \frac{1}{\alpha_{k-1}} \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_{k-1}\|} \mathbf{q}_{k+1} + \left( \frac{1}{\alpha_{k-1}} + \frac{\beta_{k-1}}{\alpha_{k-2}} \right) \mathbf{q}_k + \frac{1}{\alpha_{k-2}} \frac{\|\mathbf{r}_{k-1}\|}{\|\mathbf{r}_{k-2}\|} \mathbf{q}_{k-1},$$

where  $\alpha_k$  and  $\beta_k$  are computed in finite precision and  $\mathbf{q}_{k+1} := (-1)^k \mathbf{r}_k / \|\mathbf{r}_k\|$  is obtained from the updated residuals.

If these conditions are satisfied, then for some small  $\eta$  which depends on the machine precision and on the technical definition of “approximately” (see [17] for details), the error norms will satisfy the relaxed minimax bound<sup>1</sup>

$$(3.2) \quad \frac{\|\mathbf{e}_k\|_{\mathbf{A}}}{\|\mathbf{e}_0\|_{\mathbf{A}}} \leq \min_{p \in \mathcal{P}_k} \left[ \max_{z \in \mathcal{L}(\mathbf{A})} |p(z)| \right], \quad \mathcal{L}(\mathbf{A}) = \bigcup_{i=1}^n [\lambda_i - \eta, \lambda_i + \eta].$$

The degree to which these conditions are satisfied in finite precision directly impacts the size of  $\eta$ , with better approximations yielding smaller  $\eta$  and therefore stronger bounds. In fact, if the conditions are exactly satisfied, then the analysis will yield  $\eta = 0$  and (3.2) becomes the well-known minimax bound for exact arithmetic CG.

However, it remains to be proved that *any* of the variants discussed in this paper satisfy both of the conditions of [17] in a meaningful way. Experimentally, it seems to be the case that these variants do keep successive residuals approximately orthogonal (this has been essentially proved for HS-CG [26, Proposition 5.19]), and that, with the exception of GV-CG, they approximately satisfy the three-term Lanczos recurrence. In [20], expressions for the degree to which the updated residuals from HS-CG, CG-CG, and GV-CG satisfy (3.1) in finite precision are given in terms of roundoff errors, and it is shown that while HS-CG and CG-CG satisfy the three-term recurrence to within local rounding errors, GV-CG does not. While this analysis does not prove the degree to which any of the variants satisfy the conditions of [17], it does provide some indication that the size of  $\eta$  for HS-CG and CG-CG will likely be smaller than for GV-CG on most problems.

Experimentally, on many problems the rate of convergence of all variants is roughly the same prior to the stagnation of the true residuals. However, on other

<sup>1</sup>Greenbaum’s theory actually applies to  $\|\mathbf{A}^{-1/2}\mathbf{r}_k\|$ , which in exact arithmetic is equal to the  $\mathbf{A}$ -norm of the error  $\|\mathbf{e}_k\|_{\mathbf{A}} = \|\mathbf{A}^{-1}\mathbf{b} - \mathbf{x}_k\|_{\mathbf{A}} = \|\mathbf{A}^{-1/2}(\mathbf{b} - \mathbf{A}\mathbf{x}_k)\|$ . Thus, the relaxed minimax bound (3.2) only applies before the updated and true residuals begin to differ significantly, i.e., when the residual gap  $\Delta_{\mathbf{r}_k}$  is still small.



problems, the rate of convergence of different variants is observed to differ significantly. Both cases are shown in the experiments in section 4; see, e.g., Figure 3. In [20] it is suggested that the first case corresponds to problems where the size of  $\eta$  has little effect on the strength of the relaxed minimax bound (3.2) (such as those with a relatively uniformly distributed spectrum), while the second case corresponds to problems where the size of  $\eta$  has a large effect on the strength of this bound (such as those with large gaps in the upper spectrum). Thus, on the second type of problem where the size of  $\eta$  is important, the fact that GV-CG does not do a good job of satisfying the three-term Lanczos recurrence (3.1) means that the relaxed minimax bound will be stronger for HS-CG and CG-CG than for GV-CG. Again, while this analysis does not prove that HS-CG and CG-CG will have better rates of convergence on such problems, it is evidence supporting the observation that they do have better rates of convergence in practice.

Throughout the analysis we assume the standard model of floating point arithmetic, i.e.,

$$(3.3) \quad |\text{fp}(\alpha \circ \beta) - \alpha \circ \beta| \leq \epsilon |\alpha \circ \beta|$$

for real numbers  $\alpha$  and  $\beta$  with standard operations  $\circ \in \{+, -, \times, \div\}$ , where  $\epsilon$  is the unit roundoff of the machine and all computations inside the  $\text{fp}(\cdot)$  are performed in finite precision. From this, to first order we have the bounds

$$(3.4) \quad \|\text{fp}(\mathbf{x} + \alpha \mathbf{y}) - (\mathbf{x} + \alpha \mathbf{y})\| \leq \epsilon (\|\mathbf{x}\| + 2|\alpha|\|\mathbf{y}\|),$$

$$(3.5) \quad \|\text{fp}(\mathbf{A}\mathbf{x}) - \mathbf{A}\mathbf{x}\| \leq \epsilon c \|\mathbf{A}\|\|\mathbf{x}\|,$$

$$(3.6) \quad \|\text{fp}(\langle \mathbf{x}, \mathbf{y} \rangle) - \langle \mathbf{x}, \mathbf{y} \rangle\| \leq \epsilon n \|\mathbf{x}\|\|\mathbf{y}\|,$$

where  $n$  is the length of the vectors  $\mathbf{x}$  and  $\mathbf{y}$  and  $c$  is a constant depending on sparsity/structure of  $\mathbf{A}$  and the method of matrix multiplication used; for instance, it is common to take  $c = mn^{1/2}$ , where  $m$  is the maximum number of nonzero entries in a row of  $\mathbf{A}$ .

**3.1. Predict-and-recompute CG.** In finite precision, PR-CG generates the recurrences

$$(3.7) \quad \begin{aligned} \mathbf{x}_k &= \mathbf{x}_{k-1} + \alpha_{k-1} \mathbf{p}_{k-1} + \delta_{\mathbf{p}_k}, & \mathbf{r}_k &= \mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{s}_{k-1} + \delta_{\mathbf{r}_k}, \\ \nu'_k &= \nu_{k-1} - 2\alpha_{k-1}\sigma_{k-1} + \alpha_{k-1}^2\gamma_{k-1} + \delta_{\nu'_k}, & \beta_k &= \nu'_k/\nu_{k-1} + \delta_{\beta_k}, \\ \mathbf{p}_k &= \mathbf{r}_k + \beta_k \mathbf{p}_{k-1} + \delta_{\mathbf{p}_k}, & \mathbf{s}_k &= \mathbf{A}\mathbf{p}_k + \delta_{\mathbf{s}_k}, \\ \nu_k &= \langle \mathbf{r}_k, \mathbf{r}_k \rangle + \delta_{\nu_k}, & \sigma_k &= \langle \mathbf{r}_k, \mathbf{s}_k \rangle + \delta_{\sigma_k}, \\ \gamma_k &= \langle \mathbf{s}_k, \mathbf{s}_k \rangle + \delta_{\gamma_k}, \end{aligned}$$

where the  $\delta_{\text{quantity}}$  terms represent roundoff errors incurred at each step and are bounded by the appropriate application of (3.3)–(3.6).

It was observed by Meurant [25] that the instability caused by recursive computation of  $\nu_k$  via the formula (2.3) is due to the value of  $\nu_k$  eventually becoming negative. Our methods, like those in [36, 30, 25], also break down if  $\nu_k$  becomes negative. Meurant conjectures in [25] that using the predict-and-recompute approach appears to solve this problem, but he does not offer a rigorous theoretical explanation of this behavior.

We begin this section by filling in this gap. We first show that the predict-and-recompute scheme suggested in [25] and used in M-CG, PR-CG, PIPE-M-CG, and

PIPE-PR-CG keeps the estimate for  $\nu'_k$  to within local rounding errors of the true value  $\langle \mathbf{r}_k, \mathbf{r}_k \rangle$ . To this end, we define the  $\nu'$ -gap as  $\Delta_{\nu'_k} := \langle \mathbf{r}_k, \mathbf{r}_k \rangle - \nu'_k$ . Then

$$\begin{aligned}
 \Delta_{\nu'_k} &= \langle \mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{s}_{k-1} + \delta_{\mathbf{r}_k}, \mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{s}_{k-1} + \delta_{\mathbf{r}_k} \rangle - \nu'_k \\
 &= \langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle - 2\alpha_{k-1} \langle \mathbf{r}_{k-1}, \mathbf{s}_{k-1} \rangle + \alpha_{k-1}^2 \langle \mathbf{s}_{k-1}, \mathbf{s}_{k-1} \rangle \\
 &\quad + 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{s}_{k-1} \rangle + \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \nu'_k \\
 &= \nu_{k-1} - \delta_{\nu_{k-1}} - 2\alpha_{k-1}(\sigma_{k-1} - \delta_{\sigma_{k-1}}) + \alpha_{k-1}^2(\gamma_{k-1} - \delta_{\gamma_{k-1}}) \\
 &\quad + 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_k \rangle - \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \nu'_k \\
 &= (\nu_{k-1} - 2\alpha_{k-1}\sigma_{k-1} + \alpha_{k-1}^2\gamma_{k-1}) - (\delta_{\nu_{k-1}} - 2\alpha_{k-1}\delta_{\sigma_{k-1}} + \alpha_{k-1}^2\delta_{\gamma_{k-1}}) \\
 &\quad + 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_k \rangle - \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \nu'_k \\
 &= \nu'_k - \delta_{\nu'_k} - (\delta_{\nu_{k-1}} - 2\alpha_{k-1}\delta_{\sigma_{k-1}} + \alpha_{k-1}^2\delta_{\gamma_{k-1}}) + 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_k \rangle - \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \nu'_k \\
 (3.8) \quad &= 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_k \rangle - \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \delta_{\nu'_k} - (\delta_{\nu_{k-1}} - 2\alpha_{k-1}\delta_{\sigma_{k-1}} + \alpha_{k-1}^2\delta_{\gamma_{k-1}}).
 \end{aligned}$$

Applying (3.4)–(3.6) we have the bound

$$(3.9) \quad |\Delta_{\nu'_k}| \leq 2\|\delta_{\mathbf{r}_k}\| \|\mathbf{r}_k\| + \|\delta_{\mathbf{r}_k}\|^2 + |\delta_{\nu'_k}| + |\delta_{\nu_{k-1}}| + 2|\alpha_{k-1}||\delta_{\sigma_{k-1}}| + |\alpha_{k-1}|^2|\delta_{\gamma_{k-1}}|.$$

We will now bound the terms on the right-hand side of (3.9) using standard analysis techniques. We can rearrange the expression for  $\mathbf{r}_k$  in (3.7) to obtain

$$\mathbf{s}_{k-1} = \frac{1}{|\alpha_{k-1}|}(\mathbf{r}_{k-1} - \mathbf{r}_k + \delta_{\mathbf{r}_k}),$$

which gives the bound

$$(3.10) \quad \|\mathbf{s}_{k-1}\| \leq \frac{1}{|\alpha_{k-1}|} (\|\mathbf{r}_{k-1}\| + \|\mathbf{r}_k\| + \|\delta_{\mathbf{r}_k}\|).$$

