## RESIDUAL-BASED CHEBYSHEV FILTERED SUBSPACE ITERATION FOR SPARSE HERMITIAN EIGENVALUE PROBLEMS TOLERANT TO INEXACT MATRIX-VECTOR PRODUCTS

NIKHIL KODALI<sup>\*</sup>, KARTICK RAMAKRISHNAN<sup>†</sup>, AND PHANI MOTAMARRI<sup>††</sup>

Abstract. Chebyshev Filtered Subspace Iteration (ChFSI) has been widely adopted for computing a small subset of extreme eigenvalues in large sparse matrices. This work introduces a residualbased reformulation of ChFSI, referred to as R-ChFSI, designed to accommodate inexact matrixvector products while maintaining robust convergence properties. By reformulating the traditional Chebyshev recurrence to operate on residuals rather than eigenvector estimates, the R-ChFSI approach effectively suppresses the errors made in matrix-vector products, improving the convergence behavior for both standard and generalized eigenproblems. This ability of R-ChFSI to be tolerant to inexact matrix-vector products allows one to incorporate approximate inverses for large-scale generalized eigenproblems, making the method particularly attractive where exact matrix factorizations or iterative methods become computationally expensive for evaluating inverses. It also allows us to compute the matrix-vector products in lower-precision arithmetic allowing us to leverage modern hardware accelerators. Through extensive benchmarking, we demonstrate that R-ChFSI achieves desired residual tolerances while leveraging low-precision arithmetic. For problems with millions of degrees of freedom and thousands of eigenvalues, R-ChFSI attains final residual norms in the range of  $10^{-12}$  to  $10^{-14}$ , even with FP32 and TF32 arithmetic, significantly outperforming standard ChFSI in similar settings. In generalized eigenproblems, where approximate inverses are used, R-ChFSI achieves residual tolerances up to ten orders of magnitude lower, demonstrating its robustness to approximation errors. By efficiently utilizing modern hardware accelerators and reducing reliance on high-precision arithmetic, R-ChFSI provides a scalable and computationally efficient alternative for solving large-scale eigenproblems in high-performance computing environments.

**Key words.** generalized eigenvalue problems, inexact matrix products, Chebyshev filtered subspace iteration, mixed-precision methods

**MSC codes.** 65F15, 68W25, 65N25

**1.** Introduction. Large sparse Hermitian eigenproblems arise in a wide range of scientific and engineering applications, including data science and machine learning. Here, the goal is to compute only a small fraction of the extreme eigenpairs – often less than 0.5% of the matrix size. Specific examples include the discretization of partial differential equations using localized basis sets or grid-based approaches, as well as problems in graph theory and network analysis. The solution strategies for such eigenproblems often rely on fully iterative methods and are usually based on iterative orthogonal projection approaches. In these approaches, the large sparse matrix is orthogonally projected onto a carefully constructed smaller subspace rich in the wanted eigenvectors (Rayleigh-Ritz step), followed by subspace diagonalization of the projected matrix and a subspace rotation step to recover the desired orthogonal eigenvector estimates of the original sparse Hermitian matrix. Popular iterative approaches include Davidson [4, 8], Generalized-Davidson [27], Jacobi-Davidson [15], Chebyshev-filtered subspace iteration (ChFSI) approach [48], LOBPCG [17], and PPCG [39]. Another key class of iterative techniques is based on Krylov subspace methods, namely the Arnoldi method [1], Lanczos methods [20] and their important variants, including implicit restart Arnoldi methods [35], Krylov-Schur method [36] and block-Krylov methods [33]. Our focus in this work involves the solution of large, sparse Hermitian eigenproblems using the ChFSI approach.

Chebyshev filtered subspace iteration (ChFSI) relies on a Chebyshev filtering pro-

<sup>\*</sup>MATRIX Lab, Department of Computational and Data Sciences, Indian Institute of Science. <sup>†</sup>phanim@iisc.ac.in

cedure that constructs a subspace rich in the desired eigenvectors exploiting the fast growth property of Chebyshev polynomials outside the interval [-1,1], followed by a Rayleigh-Ritz step. Approaches based on ChFSI have become quite popular recently to solve both standard and generalized eigenproblems [23, 47, 29, 21]. Although variants of ChFSI approaches combined with the Davidson method [46, 24, 25, 40] have been proposed, the ChFSI approach remains the preferred choice in electronic structure codes based on density functional theory for solving the underlying nonlinear eigenvalue problem [7, 13, 45, 12] due to its scalability, ability to accommodate evolving subspaces and computational efficiency [6, 5]. Additionally, ChFSI is memory efficient compared to Davidson-type approaches and further offers an attractive alternative to preconditioned-type conjugate gradient approaches for problems, especially for problems where good preconditioners are unavailable. The current work fills an important gap in the ChFSI approach by developing a residual-based approach to ChFSI that is tolerant to inexact matrix-vector products during the subspace construction step. Most importantly, we demonstrate that this approach is naturally amenable to the use of inexpensive matrix inverses while employing the ChFSI method to solve large sparse matrix generalized eigenproblems without affecting the convergence behaviour. We also show that such an approach can be robust while leveraging low-precision matrixvector products. These capabilities are especially important in light of recent changes in modern heterogeneous computing architectures.

Modern hardware architectures have undergone substantial modifications in recent years due to the computational requirements of machine learning (ML) and artificial intelligence (AI) training. Owing to their high computational demands, these domains have gravitated towards the use of low-precision arithmetic for training and inferencing. In response to this demand, hardware manufacturers have been enhancing support for low-precision floating-point formats, such as tensorfloat32 and bfloat16, enabling significantly faster throughput and substantial performance improvements. For instance, NVIDIA's Blackwell GPUs, designed for AI/ML applications, demonstrate a notable decrease in peak double-precision (FP64) floating-point performance [30] compared to the previous Hopper architecture. These architectural changes underscore the necessity to modify scientific computing algorithms to efficiently utilize low-precision processes without compromising accuracy [11, 16]. For sparse eigensolvers, the ideas of mixed precision preconditioning for the LOBPCG eigensolver [18] and mixed precision orthogonalization and Rayleigh-Ritz have been explored for the LOBPCG and ChFSI eigensolvers [18, 7, 28]. We note that in these works the evaluation of matrix-vector products involved in the subspace construction still need to be performed in the higher precision arithmetic. We further note that the capacity to accommodate inexact matrix-vector products facilitated by lowprecision arithmetic can yield substantial performance enhancements. Algorithms that can accommodate inexact matrix-vector products while ensuring that errors remain within acceptable tolerances can efficiently utilize the capabilities of modern hardware, thereby reducing computational costs. The development of such algorithms becomes particularly important in light of the evolving hardware landscape.

In this work, we introduce a residual-based reformulation of the Chebyshev filtered subspace iteration (ChFSI) method [32, 47, 48, 45], referred to as the R-ChFSI method, for solving large-scale sparse Hermitian eigenvalue problems. The key novelty of R-ChFSI lies in its ability to accommodate inexact matrix-vector products while preserving convergence properties. This ability makes the proposed method particularly relevant for modern hardware architectures that favor low-precision arithmetic, significantly improving computational efficiency. The proposed R-ChFSI method is also well-suited for generalized eigenvalue problems, as it can employ approximate matrix inversion and avoid expensive matrix factorizations and iterative solvers that are otherwise typically required for such problems [21]. The core of our approach is a reformulated recurrence relation that modifies the standard ChFSI update step to operate on residuals rather than the guess of the eigenvectors. We then provide a mathematical justification demonstrating that this reformulation reduces the numerical error in the Chebyshev filtered subspace construction while employing inexact matrix-vector products compared to the traditional ChFSI recurrence relation. Our analysis establishes that the proposed R-ChFSI method effectively controls error propagation in the Chebyshev filtering process, resulting in more reliable convergence.

Through extensive benchmarking, we observe that R-ChFSI achieves comparable or superior residual tolerances while maintaining computational efficiency. Specifically, for benchmark systems with up to 7.6 million degrees of freedom (DoFs) and 14,000 desired smallest eigenpairs, the FP32 and TF32 variants of R-ChFSI achieve final residual norms within  $10^{-12}$  to  $10^{-14}$ , whereas standard ChFSI with the same precision fails to converge below  $10^{-6}$ . Additionally, in generalized eigenvalue problems, where we employ a diagonal approximation to the matrix inverse, R-ChFSI attains a residual tolerance 10 orders of magnitude lower than standard ChFSI, demonstrating superior robustness to approximation errors. The improved convergence for generalized eigenvalue problems is particularly noteworthy, as it highlights the ability of R-ChFSI to accommodate approximate inverses while maintaining accuracy. In contrast, standard ChFSI is more sensitive to such approximations, often leading to a loss in accuracy or requiring significantly higher computational costs to achieve similar residual tolerances. This advantage of R-ChFSI is especially relevant in quantum chemistry and electronic structure calculations, where generalized eigenvalue problems frequently arise, and high computational costs constitute a significant bottleneck. By reducing dependence on exact matrix factorizations and high-precision arithmetic, R-ChFSI enables efficient exploitation of modern hardware accelerators, making large-scale eigenvalue computations more feasible in high-performance computing environments.

The remainder of this article is organized as follows: Section 2 outlines the key steps of the Chebyshev filtered subspace iteration procedure and subsequently analyzes the convergence properties of ChFSI in terms of how the maximum principal angle between the current subspace and the target eigenspace evolves during the iterations. Section 3 begins by analyzing the convergence of ChFSI when the subspace is constructed approximately due to inexact matrix-vector products. Subsequently, this section introduces the proposed residual-based Chebsyshev filtered subspace iteration method (R-ChFSI) for both standard and generalized eigenvalue problems. This section discusses the convergence of R-ChFSI with approximate matrix products and demonstrates mathematically that the R-ChFSI method can converge even when the traditional ChFSI method fails. Section 4 presents a comprehensive evaluation of the proposed R-ChFSI method in terms of accuracy and computational efficiency on both CPU and GPU architectures. It compares R-ChFSI to the traditional ChFSI approach for solving real symmetric and complex Hermitian eigenproblems in the case of both standard and generalized eigenproblems while using inexact matrix-vector products.

2. Mathematical Background. The ChFSI [32, 33, 47, 48, 29, 45] approach for solving the desired eigenpairs belongs to the category of iterative orthogonal projection methods and is one of the widely used strategies to compute the smallest n

eigenvalues and their corresponding eigenvectors of large sparse matrices [9, 7, 28, 44, 13, 43, 19, 22, 42, 26, 10, 21]. To describe the proposed eigensolver strategy based on the ChFSI approach, in the current work, we consider the Hermitian generalized eigenvalue problem of the form

(2.1) 
$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{B}\mathbf{u}_i \text{ where } \mathbf{A} \in \mathbb{C}^{m \times m}, \ \mathbf{B} \in \mathbb{C}^{m \times m}$$

where **A** and **B** are Hermitian matrices with **B** being a positive-definite matrix. In addition,  $\lambda_i \in \mathbb{R}$  and  $\mathbf{u}_i \in \mathbb{C}^m : \forall i = 1, ..., n$  denote the eigenvalue-eigenvector pairs corresponding to the smallest n eigenvalues. Equation (2.1) reduces to a standard eigenvalue problem when  $\mathbf{B} = \mathbf{I}$  with  $\mathbf{I}$  denoting the  $m \times m$  identity matrix. Without loss of generality, we assume that the eigenvalues are ordered as  $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq \lambda_{n+1} \leq \cdots \leq \lambda_m$ . For completeness and to introduce notations, we now provide a brief overview of the ChFSI algorithm, traditionally used to solve the eigenvalue problem corresponding to (2.1).

2.1. Chebyshev filtered subspace iteration and convergence properties. ChFSI leverages the properties of Chebyshev polynomials to efficiently filter out the components of the unwanted eigenvectors (corresponding to the remaining m - n largest eigenvalues), thus enriching the trial subspace with the desired eigenvectors. To this end, we define the Chebyshev polynomial of degree k as  $T_k(x)$  and note that these polynomials exhibit the fastest growth in their magnitude when |x| > 1 while remaining bounded between [-1,1] when  $|x| \leq 1$ . To exploit this fast growth property of  $T_k(x)$ , an affine transformation that maps the largest m - n eigenvalues to [-1,1] is defined, and consequently, the desired smallest n eigenvalues get mapped to values lying in  $(-\infty, -1]$ . To this end, we define the center of the unwanted spectrum as  $c = (\lambda_{n+1} + \lambda_m)/2$  and the half-width of the unwanted spectrum as  $e = (\lambda_{n+1} - \lambda_m)/2$  and hence the required affine transformation can now be represented as  $\mathcal{L}(x) = (x - 1)$ .

**Algorithm 2.1** Subspace Iteration accelerated using Chebyshev polynomial of degree p (ChFSI)

*Initial Guess*: Let  $\mathbf{X}^{(0)} = \begin{bmatrix} \mathbf{x}_1^{(0)} & \mathbf{x}_2^{(0)} & \dots & \mathbf{x}_n^{(0)} \end{bmatrix}$  be the initial guess of the eigenvectors  $(\{\mathbf{u}_j\})$ .

while  $r_j^{(i+1)} = \|\mathbf{A}\mathbf{x}_j^{(i+1)} - \epsilon_j^{(i+1)}\mathbf{B}\mathbf{x}_j^{(i+1)}\| \ge \tau \text{ do}$ 

Chebyshev Filtered Subspace Construction: Construct  $\mathbf{Y}_p^{(i)} = C_p(\mathbf{H})\mathbf{X}^{(i)}$  using the recurrence:

(2.2) 
$$\mathbf{Y}_{k+1}^{(i)} = \frac{2\sigma_{k+1}}{e} \mathbf{H} \mathbf{Y}_{k}^{(i)} - \frac{2\sigma_{k+1}c}{e} \mathbf{Y}_{k}^{(i)} - \sigma_{k}\sigma_{k+1} \mathbf{Y}_{k-1}^{(i)}$$

where  $\mathbf{Y}_{0}^{(i)} = \mathbf{X}^{(i)}$  and  $\mathbf{Y}_{1}^{(i)} = \frac{\sigma_{1}}{e} (\mathbf{H} - c\mathbf{I}) \mathbf{X}^{(i)}$  and  $\sigma_{k+1} = 1/\left(\sigma_{k} - \frac{2}{\sigma_{1}}\right)$  with  $\sigma_{1} = e/(a_{low} - c)$ . Here  $\mathbf{Y}_{k}^{(i)} = C_{k}(\mathbf{H}) \mathbf{X}^{(i)}$  for  $k = 0, 1, \dots, p$ 

Rayleigh-Ritz step: Solve the smaller  $n \times n$  dense generalized eigenvalue problem,  $\mathbf{Y}_{p}^{(i)^{\dagger}} \mathbf{A} \mathbf{Y}_{p}^{(i)} \mathbf{E} = \mathbf{Y}_{p}^{(i)^{\dagger}} \mathbf{B} \mathbf{Y}_{p}^{(i)} \mathbf{E} \mathbf{\Lambda}$ , where  $\mathbf{E}$  is the eigenvector matrix and  $\mathbf{\Lambda}$  is the diagonal matrix with the eigenvalues  $\left\{ \epsilon_{j}^{(i+1)} \right\}_{j=1}^{n}$  as its entries. The Ritz vectors are now given by  $\mathbf{X}^{(i+1)} = \mathbf{Y}_{p}^{(i)} \mathbf{E}$  and the Ritz-Values are given by  $\mathbf{\Lambda}^{(i+1)} = \mathbf{\Lambda}$ . end while c)/e. Standard implementations of ChFSI also scale the Chebyshev polynomials to prevent overflow [32, 33, 47, 48, 45] and consequently the scaled and shifted Chebyshev polynomials are defined as  $C_k(x) = T_k(\mathcal{L}(x))/T_k(\mathcal{L}(a_{low}))$  where  $a_{low} \leq \lambda_1$ . We note that the ChFSI procedure for solving the eigenproblem in (2.1) is usually devised with the matrix  $\mathbf{H} = \mathbf{B}^{-1}\mathbf{A}$  that has the same eigenpairs as  $\mathbf{A}\mathbf{u}_j = \lambda_j \mathbf{B}\mathbf{u}_j$ . This procedure for computing the smallest eigenpairs of (2.1) up to a specified tolerance  $\tau$ on the eigenproblem residual norm, is summarized in Algorithm 2.1.

