B639388 Subcontract Award LLNL-Charles University: Mixed Precision Numerical Linear Algebra

Quarter I Report: Initial Exploration of the Use of Mixed Precision in Iterative Solvers

Erin Carson and Tomáš Gergelits

1 Summary of activities

The first quarter of the project was primarily spent identifying potential projects at the intersection of finite precision analysis, mixed precision computation, and Krylov subspace methods. We summarize our findings in the remainder of the document. Other activities include attending biweekly xSDK meetings as well as contributing material to the technical report and journal versions of the multiprecision landscape paper.

The subsequent quarter will be spent selecting a subset of the described projects to focus on, performing initial numerical experiments to evaluate the potential for the use of mixed precision, and developing initial theoretical analysis.

2 Introduction and motivation

Recent developments in the availability of multi-precision hardware have led to large-scale efforts to develop its use within numerical linear algebra routines. The (potentially selective) use of low precision hardware offers potentially significant performance improvements.

Our focus here is on Krylov subspace methods, which is a general class of iterative methods commonly used for solving linear systems, eigenvalue problems, and least squares problems. These methods in general rely on the construction of a well-conditioned, typically orthonormal basis of the Krylov subspace

$$\mathcal{K}_n(A, v) = \text{span}\{v, Av, \ldots, A^{n-1}v\} \tag{2.1}$$

for a given matrix A and an initial vector v. Considering the linear algebraic system

$$Ax = b, \quad A \in \mathbb{R}^{N\times N}, \tag{2.2}$$

the n-th approximation is typically determined as

$$x_n \in x_0 + \mathcal{S}_n, \quad r_n \perp \mathcal{C}_n, \tag{2.3}$$

where $r_n = b - Ax_n$ is the residual and where the n-dimensional search space $\mathcal{S}_n$ and constraint space $\mathcal{C}_n$ are the Krylov subspace (2.1) or its variant. Depending on properties of the system matrix, mathematical optimality of the resulting method, or its storage requirements, many variants of Krylov subspace methods exist.

Although the methods are mathematically finite (as they find the solution in exact arithmetic in at most N iterations), the adaptation to the data A and b built into the projection process (2.3) often leads to sufficient accuracy of the approximate solution in many fewer iterations. Thus the overall cost of iterative computations using Krylov subspace methods is determined by both

a) *the cost per iteration*, and b) *the number of iterations*,

required for the convergence to the prescribed accuracy. While the cost per iteration depends on the efficiency of the implementation, the underlying hardware, storage requirements of the method, or specific nonzero structure of the problem, the number of iterations depends on the accuracy requirements of the application, numerical properties of the data, and the properties of the used method, with a special emphasis on the finite precision behavior of the particular algorithm chosen for the given method.

Classical implementations of Krylov subspace methods require one or more matrix-vector multiplications and one or more inner product operations in each iteration. Implementation of algorithms that adaptively vary the precision used in their computation may lead to significant performance improvement per iteration, e.g., by computing and/or storing the results of sparse matrix-vector multiplication or inner products in the lower precision arithmetic. However, as different algorithms and their practical implementation for mathematically equivalent methods may have significantly different numerical behavior, one cannot simply develop high-performance Krylov subspace methods by optimizing the cost per iteration. The development of high performance implementations of Krylov subspace methods requires a thorough understanding of finite precision

behavior. It is important to both determine the level of maximal attainable accuracy and the convergence behavior prior to the level of maximal attainable accuracy in a finite (mixed) precision setting.

We therefore study the effects of finite precision, in particular mixed precision, on the convergence rate and attainable accuracy within Krylov subspace methods. We have identified a number of potential projects within this scope. For convenience, we have grouped the potential projects based on whether they apply to long-term recurrence Krylov subspace methods, such as Arnoldi and GMRES, or short-term recurrence Krylov subspace methods, such as Lanczos and the conjugate gradient method (CG).

