Check for updates

Received: 20 July 2021 | Revised: 27 November 2021 | Accepted: 11 February 2022

DOI: 10.1002/nla.2437

RESEARCH ARTICLE

WILEY

Inexact rational Krylov subspace method for eigenvalue problems

Shengjie Xu | **Fei Xue**

School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, USA

Correspondence
Fei Xue, School of Mathematical and Statistical Sciences, Clemson University, O-110 Martin Hall, Box 340975, Clemson, SC 29634, USA.
Email: fxue@clemson.edu

Funding information
U.S. National Science Foundation, Grant/Award Numbers: DMS-1819097, DMS-2111496

Abstract

An inexact rational Krylov subspace method is studied to solve large-scale nonsymmetric eigenvalue problems. Each iteration (outer step) of the rational Krylov subspace method requires solution to a shifted linear system to enlarge the subspace, performed by an iterative linear solver for large-scale problems. Errors are introduced at each outer step if these linear systems are solved approximately by iterative methods (inner step), and they accumulate in the rational Krylov subspace. In this article, we derive an upper bound on the errors introduced at each outer step to maintain the same convergence as exact rational Krylov subspace method for approximating an invariant subspace. Since this bound is inversely proportional to the current eigenresidual norm of the target invariant subspace, the tolerance of iterative linear solves at each outer step can be relaxed with the outer iteration progress. A restarted variant of the inexact rational Krylov subspace method is also proposed. Numerical experiments show the effectiveness of relaxing the inner tolerance to save computational cost.

KEYWORDS

eigenvalue and eigenvector, generalized Schur decomposition, inexact method, rational Krylov subspace

1 | INTRODUCTION

Let $A \in \mathbb{R}^{n\times n}$ be a large, sparse, and nonsymmetric matrix. In this article, we consider computing a real partial Schur form

$$AV = V\Theta, \tag{1}$$

where $V \in \mathbb{R}^{n\times p}$ has orthonormal columns, and $\Theta \in \mathbb{R}^{p\times p}$ is quasi upper triangular with 1×1 or 2×2 diagonal blocks that contain the p desired eigenvalues of A.

Variants of standard Krylov subspace methods have been widely used for eigenvalue computation (see, e.g., References 1-6), and they are most efficient for approximating dominant eigenvalues or exterior eigenvalues that are not significantly smaller in modulus than the dominant ones. For eigenvalues in other locations, the rational Krylov subspace method (RKSM), which was first proposed by Ruhe,[7] can be more effective. RKSM has been extensively investigated in recent

Funding information: U.S. National Science Foundation under Grants DMS-1819097 and DMS-2111496.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Numerical Linear Algebra with Applications* published by John Wiley & Sons Ltd.

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

years for approximating solutions to matrix equations,[8-10] actions of functions of matrices,[11,12] and algebraic nonlinear eigenvalue problems.[13-15]

Standard Krylov subspace methods construct subspaces of the form

$$\mathcal{K}_m(A, v_0) = \text{span}\left\{v_0, Av_0, A^2v_0, \ldots, A^{m-1}v_0\right\},$$

while RKSM generates subspaces

$$\mathcal{Q}_m(A, v_0) = q_{m-1}(A)^{-1}\mathcal{K}_m(A, v_0),$$

where $q_{m-1}(A)$ is a polynomial of degree $m-1$ with respect to matrix A. Expansion of such a subspace at each step needs the computation of a shift-invert matrix-vector product of the form $(\gamma - \eta A)^{-1}(\alpha I - \beta A)v$, which is equivalent to the solution of the linear system $(\gamma - \eta A)x = (\alpha I - \beta A)v$. For large-scale problems, especially those arising from discretizations of PDEs in 3D domains, iterative methods are recommended to solve the linear systems. Errors are introduced at each outer step from approximate linear solves, and they accumulate in the rational Krylov subspace.

Our goal is to relax the accuracy of the shift-invert matrix-vector product (linear solves) at each rational Krylov step, without negatively impacting the convergence of RKSM toward the desired invariant subspace. This motivation is the same as that for the investigation of inexact standard Krylov methods for eigenvalue problems,[16] but the need for relaxing the accuracy of operator-vector products is more natural and consequential in the setting of RKSM for reducing the computational cost. Similar research for inexact shift-invert Arnoldi's method can be seen in Reference 17, which uses one fixed pole at each step, whereas our RKSM uses variable poles. Also in Reference 18, the authors investigated the influence on the eigenresiduals of RKSM with a fixed uniform tolerance for errors allowed for inner linear solves. In this article, we have found that the errors allowed for solving the shifted linear system at each rational Krylov step is inversely proportional to the current eigenresidual norm of the desired invariant subspace. Therefore, the tolerance of iterative linear solves can be relaxed with the outer iteration progress. More computational cost can be saved at the later stage of the algorithm, when we are approaching convergence and having a smaller eigenresidual norm. Similar result of inexact RKSM for solving Lyapunov matrix equations can be found in Reference 19.

The rest of the article is organized as follows. In Section 2, we review the RKSM for solving eigenvalue problems, and derive the inexact Arnoldi relation and residual expressions. In Section 3, we review the perturbation theorem of invariant subspaces of matrix pairs and derive a theoretical bound on the norm of the allowable error introduced at each RKSM step, which guarantees the difference of eigenresiduals between the inexact and the exact method is below a given tolerance. In Section 4, we introduce a restarted RKSM based on Schur decomposition, and derive similar bounds on the errors allowed. In Section 5, we provide numerical results to show that the norm of the errors can be allowed to grow without affecting the convergence of the algorithm. We also compare inexact and exact RKSM to show the advantage of the inexact method. Conclusions of this article are presented in Section 6.

2 | PRELIMINARIES ABOUT RKSM FOR EIGENVALUE PROBLEMS

2.1 | Exact RKSM

To review the framework of RKSM, we begin with a starting vector v_1 with $\|v_1\|_2 = 1$. At step k, we choose parameters α_k, β_k, γ_k and η_k such that $|\alpha_k|^2 + |\beta_k|^2 \neq 0$, $|\gamma_k|^2 + |\eta_k|^2 \neq 0$, and $(\alpha_k, \beta_k) \neq (\gamma_k, \eta_k)$ up to a constant; then we compute the shift-invert matrix vector product $w = (\gamma_k I - \eta_k A)^{-1}(\alpha_k I - \beta_k A)v_k$, orthogonalize w against $v_1, v_2, \ldots, v_k$ and normalize into v_{k+1}. This can be described by Equation (2):

$$(\gamma_k I - \eta_k A)^{-1}(\alpha_k I - \beta_k A)v_k = \sum_{i=1}^{k+1} h_{ik}v_i. \tag{2}$$

Repeat the above relation for each index value $k = 1, 2, \ldots, m$. Assuming that there is no breakdown, we can get the Arnoldi relation for RKSM:

$$AV_{m+1}\underline{F}_m = V_{m+1}\underline{K}_m, \tag{3}$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $V_{m+1} = [v_1, v_2, \ldots, v_{m+1}]$ contains orthonormal basis vectors of the rational Krylov subspace

$$\mathcal{Q}_{m+1}(A, v_1) = q_m(A)^{-1}\mathcal{K}_{m+1}(A, v_1) = \left(\prod_{k=1}^{m}(\gamma_k I - \eta_k A)^{-1}\right) \text{span}\left\{v_1, Av_1, A^2v_1, \ldots, A^m v_1\right\},$$

and $\underline{H}_m$, $\underline{F}_m$, and $\underline{K}_m \in \mathbb{R}^{(m+1)\times m}$ are all upper Hessenberg matrices as follows:

$$\underline{H}_m = \begin{bmatrix} H_m \\ h_{m+1,m}e_m^* \end{bmatrix}, \quad \underline{F}_m = \begin{bmatrix} F_m \\ f_{m+1,m}e_m^* \end{bmatrix} = \begin{bmatrix} H_m\text{diag}(\eta_1, \ldots, \eta_m) - \text{diag}(\beta_1, ..., \beta_m) \\ h_{m+1,m}\eta_m e_m^* \end{bmatrix},$$
$$\underline{K}_m = \begin{bmatrix} K_m \\ g_{m+1,m}e_m^* \end{bmatrix} = \begin{bmatrix} H_m\text{diag}(\gamma_1, \ldots, \gamma_m) - \text{diag}(\alpha_1, ..., \alpha_m) \\ h_{m+1,m}\gamma_m e_m^* \end{bmatrix}. \tag{4}$$

To simplify the parameter configuration, a convenient approach is to set $\alpha_k = 1$, $\beta_k = 0$, $\gamma_k = -s_k$, and $\eta_k = -1$ where s_k is a pole of the rational Krylov subspace $\mathcal{Q}_{m+1}(A, v_1)$, that is, a zero of $q_m(t) = \prod_{k=1}^{m}(\gamma_k - \eta_k t)$. With such a choice, the Arnoldi relation represented in (3) can be written as $AV_{m+1}\underline{H}_m = V_{m+1}\underline{G}_m$, or:

$$AV_m H_m + h_{m+1,m}Av_{m+1}e_m^* = V_m G_m + s_m h_{m+1,m}v_{m+1}e_m^*, \tag{5}$$

where $G_m = H_m D_m + I_m$, $D_m = \text{diag}(s_1, ..., s_m)$, and $\underline{G}_m = \begin{bmatrix} G_m \\ s_m h_{m+1,m}e_m^* \end{bmatrix}$.

Approximate eigenvalues and eigenvectors of the matrix A can be obtained from the eigenpairs of the matrix pair (G_m, H_m):

$$G_m y_i = \lambda_i H_m y_i.$$

We call $\left(\lambda_i, V_{m+1}\underline{H}_m y_i\right)$ a Ritz pair of matrix A with respect to the subspace col (V_m); see, for example, References 14,20-22. Another definition of the Ritz pair for RKSM is given in Reference 6. For inexact rational Arnoldi methods, some references prefer to use the explicit projection $A_m = V_m^* A V_m$, instead of the derived projection matrix pair (G_m, H_m); see, for example, References 19,23. Our experiments suggest that there is no obvious advantage in the convergence rate by using the explicit projection in eigenvalue computation by RKSM. Therefore, we use derived projection in both our derivations and numerical tests.

Assume that we want to find a specific set of p $(p < m)$ eigenpairs of matrix A. Suppose that the corresponding generalized Schur decomposition[24] of (G_m, H_m) is (see, e.g., References 24,25):

$$G_m = Z_m S_m U_m^*, \; H_m = Z_m T_m U_m^*, \tag{6}$$

where $U_m, Z_m \in \mathbb{R}^{m\times m}$ are unitary matrices, and (S_m, T_m) is a pair of (quasi) upper triangular matrices of order m. We partition the above matrices into blocks:

$$U_m = \begin{bmatrix} U_m^1 & U_m^2 \end{bmatrix}, \; Z_m = \begin{bmatrix} Z_m^1 & Z_m^2 \end{bmatrix}, S_m = \begin{bmatrix} S_m^{11} & S_m^{12} \\ 0 & S_m^{22} \end{bmatrix}, \text{ and } T_m = \begin{bmatrix} T_m^{11} & T_m^{12} \\ 0 & T_m^{22} \end{bmatrix}, \tag{7}$$

where $S_m^{11}, T_m^{11} \in \mathbb{R}^{p\times p}$, $S_m^{22}, T_m^{22} \in \mathbb{R}^{(m-p)\times(m-p)}$, $U_m^1, Z_m^1 \in \mathbb{R}^{m\times p}$, $U_m^2, Z_m^2 \in \mathbb{R}^{m\times(m-p)}$, and the 1×1 or 2×2 diagonal blocks of (S_m^{11}, T_m^{11}) and (S_m^{22}, T_m^{22}) define the wanted and unwanted Ritz values, respectively. The partial generalized Schur form of (G_m, H_m) of order p is then given by:

$$G_m U_m^1 = H_m U_m^1 \Theta_m^{11}, \tag{8}$$

where $\Theta_m^{11} = \left(T_m^{11}\right)^{-1} S_m^{11} \in \mathbb{R}^{p\times p}$ is (quasi) upper triangular, and $U_m^1 \in \mathbb{R}^{m\times p}$ has orthonormal columns that are basis vectors of the invariant subspace of the matrix pair (G_m, H_m) corresponding to our desired spectrum of A.

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We are mostly interested in the eigenresidual associated with an approximate partial Schur form of A. Based on the Arnoldi relation in (3) and partial Schur form in (8), we have

$$\begin{aligned} R_m &= AV_{m+1}\underline{H}_m U_m^1 - V_{m+1}\underline{H}_m U_m^1 \Theta_m^{11} = V_{m+1}\underline{G}_m U_m^1 - V_{m+1}\underline{H}_m U_m^1 \Theta_m^{11} \\ &= V_m G_m U_m^1 + s_m h_{m+1,m} v_{m+1} e_m^* U_m^1 - V_m H_m U_m^1 \Theta_m^{11} - h_{m+1,m} v_{m+1} e_m^* U_m^1 \Theta_m^{11} \\ &= h_{m+1,m} v_{m+1} e_m^* U_m^1 \left(s_m I - \Theta_m^{11}\right), \end{aligned} \tag{9}$$

and

$$\|R_m\|_2 = |h_{m+1,m}| \, \|v_{m+1}\|_2 \left\| e_m^* U_m^1 \left(s_m I - \Theta_m^{11}\right) \right\|_2 = |h_{m+1,m}| \left\| e_m^* U_m^1 \left(s_m I - \Theta_m^{11}\right) \right\|_2 . \tag{10}$$

This means that the residual norm associated with the wanted invariant subspace approximation can be obtained easily from Θ_m^{11}, the last row of U_m^1, s_m, and $h_{m+1,m}$. Our primary interest in this article is about the conditions under which this observation still holds approximately for the inexact method. The process of RKSM for eigenvalue computation is shown in Algorithm 1.

Algorithm 1. RKSM to solve eigenvalue problems

Input: $A \in \mathbb{R}^{n\times n}$, $v_1 \in \mathbb{R}^n$ and $\|v_1\| = 1$, max iteration step m, tolerance $tol > 0$.
Output: desired p eigenvalues and the corresponding invariant subspace.

for $k = 1, 2, ..., m$ **do**
  Choose the pole s_k.
  Let $w_{k+1} = (A - s_k I)^{-1} v_k$, orthogonalize against $v_1, v_2, ..., v_k$, and normalize into v_{k+1}.
  Compute the generalized Schur decomposition of matrix pair (G_k, H_k) in (6).
  if residual $\|R_k\|_2 \le tol$, **then**
    Return the diagonal entries or the eigenvalues of the 2×2 diagonal blocks of $\left(T_k^{11}\right)^{-1} S_k^{11}$ in (7) as approximations to the desired eigenvalues of A, and $V_{k+1}\underline{H}_k U_k^1$ as approximation to the desired invariant subspace.
  end if
end for

2.2 | Inexact RKSM

For large-scale problems, the shift-invert matrix vector product $w = (A - s_k I)^{-1} v_k$ cannot be easily computed to high precision. In practice, as the linear system $(A - s_k I)w = v_k$ is solved approximately by an iterative method, errors are introduced into the solution w and hence into the basis vectors of the rational Krylov subspaces. Let the residual of this linear solve be $\xi_k = v_k - (A - s_k I)w_{k+1}$. Then (2) turns into:

$$w_{k+1} = (A - s_k I)^{-1} (v_k - \xi_k) = \sum_{i=1}^{k+1} h_{ik} v_i. \tag{11}$$

From (11), the inexact Arnoldi relation of RKSM is given by:

$$AV_{m+1}\underline{H}_m + \Xi_m = V_{m+1}\underline{G}_m, \tag{12}$$

where $\Xi_m = [\xi_1, \xi_2, \ \dots \ , \xi_m]$, and H_m, G_m have the same forms as exact RKSM in (5).