**2.1.1. Convergence analysis of ChFSI.** To ensure completeness, we now provide a mathematical justification for the convergence of ChFSI. Note that  $\|\cdot\|$  implies the 2-norm for vectors and vector-induced matrix 2-norm for matrices (spectral norm) throughout this work. In order to analyze convergence, we first define the maximum principal angle between two subspaces [3, 49, 41].

DEFINITION 2.1. Let  $\mathcal{X} \subset \mathbb{C}^{m \times m}$  and  $\mathcal{Y} \subset \mathbb{C}^{m \times m}$  be subspaces of dimension n. The maximum principal angle between the two subspaces denoted by  $\angle(\mathcal{X}, \mathcal{Y})$  is defined as

$$\sin \angle (\mathcal{X}, \mathcal{Y}) = \| \mathbf{X}_{\perp}^{\dagger} \mathbf{Y} \| = \| \mathbf{Y}_{\perp}^{\dagger} \mathbf{X} \|$$

where  $\mathbf{X} \in \mathbb{C}^{m \times n}$  and  $\mathbf{Y} \in \mathbb{C}^{m \times n}$  are matrices whose columns form an orthonormal basis for  $\mathcal{X}$  and  $\mathcal{Y}$  respectively.  $\mathbf{X}_{\perp} \in \mathbb{C}^{m \times m-n}$  and  $\mathbf{Y}_{\perp} \in \mathbb{C}^{m \times m-n}$  are matrices whose columns form an orthonormal basis spanning the orthogonal complements of  $\mathcal{X}$ and  $\mathcal{Y}$  respectively.

We also state a few results useful for further analysis

LEMMA 2.2. Let the matrices  $\mathbf{E} \in \mathbb{C}^{m \times n}$  and  $\mathbf{E}_{\perp} \in \mathbb{C}^{m \times (m-n)}$  be given by  $\mathbf{E} = \begin{bmatrix} \mathbf{I}_n \\ 0 \end{bmatrix}$  and  $\mathbf{E}_{\perp} = \begin{bmatrix} 0 \\ \mathbf{I}_{m-n} \end{bmatrix}$  where  $\mathbf{I}_n$  and  $\mathbf{I}_{m-n}$  are the  $n \times n$  and  $(m-n) \times (m-n)$  identity matrices respectively. Let  $\mathcal{E} = \mathcal{R}(\mathbf{E})$  and  $\mathcal{E}_{\perp} = \mathcal{R}(\mathbf{E}_{\perp})$  where  $\mathcal{R}(.)$  denotes range of a matrix. Then for any matrix  $\mathbf{X} \in \mathbb{C}^{m \times n}$  with  $\mathcal{X} = \mathcal{R}(\mathbf{X})$  satisfying  $\mathcal{E}_{\perp} \cap \mathcal{X} = \{0\}$  or equivalently rank  $(\mathbf{E}^{\dagger}\mathbf{X}) = n$ , we have

$$\tan \angle (\mathcal{X}, \mathcal{E}) = \| \mathbf{E}_{\perp}^{\dagger} \mathbf{X} (\mathbf{E}^{\dagger} \mathbf{X})^{-1} \|$$

*Proof.* For proof of this statement, we refer the reader to [49, 41].

For the sake of convenience in the subsequent convergence analysis, we define the Hermitian matrix  $\hat{\mathbf{H}} = \mathbf{B}^{-\frac{1}{2}} \mathbf{A} \mathbf{B}^{-\frac{1}{2}}$  allowing us to rewrite the generalized eigenproblem  $\mathbf{A}\mathbf{U} = \mathbf{B}\mathbf{U}\mathbf{\Lambda}$ , as a Hermitian standard eigenvalue problem  $\hat{\mathbf{H}}\hat{\mathbf{U}} = \hat{\mathbf{U}}\mathbf{\Lambda}$ , where  $\hat{\mathbf{U}} = \mathbf{B}^{\frac{1}{2}}\mathbf{U}$  denotes the  $m \times m$  unitary matrix with columns as eigenvectors of  $\hat{\mathbf{H}}$  and  $\boldsymbol{\Lambda}$  denotes the diagonal matrix comprising eigenvalues of  $\hat{\mathbf{H}}$ . Consider the partitioning of  $\hat{\mathbf{U}}$  as  $\begin{bmatrix} \hat{\mathbf{U}}_1 & \hat{\mathbf{U}}_2 \end{bmatrix}$  where  $\hat{\mathbf{U}}_1$  is the  $m \times n$  matrix whose columns are the eigenvectors corresponding to the lowest n eigenvalues and  $\hat{\mathbf{U}}_2$  is the  $m \times (m - n)$  matrix whose columns are the rest of the eigenvectors. We define the wanted eigenspace corresponding to the n smallest eigenvectors of  $\hat{\mathbf{H}}$  as  $\mathcal{S} = \mathcal{R}(\hat{\mathbf{U}}_1)$  and the unwanted eigenspace as  $\mathcal{S}_{\perp} = \mathcal{R}(\hat{\mathbf{U}}_2)$ . The trial subspace corresponding to  $\hat{\mathbf{H}}$  at the beginning of  $i^{th}$  iteration is denoted as  $\mathcal{S}^{(i)} = \mathcal{R}(\hat{\mathbf{X}}^{(i)})$  where  $\hat{\mathbf{X}}^{(i)} = \mathbf{B}^{\frac{1}{2}}\mathbf{X}^{(i)}$ . Further, the filtered subspace obtained at the end of the  $i^{th}$  iteration is denoted as  $\mathcal{S}^{(i+1)} = \mathcal{R}(\hat{\mathbf{Y}}_p^{(i)})$  where  $\hat{\mathbf{Y}}_p^{(i)} = \mathbf{B}^{\frac{1}{2}}\mathbf{Y}_p^{(i)}$ . We note that using the relation  $\mathbf{B}^{\frac{1}{2}}C_p(\mathbf{H})\mathbf{B}^{-\frac{1}{2}} = C_p(\hat{\mathbf{H}})\mathbf{X}^{(i)}$ , i.e.,  $\mathcal{S}^{(i+1)} = C_p(\hat{\mathbf{H}})\mathbf{S}^{(i)}$ .

THEOREM 2.3. For an n-dimensional space  $S^{(i)}$  satisfying  $S^{(i)} \cap S_{\perp} = \{0\}$  where  $S_{\perp}$  is the orthogonal complement of S and  $S^{(i+1)} = C_p(\hat{\mathbf{H}})S^{(i)}$ , we have the following inequality

$$\tan \angle (\mathcal{S}^{(i+1)}, \mathcal{S}) \leq \left| \frac{C_p(\lambda_{n+1})}{C_p(\lambda_n)} \right| \, \tan \angle (\mathcal{S}^{(i)}, \mathcal{S})$$

*Proof.* Since the maximum principal angle given by Definition 2.1 is invariant under unitary transformations, we have  $\angle(\mathcal{S}^{(i+1)}, \mathcal{S}) = \angle(\hat{\mathbf{U}}^{\dagger} \mathcal{S}^{(i+1)}, \hat{\mathbf{U}}^{\dagger} \mathcal{S})$  and also  $\angle(\mathcal{S}^{(i)}, \mathcal{S}) = \angle(\hat{\mathbf{U}}^{\dagger} \mathcal{S}^{(i)}, \hat{\mathbf{U}}^{\dagger} \mathcal{S})$ . Since the space defined by  $\hat{\mathbf{U}}^{\dagger} \mathcal{S}$  is the same as  $\mathcal{E} = \mathcal{R}(\mathbf{E})$ , we have,  $\angle(\mathcal{S}^{(i+1)}, \mathcal{S}) = \angle(\hat{\mathbf{U}}^{\dagger} \mathcal{S}^{(i+1)}, \mathcal{E})$  and  $\angle(\mathcal{S}^{(i)}, \mathcal{S}) = \angle(\hat{\mathbf{U}}^{\dagger} \mathcal{S}^{(i)}, \mathcal{E})$ . Consider a partitioning of the matrix  $\hat{\mathbf{U}}^{\dagger} \hat{\mathbf{X}}^{(i)}$  as  $\hat{\mathbf{U}}^{\dagger} \hat{\mathbf{X}}^{(i)} = \begin{bmatrix} \hat{\mathbf{Z}}_{1}^{(i)} \\ \hat{\mathbf{Z}}_{2}^{(i)} \end{bmatrix}$  where  $\hat{\mathbf{Z}}_{1}^{(i)} = \hat{\mathbf{U}}_{1}^{\dagger} \hat{\mathbf{X}}^{(i)}$ and  $\hat{\mathbf{Z}}_{2}^{(i)} = \hat{\mathbf{U}}_{2}^{\dagger} \hat{\mathbf{X}}^{(i)}$ . We note that the assumption  $\mathcal{S}^{(i)} \cap \mathcal{S}_{\perp} = \{0\}$  ensures that  $\hat{\mathbf{Z}}_{1}^{(i)}$ is invertible, and consequently, we can write

$$\hat{\mathbf{U}}^{\dagger} \mathcal{S}^{(i)} = \mathcal{R}(\hat{\mathbf{U}}^{\dagger} \hat{\mathbf{X}}^{(i)}) = \mathcal{R}\left( \begin{bmatrix} \hat{\mathbf{Z}}_{1}^{(i)} \\ \hat{\mathbf{Z}}_{2}^{(i)} \end{bmatrix} \right)$$

Subsequently from Lemma 2.2 we have  $\tan \angle (\mathcal{S}^{(i)}, \mathcal{S}) = \|\hat{\mathbf{Z}}_{2}^{(i)} \hat{\mathbf{Z}}_{1}^{(i)-1}\|$ . Further, partitioning the diagonal matrix of eigenvalues  $\mathbf{\Lambda}$ , we have  $\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{1} & 0 \\ 0 & \mathbf{\Lambda}_{2} \end{bmatrix}$  where  $\mathbf{\Lambda}_{1}$  is the  $n \times n$  diagonal matrix comprising of the wanted eigenvalues and  $\mathbf{\Lambda}_{2}$  is the  $(m-n) \times (m-n)$  diagonal matrix comprising of the unwanted eigenvalues. We can now write

$$\begin{aligned} \hat{\mathbf{U}}^{\dagger} \mathcal{S}^{(i+1)} &= \hat{\mathbf{U}}^{\dagger} C_{p}(\hat{\mathbf{H}}) \mathcal{S}^{(i)} = C_{p}(\mathbf{\Lambda}) \hat{\mathbf{U}}^{\dagger} \mathcal{S}^{(i)} \\ &= \mathcal{R} \left( \begin{bmatrix} C_{p}(\mathbf{\Lambda}_{1}) & 0\\ 0 & C_{p}(\mathbf{\Lambda}_{2}) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{Z}}_{1}^{(i)}\\ \hat{\mathbf{Z}}_{2}^{(i)} \end{bmatrix} \right) = \mathcal{R} \left( \begin{bmatrix} C_{p}(\mathbf{\Lambda}_{1}) \hat{\mathbf{Z}}_{1}^{(i)}\\ C_{p}(\mathbf{\Lambda}_{2}) \hat{\mathbf{Z}}_{2}^{(i)} \end{bmatrix} \right) \end{aligned}$$

From Lemma 2.2 we now have  $\tan \angle (\mathcal{S}^{(i+1)}, \mathcal{S}) = \|C_p(\Lambda_2)\hat{\mathbf{Z}}_2^{(i)}\hat{\mathbf{Z}}_1^{(i)^{-1}}C_p(\Lambda_1)^{-1}\|$ , allowing us to write

$$\tan \angle (\mathcal{S}^{(i+1)}, \mathcal{S}) \le \|C_p(\mathbf{\Lambda}_2)\| \| \hat{\mathbf{Z}}_2^{(i)} \hat{\mathbf{Z}}_1^{(i)^{-1}} \| \| C_p(\mathbf{\Lambda}_1)^{-1} \| = \left| \frac{C_p(\lambda_{n+1})}{C_p(\lambda_n)} \right| \tan \angle (\mathcal{S}^{(i)}, \mathcal{S}) \square$$

3. Approximate Chebyshev Filtered Subspace Construction. We note that the computationally dominant step in Chebyshev filtered subspace construction is the evaluation of the sparse-matrix multi-vector product  $\mathbf{HY}_{k}^{(i)}$  in (2.2). A straightforward way to accelerate this step is to use approximations in the computation of  $\mathbf{HY}_{k}^{(i)}$ , allowing for improved efficiency. This should, in principle, allow for the use of various efficient approximate matrix multiplication techniques, including but not limited to mixed-precision arithmetic. However, we note that this requires an understanding of the convergence properties of ChFSI when such approximations are employed. We now adapt Theorem 2.3 for the case where approximations are used in computing  $\mathbf{HY}_{k}^{(i)}$ .

**3.1.** Convergence of ChFSI with inexact subspace construction. We now define  $\underline{S}^{(i+1)}$  as the space spanned by the columns of  $\hat{\underline{Y}}_{p}^{(i)} = \underline{B}_{p}^{1} \underline{Y}_{p}^{(i)}$  with  $\underline{Y}_{p}^{(i)}$  defined as  $\underline{Y}_{p}^{(i)} = \underline{C}_{p}(\underline{H}) \underline{X}^{(i)}$  where underline here denotes that approximations are introduced during matrix multiplications in Step 2 of Algorithm 2.1. We note that the columns of  $\hat{\underline{X}}^{(i)} = \underline{B}_{p}^{1} \underline{X}^{(i)}$  form an orthonormal basis for  $S^{(i)}$ .