3 Krylov subspace methods based on long recurrences

Here we focus mainly on the GMRES method, introduced in [25] as a generalization of the method of minimized residuals (MINRES) to nonsymmetric linear algebraic systems. Given an $N \times N$ nonsingular matrix A, a right-hand side b, and an initial approximate solution x_0, GMRES computes an approximate solution x_n to the linear system $Ax = b$ in each iteration according to the projection process

$$x_n \in x_0 + \mathcal{K}_n(A, r_0), \qquad r_n \perp A\mathcal{K}_n(A, r_0),$$

where $r_0 = b - Ax_0$ is the initial residual, i.e., the projection process selects an approximate solution such that the corresponding residual is minimized over the shifted Krylov subspace. This results in the Arnoldi recurrence

$$AV_n = V_n H_{n,n} + h_{n+1,n} v_{n+1} e_n^T, \tag{3.1}$$

where, assuming exact arithmetic, V_n is an orthonormal basis for the Krylov subspace $\mathcal{K}_n(A, r_0)$ and $H_{n,n}$ is an upper Hessenberg matrix. The recurrence (3.1) is a matrix formulation of the Arnoldi algorithm [1], which can be viewed as a variant of the Gram-Schmidt orthogonalization method applied to the Krylov sequence in order to generate an orthonormal basis of $\mathcal{K}_n(A, r_0)$. The matrix $H_{n,n}$ can be viewed as the matrix representation of the orthogonal restriction of A to the invariant Krylov subspace in the basis V_n. Thus the Arnoldi algorithm can be seen as the unitary reduction of A to upper Hessenberg form.

As the matrix $H_{n,n}$ is of upper Hessenberg form, the algorithm is based on long recurrences, i.e., the computation of new vector v_{n+1} requires explicit orthogonalization of the vector Av_n against all previously computed basis vectors $v_1, \ldots, v_n$. As a consequence, the cost of one iteration of GMRES is not constant and increases significantly as the computation proceeds.

The finite precision behavior of these methods is largely dependent on the orthogonalization scheme used within construction of the Arnoldi recurrence. The potential projects described in this section thus largely revolve around studying the numerical properties of mixed precision orthogonalization schemes, as well as the numerical properties of variants of Arnoldi/GMRES developed for high-performance computations.

Summary of standard results

There are a number of existing results on the behavior of Krylov subspace methods based on long recurrences in (a single) finite precision. We briefly summarize relevant results for the Arnoldi/GMRES method, as these form the starting place for investigations into the use of mixed precision.

Much work on the finite precision analysis of GMRES relies on the fact that in exact arithmetic, the first n steps of the Arnoldi recurrence can be written as the QR factorization

$$[r_0, AV_n] = V_{n+1}[r_0 e_1, H_{n+1,n}], \qquad V_n^* V_n = I.$$

It is known, however, that the mathematical properties of the Arnoldi algorithm and GMRES method are not maintained in finite precision arithmetic. In particular, the basis V_n is no longer orthonormal, and the entries of the upper Hessenberg matrix will differ from their exact counterparts. In fact, the orthogonalization scheme used is crucial, and can result in different GMRES implementations with drastically different finite precision behavior.

Walker [27] first proposed an implementation based on Householder Arnoldi. Assuming certain restrictions on the condition number of A (which depend in turn on the machine precision used), Drkošová, Greenbaum, Rozložník and Strakoš proved in [6] that Householder Arnoldi GMRES is normwise backward stable, meaning that after at most N iterations, the normwise relative backward error is proportional to the machine precision. The proof is based on the result that the use of Householder reflections for performing orthogonalization results in the Arnoldi basis vectors maintaining orthogonality at the level of machine precision.

It is perhaps most common in practice to use a GMRES implementation based on the modified Gram-Schmidt implementation of Arnoldi (MGS-GMRES). When MGS is applied to the QR factorization of $[r_0, AV_n]$, it was shown by Greenbaum, Rozložník and Strakoš [13] that the loss of orthogonality among the MGS Arnoldi vectors in iteration n is proportional to the product of the machine precision, the condition number of A, and

the quantity $\|r_0\|/\|r_n\|$. This means that orthogonality is lost only once the relative residual norm becomes proportional to the product of the machine precision and the condition number of A. The work in [13] resulted in subsequent investigations which link the finite precision behavior of GMRES with solutions to total least squares problems; see, e.g., [22, 23, 24] and references therein. These results then eventually led to the work [21], which proves that MGS-GMRES is backward stable despite a gradual loss of orthogonality.

