Local convergence analysis of several inexact Newton-type algorithms for general nonlinear eigenvalue problems

Fei Xue and Daniel B. Szyld

Report 11-08-09
August 2011

This report is available in the World Wide Web at
http://www.math.temple.edu/~szyld

LOCAL CONVERGENCE ANALYSIS OF SEVERAL INEXACT NEWTON-TYPE ALGORITHMS FOR GENERAL NONLINEAR EIGENVALUE PROBLEMS*

DANIEL B. SZYLD† AND FEI XUE†

Abstract. We study the local convergence of several inexact numerical algorithms closely related to Newton's method for the solution of a simple eigenpair of the general nonlinear eigenvalue problem $T(\lambda)v = 0$. We investigate inverse iteration, Rayleigh quotient iteration, residual inverse iteration, and the single-vector Jacobi-Davidson method, analyzing the impact of the tolerances chosen for the approximate solution of the linear systems arising in these algorithms on the order of the local convergence rates. We show that the inexact algorithms can achieve the same order of convergence as the exact methods if appropriate sequences of tolerances are applied to the inner solves. We discuss the connections and emphasize the differences between the standard inexact Newton's method and these inexact algorithms. When the local symmetry of $T(\lambda)$ is present, the use of a nonlinear Rayleigh functional is shown to be fundamental in achieving higher order of convergence rates. The convergence results are illustrated by numerical experiments.

Key words. nonlinear eigenvalue problems, inexact Newton-type algorithms, order of convergence, Rayleigh functional

AMS subject classifications. 65F15, 15A18, 15A22.

1. Introduction. Nonlinear eigenvalue problems (NEPs) of the form $T(\lambda)v = 0$ arise naturally in a variety of science and engineering applications, such as the dynamic analysis of structures, the optimization of the acoustic emissions of high speed trains, and the solution of optimal control problems; see, for example, the survey [30] and references therein, as well as, e.g., the recent paper [29]. Among all types of NEPs, the quadratic eigenvalue problem (QEP) is of particular interest and has been studied extensively, due to its wide range of applications; see [38] and some recent references [2, 8, 16, 17, 19, 21, 27]. For quadratic, polynomial and rational eigenvalue problems, a commonly-used approach is to transform the original problem to a linear generalized eigenvalue problem of larger size and with identical spectrum by a linearization [1, 15, 26, 36], while preserving the same matrix structure as $T(\lambda)$. For general NEPs, variants of Newton's method constituent a most important class of algorithms for computing a single eigenpair, provided that a good initial eigenpair approximation is available; see [33] for a detailed discussion of a few *exact* algorithms of this type, including analyses of convergence rates. If more eigenpairs are desired, iterative projection methods (with subspace acceleration) such as the nonlinear rational Krylov method [32] and the Jacobi-Davidson (JD) method [7], are generally effective. In this paper, we study *inexact variants* of several algorithms closely related to Newton's method, where the inverse matrix-vector products are performed approximately, for computing one simple eigenpair of large scale general NEPs.

A major difficulty in using Newton's method for solving large nonlinear systems of equations is the prohibitive cost for the solution of a linear system arising at each iteration step. For applications where the matrices involved are very large and sparse, or if these matrices cannot be formed explicitly, this linear solve needs to be performed approximately by some iterative method; see, e.g., [29] for NEPs arising in frequency response problems of this kind. This leads to a variety of inexact eigenvalue algorithms with inner-outer iteration structure. For linear eigenvalue problems, the study of inexact algorithms has attracted considerable attention in the last decade (see, e.g., [5, 14, 40, 41] and references therein), and the un-

*This version dated August 9, 2011. This work was supported by the U.S. Department of Energy under grant DEFG0205ER25672, and the U.S. National Science Foundation under grant DMS-1115520.

†Department of Mathematics, Temple University (038-16), 1805 N. Broad Street, Philadelphia, PA 19122-6094, USA ({szyld,fxue}@temple.edu).

derstanding of the local convergence behavior of these methods has become gradually more mature. Our primary objective in this paper is to extend the study of the local convergence of several inexact Newton-type methods to the solution of general NEPs. One of our contributions is to show that the same order of local convergence can be achieved by the inexact variants as by the exact algorithms. In addition, since some of the methods to be discussed are single-vector versions (without subspace projection and acceleration) of the nonlinear Arnoldi method and the Jacobi-Davidson method, our study also provides some insight into the local convergence of the two iterative projection methods. The global convergence of these methods is generally complicated, and is not treated in this paper.

The rest of the paper is organized as follows. In Section 2, we briefly review the general NEP, some mathematical tools and existing preliminary results for the study of the Newton-type algorithms. In Sections 3 and 4, we present detailed local convergence analyses of the inexact versions of inverse iteration, Rayleigh quotient iteration, residual inverse iteration and the single-vector Jacobi-Davidson method. Numerical experiments are provided in Section 5 to illustrate the convergence analysis. Finally, Section 6 summarizes the results.

2. Preliminaries. In this section, we briefly introduce the problem under consideration, and we outline some preliminaries for the study of the local convergence of several Newton-type methods. Specifically, we review the definition and properties of the nonlinear Rayleigh functional, the derivation of the methods to be studied, a characterization of the resolvent and a measure of the error of eigenvector approximations.

2.1. Problem description. We are interested in computing an algebraically simple eigenpair (λ, v) of the general NEP

$$F(\lambda, v) \equiv \begin{bmatrix} T(\lambda)v \\ u^H v - 1 \end{bmatrix} = 0, \tag{2.1}$$

where $\lambda \in \mathbb{C}$ is an eigenvalue satisfying $\det(T'(\lambda)) \neq 0$, $v \in \mathbb{C}^n \setminus \{0\}$ is the corresponding right eigenvector, $T : \mathbb{C} \to \mathbb{C}^{n\times n}$ is a matrix-valued function that is entry-wise analytic in a neighborhood U of λ, and $u \in \mathbb{C}^n$ is a fixed vector used for the normalization of v. We denote by $\sigma(T)$ the set of all eigenvalues of T in the domain Λ, and by $R(T) = \Lambda \setminus \sigma(T)$ the resolvent set. We assume that $\det(T(\cdot)) \not\equiv 0$ on Λ, i.e., the resolvent set $R(T)$ is not empty. By using this general framework, we can apply our analysis to polynomial, rational and other types of NEPs. We prefer this approach to one in which only one class of problems is studied, since the analysis is the same for all problem classes.

2.2. Nonlinear Rayleigh functional. The nonlinear Rayleigh functional is a conceptually straightforward generalization of the Rayleigh quotient to the setting of NEPs. Given an approximate eigenvector $x \approx v$, a corresponding auxiliary vector p can be chosen, and the Rayleigh quotient of the linear generalized eigenvalue problem $Av = \lambda Bv$ is defined as $\rho = (p^H Ax)/(p^H Bx)$, i.e., the root of $p^H(A - \rho B)x = 0$. This definition can be directly extended to the NEP (2.1). Given a $x \approx v$, the *Rayleigh functional* $\rho_F(x)$ is the solution of

$$p^H T(\rho)x = 0.$$

If p is a left eigenvector approximation, then $\rho_F(x)$ is a *two-sided* Rayleigh functional; otherwise, it is a *one-sided* Rayleigh functional; see [34]. As we will see in Sections 3 and 4, the use of the two-sided Rayleigh functional is critical for Rayleigh quotient iteration and the single-vector Jacobi-Davidson method to achieve at least cubic local convergence.

For a large class of NEPs $T(\lambda)v = 0$, the two-sided Rayleigh functional can be computed without additional cost of solving a conjugate system for a left eigenvector approximation.

These problems are the ones where the local symmetry of $T(\lambda)$ is present, that is, if the matrix $T(\lambda)$ (instead of the matrix pencil $T(\cdot)$) is real (skew) symmetric, complex (skew) Hermitian, or complex (skew) symmetric. In the first two cases, the corresponding left eigenvector is v^H; in the third case, the left eigenvector is $v^T = \bar{v}^H$. Therefore, if the local symmetry is present, one may choose

$$\begin{aligned} &p = x \quad \text{if } T(\lambda) \text{ is real (skew) symmetric or (skew) Hermitian, or} \\ &p = \bar{x} \quad \text{if } T(\lambda) \text{ is complex (skew) symmetric.} \end{aligned} \tag{2.2}$$

The definitions and properties of several variants of nonlinear Rayleigh functionals are discussed in detail in [34]. The following theorem shows the existence and uniqueness of $\rho_F(x)$, and the difference between $\rho_F(x)$ and λ.