THEOREM 3.1. For an n-dimensional space  $S^{(i)}$  satisfying  $S^{(i)} \cap S_{\perp} = \{0\}$  and  $\underline{S}^{(i+1)} = \mathcal{R}\left(\underline{C_p(\mathbf{H})\mathbf{X}^{(i)}}\right)$  we can write

(3.1) 
$$\tan \angle (\underline{\mathcal{S}}^{(i+1)}, \mathcal{S}) \leq \left( \frac{|C_p(\lambda_{n+1})| + \|\hat{\mathbf{\Delta}}_p^{(i)}\| \csc \angle (\mathcal{S}^{(i)}, \mathcal{S})}{|C_p(\lambda_n)| - \|\hat{\mathbf{\Delta}}_p^{(i)}\| \sec \angle (\mathcal{S}^{(i)}, \mathcal{S})} \right) \tan \angle (\mathcal{S}^{(i)}, \mathcal{S})$$

where  $\hat{\boldsymbol{\Delta}}_{p}^{(i)} = \underline{C_{p}(\hat{\mathbf{H}})\hat{\mathbf{X}}^{(i)}} - C_{p}(\hat{\mathbf{H}})\hat{\mathbf{X}}^{(i)}$ 

*Proof.* Partitioning the matrix  $\hat{\mathbf{U}}^{\dagger} \hat{\mathbf{X}}^{(i)}$ , we have  $\hat{\mathbf{U}}^{\dagger} \hat{\mathbf{X}}^{(i)} = \begin{bmatrix} \hat{\mathbf{Z}}_{1}^{(i)} \\ \hat{\mathbf{Z}}_{2}^{(i)} \end{bmatrix}$  where  $\hat{\mathbf{Z}}_{1}^{(i)} = \hat{\mathbf{U}}_{2}^{\dagger} \hat{\mathbf{X}}^{(i)}$  and  $\hat{\mathbf{Z}}_{2}^{(i)} = \hat{\mathbf{U}}_{2}^{\dagger} \hat{\mathbf{X}}^{(i)}$ . We note that the assumption  $\mathcal{S}^{(i)} \cap \mathcal{S}_{\perp} = \{0\}$  ensures that  $\hat{\mathbf{Z}}_{1}^{(i)}$  is invertible. Consequently, we can write

$$\begin{split} \hat{\mathbf{U}}^{\dagger} \underline{\boldsymbol{\mathcal{S}}}^{(i+1)} &= \mathcal{R}\left(\hat{\mathbf{U}}^{\dagger} \underline{C_{p}}(\hat{\mathbf{H}}) \hat{\mathbf{X}}^{(i)}\right) = \mathcal{R}\left(\hat{\mathbf{U}}^{\dagger} C_{p}(\hat{\mathbf{H}}) \hat{\mathbf{X}}^{(i)} + \hat{\mathbf{U}}^{\dagger} \hat{\boldsymbol{\Delta}}_{p}^{(i)}\right) \\ &= \mathcal{R}\left(C_{p}(\boldsymbol{\Lambda}) \hat{\mathbf{U}}^{\dagger} \hat{\mathbf{X}}^{(i)} + \hat{\mathbf{U}}^{\dagger} \hat{\boldsymbol{\Delta}}_{p}^{(i)}\right) = \mathcal{R}\left(\begin{bmatrix}C_{p}(\boldsymbol{\Lambda}_{1}) \hat{\mathbf{Z}}_{1}^{(i)} + \hat{\mathbf{U}}_{1}^{\dagger} \hat{\boldsymbol{\Delta}}_{p}^{(i)}\\C_{p}(\boldsymbol{\Lambda}_{2}) \hat{\mathbf{Z}}_{2}^{(i)} + \hat{\mathbf{U}}_{2}^{\dagger} \hat{\boldsymbol{\Delta}}_{p}^{(i)}\end{bmatrix}\right) \end{split}$$

Consequently, from Lemma 2.2 we can write

$$\tan \angle (\underline{S}^{(i+1)}, S) = \left\| \left( C_p(\mathbf{\Lambda}_2) \hat{\mathbf{Z}}_2^{(i)} + \hat{\mathbf{U}}_2^{\dagger} \hat{\mathbf{\Delta}}_p^{(i)} \right) \left( C_p(\mathbf{\Lambda}_1) \hat{\mathbf{Z}}_1^{(i)} + \hat{\mathbf{U}}_1^{\dagger} \hat{\mathbf{\Delta}}_p^{(i)} \right)^{-1} \right\|$$

$$(3.2) \qquad \leq \frac{\| C_p(\mathbf{\Lambda}_2) \hat{\mathbf{Z}}_2^{(i)} \hat{\mathbf{Z}}_1^{(i)^{-1}} C_p(\mathbf{\Lambda}_1)^{-1} \| + \| \hat{\mathbf{\Delta}}_p^{(i)} \hat{\mathbf{Z}}_1^{(i)^{-1}} C_p(\mathbf{\Lambda}_1)^{-1} \|}{1 - \| \hat{\mathbf{\Delta}}_p^{(i)} \hat{\mathbf{Z}}_1^{(i)^{-1}} C_p(\mathbf{\Lambda}_1)^{-1} \|}$$

Note that in the last step, we have assumed that  $\|\hat{\boldsymbol{\Delta}}_{p}^{(i)}\hat{\mathbf{Z}}_{1}^{(i)-1}C_{p}(\boldsymbol{\Lambda}_{1})^{-1}\| < 1$ , and using the fact that  $\|\hat{\mathbf{Z}}_{1}^{(i)-1}\| = \sec \angle (\mathcal{S}^{(i)}, \mathcal{S})$ , a sufficient condition for this to be true is  $|C_{p}(\lambda_{n})| \cos \angle (\mathcal{S}^{(i)}, \mathcal{S}) > \|\hat{\boldsymbol{\Delta}}_{p}^{(i)}\|$ . Upon further simplification using submultiplicative property of matrix spectral norms and triangle inequality, the inequality in (3.2) can be written as

$$\tan \angle (\underline{\mathcal{S}}^{(i+1)}, \mathcal{S}) \leq \frac{|C_p(\lambda_{n+1})| \sin \angle (\mathcal{S}^{(i)}, \mathcal{S}) + \|\hat{\mathbf{\Delta}}_p^{(i)}\|}{|C_p(\lambda_n)| \cos \angle (\mathcal{S}^{(i)}, \mathcal{S}) - \|\hat{\mathbf{\Delta}}_p^{(i)}\|} \\ = \left(\frac{|C_p(\lambda_{n+1})| + \|\hat{\mathbf{\Delta}}_p^{(i)}\| \csc \angle (\mathcal{S}^{(i)}, \mathcal{S})}{|C_p(\lambda_n)| - \|\hat{\mathbf{\Delta}}_p^{(i)}\| \sec \angle (\mathcal{S}^{(i)}, \mathcal{S})}\right) \tan \angle (\mathcal{S}^{(i)}, \mathcal{S})$$

which proves the desired inequality in the theorem.

Now for convergence we demand that  $\angle(\underline{S}^{(i+1)}, S) < \angle(S^{(i)}, S)$  and consequently we require

(3.3) 
$$|C_p(\lambda_n)| - |C_p(\lambda_{n+1})| > \|\hat{\boldsymbol{\Delta}}_p^{(i)}\| \left( \sec \angle (\mathcal{S}^{(i)}, \mathcal{S}) + \csc \angle (\mathcal{S}^{(i)}, \mathcal{S}) \right) \quad \forall i = 0, \dots, \infty$$

We note that if  $\|\hat{\Delta}_{p}^{(i)}\|$  remains nearly constant with iteration *i*, the right-hand side of the inequality in (3.3) keeps increasing as we approach the exact eigenspace and beyond a certain angle  $\angle(\mathcal{S}^{(i)}, \mathcal{S})$  this inequality gets violated. Consequently, the angle stops decreasing and we cannot approach the exact eigenspace beyond that point. To this end, for robust convergence, we require that  $\|\hat{\Delta}_{p}^{(i)}\|$  also decreases as we approach the exact eigenspace, and we demonstrate that our proposed residual-based reformulation of Chebyshev filtered subspace iteration procedure (R-ChFSI) algorithm described subsequently accomplishes this. We further note that  $(\sec \angle(\mathcal{S}^{(i)}, \mathcal{S}) + \csc \angle(\mathcal{S}^{(i)}, \mathcal{S})) \ge 2\sqrt{2}$  and consequently for convergence with approximations, we obtain the following necessary condition on  $\|\hat{\Delta}_{p}^{(i)}\|$  from (3.3),

(3.4) 
$$\|\hat{\boldsymbol{\Delta}}_{p}^{(i)}\| < \frac{|C_{p}(\lambda_{n})| - |C_{p}(\lambda_{n+1})|}{2\sqrt{2}}$$

3.2. Standard Eigenvalue Problems. In this section we analyze the case of  $\mathbf{B} = \mathbf{I}$  and consequently  $\hat{\mathbf{H}} = \mathbf{B}^{-\frac{1}{2}} \mathbf{A} \mathbf{B}^{-\frac{1}{2}} = \mathbf{H} = \mathbf{A}$  is a Hermitian matrix. We now consider the specific case of utilizing a lower precision to compute the matrix product  $\mathbf{H} \mathbf{Y}_{k}^{(i)}$  in the Chebyshev recurrence relation defined by (2.2). We first evaluate the upper bound on  $\boldsymbol{\Delta}_{p}^{(i)}$  if one naively replaces the matrix product  $\mathbf{H} \mathbf{Y}_{k}^{(i)}$  in (2.2) with the approximate matrix product denoted as  $\mathbf{H} \mathbf{Y}_{k}^{(i)} = \mathbf{H} \otimes \mathbf{Y}_{k}^{(i)}$ , where  $\otimes$  represents the product evaluated with lower precision arithmetic, and argue that this method will fail to converge to the same residual tolerance that can be achieved with full precision matrix products. We then propose a residual-based reformulation of the recurrence relation and argue that the proposed reformulation allows for convergence to similar residual tolerances that can be achieved with full precision matrix products.

3.2.1. Traditional Chebyshev filtering method employing low-precision matrix-products. Employing low-precision matrix-products, the recurrence relation in (2.2) for  $\underline{\mathbf{Y}}_{k}^{(i)} = C_{k}(\mathbf{H}) \otimes \mathbf{X}^{(i)}$ , where  $k = 2, \ldots, p$  can be written as

(3.5) 
$$\underline{\mathbf{Y}}_{k+1}^{(i)} = a_k \mathbf{H} \otimes \underline{\mathbf{Y}}_k^{(i)} + b_k \underline{\mathbf{Y}}_k^{(i)} + c_k \underline{\mathbf{Y}}_{k-1}^{(i)}$$

where for convenience of notation we have defined

$$a_k = \frac{2\sigma_{k+1}}{e}$$
  $b_k = -\frac{2\sigma_{k+1}c}{e}$   $c_k = -\sigma_k\sigma_{k+1}$ 

and with the initial conditions  $\underline{\mathbf{Y}}_{0}^{(i)} = \mathbf{X}^{(i)}$  and  $\underline{\mathbf{Y}}_{1}^{(i)} = \frac{\sigma_{1}}{e} (\mathbf{H} - c\mathbf{I}) \mathbf{X}^{(i)}$ . We now state some useful results required for further analysis

LEMMA 3.2. If  $\mathbf{H} \in \mathbb{C}^{m \times m}$  and  $\mathbf{X} \in \mathbb{C}^{m \times n}$ , the spectral norm of the error in evaluating the matrix product  $\mathbf{H}\mathbf{X}$  due to floating point approximations satisfies  $\|\mathbf{H}\mathbf{X} - \mathbf{H} \otimes \mathbf{X}\| \leq \gamma_m \|\mathbf{H}\| \|\mathbf{X}\|$ 

*Proof.* This is a straightforward application of the results derived in [14, section 3.5]. We note that  $\gamma_m$  is a constant that depends on m, and the machine precision

of the floating point arithmetic (denoted as  $\varepsilon_M$ ). For the case of dense real matrices, we have  $\gamma_m = m\varepsilon_M/(1 - m\varepsilon_M)$  and for dense complex matrices, it needs to be appropriately modified as described in [14, section 3.6]. Further, we also note that in the case where **H** is sparse, significantly tighter bounds independent of m can be achieved depending on the specific implementations of the sparse matrix-dense matrix multiplication routines.

LEMMA 3.3. Consider a recurrence relation of the form

(3.6) 
$$\boldsymbol{\Delta}_{k+1}^{(i)} = a_k \mathbf{H} \boldsymbol{\Delta}_k^{(i)} + b_k \boldsymbol{\Delta}_k^{(i)} + c_k \boldsymbol{\Delta}_{k-1}^{(i)} + a_k \mathbf{e}_k^{(i)}$$

with  $\|\mathbf{e}_{k}^{(i)}\| \leq h_{0} \|\mathbf{\Delta}_{k}^{(i)}\| + h_{1}^{(i)} \ \forall k = 1, 2, \dots, p.$  We can then write

(3.7) 
$$\|\boldsymbol{\Delta}_{k+1}^{(i)}\| \le g_k^0 \left(\|\boldsymbol{\Delta}_0^{(i)}\| + \|\boldsymbol{\Delta}_1^{(i)}\|\right) + g_k^1 h_1^{(i)}$$

*Proof.* From (3.6) we can write

$$\begin{aligned} |\mathbf{\Delta}_{k+1}^{(i)}|| &\leq |a_k| \|\mathbf{H}\| \|\mathbf{\Delta}_k^{(i)}\| + |b_k| \|\mathbf{\Delta}_k^{(i)}\| + |c_k| \|\mathbf{\Delta}_{k-1}^{(i)}\| + |a_k| \|\mathbf{e}_k^{(i)}\| \\ &\leq (|a_k| \|\mathbf{H}\| + |b_k| + h_0) \|\mathbf{\Delta}_k^{(i)}\| + |c_k| \|\mathbf{\Delta}_{k-1}^{(i)}\| + |a_k| h_1^{(i)} \end{aligned}$$

We define the following

$$\mathbf{d}_{k}^{(i)} = \begin{bmatrix} \|\mathbf{\Delta}_{k}^{(i)}\| \\ \|\mathbf{\Delta}_{k-1}^{(i)}\| \end{bmatrix} \quad \mathbf{F}_{k} = \begin{bmatrix} |a_{k}| \|\mathbf{H}\| + |b_{k}| + |a_{k}| h_{0} & |c_{k}| \\ 1 & 0 \end{bmatrix} \quad \mathbf{E}_{k}^{(i)} = \begin{bmatrix} |a_{k}| h_{1}^{(i)} \\ 0 \end{bmatrix}$$

we can now write

$$\|\mathbf{d}_{k+1}^{(i)}\| \le \|\mathbf{F}_k \mathbf{d}_k^{(i)} + \mathbf{E}_k^{(i)}\| \le \|\mathbf{F}_k\| \|\mathbf{d}_k^{(i)}\| + |a_k| h_1^{(i)}$$

Consequently, we now have

$$\|\mathbf{d}_{k+1}^{(i)}\| \le \left(\prod_{j=1}^{k} \|\mathbf{F}_{j}\|\right) \|\mathbf{d}_{1}^{(i)}\| + \sum_{j=1}^{k} \left(\prod_{r=j+1}^{k} \|\mathbf{F}_{r}\|\right) |a_{j}| h_{1}^{(i)}$$

We note that  $\|\mathbf{\Delta}_{k+1}^{(i)}\| \le \|\mathbf{d}_{k+1}^{(i)}\|$  and  $\|\mathbf{d}_{1}^{(i)}\| \le \|\mathbf{\Delta}_{0}^{(i)}\| + \|\mathbf{\Delta}_{1}^{(i)}\|$ . This results in

$$\|\boldsymbol{\Delta}_{k+1}^{(i)}\| \le \left(\prod_{j=1}^{k} \|\mathbf{F}_{j}\|\right) \left(\|\boldsymbol{\Delta}_{0}^{(i)}\| + \|\boldsymbol{\Delta}_{1}^{(i)}\|\right) + \sum_{j=1}^{k} \left(\prod_{r=j+1}^{k} \|\mathbf{F}_{r}\|\right) |a_{j}| h_{1}^{(i)}$$

Defining  $g_k^0 = \prod_{j=1}^k \|\mathbf{F}_j\|$  and  $g_k^1 = \sum_{j=1}^k \prod_{r=j+1}^k \|\mathbf{F}_r\| |a_j|$  we have

$$\|\boldsymbol{\Delta}_{k+1}^{(i)}\| \le g_k^0 \left( \|\boldsymbol{\Delta}_0^{(i)}\| + \|\boldsymbol{\Delta}_1^{(i)}\| \right) + g_k^1 h_1^{(i)} \qquad \Box$$

THEOREM 3.4. The spectral norm of the error  $\mathbf{\Delta}_{k}^{(i)} = \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{Y}_{k}^{(i)}$  in the subspace construction using lower precision arithmetic in matrix-products of the recurrence relation (2.2) satisfies  $\|\mathbf{\Delta}_{k}^{(i)}\| \leq \gamma_{m}\eta_{k}$  for  $k = 0, 1, \ldots, p$  where  $\eta_{k}$  is some finite constant that depends on k

*Proof.* Using (2.2) and (3.5) we can now write a recurrence relation for  $\Delta_k^{(i)} = \underline{\mathbf{Y}}_k^{(i)} - \mathbf{Y}_k^{(i)}$  as