The eigenresidual of the inexact method is defined as:

$$\begin{aligned} \tilde{R}_m &= AV_{m+1}\underline{H}_m U_m^1 - V_{m+1}\underline{H}_m U_m^1 \Theta_m^{11} = V_{m+1}\underline{G}_m U_m^1 - \Xi_m U_m^1 - V_{m+1}\underline{H}_m U_m^1 \Theta_m^{11} \\ &= h_{m+1,m} v_{m+1} e_m^* U_m^1 \left(s_m I - \Theta_m^{11}\right) - \Xi_m U_m^1. \end{aligned} \tag{13}$$

The residual $\tilde{R}_m$ in (13) is the *true residual* of inexact RKSM, and R_m in (9) is the *derived residual*, which has the same expression as the true residual of exact RKSM and can be computed very conveniently. However, the *derived residual* R_m for the exact and the inexact method are not equal, since they have different values of entries in $\underline{H}_m$. Consequently, there

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

is no theoretical guarantee that the eigenresiduals of the inexact and the exact method are sufficiently close at any RKSM step, though they are usually close in practice. We are interested in exploring the difference between these two residuals for inexact RKSM to see how to keep it sufficiently small, so that we can disregard the impact of the error term Ξ_m. The difference between the two residuals is:

$$\Delta_m = R_m - \tilde{R}_m = \Xi_m U_m^1. \tag{14}$$

We will explore certain restrictions on Ξ_k $(1 \le k \le m)$ so that Δ_k is always below a user-specified small tolerance for us to see similar convergence behavior between the exact and the inexact method.

3 | TOLERANCE RELAXATION STRATEGY FOR INEXACT RATIONAL KRYLOV

To realize the relaxed accuracy of operator-vector product at later steps of RKSM, what we need is a $\Delta_k = \Xi_k U_m^1$ $(k \le m)$ sufficiently small in norm, where k denotes the current step of RKSM and m denotes the maximum steps of RKSM. To achieve this, it would be sufficient to have either the jth column of Ξ_k $(1 \le j \le k)$ small in norm, or the jth entry of each column of U_m^1 small in absolute value. The main observation to support inexact rational Krylov is that, as RKSM approaches convergence to the desired invariant subspace, the last $k - p$ entries in each column of U_m^1 typically decrease to zero in modulus from top to bottom. As a result, the jth column of Ξ_k $(p < j \le k)$ can be inversely proportional in norm to the entries in the jth row of U_m^1, which in turn are proportional to the eigenresidual norm of the desired invariant subspace approximation at step j. This idea is similar to that explored in Reference 16, but there are more complex technical details to handle in our problem setting.

3.1 | Perturbation theorem for regular pairs

In order to investigate the difference between the true residual and the derived residual, we first introduce the approximation theorem for regular pairs.

Definition 1. For square matrices $A, B \in \mathbb{R}^{n\times n}$, (A, B) is called a regular pair if there exists $\lambda \in \mathbb{C}$ such that $\det(\lambda A - B) \neq 0$. Define the norm $\|\cdot\|_F$ on the space of matrix pairs (P, Q), where $P, Q \in \mathbb{R}^{p\times q}$, as:

$$\|(P, Q)\|_F = \max\left\{\|P\|_F, \|Q\|_F\right\}. \tag{15}$$

Based on the $\|\cdot\|_F$ norm, the difference between regular pairs (A_1, B_1) and (A_2, B_2), where $A_1, B_1 \in \mathbb{R}^{q\times q}$ and $A_2, B_2 \in \mathbb{R}^{p\times p}$, is defined as:

$$\mathrm{dif}\left[(A_1, B_1), (A_2, B_2)\right] = \inf_{\|(P,Q)\|_F=1} \|(QA_1 + A_2P, QB_1 + B_2P)\|_F. \tag{16}$$

Note that $\mathrm{dif}\left[(A_1, B_1), (A_2, B_2)\right] > 0$ if and only if the spectra of (A_1, B_1) and (A_2, B_2) are disjoint.

With the above definition, we can introduce the approximation theorem[26,27] for the regular pair.

Theorem 1 (26 (theorem 2.13)). *Let $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n\times n}$ be nonsingular, partitioned as $\mathcal{X} = [U\ X]$, $\mathcal{Y} = [Z\ W]$, where $U, Z \in \mathbb{R}^{n\times p}$. For a regular pair (A, B), set*

$$\mathcal{Y}^* A\mathcal{X} = \begin{bmatrix} Z^* \\ W^* \end{bmatrix} A \begin{bmatrix} U\ X \end{bmatrix} = \begin{bmatrix} A_1\ H_A \\ G_A\ A_2 \end{bmatrix}, \quad \mathcal{Y}^* B\mathcal{X} = \begin{bmatrix} Z^* \\ W^* \end{bmatrix} B \begin{bmatrix} U\ X \end{bmatrix} = \begin{bmatrix} B_1\ H_B \\ G_B\ B_2 \end{bmatrix}. \tag{17}$$

Define $\gamma = \|(G_A, G_B)\|_F$, $\eta = \|(H_A, H_B)\|_F$, and $\delta = \mathrm{dif}\left[(A_1, B_1), (A_2, B_2)\right]$. Assume that the spectra of (A_1, B_1) and (A_2, B_2) are disjoint ($\delta > 0$). Then if

$$\frac{\eta\gamma}{\delta^2} < \frac{1}{4}, \tag{18}$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

there is a unique (P, Q) *satisfying:*

$$\|(P,Q)\|_F \le \frac{2\gamma}{\delta + \sqrt{\delta^2 - 4\gamma\eta}} < 2\frac{\gamma}{\delta}, \tag{19}$$

such that the column space of $\widehat{U} = U + XP$ *is an invariant subspace of* (A, B) *corresponding to the regular pairs* $(A_1 + H_A P, B_1 + H_B P)$.

Remark 1. If $\mathcal{X}$ defined in Theorem 1 is unitary, we can get a new unitary matrix

$$\hat{\mathcal{X}} = \left[\widehat{U}\ \widehat{X}\right] = \left[(U + XP)(I + P^*P)^{-\frac{1}{2}}\ \ (X - UP^*)(I + P^*P)^{-\frac{1}{2}}\right], \tag{20}$$

where the column space of $\widehat{U}$ is an invariant subspaces of (A, B). It's easy to directly verify that $\hat{\mathcal{X}}$ is unitary.

3.2 | Approximation theorem for eigenpairs computation

As explained earlier in Section 3, the fundamental observation backing the theory of inexact RKSM is that the eigenvectors of the projected matrix pair (G_m, H_m) corresponding to the desired eigenvalues (called "primitive Ritz vectors" in Reference 28) tend to have a decreasing pattern in their trailing entries (those at the bottom) as the method proceeds toward convergence. To establish such a pattern rigorously, we need to study the trailing entries of these eigenvectors at different steps of RKSM.

We can now use the result of Theorem 1 about approximate eigenpairs. Let $U_k^1, Z_k^1 \in \mathbb{R}^{k\times p}$ $(k > p)$ contain p Schur vectors of (G_k, H_k), such that $G_k U_k^1 = Z_k^1 S_k^{11}$, $H_k U_k^1 = Z_k^1 T_k^{11}$, where $S_k^{11}, T_k^{11} \in \mathbb{R}^{p\times p}$ are both order-p (quasi) upper triangular. We extend matrices U_k^1 and Z_k^1 into unitary matrices $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{m\times m}$ as follows:

$$\mathcal{X} = \left[\begin{pmatrix} U_k^1 \\ 0 \end{pmatrix}, X\right] = \begin{bmatrix} U_k^1 & X_1 \\ 0 & X_2 \end{bmatrix},\ \mathcal{Y} = \left[\begin{pmatrix} Z_k^1 \\ 0 \end{pmatrix}, W\right] = \begin{bmatrix} Z_k^1 & W_1 \\ 0 & W_2 \end{bmatrix}. \tag{21}$$

We partition G_m and H_m into 2×2 blocks:

$$G_m = \begin{bmatrix} G_k & G_a \\ s_k h_{k+1,k} e_1 e_k^* & G_b \end{bmatrix},\ H_m = \begin{bmatrix} H_k & H_a \\ h_{k+1,k} e_1 e_k^* & H_b \end{bmatrix}.$$

Then, we left multiply $\mathcal{Y}^*$ and right multiply $\mathcal{X}$ to G_m and H_m, respectively:

$$\begin{aligned} \mathcal{Y}^* G_m \mathcal{X} &= \begin{bmatrix} (Z_k^1)^* & 0 \\ W_1^* & W_2^* \end{bmatrix} \begin{bmatrix} G_k & G_a \\ s_k h_{k+1,k} e_1 e_k^* & G_b \end{bmatrix} \begin{bmatrix} U_k^1 & X_1 \\ 0 & X_2 \end{bmatrix} \\ &= \begin{bmatrix} (Z_k^1)^* G_k U_k^1 & (Z_k^1)^* G_k X_1 + (Z_k^1)^* G_a X_2 \\ W_1^* G_k U_k^1 + s_k h_{k+1,k} W_2^* e_1 e_k^* U_k^1 & W^* G_m X \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \end{aligned} \tag{22}$$

$$\begin{aligned} \mathcal{Y}^* H_m \mathcal{X} &= \begin{bmatrix} (Z_k^1)^* & 0 \\ W_1^* & W_2^* \end{bmatrix} \begin{bmatrix} H_k & H_a \\ h_{k+1,k} e_1 e_k^* & H_b \end{bmatrix} \begin{bmatrix} U_k^1 & X_1 \\ 0 & X_2 \end{bmatrix} \\ &= \begin{bmatrix} (Z_k^1)^* H_k U_k^1 & (Z_k^1)^* H_k X_1 + (Z_k^1)^* H_a X_2 \\ W_1^* H_k U_k^1 + h_{k+1,k} W_2^* e_1 e_k^* U_k^1 & W^* H_m X \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}. \end{aligned} \tag{23}$$

Note that $W_1^* G_k U_k^1 = W_1^* Z_k^1 S_k^{11} = 0$ and $W_1^* H_k U_k^1 = W_1^* Z_k^1 T_k^{11} = 0$, thanks to the partial Schur relation $G_k U_k^1 = Z_k^1 S_k^{11}$, $H_k U_k^1 = Z_k^1 T_k^{11}$ and the structure of the unitary matrices in (21). Also, W_2 has orthonormal rows, that is, W_2^* has orthonormal columns by (21), such that $W_2^* e_1$ is a unit vector in 2-norm. Besides, the Frobenius norm of a rank-1 matrix uw^* is simply $\|u\|_2\|w\|_2$. It follows that

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\|A_{21}\|_F = \left\|s_k h_{k+1,k} W_2^* e_1 e_k^* U_k^1\right\|_F = |s_k h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2, \tag{24}$$

$$\|B_{21}\|_F = \left\|h_{k+1,k} W_2^* e_1 e_k^* U_k^1\right\|_F = |h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2. \tag{25}$$

We now define γ for the partitioned matrices in (22) and (23) corresponding to that defined in Theorem 1. With the relation (10), we have

$$\begin{aligned}\gamma &= \|(A_{21}, B_{21})\|_F = \max\left\{|s_k h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2, |h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2\right\} \\ &= |h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2 \max\{1, |s_k|\} = \omega_k |h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2,\end{aligned} \tag{26}$$

where ω_k is defined as

$$\omega_k = \max\{1, |s_k|\}. \tag{27}$$

Our next step is to quantify η and δ for our partitioned matrices in (22) and (23). To this end, define $\Psi_1 = \left[\left(Z_k^1\right)^* 0\right] G_m - S_k^{11} \left[\left(U_k^1\right)^* 0\right] \in \mathbb{R}^{k\times m}$. Given the construction of $\mathcal{X}$ in (21), we have $\left[\left(U_k^1\right)^* 0\right] X = 0$. Then from (22), we have:

$$A_{12} = \left(Z_k^1\right)^* G_k X_1 + \left(Z_k^1\right)^* G_a X_2 = \left[\left(Z_k^1\right)^* 0\right] G_m X = \Psi_1 X. \tag{28}$$

Define $\Psi_2 = \left[\left(Z_k^1\right)^* 0\right] H_m - T_k^{11} \left[\left(U_k^1\right)^* 0\right] \in \mathbb{R}^{k\times m}$. Similarly, from (23):

$$B_{12} = \left(Z_k^1\right)^* H_k X_1 + \left(Z_k^1\right)^* H_a X_2 = \left[\left(Z_k^1\right)^* 0\right] H_m X = \Psi_2 X. \tag{29}$$

Note that $S_k^{11} = \left(Z_k^1\right)^* G_k U_k^1 = \left[\left(Z_k^1\right)^* 0\right] G_m \begin{bmatrix} U_k^1 \\ 0 \end{bmatrix}$. Therefore, Ψ_1 can be written as:

$$\Psi_1 = \left[\left(Z_k^1\right)^* 0\right] G_m - S_k^{11} \left[\left(U_k^1\right)^* 0\right] = \left[\left(Z_k^1\right)^* 0\right] G_m \left(I_m - \begin{bmatrix} U_k^1 \left(U_k^1\right)^* & 0 \\ 0 & 0 \end{bmatrix}\right).$$

Since $T_k^{11} = \left(Z_k^1\right)^* H_k U_k^1 = \left[\left(Z_k^1\right)^* 0\right] H_m \begin{bmatrix} U_k^1 \\ 0 \end{bmatrix}$, it follows that Ψ_2 can be written as:

$$\Psi_2 = \left[\left(Z_k^1\right)^* 0\right] H_m - T_k^{11} \left[\left(U_k^1\right)^* 0\right] = \left[\left(Z_k^1\right)^* 0\right] H_m \left(I_m - \begin{bmatrix} U_k^1 \left(U_k^1\right)^* & 0 \\ 0 & 0 \end{bmatrix}\right).$$

From (28) and (29), we can define η and derive an upper bound on this quantity for the partitioned matrices in (22) and (23), the same way as their counterpart described in Theorem 1:

$$\begin{aligned}\eta &= \|(A_{12}, B_{12})\|_F = \max\{\|\Psi_1 X\|_F, \|\Psi_2 X\|_F\} \\ &\le \max\left\{\left\|\Psi_1\left[\begin{pmatrix} U_k^1 \\ 0 \end{pmatrix}, X\right]\right\|_F, \left\|\Psi_2\left[\begin{pmatrix} U_k^1 \\ 0 \end{pmatrix}, X\right]\right\|_F\right\} = \max\{\|\Psi_1\|_F, \|\Psi_2\|_F\},\end{aligned} \tag{30}$$

where the first inequality holds following the definition of the Frobenius norm, and the last equality holds because $\left[\begin{pmatrix} U_k^1 \\ 0 \end{pmatrix}, X\right]$ is unitary, and the Frobenius norm is invariant under unitary transformations.