THEOREM 2.1. [34, Corollary 18 and Theorem 21] *Assume that (λ, v) is a simple eigenpair of $T(\cdot) : \mathbb{C} \to \mathbb{C}^{n\times n}$, and $T'(\cdot)$ is Lipschitz continuous in $\overline{B(\lambda, \tau_0)} \equiv \{\rho : |\rho - \lambda| \le \tau_0\}$ for some constant τ_0. Suppose that $(T'(\lambda)v, v) \neq 0$. Then there exists an angle $\epsilon_0 < \pi/2$, such that for all $x \in \mathbb{C}^n \setminus \{0\}$ with $\angle(x, v) \le \epsilon_0$, there is a unique scalar $\rho_F(x) \in \overline{B(\lambda, \tau_0)}$ satisfying $x^H T(\rho_F(x))\, x = 0$, and*

$$|\rho_F(x) - \lambda| \le \frac{10}{3} \frac{\|T(\lambda)\| \|v\|^2}{|v^H T'(\lambda) v|} \tan \angle(x, v). \tag{2.3}$$

If the local symmetry is present, and p is chosen as in (2.2), there is a unique $\rho_F(x) \in \overline{B(\lambda, \tau_0)}$ satisfying $p^H T(\rho_F(x))x = 0$, and

$$|\rho_F(x) - \lambda| \le \frac{8}{3} \frac{\|T(\lambda)\| \|v\|^2}{|v^H T'(\lambda) v|} \tan^2 \angle(x, v). \tag{2.4}$$

2.3. Iteration schemes. To make the exposition self-contained, we briefly review several iteration schemes based on linearizations (first order approximation) of (2.1), from which variants of Newton-type methods can be derived; see [33] for details. Since $T(\cdot)$ is analytic in a neighborhood U of λ, the Taylor expansion of $T(\lambda)$ at $\mu \in U$ gives

$$\begin{aligned} T(\lambda)v = {} & T(\mu)x + T(\mu)(v - x) + (\lambda - \mu)T'(\mu)x \quad \text{(0th and 1st order terms)} \\ & + (\lambda - \mu)T'(\mu)(v - x) + \frac{(\lambda - \mu)^2}{2} T''(\mu)x \quad \text{(2nd order terms)} \\ & + \mathcal{O}\left((\lambda - \mu)^2 (v - x) + (\lambda - \mu)^3\right). \quad \text{(higher order terms)} \end{aligned} \tag{2.5}$$

Linearization of (2.5) leads to $T(\mu)v \approx (\mu - \lambda)T'(\mu)x$. Replacing (μ, x) with $(\mu^{(i)}, x^{(i)})$, and (λ, v) with $(\mu^{(i+1)}, x^{(i+1)})$, we have the iteration scheme

$$\begin{cases} y^{(i)} = T(\mu^{(i)})^{-1} T'(\mu^{(i)}) x^{(i)} \\ x^{(i+1)} = y^{(i)} / (u^H y^{(i)}) \\ \mu^{(i+1)} = \mu^{(i)} - (u^H x^{(i+1)}) / (u^H y^{(i)}). \end{cases} \tag{2.6}$$

A different scheme can be derived from the same linearization as follows

$$\begin{aligned} v - x &\approx -(\lambda - \mu) T(\mu)^{-1} T'(\mu) x - x \\ &= -T(\mu)^{-1} \left(T(\mu) + (\lambda - \mu) T'(\mu)\right) x \\ &= -T(\mu)^{-1} T(\lambda) x + \mathcal{O}(|\lambda - \mu|^2), \end{aligned} \tag{2.7}$$

where the last equation is obtained from first order Taylor expansion of $T(\lambda)$). Let $\sigma \approx \lambda$ be a fixed shift, and $\rho^{(i)} = \rho_F(x^{(i)})$ the Rayleigh functional. Replacing (μ, x) and (λ, v), respectively, with $(\sigma, x^{(i)})$ and $(\rho^{(i)}, x^{(i+1)})$, we have

$$\begin{cases} T(\sigma)\Delta x^{(i)} = -T(\rho^{(i)})x^{(i)} \\ x^{(i+1)} = x^{(i)} + \Delta x^{(i)} \end{cases}.$$

The third scheme is based on the linearization of (2.5) and the normalization condition $u^H x = u^H v = 1$:

$$\begin{bmatrix} T(\mu) & T'(\mu)x \\ u^H & 0 \end{bmatrix} \begin{bmatrix} v - x \\ \lambda - \mu \end{bmatrix} \approx \begin{bmatrix} -T(\mu)x \\ 0 \end{bmatrix}.$$

One can replace (μ, x) and (λ, v) with $(\mu^{(i)}, x^{(i)})$ and $(\mu^{(i+1)}, x^{(i+1)})$, respectively, and zero the term $(\mu^{(i+1)} - \mu^{(i)})T'(\mu^{(i)})x^{(i)}$ arising in the first block equation by a projector $\Pi_1^{(i)}$ with $\ker(\Pi_1^{(i)}) = \operatorname{span}\{T'(\mu^{(i)})x^{(i)}\}$. In addition, we can incorporate the orthogonality condition $u^{(i)H}(x^{(i+1)} - x^{(i)}) = 0$ into the first block equation by imposing a second projector $\Pi_2^{(i)}$ with $\operatorname{range}(\Pi_2^{(i)}) = \operatorname{span}\{u^{(i)}\}^\perp$. Note that we allow the use of a variable vector $u^{(i)}$ for the normalization of $x^{(i)}$ in the ith iteration. The resulting scheme is

$$\begin{cases} \Pi_1^{(i)} T(\mu^{(i)}) \Pi_2^{(i)} \Delta x^{(i)} = -\left(T(\mu^{(i)}) - \chi^{(i)} T'(\mu^{(i)})\right) x^{(i)} \text{ with } \Delta x^{(i)} \perp u^{(i)} \\ x^{(i+1)} = x^{(i)} + \Delta x^{(i)} \end{cases}, \tag{2.8}$$

where $\Pi_1^{(i)} = I - \dfrac{T'(\mu^{(i)})x^{(i)}p^{(i)H}}{p^{(i)H}T'(\mu^{(i)})x^{(i)}}$, $\Pi_2^{(i)} = I - \dfrac{x^{(i)}u^{(i)H}}{u^{(i)H}x^{(i)}}$, $\chi^{(i)} = \dfrac{p^{(i)H}T(\mu^{(i)})x^{(i)}}{p^{(i)H}T'(\mu^{(i)})x^{(i)}}$, and $p^{(i)}$ is a vector satisfying $p^{(i)H}T'(\mu^{(i)})x^{(i)} \neq 0$. If $\mu^{(i)} = \rho_F(x^{(i)})$ such that $p^{(i)H}T(\mu^{(i)})x^{(i)} = 0$, the right-hand side of the equation in (2.8) can be simplified as $-T(\mu^{(i)})x^{(i)}$.

2.4. The resolvent and the decomposition of an eigenvector approximation. For inexact Rayleigh quotient iteration (IRQI) and the single-vector Jacobi-Davidson (JD) method, the properties of the resolvent $T(\mu)^{-1}$ near the pole $\mu = \lambda$ plays a critical role in the local convergence analysis. The following theorem characterizes the resolvent.