We now seek to bound the norms of the quantities  $\delta_{\mathbf{r}_k}$ ,  $\delta_{\nu_{k-1}}$ ,  $\delta_{\sigma_{k-1}}$ ,  $\delta_{\gamma_{k-1}}$ , and  $\delta_{\nu'_k}$ . Using (3.7), (3.10), the bounds (3.4)–(3.6), and dropping terms of order  $O(\epsilon^2)$ , we have

$$\begin{aligned}
 \|\delta_{\mathbf{r}_k}\| &\leq \epsilon (\|\mathbf{r}_k\| + 2|\alpha_{k-1}|\|\mathbf{s}_{k-1}\|) \\
 &\leq \epsilon \left( \|\mathbf{r}_k\| + 2|\alpha_{k-1}| \left( \frac{1}{|\alpha_{k-1}|} (\|\mathbf{r}_{k-1}\| + \|\mathbf{r}_k\| + \|\delta_{\mathbf{r}_k}\|) \right) \right) \\
 &\leq \epsilon (3\|\mathbf{r}_k\| + 2\|\mathbf{r}_{k-1}\|) + 2\epsilon \|\delta_{\mathbf{r}_k}\| \\
 (3.11) \quad &\leq 3\epsilon (\|\mathbf{r}_k\| + \|\mathbf{r}_{k-1}\|),
 \end{aligned}$$

$$(3.12) \quad |\delta_{\nu_{k-1}}| \leq \epsilon n \|\mathbf{r}_{k-1}\|^2,$$

$$\begin{aligned}
 |\delta_{\sigma_{k-1}}| &\leq \epsilon n \|\mathbf{r}_{k-1}\| \|\mathbf{s}_{k-1}\| \\
 &\leq \epsilon n \|\mathbf{r}_{k-1}\| \left( \frac{1}{|\alpha_{k-1}|} (\|\mathbf{r}_{k-1}\| + \|\mathbf{r}_k\| + \|\delta_{\mathbf{r}_k}\|) \right) \\
 (3.13) \quad &\leq \epsilon n \frac{1}{|\alpha_{k-1}|} (\|\mathbf{r}_{k-1}\|^2 + \|\mathbf{r}_{k-1}\| \|\mathbf{r}_k\|),
 \end{aligned}$$

$$|\delta_{\gamma_{k-1}}| \leq \epsilon n \|\mathbf{s}_{k-1}\|^2$$

$$(3.14) \quad \leq \epsilon n \frac{1}{|\alpha_{k-1}|^2} (\|\mathbf{r}_{k-1}\| + \|\mathbf{r}_k\|)^2,$$

where we have used the bound

$$\|\mathbf{s}_{k-1}\| \leq (1 + 3\epsilon) \frac{1}{|\alpha_{k-1}|} (\|\mathbf{r}_{k-1}\| + \|\mathbf{r}_k\|).$$

Finally, using (3.7), (B.1), and (3.4)–(3.6), we find

$$(3.15) \quad \begin{aligned} |\delta_{\nu'_k}| &\leq 3\epsilon (|\nu_{k-1}| + 2|\alpha_{k-1}||\sigma_{k-1}| + |\alpha_{k-1}|^2|\gamma_{k-1}|) \\ &\leq 3\epsilon (\|\mathbf{r}_{k-1}\|^2 + 2|\alpha_{k-1}||\|\mathbf{r}_{k-1}\||\|\mathbf{s}_{k-1}\| + |\alpha_{k-1}|^2\|\mathbf{s}_{k-1}\|^2) \\ &\leq 3\epsilon (\|\mathbf{r}_{k-1}\|^2 + 2\|\mathbf{r}_{k-1}\|(\|\mathbf{r}_{k-1}\| + \|\mathbf{r}_k\|) + (\|\mathbf{r}_{k-1}\| + \|\mathbf{r}_k\|)^2) \\ &= 3\epsilon (4\|\mathbf{r}_{k-1}\|^2 + 4\|\mathbf{r}_{k-1}\||\|\mathbf{r}_k\| + \|\mathbf{r}_k\|^2). \end{aligned}$$

Substituting (3.11)–(3.15) into (3.9) gives a bound on the  $\nu'_k$  gap:

$$(3.16) \quad |\Delta_{\nu'_k}| \leq \epsilon ((12 + 3n)\|\mathbf{r}_{k-1}\|^2 + (18 + 4n)\|\mathbf{r}_{k-1}\||\|\mathbf{r}_k\| + (9 + n)\|\mathbf{r}_k\|^2)$$

$$(3.17) \quad \leq \epsilon (30 + 7n) (\|\mathbf{r}_{k-1}\|^2 + \|\mathbf{r}_k\|^2).$$

Therefore, given that  $\nu'_k = \|\mathbf{r}_k\|^2 - \Delta_{\nu'_k}$ , it is clear from the above bound that

$$\nu'_k \geq \|\mathbf{r}_k\|^2 - \epsilon (30 + 7n) (\|\mathbf{r}_{k-1}\|^2 + \|\mathbf{r}_k\|^2).$$

Assuming  $\epsilon$  is small enough that the higher order terms do not have a significant impact on the derived bounds, this quantity will remain positive, provided  $\|\mathbf{r}_{k-1}\|^2/\|\mathbf{r}_k\|^2 \leq 1/(\epsilon(30 + 7n)) - 1$ . While we do not explicitly bound  $\|\mathbf{r}_{k-1}\|/\|\mathbf{r}_k\|$ , we note that in practice it is unlikely to be of size  $\sim 1/\sqrt{\epsilon n}$ , thus explaining why the recompute scheme keeps  $\nu'_k$  positive.

We now consider the case where we do not use the recomputed value  $\nu_{k-1}$ , i.e., if we were to instead compute

$$(3.18) \quad \nu'_k = \nu'_{k-1} - 2\alpha_{k-1}\sigma_{k-1} + \alpha_{k-1}^2\gamma_{k-1} + \underline{\delta}_{\nu'_k},$$

where  $\underline{\delta}_{\nu'_k}$  represents the local rounding errors in computing this recurrence. In this case we would have, letting  $\underline{\Delta}_{\nu'_k}$  represent the  $\nu'$ -gap without recomputation,

$$\begin{aligned} \underline{\Delta}_{\nu'_k} &= \langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle - 2\alpha_{k-1}(\sigma_{k-1} - \delta_{\sigma_{k-1}}) + \alpha_{k-1}^2(\gamma_{k-1} - \delta_{\gamma_{k-1}}) \\ &\quad + 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_k \rangle - \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \nu'_{k-1} + 2\alpha_{k-1}\sigma_{k-1} - \alpha_{k-1}^2\gamma_{k-1} - \underline{\delta}_{\nu'_k} \\ &= (\langle \mathbf{r}_{k-1}, \mathbf{r}_{k-1} \rangle - \nu'_{k-1}) + 2\alpha_{k-1}\delta_{\sigma_{k-1}} - \alpha_{k-1}^2\delta_{\gamma_{k-1}} + 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_k \rangle - \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \underline{\delta}_{\nu'_k} \\ &= \underline{\Delta}_{\nu'_{k-1}} + 2\alpha_{k-1}\delta_{\sigma_{k-1}} - \alpha_{k-1}^2\delta_{\gamma_{k-1}} + 2\langle \delta_{\mathbf{r}_k}, \mathbf{r}_k \rangle - \langle \delta_{\mathbf{r}_k}, \delta_{\mathbf{r}_k} \rangle - \underline{\delta}_{\nu'_k}. \end{aligned}$$

Taking the absolute value of both sides and ignoring terms of  $\mathcal{O}(\epsilon)$ , we thus have the bound

$$(3.19) \quad |\underline{\Delta}_{\nu'_k}| \leq |\underline{\Delta}_{\nu'_{k-1}}| + 2|\alpha_{k-1}||\delta_{\sigma_{k-1}}| + \alpha_{k-1}^2|\delta_{\gamma_{k-1}}| + 2\|\delta_{\mathbf{r}_k}\||\|\mathbf{r}_k\| + |\underline{\delta}_{\nu'_k}|.$$

The only term in (3.19) that remains to bound is  $|\underline{\delta}_{\nu'_k}|$ . From (3.18) we have

$$|\underline{\delta}_{\nu'_k}| \leq 3\epsilon (|\nu'_{k-1}| + 2|\alpha_{k-1}||\sigma_{k-1}| + \alpha_{k-1}^2|\gamma_{k-1}|).$$

By definition,  $\nu'_{k-1} = \|\mathbf{r}_{k-1}\|^2 - \Delta_{\nu'_{k-1}}$ , so

$$|\delta_{\nu'_k}| \leq 3\epsilon \left( \|\mathbf{r}_{k-1}\|^2 + |\Delta_{\nu'_{k-1}}| + 2|\alpha_{k-1}||\sigma_{k-1}| + \alpha_{k-1}^2|\gamma_{k-1}| \right).$$

Thus, using an approach similar to the computation in (3.15), we have

$$(3.20) \quad |\delta_{\nu'_k}| \leq 3\epsilon \left( 4\|\mathbf{r}_{k-1}\|^2 + 4\|\mathbf{r}_{k-1}\|\|\mathbf{r}_k\| + \|\mathbf{r}_k\|^2 + |\Delta_{\nu'_{k-1}}| \right).$$

Plugging (3.20), (3.13), (3.14), and (3.11) into (3.19), we obtain

$$|\Delta_{\nu'_k}| \leq (1 + 3\epsilon)|\Delta_{\nu'_{k-1}}| + \epsilon\mathcal{O}(n) (\|\mathbf{r}_{k-1}\|^2 + \|\mathbf{r}_k\|^2).$$

Using that  $|\Delta_{\nu'_0}| \leq \epsilon n \|\mathbf{r}_0\|^2$ , this can be written

$$(3.21) \quad |\Delta_{\nu'_k}| \leq \epsilon\mathcal{O}(n) \sum_{i=0}^k \|\mathbf{r}_i\|^2,$$

Thus for the  $\nu'$ -gap without recomputation, we have the bound

$$(3.22) \quad \nu'_k \geq \|\mathbf{r}_k\|^2 - \epsilon\mathcal{O}(n) \sum_{i=0}^k \|\mathbf{r}_i\|^2.$$

Since now the term subtracted involves the sum of the squares of *all* previous residuals, which can be large especially at the beginning of the iterations, it is entirely possible that  $\nu'_k$  can become negative at some point during the iterations. While (3.22) is only an upper bound, it is more or less clear that this is the cause of instability in the case without recomputation, which explains the observations of Meurant [25]. In fact, if we assume that early residuals have norm  $\approx 1$ , the form of the expression (3.22) suggests that  $\nu'_k$  can no longer be guaranteed to be positive once  $\|\mathbf{r}_k\| \approx \sqrt{\epsilon}$ . This aligns with what we have observed in practice; see, for instance, Figure 1.