Backward stability of GMRES with mixed precision computations

In [3] and [4], it is shown that mixed-precision iterative refinement can achieve certain desired bounds on forward and backward error assuming some constraint on condition number of the coefficient matrix. The paper [3] also introduces GMRES-based iterative refinement; when preconditioned (mixed precision) GMRES is used as the correction solver within iterative refinement, the aforementioned constraint on condition number can be loosened. The analysis assumes that within GMRES, the preconditioned system is applied to a vector (implicitly) at double the working precision. This assumption is necessary in order to apply the existing GMRES backward error analysis in [21] to the preconditioned case.

Clearly, requiring that preconditioned solves in each iteration are performed in double the working precision is undesirable in practice for performance reasons. We therefore ask what are the implications of applying the preconditioner instead at the working precision. Higham has given a preliminary answer in [15], although the result is rather unsatisfying in that the analysis is quite loose. Improving the results in [15], if possible, would require revisiting and modifying the backward error analysis in [21].

This is a potential collaboration with Nicholas Higham and his team at the University of Manchester. The primary challenges are technical in nature, involving thoroughly revising and expanding the analysis of [21]; further, it is not clear that significant theoretical results are attainable, although experimental results seem to indicate that there is room for improvement.

The equivalence of classical KSMs in mixed precision arithmetic to high-performance variants

There have been many efforts to reformulate Krylov subspace methods in order to make them more suitable for use on high-performance computers. These include, for example, pipelined Krylov subspace methods (e.g., [8, 9]) and s-step Krylov subspace methods (e.g., [16]). Although such methods can effectively reduce the time per iteration, it is experimentally observed that in finite precision, both the number of iterations is increased and the attainable accuracy is decreased relative to the classical algorithm in many cases.

The understanding of finite precision behavior of high-performance variants in the literature is thus far incomplete. We pose the following questions:

- Can high performance variants such as pipelined GMRES and s-step GMRES be seen as being equivalent to classical GMRES in selective mixed precisions?
- Can higher precision be selectively used within pipelined GMRES and s-step GMRES such that the numerical behavior is close to that of classical GMRES? If so, can this be done without sacrificing performance (i.e., while still maintaining the benefits of reduced synchronization)?

Exploring this topic could allow us to apply some existing results on finite precision analysis and stability to the pipelined and s-step cases. The primary challenge is the framing of pipelined and s-step GMRES methods as classical methods, and extending existing finite precision analyses, incorporating both mixed precision as well as elements of the different algorithmic variants; it is not necessarily clear how to do this. Initial numerical experiments are needed to gauge whether the above conjectures are true and can lead to theoretical results. We note that although we have framed this project in terms of GMRES, analogous investigations could also be performed for CG.

Numerical stability of mixed precision low-synch orthogonalization methods

The recent paper [26] details low-synchronization variants of Gram-Schmidt orthogonalization and their use within GMRES. These variants have the benefit of requiring little communication (like classical Gram-Schmidt) but improved numerical behavior (closer to that of modified Gram-Schmidt). The results on the numerical behavior, however, are thus far only experimental.

This project involves performing error analysis to prove bounds on the loss of orthogonality and residual in these new approaches, and incorporating the use of mixed precision within low-synch orthogonalization schemes. Note that the methods presented in [26] are similar (some perhaps equivalent) to those analyzed by Barlow [2], so his work may provide a useful starting point.

This is a potential collaboration with Stephen Thomas from NREL (also on the multiprecision project), with whom we have had initial discussions on this topic. The primary challenges are understanding the work of Barlow and potentially extending or applying his results, as well as the theoretical analysis of whether mixed precision could be beneficial within low-synch methods; further numerical experiments are needed to confirm this.

Potential for mixed precision in block Gram-Schmidt orthogonalization

The description of a block Gram-Schmidt orthogonalization method involves the selection of 1) the choice of an inter-block orthogonalization method and 2) the choice of an intra-block orthogonalization method. These need not be the same method; for example, one could use classical Gram-Schmidt between blocks with TSQR used within blocks (this is what is commonly-used in s-step GMRES).