Finally, we define δ for our partitioned matrices as defined in Theorem 1. We note that A_{11} in (22) can be simplified, since $A_{11} = \left(Z_k^1\right)^* G_k U_k^1 = S_k^{11}$. Also, from (23), we have $B_{11} = \left(Z_k^1\right)^* H_k U_k^1 = T_k^{11}$. Therefore, we define δ as:

$$\delta_{m,k} = \mathrm{dif}\left[(A_{11}, B_{11}), (A_{22}, B_{22})\right] = \mathrm{dif}\left[\left(S_k^{11}, T_k^{11}\right), (W^* G_m X, W^* H_m X)\right]. \tag{31}$$

Proposition 1. *Consider an m-step RKSM, which generates the Arnoldi relation (3), for computing a specific set of p desired eigenvalues of matrix $A \in \mathbb{R}^{n\times n}$. At the kth step, where $p \le k \le m$, let the columns of $U_k^1 \in \mathbb{R}^{k\times p}$ contain an orthonormal*

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

basis for a simple invariant subspace of the matrix pair (G_k, H_k) *such that* $G_k U_k^1 = Z_k^1 S_k^{11}$, $H_k U_k^1 = Z_k^1 T_k^{11}$, *where* S_k^{11} *and* T_k^{11} *are both order-p (quasi) upper triangular and the column space of* $Z_k^1 \in \mathbb{R}^{k\times p}$ *is corresponding left invariant subspace. Define the simplified eigenresidual* $\mathcal{R}_k$ *as:*

$$\mathcal{R}_k = h_{k+1,k} v_{k+1} e_k^* U_k^1. \tag{32}$$

With the quantities γ, ω_k, η, *and* $\delta_{m,k}$ *defined in (26), (27), (30), and (31), respectively, if:*

$$\|\mathcal{R}_k\|_2 < \frac{\delta_{m,k}^2}{4\eta\omega_k}, \tag{33}$$

then there exists a matrix $\widehat{U} = \begin{bmatrix}\widehat{U}_1\\ \widehat{U}_2\end{bmatrix}$, $\widehat{U}^*\widehat{U} = I$ *with* $\widehat{U}_1 \in \mathbb{R}^{k\times p}$ *and* $\widehat{U}_2 \in \mathbb{R}^{(m-k)\times p}$, *such that the columns of* $\widehat{U}$ *span a simple invariant subspace of* (G_m, H_m) *with*

$$\left\|\widehat{U}_2\right\|_F \le \|P\|_F, \text{ where } \quad 0 \le \|P\|_F < 2\frac{\omega_k \|\mathcal{R}_k\|_2}{\delta_{m,k}}.$$

Proof. Based on the expression of $\mathcal{R}_m$ in (32), we have:

$$\|\mathcal{R}_k\|_2 = |h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2.$$

From (26), if (33) holds, then

$$\frac{\gamma\eta}{\delta_{m,k}^2} = \frac{\omega_k |h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2 \eta}{\delta_{m,k}^2} = \frac{\omega_k \|\mathcal{R}_k\|_2 \eta}{\delta_{m,k}^2} < \frac{1}{4}.$$

By Theorem 1, we conclude that there is a unique (P, Q) satisfying:

$$\|(P,Q)\|_F < 2\frac{\gamma}{\delta_{m,k}} = 2\frac{\omega_k |h_{k+1,k}| \left\|e_k^* U_k^1\right\|_2}{\delta_{m,k}} \le 2\frac{\omega_k \|\mathcal{R}_k\|_2}{\delta_{m,k}},$$

such that $\widehat{U} = \begin{bmatrix}\widehat{U}_1\\ \widehat{U}_2\end{bmatrix} = \left(\begin{bmatrix}U_k^1\\ 0\end{bmatrix} + \begin{bmatrix}X_1\\ X_2\end{bmatrix} P\right)(I + P^*P)^{-\frac{1}{2}}$ contains the p right Schur vectors of (G_m, H_m) corresponding to the desired spectrum. In addition,

$$\left\|\widehat{U}_2\right\|_2 = \left\|X_2 P (I + P^*P)^{-\frac{1}{2}}\right\|_2 \le \left\|P(I + P^*P)^{-\frac{1}{2}}\right\|_2 \le \|P\|_2 \le \|P\|_F \le \|(P,Q)\|_F.$$

Here, we used the fact $\|X_2\|_2 \le \|X\|_2 = 1$, and the inequality $\left\|(I + P^*P)^{-\frac{1}{2}}\right\|_2 \le 1$ which can be shown by the singular value decomposition of P without difficulty. ∎

Proposition 1 shows that if $\|\mathcal{R}_k\|_2$ satisfies (33), then for any i, $k+1 \le i \le m$,

$$\left\|e_i^* \widehat{U}\right\|_2 \le \left\|\widehat{U}_2\right\|_2 \le \|P\|_F < 2\frac{\omega_k \|\mathcal{R}_k\|_2}{\delta_{m,k}}. \tag{34}$$

The definition of the simplified eigenresidual $\mathcal{R}_m$ in (32) is slightly different from the derived residual R_m in (9) up to a matrix factor of $s_m I - \Theta_m^{11}$. We used $\mathcal{R}_m$ instead of R_m in Proposition 1 to simplify the discussion of the impact of the extra factor $s_m I - \Theta_m^{11}$, which may yield an excessively stringent relaxation estimate in practice for RKSM. Based on the definition of $\mathcal{R}_m$ in (32), it can be easily acquired once the mth RKSM step is completed. Both $\|\mathcal{R}_k\|_2$ and $\|R_k\|_2$ tend to zero with the outer iteration progress of RKSM, if the algorithm converges to the desired invariant subspace. Therefore, $\|\mathcal{R}_k\|_2$ can be a monitor for the convergence of RKSM.

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We have just established a critical foundation for inexact RKSM, namely, if $\frac{\omega_k}{\delta_{m,k}}$ is bounded from above, the trailing entries of the wanted eigenvectors of (G_k, H_k) indeed have a decreasing pattern as $\|\mathcal{R}_k\|_2 \to 0$ with step k, that is, as RKSM proceeds to convergence towards the desired invariant subspace of A.

3.3 | Error bounds for inexact RKSM solving eigenvalue problems

As shown in (12), if the linear operator at each step is not applied exactly, the Arnoldi relation from RKSM involves a matrix of errors accumulated at the accomplished steps. If the norm of Δ_m in (14) is sufficiently small, we can conclude that the difference between the true and the derived residuals remains small.

The following theorem shows that if the norm of the error introduced at each step of RKSM is properly bounded, inexact RKSM can deliver the true eigenresiduals of the desired invariant subspace that are close to the derived eigenresiduals.

Theorem 2. *Assume that inexact RKSM is used for computing p eigenpairs of the matrix A, and m (m > p) steps are taken such that the Arnoldi relation (12) holds. With the quantities R_k, ξ_k, $\tilde{R}_k$, ω_k, η, $\delta_{m,k}$, and $\mathcal{R}_m$ defined in (9), (11), (13), (27), (30), (31), and (32), respectively, given any $\epsilon > 0$, assume that for each step k ($1 \le k \le m$),*

$$\|\xi_k\|_2 = \begin{cases} \frac{\delta_{m,k-1}\epsilon}{2m\omega_{k-1}\|\mathcal{R}_{k-1}\|_2}, & \text{if } k > p \text{ and } \|\mathcal{R}_{k-1}\|_2 < \frac{\delta^2_{m,k-1}}{4\eta\omega_{k-1}}, \\ \frac{\epsilon}{m}, & \text{otherwise.} \end{cases} \tag{35}$$

Then

$$\|\Delta_m\|_2 = \left\|R_m - \tilde{R}_m\right\|_2 \le \epsilon.$$

Proof. When $k > p$, the simplified eigenresidual at the end of step $k-1$ of RKSM can be written as $\|\mathcal{R}_{k-1}\|_2 = |h_{k,k-1}| \left\|e^*_{k-1}U^1_{k-1}\right\|_2$, where $U^1_{k-1} \in \mathbb{R}^{(k-1)\times p}$ contains the desired Schur vectors of (G_{k-1}, H_{k-1}). Assume that Ω_1 is the subset of $\{1, 2, \dots, m\}$ such that for each $k \in \Omega_1$, $k > p$ and $\|\mathcal{R}_{k-1}\|_2 < \frac{\delta^2_{m,k-1}}{4\eta w_{k-1}}$, and Ω_2 is the complement of Ω_1. With the bound on ξ_k given in (35) and the conclusion (34) as a result of Proposition 1, we have from (14):

$$\begin{aligned} \|\Delta_m\|_2 = \left\|\Xi_m U^1_m\right\|_2 = \left\|\sum_{k=1}^{m} \xi_k e^*_k U^1_m\right\|_2 &\le \sum_{k\in\Omega_1} \|\xi_k\|_2 \left\|e^*_k U^1_m\right\|_2 + \sum_{k\in\Omega_2} \|\xi_k\|_2 \left\|e^*_k U^1_m\right\|_2 \\ &\le \sum_{k\in\Omega_1} \frac{\delta_{m,k-1}\epsilon}{2m\omega_{k-1}\|\mathcal{R}_{k-1}\|_2} \left\|e^*_k U^1_m\right\|_2 + \sum_{k\in\Omega_2} \frac{\epsilon}{m} \left\|e^*_k U^1_m\right\|_2 \\ &\le \sum_{k\in\Omega_1} \frac{\delta_{m,k-1}\epsilon}{2m\omega_{k-1}\|\mathcal{R}_{k-1}\|_2} 2\frac{\omega_{k-1}\|\mathcal{R}_{k-1}\|_2}{\delta_{m,k-1}} + \sum_{k\in\Omega_2} \frac{\epsilon}{m} = \frac{\epsilon|\Omega_1|}{m} + \frac{\epsilon|\Omega_2|}{m} = \epsilon. \end{aligned}$$

The proof is established. ∎

Note that $\delta_{m,k-1}$ depends on G_m and H_m, which are not available yet at step k. Therefore, Theorem 2 in its original form is mostly of theoretical interest. In practice, we need to find reasonable approximations to $\delta_{m,k-1}$ that are available at each step k to effectively run inexact RKSM.

3.4 | Evaluation of the difference between two regular pairs

It is impossible to know the exact value of $\delta_{m,k-1}$ at the kth step from its definition in (31) before finishing all m steps. Upon completing the $(k-1)$th step, we already get the matrix pair (G_{k-1}, H_{k-1}), and we suggest to use it to approximate $\delta_{m,k-1}$. Let $\mathcal{X}_{k-1}, \mathcal{Y}_{k-1} \in \mathbb{R}^{(k-1)\times(k-1)}$ be the Schur vectors of (G_{k-1}, H_{k-1}), such that

$$\mathcal{Y}^*_{k-1} G_{k-1} \mathcal{X}_{k-1} = \begin{bmatrix} \tilde{G}^{11}_{k-1} & \times \\ 0 & \tilde{G}^{22}_{k-1} \end{bmatrix}, \ \mathcal{Y}^*_{k-1} H_{k-1} \mathcal{X}_{k-1} = \begin{bmatrix} \tilde{H}^{11}_{k-1} & \times \\ 0 & \tilde{H}^{22}_{k-1} \end{bmatrix},$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $\tilde{G}_{k-1}^{11}, \tilde{H}_{k-1}^{11} \in \mathbb{R}^{p\times p}$ are upper triangular matrices whose diagonal entries define the p desired Ritz values. Then $\delta_{m,k-1}$ in (31) can be approximated by

$$\tilde{\delta}_{m,k-1} = \operatorname{dif}\left[\left(\tilde{G}_{k-1}^{11}, \tilde{H}_{k-1}^{11}\right), \left(\tilde{G}_{k-1}^{22}, \tilde{H}_{k-1}^{22}\right)\right]. \tag{36}$$

To approximate $\tilde{\delta}_{m,k-1}$, the following lemma gives relatively tight lower and upper bounds on $\tilde{\delta}_{m,k-1}$, such that the upper and the lower bounds differ only by a factor of 2.

Lemma 1. *For two regular pairs* (A_1, B_1) *and* (A_2, B_2)*, where* A_1, $B_1 \in \mathbb{R}^{m\times m}$*, and* A_2, $B_2 \in \mathbb{R}^{n\times n}$*, and* $\delta = \operatorname{dif}[(A_1, B_1), (A_2, B_2)]$*, we have:*

$$\frac{1}{\sqrt{2}}\sigma_{\min}\left(\begin{bmatrix} A_1^*\otimes I_n & I_m\otimes A_2 \\ B_1^*\otimes I_n & I_m\otimes B_2\end{bmatrix}\right) \le \delta \le \sqrt{2}\sigma_{\min}\left(\begin{bmatrix} A_1^*\otimes I_n & I_m\otimes A_2 \\ B_1^*\otimes I_n & I_m\otimes B_2\end{bmatrix}\right),$$

where $\sigma_{\min}$ *refers to the smallest singular value of the matrix involved.*

Proof. We begin with the first matrix of the pair in the definition of dif (16):

$$\begin{aligned}\|QA_1 + A_2P\|_F &= \|\operatorname{vec}(QA_1 + A_2P)\|_2 = \left\|\left(A_1^*\otimes I_n\right)\operatorname{vec}(Q) + (I_m\otimes A_2)\operatorname{vec}(P)\right\|_2 \\ &= \left\|\begin{bmatrix} A_1^*\otimes I_n & I_m\otimes A_2\end{bmatrix}\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}\right\|_2.\end{aligned}$$

Similar equation holds for B_1 and B_2. Also, for any vectors v_1, v_2:

$$\max\{\|v_1\|_2, \|v_2\|_2\} \ge \frac{1}{\sqrt{2}}\sqrt{\|v_1\|_2^2 + \|v_2\|_2^2} = \frac{1}{\sqrt{2}}\left\|\begin{bmatrix}v_1\\ v_2\end{bmatrix}\right\|_2.$$

Based on this inequality,

$$\begin{aligned}\max\{\|QA_1 + A_2P\|_F, \|QB_1 + B_2P\|_F\} &= \max\left\{\left\|\begin{bmatrix} A_1^*\otimes I_n & I_m\otimes A_2\end{bmatrix}\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}\right\|_2, \left\|\begin{bmatrix} B_1^*\otimes I_n & I_m\otimes B_2\end{bmatrix}\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}\right\|_2\right\} \\ &\ge \frac{1}{\sqrt{2}}\left\|\begin{bmatrix} A_1^*\otimes I_n & I_m\otimes A_2\\ B_1^*\otimes I_n & I_m\otimes B_2\end{bmatrix}\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}\right\|_2.\end{aligned}$$

Define $\mathcal{M} = \begin{bmatrix} A_1^*\otimes I_n & I_m\otimes A_2\\ B_1^*\otimes I_n & I_m\otimes B_2\end{bmatrix}$, and then:

$$\begin{aligned}\delta &= \inf_{\|(P,Q)\|_F=1}\|(QA_1 + A_2P, QB_1 + B_2P)\|_F \ge \frac{1}{\sqrt{2}}\inf_{\|(P,Q)\|_F=1}\left\|\mathcal{M}\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}\right\|_2 \\ &= \frac{1}{\sqrt{2}}\sigma_{\min}(\mathcal{M})\left\|\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}\right\|_2 \ge \frac{1}{\sqrt{2}}\sigma_{\min}(\mathcal{M}),\end{aligned}$$

where the particular $\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}$ in the second equality is the right singular vector of $\mathcal{M}$ up to a scaling factor, corresponding to $\sigma_{\min}(\mathcal{M})$. The last inequality holds because $\left\|\begin{bmatrix}\operatorname{vec}(Q)\\ \operatorname{vec}(P)\end{bmatrix}\right\|_2 \ge 1$ for $\|(P, Q)\|_F = 1$.