THEOREM 2.2. [25, Theorem A.10.1] *Let Λ be an open subset of $\mathbb{C}$ and $T : \Lambda \to \mathbb{C}^{n\times n}$ be an analytic matrix-valued function with $\det(T(\cdot)) \not\equiv 0$. Then, every eigenvalue λ of T is isolated, i.e., there is some neighborhood U of λ such that $U\setminus\{\lambda\} \subset R(T)$. For an algebraically simple eigenpair (λ, v), there exists a unique left eigenvector $w \in \ker\left(T(\lambda)^H\right)$ such that*

$$T(\mu)^{-1} = \frac{1}{\mu - \lambda} v w^H + F(\mu), \qquad \text{for all } \mu \in U \setminus \{\lambda\}, \tag{2.9}$$

where $F : U \to \mathbb{C}^{n\times n}$ is analytic. Moreoever, the following relations hold:

$$(T'(\lambda)v, w) = 1, \quad \text{and} \tag{2.10}$$

$$T(\lambda)F(\lambda)T'(\lambda)v = 0. \tag{2.11}$$

Note that $\|w\|$ is implicitly determined by the conditions $(T'(\lambda)v, w) = 1$ and $u^H v = 1$. With the left eigenvector w, we propose a decomposition of an eigenvector approximation x that is close to v in direction. Let $W_{n-1} \in \mathbb{C}^{n\times(n-1)}$ be a matrix whose columns are

orthonormal vectors that form a basis of the *orthogonal complement* of $\mathrm{span}\{T'(\lambda)v\}$. Then for a given nonzero vector $x \notin \mathrm{span}\{v\}$, define

$$\gamma = \left\| \begin{bmatrix} w^H \\ W_{n-1}^H \end{bmatrix} T'(\lambda)x \right\|, \quad s = \frac{\|W_{n-1}^H T'(\lambda)x\|}{\gamma}, \quad c = \frac{w^H T'(\lambda)x}{\gamma}, \tag{2.12}$$

where γ is a *generalized norm* of x, and $|s|^2 + |c|^2 = 1$. To decompose x in terms of v and another direction, define the vector

$$g = \frac{1}{s}\left(\frac{1}{\gamma}x - cv\right). \tag{2.13}$$

It can be shown that the vector g defined above has the following properties:

$$w^H T'(\lambda)g = 0 \quad \text{and} \quad \|W_{n-1}^H T'(\lambda)g\| = 1. \tag{2.14}$$

Therefore $\|g\|$ is bounded as $\dfrac{1}{\sigma_{max}(W_{n-1}^H T'(\lambda))} \le \|g\| \le \dfrac{1}{\sigma_{min}(W_{n-1}^H T'(\lambda))}$. With g defined in (2.13), x can be decomposed as

$$x = \gamma(cv + sg). \tag{2.15}$$

That is, s and c can be interpreted as generalized sine and cosine of $\angle(x, v)$, respectively. The generalized tangent, therefore, can be defined as $t = s/c$.

The generalized sine and tangent can be used to measure the error of x as an eigenvector approximation. In fact, under some mild assumptions (as discussed below), it can be shown that x converges to v in direction if and only if $t \to 0$. To see this, it is enough to show that t is proportional to $\tan\angle(x, v)$. Without loss of generality, suppose that $\gamma = 1$, and consider a ball centered at the terminal point of cv of radius $\|sg\|$. Suppose that s is small such that $\|sg\| < \|cv\|$. Since the terminal point of x is on the boundary of the ball, it is not hard to see from elementary geometry that the maximum value of $\angle(x, v)$ is achieved if and only if x is tangent to the boundary of the ball, i.e., $x \perp sg$. Therefore,

$$\sin\angle(x, v) \le \frac{\|sg\|}{\|cv\|} = t\frac{\|g\|}{\|v\|}. \tag{2.16}$$

Suppose that $\|g\|$ is a moderate multiple of $\|v\|$. If $|s| \ll |c|$, then

$$\tan\angle(x, v) = \frac{\|sg_\perp\|}{\|cv + sg_\parallel\|} = \frac{\|(s\sin\varphi)g\|}{\|cv + (s\cos\varphi)g\|} \approx t\frac{\|g\|}{\|v\|}\sin\varphi, \tag{2.17}$$

where $g_\parallel$ is the orthogonal projection of g onto v, $g_\perp \equiv g - g_\parallel$ with $g_\perp \perp v$, and $\varphi \equiv \angle(v, g)$. It follows from (2.16) and (2.17) that if $\varphi = \angle(v, g)$ is not very small, there exist some modest constants C_1 and C_2 independent of t, such that

$$C_1 t \le \tan\angle(x, v) \le C_2 t \tag{2.18}$$

for all sufficiently small $\angle(x, v)$. The equivalence of the regular and the generalized tangent of $\angle(x, v)$ shows that both s and t can be used as error estimates of the eigenvector approximation. In Section 3.2, we will use the decomposition (2.15) and the property of the resolvent $T^{-1}(\mu)$ (Theorem 2.2) to analyze the local convergence of IRQI.

3. Variants of inexact inverse iteration. In this section, we present a local convergence analysis of the standard inexact inverse iteration and inexact Rayleigh quotient iteration (IRQI). We begin with the description of the two algorithms.

In the scheme (2.6), suppose that $T(\mu^{(i)})^{-1}T'(\mu^{(i)})x^{(i)}$ is computed approximately. We have the standard inexact inverse iteration as follows.

ALGORITHM 3.1. **Standard inexact inverse iteration**

Start with $(\mu^{(0)}, x^{(0)})$ with $u^H x^{(0)} = 1$ for some fixed $u \in \mathbb{C}^n$

For $i = 0, 1, 2, ...,$ until convergence

1. Choose a relative tolerance $\tau^{(i)}$, and solve $T(\mu^{(i)})y^{(i)} = T'(\mu^{(i)})x^{(i)}$ approximately such that the residual $res^{(i)} \equiv T'(\mu^{(i)})x^{(i)} - T(\mu^{(i)})y^{(i)}$ satisfies $\|res^{(i)}\| \leq \tau^{(i)}\|T'(\mu^{(i)})x^{(i)}\|$
2. Normalize $x^{(i+1)} = y^{(i)}/(u^H y^{(i)})$
3. $\mu^{(i+1)} = \mu^{(i)} - 1/(u^H y^{(i)})$ and test for convergence.

End For

In Step 3, Algorithm 3.1 evaluates the new eigenvalue approximation directly derived from the linearized equation $T(\mu)v \approx (\mu - \lambda)T'(\mu)x$. This new approximation can be replaced by some alternative values. In addition, if the new eigenvalue approximation does not depend on the normalization vector u, then x can be normalized in any appropriate manner. The resulting flexible algorithm is as follows:

ALGORITHM 3.2. **Flexible inexact inverse iteration**

Start with $(\mu^{(0)}, x^{(0)})$ with normalized $x^{(0)}$, and choose $\mu_1^{(0)}$ and $\mu_2^{(0)}$ based on $(\mu^{(0)}, x^{(0)})$

For $i = 0, 1, 2, ...,$ until convergence

1. Choose a relative tolerance $\tau^{(i)}$, and solve $T(\mu_1^{(i)})y^{(i)} = T'(\mu_2^{(i)})x^{(i)}$ approximately such that the residual $res^{(i)} \equiv T'(\mu_2^{(i)})x^{(i)} - T(\mu_1^{(i)})y^{(i)}$ satisfies $\|res^{(i)}\| \leq \tau^{(i)}\|T'(\mu_2^{(i)})x^{(i)}\|$
2. Normalize $y^{(i)}$ into $x^{(i+1)}$
3. Compute $\mu_1^{(i+1)}$ and $\mu_2^{(i+1)}$ based on $x^{(i+1)}$, and test for convergence

End For

$\mu^{(i+1)} = \mu_1^{(i+1)}$ or $\mu_2^{(i+1)}$, and $x^{(i+1)} = x^{(i+1)}/(u^H x^{(i+1)})$.

The difference between the standard and the flexible versions lies in the computation of the new eigenvalue approximation and the choice of $\mu_1^{(i)}$ and $\mu_2^{(i)}$. In particular, the flexible inexact inverse iteration with $\mu_1^{(i)} = \mu_2^{(i)} = \rho_F(x^{(i)})$ is called the inexact Rayleigh quotient iteration (IRQI). We will see that the use of the Rayleigh functional is critical for IRQI to achieve higher order convergence if the local symmetry of $T(\lambda)$ is present. As a result, different tools are needed to analyze the local convergence of the two algorithms. Specifically, we will show that the standard version is essentially a modified Newton's method, whereas IRQI may be interpreted as a Newton's method enhanced with more accurate eigenvalue approximation, which can be analyzed by eigenvector decompositions and the property of the resolvent.

3.1. The standard inexact inverse iteration. In this section, we give a local convergence analysis of the standard inexact inverse iteration (Algorithm 3.1). To this end, we need the following Theorem on the convergence of a modified Newton's method [10, Ch. 5].