**3.1.1. Maximum attainable accuracy.** We now compute an expression for the residual gap in Algorithm 2.1. Substituting in our recurrences for  $\mathbf{x}_k$  and  $\mathbf{r}_k$  computed in finite precision, we find

$$\begin{aligned} \Delta_{\mathbf{r}_k} &= \mathbf{b} - \mathbf{A}(\mathbf{x}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1} + \delta_{\mathbf{x}_k}) - (\mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1} + \delta_{\mathbf{r}_k}) \\ &= (\mathbf{b} - \mathbf{A}\mathbf{x}_{k-1} - \mathbf{r}_{k-1}) + \alpha_{k-1}(\mathbf{s}_{k-1} - \mathbf{A}\mathbf{p}_{k-1}) - \mathbf{A}\delta_{\mathbf{x}_k} - \delta_{\mathbf{r}_k} \\ &= \Delta_{\mathbf{r}_{k-1}} + \alpha_{k-1}\delta_{\mathbf{s}_{k-1}} - \mathbf{A}\delta_{\mathbf{x}_k} - \delta_{\mathbf{r}_k}. \end{aligned}$$

Therefore the residual gap can be written

$$\Delta_{\mathbf{r}_k} = \Delta_{\mathbf{r}_0} + \sum_{i=1}^k (\alpha_{i-1}\delta_{\mathbf{s}_{i-1}} - \mathbf{A}\delta_{\mathbf{x}_i} - \delta_{\mathbf{r}_i}).$$

This expression is a simple accumulation of local rounding errors. We note that the residual gaps for HS-CG and CG-CG are both given by

$$\Delta_{\mathbf{r}_k} = \Delta_{\mathbf{r}_0} - \sum_{i=1}^k (\mathbf{A}\delta_{\mathbf{x}_i} + \delta_{\mathbf{r}_i}),$$

where the  $\delta_{\text{quantity}}$  terms correspond to the round off errors made by those algorithms, respectively [8].

**3.1.2. Towards understanding convergence.** We now derive an expression for the extent to which the recurrence (3.1) is satisfied by PR-CG in finite precision. We express  $\mathbf{s}_k = \mathbf{A}\mathbf{p}_k + \delta_{\mathbf{s}_k}$  in terms of  $\mathbf{r}_k$  and  $\mathbf{s}_{k-1}$ , writing

$$\begin{aligned} \mathbf{s}_k &= \mathbf{A}(\mathbf{r}_k + \beta_k \mathbf{p}_{k-1} + \delta_{\mathbf{p}_k}) + \delta_{\mathbf{s}_k} \\ &= \mathbf{A}\mathbf{r}_k + \beta_k \mathbf{A}\mathbf{p}_{k-1} + \mathbf{A}\delta_{\mathbf{p}_k} + \delta_{\mathbf{s}_k} \\ (3.23) \quad &= \mathbf{A}\mathbf{r}_k + \beta_k(\mathbf{s}_{k-1} - \delta_{\mathbf{s}_{k-1}}) + \mathbf{A}\delta_{\mathbf{p}_k} + \delta_{\mathbf{s}_k}. \end{aligned}$$

By rearranging the equation  $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1} + \delta_{\mathbf{r}_k}$ , we can write  $\mathbf{s}_{k-1} = (1/\alpha_{k-1})(\mathbf{r}_{k-1} - \mathbf{r}_k + \delta_{\mathbf{r}_k})$ . From this we compute

$$\begin{aligned} \mathbf{s}_k - \beta_k \mathbf{s}_{k-1} &= \frac{1}{\alpha_k} (\mathbf{r}_k - \mathbf{r}_{k+1} + \delta_{\mathbf{r}_{k+1}}) - \frac{\beta_k}{\alpha_{k-1}} (\mathbf{r}_{k-1} - \mathbf{r}_k + \delta_{\mathbf{r}_k}) \\ &= -\frac{1}{\alpha_k} \mathbf{r}_{k+1} + \left( \frac{1}{\alpha_k} + \frac{\beta_k}{\alpha_{k-1}} \right) \mathbf{r}_k - \frac{\beta_k}{\alpha_{k-1}} \mathbf{r}_{k-1} + \frac{1}{\alpha_k} \delta_{\mathbf{r}_{k+1}} - \frac{\beta_k}{\alpha_{k-1}} \delta_{\mathbf{r}_k}. \end{aligned}$$

Thus, shifting the indexing down by one, (3.23) becomes

$$(3.24) \quad \mathbf{A}\mathbf{r}_{k-1} = -\frac{1}{\alpha_{k-1}} \mathbf{r}_k + \left( \frac{1}{\alpha_{k-1}} + \frac{\beta_{k-1}}{\alpha_{k-2}} \right) \mathbf{r}_{k-1} - \frac{\beta_{k-1}}{\alpha_{k-2}} \mathbf{r}_{k-2} + \mathbf{f}_k,$$

where

$$\mathbf{f}_k = -\frac{\beta_{k-1}}{\alpha_{k-2}} \delta_{\mathbf{r}_{k-1}} + \beta_{k-1} \delta_{\mathbf{s}_{k-2}} - \mathbf{A}\delta_{\mathbf{p}_{k-1}} - \delta_{\mathbf{s}_{k-1}} + \frac{1}{\alpha_{k-1}} \delta_{\mathbf{r}_k}.$$

Defining  $\mathbf{q}_{k+1} := (-1)^k \mathbf{r}_k / \|\mathbf{r}_k\|$ , we obtain the three-term recurrence

$$(3.25) \quad \mathbf{A}\mathbf{q}_k = \frac{1}{\alpha_{k-1}} \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_{k-1}\|} \mathbf{q}_{k+1} + \left( \frac{1}{\alpha_{k-1}} + \frac{\beta_{k-1}}{\alpha_{k-2}} \right) \mathbf{q}_k + \frac{\beta_{k-1}}{\alpha_{k-2}} \frac{\|\mathbf{r}_{k-2}\|}{\|\mathbf{r}_{k-1}\|} \mathbf{q}_{k-1} - \frac{(-1)^{k-1}}{\|\mathbf{r}_{k-1}\|} \mathbf{f}_k.$$

In order for Greenbaum's analysis to apply, we must write this as a symmetric three-term recurrence with some perturbation term. Using our definition of the  $\nu'$ -gap,  $\Delta_{\nu'_k} := \langle \mathbf{r}_k, \mathbf{r}_k \rangle - \nu'_k$ , we have

$$\beta_k = \frac{\nu'_k}{\nu_{k-1}} + \delta_{\beta_k} = \frac{\|\mathbf{r}_k\|^2 + \Delta_{\nu'_k}}{\|\mathbf{r}_{k-1}\|^2 + \delta_{\nu'_{k-1}}} + \delta_{\beta_k} = \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2} + \Delta_{\beta_k},$$

where  $\Delta_{\beta_k} := \beta_k - \|\mathbf{r}_k\|^2 / \|\mathbf{r}_{k-1}\|^2$  can be written explicitly as

$$\Delta_{\beta_k} = \frac{\Delta_{\nu'_k} + \|\mathbf{r}_{k-1}\|^2 \delta_{\beta_k} + \delta_{\nu'_{k-1}} \delta_{\beta_k}}{\|\mathbf{r}_{k-1}\|^2 + \delta_{\nu'_{k-1}}} + \frac{-\|\mathbf{r}_k\|^2 \delta_{\nu'_{k-1}}}{\|\mathbf{r}_{k-1}\|^2 (\|\mathbf{r}_{k-1}\|^2 + \delta_{\nu'_{k-1}})}.$$

Then, by plugging this expression for  $\beta_{k-1}$  into (3.25), we obtain the approximately symmetric three-term recurrence for  $\mathbf{q}_k$ ,

$$\mathbf{A}\mathbf{q}_k = \frac{1}{\alpha_{k-1}} \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_{k-1}\|} \mathbf{q}_{k+1} + \left( \frac{1}{\alpha_{k-1}} + \frac{\beta_{k-1}}{\alpha_{k-2}} \right) \mathbf{q}_k + \frac{1}{\alpha_{k-2}} \frac{\|\mathbf{r}_{k-1}\|}{\|\mathbf{r}_{k-2}\|} \mathbf{q}_{k-1} + \mathbf{F}_k,$$

so that the error in the three-term Lanczos recurrence is given by

$$\mathbf{F}_k = \frac{\Delta_{\beta_{k-1}}}{\alpha_{k-2}} \frac{\|\mathbf{r}_{k-2}\|^2}{\|\mathbf{r}_{k-1}\|^2} \mathbf{q}_{k-1} - \frac{(-1)^k}{\|\mathbf{r}_{k-1}\|} \mathbf{f}_k.$$

Hence, the amount by which the updated residuals in PR-CG fail to satisfy (3.1) depends only on local rounding errors, which was also observed for HS-CG and CG-CG in [20]. While  $\|\mathbf{r}_{k-1}\|$  may become quite small, we note that  $\mathbf{f}_k$  consists of roundoff errors made in iterations  $k-1$  and  $k-2$  and corresponds to vectors which are of comparable size to  $\mathbf{r}_{k-1}$ ; i.e., it is reasonable to assume that the ratio of the norms of these vectors to  $\mathbf{r}_{k-1}$  will be much less than  $1/\epsilon$ . Based on this heuristic,  $\mathbf{f}_k/\|\mathbf{r}_{k-1}\|$  can be expected to stay small, which provides some indication for why the rate of convergence of PR-CG tends to be similar to HS-CG and CG-CG.

**3.2. Pipelined predict-and-recompute CG.** We now provide for PIPE-PR-CG a similar rounding error analysis for the residual gap and Lanczos recurrence as we did for PR-CG. Of course, the  $\delta_{\text{quantity}}$  error terms in this section are different from those in the previous section, even though we use many of the same symbols for notational convenience. In particular, our expression for  $\mathbf{s}_k$  in finite precision is now  $\mathbf{s}_k = \mathbf{w}'_k + \beta_k \mathbf{s}_{k-1} + \delta_{\mathbf{s}_k}$  and we add the recurrences

$$\begin{aligned} \mathbf{w}'_k &= \mathbf{w}_{k-1} - \alpha_{k-1} \mathbf{u}_{k-1} + \delta_{\mathbf{w}'_k}, & \mathbf{w}_k &= \mathbf{A} \mathbf{r}_k + \delta_{\mathbf{w}_k}, \\ \mathbf{p}_k &= \mathbf{r}_k + \beta_k \mathbf{p}_{k-1} + \delta_{\mathbf{p}_k}, & \mathbf{u}_k &= \mathbf{A} \mathbf{s}_k + \delta_{\mathbf{u}_k}. \end{aligned}$$

We note that the same expression holds for  $\Delta_{\nu'_k}$  as with PR-CG; see (3.8).