Ongoing work with Kathryn Lund at Charles University involves determining the numerical properties (in terms of both the loss of orthogonality and the residual) resulting from different combinations of inter- and intra-block routines, and determining what constraints on the intra-block orthogonalization routine result in a stable overall block orthogonalization routine. These questions are relevant especially in the context of Krylov subspace method variants like s-step GMRES, which requires the use of a low-synchronization block orthogonalization method in order to avoid communication; see [16]. The primary question here is whether low precision can be selectively used in parts of the block orthogonalization routine, and whether this has any benefit to performance.

This is a potential collaboration with Kathryn Lund. The primary challenge here is technical and involves extending existing results to the mixed precision case. Numerical experiments should first be performed to gauge potential benefit.

Relation between loss of orthogonality and normwise backward error in finite precision GMRES

This topic involves exploration of an open problem introduced in [22] involving the proof of a finite precision version of the relationship between the normwise relative backward error and the condition number of the augmented matrix describing the least squares problem solved in each iteration of GMRES.

It has previously been noted in [22] that, in exact arithmetic, we have the relationship

$$\beta(x_k)\kappa([v_1\rho_0, AV_k]) = O(1), \tag{3.2}$$

where the condition number κ denotes the ratio of largest to smallest singular values and $\beta(x_k)$ denotes the normwise relative backward error

$$\beta(x_k) = \frac{\|r_k\|}{\|b\| + \|A\|\|x_k\|}.$$

In exact arithmetic, we have $V_{k+1}^T V_{k+1} = I$. In finite precision, however, this ceases to hold, as the Arnoldi vectors will lose orthogonality due to rounding errors. The rate at which $\|I - V_{k+1}^T V_{k+1}\|$ grows will depend on the particular orthogonalization algorithm used; e.g., Householder, modified Gram-Schmidt, or classical Gram-Schmidt. Bounds on the loss of orthogonality due to finite precision have been previously derived. For example, it has been shown (see [21]) that with some restrictions on A, for modified Gram-Schmidt with unit roundoff ϵ, we have the bound

$$\|I - V_{k+1}^T V_{k+1}\|_F \leq \kappa([v_1\rho_0, AV_k])O(\epsilon).$$

If a finite precision version of (3.2) existed, we could then show that

$$\frac{\|r_k\|\|I - V_{k+1}^T V_{k+1}\|_F}{\|b\| + \|A\|\|x_k\|} \leq \beta(x_k)\kappa([v_1\rho_0, AV_k])O(\epsilon) = O(\epsilon).$$

In other words, there is no significantly loss of orthogonality until the backward error is small, and significant loss of orthogonality implies convergence (with no significant deterioration of convergence rate) and backward stability.

The first goal of this project involves proving a version of (3.2) with finite precision error taken into account. Experimental results presented in [22] suggest that such a relation exists, but its proof will require rigorous rounding error analysis, particularly in regard to the effect of the loss of orthogonality on the size of $\|b - Ax_k\|$. This was originally planned for the paper [21], but did not appear there.

This first result would prove that the basis vectors can lose orthogonality at a rate proportional to rate of residual decrease without sacrificing backward stability. The second goal of this project is to use this result to say in what precision a certain orthogonalization scheme must be performed in order to guarantee backward stability of GMRES. For example, it could be the case that we could use Householder orthogonalization in increasingly low precision and still obtain a backward stable GMRES method. This project is closely related to the work in [10], which, although incomplete from a theoretical perspective, could be a useful starting point.

4 Krylov subspace methods based on short recurrences

Here we focus on the method of conjugate gradients (CG) introduced in [14] or the mathematically equivalent method based on the Lanczos algorithm introduced in [17]. Given an $N \times N$ symmetric and positive definite matrix A, a right-hand side b, and an initial approximation x_0, CG computes an approximate solution to the linear system (2.2) according to the projection process

$$x_n \in x_0 + \mathcal{K}_n(A, r_0), \qquad r_n \perp \mathcal{K}_n(A, r_0),$$

where r_0 is the initial residual. In other words, in each iteration, CG adds to the initial approximation x_0 the vector from $\mathcal{K}_n(A, r_0)$ that minimizes the energy norm of the error. Equivalently, the n-th CG approximation is determined by $x_n = V_n t_n$, where t_n is the solution of the $n \times n$ problem with the Jacobi matrix T_n from the matrix formulation of the Lanczos algorithm

$$AV_n = V_n T_n + \beta_{n+1} v_{n+1} e_n^T.$$

The matrix V_n of Lanczos vectors $v_1, \ldots, v_n$ is an orthonormal basis of the Krylov subspace $\mathcal{K}_n(A, r_0)$ and the tridiagonal matrix T_n can be viewed as the matrix representation of the orthogonal restriction of A to $\mathcal{K}_n(A, r_0)$. Its eigenvalues are called Ritz values and they tend to approximate the eigenvalues of A.