THEOREM 3.3. *Let $F(z) : \mathbb{C}^m \to \mathbb{C}^m$ be a differentiable function, z_* a simple root of $F(z)$, $\eta > 0$ an appropriate constant, and J_F the Jacobian of F such that $\|J_F(z_*)^{-1}\| \leq \eta$. Suppose that $\|J_F(z_1) - J_F(z_2)\| \leq \kappa\|z_1 - z_2\|$ for all $z_1, z_2 \in B(z_*, r)$ for some $r > 0$. Let $\widetilde{J}(z)$ be an approximation to J_F, such that for all $z \in B(z_*, r)$, $\left\|J_F(z_*)^{-1}\left(\widetilde{J}(z) - J_F(z_*)\right)\right\| \leq \delta$*

uniformly for some $\delta \in [0,1)$. *Then* $\widetilde{J}(z)^{-1}$ *exists in* $B(z_*,r)$, *and* $\|\widetilde{J}(z)^{-1}\| \le \dfrac{\eta}{1-\delta}$. *If* $\dfrac{\kappa\eta r}{2(1-\delta)} + \delta < 1$, *then the modified Newton's method*

$$z^{(i+1)} = z^{(i)} - \widetilde{J}(z^{(i)})^{-1}F(z^{(i)}) \quad \text{with } z^{(0)} \in B(z_*,r)$$

converges to z_* *at least linearly with the error bound*

$$\|e^{(i+1)}\| \le \frac{\eta}{1-\delta}\left(\frac{\kappa}{2}\|e^{(i)}\| + \|J_F(z^{(i)}) - \widetilde{J}(z^{(i)})\|\right)\|e^{(i)}\|,$$

where $e^{(i)} = z^{(i)} - z_*$. *If, in addition,* $\left\|\left(J_F(z^{(i)}) - \widetilde{J}(z^{(i)})\right)e^{(i)}\right\| \le C\|e^{(i)}\|^2$ *for some constant* C *independent of* i, *then the modified Newton's method converges at least quadratically.*

To simplify the notation in using Theorem 3.3 to study Algorithm 3.1, let $z = \begin{bmatrix} x \\ \mu \end{bmatrix}$, $z_* = \begin{bmatrix} v \\ \lambda \end{bmatrix}$ and $\Delta z^{(i)} = \begin{bmatrix} \Delta x^{(i)} \\ \Delta\mu^{(i)} \end{bmatrix}$. To prepare for the complete proof, we first show that the Jacobian of F defined in (2.1) satisfies $\|J_F(z_*)^{-1}\| < \eta$ for some η, i.e., $J_F(z_*)$ is nonsingular.

LEMMA 3.4. *If* (λ, v) *is a simple eigenpair of* (2.1), *then* $J_F(z_*) = \begin{bmatrix} T(\lambda)v & T'(\lambda)v \\ u^H & 0 \end{bmatrix}$ *is nonsingular.*

Proof. Since λ is an algebraically simple root of $\det(T(\mu)) = 0$, Proposition 1 in [31] shows that $\operatorname{rank}(T(\lambda)) = n-1$ and $w^H T'(\lambda)v \neq 0$. Lemma 2.8 in [24] shows that if, in addition, $u^H v = 1$, then $J_F(z_*)$ is nonsingular. □

We are now ready to show our result on the convergence of Algorithm 3.1. The idea of this proof comes from [12, Theorem 3.1].

THEOREM 3.5. [Local convergence of the standard inexact inverse iteration] *Let* (λ, v) *be a simple eigenpair of the* NEP (2.1). *Let* $\eta > 0$, *such that* $\|J_F(z_*)\| \le \eta$. *Then, for some sufficiently small* r *and* τ_{max}, $z^{(0)} \in B(z_*,r)$ *and* $\tau^{(i)} \le \tau_{max}$, Algorithm 3.1 *converges to* $z_* = \begin{bmatrix} v \\ \lambda \end{bmatrix}$ *at least linearly. If* $\tau^{(i)} \le C\|e^{(i)}\| \le \tau_{max}$, *where* C *is a constant independent of* i, *then* Algorithm 3.1 *converges at least quadratically.*

Proof. By assumption, $T(\mu)$ is analytic in a neighborhood U of λ. Therefore, there exists a small $r > 0$ such that $\overline{B(\lambda,r)} \subset U \subset \Lambda$, and all the derivatives of $T(\mu)$ are bounded in $\overline{B(\lambda,r)}$. Given such r, consider any $z_1, z_2 \in B(z_*,r) \subset \mathbb{C}^{n+1}$, such that $\mu_j = e_{n+1}^T z_j \in B(\lambda,r)$ $(j = 1,2)$. Since $J_F(z) = \begin{bmatrix} T(\mu) & T'(\mu)x \\ u^H & 0 \end{bmatrix}$, we have

$$\begin{aligned}
\|J_F(z_1) - J_F(z_2)\| &= \left\|\begin{bmatrix} T(\mu_1) - T(\mu_2) & T'(\mu_1)x_1 - T'(\mu_2)x_2 \\ 0 & 0 \end{bmatrix}\right\| \\
&\le \|T(\mu_1) - T(\mu_2)\| + \|T'(\mu_1)x_1 - T'(\mu_2)x_2\| \\
&\le \|T(\mu_1) - T(\mu_2)\| + \|T'(\mu_2)(x_1 - x_2)\| + \|\left(T'(\mu_1) - T'(\mu_2)\right)x_1\| \\
&= |\mu_1 - \mu_2|\|T'(\mu_2)\| + \|T'(\mu_2)(x_1 - x_2)\| + |\mu_1 - \mu_2|\|T''(\mu_2)x_1\| + \mathcal{O}(|\mu_1 - \mu_2|^2) \\
&\le (2\|T'(\mu_2)\| + \|T''(\mu_2)x_1\|)\|z_1 - z_2\| + \mathcal{O}(|\mu_1 - \mu_2|^2) \le \kappa\|z_1 - z_2\|.
\end{aligned}$$

In other words, there exists a constant κ such that the Lipschitz condition $\|J_F(z_1) - J_F(z_2)\| \le \kappa\|z_1 - z_2\|$ holds for all $z_1, z_2 \in B(z_*,r)$.

To define the modified Jacobian $\widetilde{J}(z)$ as an approximation to $J_F(z)$, note from Step 3 in Algorithm 3.1 that $u^H x^{(i)} = 1$ for all i. Therefore, we have

$$x^{(i+1)} = \frac{1}{u^H y^{(i)}} y^{(i)} = (\mu^{(i)} - \mu^{(i+1)}) y^{(i)}.$$

Pre-multiplying both sides by $T(\mu^{(i)})$, and noting that $x^{(i+1)} = x^{(i)} + \Delta x^{(i)}$, we have

$$T(\mu^{(i)})\Delta x^{(i)} + \Delta\mu^{(i)}(T'(\mu^{(i)})x^{(i)} - res^{(i)}) = -T(\mu^{(i)})x^{(i)}.$$

Since $u^H \Delta x^{(i)} = u^H x^{(i+1)} - u^H x^{(i)} = 0$, we have

$$\widetilde{J}(z^{(i)})\Delta z^{(i)} \equiv \begin{bmatrix} T(\mu^{(i)}) & T'(\mu^{(i)})x^{(i)} - res^{(i)} \\ u^H & 0 \end{bmatrix} \begin{bmatrix} \Delta x^{(i)} \\ \Delta\mu^{(i)} \end{bmatrix} = \begin{bmatrix} -T(\mu^{(i)})x^{(i)} \\ 0 \end{bmatrix}.$$

To simplify the notation for the proof, we drop the superscripts (i). This omission should not cause any confusion. It follows that for any $z \in B(z_*, r)$,

$$\begin{aligned}
\|\widetilde{J}(z) - J_F(z_*)\| &= \left\| \begin{bmatrix} T(\mu) - T(\lambda) & T'(\mu)x - T'(\lambda)v - res \\ 0 & 0 \end{bmatrix} \right\| \\
&\le \|T(\mu) - T(\lambda)\| + \|T'(\mu)x - T'(\lambda)v - res\| \\
&\le |\mu - \lambda|\|T'(\lambda)\| + \|T'(\lambda)\gamma(cv + sg) + (\mu - \lambda)T''(\lambda)x - T'(\lambda)v\| \\
&\quad + \|res\| + \mathcal{O}(|\mu - \lambda|^2) \\
&\le |\mu - \lambda|(\|T'(\lambda)\| + \|T''(\lambda)x\|) + |\gamma c - 1|\|T'(\lambda)v\| + \gamma s\|T'(\lambda)g\| \\
&\quad + \tau_{max}\|T'(\mu)x\| + \mathcal{O}(|\mu - \lambda|^2).
\end{aligned}$$

Note that if r is sufficiently small, then for any $z = \begin{bmatrix} x \\ \mu \end{bmatrix} \in B(z_*, r)$, $|\mu - \lambda|$ must also be sufficiently small; in addition, since $u^H x = \gamma(c + s(u^H g)) = 1$, we see that $|\gamma c - 1|$ must be sufficiently small for all x sufficiently close to v in direction (for which s is small enough).