On the other hand, for any vectors v_1, v_2:

$$\max\{\|v_1\|_2, \|v_2\|_2\} \le \sqrt{\|v_1\|_2^2 + \|v_2\|_2^2} = \left\|\begin{bmatrix}v_1\\ v_2\end{bmatrix}\right\|_2.$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Similarly, based on this inequality,

$$\max\left\{\|QA_1+A_2P\|_F,\|QB_1+B_2P\|_F\right\}\le\left\|\mathcal{M}\begin{bmatrix}\mathrm{vec}(Q)\\ \mathrm{vec}(P)\end{bmatrix}\right\|_2,$$

and then

$$\begin{aligned}\delta&=\inf_{\|(P,Q)\|_F=1}\|(QA_1+A_2P,QB_1+B_2P)\|_F\le\inf_{\|(P,Q)\|_F=1}\left\|\mathcal{M}\begin{bmatrix}\mathrm{vec}(Q)\\ \mathrm{vec}(P)\end{bmatrix}\right\|_2=\sigma_{\min}(\mathcal{M})\left\|\begin{bmatrix}\mathrm{vec}(Q)\\ \mathrm{vec}(P)\end{bmatrix}\right\|_2\\&\le\sqrt{2}\sigma_{\min}(\mathcal{M})\max\left\{\|\mathrm{vec}(Q)\|_2,\|\mathrm{vec}(P)\|_2\right\}\le\sqrt{2}\sigma_{\min}(\mathcal{M}).\end{aligned}$$

The proof is established. ∎

3.5 | A necessary condition for inexact RKSM to track exact RKSM

In Section 3.3, we derived a sufficient condition for $\|\xi_k\|_2$ at each step k, to make the difference Δ_m between the true residual and the derived residual in (14) smaller than a given tolerance. In this section, we will derive another upper bound on $\|\xi_{p+1}\|_2$ which is a necessary condition to make $\|\Delta_m\|_2$ smaller than a given tolerance.

For an m-step RKSM, we notice that $\|\Delta_m\|_2=\|\Xi_m U_m^1\|_2$, where the columns of $U_m^1\in\mathbb{R}^{m\times p}$ are partial general Schur vectors of the projection matrix pair (G_m,H_m) in (8). We partition U_m and Z_m in (6) into 2×2 blocks:

$$U_m=\begin{bmatrix}U_m^1 & U_m^2\end{bmatrix}=\begin{bmatrix}U_m^{11} & U_m^{12}\\ U_m^{21} & U_m^{22}\end{bmatrix},\ Z_m=\begin{bmatrix}Z_m^1 & Z_m^2\end{bmatrix}=\begin{bmatrix}Z_m^{11} & Z_m^{12}\\ Z_m^{21} & Z_m^{22}\end{bmatrix}.\tag{37}$$

It follows that the difference $\|\Delta_m\|_2$ can be divided into two parts:

$$\|\Delta_m\|_2=\left\|\Xi_pU_m^{11}+[\xi_{p+1},\dots,\xi_m]U_m^{21}\right\|_2\le\|\Xi_p\|_2+\|[\xi_{p+1},\dots,\xi_m]\|_2\left\|U_m^{21}\right\|_2.\tag{38}$$

Suppose that $\|U_m^{21}\|_2$ is not small. If $\|[\xi_{p+1},\dots,\xi_m]\|_2$ is not sufficiently small, then $\|\Delta_m\|_2$ may not be small, even if $\|\Xi_p\|_2=0$. This observation would generate another upper bound on the errors at later steps, for example, $\|\xi_{p+1}\|_2$, which is necessary (may not be sufficient) to keep $\|\Delta_m\|_2$ small. To this end, we first derive a lower bound on $\|U_m^{21}\|_2$ in the next theorem, which is similar to theorem 4.1 in Reference 29.

Theorem 3. *Let $AV_pH_p+h_{p+1,p}Av_{p+1}e_p^*+\Xi_p=V_pG_p+s_ph_{p+1,p}v_{p+1}e_p^*$ be a p-step inexact Arnoldi decomposition for RKSM, where the general Schur form of the matrix pair (G_p,H_p) is $G_p=Z_pS_pU_p^*$ and $H_p=Z_pT_pU_p^*$. Let $m-p$ additional inexact RKSM steps be performed, giving $AV_mH_m+h_{m+1,m}Av_{m+1}e_m^*+\Xi_m=V_mG_m+s_mh_{m+1,m}v_{m+1}e_m^*$. Let $\mathcal{R}_p=h_{p+1,p}v_{p+1}e_p^*U_p$ be the simplified residual at RKSM step p. Given the generalized Schur decomposition of (G_m,H_m) in (6) and partitions in (37), then*

$$\left\|\left(U_m^{12},Z_m^{12}\right)\right\|_F\ge\frac{\|\mathcal{R}_p\|_2}{\|\mathcal{R}_p\|_2+\|\mathcal{G}_m\|_F},\tag{39}$$

where $\mathcal{G}_m$ is the operator $(X,Y)\to\mathcal{G}_m(X,Y):\left(\frac{1}{s_p}S_m^{22}X+\frac{1}{s_p}YS_p,T_m^{22}X+YT_p\right)$, and $\|\mathcal{G}_m\|_F=\max_{\|(X,Y)\|_F=1}\|\mathcal{G}_m(X,Y)\|_F$.

Proof. The simplified eigenresidual norm at RKSM step p is

$$\|\mathcal{R}_p\|_2=\|h_{p+1,p}v_{p+1}e_p^*U_p\|_2=|h_{p+1,p}|\,\|e_p^*U_p\|_2.$$

Define

$$\Upsilon_1=G_m\begin{bmatrix}U_p\\0\end{bmatrix}-\begin{bmatrix}Z_p\\0\end{bmatrix}S_p=\begin{bmatrix}G_p & G_m^{12}\\ s_ph_{p+1,p}e_1e_p^* & G_m^{22}\end{bmatrix}\begin{bmatrix}U_p\\0\end{bmatrix}-\begin{bmatrix}Z_p\\0\end{bmatrix}S_p=\begin{bmatrix}G_pU_p-Z_pS_p\\ s_ph_{p+1,p}e_1e_p^*U_p\end{bmatrix}=\begin{bmatrix}0\\ s_ph_{p+1,p}e_1e_p^*U_p\end{bmatrix}.$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Similarly, define:

$$\Upsilon_2 = H_m \begin{bmatrix} U_p \\ 0 \end{bmatrix} - \begin{bmatrix} Z_p \\ 0 \end{bmatrix} T_p = \begin{bmatrix} H_p & H_m^{12} \\ h_{p+1,p} e_1 e_p^* & H_m^{22} \end{bmatrix} \begin{bmatrix} U_p \\ 0 \end{bmatrix} - \begin{bmatrix} Z_p \\ 0 \end{bmatrix} T_p = \begin{bmatrix} H_p U_p - Z_p T_p \\ h_{p+1,p} e_1 e_p^* U_p \end{bmatrix} = \begin{bmatrix} 0 \\ h_{p+1,p} e_1 e_p^* U_p \end{bmatrix}.$$

We can see that $\left\| \frac{1}{s_p} \Upsilon_1 \right\|_2 = \|\Upsilon_2\|_2 = \left\| h_{p+1,p} e_1 e_p^* U_p \right\|_2 = |h_{p+1,p}| \left\| e_p^* U_p \right\|_2 = \|\mathcal{R}_p\|_2$.

Using the partition of Z_m in (37), and left multiplying Z_m^* to Υ_1, we get

$$Z_m^* \Upsilon_1 = \begin{bmatrix} (Z_m^{11})^* & (Z_m^{21})^* \\ (Z_m^{12})^* & (Z_m^{22})^* \end{bmatrix} \begin{bmatrix} 0 \\ s_p h_{p+1,p} e_1 e_p^* U_p \end{bmatrix} = \begin{bmatrix} s_p h_{p+1,p} (Z_m^{21})^* e_1 e_p^* U_p \\ s_p h_{p+1,p} (Z_m^{22})^* e_1 e_p^* U_p \end{bmatrix}. \tag{40}$$

On the other hand, using the general Schur decomposition of H_m, we get

$$\begin{aligned} Z_m^* \Upsilon_1 &= Z_m^* G_m \begin{bmatrix} U_p \\ 0 \end{bmatrix} - Z_m^* \begin{bmatrix} Z_p \\ 0 \end{bmatrix} S_p = S_m U_m^* \begin{bmatrix} U_p \\ 0 \end{bmatrix} - Z_m^* \begin{bmatrix} Z_p \\ 0 \end{bmatrix} S_p = \begin{bmatrix} S_m^{11} & S_m^{12} \\ 0 & S_m^{22} \end{bmatrix} \begin{bmatrix} (U_m^{11})^* U_p \\ (U_m^{12})^* U_p \end{bmatrix} - \begin{bmatrix} (Z_m^{11})^* Z_p \\ (Z_m^{12})^* Z_p \end{bmatrix} S_p \\ &= \begin{bmatrix} S_m^{11} (U_m^{11})^* U_p + S_m^{12} (U_m^{12})^* U_p - (Z_m^{11})^* Z_p S_p \\ S_m^{22} (U_m^{12})^* U_p - (Z_m^{12})^* Z_p S_p \end{bmatrix}. \end{aligned} \tag{41}$$

Using the upper block from (40) and lower block from (41), we have

$$Z_m^* \Upsilon_1 = \begin{bmatrix} s_p h_{p+1,p} (Z_m^{21})^* e_1 e_p^* U_p \\ S_m^{22} (U_m^{12})^* U_p - (Z_m^{12})^* Z_p S_p \end{bmatrix}. \tag{42}$$

Similarly, if we left multiply Z_m^* to Υ_2, we will eventually have

$$Z_m^* \Upsilon_2 = \begin{bmatrix} h_{p+1,p} (Z_m^{21})^* e_1 e_p^* U_p \\ T_m^{22} (U_m^{12})^* U_p - (Z_m^{12})^* Z_p T_p \end{bmatrix}. \tag{43}$$

We have already shown that $\left\| \frac{1}{s_p} \Upsilon_1 \right\|_2 = \|\Upsilon_2\|_2 = \|\mathcal{R}_p\|_2$, and since $\frac{1}{s_p} \Upsilon_1$, Υ_2, and $\mathcal{R}_p$ are all rank-1 matrices, their Frobenius norms are also equal. It follows that:

$$\begin{aligned} \|\mathcal{R}_p\|_2 &= \max \left\{ \left\| \frac{1}{s_p} \Upsilon_1 \right\|_F, \|\Upsilon_2\|_F \right\} = \max \left\{ \left\| \frac{1}{s_p} Z_m^* \Upsilon_1 \right\|_F, \|Z_m^* \Upsilon_2\|_F \right\} \\ &\le \left\| h_{p+1,p} (Z_m^{21})^* e_1 e_p^* U_p \right\|_F + \left\| \mathcal{G}_m \left((U_m^{12})^* U_p, -(Z_m^{12})^* Z_p \right) \right\|_{\mathcal{F}} \\ &\le \|\mathcal{R}_p\|_2 \left\| (Z_m^{21})^* e_1 \right\|_2 + \|\mathcal{G}_m\|_{\mathcal{F}} \left\| \left((U_m^{12})^* U_p, -(Z_m^{12})^* Z_p \right) \right\|_{\mathcal{F}} \\ &\le \|\mathcal{R}_p\|_2 \left\| Z_m^{21} \right\|_2 + \|\mathcal{G}_m\|_{\mathcal{F}} \left\| (U_m^{12}, Z_m^{12}) \right\|_{\mathcal{F}}. \end{aligned} \tag{44}$$

Note that $\left\| (Z_m^{12})^* Z_p \right\|_2 = \|Z_m^{12}\|_2 = \|Z_m^{21}\|_2$; see, for example, Reference 24. It follows that $\|Z_m^{21}\|_2 = \|Z_m^{12}\|_2 \le \|Z_m^{12}\|_F \le \left\| (U_m^{12}, Z_m^{12}) \right\|_{\mathcal{F}}$. Based on (44), we get

$$\|\mathcal{R}_p\|_2 \le \left(\|\mathcal{R}_p\|_2 + \|\mathcal{G}_m\|_{\mathcal{F}} \right) \left\| (U_m^{12}, Z_m^{12}) \right\|_{\mathcal{F}}.$$

The lower bound on $\left\| (U_m^{12}, Z_m^{12}) \right\|_{\mathcal{F}}$ in (39) is thus established. ∎

As shown in Theorem 3, there exists a lower bound on $\left\| (U_m^{12}, Z_m^{12}) \right\|_{\mathcal{F}}$. We are most interested in $\|U_m^{12}\|_2$ for $m = p + 1$, which corresponds to the first RKSM step after each restart. The following lemma shows a lower bound on $\left\| U_{p+1}^{12} \right\|_2$.

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Lemma 2. *With the notation in Theorem 3,*

$$\left\|U_{p+1}^{12}\right\|_2 \geq \frac{\left\|\mathcal{R}_p\right\|_2}{\left\|\mathcal{R}_p\right\|_2 + \left\|\mathcal{G}_m\right\|_{\mathcal{F}}} \frac{1}{\theta}, \tag{45}$$

where $\theta = \max\left\{1, \left\|\frac{1}{s_p}S_{p+1}^{22} - T_{p+1}^{22}\right\|_2 \left\|\left(\frac{1}{s_p}S_p - T_p\right)^{-1}\right\|_2\right\}$.

Proof. It's easy to show that $\frac{1}{s_p}\Upsilon_1 = \Upsilon_2$, and therefore $\frac{1}{s_p}Z_m^*\Upsilon_1 = Z_m^*\Upsilon_2$. Based on the lower blocks in (42) and (43), we get

$$\begin{aligned}
&\frac{1}{s_p}S_m^{22}\left(U_m^{12}\right)^* U_p - \frac{1}{s_p}\left(Z_m^{12}\right)^* Z_p S_p = T_m^{22}\left(U_m^{12}\right)^* U_p - \left(Z_m^{12}\right)^* Z_p T_p \\
&\Leftrightarrow \left(Z_m^{12}\right)^* = \left(\frac{1}{s_p}S_m^{22} - T_m^{22}\right)\left(U_m^{12}\right)^* U_p \left(\frac{1}{s_p}S_p - T_p\right)^{-1}\left(Z_p\right)^{-1}.
\end{aligned}$$

Let $m = p+1$, and it follows that

$$\begin{aligned}
\left\|Z_{p+1}^{12}\right\|_2 &\leq \left\|\frac{1}{s_p}S_{p+1}^{22} - T_{p+1}^{22}\right\|_2 \left\|U_{p+1}^{12}\right\|_2 \left\|U_p\right\|_2 \left\|\left(\frac{1}{s_p}S_p - T_p\right)^{-1}\right\|_2 \left\|\left(Z_p\right)^{-1}\right\|_2 \\
&= \left\|\frac{1}{s_p}S_{p+1}^{22} - T_{p+1}^{22}\right\|_2 \left\|\left(\frac{1}{s_p}S_p - T_p\right)^{-1}\right\|_2 \left\|U_{p+1}^{12}\right\|_2,
\end{aligned}$$

Based on the definition of $\|\cdot\|_{\mathcal{F}}$ norm in (15), we immediately get:

$$\left\|\left(U_{p+1}^{12}, Z_{p+1}^{12}\right)\right\|_{\mathcal{F}} = \max\left\{\left\|U_{p+1}^{12}\right\|_F, \left\|Z_{p+1}^{12}\right\|_F\right\} \leq \left\|U_{p+1}^{12}\right\|_2 \theta,$$

where the last inequality holds because both U_{p+1}^{12} and Z_{p+1}^{12} are column vectors. Together with the lower bound on $\left\|\left(U_m^{12}, Z_m^{12}\right)\right\|_{\mathcal{F}}$ in Theorem 3, the lower bound on $\left\|U_m^{12}\right\|_2$ is thus established. ∎

As shown in Lemma 2, there exists a lower bound on $\left\|U_m^{21}\right\|_2$. To make $\|\Delta_m\|_2$ in (38) sufficient small, $\left\|[\xi_{p+1}, \ldots, \xi_m]\right\|_2$ cannot be too large. In particular, The following theorem gives an upper bound on $\left\|\xi_{p+1}\right\|_2$, similar to theorem 4.2 in Reference 29.