**3.2.1. Maximum attainable accuracy.** Similarly as for PR-CG, we can write the residual gap in PIPE-PR-CG as

$$\begin{aligned} \Delta_{\mathbf{r}_k} &= \mathbf{b} - \mathbf{A} \mathbf{x}_k - \mathbf{r}_k \\ &= \Delta_{\mathbf{r}_{k-1}} + \alpha_{k-1} (\mathbf{s}_{k-1} - \mathbf{A} \mathbf{p}_{k-1}) - \mathbf{A} \delta_{\mathbf{x}_k} - \delta_{\mathbf{r}_k} \\ &= \Delta_{\mathbf{r}_{k-1}} - \alpha_{k-1} \Delta_{\mathbf{s}_{k-1}} - \mathbf{A} \delta_{\mathbf{x}_k} - \delta_{\mathbf{r}_k} \\ &= \Delta_{\mathbf{r}_0} - \sum_{i=1}^k [\alpha_{i-1} \Delta_{\mathbf{s}_{i-1}} + \mathbf{A} \delta_{\mathbf{x}_i} + \delta_{\mathbf{r}_i}], \end{aligned}$$

where  $\Delta_{\mathbf{s}_i} := \mathbf{A} \mathbf{p}_i - \mathbf{s}_i$  represents the gap between  $\mathbf{A} \mathbf{p}_i$  and the recursively computed quantity  $\mathbf{s}_i$ . We can write this  $\mathbf{s}$ -gap as

$$\begin{aligned} \Delta_{\mathbf{s}_i} &= \mathbf{A}(\mathbf{r}_i + \beta_i \mathbf{p}_{i-1} + \delta_{\mathbf{p}_i}) - (\mathbf{w}'_i + \beta_i \mathbf{s}_{i-1} + \delta_{\mathbf{s}_i}) \\ &= (\mathbf{A} \mathbf{r}_i - \mathbf{w}'_i) + \beta_i (\mathbf{A} \mathbf{p}_{i-1} - \mathbf{s}_{i-1}) + \mathbf{A} \delta_{\mathbf{p}_i} - \delta_{\mathbf{s}_i} \\ &= \Delta_{\mathbf{w}'_i} + \beta_i \Delta_{\mathbf{s}_{i-1}} + \mathbf{A} \delta_{\mathbf{p}_i} - \delta_{\mathbf{s}_i} \\ &= \Delta_{\mathbf{s}_0} \prod_{j=1}^i \beta_j + \sum_{j=1}^i \left[ (\Delta_{\mathbf{w}'_j} + \mathbf{A} \delta_{\mathbf{p}_j} - \delta_{\mathbf{s}_j}) \prod_{\ell=j+1}^i \beta_\ell \right], \end{aligned}$$

where  $\Delta_{\mathbf{w}'_j}$  denotes the gap  $\Delta_{\mathbf{w}'_j} := \mathbf{A} \mathbf{r}_j - \mathbf{w}'_j$ . Now we look at how to bound this  $\mathbf{w}'$ -gap. In PIPE-PR-CG, we have

$$\begin{aligned} \Delta_{\mathbf{w}'_j} &= \mathbf{A}(\mathbf{r}_{j-1} - \alpha_{j-1} \mathbf{s}_{j-1} + \delta_{\mathbf{r}_j}) - (\mathbf{w}_{j-1} - \alpha_{j-1} \mathbf{u}_{j-1} + \delta_{\mathbf{w}'_j}) \\ &= \mathbf{A} \mathbf{r}_{j-1} - \mathbf{w}_{j-1} - \alpha_{j-1} (\mathbf{A} \mathbf{s}_{j-1} - \mathbf{u}_{j-1}) - \delta_{\mathbf{w}'_j} + \mathbf{A} \delta_{\mathbf{r}_j} \\ (3.26) \quad &= -\delta_{\mathbf{w}_{j-1}} + \alpha_{j-1} \delta_{\mathbf{u}_{j-1}} - \mathbf{A} \delta_{\mathbf{r}_j} - \delta_{\mathbf{w}'_j}, \end{aligned}$$

which is the sum of local rounding errors.

Notice that without recomputation, i.e., if we were to instead compute

$$\mathbf{w}'_j - \mathbf{w}'_{j-1} - \alpha_{j-1} \mathbf{u}_{j-1} + \delta_{\mathbf{w}'_j},$$

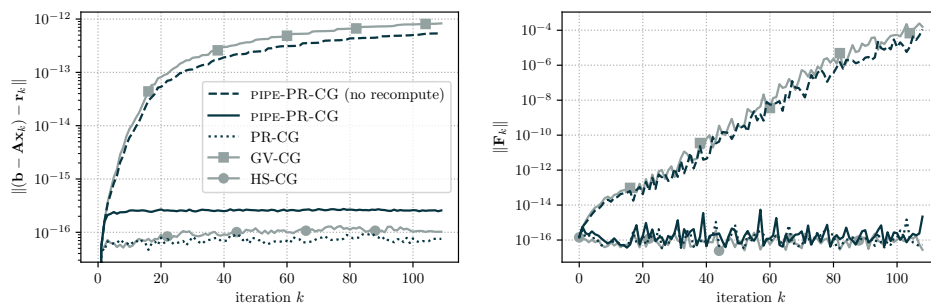


FIG. 2. Residual gap (left) and 2-norm of error in three-term Lanczos recurrence (right) on model\_48-8-3 without a preconditioner. In the legend, “no recompute” indicates that  $\mathbf{w}_k$  is not recomputed.

where  $\delta_{\mathbf{w}'_j}$  denotes the local rounding errors in computing this recurrence, then we would have, letting  $\underline{\Delta}_{\mathbf{w}'_j}$  denote the  $\mathbf{w}'$ -gap in the case of no recomputation,

$$\begin{aligned}
 \underline{\Delta}_{\mathbf{w}'_j} &= \mathbf{A}\mathbf{r}_{j-1} - \mathbf{w}'_{j-1} - \alpha_{j-1}(\mathbf{A}\mathbf{s}_{j-1} - \mathbf{u}_{j-1}) + \mathbf{A}\delta_{\mathbf{r}_j} - \delta_{\mathbf{w}'_j} \\
 &= \underline{\Delta}_{\mathbf{w}'_{j-1}} + \alpha_{j-1}\delta_{\mathbf{u}_{j-1}} + \mathbf{A}\delta_{\mathbf{r}_j} - \delta_{\mathbf{w}'_j} \\
 (3.27) \quad &= \underline{\Delta}_{\mathbf{w}'_0} + \sum_{\ell=1}^j \left[ \alpha_{\ell-1}\delta_{\mathbf{u}_{\ell-1}} + \mathbf{A}\delta_{\mathbf{r}_\ell} + \delta_{\mathbf{w}'_\ell} \right].
 \end{aligned}$$

Thus, comparing (3.26) with (3.27), we see that if we use recomputation,  $\underline{\Delta}_{\mathbf{w}'_j}$  is essentially “reset” in each iteration; the errors made do not accumulate and  $\underline{\Delta}_{\mathbf{w}'_j}$  remains a small multiple of the machine precision. Without recomputation, however, a bound on size of the quantity  $\underline{\Delta}_{\mathbf{w}'_j}$  will be monotonically increasing, growing larger in each iteration due to the accumulation of rounding errors. We note that the expression for the maximum attainable accuracy for PIPE-PR-CG in (3.27), *without the recomputation of  $\mathbf{w}_j$* , closely resembles the expression derived for the maximum attainable accuracy for GV-CG derived in [8]. This is demonstrated numerically in Figure 2. Indeed, without recomputation, it is easily observed that PIPE-PR-CG and GV-CG typically behave quite similarly.

If we look at the expression for  $\Delta_{\mathbf{s}_i}$ , we see that  $\Delta_{\mathbf{w}'_j}$  (or in the case of no recomputation,  $\underline{\Delta}_{\mathbf{w}'_j}$ ), along with local rounding errors  $\mathbf{A}\delta_{\mathbf{p}_j}$  and  $\delta_{\mathbf{s}_j}$ , is amplified by a product of  $\beta$  terms,  $\prod_{\ell=j+1}^i \beta_\ell$ . In exact arithmetic at least, we have

$$\prod_{\ell=j+1}^i \beta_\ell = \beta_{j+1}\beta_{j+2}\cdots\beta_i = \frac{\|\mathbf{r}_i\|^2}{\|\mathbf{r}_j\|^2},$$

and we expect this quantity to be greater than one in cases where CG exhibits large oscillation in the residual norm.

Thus it is clear why recomputation can improve the maximal attainable accuracy: the term  $\Delta_{\mathbf{w}'_j}$  should remain small (some multiple of  $\epsilon$ ) in this case, whereas without recomputation, the analogous term can grow larger and larger throughout the iterations. It is also clear why the PIPE-PR-CG method can exhibit potentially decreased accuracy versus the HS-CG method; the local rounding errors (including the term

$\Delta_{\mathbf{w}'_j}$ ) are amplified by quantities related to the ratios of residual norms, which can be large for certain problems.

Finally, we note that the same theory holds for PIPE-M-CG, as the expressions for the relevant recurrences are the same as in PIPE-PR-CG.

**3.2.2. Towards understanding convergence.** Following the same procedure as for PR-CG, we can write

$$\mathbf{s}_k = \mathbf{w}'_k + \beta_k \mathbf{s}_{k-1} + \delta_{\mathbf{s}_k} = \mathbf{A}\mathbf{r}_k + \beta_k \mathbf{s}_{k-1} - \Delta_{\mathbf{w}'_k} + \delta_{\mathbf{s}_k}.$$

This is almost the same as (3.23) except that  $\mathbf{A}\delta_{\mathbf{p}_k} - \beta_k \delta_{\mathbf{s}_{k-1}}$  has been replaced by  $-\Delta_{\mathbf{w}'_k}$ . Thus, through similar computations as above, we find

$$\mathbf{A}\mathbf{q}_k = \frac{1}{\alpha_{k-1}} \frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_{k-1}\|} \mathbf{q}_{k+1} + \left( \frac{1}{\alpha_{k-1}} + \frac{\beta_{k-1}}{\alpha_{k-2}} \right) \mathbf{q}_k + \frac{1}{\alpha_{k-2}} \frac{\|\mathbf{r}_{k-1}\|}{\|\mathbf{r}_{k-2}\|} \mathbf{q}_{k-1} + \mathbf{F}_k,$$

where the error term is given by

$$\mathbf{F}_k = \frac{\Delta_{\beta_{k-1}}}{\alpha_{k-2}} \frac{\|\mathbf{r}_{k-2}\|^2}{\|\mathbf{r}_{k-1}\|^2} \mathbf{q}_{k-1} - \frac{(-1)^k}{\|\mathbf{r}_{k-1}\|} \mathbf{f}_k$$

and

$$\mathbf{f}_k = -\frac{\beta_{k-1}}{\alpha_{k-2}} \delta_{\mathbf{r}_{k-1}} + \Delta_{\mathbf{w}'_k} - \delta_{\mathbf{s}_{k-1}} + \delta_{\mathbf{r}_k}.$$

While the three-term Lanczos recurrence generated by PIPE-PR-CG is satisfied to within local rounding errors, the error term  $\mathbf{F}_k$  for PIPE-PR-CG depends on  $\Delta_{\mathbf{w}'_k}$  rather than  $\mathbf{A}\delta_{\mathbf{p}_k}$ . Thus, it is reasonable to expect it to be larger than the error term for PR-CG. On the other hand, the expression for the degree to which GV-CG satisfies the three-term Lanczos recurrence depends on errors made in all previous steps [20]. This provides some intuition for why the rate of convergence of PIPE-PR-CG is observed to be better than GV-CG. Again, a numerical comparison is shown in Figure 2.

**4. Numerical experiments.** In this section we present the results of numerical experiments intended to give insight into the numerical behavior of the variants introduced in this paper. We run experiments on a range of matrices from the Matrix Market [3]. In addition we include a model problem which was introduced in [33] and has since been considered in [34, 12, 20]. This model problem has eigenvalues  $\lambda_1 = 1/\kappa$ ,  $\lambda_n = 1$ , and for  $i = 2, 3, \dots, n-1$ ,

$$(4.1) \quad \lambda_i = \lambda_1 + \left( \frac{i-1}{n-1} \right) \cdot (\lambda_n - \lambda_1) \cdot \rho^{n-i}$$

for  $n = 48$ ,  $\rho = 0.8$ , and  $\kappa = 10^3$ , with eigenvectors chosen uniformly from the set of unitary matrices. Since the spacing between eigenvalues grows exponentially, this is a particularly difficult problem in finite precision arithmetic.