The standard implementation of CG includes three two-term recurrences. Similarly, in the Lanczos algorithm the new Lanczos vector v_{n+1} is computed by orthogonalization of Av_n against only two previous vectors v_n and v_{n-1}, i.e., the implementation is based on short recurrences and the computational cost per iteration is constant. However, while in exact arithmetic the use of short recurrences ensures the global orthogonality of the computed basis vectors, the presence of rounding errors in practical computations leads to the loss of the global orthogonality and to significantly different numerical behavior. Thus, the methods do not fulfill their theoretical properties in practical computations.

In this part of the project, the aim is to study numerical behavior of these methods implemented using some variant of mixed precision arithmetic. As the actual convergence behavior substantially depends on rounding errors, their size and their propagation throughout the computations, without proper rounding error analysis it is unclear how the numerical behavior of these methods implemented using mixed precision arithmetic differs from standard implementations and/or their theoretical properties. A crucial practical question is whether and how additional inexactness affects the convergence behavior before the maximal attainable accuracy is reached. In particular:

- Is the possible effect of low precision on convergence rate too detrimental such that any advantages of algorithmically efficient implementations are lost?
- Viceversa, can the (partial) use of extended precision arithmetic lead to a significant reduction in the number of iterations needed to achieve the desired accuracy?

Investigation of these topics is important for relevant evaluation and comparison of different implementations. After a brief summary of the most important results on the finite precision behavior of Lanczos and CG methods, we formulate the studied questions in more detail. We believe that these existing results can be extended to the analysis of computations in mixed precision arithmetic.

Summary of standard results

There is a huge literature on the finite precision behavior of Lanczos-based methods; see, e.g., the extensive review [19] and reference therein. Here we focus on the important results achieved in particular by Paige and Greenbaum.

In finite precision computations, the orthogonality of computed Lanczos vectors is usually very quickly lost and they often become even (numerically) highly linearly dependent. In the Lanczos algorithm the loss of global orthogonality of the computed Lanczos vectors results in multiple approximations of the individual eigenvalues of A by the eigenvalues of the computed matrix T_n, which leads to *delay* in the approximation of some other eigenvalues. Subsequently, in finite precision CG computations the computed residual vectors then span a subspace of dimension smaller than the given iteration step. This *rank-deficiency* of computed Krylov subspace bases thus determines *delay of convergence* of finite precision computations, which can be defined as the difference between the number of iterations required to attain a prescribed accuracy in finite precision computations and the number of iterations required to attain the same accuracy assuming exact arithmetic.

Nearly all developments in the analysis of rounding errors in the Lanczos algorithm and CG are based on the results of Paige; see e.g., [20]. He starts with bounds on the elementary round-off errors at each iteration, where he defines different quantities ε_0 and ε_1 used for errors in computation of inner products and of matrix-vector products, respectively. His analysis concludes with elegant mathematical results which link convergence of the

computed eigenvalue approximations to the loss of orthogonality. In particular, he shows that the orthogonality of the computed Lanczos vectors can be lost only in the direction of vectors associated with converged Ritz values.

Building on the results of Paige, the seminal paper of Greenbaum [11] develops a backward-like analysis of the Lanczos algorithm (and also of the closely related CG). In short, she proves that the finite precision CG computation behaves like the exact CG computations for a larger matrix having its eigenvalues located in tight clusters around the eigenvalues of the original matrix. The diameter of these clusters depends on the matrix properties and the machine precision used in the computations. Greenbaum also proved results on the maximum attainable accuracy in finite precision for the CG method and other methods based on the Lanczos algorithm [12].