Theorem 4. *Given* $\epsilon_1 > 0$*, let* $AV_pH_p + h_{p+1,p}Av_{p+1}e_p^* + \Xi_p = V_pG_p + s_ph_{p+1,p}v_{p+1}e_p^*$ *be a p-step inexact Arnoldi relation for RKSM, where* $\left\|\Xi_p\right\|_2 \leq \epsilon_1$*. Let* $\mathcal{G}_{p+1}$ *be defined as in Theorem 3, and* θ *be defined as in Lemma 2. Then for the next RKSM step,*

$$\left\|\xi_{p+1}\right\|_2 \leq \left(\frac{\left\|\mathcal{R}_p\right\|_2 + \left\|\mathcal{G}_{p+1}\right\|_{\mathcal{F}}}{\left\|\mathcal{R}_p\right\|_2}\right)\theta\left(\epsilon_1 + \epsilon\right) \tag{46}$$

is a necessary condition to make $\left\|\Delta_{p+1}\right\|_2 \leq \epsilon$*, where* Δ_{p+1} *defined in (38) is the difference between the true and the derived residuals at the* $p+1$*st step of RKSM.*

Proof. Let $m = p+1$ for (38). Note that ξ_{p+1} and U_{p+1}^{21} are, respectively, a column vector and row vector, so $\left\|\xi_{p+1}U_{p+1}^{21}\right\|_2 = \left\|\xi_{p+1}\right\|_2\left\|U_{p+1}^{21}\right\|_2$. We get

$$\begin{aligned}
\left\|\Delta_{p+1}\right\|_2 &= \left\|\Xi_pU_{p+1}^{11} + \xi_{p+1}U_{p+1}^{21}\right\|_2 \geq \left\|\xi_{p+1}U_{p+1}^{21}\right\|_2 - \left\|\Xi_pU_{p+1}^{11}\right\|_2 \\
&\geq \left\|\xi_{p+1}\right\|_2\left\|U_{p+1}^{21}\right\|_2 - \left\|\Xi_p\right\|_2\left\|U_{p+1}^{11}\right\|_2 \geq \left\|\xi_{p+1}\right\|_2 \frac{\left\|\mathcal{R}_p\right\|_2}{\left\|\mathcal{R}_p\right\|_2 + \left\|\mathcal{G}_{p+1}\right\|_{\mathcal{F}}}\frac{1}{\theta} - \epsilon_1.
\end{aligned} \tag{47}$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Note that $\left\|\left(U_{p+1}^{12}\right)^* U_p\right\|_2 = \left\|U_{p+1}^{12}\right\|_2 = \left\|U_{p+1}^{21}\right\|_2$; see, for example, Reference 24. It follows immediately that (47) is greater than ϵ if $\left\|\xi_{p+1}\right\|_2 > \left(\frac{\|\mathcal{R}_p\|_2 + \|\mathcal{G}_{p+1}\|_F}{\|\mathcal{R}_p\|_2}\right)\theta\,(\epsilon_1 + \epsilon)$. Note that the bound (46) on the error at step $p+1$ is also inversely proportional to the residual norm $\left\|\mathcal{R}_p\right\|_2$ (assuming it is sufficiently small) as the bound in (45) for this step. ∎

The difference between Theorems 2 and 4 is that the former derived a sufficient condition to make $\|\Delta_m\|_2 \le \epsilon$, while the later derived a necessary condition to make $\left\|\Delta_{p+1}\right\|_2 \le \epsilon$. To be specific, if we follow the upper bounds on $\|\xi_k\|_2$ $(1 \le k \le m)$ in (35), it is guaranteed that $\|\Delta_m\|_2 \le \epsilon$. If $\left\|\xi_{p+1}\right\|_2$ is greater than the upper bound in (46), we have $\left\|\Delta_{p+1}\right\|_2 > \epsilon$, which means that the eigenresidual of inexact RKSM and that of the exact method differ more than the prescribed tolerance.

4 | RESTARTED RKSM WITH SCHUR DECOMPOSITION

In this section, we consider extending the observation we developed for unrestarted RKSM in Section 3 to the restarted variant,[30,31] which aims at saving storage and orthogonalization cost.

We first provide the formulation of inexact restarted RKSM, with the relationship between the eigenresiduals at the end of the current cycle and the beginning of the restarted cycle. We assume that p eigenpairs of A are wanted, and the largest dimension of rational Krylov subspaces used for projection is m $(m > p)$. For our problem setting, we adopt the widely-used Krylov-Schur restarting.[5,15] Assume that at the mth step of inexact RKSM, we end up with the Arnoldi decomposition:

$$AV_m^{(j)}H_m^{(j)} + h_{m+1,m}^{(j)}Av_{m+1}^{(j)}e_m^* + \Xi_m^{(j)} = V_m^{(j)}G_m^{(j)} + s_m^{(j)}h_{m+1,m}^{(j)}v_{m+1}^{(j)}e_m^*, \tag{48}$$

where the superscript $j \ge 0$ denotes the number of restarted cycles.

We apply the partial generalized Schur decomposition[25] to $\left(G_m^{(j)}, H_m^{(j)}\right)$ and get $G_m^{(j)}U_m^{1(j)} = Z_m^{1(j)}S_m^{11(j)}$ and $H_m^{(j)}U_m^{1(j)} = Z_m^{1(j)}T_m^{11(j)}$, where $U_m^{1(j)}, Z_m^{1(j)} \in \mathbb{R}^{m\times p}$ have orthonormal columns, and $\left(S_m^{11(j)}, T_m^{11(j)}\right)$ is a pair of (quasi) upper triangular matrices of order p, whose diagonal entries define the desired Ritz values obtained from $\left(G_m^{(j)}, H_m^{(j)}\right)$.

We post-multiply both sides of (48) by $U_m^{1(j)}$ and obtain

$$A\left[V_m^{(j)}Z_m^{1(j)}, v_{m+1}^{(j)}\right]\begin{bmatrix} T_m^{11(j)} \\ h_{m+1,m}^{(j)}e_m^*U_m^{1(j)}\end{bmatrix} + \Xi_m^{(j)}U_m^{1(j)} = \left[V_m^{(j)}Z_m^{1(j)}, v_{m+1}^{(j)}\right]\begin{bmatrix} S_m^{11(j)} \\ s_m^{(j)}h_{m+1,m}^{(j)}e_m^*U_m^{1(j)}\end{bmatrix}.$$

Then for next cycle of RKSM, it begins with the Arnoldi relation

$$AV_{p+1}^{(j+1)}\underline{H}_p^{(j+1)} + \Xi_p^{(j+1)} = V_{p+1}^{(j+1)}\underline{G}_p^{(j+1)},$$

where

$$\begin{cases} \underline{H}_p^{(j+1)} = \begin{bmatrix} T_m^{11(j)} \\ h_{m+1,m}^{(j)}e_m^*U_m^{1(j)}\end{bmatrix}, & \underline{G}_p^{(j+1)} = \begin{bmatrix} S_m^{11(j)} \\ s_m^{(j)}h_{m+1,m}^{(j)}e_m^*U_m^{1(j)}\end{bmatrix}, \\ V_{p+1}^{(j+1)} = \begin{bmatrix} V_m^{(j)}Z_m^{1(j)} & v_{m+1}^{(j)}\end{bmatrix}, & \Xi_p^{(j+1)} = \Xi_m^{(j)}U_m^{1(j)}. \end{cases} \tag{49}$$

Since $G_p^{(j+1)}$ and $H_p^{(j+1)}$ are both already (quasi) upper triangular, if we follow the generalized Schur decomposition in (6) for $\left(G_p^{(j+1)}, H_p^{(j+1)}\right)$, we immediately get $U_p^{(j+1)} = Z_p^{(j+1)} = I_p$, $S_p^{(j+1)} = G_p^{(j+1)}$, and $T_p^{(j+1)} = H_p^{(j+1)}$. Let $R_m^{(j)}$ and $R_p^{(j+1)}$ be the derived residual corresponding to $U_m^{(j)}$ and $U_p^{(j+1)}$, respectively. Then:

$$R_p^{(j+1)} = h_{m+1,m}^{(j)}v_{m+1}^{(j)}e_m^*U_m^{1(j)}U_p^{(j+1)}\left(s_m^{(j)}I - \Theta_p\right) = h_{m+1,m}^{(j)}v_{m+1}^{(j)}e_m^*U_m^{1(j)}\left(s_m^{(j)}I - \Theta_p\right) = R_m^{(j)}. \tag{50}$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Similarly, the simplified eigenresidual in (32) satisfies:

$$\mathcal{R}_p^{(j+1)} = h_{m+1,m}^{(j)} v_{m+1}^{(j)} e_m^* U_m^{1(j)} U_p^{(j+1)} = h_{m+1,m}^{(j)} v_{m+1}^{(j)} e_m^* U_m^{1(j)} = \mathcal{R}_m^{(j)}. \tag{51}$$

Our next step is to quantify the allowable errors introduced at each RKSM step. Except the first cycle, errors that occur at each of the $m-p$ steps during the jth cycle are denoted as $\xi_{p+1}^{(j)}, \xi_{p+2}^{(j)}, ..., \xi_m^{(j)}$ $(j>1)$. Additional errors are inherited from the previous cycle, represented by $\Xi_m^{(j-1)} U_m^{1(j-1)}$. The algorithm of restarted RKSM is shown in Algorithm 2.

Algorithm 2. Restarted RKSM for eigenvalue computation

Input: $A \in \mathbb{R}^{n\times n}$, $v_1 \in \mathbb{R}^n$ and $\|v_1\| = 1$, max subspace dimension $m+1$, max restarts J, tolerance tol>0.
Output: desired p eigenvalues and the corresponding invariant subspace.

Follow Algorithm 1 to go through the first cycle of RKSM; obtain the Arnoldi relation at the end of this cycle $AV_{m+1}^{(1)} \underline{H}_m^{(1)} = V_{m+1}^{(1)} \underline{G}_m^{(1)}$.
for $j = 2, 3, \dots, J$ **do**
  Compute the generalized Schur decomposition $G_m^{(j-1)} U_m^{1(j-1)} = Z_m^{1(j-1)} S_m^{11(j-1)}$ and $H_m^{(j-1)} U_m^{1(j-1)} = Z_m^{1(j-1)} T_m^{11(j-1)}$.
  Obtain $\underline{H}_p^{(j)}$, $\underline{G}_p^{(j)}$, and $V_{p+1}^{(j)}$ from (49) to shrink the dimension of subspace.
  for $k = p+1, p+2, \dots, m$ **do**
    Choose the pole $s_k^{(j)}$ at step k.
    Let $w_{k+1} = (A - s_k^{(j)} I)^{-1} v_k$, orthogonalize against $V_k^{(j)}$ and normalize into $v_{k+1}^{(j)}$; update $G_k^{(j)}$ and $F_k^{(j)}$.
    Compute the generalized Schur decomposition of matrix pair $\left(G_k^{(j)}, F_k^{(j)}\right)$.
    if $\left\|R_k^{(j)}\right\|_F < tol$, **then**
      Return the eigenvalues of $\left(T_k^{11(j)}\right)^{-1} S_k^{11(j)}$ in (7) as approximations to the desired eigenvalues of A, and $V_{k+1}^{(j)} \underline{H}_k^{(j)} U_k^{1(j)}$ as approximations to the desired invariant subspace.
    end if
  end for
end for

Similar to (14), the difference between the true and the derived eigenresidual of inexact RKSM at the end of the jth cycle is:

$$\Delta_m^{(j)} = \Xi_m^{(j)} U_m^{(j)}. \tag{52}$$

As we did for RKSM without restart, we want to derive a bound on the errors allowed at each rational Krylov step, so that the inexact restarted RKSM can achieve the desired eigenresiduals sufficiently close to those obtained by the exact counterpart. To this end, we first present an inequality in linear algebra, then give the main result on inexact restarted RKSM.

Lemma 3. *For matrix $A \in \mathbb{R}^{n\times m}$ $(n \geq m)$, any matrix $Q \in \mathbb{R}^{m\times k}$ $(m \geq k)$ with orthonormal columns satisfies*

$$\inf_{Q\in\mathbb{R}^{m\times k}, Q^*Q=I} \sum_{i=1}^{k} \|AQe_i\|_2 \leq \sqrt{k} \sqrt{\sum_{i=1}^{k} \sigma_i^2},$$

where $\sigma_1, \sigma_2, \dots, \sigma_k$ are the k largest singular values of A.