The experiments are implemented in NumPy using IEEE double precision floating point arithmetic. The results are outlined in Table 2. In this table we give two summary statistics: (i) the number of iterations required to decrease the  $\mathbf{A}$ -norm of the error by a factor of  $10^5$ , and (ii) the minimum error reached. For a given problem, these two quantities give a rough indication of the rate of convergence and ultimately

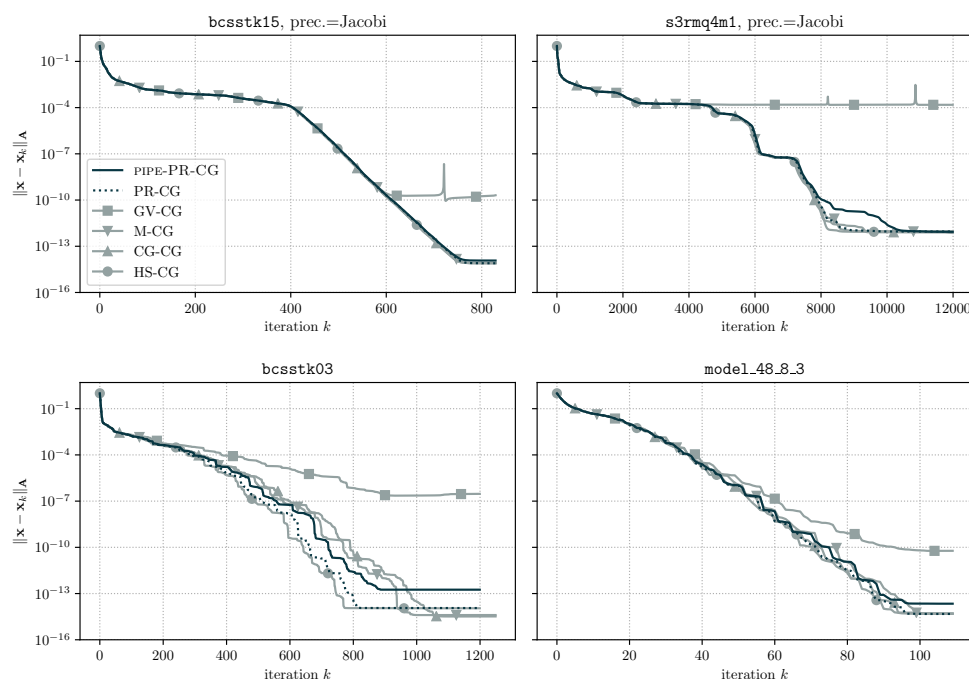


TABLE 2

Summary statistics of convergence behavior on problems from the Matrix Market. Preconditioners are applied using preconditioned variants rather than constructing an explicitly preconditioned system. Values are bold if they differ from HS-CG by more than 10 percent, and dashes indicate that a method failed to reach the specified accuracy. Note that  $\mathbf{e}_k = \mathbf{A}^{-1}\mathbf{b} - \mathbf{x}_k$  is the error at step  $k$ .

Matrix	Prec.	$n$	nnz	Iterations: $\min\{k : \ \mathbf{e}_k\ _{\mathbf{A}}/\ \mathbf{e}_0\ _{\mathbf{A}} < 10^{-5}\}$								Minimum error: $\min_k \log_{10}(\ \mathbf{e}_k\ _{\mathbf{A}}/\ \mathbf{e}_0\ _{\mathbf{A}})$							
				HS	CG	M	PR	GV	PM	PPR	HS	CG	M	PR	GV	PM	PPR	HS	CG
1138_bus	-	1138	4054	1721	1753	1797	1727	1870	1799	1733	-12.69	-12.52	-12.70	-12.73	-6.54	-11.85	-11.85		
494_bus	-	494	1666	898	917	941	899	<b>1040</b>	957	909	-13.14	-12.48	-13.11	-13.11	-6.89	-12.24	-12.16		
662_bus	-	662	2474	443	447	451	444	464	451	444	-13.93	-13.21	-13.93	-13.95	-8.78	-12.67	-13.35		
685_bus	-	685	3249	437	456	455	439	<b>485</b>	471	445	-14.36	-14.10	-14.31	-14.36	-9.53	-13.15	-13.06		
bcsstk03	-	112	640	364	<b>439</b>	<b>425</b>	380	<b>598</b>	<b>492</b>	<b>411</b>	-14.55	-14.49	-14.40	-14.43	-6.86	-12.65	-12.96		
bcsstk14	-	1806	63454	3982	4060	4045	4003	4212	4096	4014	-14.31	-14.19	-14.30	-14.29	-5.32	-14.24	-14.26		
bcsstk15	-	3948	117816	5702	5777	5782	5721	5951	5820	5738	-13.77	-13.40	-13.79	-13.79	-5.60	-13.61	-13.74		
bcsstk16	-	4884	290378	429	430	430	430	<b>818</b>	431	430	-14.48	-14.34	-14.47	-14.47	-5.07	-14.18	-14.20		
bcsstk17	-	10974	428650	17568	18103	18174	17590	-	18497	17795	-13.42	-12.85	-13.42	-13.45	-4.25	-12.06	-12.30		
bcsstk18	-	11948	149090	42525	43366	43118	42652	<b>49210</b>	43704	42908	-13.18	-13.15	-13.16	-13.18	-5.00	-13.15	-13.18		
bcsstk27	-	1224	56126	519	521	523	520	531	523	520	-14.42	-14.26	-14.42	-14.43	-9.85	-14.04	-14.10		
bcsstm19	-	817	817	274	287	286	274	<b>334</b>	290	277	-14.84	-14.84	-14.82	-14.82	-9.63	-14.59	-14.67		
bcsstm20	-	485	485	203	219	221	205	<b>239</b>	<b>228</b>	208	-15.06	-14.95	-15.09	-15.13	-9.81	-14.92	-14.89		
bcsstm21	-	3600	3600	3	3	3	3	3	3	3	-15.69	-15.68	-15.69	-15.67	-14.58	-14.92	-14.92		
bcsstm22	-	138	138	43	43	43	43	43	43	43	-15.43	-15.43	-15.48	-15.43	-12.51	-15.26	-15.07		
bcsstm23	-	3134	3134	1325	1376	1360	1342	1434	1367	1346	-14.35	-14.29	-14.31	-14.36	-6.87	-14.30	-14.34		
bcsstm24	-	3562	3562	1573	1689	1686	1595	<b>19411</b>	1698	1605	-14.14	-13.97	-14.05	-14.13	-5.00	-13.83	-14.04		
bcsstm25	-	15439	15439	10089	10948	10963	10293	<b>12736</b>	<b>11245</b>	10400	-13.84	-13.74	-13.68	-13.81	-5.50	-13.62	-13.76		
model_48.8.3	-	48	2304	43	42	45	44	45	43	44	-14.32	-14.28	-14.31	-14.32	-10.23	-13.67	-13.66		
nos1	-	237	1017	1846	1895	2008	1843	<b>2305</b>	1999	1870	-12.81	-13.07	-12.83	-12.80	-5.81	-12.15	-11.82		
nos2	-	957	4137	29829	30672	32717	29706	-	32157	29744	-11.29	-11.21	-11.30	-11.29	-3.23	-10.99	-10.99		
nos3	-	960	15844	221	221	221	221	221	221	221	-13.39	-13.58	-13.39	-13.39	-9.86	-13.36	-13.22		
nos4	-	100	594	72	72	72	72	72	72	72	-14.33	-14.41	-14.32	-14.33	-11.47	-14.19	-14.19		
nos5	-	468	5172	315	315	316	316	317	316	316	-14.98	-14.97	-15.00	-14.99	-10.97	-14.91	-14.89		
nos6	-	675	3255	551	555	582	555	<b>672</b>	601	589	-12.21	-12.28	-12.23	-12.22	-6.67	-10.21	-10.21		
nos7	-	729	4617	2869	2798	<b>3536</b>	2874	-	<b>3416</b>	2899	-9.01	-8.74	-8.97	-9.01	-0.65	-6.80	-7.24		
s1rmq4m1	-	5489	281111	3406	3447	3432	3410	3603	3442	3434	-13.54	-13.53	-13.55	-13.56	-8.23	-13.47	-13.43		
s1rmt3m1	-	5489	219521	3890	3932	3910	3895	4076	3916	3908	-13.39	-13.28	-13.39	-13.40	-7.35	-13.34	-13.22		
s2rmq4m1	-	5489	281111	10476	10699	10651	10491	<b>11622</b>	10693	10615	-13.09	-13.32	-13.11	-13.12	-6.07	-12.78	-13.19		
s2rmt3m1	-	5489	219521	14484	14727	14655	14533	-	14679	14620	-12.83	-12.70	-12.82	-12.81	-4.55	-12.59	-12.78		
s3rmq4m1	-	5489	281111	26628	<b>29395</b>	28004	26937	-	28822	28161	-12.06	-11.94	-12.06	-12.10	-4.20	-11.69	-11.59		
s3rmt3m1	-	5489	219521	38459	41037	40188	38471	-	40839	40105	-12.04	-11.86	-12.08	-12.07	-4.07	-11.54	-11.28		
s3rmt3m3	-	5357	207695	69095	72598	71471	69051	-	72258	70852	-12.60	-11.72	-12.59	-12.71	-4.20	-11.65	-11.13		
1138_bus	Jac.	1138	4054	734	734	734	734	734	734	734	-12.69	-12.75	-12.67	-12.70	-8.62	-12.66	-12.65		
494_bus	Jac.	494	1666	371	371	371	371	371	371	371	-13.15	-13.09	-13.09	-13.15	-9.84	-13.14	-13.16		
662_bus	Jac.	662	2474	166	166	166	166	166	166	166	-14.16	-14.12	-14.15	-14.19	-10.94	-14.11	-13.76		
685_bus	Jac.	685	3249	192	192	192	192	192	192	192	-14.48	-14.36	-14.59	-14.46	-11.32	-14.51	-14.33		
bcsstk03	Jac.	112	640	118	118	120	120	120	120	121	-14.10	-14.11	-14.10	-14.05	-9.48	-13.48	-13.50		
bcsstk14	Jac.	1806	63454	198	198	198	198	198	198	198	-14.78	-14.73	-14.66	-14.67	-11.18	-14.35	-14.36		
bcsstk15	Jac.	3948	117816	442	442	443	444	444	444	444	-14.10	-14.08	-14.11	-14.10	-10.05	-13.95	-13.93		
bcsstk16	Jac.	4884	290378	132	132	132	132	132	132	132	-14.61	-14.51	-14.60	-14.60	-10.98	-14.24	-14.20		
bcsstk17	Jac.	10974	428650	2203	2205	2210	2212	2218	2214	2216	-13.98	-13.71	-13.98	-14.00	-7.99	-13.40	-13.32		
bcsstk18	Jac.	11948	149090	536	537	539	541	542	541	542	-14.57	-14.54	-14.57	-14.55	-10.12	-14.30	-14.30		
bcsstk27	Jac.	1224	56126	173	173	173	174	174	174	174	-14.67	-14.41	-14.67	-14.70	-10.45	-13.99	-14.03		
model_48.8.3	Jac.	48	2304	49	48	50	50	52	50	50	-14.30	-14.25	-14.28	-14.29	-10.66	-13.70	-13.72		
nos1	Jac.	237	1017	306	314	322	312	<b>346</b>	323	326	-12.98	-12.79	-12.93	-12.96	-6.70	-12.67	-12.28		
nos2	Jac.	957	4137	3047	3184	3197	3097	-	3326	3303	-11.27	-11.32	-11.30	-11.27	-3.44	-11.23	-11.12		
nos3	Jac.	960	15844	186	186	186	186	186	186	186	-13.38	-13.42	-13.37	-13.39	-9.62	-13.55	-13.53		
nos4	Jac.	100	594	67	67	67	67	67	67	67	-14.30	-14.39	-14.34	-14.36	-11.76	-14.22	-14.14		
nos5	Jac.	468	5172	136	136	136	136	136	136	136	-15.07	-14.99	-15.11	-15.08	-12.02	-14.82	-14.89		
nos6	Jac.	675	3255	71	71	71	71	71	71	71	-12.17	-12.00	-12.19	-12.20	-9.10	-12.14	-12.14		
nos7	Jac.	729	4617	67	67	67	67	67	67	67	-8.91	-9.21	-8.90	-8.88	-6.41	-9.42	-9.41		
s1rmq4m1	Jac.	5489	281111	595	595	596	596	597	597	597	-13.95	-13.86	-13.97	-13.90	-8.49	-13.67	-13.65		
s1rmt3m1	Jac.	5489	219521	674	674	674	675	675	675	676	-13.58	-13.76	-13.60	-13.62	-8.76	-13.71	-13.20		
s2rmq4m1	Jac.	5489	281111	1437	1437	1438	1439	1439	1439	1440	-13.16	-12.65	-13.15	-13.23	-6.65	-13.10	-12.81		
s2rmt3m1	Jac.	5489	219521	2030	2028	2033	2034	2040	2037	2039	-12.84	-12.65	-12.83	-12.81	-6.19	-12.47	-12.19		
s3dkq4m2	Jac.	90449	4820891	25527	25513	25548	25553	-	25576	25582	-11.09	-11.17	-11.10	-11.09	-4.03	-11.12	-11.21		
s3dkt3m2	Jac.	90449	3753461	36195	36152	36247	36263	-	36327	36348	-11.39	-10.91	-11.39	-11.39	-3.97	-11.09	-10.82		
s3rmq4m1	Jac.	5489	281111	5743	5726	5775	5780	-	5800	5806	-12.06	-12.14	-12.03	-12.04	-3.82	-12.00	-12.06		
s3rmt3m1	Jac.	5489	219521	8827	8806	8867	8871	-	8908	8917	-12.07	-11.78	-12.09	-12.10	-3.78	-11.57	-11.41		
s3rmt3m3	Jac.	5357	207695	10251	10248	10317	10324	-	10385	10404	-12.84	-11.77	-12.58	-12.93	-4.43	-11.66	-11.95		