(3.8) 
$$\boldsymbol{\Delta}_{k+1}^{(i)} = a_k \mathbf{H} \boldsymbol{\Delta}_k^{(i)} + b_k \boldsymbol{\Delta}_k^{(i)} + c_k \boldsymbol{\Delta}_{k-1}^{(i)} + a_k \left( \mathbf{H} \otimes \underline{\mathbf{Y}}_k^{(i)} - \mathbf{H} \underline{\mathbf{Y}}_k^{(i)} \right)$$

with the initial conditions  $\mathbf{\Delta}_{0}^{(i)} = \mathbf{\Delta}_{1}^{(i)} = 0$ . We note that this recurrence relation is of the same form as that of (3.6) in Lemma 3.3 with  $\mathbf{e}_{k}^{(i)} = \mathbf{H} \otimes \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{H}\underline{\mathbf{Y}}_{k}^{(i)}$ . From Lemma 3.2 we have  $\|\mathbf{e}_{k}^{(i)}\| = \|\mathbf{H} \otimes \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{H}\underline{\mathbf{Y}}_{k}^{(i)}\| \leq \gamma_{m}\|\mathbf{H}\| \|\underline{\mathbf{Y}}_{k}^{(i)}\|$ . We further note that

$$\begin{aligned} \|\underline{\mathbf{Y}}_{k}^{(i)}\| &= \|\mathbf{\Delta}_{k}^{(i)} + \mathbf{Y}_{k}^{(i)}\| \leq \|\mathbf{\Delta}_{k}^{(i)}\| + \|\mathbf{Y}_{k}^{(i)}\| \leq \|\mathbf{\Delta}_{k}^{(i)}\| + \|C_{k}(\mathbf{H})\mathbf{X}^{(i)}\| \\ &\leq \|\mathbf{\Delta}_{k}^{(i)}\| + \|C_{k}(\mathbf{H})\| \end{aligned}$$

Thus we have  $\|\mathbf{e}_{k}^{(i)}\| \leq \gamma_{m} \|\mathbf{H}\| \|\mathbf{\Delta}_{k}^{(i)}\| + \gamma_{m} \|\mathbf{H}\| \|C_{k}(\mathbf{H})\|$ . Using Lemma 3.3 with  $\|\mathbf{\Delta}_{0}^{(i)}\| = \|\mathbf{\Delta}_{1}^{(i)}\| = 0, h_{0} = \gamma_{m} \|\mathbf{H}\|$  and  $h_{1}^{(i)} = \gamma_{m} \|\mathbf{H}\| \|C_{k}(\mathbf{H})\|$  we obtain  $\|\mathbf{\Delta}_{k}^{(i)}\| \leq \gamma_{m} \eta_{k}$  with  $\eta_{k} = \|\mathbf{H}\| \|C_{k}(\mathbf{H})\| g_{k-1}^{1}$ .

We now discuss the implications of the Theorem 3.4 within the context of (3.3). Without loss of generality, we assume that our initial guess of trial subspace  $S^{(i)}$  at i = 0 satisfies the inequality below as stated in (3.3).

(3.9) 
$$|C_p(\lambda_n)| - |C_p(\lambda_{n+1})| > \|\hat{\mathbf{\Delta}}_p^{(i)}\| \left(\sec \angle (\mathcal{S}^{(i)}, \mathcal{S}) + \csc \angle (\mathcal{S}^{(i)}, \mathcal{S})\right)$$

While the above inequality is satisfied,  $\angle(\mathcal{S}^{(i)}, \mathcal{S})$  reduces as *i* increases (from Theorem 3.1). However, we note that the RHS of this equation does not monotonically decrease with a decrease in  $\angle(\mathcal{S}^{(i)}, \mathcal{S})$  and increases as  $\angle(\mathcal{S}^{(i)}, \mathcal{S})$  approaches 0. Thus, it stands to reason that beyond a certain value of  $\angle(\mathcal{S}^{(i)}, \mathcal{S})$ , the inequality in (3.9) no longer holds, and we cannot approach the wanted eigenspace beyond this point. Under the condition that  $\gamma_m \eta_p << |C_p(\lambda_n)| - |C_p(\lambda_{n+1})|$ , we can estimate the closest angle that can be achieved and is given by the following expression

(3.10) 
$$\angle(\mathcal{S}^{(i)}, \mathcal{S}) \approx \frac{\gamma_m \eta_p}{|C_p(\lambda_n)| - |C_p(\lambda_{n+1})|}$$

We now propose a residual-based reformulation of ChFSI and argue that the proposed method does not suffer from this stagnation behavior even when employing lower precision arithmetic in matrix products.

**3.2.2. Proposed residual-based Chebyshev filtering approach.** We now propose a residual-based reformulation of the recurrence relation defined in (2.2). To this end we define the residual  $\mathbf{R}_k^i$  in a given iteration *i* for  $k = 0, \ldots, p$  in the following way:

(3.11) 
$$\mathbf{R}_{k}^{(i)} = C_{k}(\mathbf{H})\mathbf{X}^{(i)} - \mathbf{X}^{(i)}C_{k}(\mathbf{\Lambda}^{(i)}) = \mathbf{Y}_{k}^{(i)} - \mathbf{X}^{(i)}\mathbf{\Lambda}_{k}^{(i)}$$
 for  $k = 0, \dots, p$ 

where we have defined  $\mathbf{\Lambda}_{k}^{(i)} = C_{k}(\mathbf{\Lambda}^{(i)})$  and recall p is the maximum Chebyshev polynomial degree used in the subspace construction step.

PROPOSITION 3.5. The recurrence relation given by (2.2) can be reformulated in terms of the residuals defined by  $\mathbf{R}_k^{(i)} = \mathbf{Y}_k^{(i)} - \mathbf{X}^{(i)} \mathbf{\Lambda}_k^{(i)}$  as

(3.12) 
$$\mathbf{R}_{k+1}^{(i)} = a_k \mathbf{H} \mathbf{R}_k^{(i)} + b_k \mathbf{R}_k^{(i)} + c_k \mathbf{R}_{k-1}^{(i)} + a_k \mathbf{R}^{(i)} \mathbf{\Lambda}_k^{(i)}$$

(3.13) 
$$\mathbf{\Lambda}_{k+1}^{(i)} = a_k \mathbf{\Lambda}_k^{(i)} \mathbf{\Lambda}^{(i)} + b_k \mathbf{\Lambda}_k^{(i)} + c_k \mathbf{\Lambda}_{k-1}^{(i)}$$

where  $\mathbf{R}_{0}^{(i)} = 0$  and  $\mathbf{R}_{1}^{(i)} = \frac{\sigma_{1}}{e} \mathbf{R}^{(i)}$  with  $\mathbf{R}^{(i)} = \mathbf{H}\mathbf{X}^{(i)} - \mathbf{X}^{(i)}\mathbf{\Lambda}^{(i)}$  and further we have  $a_{k} = 2\sigma_{k+1}/e$ ,  $b_{k} = -2\sigma_{k+1}c/e$ , and  $c_{k} = -\sigma_{k}\sigma_{k+1}$ 

*Proof.* The recurrence relation for  $\Lambda_k^{(i)} = C_k(\Lambda^{(i)})$  can be written as

$$\boldsymbol{\Lambda}_{k+1}^{(i)} = a_k \boldsymbol{\Lambda}_k^{(i)} \boldsymbol{\Lambda}^{(i)} + b_k \boldsymbol{\Lambda}_k^{(i)} + c_k \boldsymbol{\Lambda}_{k-1}^{(i)}$$

Multiplying with  $\mathbf{X}^{(i)}$  and subtracting from (2.2), we have

$$\mathbf{R}_{k+1}^{(i)} = a_k \mathbf{H} \mathbf{R}_k^{(i)} + b_k \mathbf{R}_k^{(i)} + c_k \mathbf{R}_{k-1}^{(i)} + a_k \mathbf{R}^{(i)} \mathbf{\Lambda}_k^{(i)} \qquad \Box$$

After computing  $\mathbf{R}_{p}^{(i)}$  using this recurrence relation we can now recover the desired subspace  $\mathbf{Y}_{p}^{(i)}$  using the relation in (3.11) i.e.,  $\mathbf{Y}_{p}^{(i)} = \mathbf{R}_{p}^{(i)} + \mathbf{X}_{p}^{(i)} \mathbf{\Lambda}_{p}^{(i)}$ . We note that using approximate matrix multiplication for evaluating  $\mathbf{HR}_{k}^{(i)}$  we can write the following recurrence relation

(3.14) 
$$\underline{\mathbf{R}}_{k+1}^{(i)} = a_k \mathbf{H} \otimes \underline{\mathbf{R}}_k^{(i)} + b_k \underline{\mathbf{R}}_k^{(i)} + c_k \underline{\mathbf{R}}_{k-1}^{(i)} + a_k \mathbf{R}^{(i)} \mathbf{\Lambda}_k^{(i)}$$

and consequently, we have  $\underline{\mathbf{Y}}_{p}^{(i)} = \underline{\mathbf{R}}_{p}^{(i)} + \mathbf{X}^{(i)} \mathbf{\Lambda}_{p}^{(i)}$ . We now prove a result useful for further analysis

LEMMA 3.6. The spectral norm of the residual,  $\mathbf{R}_{k}^{(i)} = C_{k}(\mathbf{H})\mathbf{X}^{(i)} - \mathbf{X}^{(i)}C_{k}(\mathbf{\Lambda}^{(i)})$ for k = 0, ..., p, is bounded by  $\|\mathbf{R}_{k}^{(i)}\| \leq f_{k} \|\mathbf{R}^{(i)}\|$  where  $f_{k}$  is a finite constant and  $\mathbf{R}^{(i)} = \mathbf{H}\mathbf{X}^{(i)} - \mathbf{X}^{(i)}\mathbf{\Lambda}^{(i)}$ 

*Proof.* Let  $C_k(x) = \sum_{j=0}^k \alpha_j x^j$ , we can now write

$$\mathbf{R}_{k}^{(i)} = \sum_{j=1}^{k} \alpha_{j} \left( \mathbf{H}^{j} \mathbf{X}^{(i)} - \mathbf{X}^{(i)} \mathbf{\Lambda}^{(i)}^{j} \right) = \sum_{j=1}^{k} \alpha_{j} \sum_{r=0}^{j-1} \mathbf{H}^{j-r-1} \left( \mathbf{H} \mathbf{X}^{(i)} - \mathbf{X}^{(i)} \mathbf{\Lambda}^{(i)} \right) \mathbf{\Lambda}^{(i)^{r}}$$

Using submultiplicativity of the spectral norm we can now write

$$\|\mathbf{R}_{k}^{(i)}\| \leq \sum_{j=0}^{k} \sum_{r=0}^{j-1} \|\alpha_{j}\mathbf{H}^{j-r-1}\| \|\mathbf{H}\mathbf{X}^{(i)} - \mathbf{X}^{(i)}\mathbf{\Lambda}^{(i)}\| \|\mathbf{\Lambda}^{(i)^{r}}\| = f_{k}\|\mathbf{R}^{(i)}\|$$

where we have defined  $f_k = \sum_{j=0}^k \sum_{r=0}^{j-1} \|\alpha_j \mathbf{H}^{j-r-1}\| \| \mathbf{\Lambda}^{(i)^r} \|$ .

LEMMA 3.7. The spectral norm of the error  $\boldsymbol{\Delta}_{k}^{(i)} = \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{Y}_{k}^{(i)}$  in the subspace construction using lower precision arithmetic in matrix-products of the recurrence relation (3.14) satisfies  $\|\boldsymbol{\Delta}_{k}^{(i)}\| \leq \gamma_{m} \tilde{\eta}_{k} \|\boldsymbol{R}^{(i)}\|$  for  $k = 0, 1, \ldots, p$  where  $\tilde{\eta}_{k}$  is some finite constant that depends on k.

*Proof.* Using (3.12) and (3.14) we can now write a recurrence relation for  $\Delta_k^{(i)} = \underline{\mathbf{Y}}_k^{(i)} - \mathbf{Y}_k^{(i)}$  as

(3.15) 
$$\boldsymbol{\Delta}_{k+1}^{(i)} = a_k \mathbf{H} \boldsymbol{\Delta}_k^{(i)} + b_k \boldsymbol{\Delta}_k^{(i)} + c_k \boldsymbol{\Delta}_{k-1}^{(i)} + a_k \left( \mathbf{H} \otimes \underline{\mathbf{R}}_k^{(i)} - \mathbf{H} \underline{\mathbf{R}}_k^{(i)} \right)$$

with the initial conditions  $\mathbf{\Delta}_{0}^{(i)} = \mathbf{\Delta}_{1}^{(i)} = 0$ . We note that this recurrence relation is of the same form as that of (3.6) in Lemma 3.3 with  $\mathbf{e}_k^{(i)} = \mathbf{H} \otimes \underline{\mathbf{R}}_k^{(i)} - \mathbf{H}\underline{\mathbf{R}}_k^{(i)}$ . From

Lemma 3.2 we have  $\|\mathbf{e}_{k}^{(i)}\| = \|\mathbf{H} \otimes \underline{\mathbf{R}}_{k}^{(i)} - \mathbf{H}\underline{\mathbf{R}}_{k}^{(i)}\| \leq \gamma_{m} \|\mathbf{H}\| \|\underline{\mathbf{R}}_{k}^{(i)}\|$ . We further note that

$$\|\underline{\mathbf{R}}_{k}^{(i)}\| = \|\mathbf{\Delta}_{k}^{(i)} + \mathbf{R}_{k}^{(i)}\| \le \|\mathbf{\Delta}_{k}^{(i)}\| + \|\mathbf{R}_{k}^{(i)}\| \le \|\mathbf{\Delta}_{k}^{(i)}\| + f_{k}\|\mathbf{R}^{(i)}\|$$

Thus we have  $\|\mathbf{e}_{k}^{(i)}\| \leq \gamma_{m} \|\mathbf{H}\| \|\mathbf{\Delta}_{k}^{(i)}\| + \gamma_{m} f_{k} \|\mathbf{H}\| \|\mathbf{R}^{(i)}\|$ . Using Lemma 3.3 with  $\|\mathbf{\Delta}_{0}^{(i)}\| = \|\mathbf{\Delta}_{1}^{(i)}\| = 0, h_{0} = \gamma_{m} \|\mathbf{H}\|$  and  $h_{1}^{(i)} = \gamma_{m} f_{k} \|\mathbf{H}\| \|\mathbf{R}^{(i)}\|$  we obtain  $\|\mathbf{\Delta}_{k}^{(i)}\| \leq \gamma_{m} \tilde{\eta}_{k} \|\mathbf{R}^{(i)}\|$  with  $\tilde{\eta}_{k} = f_{k} \|\mathbf{H}\| g_{k-1}^{1}$ .