Proof. Assume A has a singular value decomposition $A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^*$, where $U \in \mathbb{C}^{n\times n}$, $V \in \mathbb{C}^{m\times m}$, $\Sigma \in \mathbb{C}^{m\times m}$, and Σ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m$ on its diagonal. With $W = V^*Q$, it's easy to show that

$$\inf_{Q\in\mathbb{R}^{m\times k}, Q^*Q=I} \sum_{i=1}^{k} \|AQe_i\|_2 = \inf_{W\in\mathbb{R}^{m\times k}, W^*W=I} \sum_{i=1}^{k} \left\| \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} We_i \right\|_2.$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The object function is the sum of 2-norm of each column of the matrix $\begin{bmatrix}\Sigma\\0\end{bmatrix} W$. We assume that the 2-norm of each column is $a_1, a_2, ..., a_k$. Then

$$\sum_{i=1}^{k}\left\|\begin{bmatrix}\Sigma\\0\end{bmatrix} We_i\right\|_2 = \sum_{i=1}^{k} a_i \le \sqrt{k}\sqrt{\sum_{i=1}^{k} a_i^2} = \sqrt{k}\left\|\begin{bmatrix}\Sigma\\0\end{bmatrix} W\right\|_F.$$

Denote $W = [w_1, w_2, ..w_k]$, then

$$\left\|\begin{bmatrix}\Sigma\\0\end{bmatrix} W\right\|_F = \sqrt{\text{trace}\left(W^*\Sigma^2 W\right)} = \sqrt{\sum_{i=1}^{k} w_i^*\Sigma^2 w_i}.$$

By using Ky Fan's Maximum Principle[32(p. 24, problem I.6.15)], we get

$$\inf_{\substack{Q\in\mathbb{R}^{m\times k},\\ Q^*Q=I}} \sum_{i=1}^{k}\|AQe_i\|_2 \le \inf_{\substack{W\in\mathbb{R}^{m\times k},\\ W^*W=I}} \sqrt{k}\sqrt{\sum_{i=1}^{k} w_i^*\Sigma^2 w_i} = \sqrt{k}\sqrt{\sum_{i=1}^{k}\sigma_i^2}.$$

The proof is established. ∎

Theorem 5. *Consider using inexact restarted RKSM for computing a set of p eigenvalues of matrix $A \in \mathbb{R}^{n\times n}$, with maximum subspace dimension $m+1$, and maximum cycle J. Let all quantities with superscript j refer to those in the jth cycle previously defined for the unrestarted RKSM. Given $\epsilon > 0$, assume that for all $1 \le k \le m$ and $1 \le j \le J$,*

$$\left\|\xi_k^{(j)}\right\|_2 = \begin{cases} \dfrac{\delta_{m,k-1}^{(j)}\epsilon}{2Jm\omega_{k-1}^{(j)}\left\|\mathcal{R}_{k-1}^{(j)}\right\|_2}, & \text{if } k > p \text{ and } \left\|\mathcal{R}_{k-1}^{(j)}\right\|_2 < \dfrac{(\delta_{m,k-1}^{(j)})^2}{4\eta\omega_{k-1}^{(j)}} \\ \dfrac{\epsilon}{Jm}, & \text{otherwise.}\end{cases} \tag{53}$$

Let $\sigma_1^{(j)} \ge \sigma_2^{(j)} \ge \cdots \ge \sigma_m^{(j)}$ be the singular values of the error matrix $\Xi_m^{(j)}$. If

$$\sqrt{\sum_{i=1}^{p}\sigma_i^{(j)2}} \le \sqrt{p}\frac{\delta_{m,p}^{(j+1)}\epsilon}{2Jm\omega_p^{(j+1)}\left\|\mathcal{R}_m^{(j)}\right\|_2}, \tag{54}$$

then

$$\left\|\Delta_m^{(J)}\right\|_2 = \left\|R_m^{(J)} - \tilde{R}_m^{(J)}\right\|_2 \le \epsilon.$$

Proof. To prove this theorem, it is sufficient to prove that for $1 \le j \le J$,

$$\left\|\Delta_m^{(j)}\right\|_2 = \left\|R_m^{(j)} - \tilde{R}_m^{(j)}\right\|_2 \le \frac{j}{J}\epsilon, \tag{55}$$

which can be shown by mathematical induction. For $j = 1$, (55) holds directly by Theorem 2. Now assume when $j = i \ge 1$,

$$\left\|\Delta_m^{(i)}\right\|_2 = \left\|R_m^{(i)} - \tilde{R}_m^{(i)}\right\|_2 = \left\|\Xi_m^{(i)}U_m^{1(i)}\right\|_2 \le \frac{i}{J}\epsilon, \tag{56}$$

and we want to show that (56) also holds for superscript $j = i+1$.

By the assumption of the theorem, for $j = i+1$, if $\mathcal{R}_{k-1}^{(i+1)}$ satisfies $\left\|\mathcal{R}_{k-1}^{(i+1)}\right\|_2 < \frac{(\delta_{m,k-1}^{(i+1)})^2}{4\eta\omega_{k-1}^{(i+1)}}$, then from Proposition 1:

$$\widehat{U}_m^{(i+1)} = \begin{bmatrix}\widehat{U}_a^{(i+1)}\\ \widehat{U}_b^{(i+1)}\end{bmatrix} = \left(\begin{bmatrix}U_{k-1}^{1(i+1)}\\0\end{bmatrix} + \begin{bmatrix}X_1\\X_2\end{bmatrix}P\right)(I+P^*P)^{-\frac{1}{2}},$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $0 \le \|P\|_F < 2\frac{\omega_{k-1}^{(i+1)}\left\|\mathcal{R}_{k-1}^{(i+1)}\right\|_2}{\delta_{m,k-1}^{(i+1)}}$, and $\widehat{U}_m^{(i+1)} \in \mathbb{R}^{m\times p}$ and $U_{k-1}^{1(i+1)} \in \mathbb{R}^{(k-1)\times p}$ contain the Schur vectors of $\left(G_m^{(i+1)}, H_m^{(i+1)}\right)$ and $\left(G_{k-1}^{(i+1)}, H_{k-1}^{(i+1)}\right)$ corresponding to the desired spectra, respectively. Besides, for any $k \le l \le m$:

$$\left\|e_l^*\widehat{U}_m^{(i+1)}\right\|_2 \le \left\|\widehat{U}_b^{(i+1)}\right\|_2 \le \|P\|_F < 2\frac{\omega_{k-1}^{(i+1)}\left\|\mathcal{R}_{k-1}^{(i+1)}\right\|_2}{\delta_{m,k-1}^{(i+1)}}, \tag{57}$$

and

$$\left\|\begin{bmatrix} I_{k-1} & 0\end{bmatrix}\widehat{U}_m^{(i+1)}\right\|_2 = \left\|\widehat{U}_a^{(i+1)}\right\|_2 = \left\|\left(U_{k-1}^{1(i+1)} + X_1P\right)(I+P^*P)^{-\frac{1}{2}}\right\|_2. \tag{58}$$

In particular, when $k = p+1$, the definition of $\omega_{k-1}^{(i+1)} = \omega_p^{(i+1)}$ is slightly different from other $\omega_{k-1}^{(i+1)}$ with $k > p+1$. To be specific, note that $\underline{G}_p^{(i+1)}$ and $\underline{H}_p^{(i+1)}$ in (49) is not an upper Hessenberg matrix. From (50),

$$\left\|R_p^{(i+1)}\right\|_2 = \left|h_{m+1,m}^{(i)}\right|\left\|e_m^*U_m^{1(i)}\left(s_m^{(i)}I - \Theta_p\right)\right\|_2.$$

Based on (49), the (2, 1) blocks of $H_m^{(i+1)}$ and $G_m^{(i+1)}$ are $h_{m+1,m}^{(i)}e_1e_m^*U_m^{1(i)}$ and $s_m^{(i)}h_{m+1,m}^{(i)}e_1e_m^*U_m^{1(i)}$, respectively. Following the procedure in (24) and (25),

$$\begin{aligned}\|A_{21}\|_F &= \left\|s_m^{(i)}h_{m+1,m}^{(i)}W_2^*e_1e_m^*U_m^{1(i)}\right\|_F = |s_m^{(i)}h_{m+1,m}^{(i)}|\left\|e_m^*U_m^{1(i)}\right\|_2,\\ \|B_{21}\|_F &= \left\|h_{m+1,m}^{(i)}W_2^*e_1e_m^*U_m^{1(i)}\right\|_F = |h_{m+1,m}^{(i)}|\left\|e_m^*U_m^{1(i)}\right\|_2.\end{aligned}$$

Therefore, we can still write $\gamma = \|(A_{21}, B_{21})\|_F = \omega_p^{(i+1)}|h_{m+1,m}^{(i)}|\left\|e_m^*U_m^{1(i)}\right\|_2$ with $\omega_p^{(i+1)} := \max\left\{1, |s_m^{(i)}|\right\}$.

Let $\Omega_1^{(i+1)}$ be a subset of $\{p+1, p+2, \dots, m\}$, such that for each $k \in \Omega_1^{(i+1)}$, the condition $\left\|\mathcal{R}_{k-1}^{(j)}\right\|_2 < \frac{(\delta_{m,k-1}^{(j)})^2}{4\eta\omega_{k-1}^{(j)}}$ is satisfied with the cycle number $j = i+1$. Also, define $\Omega_2^{(i+1)} = \{p+1, p+2, \dots, m\} \setminus \Omega_1^{(i+1)}$. Then

$$\begin{aligned}\left\|\Delta_m^{(i+1)}\right\|_2 &= \left\|\Xi_m^{(i+1)}\widehat{U}_m^{(i+1)}\right\|_2 = \left\|\Xi_m^{(i+1)}\begin{bmatrix} I_p \\ 0\end{bmatrix}\begin{bmatrix} I_p & 0\end{bmatrix}\widehat{U}_m^{(i+1)} + \sum_{k=p+1}^{m}\xi_k^{(i+1)}e_k^*\widehat{U}_m^{(i+1)}\right\|_2\\ &\le \left\|\Xi_p^{(i+1)}\begin{bmatrix} I_p & 0\end{bmatrix}\widehat{U}_m^{(i+1)}\right\|_2 + \sum_{k=p+1}^{m}\left\|\xi_k^{(i+1)}\right\|_2\left\|e_k^*\widehat{U}_m^{(i+1)}\right\|_2.\end{aligned} \tag{59}$$

Similar to the proof of Theorem 2, if we refer to (53) and (57), the second term in the last line of (59) can be bounded as follows:

$$\begin{aligned}&\sum_{k=p+1}^{m}\left\|\xi_k^{(i+1)}\right\|_2\left\|e_k^*\widehat{U}_m^{(i+1)}\right\|_2 \le \sum_{k\in\Omega_1^{(i+1)}}\frac{\delta_{m,k-1}^{(i+1)}\epsilon}{2Jm\omega_{k-1}^{(i+1)}\left\|\mathcal{R}_{k-1}^{(i+1)}\right\|_2}2\frac{\omega_{k-1}^{(i+1)}\left\|\mathcal{R}_{k-1}^{(i+1)}\right\|_2}{\delta_{m,k-1}^{(i+1)}} + \sum_{k\in\Omega_2^{(i+1)}}\frac{\epsilon}{Jm}\left\|e_k^*\widehat{U}_m^{(i+1)}\right\|_2\\ &= \frac{\epsilon|\Omega_1^{(i+1)}|}{Jm} + \frac{\epsilon|\Omega_2^{(i+1)}|}{Jm} = \frac{\epsilon(m-p)}{Jm}.\end{aligned}$$

Considering (49), (51), (54), (56), and (58) with $k = p+1$, together with Lemma 3, we can bound the first term in the last expression in (59):

$$\begin{aligned}\left\|\Xi_p^{(i+1)}\begin{bmatrix} I_p & 0\end{bmatrix}\widehat{U}_m^{(i+1)}\right\|_2 &= \left\|\Xi_m^{(i)}U_m^{1(i)}\left(U_p^{(i+1)} + X_1P\right)(I+P^*P)^{-\frac{1}{2}}\right\|_2 \le \left\|\Xi_m^{(i)}U_m^{1(i)}\left(I_p + X_1P\right)\right\|_2\\ &\le \left\|\Xi_m^{(i)}U_m^{1(i)}\right\|_2 + \sum_{k=1}^{p}\left\|\Xi_m^{(i)}U_m^{1(i)}e_k\right\|_2\left\|e_k^*X_1P\right\|_2 \le \frac{i}{J}\epsilon + \sqrt{p}\sqrt{\sum_{j=1}^{p}\sigma_j^{(i)2}}\|P\|_2\\ &\le \frac{i}{J}\epsilon + \sqrt{p}\frac{\sqrt{p}\delta_{m,p}^{(i+1)}\epsilon}{2Jm\omega_p^{(i+1)}\left\|\mathcal{R}_m^{(i)}\right\|_2}2\frac{\omega_p^{(i+1)}\left\|\mathcal{R}_p^{(i+1)}\right\|_2}{\delta_{m,p}^{(i+1)}} = \frac{i}{J}\epsilon + \frac{p}{Jm}\epsilon.\end{aligned}$$

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Combining the bounds on the two terms in the last line of (59), we get:

$$\left\|\Delta_m^{(i+1)}\right\|_2 \leq \frac{\epsilon(m-p)}{Jm} + \frac{i}{J}\epsilon + \frac{p}{Jm}\epsilon = \frac{i+1}{J}\epsilon.$$

This concludes our mathematical induction for (55). ∎

Similar to the situation for the non-restarted inexact RKSM, it is impossible to know the exact value of $\delta_{m,k-1}^{(j)}$ in the jth cycle before end of this cycle. For the first cycle, we can follow Section 3.4 for the non-restarted method to evaluate δ. For the jth cycle ($j > 1$), we can use $\delta_{m,m}^{(j-1)}$ to approximate $\delta_{m,k-1}^{(j)}$, for any values of k ($k \leq m$), so that this quantity is fixed for the entire jth cycle. To be specific, at the beginning of the jth cycle, we can use the generalized Schur decomposition of $\left(G_m^{(j-1)}, H_m^{(j-1)}\right)$ from the last cycle. Let $\mathcal{X}^{(j-1)}, \mathcal{Y}^{(j-1)} \in \mathbb{R}^{m\times m}$ contain the Schur vectors of $\left(G_m^{(j-1)}, H_m^{(j-1)}\right)$, such that

$$\mathcal{Y}^{(j-1)*} G_m^{(j-1)} \mathcal{X}^{(j-1)} = \begin{bmatrix} \tilde{G}_1^{(j-1)} & \times \\ 0 & \tilde{G}_2^{(j-1)} \end{bmatrix}, \ \mathcal{Y}^{(j-1)*} H_m^{(j-1)} \mathcal{X}^{(j-1)} = \begin{bmatrix} \tilde{H}_1^{(j-1)} & \times \\ 0 & \tilde{H}_2^{(j-1)} \end{bmatrix},$$

where $\tilde{G}_1^{(j-1)}, \tilde{H}_1^{(j-1)} \in \mathbb{R}^{p\times p}$ are (quasi) upper triangular matrices whose diagonal entries define the p desired Ritz values. Then $\delta_{m,k-1}^{(j)}$ can be approximated by

$$\tilde{\delta}^{(j)} = \mathrm{dif}\left[\left(\tilde{G}_1^{(j-1)}, \tilde{H}_1^{(j-1)}\right), \left(\tilde{G}_2^{(j-1)}, \tilde{H}_2^{(j-1)}\right)\right]. \tag{60}$$

This quantity can be then approximated by the lower bound given in Lemma 1.

5 | NUMERICAL TEST

To support the strategy of relaxing the accuracy for solving the linear system at each RKSM step suggested by Theorem 2 for inexact non-restarted RKSM and Theorem 5 for inexact restarted RKSM, we report numerical experiment results in this section.

In all numerical experiments, we assume that the condition $\|\mathcal{R}_k\|_2 < \frac{\delta_{m,k}^2}{4\eta\omega_k}$ is always satisfied for all k. Actually, if our desired eigenvalues are well-separated from other eigenvalues, $\delta_{m,k}$ should not be too small so that this condition can be satisfied in practice. For the restarted method, we also assume that condition (54) is always satisfied. All experiments were carried out in MATLAB R2019b in Windows 10 on a laptop with a 16GB DDR4 2400MHz memory, and a 2.81 GHz Intel dual Core CPU.

5.1 | Compare exact and inexact RKSM with artificial errors

We first show that the convergence of inexact methods can match that of exact methods if we artificially introduce properly controlled errors to the shift-invert matrix vector product at each step of RKSM. Let us recall that the number of desired eigenvalues is p, the maximum subspace dimension is $m + 1$, and the maximum number of restarted cycles is J. For exact RKSM, the shift-invert matrix-vector product is solved by *backslash* in MATLAB. For inexact unrestarted RKSM, we still use *backslash* for the first p steps, and after that we artificially add a random error term ξ_k to the right-hand side of the linear system at each step based on Theorem 2, which satisfies:

$$\|\xi_k\|_2 \leq \frac{\delta_{m,k-1}\epsilon}{2m\omega_{k-1}\|\mathcal{R}_{k-1}\|_2}, \tag{61}$$

where ϵ is the tolerance of the difference between the residuals obtained by the exact and the inexact methods.