attainable accuracy. Plots of convergence for all experiments appearing in Table 2 can be found online in the repository linked in Appendix A. As was done in [16], the right-hand side  $\mathbf{b}$  is chosen so that  $\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{b}$  has entries  $1/\sqrt{n}$ , and the initial guess  $\mathbf{x}_0$  is the zero vector. For most problems we selected, we run tests without a preconditioner, and then with a simple Jacobi (diagonal) preconditioner. For each


 FIG. 3. Error  $A$ -norm of conjugate gradient variants on selected problems.

problem, we run all variants for a sufficient number of iterations such that the true residual stagnates.

Figure 3 shows the results of four of the numerical tests contained in Table 2. These problems were chosen to highlight some of the types of behavior observed for finite precision CG variants. On some problems such as `bcsstk03` and the model problem `model_48.8.3`, the rate of convergence and final accuracy of each variant may differ, due primarily to the large gaps in the spectrum [4]. Alternatively, on many other problems, such as `bcsstk15` with Jacobi preconditioning, the rate of convergence for all variants is the same until the final accuracy is reached. However, even on such problems, it may be possible for the final accuracy of a variant to be significantly worse than other variants. For instance, on `s3rmq4m1` with Jacobi preconditioning, the final accuracy of GV-CG is *8 orders of magnitude* worse than HS-CG even though the rates of convergence are initially the same.

We note that on problems where CG-CG encounters a delay of convergence, such as `bcsstk03`, PR-CG converges more quickly. More notably, for the experiments in Table 2, the pipelined predict-and-recompute variants PIPE-M-CG and PIPE-PR-CG show significantly better convergence than GV-CG, frequently exhibiting convergence similar to that of HS-CG. In particular, on all the problems tested, PIPE-M-CG and PIPE-PR-CG converge to a final accuracy within 10 percent (on a log scale) of that of HS-CG if Jacobi preconditioning is used, and on some problems, these two variants actually converge to a *better* final accuracy than HS-CG.

**5. Parallel performance.** We implement the unpreconditioned variants discussed in this paper using `mpi4py` [11] and perform a strong scaling experiment on the Hyak supercomputer at the University of Washington. In this experiment we

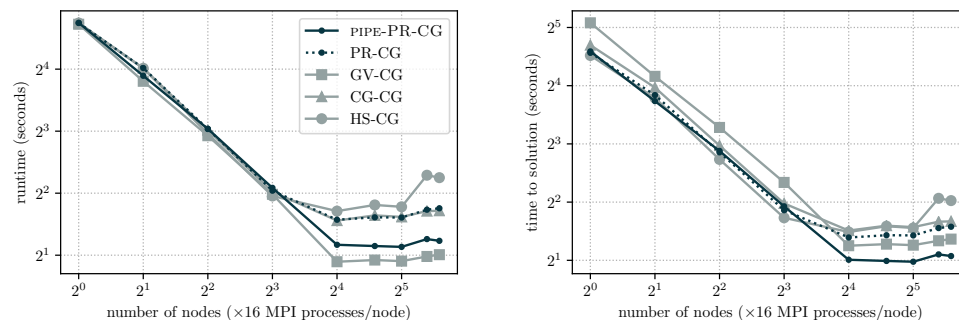


FIG. 4. Strong scaling experiment on (dense) model problem with  $n = 10752$  unknowns and 1000 iterations. Time to solution is approximate time to decrease the  $\mathbf{A}$ -norm of error by a factor of  $10^5$ .

solve the model problem (4.1) with parameters  $n = 10752$ ,  $\rho = 0.9$ ,  $\kappa = 10^6$ . In order to demonstrate that the additional computation costs associated with the extra matrix-vector product required by PIPE-PR-CG are not particularly important, we represent our model problem as a dense matrix. In each trial, we first allocate 48 nodes with 16 processors each and then iterate over a selection of node counts. On each node count we run each variant for 1000 iterations, timing only the main loop of each variant and not any costs corresponding to the setup. In order to account for effects such as cluster topology and system noise we repeat this process 4 times and report the minimum runtimes and accuracy of these runs.

Figure 4 shows the results of the experiment. In our implementation, we compute  $\mathbf{A}\mathbf{s}_k$  and  $\mathbf{A}\mathbf{r}_k$  simultaneously using a block vector  $[\mathbf{s}_k, \mathbf{r}_k]$  so that  $\mathbf{A}$  is accessed only once per iteration. While the number of floating point operations is nearly doubled in PIPE-PR-CG, the memory access pattern is more or less unchanged. Specifically,  $\mathbf{A}$  needs to only be read/communicated a single time per iteration. For this reason, the runtime of PIPE-PR-CG is almost the same as GV-CG. If we compute the two matrix products separately, then the runtime of PIPE-PR-CG is approximately doubled on low node counts.

In both cases, as a higher number of nodes are used, and communication costs begin to come into play, PIPE-PR-CG becomes faster than the nonpipelined variants and exhibits similar scaling properties to GV-CG. In the high node limit, GV-CG does perform marginally better than PIPE-PR-CG. This is because our test involves a dense matrix, so due to the two matrix-vector products, the amount of data communicated between nodes in PIPE-PR-CG is nearly double that of GV-CG. However, it is important to keep in mind that the actual “time to solution” is given by the product of the time per iteration and the number of iterations required.

**6. Future work.** First, while we have provided an analysis of the maximal attainable accuracy for our new variants, the analysis of the resulting convergence rate is far from complete. This is a difficult area of study and deserves further rigorous treatment.

Second, since the maximum reduction in communication costs of PIPE-PR-CG over HS-CG is only a factor of three, potential ways of further decreasing communication costs should be explored. Recently, there has been work on “deep pipelined” CG variants where more matrix-vector products are overlapped with global communication; see, for instance, [16, 10, 7]. This approach is similar to the “look ahead”

strategy suggested in [36] and may be compatible with the schemes introduced here.

Alternatively, it may be possible to either incorporate predict-and-recompute strategies into  $s$ -step methods or to develop new  $s$ -step methods which are built upon PR-CG. As CG-CG is the  $s = 1$  case of the  $s$ -step method from [6], we expect it may be possible to develop an  $s$ -step method based on PR-CG which has slightly better numerical properties than CG-CG. Finding such a method which is usable in practice would be of great practical interest.

Finally, the predict-and-recompute variants presented here can be naturally extended to other related methods such as conjugate residual and conjugate gradient squared.

**7. Conclusion.** In this paper we extend the predict-and-recompute idea of Meurant [25] and derive new pipelined CG variants, PIPE-M-CG and PIPE-PR-CG. These variants exhibit better theoretical scaling properties than the standard HS-CG algorithm as well as improved numerical behavior compared to their previously studied counterpart, GV-CG, on *every* numerical experiment we ran. We provide an analysis of the maximum attainable accuracy, applicable to both PIPE-M-CG and PIPE-PR-CG, which explains the benefit of predict-and-recompute with regard to avoiding the breakdown observed in [25], as well as to the improved final accuracy and rate of convergence observed by our pipelined variants. We additionally provide a strong scaling experiment which confirms the potential performance benefits of our approach. Our predict-and-recompute variants require exactly the same input parameters as HS-CG, and therefore have the potential to be used wherever HS-CG is used without any additional parameter selection. Despite these advances, there is still significant room for future work on high performance CG variants, especially in the direction of further decreasing the communication costs.

**Appendix A. Additional resources.** A repository with the code necessary to reproduce all of the figures and results in this paper is available at [https://github.com/tchen01/new\\_cg\\_variants](https://github.com/tchen01/new_cg_variants) and released to the public domain under the MIT License. The repository also contains convergence data and plots for all the matrices listed in Table 2.

We are committed to facilitating the reproducibility process and encourage questions and inquiries into the methods used in this paper.

**Appendix B. Bound on sum of products.** Define

$$C = \sum_{i=1}^n \prod_{j=1}^{b_i} c_{i,j}.$$

Then

$$\left| \text{fp} \left( \prod_{j=1}^{b_i} c_{i,j} \right) - \prod_{j=1}^{b_i} c_{i,j} \right| \leq (1 + \epsilon)^{b_i} \prod_{j=1}^{b_i} |c_{i,j}|,$$

so

$$(B.1) \quad |\text{fp}(C) - C| \leq n \epsilon \sum_{i=1}^n (1 + \epsilon)^{b_i} \prod_{j=1}^{b_i} |c_{i,j}| = n \epsilon \sum_{i=1}^n \prod_{j=1}^{b_i} |c_{i,j}| + \mathcal{O}(\epsilon^2).$$

**Appendix C. A note on other variants studied.** Many of the scalar quantities appearing in the variants discussed in this paper can be replaced with equivalent

expressions. For instance, the difference between M-CG and PR-CG (as well as PIPE-M-CG and PIPE-PR-CG) is the choice of expression for  $\nu_k$ , which in exact arithmetic is equal to  $\langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ . However, it is not typically clear which scalar expression should be used.

We first discuss our simplified expression for (2.2) which is used in PR-CG and PIPE-PR-CG. Recall that in exact arithmetic,  $\langle \tilde{\mathbf{s}}_k, \mathbf{r}_k \rangle = \langle \mathbf{s}_k, \tilde{\mathbf{r}}_k \rangle$ . This cannot be expected to be true in finite precision.