LEMMA 3.8. The norm of the residual  $\mathbf{R}^{(i)}$  satisfies  $\|\mathbf{R}^{(i)}\| \leq 2\|\mathbf{H}\| \sin \angle (\mathcal{S}^{(i)}, \mathcal{S})$ 

*Proof.* We have,  $\mathbf{R}^{(i)} = \mathbf{H}\mathbf{X}^{(i)} - \mathbf{X}^{(i)}\mathbf{\Lambda}^{(i)}$  and the subspace diagonalization performed in the Rayleigh-Ritz projection step associated with  $(i-1)^{th}$  iteration allows one to write  $\mathbf{\Lambda}^{(i)}$  as  $\mathbf{\Lambda}^{(i)} = \mathbf{X}^{(i)\dagger}\mathbf{H}\mathbf{X}^{(i)}$ . Subsequently we can write the residual  $\mathbf{R}^{(i)} = (\mathbf{I} - \mathbf{X}^{(i)}\mathbf{X}^{(i)\dagger})\mathbf{H}\mathbf{X}^{(i)}$ . We note that  $\mathbf{H} = \hat{\mathbf{U}}_{1}\mathbf{\Lambda}_{1}\hat{\mathbf{U}}_{1}^{\dagger} + \hat{\mathbf{U}}_{2}\mathbf{\Lambda}_{2}\hat{\mathbf{U}}_{2}^{\dagger}$  and consequently we can write

$$\begin{aligned} \|\mathbf{R}^{(i)}\| &= \|(\mathbf{I} - \mathbf{X}^{(i)}\mathbf{X}^{(i)^{\dagger}})\mathbf{H}\mathbf{X}^{(i)}\| = \|(\mathbf{I} - \mathbf{X}^{(i)}\mathbf{X}^{(i)^{\dagger}})(\hat{\mathbf{U}}_{1}\mathbf{\Lambda}_{1}\hat{\mathbf{U}}_{1}^{\dagger} + \hat{\mathbf{U}}_{2}\mathbf{\Lambda}_{2}\hat{\mathbf{U}}_{2}^{\dagger})\mathbf{X}^{(i)}\| \\ &\leq \|(\mathbf{I} - \mathbf{X}^{(i)}\mathbf{X}^{(i)^{\dagger}})\hat{\mathbf{U}}_{1}\| \|\mathbf{\Lambda}_{1}\| \|\hat{\mathbf{U}}_{1}^{\dagger}\mathbf{X}^{(i)}\| + \|(\mathbf{I} - \mathbf{X}^{(i)}\mathbf{X}^{(i)^{\dagger}})\hat{\mathbf{U}}_{2}\| \|\mathbf{\Lambda}_{2}\| \|\hat{\mathbf{U}}_{2}^{\dagger}\mathbf{X}^{(i)}\| \end{aligned}$$

Using the fact that  $\sin \angle (\mathcal{S}^{(i)}, \mathcal{S}) = \| (\mathbf{I} - \mathbf{X}^{(i)} \mathbf{X}^{(i)^{\dagger}}) \hat{\mathbf{U}}_1 \| = \| \hat{\mathbf{U}}_2^{\dagger} \mathbf{X}^{(i)} \|$  and the inequalities  $\| \hat{\mathbf{U}}_1^{\dagger} \mathbf{X}^{(i)} \| \leq 1$ ,  $\| (\mathbf{I} - \mathbf{X}^{(i)^{\dagger}} \mathbf{X}^{(i)}) \hat{\mathbf{U}}_2 \| \leq 1$  we can write

(3.16)  $\|\mathbf{R}^{(i)}\| \leq \sin \angle (\mathcal{S}^{(i)}, \mathcal{S})(\|\mathbf{\Lambda}_1\| + \|\mathbf{\Lambda}_2\|) \leq 2\|\mathbf{H}\| \sin \angle (\mathcal{S}^{(i)}, \mathcal{S}) \qquad \Box$ 

THEOREM 3.9. The necessary condition for  $\angle(\underline{S}^{(i+1)}, S) < \angle(S^{(i)}, S)$  for the case of residual-based Chebyshev filtering approach can be written as

(3.17) 
$$|C_p(\lambda_n)| - |C_p(\lambda_{n+1})| > 2 \|\mathbf{H}\| \gamma_m \tilde{\eta}_p \left(1 + \tan \angle (\mathcal{S}^{(i)}, \mathcal{S})\right)$$

and if this inequality is satisfied for  $i = i_0$  then it holds for all  $i > i_0$ 

*Proof.* From Lemmas 3.6 to 3.8 we have  $\|\hat{\boldsymbol{\Delta}}_{p}^{(i)}\| \leq 2\|\mathbf{H}\| \gamma_{m} \tilde{\eta}_{p} \sin \angle(\mathcal{S}^{(i)}, \mathcal{S})$  and in order to have  $\angle(\underline{\mathcal{S}}^{(i+1)}, \mathcal{S}) < \angle(\mathcal{S}^{(i)}, \mathcal{S})$  from (3.1) we require

(3.18) 
$$|C_p(\lambda_n)| - |C_p(\lambda_{n+1})| > 2 \|\mathbf{H}\| \gamma_m \tilde{\eta}_p \left(1 + \tan \angle (\mathcal{S}^{(i)}, \mathcal{S})\right) \qquad \Box$$

We note that if this inequality is satisfied for some  $i = i_0$  then it holds for all  $i > i_0$ as the RHS decreases with decreasing  $\angle(\mathcal{S}^{(i)}, \mathcal{S})$  and hence we conclude that the R-ChFSI method can converge under approximations where the ChFSI method fails.

**3.3. Generalized Eigenvalue Problems.** We now analyze the case of a Hermitian generalized eigenproblem in (2.1) given by  $\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{B}\mathbf{u}_i$  or equivalently  $\mathbf{H}\mathbf{u}_i = \lambda_i \mathbf{u}_i$  with  $\mathbf{H} = \mathbf{B}^{-1}\mathbf{A}$ . In constructing a subspace rich in the desired eigenvectors using Chebyshev filtered subspace iteration approach, we consider the specific case of approximating  $\mathbf{B}^{-1}$  with a matrix  $\mathbf{D}^{-1}$  such that  $\mathbf{D}^{-1}\mathbf{B} \approx \mathbf{I}$  in evaluating the matrix product  $\mathbf{H}\mathbf{Y}_k^{(i)}$  within the Chebyshev recurrence relation defined by (2.2). We note that specific forms of the approximation, such as choosing  $\mathbf{D}$  to be diagonal or block diagonal, can give significant performance benefits.

3.3.1. Traditional Chebyshev filtering approach with inexact matrixproducts involving  $\mathbf{D}^{-1}$ . Using the approximation  $\mathbf{D}^{-1}$  for  $\mathbf{B}^{-1}$ , the construction of filtered subspace using the recurrence relation in (2.2) is now accomplished by defining  $\underline{\mathbf{Y}}_{k}^{(i)} = C_{k}(\mathbf{D}^{-1}\mathbf{A})\mathbf{X}^{(i)}$ , where  $k = 2, \ldots, p$ , and subsequently the 2-term recurrence relation can be written as

(3.19) 
$$\underline{\mathbf{Y}}_{k+1}^{(i)} = a_k \mathbf{D}^{-1} \mathbf{A} \underline{\mathbf{Y}}_k^{(i)} + b_k \underline{\mathbf{Y}}_k^{(i)} + c_k \underline{\mathbf{Y}}_{k-1}^{(i)}$$

with the initial conditions  $\underline{\mathbf{Y}}_{0}^{(i)} = \mathbf{X}^{(i)}$  and  $\underline{\mathbf{Y}}_{1}^{(i)} = \sigma_{1} \frac{\mathbf{D}^{-1} \mathbf{A} - c \mathbf{I}}{e} \mathbf{X}^{(i)}$ .

THEOREM 3.10. The spectral norm of the error  $\mathbf{\Delta}_{k}^{(i)} = \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{Y}_{k}^{(i)}$  in the subspace construction using inexact matrix products of the recurrence relation (2.2) involving  $\mathbf{D}^{-1}$  satisfies  $\|\mathbf{\Delta}_{k}^{(i)}\| \leq \zeta \eta_{k}$  for  $k = 0, 1, \ldots, p$  where  $\eta_{k}$  is some finite constant that depends on k and  $\zeta = \|\mathbf{D}^{-1} - \mathbf{B}^{-1}\|$ 

*Proof.* Using (2.2) and (3.19) we can now write a recurrence relation for  $\Delta_k^{(i)} = \underline{\mathbf{Y}}_k^{(i)} - \mathbf{Y}_k^{(i)}$  as

(3.20) 
$$\boldsymbol{\Delta}_{k+1}^{(i)} = a_k \mathbf{H} \boldsymbol{\Delta}_k^{(i)} + b_k \boldsymbol{\Delta}_k^{(i)} + c_k \boldsymbol{\Delta}_{k-1}^{(i)} + a_k \left( \mathbf{D}^{-1} \mathbf{A} \underline{\mathbf{Y}}_k^{(i)} - \mathbf{B}^{-1} \mathbf{A} \underline{\mathbf{Y}}_k^{(i)} \right)$$

with the initial conditions  $\mathbf{\Delta}_{0}^{(i)} = 0$  and  $\mathbf{\Delta}_{1}^{(i)} = \frac{\sigma_{1}}{e} (\mathbf{D}^{-1} - \mathbf{B}^{-1}) \mathbf{X}^{(i)}$ . We note that this recurrence relation is of the same form as that of (3.6) in Lemma 3.3 with  $\mathbf{e}_{k}^{(i)} = \mathbf{D}^{-1} \mathbf{A} \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{B}^{-1} \mathbf{A} \underline{\mathbf{Y}}_{k}^{(i)}$ . We now have  $\|\mathbf{e}_{k}^{(i)}\| = \|\mathbf{D}^{-1} \mathbf{A} \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{B}^{-1} \mathbf{A} \underline{\mathbf{Y}}_{k}^{(i)}\| \leq \zeta \|\mathbf{A}\| \|\underline{\mathbf{Y}}_{k}^{(i)}\|$ . We further note that

$$\|\underline{\mathbf{Y}}_{k}^{(i)}\| = \|\mathbf{\Delta}_{k}^{(i)} + \mathbf{Y}_{k}^{(i)}\| \le \|\mathbf{\Delta}_{k}^{(i)}\| + \|C_{k}(\mathbf{H})\mathbf{X}^{(i)}\| \le \|\mathbf{\Delta}_{k}^{(i)}\| + \|C_{k}(\mathbf{H})\|$$

Thus we have  $\|\mathbf{e}_{k}^{(i)}\| \leq \zeta \|\mathbf{A}\| \|\mathbf{\Delta}_{k}^{(i)}\| + \zeta \|\mathbf{A}\| \|C_{k}(\mathbf{H})\|$ . Using Lemma 3.3 with  $\|\mathbf{\Delta}_{0}^{(i)}\| = 0$  and  $\|\mathbf{\Delta}_{1}^{(i)}\| = \|\frac{\sigma_{1}}{e}(\mathbf{D}^{-1} - \mathbf{B}^{-1})\mathbf{X}^{(i)}\| \leq |\frac{\sigma_{1}}{e}|\zeta, h_{0} = \zeta \|\mathbf{A}\|$  and  $h_{1}^{(i)} = \zeta \|\mathbf{A}\| \|C_{k}(\mathbf{H})\|$  we obtain  $\|\mathbf{\Delta}_{k}^{(i)}\| \leq \zeta \eta_{k}$  with  $\eta_{k} = g_{k-1}^{0}|\frac{\sigma_{1}}{e}| + \|\mathbf{A}\| \|C_{k}(\mathbf{H})\| g_{k-1}^{1}$ .

We now discuss this result in the context of (3.3). Without loss of generality, we assume that our initial guess at i = 0 satisfies (3.3) and this allows for  $\angle(S^{(i)}, S)$  to decrease as *i* increases. However, we note that the RHS of (3.3) does not monotonically decrease with  $\angle(S^{(i)}, S)$  and in fact increases as  $\angle(S^{(i)}, S)$  approaches 0. Thus, it stands to reason that beyond a certain value of  $\angle(S^{(i)}, S)$  (3.3) no longer holds, and we cannot approach the wanted eigenspace beyond this point. Assuming that  $\zeta \eta_p << |C_p(\lambda_n)| - |C_p(\lambda_{n+1})|$  we estimate that the closest angle that can be achieved is

(3.21) 
$$\angle(\mathcal{S}^{(i)}, \mathcal{S}) \approx \frac{\zeta \eta_p}{|C_p(\lambda_n)| - |C_p(\lambda_{n+1})|}$$

We now propose a residual-based reformulation of ChFSI for the generalized eigenvalue problem and argue that the proposed method does not suffer from this limitation.

3.3.2. Proposed residual-based Chebyshev filtering approach for generalized eigenproblem. We now propose a residual-based reformulation of the recurrence relation defined in (2.2). To this end we redefine the residual  $\mathbf{R}_{k}^{(i)} = \mathbf{B}^{-1}\mathbf{D}(C_{k}(\mathbf{H})\mathbf{X}^{(i)} - \mathbf{X}^{(i)}C_{k}(\mathbf{\Lambda}^{(i)})) = \mathbf{B}^{-1}\mathbf{D}(\mathbf{Y}_{k}^{(i)} - \mathbf{X}^{(i)}\mathbf{\Lambda}_{k}^{(i)})$  for  $k = 0, \ldots, p$ , where we have defined  $\mathbf{\Lambda}_{k}^{(i)} = C_{k}(\mathbf{\Lambda}^{(i)})$ .

PROPOSITION 3.11. The recurrence relation given by (2.2) can be reformulated in terms of the  $\mathbf{BR}_k^{(i)} = \mathbf{D}(\mathbf{Y}_k^{(i)} - \mathbf{X}^{(i)}\mathbf{\Lambda}_k^{(i)})$  as

(3.22) 
$$\mathbf{BR}_{k+1}^{(i)} = a_k \mathbf{DHD}^{-1} \mathbf{BR}_k^{(i)} + b_k \mathbf{BR}_k^{(i)} + c_k \mathbf{BR}_{k-1}^{(i)} + a_k \mathbf{DB}^{-1} \mathbf{R}^{(i)} \mathbf{\Lambda}_k^{(i)}$$

where  $\mathbf{BR}_{0}^{(i)} = 0$  and  $\mathbf{BR}_{1}^{(i)} = \frac{\sigma_{1}}{e} \mathbf{DR}^{(i)}$  with  $\mathbf{R}^{(i)} = \mathbf{HX}^{(i)} - \mathbf{X}^{(i)} \mathbf{\Lambda}^{(i)}$ .

*Proof.* The recurrence relation for  $\mathbf{\Lambda}_k^{(i)} = C_k(\mathbf{\Lambda}^{(i)})$  can be written as

$$\mathbf{\Lambda}_{k+1}^{(i)} = a_k \mathbf{\Lambda}_k^{(i)} \mathbf{\Lambda}^{(i)} + b_k \mathbf{\Lambda}_k^{(i)} + c_k \mathbf{\Lambda}_{k-1}^{(i)}$$

Multiplying with  $\mathbf{X}^{(i)}$  and subtracting from (2.2), we have

$$\mathbf{BR}_{k+1}^{(i)} = a_k \mathbf{DHD}^{-1} \mathbf{BR}_k^{(i)} + b_k \mathbf{BR}_k^{(i)} + c_k \mathbf{BR}_{k-1}^{(i)} + a_k \mathbf{DR}^{(i)} \mathbf{\Lambda}_k^{(i)} \qquad \Box$$

After computing  $\mathbf{BR}_{p}^{(i)}$  using this recurrence relation we can now evaluate  $\mathbf{Y}_{p}^{(i)}$ using the relation  $\mathbf{Y}_{p}^{(i)} = \mathbf{D}^{-1}\mathbf{BR}_{p}^{(i)} + \mathbf{X}^{(i)}\mathbf{\Lambda}_{p}^{(i)}$ . We note that using approximation  $\mathbf{DB}^{-1} \approx \mathbf{I}$  we can write the following recurrence relation

(3.23) 
$$\mathbf{B}\underline{\mathbf{R}}_{k+1}^{(i)} = a_k \mathbf{A} \mathbf{D}^{-1} \mathbf{B}\underline{\mathbf{R}}_k^{(i)} + b_k \mathbf{B}\underline{\mathbf{R}}_k^{(i)} + c_k \mathbf{B}\underline{\mathbf{R}}_{k-1}^{(i)} + a_k \mathbf{B}\mathbf{R}^{(i)} \mathbf{\Lambda}_k^{(i)}$$

with  $\mathbf{BR}_{0}^{(i)} = 0$  and  $\mathbf{BR}_{1}^{(i)} = \frac{\sigma_{1}}{e} \mathbf{BR}^{(i)}$  and consequently, we have  $\underline{\mathbf{Y}}_{k}^{(i)} = \mathbf{D}^{-1} \mathbf{B} \underline{\mathbf{R}}_{k}^{(i)} + \mathbf{X}^{(i)} \mathbf{\Lambda}_{k}^{(i)}$ . Note that this recurrence relation does not require the evaluation of  $\mathbf{B}^{-1}$ .