Therefore, for a given $\delta \in [0, 1)$, there exist some sufficiently small r and τ_{max}, such that for all $z \in B(z_*, r)$ and $\tau \le \tau_{max}$, $\|\widetilde{J}(z) - J_F(z_*)\| \le \delta/\eta$ holds. It follows that $\left\| J_F(z_*)^{-1}\left(\widetilde{J}(z) - J_F(z_*)\right)\right\| \le \|J_F(z_*)^{-1}\|\|\widetilde{J}(z) - J_F(z_*)\| \le \delta$. Moreover, $\dfrac{\kappa\gamma r}{2(1-\delta)} + \delta < 1$ is also satisfied for sufficiently small r. It then follows from Theorem 3.3 that with $z^{(0)} \in B(z_*, r)$ and $\tau \le \tau_{max}$, Algorithm 3.1 converges at least linearly to z_*. In addition, note that

$$\|\widetilde{J}(z) - J_F(z)\| = \left\| \begin{bmatrix} 0 & res \\ 0 & 0 \end{bmatrix} \right\| \le \|res\| \le \tau\|T'(\mu)x\|,$$

and $\|T'(\mu)x\|$ is bounded for all $z \in B(z_*, r)$. Let $e = z - z_* = \begin{bmatrix} x - v \\ \mu - \lambda \end{bmatrix}$. We see that if $\tau \le C\|e\| \le \tau_{max}$ for some constant C independent of $\|e\|$, then $\left\|\left(\widetilde{J}(z) - J_F(z)\right)e\right\| \le C\|e\|^2$, and Algorithm 3.1 with $z^{(0)} \in B(z_*, r)$ converges to z_* at least quadratically. □

3.2. Inexact Rayleigh quotient iteration (IRQI). In this section, we give a local convergence analysis of IRQI, i.e., Algorithm 3.2 where $\mu_1^{(i)} = \mu_2^{(i)} = \rho_F(x^{(i)})$. We show that the use of Rayleigh functional plays a critical role in achieving higher order convergence rates if the local symmetry of $T(\lambda)$ is present.

An analysis of IRQI can be done naturally by eigenvector decompositions. For simplicity, we drop the superscripts (i) when this does not lead to confusion. Since the scaling of x

does not affect the behavior of the algorithm, we assume without loss of generality that $x = cv + sg$; see (2.15). Let $res = T'(\mu_2)x - T(\mu_1)y$ be the residual of the linear system in Step 1 of Algorithm 3.2. We want to decompose the approximate solution y in terms of v and another direction. To that end, consider the decompositions

$$F(\lambda)T'(\lambda)g = \gamma_1(c_1v + s_1g_1), \tag{3.1}$$

$$F'(\lambda)T'(\mu_2)x = \gamma_2(c_2v + s_2g_2), \tag{3.2}$$

$$F(\mu_1)T''(\lambda)x = \gamma_3(c_3v + s_3g_3), \quad \text{and} \tag{3.3}$$

$$F(\mu_1)res = \gamma_4(c_4v + s_4g_4), \tag{3.4}$$

where F is the analytic function defined in Theorem 2.2, $(\gamma_j, c_j, s_j, g_j)$ ($j = 1$ to 4) can be obtained by substituting the left-hand sides in (3.1) – (3.4), respectively, for x in (2.15); in particular, we have $w^HT'(\lambda)g_j = 0$. Here, we assume that $\mu_1, \mu_2 \in U$, where U is a neighborhood of λ in which F is analytic. It follows from (3.1) – (3.4) that $\gamma_j = \mathcal{O}(1)$; that is, the generalized norms of these vectors are bounded. In addition, we have from (2.11) that $F(\lambda)T'(\lambda)v \in \ker(T(\lambda)) = \text{span}\{v\}$, i.e., $F(\lambda)T'(\lambda)v = \alpha v$ for some bounded scalar α. Then the approximate solution obtained in Step 1 of Algorithm 3.2 is

$$\begin{aligned}
y &= T(\mu_1)^{-1}\left(T'(\mu_2)x - res\right) \\
&= \left(\frac{1}{\mu_1-\lambda}vw^H + F(\mu_1)\right)\left(T'(\mu_2)x - res\right) && \text{(see (2.9))} \\
&= \frac{vw^H}{\mu_1-\lambda}\left(T'(\lambda)(cv+sg) + (\mu_2-\lambda)T''(\lambda)x + \frac{1}{2}(\mu_2-\lambda)^2T'''(\lambda)x - res\right) \\
&\qquad + F(\mu_1)(T'(\mu_2)x - res) + \mathcal{O}\left(\frac{|\mu_2-\lambda|^3}{|\mu_1-\lambda|}\right) \\
&= \left(\frac{c - w^Hres}{\mu_1-\lambda} + \frac{\mu_2-\lambda}{\mu_1-\lambda}w^HT''(\lambda)x + \frac{(\mu_2-\lambda)^2}{2(\mu_1-\lambda)}w^HT'''(\lambda)x\right)v \\
&\qquad + F(\mu_1)T'(\mu_2)x - F(\mu_1)res + \mathcal{O}\left(\frac{|\mu_2-\lambda|^3}{|\mu_1-\lambda|}\right) \quad \text{(see (2.10) and (2.14))} \\
&= \left(\frac{c - w^Hres}{\mu_1-\lambda} + \frac{\mu_2-\lambda}{\mu_1-\lambda}w^HT''(\lambda)x + \frac{(\mu_2-\lambda)^2}{2(\mu_1-\lambda)}w^HT'''(\lambda)x\right)v \\
&\qquad + F(\lambda)T'(\lambda)(cv+sg) + (\mu_1-\lambda)F'(\lambda)T'(\mu_2)x \\
&\qquad + (\mu_2-\lambda)F(\mu_1)T''(\lambda)x - F(\mu_1)res + \mathcal{O}\left(\frac{|\mu_2-\lambda|^3}{|\mu_1-\lambda|}\right) \\
&= \left(\frac{c - w^Hres}{\mu_1-\lambda} + \frac{\mu_2-\lambda}{\mu_1-\lambda}w^HT''(\lambda)x + \frac{(\mu_2-\lambda)^2}{2(\mu_1-\lambda)}w^HT'''(\lambda)x\right)v \\
&\qquad - F(\mu_1)res + c\alpha v + s\gamma_1(c_1v + s_1g_1) + (\mu_1-\lambda)\gamma_2(c_2v + s_2g_2) \\
&\qquad + (\mu_2-\lambda)\gamma_3(c_3v + s_3g_3) + \mathcal{O}\left(\frac{|\mu_2-\lambda|^3}{|\mu_1-\lambda|}\right) \\
&\equiv \beta v + g_y + \mathcal{O}\left(\frac{|\mu_2-\lambda|^3}{|\mu_1-\lambda|}\right),
\end{aligned} \tag{3.5}$$

where

$$\begin{aligned}
\beta = \frac{c - w^Hres}{\mu_1-\lambda} + \frac{\mu_2-\lambda}{\mu_1-\lambda}w^HT''(\lambda)x + \frac{(\mu_2-\lambda)^2}{2(\mu_1-\lambda)}w^HT'''(\lambda)x + c\alpha \\
+ s\gamma_1c_1 + (\mu_1-\lambda)\gamma_2c_2 + (\mu_2-\lambda)\gamma_3c_3 - \gamma_4c_4,
\end{aligned} \tag{3.6}$$

and

$$g_y = s\gamma_1 s_1 g_1 + (\mu_1 - \lambda)\gamma_2 s_2 g_2 + (\mu_2 - \lambda)\gamma_3 s_3 g_3 - \gamma_4 s_4 g_4, \tag{3.7}$$

which satisfies $w^H T'(\lambda) g_y = 0$.