Writing the finite precision recurrences

$$\tilde{\mathbf{s}}_k = \mathbf{M}^{-1} \mathbf{s}_k + \delta_{\tilde{\mathbf{s}}_k}, \quad \tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_{k-1} - \alpha_{k-1} \tilde{\mathbf{s}}_{k-1} + \delta_{\tilde{\mathbf{r}}_k},$$

we have

$$\begin{aligned} \langle \tilde{\mathbf{s}}_k, \mathbf{r}_k \rangle &= \langle \mathbf{M}^{-1} \mathbf{s}_k + \delta_{\tilde{\mathbf{s}}_k}, \mathbf{r}_k \rangle \\ &= \langle \mathbf{s}_k, \mathbf{M}^{-1} \mathbf{r}_k \rangle + \langle \delta_{\tilde{\mathbf{s}}_k}, \mathbf{r}_k \rangle \\ &= \langle \mathbf{s}_k, \tilde{\mathbf{r}}_k \rangle + \langle \mathbf{s}_k, \Delta_{\tilde{\mathbf{r}}_k} \rangle + \langle \delta_{\tilde{\mathbf{s}}_k}, \mathbf{r}_k \rangle, \end{aligned}$$

where we have defined the  $\tilde{\mathbf{r}}$ -gap as  $\Delta_{\tilde{\mathbf{r}}_k} := \mathbf{M}^{-1} \mathbf{r}_k - \tilde{\mathbf{r}}_k$ . For the  $\tilde{\mathbf{r}}$ -gap, we have

$$\begin{aligned} \Delta_{\tilde{\mathbf{r}}_k} &= \mathbf{M}^{-1}(\mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{s}_{k-1} + \delta_{\mathbf{r}_k}) - (\tilde{\mathbf{r}}_{k-1} - \alpha_{k-1} \tilde{\mathbf{s}}_{k-1} + \delta_{\tilde{\mathbf{r}}_k}) \\ &= (\mathbf{M}^{-1} \mathbf{r}_{k-1} - \tilde{\mathbf{r}}_{k-1}) - \alpha_{k-1} (\mathbf{M}^{-1} \mathbf{s}_{k-1} - \tilde{\mathbf{s}}_{k-1}) + \mathbf{M}^{-1} \delta_{\mathbf{r}_k} - \delta_{\tilde{\mathbf{r}}_k} \\ &= \Delta_{\tilde{\mathbf{r}}_{k-1}} + \alpha_{k-1} \delta_{\tilde{\mathbf{s}}_{k-1}} + \mathbf{M}^{-1} \delta_{\mathbf{r}_k} - \delta_{\tilde{\mathbf{r}}_k}. \\ &= \Delta_{\tilde{\mathbf{r}}_0} + \sum_{i=1}^k [\alpha_{i-1} \delta_{\tilde{\mathbf{s}}_{i-1}} + \mathbf{M}^{-1} \delta_{\mathbf{r}_i} - \delta_{\tilde{\mathbf{r}}_i}]. \end{aligned}$$

We note that for preconditioners with a simple structure, the errors induced by the preconditioner application may not be that large. In such cases it is reasonable to use  $\langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle$  and  $\langle \mathbf{r}_k, \tilde{\mathbf{s}}_k \rangle$  interchangeably. On the other hand, if the structure of the preconditioner is such that the size of the rounding errors associated with its application may be large, then it may be advantageous to use the original expression (2.2) rather than the reduced expression (2.3).

We now discuss the simplification of (2.2) used in M-CG and PIPE-M-CG. This expression relies on the exact arithmetic relation  $\langle \mathbf{s}_k, \tilde{\mathbf{r}}_k \rangle = \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ . Using the finite precision recurrence  $\alpha_k = \nu_k / \mu_k + \delta_{\alpha_k}$ , we compute

$$\alpha_k \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle = \left( \frac{\nu_k}{\mu_k} + \delta_{\alpha_k} \right) \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle = \left( \frac{\langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle + \delta_{\nu_k}}{\langle \mathbf{p}_k, \mathbf{s}_k \rangle + \delta_{\mu_k}} + \delta_{\alpha_k} \right) \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle.$$

In exact arithmetic,  $\langle \mathbf{p}_k, \mathbf{s}_k \rangle = \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle$ . However, in finite precision

$$\begin{aligned} \langle \mathbf{p}_k, \mathbf{s}_k \rangle &= \langle \tilde{\mathbf{r}}_k + \beta_k \mathbf{p}_{k-1} + \delta_{\mathbf{p}_k}, \mathbf{s}_k \rangle \\ &= \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle + \beta_k \langle \mathbf{p}_{k-1} + \delta_{\mathbf{p}_k}, \mathbf{A} \mathbf{p}_k + \delta_{\mathbf{s}_k} \rangle \\ &= \langle \tilde{\mathbf{r}}_k, \mathbf{s}_k \rangle + \beta_k \langle \mathbf{p}_{k-1}, \mathbf{A} \mathbf{p}_k \rangle + \langle \delta_{\mathbf{p}_k}, \mathbf{A} \mathbf{p}_k + \delta_{\mathbf{s}_k} \rangle + \langle \delta_{\mathbf{p}_k}, \delta_{\mathbf{s}_k} \rangle. \end{aligned}$$

This differs from the expression for  $\langle \tilde{\mathbf{s}}_k, \mathbf{r}_k \rangle$  in that it relies on understanding the size of  $\langle \mathbf{p}_{k-1}, \mathbf{A} \mathbf{p}_k \rangle$ , which should be small due to the orthogonality of certain vectors. While expressions for local orthogonality of such vectors exist for HS-CG [26, Proposition 5.19], these vectors need not be particularly orthogonal in other variants. Indeed, it is easy to verify that in these variants such local orthogonality can be quite weak.

Finally, we make two remarks about previously studied communication-hiding variants. First, the formula for  $\mu_k$  in CG-CG and GV-CG is derived in a similar way to (2.2) but then simplified. If left unsimplified, convergence of CG-CG appears to be slightly improved on difficult problems. However, using the unsimplified expression in GV-CG does not seem to lead to improved convergence.

Second, the predict-and-recompute idea may be applied to GV-CG by recomputing  $\mathbf{s}_k = \mathbf{A}\mathbf{p}_k$  and/or  $\mathbf{u}_k = \mathbf{A}\mathbf{s}_k$  or any of the recursively computed inner products. While recomputing both  $\mathbf{s}_k$  and  $\mathbf{u}_k$  seems to result in even worse final accuracy, recomputing one or the other leads to a better rate of convergence and final accuracy than GV-CG on some problems (although neither is as good as PIPE-PR-CG). Recomputing inner products was not observed to have a significant effect. Further study may be of value.

As an afterthought, we suggest that it may be possible to procedurally generate mathematically equivalent CG variants and then automatically check if they have improved convergence properties. Perhaps by finding many variants which work well, the similarities between them could provide insights into necessary properties for a good finite precision CG variant.

**Appendix D. Previously studied communication-hiding variants.** While we omitted the full descriptions of M-CG [25], PIPE-M-CG, CG-CG [6], and GV-CG [16] in the main paper, we include them here for completeness. We additionally include the initialization required to fully implement these variants. Note that not all variants require all of the variables defined in Algorithm D.1. For instance, HS-CG does not use any of the variables after  $\alpha_0$ .

---

**Algorithm D.1** Initializations
 

---

```

1: procedure INITIALIZE
2:    $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ ,  $\tilde{\mathbf{r}}_0 = \mathbf{M}^{-1}\mathbf{r}_0$ ,  $\nu_0 = \langle \mathbf{r}_0, \tilde{\mathbf{r}}_0 \rangle$ ,  $\mathbf{p}_0 = \tilde{\mathbf{r}}_0$ ,  $\mathbf{s}_0 = \mathbf{A}\mathbf{p}_0$ ,  $\alpha_0 = \nu_0 / \langle \mathbf{p}_0, \mathbf{s}_0 \rangle$ ,
    $\mathbf{w}_0 = \mathbf{A}\tilde{\mathbf{r}}_0$ ,  $\tilde{\mathbf{w}}_0 = \mathbf{M}^{-1}\mathbf{w}_0$ ,  $\tilde{\mathbf{s}}_0 = \mathbf{M}^{-1}\mathbf{s}_0$ ,  $\mathbf{u}_0 = \mathbf{A}\tilde{\mathbf{s}}_0$ ,  $\tilde{\mathbf{u}}_0 = \mathbf{M}^{-1}\mathbf{u}_0$ ,  $\tilde{\mathbf{s}}_0 = \mathbf{M}^{-1}\mathbf{s}_0$ ,
    $\alpha_0 = \nu_0 / \langle \mathbf{p}_0, \mathbf{s}_0 \rangle$ ,  $\sigma_0 = \langle \tilde{\mathbf{r}}_0, \mathbf{s}_0 \rangle$ ,  $\gamma_0 = \langle \tilde{\mathbf{s}}_0, \mathbf{s}_0 \rangle$ 
3: end procedure
    
```

---



---

**Algorithm D.2** Meurant Conjugate Gradient
 

---

```

1: procedure M-CG( $\mathbf{A}$ ,  $\mathbf{M}$ ,  $\mathbf{b}$ ,  $\mathbf{x}_0$ )
2:   INITIALIZE()
3:   for  $k = 1, 2, \dots$  do
4:      $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1}$ 
5:      $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_{k-1} - \alpha_{k-1}\tilde{\mathbf{s}}_{k-1}$ 
6:      $\nu'_k = -\nu_{k-1} + \alpha_{k-1}^2\gamma_{k-1}$ 
7:      $\beta_k = \nu'_k / \nu_{k-1}$ 
8:      $\mathbf{p}_k = \tilde{\mathbf{r}}_k + \beta_k\mathbf{p}_{k-1}$ 
9:      $\mathbf{s}_k = \mathbf{A}\mathbf{p}_k$ ,  $\tilde{\mathbf{s}}_k = \mathbf{M}^{-1}\mathbf{s}_k$ 
10:     $\mu_k = \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ ,  $\gamma_k = \langle \tilde{\mathbf{s}}_k, \mathbf{s}_k \rangle$ ,  $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ 
11:     $\alpha_k = \nu_k / \mu_k$ 
12:   end for
13: end procedure
    
```

---

**Algorithm D.3** Chronopoulos and Gear Conjugate Gradient

---

```

1: procedure CG-CG( $\mathbf{A}$ ,  $\mathbf{M}$ ,  $\mathbf{b}$ ,  $\mathbf{x}_0$ )
2:   INITIALIZE()
3:   for  $k = 1, 2, \dots$  do
4:      $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1}$ 
5:      $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{r}}_k = \mathbf{M}^{-1}\mathbf{r}_k$ 
6:      $\mathbf{w}_k = \mathbf{A}\tilde{\mathbf{r}}_k$ 
7:      $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ ,  $\eta_k = \langle \tilde{\mathbf{r}}_k, \mathbf{w}_k \rangle$ 
8:      $\beta_k = \nu_k / \nu_{k-1}$ 
9:      $\mathbf{p}_k = \tilde{\mathbf{r}}_k + \beta_k\mathbf{p}_{k-1}$ 
10:     $\mathbf{s}_k = \mathbf{w}_k + \beta_k\mathbf{s}_{k-1}$ 
11:     $\mu_k = \eta_k - (\beta_k / \alpha_{k-1})\nu_k$ 
12:     $a_k = \nu_k / \mu_k$ 
13:  end for
14: end procedure