In Section 3.4, we have discussed a practical approximation to $\delta_{m,k-1}$ by $\tilde{\delta}_{m,k-1}$ in (36), which can be computed when step $k-1$ is done. But in practice, it's not advisable to use $\tilde{\delta}_{m,k-1}$ to approximate $\delta_{m,k-1}$ when k is small. To be specific,

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

when k is small, the spectrum of the projection matrix pair (G_k, H_k) varies with k significantly and cannot be a good approximation to the spectrum of the original matrix A. Consequently, the value of $\delta_{m,k-1}$ defined in (31) and $\tilde{\delta}_{m,k-1}$ defined in (36) are so different, that it is not reliable to approximate $\delta_{m,k-1}$ by $\tilde{\delta}_{m,k-1}$ at the first few steps of RKSM. In our test, $\tilde{\delta}_{m,k-1}$ tends to be relative large when k is small, and it usually decreases significantly when k increases. Therefore, we may let $\delta_{m,k-1} = \frac{\epsilon}{m}$ when k is small, then let $\delta_{m,k-1} = \tilde{\delta}_{m,k-1}$ when $\tilde{\delta}_{m,k-1}$ becomes stable. In addition, as we use Lemma 1 to compute the approximate value of δ, it takes time to compute the singular values of a large matrix constructed by Kronecker product, especially when the dimension of subspace is large. Our experience suggests that when the dimension of subspace k increases to a certain level, $\tilde{\delta}_{m,k-1}$ tends to become stable. Therefore it's sufficient to keep $\tilde{\delta}_{m,k-1}$ fixed after several iterations. In our test, for $k > 5p$, we fix $\delta_{m,k-1} = \min_{p<j\le 5p}\left\{\tilde{\delta}_{m,j-1}\right\}$. Also, whenever $\tilde{\delta}_{m,k-1}$ in (36) is less than $\frac{\epsilon}{m}$, it is sufficient to set $\delta_{m,k-1} = \frac{\epsilon}{m}$.

For inexact restarted RKSM, the first cycle can follow the evaluation of $\delta_{m,k-1}$ for the non-restarted method. For the jth cycle ($j > 1$), we can use $\tilde{\delta}^{(j)}$ in (60) to approximate $\delta^{(j)}_{m,k-1}$. For each restarted cycle, we begin with shrinking the dimension of the subspace from $m+1$ to $q = \min\{2p, p+5\}$ by generalized Schur decomposition. We let the dimension of the subspace right after restart be larger than the number of desired eigenvalues, which makes it more likely that our desired eigenpairs are kept in the rational Krylov subspace upon restart. We also artificially add a random error term $\xi_k^{(j)}$ to the right-hand side of the linear system at each step, which satisfies (see Theorem 5):

$$\left\|\xi_k^{(j)}\right\|_2 \le \frac{\delta^{(j)}_{m,k-1}\epsilon}{2Jm\omega^{(j)}_{k-1}\left\|\mathcal{R}^{(j)}_{k-1}\right\|_2}. \tag{62}$$

We hope that the above configuration helps satisfy the conditions specified in Theorems 2 and 5, so that we can see the inexact methods converge as rapidly as their exact counterpart.

Example 1. We consider a scaled 2D discrete Laplacian matrix of order $127^2 = 16{,}129$ based on standard 5-point stencils on a square, which is a classical symmetric positive definite (SPD) matrix, and we generate it by the function "`delsq`" in MATLAB. This matrix can be written as $A_0 \otimes I + I \otimes A_0$, where A_0 is the 1-D discrete Laplacian based on the 2nd order centered finite differences, with 2's on the diagonal, and -1's on its super-diagonal and subdiagonal entries. In this example, we perform two experiments to find the smallest eigenvalue alone and the 3 smallest eigenvalues, respectively. We choose $m = 50$, and $\epsilon = 10^{-11}$. To compute the smallest eigenvalue, we set repeated poles $s_1 = -1$, $s_2 = -0.1$, and $s_3 = 0$, and we get the approximate smallest eigenvalue to be 1.2047×10^{-3}. Then we set repeated poles $s_1 = 1.5 \times 10^{-3}$, and $s_2 = 2.5 \times 10^{-3}$ to compute the 3 smallest eigenvalues.

Figure 1 shows that inexact RKSM can converge within the given tolerance ϵ. We also notice that in Figure 1B, as we compute multiple eigenvalues, there are significant differences between the residuals of exact and inexact methods at the same step. Theoretically, based on Theorem 2, it only guarantees that the difference between the true and the derived residuals for inexact RKSM is no greater than ϵ, not the difference between the true residuals for the exact and the inexact methods. After sufficient number of steps, both methods nearly converge, and they tend to find the same set of the desired eigenvalues, with similar true eigenresiduals no greater than ϵ in the norm of their difference.

Restarted RKSM is not considered in this example, since this problem only takes a very small number of steps to converge.

Example 2. We consider a matrix $A \in \mathbb{R}^{16388\times 16388}$ from the *aerofoilA* problem. All eigenvalues of A are plotted in Figure 2.

Our test is to find two eigenvalues of matrix A that are closest to the imaginary axis. We set three repeated poles $s_1 = -1$, $s_2 = -0.1$, and $s_3 = 0$, and apply them cyclically for RKSM. For this example, we set $m = 100$, and $\epsilon = 10^{-12}$.

We apply two different strategies for setting the value of δ: the first one is to approximate it by using the lower bound in Lemma 1 on (36), which is denoted by δ_1; the second one is denoted by

$$\delta_2 = d_{k-1}\min\left\{\left\|\tilde{G}^{11}_{k-1}\right\|_F, \left\|\tilde{G}^{22}_{k-1}\right\|_F\right\} \tag{63}$$

at step k, where d_{k-1} is the distance between the inverse of desired eigenvalues and the undesired eigenvalues of the current step. The motivation of using δ_2 is from our observation of the definition in Figure 3B, and similar approaches

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

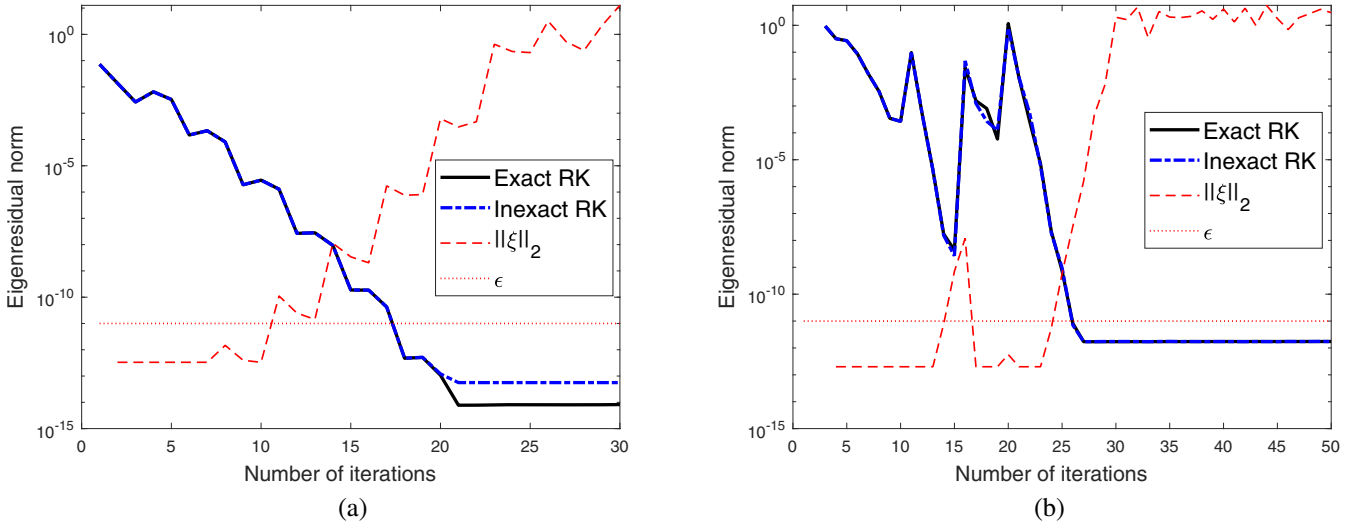

FIGURE 1 Performance of exact and inexact RKSM for Laplacian matrix. (A) Computing the smallest eigenvalue; (B) computing the 3 smallest eigenvalues

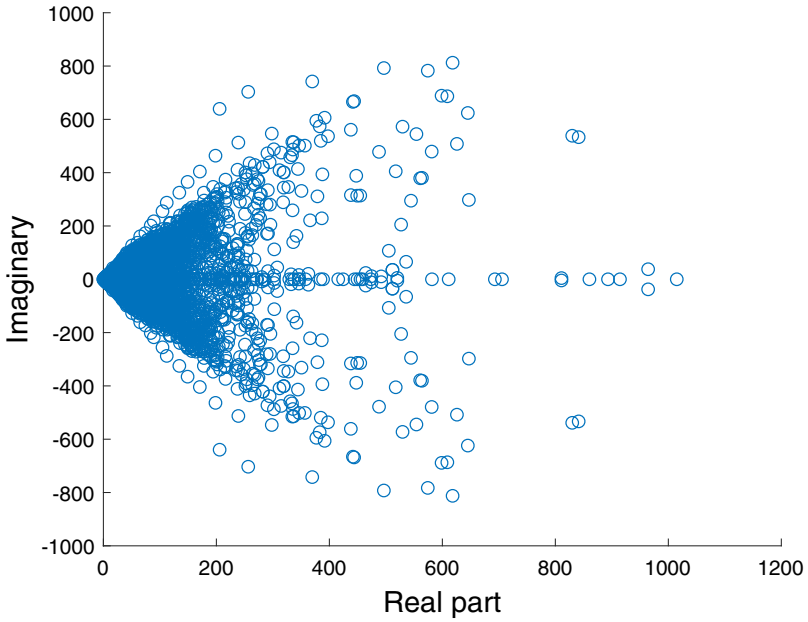

FIGURE 2 Eigenvalues of aerofoilA problem

can be seen in References 16,17. We claim that δ should be proportional to the difference between the eigenvalues of two matrix pairs and also to the norm of those matrices. We find that this strategy works for inexact RKSM in most tests.

For both strategies, when δ becomes stable, the value of δ_1 is around 3.96×10^{-3}, whereas the value of δ_2 is around 2.69. As can be seen from Figure 3B, if we increase δ_1 to the value of δ_2, the difference in eigenresiduals between the exact and inexact methods is still less than the tolerance we set.

For restarted RKSM, we let $M = 50, J = 4$, and $\epsilon = 10^{-12}$. We also apply the two different strategies for setting the value of δ. The result is shown in Figure 4.

Similar with the unrestarted method, inexact restarted RKSM still performs well even though we set a relatively large value $\delta = \delta_2$. To sum up, Lemma 1 provides an approximation to $\delta_{m,k-1}$ that guarantees near identical convergence behavior of the exact and the inexact methods, but it might be excessively conservative for nonsymmetric matrices. This approximation could be relaxed considerably without deteriorating the convergence of the inexact method.

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

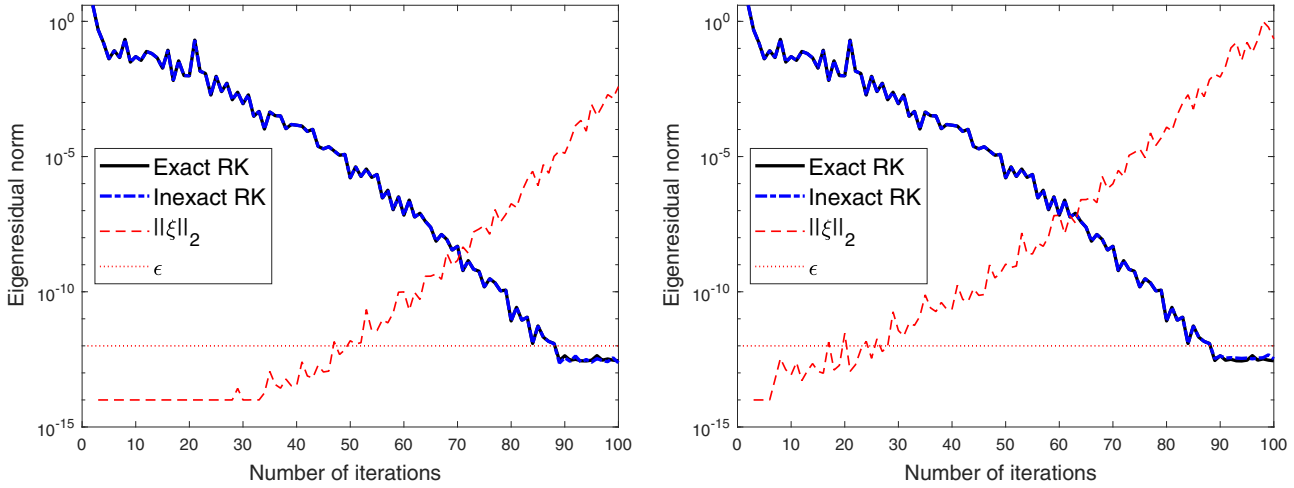

FIGURE 3 Performance of exact and inexact unrestarted RKSM for *aerofoilA* problem. (A) Adaptive δ_1; (B) δ_2

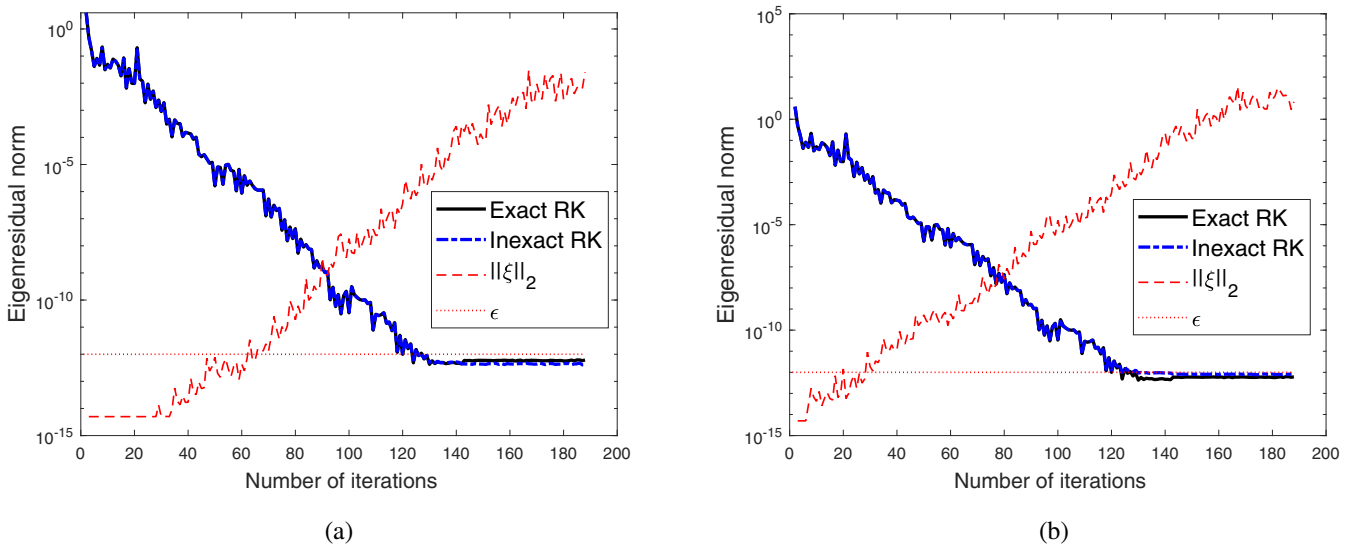

FIGURE 4 Performance of exact and inexact restarted RKSM for *aerofoil* problem. (A) Adaptive δ_1; (B) δ_2

5.2 | Inexact RKSM with GMRES as inner linear solver

For large-scale practical applications, and those arising from PDEs in 3D domains in particular, iterative methods are recommended to solve the linear system at each RKSM step. In this section, we demonstrate the behavior of inexact RKSM where the inner linear systems are solved by GMRES.