Analysis of selective mixed precision in the Lanczos algorithm and the CG method

In the PhD thesis [5] it is shown that the results of Paige and Greenbaum can be extended to the so called s-step variants of the Lanczos algorithm and the CG method. In particular, it is shown that s-step Lanczos algorithm in finite precision arithmetic behaves like classical Lanczos run in a lower "effective" precision, where this "effective" precision depends on the conditioning of the polynomials used to generate the s-step bases.

We believe that a similar approach can be used to study in detail the rounding errors and their propagation in implementations using mixed precision arithmetic. Moreover, we would like to investigate in detail the rounding errors in the pipelined variant of CG. Detail and rigorous rounding error analysis for mixed precision implementations of CG and the Lanczos algorithm could provide (partial) answers to the following questions:

- How does the maximal attainable accuracy depend on the use of lower precision in computations of matrix-vector products or inner products?
- How does the use of mixed precision affect the aforementioned results of Paige? In particular, what are the relevant criteria for convergence of Ritz values and what is the structure of the loss of orthogonality?
- Could the (partial) use of extended precision in sparse matrix-vector multiplications or in the computation of inner products lead to significant reduction of the loss of orthogonality and the consequent delay of convergence?
- Could we obtain the extension of Greenbaum's backward stability-like results for the CG method in mixed precision? How do the diameters of the Ritz value clusters depend on the various precisions used?

As a first step, we will need to distinguish between different sizes of local rounding errors associated with computing matrix-vector and inner products. Thus, the plan is to revisit Paige's approach and work with both aforementioned quantities ε_0 and ε_1 throughout the whole rounding error analysis. Our theoretical investigations will be accompanied by the numerical experiments.

Searching for relationships between implementations in different precisions

Motivated by the results of Greenbaum and by the goal of making links between seemingly different objects, which often leads to their better understanding, we would like to study the possible relationship between different implementations of the Lanczos algorithm and the CG method. We can ask the same questions formulated above for the GMRES method.

- Can high performance variants such as pipelined or s-step CG be seen as being equivalent to classical CG in selective mixed precisions? Can the (selective) use of higher precision within pipelined or s-step CG lead to numerical behavior close to classical CG?

Although the results of Greenbaum do not allow for direct comparison of the approximation or residual vectors (as they involve extended matrices), they give sufficient reasoning to relate the energy norm of the error in the n-th iteration of finite precision CG computations with an earlier l-th iteration of the exact CG computations with the *same data*; see, e.g., [18, Section 5.9]. Taking into the account the gap $n - l$, which represents the delay of convergence corresponding to the rank-deficiency of the computed Krylov subspace, we may ask the following:

- How do the approximation vectors computed in lower/mixed precision resemble their (earlier) counterparts from exact CG computations with the same matrix, right-hand side, and initial approximation?
- Related to the above, how do the subspaces generated by the CG residuals computed in different precisions (lower/standard/mixed) differ from the exact Krylov subspaces?

The recent conference proceedings contribution [7] suggests that even the computed CG approximation vectors as well as residual vectors are comparable to their exact arithmetic counterparts. A similar approach might be used to relate entities from computations in mixed precision to their exact arithmetic counterparts or to entities from computations in a standard double precision.

As it is not completely clear at the moment how to approach these questions rigorously from a theoretical perspective, the emphasis will be in the beginning on performing a variety of numerical experiments.

References

[1] W. E. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.

[2] J. L. Barlow. Block modified Gram-Schmidt algorithms and their analysis. *SIAM J. Matrix Anal. Appl.*, 40(4):1257–1290, 2019.

[3] E. Carson and N. J. Higham. A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems. *SIAM J. Sci. Comput.*, 39(6):A2834–2856, 2017.

[4] E. Carson and N. J. Higham. Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM J. Sci. Comput.*, 40(2):A817–A847, 2018.

[5] E. C. Carson. *Communication-avoiding Krylov subspace methods in theory and practice.* PhD thesis, UC Berkeley, 2015.

[6] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical stability of GMRES. *BIT Numer. Math.*, 35(3):309–330, 1995.