Insights into the local convergence of IRQI can be obtained by studying the generalized tangent of $\angle(y, v)$, which depends on the magnitude of βv and g_y. To begin the analysis, we assume that for a given $d \in (0, |c|)$, the linear system $T(\mu_1)y = T'(\mu_2)x$ is solved to an accuracy such that

$$\|res\| \leq \tau \|T'(\mu_2)x\| \leq \frac{|c| - d}{\|w\|}, \tag{3.8}$$

and therefore

$$|c - w^H res| \geq |c| - \|w\|\|res\| \geq d.$$

With the above assumption on the error of the inner solves, we can show the convergence of IRQI. Let $\mu_1 = \mu_2 = \rho_F(x)$ in (3.6) and (3.7). If (3.8) holds, we see from (3.6) that $\dfrac{c - w^H res}{\mu_1 - \lambda}$ is the unique dominant term in β, and thus

$$\beta = \mathcal{O}\left(\frac{d}{\rho_F(x) - \lambda}\right) + \mathcal{O}(1) + \mathcal{O}(\rho_F(x) - \lambda).$$

Therefore, for all x close to v in direction, $|\rho_F(x) - \lambda|$ must be small, and there exists a constant C_3 independent of t, such that

$$|\beta| \geq \frac{C_3 d}{|\rho_F(x) - \lambda|}.$$

In addition, given the definition of the generalized norm (2.12), we see from (3.4) that the generalized norm of $F(\mu_1)res$ satisfies

$$\gamma_4 = \left\| \begin{bmatrix} w^H \\ W_{n-1}^H \end{bmatrix} F(\mu_1)\, res \right\| \leq \left\| \begin{bmatrix} w^H \\ W_{n-1}^H \end{bmatrix} F(\mu_1) \right\| \|res\| \leq C\|T'(\mu_2)x\|\tau.$$

Therefore, from (3.7), we can see that there exists constants C_4 and C_5 independent of t, such that

$$\text{if } \tau = \mathcal{O}(1), \text{ then } \gamma_4 = \mathcal{O}(1), \text{and } \|g_y\| \leq C_4, \tag{3.9}$$

$$\text{if } \tau = \mathcal{O}(t), \text{ then } \gamma_4 = \mathcal{O}(t), \text{ and } \|g_y\| \leq \mathcal{O}(s) + \mathcal{O}(\rho_F(x) - \lambda) \leq C_5 t. \tag{3.10}$$

It then follows from the last equality of (3.5) that, there exists constants C_6 and C_7, such that the generalized tangent of $\angle(y, v)$ is bounded *above* by

$$\begin{aligned} \frac{\|g_y\| + \mathcal{O}\left(|\rho_F(x) - \lambda|^2\right)}{\|\beta v\| - \mathcal{O}\left(|\rho_F(x) - \lambda|^2\right)} &\leq \frac{|\rho_F(x) - \lambda|\|g_y\| + \mathcal{O}(|\rho_F(x) - \lambda|^3)}{C_3 d\|v\| - \mathcal{O}(|\rho_F(x) - \lambda|^3)} \\ &\leq \begin{cases} C_6|\rho_F(x) - \lambda| & \text{if } \tau = \mathcal{O}(1) \\ C_7|\rho_F(x) - \lambda|t & \text{if } \tau = \mathcal{O}(t) \end{cases}. \end{aligned} \tag{3.11}$$

Note from (3.7) that for all x close to v in direction, s and $|\rho_F(x) - \lambda|$ are both small. Therefore, for an appropriately small fixed tolerance τ, C_4 in (3.9), and C_6 in (3.11) must

also be small. Assuming that $|\rho_F(x) - \lambda| \leq \mathcal{O}(t)$ (see Theorem 2.1), the small magnitude of C_6 guarantees at least linear convergence of IRQI, i.e., the generalized tangent of $\angle(y, v)$ is smaller than that of $\angle(x, v)$. The analysis can be summarized as follows.

THEOREM 3.6. [Local convergence of IRQI] *Let (λ, v) be a simple eigenpair of (2.1), and w^H the corresponding left eigenvector. Assume that there exists a small $r > 0$ and $\zeta > 0$ such that $\|T'(\mu)x\| \leq \zeta$ for all $(\mu, x) \in \overline{B(z_*, r)}$ where $z_* = \begin{bmatrix} v \\ \lambda \end{bmatrix}$. Let $x^{(0)} = \gamma^{(0)}(c^{(0)}v + s^{(0)}g^{(0)})$ (see (2.15)) be a vector such that $\left(\rho_F(x^{(0)}), x^{(0)}\right) \in \overline{B(z_*, r)}$. For a given $d \in (0, |c^{(0)}|)$, let $\tau_{max} < \dfrac{|c^{(0)}| - d}{\zeta\|w\|}$ be an upper bound of the tolerance for the inner solve of Algorithm 3.2. Then, if $x^{(0)}$ is close to v in direction, and $\tau^{(i)} = \tau \leq \tau_{max}$ is an appropriately small fixed tolerance, Algorithm 3.2 with $\mu_1^{(i)} = \mu_2^{(i)} = \rho_F(x^{(i)})$ converges at least linearly to (λ, v), and it converges at least quadratically if the local symmetry of $T(\lambda)$ is present and $p^{(i)}$ is chosen as in (2.2) for $\rho_F(x^{(i)})$. In addition, if $\tau^{(i)} \leq Ct^{(i)} \leq \tau_{max}$ for some C independent of i, this algorithm converges at least quadratically and at least cubically, respectively, if the local symmetry of $T(\lambda)$ is absent, or if it is present with $p^{(i)}$ chosen as in (2.2) for $\rho_F(x^{(i)})$.*

Theorem 3.6 shows that, the impact of the tolerances for the inner solves on the order of local convergence of IRQI for general NEPs is the same as that for linear eigenvalue problems $Av = \lambda Bv$; in particular, with a decreasing sequence of tolerances proportional to the error, IRQI achieves the same order of local convergence as exact RQI does; see [3, 4, 11, 12, 13]. Note that the convergence of IRQI depends on the availability of a good initial eigenvector approximation, which can be computed by Algorithm 3.2 with $\mu_1^{(i)} = \mu_2^{(i)} = \sigma \approx \lambda$ for linear eigenvalue problems. Unfortunately, this approach is not applicable to general NEPs; in fact, Algorithm 3.2 with a fixed $\mu_1^{(i)} = \sigma$ generally fails to converge. We will see in the appendix that this difference is caused by a special property of linear eigenvalue problems that does not exist for general NEPs. The difficulty in obtaining a good initial eigenvector approximation can be resolved by residual inverse iteration, which we now discuss in Section 4.1.

4. Variants of the Davidson-type method. In this section, we present a local convergence analysis of the inexact residual inverse iteration [31] and the single-vector version of the Jacobi-Davidson method [7] for problem (2.1). Given an eigenpair approximation (μ, x), both methods solve a variant of the correction equation approximately for a correction direction Δx, and the new outer iterate $x + \Delta x$ generally becomes closer to v in direction.

4.1. Inexact residual inverse iteration. From the iteration scheme (2.7), the inexact residual inverse iteration can be described as follows.

ALGORITHM 4.1. **Inexact residual inverse iteration**

Start with $(\sigma, x^{(0)})$, where $\sigma \approx \lambda$ is fixed, and $u^H x^{(0)} = 1$

For $i = 0, 1, 2, ...,$ until convergence

1. Choose an auxiliary vector $p^{(i)}$, and compute $\rho^{(i)} = \rho_F(x^{(i)})$
2. Choose a relative tolerance $\tau^{(i)}$, and solve $T(\sigma)\Delta x^{(i)} = -T(\rho^{(i)})x^{(i)}$ approximately such that the residual $res^{(i)} \equiv -T(\rho^{(i)})x^{(i)} - T(\sigma)\Delta x^{(i)}$ satisfies $\|res^{(i)}\| \leq \tau^{(i)}\|T(\rho^{(i)})x^{(i)}\|$
3. $y^{(i)} = x^{(i)} + \Delta x^{(i)}$
4. $x^{(i+1)} = y^{(i)}/(u^H y^{(i)})$ and test for convergence.

End For

The matrix $T(\sigma)$ arising in Step 2 can be considered as a preconditioner that applies to the eigenvalue residual vector $T(\rho^{(i)})x^{(i)}$. It is shown in [31] that if σ is sufficiently close to λ,

the *exact* residual inverse iteration converges to (λ, v) linearly, and we have the estimate

$$\frac{\|x^{(i+1)} - v\|}{\|x^{(i)} - v\|} = \mathcal{O}(|\sigma - \lambda|).$$

A quantitative analysis of the convergence factor is established in [22]. If a good initial approximate eigenvector is not available, one can use the inexact residual inverse iteration to obtain an approximate eigenpair of moderate accuracy, and then use the standard inexact inverse iteration or IRQI, which converge quadratically or cubically, to refine the eigenvector approximation; see [37] for a criterion to switch from inverse iteration with a fixed shift to Rayleigh quotient iteration for linear eigenvalue problems.