LEMMA 3.12. The spectral norm of the error  $\mathbf{\Delta}_{k}^{(i)} = \underline{\mathbf{Y}}_{k}^{(i)} - \mathbf{Y}_{k}^{(i)}$  in the subspace construction using recurrence relation given by (3.23) satisfies  $\|\mathbf{\Delta}_{k}^{(i)}\| \leq \zeta \tilde{\eta}_{k} \|\mathbf{R}^{(i)}\|$  for  $k = 0, 1, \ldots, p$  where  $\tilde{\eta}_{k}$  is some finite constant that depends on k and  $\zeta = \|\mathbf{D}^{-1} - \mathbf{B}^{-1}\|$ .

*Proof.* Using (3.22) and (3.23) we can now write a recurrence relation for  $\Delta_k^{(i)} = \underline{\mathbf{Y}}_k^{(i)} - \mathbf{Y}_k^{(i)} = \mathbf{D}^{-1} \mathbf{B}(\underline{\mathbf{R}}_k^{(i)} - \mathbf{R}_k^{(i)})$  as

$$\boldsymbol{\Delta}_{k+1}^{(i)} = a_k \mathbf{H} \boldsymbol{\Delta}_k^{(i)} + b_k \boldsymbol{\Delta}_k^{(i)} + c_k \boldsymbol{\Delta}_{k-1}^{(i)} + a_k \left( \mathbf{D}^{-1} - \mathbf{B}^{-1} \right) \left( \mathbf{B} \mathbf{R}^{(i)} \boldsymbol{\Lambda}_k^{(i)} + \mathbf{A} \mathbf{D}^{-1} \mathbf{B} \mathbf{R}_k^{(i)} \right)$$

with the initial conditions  $\mathbf{\Delta}_{0}^{(i)} = 0$  and  $\mathbf{\Delta}_{1}^{(i)} = \frac{\sigma_{1}}{e} (\mathbf{D}^{-1} - \mathbf{B}^{-1}) \mathbf{B} \mathbf{R}^{(i)}$ . We note that this recurrence relation is of the same form as that of (3.6) in Lemma 3.3 with

$$\mathbf{e}_{k}^{(i)} = \left(\mathbf{D}^{-1} - \mathbf{B}^{-1}\right) \left(\mathbf{B}\mathbf{R}^{(i)}\mathbf{\Lambda}_{k}^{(i)} + \mathbf{A}\mathbf{D}^{-1}\mathbf{B}\underline{\mathbf{R}}_{k}^{(i)}\right)$$
$$\implies \|\mathbf{e}_{k}^{(i)}\| \leq \zeta (\|\mathbf{B}\| \|\mathbf{R}^{(i)}\| \|\mathbf{\Lambda}_{k}^{(i)}\| + \|\mathbf{A}\| \|\mathbf{D}^{-1}\mathbf{B}\underline{\mathbf{R}}_{k}^{(i)}\|)$$

We further note that using Lemma 3.6 we can write

$$\|\mathbf{D}^{-1}\mathbf{B}\underline{\mathbf{R}}_{k}^{(i)}\| = \|\mathbf{\Delta}_{k}^{(i)} + \mathbf{D}^{-1}\mathbf{B}\mathbf{R}_{k}^{(i)}\| \le \|\mathbf{\Delta}_{k}^{(i)}\| + f_{k}\|\mathbf{D}^{-1}\mathbf{B}\| \|\mathbf{R}^{(i)}\|$$

Thus we have  $\|\mathbf{e}_{k}^{(i)}\| \leq \zeta \|\mathbf{A}\| \|\mathbf{\Delta}_{k}^{(i)}\| + \zeta (f_{k} \|\mathbf{A}\| \|\mathbf{D}^{-1}\mathbf{B}\| + \|\mathbf{B}\| \|\mathbf{\Lambda}_{k}^{(i)}\|$ ). Now, using Lemma 3.3 with  $\|\mathbf{\Delta}_{0}^{(i)}\| = 0$  and  $\|\mathbf{\Delta}_{1}^{(i)}\| = \|\frac{\sigma_{1}}{e}(\mathbf{D}^{-1}-\mathbf{B}^{-1})\mathbf{B}\mathbf{R}^{(i)}\| \leq |\frac{\sigma_{1}}{e}|\zeta\|\mathbf{B}\| \|\mathbf{R}^{(i)}\|$ ,  $h_{0} = \zeta \|\mathbf{A}\|$  and  $h_{1}^{(i)} = \zeta (f_{k} \|\mathbf{A}\| \|\mathbf{D}^{-1}\mathbf{B}\| + \|\mathbf{B}\| \|\mathbf{\Lambda}_{k}^{(i)}\|)\|\mathbf{R}^{(i)}\|$  we obtain  $\|\mathbf{\Delta}_{k}^{(i)}\| \leq \zeta \tilde{\eta}_{k} \|\mathbf{R}^{(i)}\|$  with  $\tilde{\eta}_{k} = |\frac{\sigma_{1}}{e}| \|\mathbf{B}\| g_{k-1}^{0} + (f_{k} \|\mathbf{A}\| \|\mathbf{D}^{-1}\mathbf{B}\| + \|\mathbf{B}\| \|\mathbf{\Lambda}_{k}^{(i)}\|)g_{k-1}^{1}$ . THEOREM 3.13. The necessary condition for  $\angle(\underline{S}^{(i+1)}, S) < \angle(S^{(i)}, S)$  for the case of residual-based Chebyshev filtering approach can be written as

(3.24) 
$$|C_p(\lambda_n)| - |C_p(\lambda_{n+1})| > 2 \|\mathbf{H}\| \zeta \tilde{\eta}_p \left(1 + \tan \angle (\mathcal{S}^{(i)}, \mathcal{S})\right)$$

and if this inequality is satisfied for at  $i = i_0$  then it holds for all  $i > i_0$ 

*Proof.* As a consequence of Lemmas 3.6, 3.8, and 3.12 we can now write  $\|\hat{\boldsymbol{\Delta}}_{p}^{(i)}\| \leq 2\|\boldsymbol{H}\| \zeta \tilde{\eta}_{p} \sin \angle (\mathcal{S}^{(i)}, \mathcal{S})$  and in order to have  $\angle (\underline{\mathcal{S}}^{(i+1)}, \mathcal{S}) < \angle (\mathcal{S}^{(i)}, \mathcal{S})$  from (3.1) we require

(3.25) 
$$|C_p(\lambda_n)| - |C_p(\lambda_{n+1})| > 2 \|\mathbf{H}\| \zeta \tilde{\eta}_p \left(1 + \tan \angle (\mathcal{S}^{(i)}, \mathcal{S})\right) \qquad \Box$$

We note that if this inequality is satisfied for some  $i = i_0$  then it holds for all  $i > i_0$ as the RHS decreases with decreasing  $\angle(\mathcal{S}^{(i)}, \mathcal{S})$  and hence we conclude that the R-ChFSI method can converge under approximations where the ChFSI method fails even for generalized eigenvalue problems.

4. Results and Discussion. We present a detailed evaluation of the accuracy and efficiency achieved by the proposed residual-based Chebyshev filtered subspace iteration (R-ChFSI) method for solving real symmetric and complex Hermitian eigenvalue problems both for standard and generalized eigenproblems. We compare the performance of the R-ChFSI method with that of traditional ChFSI approaches, demonstrating its improved accuracy even when employing inexact matrix-vector products. Additionally, we investigate its performance on "Eos" NVIDIA DGX SuperPOD with NVIDIA H100 GPUs and Intel Xeon Platinum 8480C CPUs, demonstrating its ability to maintain high accuracy while achieving greater computational efficiency. For generalized eigenvalue problems of the form  $\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x}$ , we explore the use of diagonal approximations of **B** to compute its inverse for the subspace construction step within both the ChFSI and R-ChFSI methods, showing that the latter can achieve significantly lower residual tolerances. For all our benchmarks studies reported in this work, we consider the sparse matrix eigenvalue problems that arise from higher-order finite-element discretization of Kohn-Sham density functional theory (DFT) [2, 34, 38, 37, 7, 28, 31, 29].

In the following subsections, we focus on specific problem classes, starting with the results for standard eigenvalue problems, including the cases of real symmetric and complex Hermitian matrices, followed by generalized eigenvalue problems, where we assess the convergence behavior and the performance of the ChFSI and the R-ChFSI methods. To this end, we consider a set of three benchmark systems whose dimensions are summarized in Table 1. These eigenvalue problems are obtained from the finite-element discretization of the Kohn-Sham DFT equations for the periodic supercells of various sizes for body-centered cubic Molybdenum systems with a single vacancy. For all the benchmarks we utilize a Chebyshev polynomial degree of 75.

4.1. Standard Eigenvalue Problems. We now evaluate the accuracy and performance of the proposed R-ChFSI method as described in Algorithm 4.1 for solving the standard eigenvalue problems. We choose the eigenproblem encountered in using the higher-order finite-element discretization of Kohn-Sham DFT as outlined in [7, 28]. We begin by comparing the R-ChFSI method to the ChFSI method using FP32 and TF32 arithmetic. We demonstrate below that the R-ChFSI method with FP32 and TF32 arithmetic converges to the desired residual tolerance in a similar

## TABLE 1

Dimensions of the benchmark problems considered. These correspond to 3 material systems comprising  $6 \times 6 \times 6$ ,  $8 \times 8 \times 8$  and  $10 \times 10 \times 10$  body-centered-cubic periodic supercells of Molybdenum with a single vacancy.

System	# of DoFs (m)	# of wanted eigenvectors (n)	Subspace dimension
(1)	1728000	3000	3600
(2)	4251528	7000	8400
(3)	7645373	14000	16800

number of iterations as that of the ChFSI method employing FP64 arithmetic. In contrast, the ChFSI method with FP32 and TF32 arithmetic fails to reach the same residual tolerance as that of the ChFSI method with FP64 arithmetic.

Algorithm 4.1 R-ChFSI procedure for standard Hermitian eigenvalue problems

**INPUTS:** Chebyshev polynomial order p, estimates of the bounds of the eigenspectrum  $\lambda_{max}, \lambda_{min}$ , estimate of the upper bound of the wanted spectrum  $\lambda_T$  and the initial guess of eigenvectors  $\mathbf{X}^{(i)}$  and eigenvalues  $\mathbf{\Lambda}^{(i)}$  **OUTPUT:** The filtered subspace  $\mathbf{Y}_p^{(i)}$  **TEMPORARY VARIABLES:X**,  $\mathbf{Y}, \mathbf{R}_{\mathbf{X}}, \mathbf{R}_{\mathbf{Y}}, \mathbf{\Lambda}_{\mathbf{X}}$  and  $\mathbf{\Lambda}_{\mathbf{Y}}$   $e \leftarrow \frac{\lambda_{max} - \lambda_T}{2}; c \leftarrow \frac{\lambda_{max} + \lambda_T}{2}; \sigma \leftarrow \frac{e}{\lambda_{min} - c}; \sigma_1 \leftarrow \sigma; \gamma \leftarrow \frac{2}{\sigma_1}$  $\mathbf{X} \leftarrow \mathbf{X}^{(i)}; \mathbf{Y} \leftarrow \mathbf{A}\mathbf{X}^{(i)} - \mathbf{X}^{(i)} \mathbf{\Lambda}^{(i)}$ 

$$\begin{aligned} \mathbf{X} \leftarrow \mathbf{X}^{(t)}; \ \mathbf{Y} \leftarrow \mathbf{A}\mathbf{X}^{(t)} - \mathbf{X}^{(t)}\mathbf{\Lambda}^{(t)} \\ \mathbf{R}_{\mathbf{X}} \leftarrow 0; \ \mathbf{R}_{\mathbf{Y}} \leftarrow \frac{2\sigma_1}{e}\mathbf{Y} \\ \mathbf{\Lambda}_{\mathbf{X}} \leftarrow \mathbf{I}; \ \mathbf{\Lambda}_{\mathbf{Y}} \leftarrow \frac{2\sigma_1}{e}\left(\mathbf{\Lambda}^{(i)} - c\mathbf{I}\right) \\ \text{for } k \leftarrow 2 \text{ to } p \text{ do} \\ \sigma_2 \leftarrow \frac{1}{\gamma - \sigma} \\ \mathbf{R}_{\mathbf{X}} \leftarrow \frac{2\sigma_2}{e}\mathbf{A}\mathbf{R}_{\mathbf{Y}} - \frac{2\sigma_2}{e}c\mathbf{R}_{\mathbf{Y}} - \sigma\sigma_2\mathbf{R}_{\mathbf{X}} + \frac{2\sigma_2}{e}\mathbf{Y}\mathbf{\Lambda}_{\mathbf{Y}} \\ \mathbf{\Lambda}_{\mathbf{X}} \leftarrow \frac{2\sigma_2}{e}\mathbf{\Lambda}_{\mathbf{Y}}\mathbf{\Lambda}^{(i)} - \frac{2\sigma_2}{e}c\mathbf{\Lambda}_{\mathbf{Y}} - \sigma\sigma_2\mathbf{\Lambda}_{\mathbf{X}} \\ \text{swap}(\mathbf{X}, \mathbf{Y}); \text{ swap}(\mathbf{\Lambda}_{\mathbf{X}}, \mathbf{\Lambda}_{\mathbf{Y}}); \ \sigma = \sigma_2 \\ \text{end for} \\ \mathbf{X} \leftarrow \mathbf{R}_{\mathbf{Y}} + \mathbf{X}\mathbf{\Lambda}_{\mathbf{Y}} \\ \text{return } \mathbf{X} \end{aligned}$$

4.1.1. Symmetric Eigenvalue Problems. We now consider the finite-element discretization of the Kohn-Sham DFT equations sampled at the origin (Gamma point) of the Brillouin zone for the material systems described in Table 1. This ensures that the resulting discretized equation is a real symmetric eigenvalue problem. The computation of the matrix multi-vector products with FP64 arithmetic is performed according to the methodology prescribed by [7, 28]. When using FP32 arithmetic to evaluate  $AR_Y$  in Algorithm 4.1, the data structures storing the matrix A and multi-vectors  $R_X$  and  $R_Y$  are in FP32 format and the BLAS calls used to perform  $AR_Y$  are replaced with corresponding FP32 variants while the addition and scaling of multi-vectors is done using FP64 arithmetic. For TF32 arithmetic on NVIDIA GPUs we replace the cublasSgemm and its strided/batched variants with cublasGemmEx and its strided/batched variants with computeType set to CUBLAS\_COMPUTE\_32F\_FAST\_TF32.

From Figure 1, we observe that for all the benchmark systems summarized in Table 1, the accuracy achieved by the FP32 and TF32 variants for the R-ChFSI method is comparable to that achieved by the FP64 variant of the ChFSI method, whereas the FP32 and TF32 variants of the ChFSI method show significant reduction



FIG. 1. Plot of  $\max_j r_j^{(i)} = \max_j \|\mathbf{A} \mathbf{x}_j^{(i)} - \epsilon_j^{(i)} \mathbf{x}_j^{(i)}\|$  as the iterations progress to solve the symmetric standard eigenvalue problem with various precisions for the benchmark systems described in Table 1.

in achievable residual tolerance. We also note that the rate of convergence that is achieved by the FP32 and TF32 variants for the R-ChFSI method is comparable to that achieved by the FP64 variant of the ChFSI method. We also report the



FIG. 2. Speedups of lower precision ChFSI and R-ChFSI methods over the FP64 ChFSI method for subspace construction to solve the symmetric standard eigenvalue problem on GPUs for the benchmark systems described in Table 1.

speedups achieved by the FP32 and TF32 variants of ChFSI and R-ChFSI over the FP64 variant of ChFSI in Figure 2 on NVIDIA H100 GPU accelerators. We note that we obtain similar speedups for both the ChFSI and R-ChFSI methods allowing us to conclude that the additional overheads in the R-ChFSI method are negligible. We have achieved speedups of up to 1.75x for the TF32 variant of the R-ChFSI method over the baseline implementation of the ChFSI method with FP64 arithmetic. We also report the speedups achieved by the FP32 variant of ChFSI and R-ChFSI over the FP64 variant of ChFSI in Figure 3 on Intel Xeon Platinum 8480C CPUs. We note that the speedups of the R-ChFSI method are slightly lower than ChFSI method due to the additional overheads in the R-ChFSI method. We achieve speedups of up to 2.1x for the FP32 variant of the R-ChFSI method over the baseline implementation of the ChFSI method over the baseline implementation of the R-ChFSI method. We achieve speedups of up to 2.1x for the FP32 variant of the R-ChFSI method over the baseline implementation of the ChFSI method over the baseline implementation of the R-ChFSI method over the baseline implementation of the ChFSI method with FP64 arithmetic.