[7] T. Gergelits, I. Hnětynková, and M. Kubínová. Relating computed and exact entities in methods based on Lanczos tridiagonalization. In et al. Kozubek, T., editor, *High Performance Computing in Science and Engineering*, volume 11087 of *Lecture Notes in Computer Science*, pages 73–87, Cham, 2018. Springer.

[8] P. Ghysels, T. J. Ashby, K. Meerbergen, and W. Vanroose. Hiding global communication latency in the GMRES algorithm on massively parallel machines. *SIAM J. Sci. Comput.*, 35(1):C48–C71, 2013.

[9] P. Ghysels and W. Vanroose. Hiding global synchronization latency in the preconditioned conjugate gradient algorithm. *Parallel Computing*, 40(7):224–238, 2014.

[10] S. Gratton, E. Simon, D. Titley-Peloquin, and P. Toint. Exploiting variable precision in GMRES. *arXiv preprint arXiv:1907.10550*, 2019.

[11] A. Greenbaum. Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Lin. Alg. Appl.*, 113:7–63, 1989.

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

[13] A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical behaviour of the modified Gram-Schmidt GMRES implementation. *BIT Numer. Math.*, 37(3):706–719, 1997.

[14] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436, 1952.

[15] N. J. Higham. Error analysis for standard and GMRES-based iterative refinement in two and three-precisions. Technical Report MIMS EPrint 2019.19, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, 2019.

[16] M. Hoemmen. *Communication-avoiding Krylov subspace methods.* PhD thesis, UC Berkeley, 2010.

[17] C. Lanczos. Solution of systems of linear equations by minimized iterations. *J. Research Nat. Bur. Standards*, 49:33–53, 1952.

[18] J. Liesen and publisher = Oxford University Press title = Krylov subspace methods: principles and analysis year = 2013 address = Oxford isbn = 978-0-19-965541-0 series = Numerical Mathematics and Scientific Computation doi = 10.1093/acprof:oso/9780199655410.001.0001 groups = Krylov subspace methods mrclass = 65F10 (65F15) mrnumber = 3024841 mrreviewer = Melina A. Freitag pages = xvi+391 Strakoš, Z.

[19] G. Meurant and journal = Acta Numer. title = The Lanczos and conjugate gradient algorithms in finite precision arithmetic year = 2006 issn = 0962-4929 pages = 471–542 volume = 15 doi = 10.1017/S096249290626001X file = :2006MeuraStra.pdf:PDF fjournal = Acta Numerica groups = CG Lanczos isbn = 0-521-86815-7 mrclass = 65F15 (65F10 65G50) mrnumber = 2269746 (2007m:65031) mrreviewer = A. Bultheel Strakoš, Z.

[20] C. C. Paige. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Lin. Alg. Appl.*, 34:235–258, 1980.

[21] C. C. Paige, M. Rozložník, and Z. Strakoš. Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES. *SIAM J. Matrix Anal. Appl.*, 28(1):264–284, 2006.

[22] C. C. Paige and Z. Strakoš. Residual and backward error bounds in minimum residual Krylov subspace methods. *SIAM J. Sci. Comput.*, 23(6):1898–1923, 2002.

[23] C. C. Paige and Z. Strakoš. Residual and backward error bounds in minimum residual Krylov subspace methods. *SIAM J. Sci. Comput.*, 23(6):1898–1923, 2002.

[24] C. C. Paige and Z. Strakoš. Core problems in linear algebraic systems. *SIAM J. Matrix Anal. Appl.*, 27(3):861–875, 2005.

[25] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving] nonsymmetric linear systems, year = 1986, issn = 0196-5204, number = 3, pages = 856–869, volume = 7, coden = SIJCD4, fjournal = Society for Industrial and Applied Mathematics. Journal on Scientific and Statistical Computing, mrclass = 65F50 (65F10), mrnumber = 87g:65064, publisher = SIAM,. *SIAM J. Sci. Statist. Comput.*

[26] K. Swirydowicz, J. Langou, S. Ananthan, U. Yang, and S. Thomas. Low synchronization GMRES algorithms. *arXiv preprint arXiv:1809.05805*, 2018.

[27] H. F. Walker. Implementation of the GMRES method using Householder transformations. *SIAM J. Sci. Stat. Comput.*, 9(1):152–163, 1988.