Here, we show that a linear convergence factor proportional to $|\sigma - \lambda|$ can be achieved by the inexact variant as well. Our derivation closely follows that given in [31]. To begin the analysis, assume that $T(\cdot)$ is twice continuously differentiable (less stringent than analyticity) in a neighborhood U of λ. Define the *divided difference*

$$T[\mu_1, \mu_2] \equiv \begin{cases} \dfrac{T(\mu_2) - T(\mu_1)}{\mu_2 - \mu_1} & \text{if } \mu_1 \neq \mu_2, \\ \\ T'(\mu_1) & \text{if } \mu_1 = \mu_2. \end{cases}$$

Applying the mean-value theorem for each entry, we have

$$T[\mu_1, \mu_2] = T'(\xi_{ij}) \equiv \left[T'_{ij}(\xi_{ij})\right],$$

where ξ_{ij} depends on μ_1, μ_2 and the index (i, j).

To analyze the convergence of the inexact residual inverse iteration, we need to study the difference between the approximate solution y (in Step 2 of Algorithm 4.1) and the desired eigenvector v. Since λ is a simple eigenvalue of $T(\cdot)$, it is isolated (see Theorem 2.2), i.e., there exists a small δ such that for all σ satisfying $0 < |\sigma - \lambda| < \delta$, $T(\sigma)$ is nonsingular. Define

$$Q \equiv T(\lambda) + T'(\lambda) v u^H,$$

which behaves the same way as $T'(\lambda)$ on v, and is shown to be nonsingular in [31]. Next, we define a matrix S_σ closely-related to Q, and we will see that the error of the normalized y (Step 4 of Algorithm 4.1), i.e., $\|y/(u^H y) - v\|$, can be represented in terms of S_σ. Consider

$$\begin{aligned} S_\sigma &\equiv T(\sigma) + (1 - \sigma + \lambda) T[\sigma, \lambda] v u^H \\ &= Q + T(\sigma) - T(\lambda) + (T[\sigma, \lambda] - T'(\lambda) - (\sigma - \lambda) T[\sigma, \lambda])\, v u^H \\ &= Q + T(\sigma) - T(\lambda) + \left(\frac{T(\sigma) - T(\lambda)}{\sigma - \lambda} - T'(\lambda) - T(\sigma) + T(\lambda) \right) v u^H \\ &= Q + (\sigma - \lambda) T'(\lambda) + (\sigma - \lambda) \left(T''(\lambda)/2 - T'(\lambda)\right) v u^H + \mathcal{O}((\sigma - \lambda)^2). \end{aligned} \tag{4.1}$$

It follows that $S_\sigma \to Q$ as $\sigma \to \lambda$. Suppose that σ is close to λ such that $|\sigma - \lambda| < \delta \ll \|Q\|$. The Neumann expansion of S_σ^{-1} gives

$$S_\sigma^{-1} = Q^{-1} - (\sigma - \lambda) Q^{-1} \left(T'(\lambda) + (T''(\lambda)/2 - T'(\lambda)) v u^H \right) Q^{-1} + \mathcal{O}((\sigma - \lambda)^2),$$

and thus $\|S_\sigma^{-1}\|$ is uniformly bounded for all σ satisfying $|\sigma - \lambda| < \delta$.

To investigate $\|y/(u^H y) - v\|$, we need to study the behavior of S_σ. Since $T(\lambda) v = 0$, it follows from (4.1) that

$$\begin{aligned} S_\sigma v &= (T(\sigma) - T(\lambda))\, v + (1 - \sigma + \lambda) T[\sigma, \lambda] v u^H v \\ &= (\sigma - \lambda) T[\sigma, \lambda] v + (1 - \sigma + \lambda) T[\sigma, \lambda] v = T[\sigma, \lambda] v. \end{aligned} \tag{4.2}$$

Assume without loss of generality that $x = cv + sg$ ($\gamma = 1$). Note that for any small angle φ, $\cos\varphi = \cos 2\left(\frac{\varphi}{2}\right) = 1 - 2\sin^2\left(\frac{\varphi}{2}\right) = 1 - \frac{1}{2}\sin^2\varphi + \mathcal{O}(\varphi^4)$. Then for a fixed σ and a Rayleigh functional $\rho = \rho_F(x)$ satisfying $|\rho - \lambda| \ll |\sigma - \lambda|$, we have

$$
\begin{aligned}
(4.3)\quad T[\sigma,\rho]x = \frac{T(\sigma)-T(\rho)}{\sigma-\rho}x &= \frac{T(\sigma)-T(\lambda)+T(\lambda)-T(\rho)}{\sigma-\lambda}\frac{\sigma-\lambda}{\sigma-\rho}(cv+sg)\\
&= \left(T[\sigma,\lambda] - T'(\lambda)\frac{\rho-\lambda}{\sigma-\lambda} + \mathcal{O}\left(\frac{(\rho-\lambda)^2}{\sigma-\lambda}\right)\right)\cdot\\
&\qquad\qquad \left(1+\frac{\rho-\lambda}{\sigma-\rho}\right)\left(v - \frac{s^2}{2}v + sg + \mathcal{O}(s^4)\right)\\
&= T[\sigma,\lambda]v + \mathcal{O}(\rho-\lambda) + \mathcal{O}(s) = T[\sigma,\lambda]v + \mathcal{O}(s). \text{ (from Theorem 2.1)}
\end{aligned}
$$

From Step 2 of Algorithm 4.1, we have

$$
(4.4)\qquad y = x - T(\sigma)^{-1}(T(\rho)x - res).
$$

Since $u^H(y - (u^Hy)v) = u^Hy - (u^Hy)u^Hv = 0$, it follows from the definition of S_σ that

$$
\begin{aligned}
(4.5)\quad & S_\sigma(y-(u^Hy)v) = T(\sigma)(y-(u^Hy)v) = T(\sigma)y - (u^Hy)T(\sigma)v\\
&= (T(\sigma)-T(\rho))\,x + res - (u^Hy)\,(T(\sigma)-T(\lambda))\,v \qquad \text{(see (4.4))}\\
&= (\sigma-\rho)T[\sigma,\rho]x - (u^Hy)(\sigma-\lambda)T[\sigma,\lambda]v + res\\
&= (\sigma-\rho)\,(T[\sigma,\lambda]v + \mathcal{O}(s)) - (u^Hy)(\sigma-\lambda)S_\sigma v + res \quad \text{(see (4.2) and (4.3))}\\
&= (\sigma-\rho)\left(S_\sigma v + \frac{res}{\sigma-\rho} + \mathcal{O}(s)\right) - (u^Hy)(\sigma-\lambda)S_\sigma v. \text{ (see (4.2))}
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
(4.6)\quad & y - (u^Hy)v = (\sigma-\rho)\left(v + \frac{S_\sigma^{-1}res}{\sigma-\rho} + \mathcal{O}(s)\right) - (u^Hy)(\sigma-\lambda)v\\
\Rightarrow\quad & 0 = u^H(y-(u^Hy)v) = (\sigma-\rho)\left(1 + \frac{u^HS_\sigma^{-1}res}{\sigma-\rho} + \mathcal{O}(s)\right) - (u^Hy)(\sigma-\lambda)\\
\Leftrightarrow\quad & \sigma-\rho = u^Hy(\sigma-\lambda)\left(1 + \frac{u^HS_\sigma^{-1}res}{\sigma-\rho} + \mathcal{O}(s)\right)^{-1}.
\end{aligned}
$$

Since $|\rho-\lambda| \ll |\sigma-\lambda|$, we have $|\sigma-\rho| \approx |\sigma-\lambda|$. To see the magnitude of res, note that

$$
\begin{aligned}
(4.7)\quad \|T(\rho)x\| &= \|T(\lambda)\gamma(cv+sg) + (\rho-\lambda)T'(\lambda)x\| + \mathcal{O}(|\rho-\lambda|^2)\\
&= \mathcal{O}(s) + \mathcal{O}(\rho-\lambda) = \mathcal{O}(t).
\end{aligned}
$$

Therefore, if a *fixed* small relative tolerance $\tau = \mathcal{O}(\sigma-\lambda)$ is used in Algorithm 4.1, such that $\|res\| \le \tau\|T(\rho)x\| = \mathcal{O}\left((\sigma-\rho)s\right)$, then

$$
(4.8)\qquad \sigma-\rho = u^Hy(\sigma-\lambda)(1+\mathcal{O}(s)).
$$

Substituting (4.8) into the first line of (4.6), we see that the difference between the new eigenvector approximation (Step 4 of Algorithm 4.1) and the true eigenvector is

$$
\begin{aligned}
&\left\|\frac{1}{u^Hy}y - v\right\| = \left\|\frac{1}{u^Hy}\left(y-(u^Hy)v\right)\right\|\\
&= \|(\sigma-\lambda)(1+\mathcal{O}(s))\,(v+\mathcal{O}(s)) - (\sigma-\lambda)v\| = (\sigma-\lambda)\mathcal{O}(s).
\end{aligned}
$$

If the fixed shift σ is sufficiently close to λ, such that

$$\left\| \frac{1}{u^H y} y - v \right\| = (\sigma - \lambda)\mathcal{O}(s) < \|x - v\| = \left\| -\frac{s^2}{2} v + sg \right\| = \mathcal{O}(s),$$

Algorithm 4.1 converges linearly, and the convergence factor is proportional to $|\sigma - \lambda|$. The above convergence analysis is summarized in the following theorem.