Instead of artificially adding an error term as we did in the previous examples, we set $\|\xi\|_2$ proportional to the tolerance for solving linear systems by GMRES at each inner step of RKSM. This value is bounded by (61) and (62) for the unrestarted and the restarted RKSM, respectively. We set $\delta_{m,k-1}$ to be δ_2 in (63).

We test 6 nonsymmetric real matrices, and these matrices are found in the SuiteSparse Matrix Collection.[33] *Epb1*, *Goodwin054*, *poli3*, and *aerofoilB* have all eigenvalues on the right half complex plane, while the other matrices have most eigenvalues on the right half complex plane. The matrix in *matRE500C* is in the form $A = M^{-1}K$ where both M and K are sparse, but A is not formed explicitly. Our test aims to compute several eigenvalues closest to the imaginary axis, and they either form complex conjugate pairs or are real eigenvalues. In Table 1, we reports some properties and parameters setting for each matrix: the matrix size n, the number of desired eigenvalues p, the maximum dimension of

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Information and parameters setting for test problems

Problem	Size n	p	$\{M, J\}$	Cyclic poles	Preconditioner	Tol
*epb*1	14734	5	$\{70, 1\}$	$0, -0.01, -0.1$	ILUTP, 10^{-3}	10^{-13}
*Goodwin*054	32510	5	$\{70, 1\}$	$-10^{-4}, -10^{-3}, -10^{-2}$	ILUTP, 10^{-2}	10^{-12}
big	13209	5	$\{70, 1\}$	$0, -10, -100$	ILUTP, 10^{-3}	10^{-13}
*poli*3	16955	20	$\{70, 4\}$	$0, -0.1, -1$	ILUTP, 10^{-3}	10^{-13}
aerofoilB	23560	20	$\{70, 5\}$	$0, -1, -10$	ILUTP, 10^{-3}	10^{-12}
*matRE*500*C*	22385	20	$\{70, 6\}$	$-1, -5, -10$	ILUTP, 10^{-3}	10^{-16}

TABLE 2 Performance of the exact and the inexact unrestarted RKSM

	Time (s)		GMRES iteration counts		
Problem	Exact	Inexact	Exact	Inexact	RSKM steps
*epb*1	9.10	2.88	2383	831	61
*Goodwin*054	36.71	22.92	3625	2231	64
big	11.04	4.18	2585	914	49

TABLE 3 Performance of the exact and the inexact restarted RKSM

	Time (s)		GMRES iteration counts		
Problem	Exact	Inexact	Exact	Inexact	RSKM steps
*poli*3	46.03	8.91	11689	2015	191
aerofoilB	59.83	37.31	6142	3790	225
*matRE*500*C*	766.21	602.81	19927	10375	273

approximation subspace M, the maximum restarted cycle J($J = 1$ for unrestarted RKSM), cyclic poles for RKSM, and the residual tolerance.

We use the right-sided preconditioned GMRES(m) as the inner linear solver of RKSM for nonsymmetric linear systems. The maximum dimension of the subspace for GMRES is set to be $m_{\text{inner}} = 70$ and the maximum number of GMRES restart cycles is set to be $J_{\text{inner}} = 20$. We use incomplete LU factorization with threshold and pivoting (ILUTP) preconditioners[34(section 10.4.4, p. 327)], and Table 1 also reports the drop tolerance for ILUTP preconditioners. To simulate the behavior of the "*exact*" RKSM with iterative inner linear solves, we set the tolerance of GMRES to be 10 times smaller than the bound $\frac{\epsilon}{m}$ in (35) and $\frac{\epsilon}{JM}$ in (53). A tuned preconditioner is used reduce the number of inner iterations in different eigenvalue algorithms; see, for example, References 17,29,35,36. We have tested the tuned versions of ILU preconditioners based on these references, and found that it does not decrease the number of inner iterations in general. Therefore, we did not use tuned preconditioners in the following tests.

For the first three problems, we use the exact and the inexact unrestarted RKSM, and record runtime and the total number of GMRES iterations for all methods. The results are summarized in Table 2.

For the last three problems, we use the exact and the inexact restarted RKSM, and the results are summarized in Table 3.

From the results in Tables 2 and 3, inexact methods need less time to converge, because they require less accurate GMRES linear solves and hence fewer GMRES steps. If the direct solve of linear system is costly, it would be better to use inexact RKSM for eigenvalue computation.

To understand how inexact methods save the computation time, we take an example of *Goodwin054*. The convergence of both methods in this example is shown in Figure 5. We can see that as we relax the tolerance of inner linear system solves with outer iteration progress, the number of GMRES iteration decreases. In Table 4, we

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

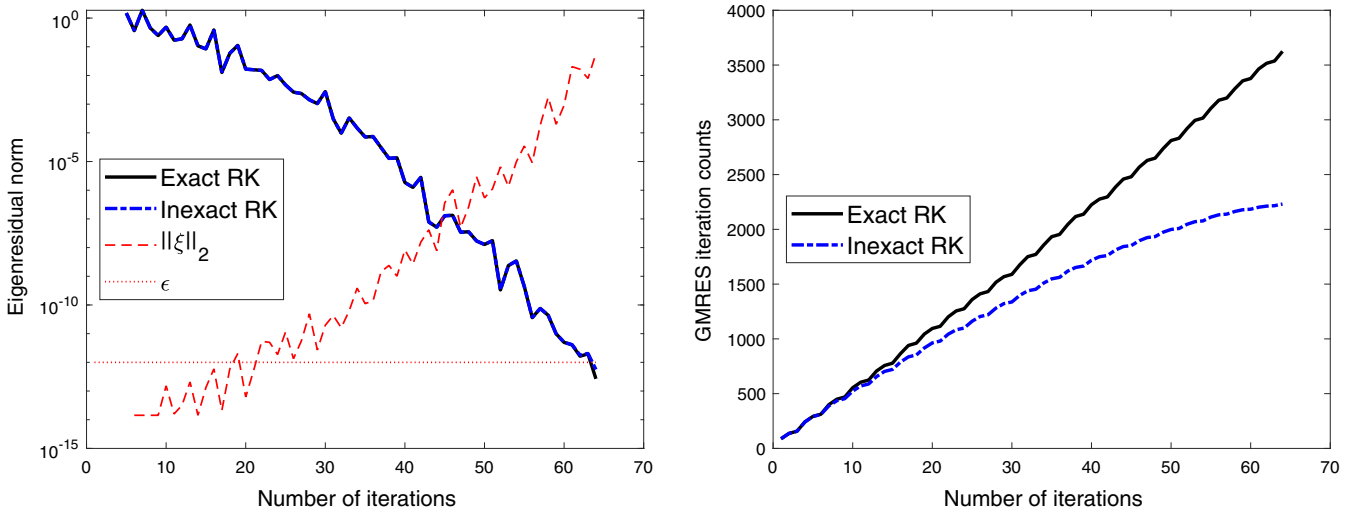

FIGURE 5 Performance of exact and inexact RKSM for *Goodwin054* problem

TABLE 4 Itemized computation time (s) for *Goodwin054*

Item	Exact	Inexact
Preconditioner construction	2.66	2.67
GMRES	33.00	19.23
Orthogonalization (outer step)	0.26	0.24
Total time	36.71	22.92

record the computation time for the processes of constructing preconditioners, applying GMRES, and orthogonalization for RKSM new basis vectors, all of which are majority of time consumption processes for RKSM. We can see that in this example, applying GMRES takes majority of time. With the inexact method, the total computation time is saved significantly. In this example, inexact RKSM only take 62.42% of the time used by the exact method.

6 | CONCLUSION

In this article, we studied inexact RKSM and inexact restarted RKSM for eigenvalue computation. For large-scale problems, errors are introduced by iterative solutions of the inner linear systems at each RKSM step to enlarge the rational Krylov subspaces. We reviewed the invariant subspace perturbation result and derived a theoretical upper bound on the norm of allowable errors in the shift-invert matrix-vector product at each RKSM step in an effort to keep the convergence behavior of inexact RKSM similar to that of exact RKSM. Since the theoretical bound is inversely proportional to the current eigenresidual norm, it is possible to relax the tolerance of inner linear system solves with the outer iteration progress. Numerical experiments show that inexact methods have similar convergence to exact methods, but the former entails lower computational cost thanks to the relaxed accuracy for solving the inner linear systems.

CONFLICT OF INTEREST

This study does not have any conflicts to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ORCID

Shengjie Xu https://orcid.org/0000-0003-2025-4043
Fei Xue https://orcid.org/0000-0002-2491-9359

REFERENCES

1. Lanczos C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J Res Nat Bur Stand. 1950;45:255–82.
2. Saad Y. Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices. Linear Algebra Appl. 1980;34:269–95.
3. Sorensen DC. Implicit application of polynomial filters in a *k*-step Arnoldi method. SIAM J Matrix Anal Appl. 1992;13(1):357–85.
4. Wu K, Simon H. Thick-restart Lanczos method for large symmetric eigenvalue problems. SIAM J Matrix Anal Appl. 2000;22(2):602–16.
5. Stewart GW. A Krylov-Schur algorithm for large eigenproblems. SIAM J Matrix Anal Appl. 2001;23(3):601–14.
6. Bai Z, Demmel J, Dongarra J, Ruhe A, van der Vorst H, editors. Templates for the solution of algebraic eigenvalue problems: a practical guide. Volume 11 of software, environments, and tools. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM); 2000.
7. Ruhe A. Rational Krylov sequence methods for eigenvalue computation. Linear Algebra Appl. 1984;58:391–405.
8. Druskin V, Knizhnerman L, Simoncini V. Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation. SIAM J Numer Anal. 2011;49(5):1875–98.
9. Druskin V, Simoncini V. Adaptive rational Krylov subspaces for large-scale dynamical systems. Systems Control Lett. 2011;60(8):546–60.
10. Benner P, Saak J. Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey. GAMM-Mitt. 2013;36(1):32–52.
11. Güttel S. Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection. GAMM-Mitt. 2013;36(1):8–31.
12. Druskin V, Lieberman C, Zaslavsky M. On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. SIAM J Sci Comput. 2010;32(5):2485–96.
13. Jarlebring E, Voss H. Rational Krylov for nonlinear eigenproblems, an iterative projection method. Appl Math. 2005;50(6):543–54.
14. Güttel S, Van Beeumen R, Meerbergen K, Michiels W. NLEIGS: a class of fully rational Krylov methods for nonlinear eigenvalue problems. SIAM J Sci Comput. 2014;36(6):A2842–64.
15. Van Beeumen R, Meerbergen K, Michiels W. Compact rational Krylov methods for nonlinear eigenvalue problems. SIAM J Matrix Anal Appl. 2015;36(2):820–38.
16. Simoncini V. Variable accuracy of matrix-vector products in projection methods for eigencomputation. SIAM J Numer Anal. 2005;43(3):1155–74.
17. Freitag MA, Spence A. Shift-invert Arnoldi's method with preconditioned iterative solves. SIAM J Matrix Anal Appl. 2009;31(3):942–69.
18. Lehoucq RB, Meerbergen K. Using generalized Cayley transformations within an inexact rational Krylov sequence method. SIAM J Matrix Anal Appl. 1999;20(1):131–48.
19. Kürschner P, Freitag MA. Inexact methods for the low rank solution to large scale Lyapunov equations. BIT. 2020;60(4):1221–59.
20. Van Beeumen R, Meerbergen K, Michiels W. A rational Krylov method based on Hermite interpolation for nonlinear eigenvalue problems. SIAM J Sci Comput. 2013;35(1):A327–50.
21. Ruhe A. Rational Krylov: a practical algorithm for large sparse nonsymmetric matrix pencils. SIAM J Sci Comput. 1998;19(5):1535–51.
22. Ruhe A. Rational Krylov algorithms for nonsymmetric eigenvalue problems. II. Matrix pairs. Linear Algebra Appl. 1994;197:283–95.
23. Güttel S. Rational Krylov methods for operator functions [PhD thesis]. Technische Universität Bergakademie Freiberg; 2010.
24. Golub GH, Van Loan CF. Matrix computations. 3rd ed. Baltimore, Maryland: Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press; 1996.
25. Demmel J, Kågström B. The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: robust software with error bounds and applications. I. Theory and algorithms. ACM Trans Math Softw. 1993;19(2):160–74.
26. Stewart GW, Sun JG. Matrix perturbation theory. Boston, MA: Computer Science and Scientific Computing. Academic Press, Inc; 1990.
27. Stewart GW. On the sensitivity of the eigenvalue problem Ax = λBx. SIAM J Numer Anal. 1972;9:669–86.
28. Stewart GW. Matrix algorithms. Volume II. Eigensystems. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM); 2001.
29. Xue F, Elman HC. Fast inexact implicitly restarted Arnoldi method for generalized eigenvalue problems with spectral transformation. SIAM J Matrix Anal Appl. 2012;33(2):433–59.
30. Mehrmann V, Schröder C, Simoncini V. An implicitly-restarted Krylov subspace method for real symmetric/skew-symmetric eigenproblems. Linear Algebra Appl. 2012;436(10):4070–87.
31. Sorensen DC. Implicitly restarted Arnoldi/Lanczos methods for large scale eigenvalue calculations. In: DE Keyes, A Sameh & V Venkatakrishnan, eds. Parallel numerical algorithms (Hampton, VA, 1994). ICASE/LaRC Interdisciplinary Series in Science and Engineering. Volume 4. Dordrecht: Kluwer Academic Publishers; 1997. p. 119–65.
32. Bhatia R. Matrix analysis. Graduate Texts in Mathematics. Vol 169. New York, NY: Springer-Verlag; 1997.
33. Davis TA, Hu Y. The university of Florida sparse matrix collection. ACM Trans Math Softw. 2011;38(1):1:1–1:25.
34. Saad Y. Iterative methods for sparse linear systems. 2nd ed. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2003.
35. Freitag MA, Spence A. A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems. IMA J Numer Anal. 2008;28(3):522–51.

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

36. Robbé M, Sadkane M, Spence A. Inexact inverse subspace iteration with preconditioning applied to non-Hermitian eigenvalue problems. SIAM J Matrix Anal Appl. 2009;31(1):92–113.

How to cite this article: Xu S, Xue F. Inexact rational Krylov subspace method for eigenvalue problems. Numer Linear Algebra Appl. 2022;29(5):e2437. https://doi.org/10.1002/nla.2437

10991506, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/nla.2437, Wiley Online Library on [29/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License