```

---

**Algorithm D.4** Pipelined Predict-and-Recompute Meurant Conjugate Gradient

---

```

1: procedure PIPE-M-CG( $\mathbf{A}$ ,  $\mathbf{M}$ ,  $\mathbf{b}$ ,  $\mathbf{x}_0$ )
2:   INITIALIZE()
3:   for  $k = 1, 2, \dots$  do
4:      $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1}$ 
5:      $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_{k-1} - \alpha_{k-1}\tilde{\mathbf{s}}_{k-1}$ 
6:      $\mathbf{w}'_k = \mathbf{w}_{k-1} - \alpha_{k-1}\mathbf{u}_{k-1}$ ,  $\tilde{\mathbf{w}}'_k = \tilde{\mathbf{w}}_{k-1} - \alpha_{k-1}\tilde{\mathbf{u}}_{k-1}$ 
7:      $\nu'_k = -\nu_{k-1} + \alpha_{k-1}^2\gamma_{k-1}$ 
8:      $\beta_k = \nu'_k / \nu_{k-1}$ 
9:      $\mathbf{p}_k = \tilde{\mathbf{r}}_k + \beta_k\mathbf{p}_{k-1}$ 
10:     $\mathbf{s}_k = \mathbf{w}'_k + \beta_k\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{s}}_k = \tilde{\mathbf{w}}'_k + \beta_k\tilde{\mathbf{s}}_{k-1}$ 
11:     $\mathbf{u}_k = \mathbf{A}\tilde{\mathbf{s}}_k$ ,  $\tilde{\mathbf{u}}_k = \mathbf{M}^{-1}\mathbf{u}_k$ 
12:     $\mathbf{w}_k = \mathbf{A}\tilde{\mathbf{r}}_k$ ,  $\tilde{\mathbf{w}}_k = \mathbf{M}^{-1}\mathbf{w}_k$ 
13:     $\mu_k = \langle \mathbf{p}_k, \mathbf{s}_k \rangle$ ,  $\gamma_k = \langle \tilde{\mathbf{s}}_k, \mathbf{s}_k \rangle$ ,  $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ 
14:     $\alpha_k = \nu_k / \mu_k$ 
15:  end for
16: end procedure

```

---

**Algorithm D.5** Ghysels and Vanroose Conjugate Gradient

---

```

1: procedure GV-CG( $\mathbf{A}$ ,  $\mathbf{M}$ ,  $\mathbf{b}$ ,  $\mathbf{x}_0$ )
2:   INITIALIZE()
3:   for  $k = 1, 2, \dots$  do
4:      $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_{k-1}\mathbf{p}_{k-1}$ 
5:      $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1}\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{r}}_k = \tilde{\mathbf{r}}_{k-1} - a_{k-1}\tilde{\mathbf{s}}_{k-1}$ 
6:      $\mathbf{w}_k = \mathbf{w}_{k-1} - a_{k-1}\mathbf{u}_{k-1}$ ,  $\tilde{\mathbf{w}}_k = \mathbf{M}^{-1}\mathbf{w}_k$ 
7:      $\nu_k = \langle \tilde{\mathbf{r}}_k, \mathbf{r}_k \rangle$ ,  $\eta_k = \langle \tilde{\mathbf{r}}_k, \mathbf{w}_k \rangle$ 
8:      $\mathbf{t}_k = \mathbf{A}\tilde{\mathbf{w}}_k$ 
9:      $\beta_k = \nu_k / \nu_{k-1}$ 
10:     $\mathbf{p}_k = \tilde{\mathbf{r}}_k + \beta_k\mathbf{p}_{k-1}$ 
11:     $\mathbf{s}_k = \mathbf{w}_k + \beta_k\mathbf{s}_{k-1}$ ,  $\tilde{\mathbf{s}}_k = \tilde{\mathbf{w}}_k + \beta_k\tilde{\mathbf{s}}_{k-1}$ 
12:     $\mathbf{u}_k = \mathbf{t}_k + \beta_k\mathbf{u}_{k-1}$ 
13:     $\mu_k = \eta_k - (\beta_k / \alpha_{k-1})\nu_k$ 
14:     $\alpha_k = \nu_k / \mu_k$ 
15:  end for
16: end procedure

```

---

**Acknowledgments.** The first author expresses gratitude to their advisor, Anne Greenbaum, for encouraging them to pursue the side projects which gave rise to the results of this paper. We also thank everyone who provided feedback on early drafts, as well as the referees, whose feedback significantly improved several aspects of this paper.

## REFERENCES

- [1] T. J. ASHBY, P. GHYSELS, W. HEIRMAN, AND W. VANROOSE, *The impact of global communication latency at extreme scales on Krylov methods*, in Algorithms and Architectures for Parallel Processing, Y. Xiang, I. Stojmenovic, B. O. Apduhan, G. Wang, K. Nakano, and A. Zomaya, eds., Springer, Berlin, 2012, pp. 428–442.
- [2] G. BALLARD, E. C. CARSON, J. DEMMEL, M. HOEMMEN, N. KNIGHT, AND O. SCHWARTZ, *Communication lower bounds and optimal algorithms for numerical linear algebra*, Acta Numer., 23 (2014), p. 1–155.
- [3] R. F. BOISVERT, R. POZO, K. REMINGTON, R. F. BARRETT, AND J. J. DONGARRA, *Matrix Market: A web resource for test matrix collections*, in Quality of Numerical Software, Springer, Berlin, 1997, pp. 125–137.
- [4] E. C. CARSON, M. ROZLOŽNÍK, Z. STRAKOŠ, P. TICHÝ, AND M. TŮMA, *The numerical stability analysis of pipelined conjugate gradient methods: Historical context and methodology*, SIAM J. Sci. Comput., 40 (2018), pp. A3549–A3580, <https://doi.org/10.1137/16M1103361>.
- [5] A. CHRONOPOULOS, *A Class of Parallel Iterative Methods Implemented on Multiprocessors*, Ph.D. thesis, University of Illinois, Chicago, IL, 1987.
- [6] A. CHRONOPOULOS AND C. W. GEAR, *s-step iterative methods for symmetric linear systems*, J. Comput. Appl. Math., 25 (1989), pp. 153–168.
- [7] S. COOLS, J. CORNELIS, AND W. VANROOSE, *Numerically stable recurrence relations for the communication hiding pipelined conjugate gradient method*, IEEE Trans. Parallel Distrib. Syst., 30 (2019), pp. 2507–2522.
- [8] S. COOLS, E. F. YETKIN, E. AGULLO, L. GIRAUD, AND W. VANROOSE, *Analyzing the effect of local rounding error propagation on the maximal attainable accuracy of the pipelined conjugate gradient method*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 426–450, <https://doi.org/10.1137/17M1117872>.
- [9] S. COOLS AND W. VANROOSE, *Numerically Stable Variants of the Communication-Hiding Pipelined Conjugate Gradients Algorithm for the Parallel Solution of Large Scale Symmetric Linear Systems*, preprint, <https://arxiv.org/abs/1706.05988>, 2017.
- [10] J. CORNELIS, S. COOLS, AND W. VANROOSE, *The Communication-Hiding Conjugate Gradient Method with Deep Pipelines*, preprint, <https://arxiv.org/abs/1801.04728>, 2018.
- [11] L. D. DALCIN, R. R. PAZ, P. A. KLER, AND A. COSIMO, *Parallel distributed computing using Python*, Adv. Water Res., 34 (2011), pp. 1124–1139.
- [12] E. F. D’AZEVEDO AND C. H. ROMINE, *Reducing Communication Costs in the Conjugate Gradient Algorithm on Distributed Memory Multiprocessors*, Tech. report, Office of Scientific and Technical Information (OSTI), 1992, <https://doi.org/10.2172/10176473>.
- [13] J. DEMMEL, M. F. HOEMMEN, M. MOHIYUDDIN, AND K. A. YELICK, *Avoiding Communication in Computing Krylov Subspaces*, Tech. Report UCB/EECS-2007-123, EECS Department, University of California, Berkeley, CA, 2007.
- [14] J. DONGARRA, M. A. HEROUX, AND P. LUSZCZEK, *High-performance conjugate-gradient benchmark: A new metric for ranking high-performance computing systems*, Internat. J. High Performance Comput. Appl., 30 (2016), pp. 3–10.
- [15] P. R. ELLER AND W. GROPP, *Scalable non-blocking preconditioned conjugate gradient methods*, in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC’16, IEEE Press, Piscataway, NJ, 2016, pp. 18:1–18:12.
- [16] P. GHYSELS AND W. VANROOSE, *Hiding global synchronization latency in the preconditioned conjugate gradient algorithm*, Parallel Comput., 40 (2014), pp. 224–238.
- [17] A. GREENBAUM, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.
- [18] A. GREENBAUM, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551, <https://doi.org/10.1137/S0895479895284944>.
- [19] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997, <https://doi.org/10.1137/1.9781611970937>.



- [20] A. GREENBAUM, H. LIU, AND T. CHEN, *On the Convergence Rate of Variants of the Conjugate Gradient Algorithm in Finite Precision Arithmetic*, preprint, <https://arxiv.org/abs/1905.05874>, 2019.
- [21] M. H. GUTKNECHT AND Z. STRAKOŠ, *Accuracy of two three-term and three two-term recurrences for Krylov space solvers*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 213–229, <https://doi.org/10.1137/S0895479897331862>.
- [22] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
- [23] L. JOHNSON, *Highly concurrent algorithms for solving linear systems of equations*, in Elliptic Problem Solvers, Academic Press, Orlando, FL, 1984, pp. 105–126.
- [24] K. MCMANUS, S. JOHNSON, AND M. CROSS, *Communication latency hiding in a parallel conjugate gradient method*, in Eleventh International Conference on Domain Decomposition Methods (London, 1998), DDM.org, Augsburg, 1999, pp. 306–313.
- [25] G. MEURANT, *Multitasking the conjugate gradient method on the Cray X-MP/48*, Parallel Comput., 5 (1987), pp. 267–280.
- [26] G. MEURANT, *The Lanczos and Conjugate Gradient Algorithms: From Theory to Finite Precision Computations*, SIAM, Philadelphia, 2006, <https://doi.org/10.1137/1.9780898718140>.
- [27] G. MEURANT AND Z. STRAKOŠ, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542.
- [28] C. C. PAIGE, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, Ph.D. thesis, University of London, 1971.
- [29] C. C. PAIGE, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra Appl., 34 (1980), pp. 235–258.
- [30] Y. SAAD, *Practical use of polynomial preconditionings for the conjugate gradient method*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 865–881, <https://doi.org/10.1137/0906059>.
- [31] Y. SAAD, *Krylov subspace methods on supercomputers*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1200–1232, <https://doi.org/10.1137/0910073>.
- [32] G. L. G. SLEIJPEN, H. A. VAN DER VORST, AND D. R. FOKKEMA, *BiCGstab(l) and other hybrid Bi-CG methods*, Numer. Algorithms, 7 (1994), pp. 75–109.
- [33] Z. STRAKOŠ, *On the real convergence rate of the conjugate gradient method*, Linear Algebra Appl., 154-156 (1991), pp. 535–549.
- [34] Z. STRAKOŠ AND A. GREENBAUM, *Open Questions in the Convergence Analysis of the Lanczos Process for the Real Symmetric Eigenvalue Problem*, preprint, University of Minnesota, 1992.
- [35] R. STRZODKA AND D. GODDEKE, *Pipelined mixed precision algorithms on FPGAs for fast and accurate PDE solvers from low precision components*, in 2006 14th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, 2006, pp. 259–270.
- [36] J. VAN ROSENDALE, *Minimizing inner product data dependencies in conjugate gradient iteration*, in International Conference on Parallel Processing (ICPP), 1983, pp. 44–46.