**4.1.2. Hermitian Eigenvalue Problems.** We now consider the finite-element discretization of the Kohn-Sham DFT equations at a non-zero k-point [7] in the Brillouin zone for the systems described in Table 1. The resulting discretized equation is now a complex Hermitian eigenvalue problem. The computation of the matrix multi-vector products with FP64 arithmetic is done according to the methodology



FIG. 3. Speedups of lower precision ChFSI and R-ChFSI methods over the FP64 ChFSI method for subspace construction to solve the symmetric standard eigenvalue problem on CPUs for the benchmark systems described in Table 1.

prescribed by [7, 28]. When using FP32 arithmetic to evaluate  $AR_Y$  in Algorithm 4.1, the data structures storing the matrix A and multi-vectors  $R_X$  and  $R_Y$  are in FP32 format and the BLAS calls used to perform  $AR_Y$  are replaced with corresponding FP32 variants while the addition and scaling of multi-vectors is done using FP64 arithmetic. For TF32 arithmetic on NVIDIA GPUs we replace the cublasCgemm and its strided/batched variants with cublasGemmEx a



FIG. 4. Plot of  $\max_j r_j^{(i)} = \max_j \|\mathbf{A}\mathbf{x}_j^{(i)} - \epsilon_j^{(i)}\mathbf{x}_j^{(i)}\|$  as the iterations progress to solve the Hermitian standard eigenvalue problem for various precisions to solve the Hermitian standard eigenvalue problem for the benchmark systems described in Table 1.

for all the benchmark systems summarized in Table 1, the accuracy achieved by the FP32 and TF32 variants for the R-ChFSI method is comparable to that achieved by the FP64 variant of the ChFSI method, whereas the FP32 and TF32 variants of the R-ChFSI method show a significant reduction in achievable residual tolerance. We also note that the rate of convergence that is achieved by the FP32 and TF32 variants for the R-ChFSI method is comparable to that achieved by the FP64 variant of the ChFSI.

We also report the speedups achieved by the FP32 and TF32 variants of ChFSI and R-ChFSI over the FP64 variant of ChFSI in Figure 5 on NVIDIA H100 GPU accelerators. We note that we obtain similar speedups for both the ChFSI and R-ChFSI methods allowing us to conclude that the additional overheads in the R-ChFSI method are negligible. We achieve speedups of up to 2.0x for the TF32 variant of the R-ChFSI method over the baseline implementation of the ChFSI method with FP64



FIG. 5. Speedups of lower precision ChFSI and R-ChFSI methods over the FP64 ChFSI method for subspace construction to solve the Hermitian standard eigenvalue problem on GPUs for the benchmark systems described in Table 1.

arithmetic.



FIG. 6. Speedups of lower precision ChFSI and R-ChFSI methods over the FP64 ChFSI method for subspace construction to solve the Hermitian standard eigenvalue problem on CPUs for the benchmark systems described in Table 1.

We also report the speedups achieved by the FP32 variant of ChFSI and R-ChFSI over the FP64 variant of ChFSI in Figure 6 on Intel Xeon Platinum 8480C CPUs. We note that the speedups of the R-ChFSI method are slightly lower due to the additional overheads in the R-CHFSI method. We have achieved speedups of up to 2.0x for the FP32 variant of the R-ChFSI method over the baseline implementation of the ChFSI method with FP64 arithmetic.

4.2. Generalized Eigenvalue Problems. We now evaluate the accuracy and performance of the proposed R-ChFSI method for generalized eigenvalue problems as described in Algorithm 4.2 resulting from the higher-order finite-element discretized Kohn-Sham DFT equations. The generalized eigen problem arises in the finite-element discretization of DFT equations because of the non-orthogonality of the FE basis functions. We begin by comparing the R-ChFSI method to the ChFSI method using a diagonal approximation of the finite-element overlap matrix. To this end, we consider the set of three benchmark systems whose dimensions are summarized in Table 1. We demonstrate that the R-ChFSI method converges to a significantly lower residual tolerance than the ChFSI method when the diagonal approximation is employed for approximating the inverse overlap matrix during the subspace construction.

Algorithm 4.2 R-ChFSI procedure for generalized Hermitian eigenvalue problems

**INPUTS:** Chebyshev polynomial order p, estimates of the bounds of the eigenspectrum  $\lambda_{max}, \lambda_{min}$ , estimate of the upper bound of the wanted spectrum  $\lambda_T$  and the initial guess of eigenvectors  $\mathbf{X}^{(i)}$  and eigenvalues  $\mathbf{\Lambda}^{(i)}$  **OUTPUT:** The filtered subspace  $\mathbf{Y}_p^{(i)}$  **TEMPORARY VARIABLES:**  $\mathbf{X}, \mathbf{Y}, \mathbf{R}_{\mathbf{X}}, \mathbf{R}_{\mathbf{Y}}, \mathbf{\Lambda}_{\mathbf{X}}$  and  $\mathbf{\Lambda}_{\mathbf{Y}}$   $e \leftarrow \frac{\lambda_{max} - \lambda_T}{2}; c \leftarrow \frac{\lambda_{max} + \lambda_T}{2}; \sigma \leftarrow \frac{e}{\lambda_{min} - c}; \sigma_1 \leftarrow \sigma; \gamma \leftarrow \frac{2}{\sigma_1}$   $\mathbf{X} \leftarrow \mathbf{X}^{(i)}; \mathbf{Y} \leftarrow \mathbf{A}\mathbf{X}^{(i)} - \mathbf{B}\mathbf{X}^{(i)}\mathbf{\Lambda}^{(i)}$   $\mathbf{R}_{\mathbf{X}} \leftarrow 0; \mathbf{R}_{\mathbf{Y}} \leftarrow \frac{2\sigma_1}{e} \mathbf{Y}$   $\mathbf{\Lambda}_{\mathbf{X}} \leftarrow \mathbf{I}; \mathbf{\Lambda}_{\mathbf{Y}} \leftarrow \frac{2\sigma_1}{e} \left(\mathbf{\Lambda}^{(i)} - c\mathbf{I}\right)$ for  $k \leftarrow 2$  to p do  $\sigma_2 \leftarrow \frac{1}{\gamma - \sigma}$   $\mathbf{R}_{\mathbf{X}} \leftarrow \frac{2\sigma_2}{e} \mathbf{A}\mathbf{D}^{-1}\mathbf{R}_{\mathbf{Y}} - \frac{2\sigma_2}{e} c\mathbf{R}_{\mathbf{Y}} - \sigma\sigma_2\mathbf{R}_{\mathbf{X}} + \frac{2\sigma_2}{e} \mathbf{Y}\mathbf{\Lambda}_{\mathbf{Y}}$   $\mathbf{\Lambda}_{\mathbf{X}} \leftarrow \frac{2\sigma_2}{e} \mathbf{A}\mathbf{Y}\mathbf{\Lambda}^{(i)} - \frac{2\sigma_2}{e} c\mathbf{\Lambda}_{\mathbf{Y}} - \sigma\sigma_2\mathbf{\Lambda}_{\mathbf{X}}$ swap $(\mathbf{X}, \mathbf{Y}); \operatorname{swap}(\mathbf{\Lambda}_{\mathbf{X}}, \mathbf{\Lambda}_{\mathbf{Y}}); \sigma = \sigma_2$ end for  $\mathbf{X} \leftarrow \mathbf{D}^{-1}\mathbf{R}_{\mathbf{Y}} + \mathbf{X}\mathbf{\Lambda}_{\mathbf{Y}}$ return  $\mathbf{X}$ 

**4.2.1. Symmetric Eigenvalue Problems.** We now consider the finite-element discretization of the Kohn-Sham DFT equations at the origin (Gamma point) in the Brillouin zone for the systems described in Table 1. This ensures that the resulting discretized equation is a real symmetric eigenvalue problem. All the computations are done with FP64 according to the methodology prescribed by [7, 28].



FIG. 7. Plot of  $\max_j r_j^{(i)} = \max_j \|\mathbf{A}\mathbf{x}_j^{(i)} - \epsilon_j^{(i)} \mathbf{B}\mathbf{x}_j^{(i)}\|$  as the iterations progress for the ChFSI method and the R-ChFSI method to solve the symmetric generalized eigenvalue problem for the benchmark systems described in Table 1.

From Figure 7, we observe that for all the benchmark systems summarized in Table 1, the residual tolerance than can be achieved by the R-ChFSI method is nearly 10 orders of magnitude lower than what can be achieved using the ChFSI method when employing the diagonal approximation for the inverse overlap matrix. We also note that our proposed R-ChFSI algorithm allows us to construct the filtered subspace using lower precision and we report the residuals achieved by the FP32 and TF32 variants of R-ChFSI in Figure 8. We also report the speedups achieved by the FP32 and TF32 variants of R-ChFSI over the FP64 variant of R-ChFSI in Figure 9 on NVIDIA H100 GPU accelerators. We achieve speedups of up to 2.3x for



FIG. 8. Plot of  $\max_j r_j^{(i)} = \max_j \|\mathbf{A}\mathbf{x}_j^{(i)} - \epsilon_j^{(i)} \mathbf{B}\mathbf{x}_j^{(i)}\|$  as the iterations progress for the R-ChFSI method to solve the symmetric generalized eigenvalue problem with various precisions for the benchmark systems described in Table 1.



FIG. 9. Speedups of lower precision R-ChFSI methods over the FP64 R-ChFSI method for subspace construction to solve the symmetric generalized eigenvalue problem for the benchmark systems described in Table 1.

the TF32 variant of the R-ChFSI method over the baseline implementation of the ChFSI method with FP64 arithmetic.

4.2.2. Hermitian Eigenvalue Problems. We now consider the finite-element discretization of the Kohn-Sham DFT equations at a non-zero k-point [7] in the Brillouin zone for the systems described in Table 1. The resulting discretized equation is now a complex Hermitian eigenvalue problem. All the computations are done with FP64 according to the methodology prescribed by [7, 28]. From Figure 10, we observe that for all the benchmark systems summarized in Table 1, even for Hermitian eigenvalue problems the residual tolerance than can be achieved by the R-ChFSI method is nearly 10 orders of magnitude lower than what can be achieved using the ChFSI method when employing the diagonal approximation for the inverse overlap matrix. We also note that our proposed R-ChFSI algorithm allows for us to construct the filtered subspace using lower precision and we report the residuals achieved by the FP32 and TF32 variants of R-ChFSI in Figure 11. We also report the speedups achieved by the FP32 and TF32 variants of R-ChFSI over the FP64 variant of R-ChFSI in Figure 12 on NVIDIA H100 GPU accelerators. We achieve speedups of up to 2.7x for the TF32 variant of the R-ChFSI method over the baseline implementation of the R-ChFSI method with FP64 arithmetic.

5. Conclusion. In this work, we presented a residual-based Chebyshev filtered subspace iteration (R-ChFSI) method tailored for large sparse Hermitian eigenvalue



FIG. 10. Plot of  $\max_j r_j^{(i)} = \max_j \|A x_j^{(i)} - \epsilon_j^{(i)} B x_j^{(i)}\|$  as the iterations progress for the ChFSI method and the *R*-ChFSI method to solve the Hermitian generalized eigenvalue problem for the benchmark systems described in Table 1.



FIG. 11. Plot of  $\max_j r_j^{(i)} = \max_j \|\mathbf{A}\mathbf{x}_j^{(i)} - \epsilon_j^{(i)} \mathbf{B}\mathbf{x}_j^{(i)}\|$  as the iterations progress for the R-ChFSI method to solve the Hermitian generalized eigenvalue problem with various precisions for the benchmark systems described in Table 1.



FIG. 12. Speedups of lower precision R-ChFSI methods over the FP64 R-ChFSI method for subspace construction to solve the Hermitian generalized eigenvalue problem for the benchmark systems described in Table 1.

problems while being tolerant to inexact matrix-vector products. We have demonstrated that the proposed method significantly outperforms the standard Chebyshev filtered subspace iteration (ChFSI) approach through extensive analysis and numerical experiments. We have provided a mathematical justification demonstrating that this reformulation reduces the numerical error in the Chebyshev filtered subspace construction while employing inexact matrix-vector products compared to the traditional ChFSI recurrence relation. We have studied the performance and accuracy of R-ChFSI under different computational precisions, demonstrating that it maintains robust convergence properties even when using lower-precision arithmetic. This characteristic makes it particularly suitable for modern hardware architectures that prioritize low-precision operations for efficiency gains in AI training. The results indicate that R-ChFSI achieves similar residual tolerances as full-precision ChFSI while retaining the performance benefits of reduced precision arithmetic.

For standard eigenvalue problems, we demonstrate that the proposed R-ChFSI method leverages the advantages of low-precision arithmetic. This capability is critical in today's computational landscape, where the use of low-precision arithmetic (e.g., FP32, TF32) on modern hardware accelerators can dramatically reduce computational costs while the proposed R-ChFSI method still allows us to achieve high accuracy. Our benchmarks demonstrate that, for large-scale eigenproblems, R-ChFSI attains residual norms in the range of  $10^{-12}$  to  $10^{-14}$  which is a significant improvement over traditional ChFSI methods, that stagnate at residuals around  $10^{-6}$  when using lower precision. In the context of generalized eigenvalue problems of the form  $Ax = \lambda Bx$ , the proposed R-ChFSI method is particularly impactful due to its ability to incorporate approximate inverses of **B**. By replacing the exact computation of matrix inverses with computationally inexpensive approximate inverses (e.g., diagonal or block-diagonal approximations of **B**), the method not only reduces the overall computational burden but also maintains convergence robustness. This results in substantial performance gains, as the proposed approach achieves residual tolerances that are up to ten orders of magnitude lower than those obtainable with traditional ChFSI when solving generalized problems. We further note that even in the context of generalized eigenvalue problems, the proposed R-ChFSI method allows us to utilize low-precision arithmetic alongside approximate inverses, further enhancing the computational gains.

The proposed R-ChFSI method is particularly relevant for quantum mechanics and material science applications, where we encounter large-scale sparse Hermitian eigenvalue problems. By enabling the use of lower-precision arithmetic without compromising accuracy, the proposed method provides an efficient alternative for largescale simulations that require extensive computational resources. The ability to retain accuracy comparable to full precision solvers with reduced precision arithmetic further facilitates integration with modern accelerators, such as GPUs and specialized tensorprocessing hardware. The findings in this study open avenues for further optimization and application of R-ChFSI in large-scale scientific computing tasks, including electronic structure calculations and other domains requiring efficient sparse eigenvalue solvers.

6. Acknowledgements. The authors gratefully acknowledge the seed grant from the Indian Institute of Science and the SERB Startup Research Grant from the Department of Science and Technology India (Grant Number: SRG/2020/002194). P.M. acknowledges Prathu Tiwari at Nvidia, Bangalore, for helping us run the benchmarks on NVIDIA DGX SuperPOD Eos system. The research also used the resources of PARAM Pravega at the Indian Institute of Science, supported by the National Supercomputing Mission (NSM) for testing few of the benchmarks on CPUs. N.K and K.R would like to acknowledge the Prime Minister Research Fellowship(PMRF) from the Ministry of Education India for financial support. P.M. acknowledges the Google India Research Award 2023 for financial support during the course of this work.

REFERENCES

- W. E. ARNOLDI, The principle of minimized iterations in the solution of the matrix eigenvalue problem, Quarterly of Applied Mathematics, 9 (1951), pp. 17–29, https://doi.org/10.1090/ qam/42792, https://www.ams.org/qam/1951-09-01/S0033-569X-1951-42792-9/ (accessed 2025-03-28).
- [2] H. CHEN, X. DAI, X. GONG, L. HE, AND A. ZHOU, Adaptive Finite Element Approximations for Kohn-Sham Models, Multiscale Modeling & Simulation, 12 (2014), pp. 1828–1869, https://doi.org/10.1137/130916096.
- [3] Y. CHEN, Y. CHI, J. FAN, AND C. MA, Spectral Methods for Data Science: A Statistical Perspective, Foundations and Trends® in Machine Learning, 14 (2021), pp. 566–806, https: //doi.org/10.1561/2200000079.
- [4] M. CROUZEIX, B. PHILIPPE, AND M. SADKANE, The Davidson Method, SIAM Journal on Scientific Computing, 15 (1994), pp. 62–76, https://doi.org/10.1137/0915004, http://epubs. siam.org/doi/10.1137/0915004 (accessed 2025-03-28).
- [5] S. DAS, B. KANUNGO, V. SUBRAMANIAN, G. PANIGRAHI, P. MOTAMARRI, D. ROGERS, P. M. ZIMMERMAN, AND V. GAVINI, Large-Scale Materials Modeling at Quantum Accuracy: Ab Initio Simulations of Quasicrystals and Interacting Extended Defects in Metallic Alloys, International Conference for High Performance Computing, Networking, Storage and Analysis, SC, (2023), https://doi.org/10.1145/3581784.3627037. Publisher: IEEE Computer Society ISBN: 9798400701092.
- [6] S. DAS, P. MOTAMARRI, V. GAVINI, B. TURCKSIN, Y. W. LI, AND B. LEBACK, Fast, Scalable and Accurate Finite-Element Based Ab Initio Calculations Using Mixed Precision Computing: 46 PFLOPS Simulation of a Metallic Dislocation System, in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, vol. 11, New York, NY, USA, Nov. 2019, ACM, pp. 1–11, https://doi.org/10.1145/3295500. 3357157, https://dl.acm.org/doi/10.1145/3295500.3357157 (accessed 2022-11-10).
- S. DAS, P. MOTAMARRI, V. SUBRAMANIAN, D. M. ROGERS, AND V. GAVINI, DFT-FE 1.0: A massively parallel hybrid CPU-GPU density functional theory code using finite-element discretization, Computer Physics Communications, 280 (2022), p. 108473, https://doi.org/10.1016/j.cpc.2022.108473, https://linkinghub.elsevier.com/ retrieve/pii/S0010465522001928 (accessed 2022-08-15). Publisher: North-Holland.
- [8] E. R. DAVIDSON, The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices, Journal of Computational Physics, 17 (1975), pp. 87–94, https://doi.org/10.1016/0021-9991(75)90065-0, https://linkinghub.elsevier.com/retrieve/pii/0021999175900650 (accessed 2025-03-28).
- [9] A. DEKTOR, P. DELMASTRO, E. YE, R. V. BEEUMEN, AND C. YANG, Inexact subspace projection methods for low-rank tensor eigenvalue problems, Mar. 2025, https://doi. org/10.48550/arXiv.2502.19578, http://arxiv.org/abs/2502.19578 (accessed 2025-03-31). arXiv:2502.19578 [math].
- M. DOGAN, K.-H. LIOU, AND J. R. CHELIKOWSKY, Solving the electronic structure problem for over 100 000 atoms in real space, Physical Review Materials, 7 (2023), p. L063001, https://doi.org/10.1103/PhysRevMaterials.7.L063001, https://link.aps.org/ doi/10.1103/PhysRevMaterials.7.L063001.
- [11] J. DONGARRA, J. GUNNELS, H. BAYRAKTAR, A. HAIDAR, AND D. ERNST, Hardware Trends Impacting Floating-Point Computations In Scientific Applications, 2024, https://arxiv. org/abs/2411.12090. arXiv: 2411.12090.
- [12] J. FENG, L. WAN, J. LI, S. JIAO, X. CUI, W. HU, AND J. YANG, Massively parallel implementation of iterative eigensolvers in large-scale plane-wave density functional theory, Computer Physics Communications, 299 (2024), p. 109135, https://doi.org/10.1016/j.cpc. 2024.109135.
- [13] S. GHOSH AND P. SURYANARAYANA, SPARC: Accurate and efficient finite-difference formulation and parallel implementation of Density Functional Theory: Extended systems, Computer Physics Communications, 216 (2017), pp. 109–125, https://doi.org/10.1016/j.cpc.2017.02. 019.
- [14] N. J. HIGHAM, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics, Jan. 2002, https://doi.org/10.1137/1.9780898718027.
- [15] M. HOCHSTENBACH AND Y. NOTAY, The Jacobi–Davidson method, GAMM-Mitteilungen, 29 (2006), pp. 368–382, https://doi.org/10.1002/gamm.201490038, https://onlinelibrary. wiley.com/doi/10.1002/gamm.201490038 (accessed 2025-03-28).
- [16] A. KASHI, H. LU, W. BREWER, D. ROGERS, M. MATHESON, M. SHANKAR, AND F. WANG, Mixed-precision numerics in scientific applications: survey and perspectives, 2025, https: //arxiv.org/abs/2412.19322. arXiv: 2412.19322.
- [17] A. V. KNYAZEV, Toward the Optimal Preconditioned Eigensolver: Locally Optimal Block Pre-

24

conditioned Conjugate Gradient Method, SIAM Journal on Scientific Computing, 23 (2001), pp. 517–541, https://doi.org/10.1137/S1064827500366124, http://epubs.siam.org/doi/10.1137/S1064827500366124 (accessed 2025-03-28).

- [18] D. KRESSNER, Y. MA, AND M. SHAO, A mixed precision LOBPCG algorithm, Numerical Algorithms, 94 (2023), pp. 1653–1671, https://doi.org/10.1007/s11075-023-01550-9.
- [19] L. KRONIK, A. MAKMAL, M. L. TIAGO, M. M. G. ALEMANY, M. JAIN, X. HUANG, Y. SAAD, AND J. R. CHELIKOWSKY, PARSEC – the pseudopotential algorithm for real-space electronic structure calculations: recent advances and novel applications to nano-structures, physica status solidi (b), 243 (2006), pp. 1063–1079, https://doi.org/10.1002/pssb.200541463.
- [20] C. LANCZOS, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, Journal of Research of the National Bureau of Standards, 45 (1950), p. 255, https://doi.org/10.6028/jres.045.026, https://nvlpubs.nist.gov/nistpubs/jres/045/ jresv45n4p255\_A1b.pdf (accessed 2025-03-28).
- [21] A. LEVITT AND M. TORRENT, Parallel eigensolvers in plane-wave Density Functional Theory, Computer Physics Communications, 187 (2015), pp. 98–105, https://doi.org/10.1016/j. cpc.2014.10.015.
- [22] K.-H. LIOU, C. YANG, AND J. R. CHELIKOWSKY, Scalable implementation of polynomial filtering for density functional theory calculation in PARSEC, Computer Physics Communications, 254 (2020), p. 107330, https://doi.org/10.1016/j.cpc.2020.107330.
- [23] C. H. LOH, Chebyshev filtering Lanczos' process in the subspace iteration method, International Journal for Numerical Methods in Engineering, 20 (1984), pp. 182–186, https://doi.org/10. 1002/nme.1620200114, https://onlinelibrary.wiley.com/doi/10.1002/nme.1620200114 (accessed 2025-03-28).
- [24] C.-Q. MIAO, On Chebyshev-Davidson Method for Symmetric Generalized Eigenvalue Problems, Journal of Scientific Computing, 85 (2020), p. 53, https://doi.org/ 10.1007/s10915-020-01360-4, http://link.springer.com/10.1007/s10915-020-01360-4 (accessed 2025-03-28).
- [25] C.-Q. MIAO AND L. CHENG, On flexible block Chebyshev-Davidson method for solving symmetric generalized eigenvalue problems, Advances in Computational Mathematics, 49 (2023), p. 78, https://doi.org/10.1007/s10444-023-10078-4, https://link.springer.com/10. 1007/s10444-023-10078-4 (accessed 2025-03-28).
- [26] V. MICHAUD-RIOUX, L. ZHANG, AND H. GUO, RESCU: A real space electronic structure method, Journal of Computational Physics, 307 (2016), pp. 593–613, https://doi.org/10.1016/j.jcp. 2015.12.014.
- [27] R. B. MORGAN AND D. S. SCOTT, Generalizations of Davidson's Method for Computing Eigenvalues of Sparse Symmetric Matrices, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 817–825, https://doi.org/10.1137/0907054, http://epubs.siam.org/doi/ 10.1137/0907054 (accessed 2025-03-28).
- [28] P. MOTAMARRI, S. DAS, S. RUDRARAJU, K. GHOSH, D. DAVYDOV, AND V. GAVINI, DFT-FE – A massively parallel adaptive finite-element code for large-scale density functional theory calculations, Computer Physics Communications, 246 (2020), p. 106853, https://doi.org/10.1016/j.cpc.2019.07.016, https://linkinghub.elsevier.com/ retrieve/pii/S0010465519302309 (accessed 2022-09-20). Publisher: North-Holland.
- [29] P. MOTAMARRI, M. NOWAK, K. LEITER, J. KNAP, AND V. GAVINI, Higher-order adaptive finite-element methods for Kohn-Sham density functional theory, Journal of Computational Physics, 253 (2013), pp. 308-343, https://doi.org/10.1016/j.jcp.2013.06.042, https://linkinghub.elsevier.com/retrieve/pii/S0021999113004774.
- [30] NVIDIA, HGX AI Supercomputing Platform, 2025, https://www.nvidia.com/en-us/ data-center/hgx/ (accessed 2025-02-07).
- [31] J. E. PASK AND P. A. STERNE, Finite element methods in ab initio electronic structure calculations, Modelling and Simulation in Materials Science and Engineering, 13 (2005), pp. R71– R96, https://doi.org/10.1088/0965-0393/13/3/R01.
- [32] Y. SAAD, Chebyshev Acceleration Techniques for Solving Nonsymmetric Eigenvalue Problems, Mathematics of Computation, 42 (1984), p. 567, https://doi.org/10.2307/2007602.
- [33] Y. SAAD, Numerical Methods for Large Eigenvalue Problems, Society for Industrial and Applied Mathematics, Jan. 2011, https://doi.org/10.1137/1.9781611970739.
- [34] V. SCHAUER AND C. LINDER, All-electron Kohn-Sham density functional theory on hierarchic finite element spaces, Journal of Computational Physics, 250 (2013), pp. 644–664, https: //doi.org/10.1016/j.jcp.2013.04.020.
- [35] D. C. SORENSEN, Implicit Application of Polynomial Filters in a k-Step Arnoldi Method, SIAM Journal on Matrix Analysis and Applications, 13 (1992), pp. 357–385, https://doi.org/10. 1137/0613025, http://epubs.siam.org/doi/10.1137/0613025 (accessed 2025-03-28).

- [36] G. W. STEWART, A Krylov-Schur Algorithm for Large Eigenproblems, SIAM Journal on Matrix Analysis and Applications, 23 (2002), pp. 601–614, https://doi.org/ 10.1137/S0895479800371529, http://epubs.siam.org/doi/10.1137/S0895479800371529 (accessed 2025-03-28).
- [37] E. TSUCHIDA AND M. TSUKADA, Electronic-structure calculations based on the finite-element method, Physical Review B, 52 (1995), pp. 5573–5578, https://doi.org/10.1103/PhysRevB. 52.5573.
- [38] E. TSUCHIDA AND M. TSUKADA, Adaptive finite-element method for electronic-structure calculations, Physical Review B - Condensed Matter and Materials Physics, 54 (1996), pp. 7602– 7605, https://doi.org/10.1103/PhysRevB.54.7602.
- [39] E. VECHARYNSKI, C. YANG, AND J. E. PASK, A projected preconditioned conjugate gradient algorithm for computing many extreme eigenpairs of a Hermitian matrix, Journal of Computational Physics, 290 (2015), pp. 73–89, https://doi.org/10.1016/j.jcp.2015.02.030, https://linkinghub.elsevier.com/retrieve/pii/S002199911500100X (accessed 2025-03-28).
- [40] B. WANG, H. AN, H. XIE, AND Z. MO, A new subspace iteration algorithm for solving generalized eigenvalue problems in vibration analysis, Journal of Computational and Applied Mathematics, 468 (2025), p. 116622, https://doi.org/10.1016/j.cam.2025.116622, https://linkinghub.elsevier.com/retrieve/pii/S0377042725001372 (accessed 2025-03-28).
- [41] D. S. WATKINS, The Matrix Eigenvalue Problem, Society for Industrial and Applied Mathematics, Jan. 2007, https://doi.org/10.1137/1.9780898717808.
- [42] J. WINKELMANN, P. SPRINGER, AND E. D. NAPOLI, ChASE, ACM Transactions on Mathematical Software, 45 (2019), pp. 1–34, https://doi.org/10.1145/3313828.
- [43] Q. XU, A. SHARMA, B. COMER, H. HUANG, E. CHOW, A. J. MEDFORD, J. E. PASK, AND P. SURYANARAYANA, SPARC: Simulation Package for Ab-initio Real-space Calculations, SoftwareX, 15 (2021), p. 100709, https://doi.org/10.1016/j.softx.2021.100709.
- [44] B. ZHANG, X. JING, Q. XU, S. KUMAR, A. SHARMA, L. ERLANDSON, S. J. SAHOO, E. CHOW, A. J. MEDFORD, J. E. PASK, AND P. SURYANARAYANA, SPARC v2.0.0: Spin-orbit coupling, dispersion interactions, and advanced exchange-correlation functionals, Software Impacts, 20 (2024), p. 100649, https://doi.org/10.1016/j.simpa.2024.100649.
- [45] Y. ZHOU, J. R. CHELIKOWSKY, AND Y. SAAD, Chebyshev-filtered subspace iteration method free of sparse diagonalization for solving the Kohn-Sham equation, Journal of Computational Physics, 274 (2014), pp. 770–782, https://doi.org/https://doi.org/10.1016/j.jcp.2014.06. 056, https://www.sciencedirect.com/science/article/pii/S0021999114004744.
- [46] Y. ZHOU AND Y. SAAD, A Chebyshev-Davidson Algorithm for Large Symmetric Eigenproblems, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 954–971, https://doi. org/10.1137/050630404, http://epubs.siam.org/doi/10.1137/050630404 (accessed 2025-03-28).
- [47] Y. ZHOU, Y. SAAD, M. L. TIAGO, AND J. R. CHELIKOWSKY, Parallel self-consistentfield calculations via Chebyshev-filtered subspace acceleration, Physical Review E, 74 (2006), p. 066704, https://doi.org/10.1103/PhysRevE.74.066704, https://link.aps.org/doi/ 10.1103/PhysRevE.74.066704 (accessed 2023-02-24). Publisher: American Physical Society.
- [48] Y. ZHOU, Y. SAAD, M. L. TIAGO, AND J. R. CHELIKOWSKY, Self-consistent-field calculations using Chebyshev-filtered subspace iteration, Journal of Computational Physics, 219 (2006), pp. 172–184, https://doi.org/10.1016/J.JCP.2006.03.017. Publisher: Academic Press.
- [49] P. ZHU AND A. KNYAZEV, Angles between subspaces and their tangents, Journal of Numerical Mathematics, 21 (2013), https://doi.org/10.1515/jnum-2013-0013.

26