THEOREM 4.2. [Local convergence of inexact residual inverse iteration] *Suppose that Algorithm 4.1 with a fixed shift σ is used to solve (2.1), where $x^{(0)}$ is sufficiently close to v in direction such that the Rayleigh functional $\rho_F(x^{(0)})$ satisfies $|\rho_F(x^{(0)}) - \lambda| \ll |\sigma - \lambda|$. If σ is close enough to λ, and an appropriately small fixed relative tolerance $\tau = \mathcal{O}(\sigma - \lambda)$ is used in Step 2, then Algorithm 4.1 converges at least linearly such that $\dfrac{\|x^{(i+1)} - v\|}{\|x^{(i)} - v\|} = \mathcal{O}(\sigma - \lambda)$.*

4.2. Single-vector Jacobi-Davidson (JD) method. A potential problem for the inexact residual inverse iteration is that it may suffer from stagnation (little improvement in eigenvector approximation) in the ith iteration, if $\rho^{(i)}$ is close to σ. In particular, if $\rho^{(i)} = \sigma$, it follows from Step 2 of Algorithm 4.1 that $\Delta x^{(i)} = -x^{(i)}$, and thus $y^{(i)} = 0$, if the correction equation is solved exactly. Fortunately, this stagnation can be avoided by the Jacobi-Davidson method, which imposes some orthogonality condition on the correction vector Δx to x.

From the iteration scheme (2.8), the single-vector JD method with the Rayleigh functional shift can be described as follows.

ALGORITHM 4.3. **Single-vector JD method**

Start with an approximate pair $(\mu^{(0)}, x^{(0)})$ where $x^{(0)}$ is normalized

For $i = 0, 1, 2, ...,$ until convergence

1. Choose an auxiliary vector $p^{(i)}$, and compute the Rayleigh functional $\mu_2^{(i)} = \rho_F(x^{(i)})$; choose $\mu_1^{(i)}$ accordingly

2. Choose a relative tolerance $\tau_{JD}^{(i)}$, and solve $\Pi_1^{(i)} T(\mu_1^{(i)}) \Pi_2^{(i)} \Delta x^{(i)} = -T(\mu_2^{(i)}) x^{(i)}$ approximately, such that the residual $res_{JD}^{(i)} \equiv -T(\mu_2^{(i)}) x^{(i)} - \Pi_1^{(i)} T(\mu_1^{(i)}) \Pi_2^{(i)} \Delta x^{(i)}$ satisfies $\|res_{JD}^{(i)}\| \le \tau_{JD}^{(i)} \|T(\mu_2^{(i)}) x^{(i)}\|$

3. $y^{(i+1)} = x^{(i)} + \Delta x^{(i)}$

4. Normalize $y^{(i+1)}$ into $x^{(i+1)}$ and test for convergence

End For

$\mu^{(i+1)} = \mu_2^{(i+1)}$, and $x^{(i+1)} = x^{(i+1)} / (u^H x^{(i+1)})$.

In this section, we consider single-vector JD with $\mu_1^{(i)} = \mu_2^{(i)} = \rho_F(x^{(i)})$, and we will discuss in the appendix the case where $\mu_1^{(i)} = \sigma$ is fixed. Here, we derive a convergence analysis of single-vector JD from the convergence of IRQI (Theorem 3.6) by exploring close connections between the two algorithms. This approach is used in [11] in the setting of linear generalized eigenvalue problems. To simplify the notation, we again drop the superscripts (i). Letting $\mu_1 = \mu_2 = \rho = \rho_F(x)$, from Step 2 of Algorithm 4.3, we have

$$\Pi_1 T(\rho) \Pi_2 \Delta x = -T(\rho) x - res_{JD}.$$

Note that $T(\rho)x = \Pi_1 T(\rho)x$, and $\Pi_2 \Delta x = \Delta x$. Therefore,

$$\Pi_1 T(\rho) \Pi_2 \Delta x + T(\rho) x = \Pi_1 T(\rho)(x + \Delta x) = -res_{JD},$$

and since $\Pi_1 = I - \dfrac{T'(\rho)xp^H}{p^HT'(\rho)x}$, we have, equivalently,

$$T(\rho)(x+\Delta x) = -res_{JD} + \nu T'(\rho)x, \tag{4.9}$$

where ν is a scalar for which $\Delta x \perp u$. Premultiplying both sides by $T(\rho)^{-1}$, subtracting x, and then premultiplying by u^H, we have

$$\nu = \frac{u^H\left(T(\rho)^{-1}res_{JD} + x\right)}{u^HT(\rho)^{-1}T'(\rho)x}. \tag{4.10}$$

To study the numerator of ν in (4.10), recall from (4.7) that $\|T(\rho)x\| = \mathcal{O}(s)$. Therefore, if $\|res_{JD}\| \le \tau_{JD}\|T(\rho)x\|$ with $\tau_{JD} < 1$, we have $\|res_{JD}\| = \mathcal{O}(t)$.

With this observation, it can be shown that the numerator of ν is bounded away from zero with some appropriate choice of the tolerance τ_{JD}. In fact, note from Theorem 2.1 and (2.9) that $\|T(\rho)^{-1}\| = \mathcal{O}(t^{-1})$ or $\mathcal{O}(t^{-2})$, depending on the presence of the local symmetry of $T(\lambda)$. If $\|T(\rho)^{-1}\| = \mathcal{O}(t^{-1})$, and $\tau_{JD} < 1$ is an appropriately small fixed tolerance, then there exists a small constant C_7 such that $\|res_{JD}\| \le C_7 t$, and thus $|u^HT(\rho)^{-1}res_{JD}|$ can be bounded above by a constant independent of t and smaller than 1. Next, suppose that $\|T(\rho)^{-1}\| = \mathcal{O}(t^{-2})$ due to the presence of the local symmetry. If $\tau_{JD} < 1$ is a fixed tolerance such that $\|res_{JD}\| = \mathcal{O}(t)$, then for all sufficiently small t, we have $|u^HT(\rho)^{-1}res_{JD}| = \mathcal{O}(t^{-1}) \gg 1$; if $\tau_{JD} \le C_8 t$ for an appropriately small constant C_8, then there exists a small constant C_9 such that $\|res_{JD}\| \le C_9 t^2$, and thus $|u^HT(\rho)^{-1}res_{JD}|$ can be bounded above by a constant independent of t and smaller than 1. In all these circumstances, the numerator of ν

$$u^H(T(\rho)^{-1}res_{JD} + x) = u^HT(\rho)^{-1}res_{JD} + 1$$

is bounded away from zero.

Now consider the denominator of ν in (4.10). By letting the residual be zero in (3.5), and noting that $\dfrac{c}{\rho-\lambda}v$ is the unique dominant term in $T(\rho)^{-1}T'(\rho)x$, we see there exists some positive constant C_{10} such that the denominator of ν satisfies

$$\left|u^HT(\rho)^{-1}T'(\rho)x\right| \le \frac{C_{10}}{|\rho-\lambda|}.$$

Therefore, since the numerator of ν is bounded away from zero with some appropriate τ_{JD}, we have $|\nu| \ge C_{11}|\rho-\lambda|$ for some positive constant C_{11}.

We are now ready to give the connection between the approximate solution to the linear systems arising in single-vector JD and IRQI. Note that (4.9) can be written as

$$T(\rho)\frac{x+\Delta x}{\nu} = T'(\rho)x - \frac{res_{JD}}{\nu}, \qquad \text{or, equivalently,}$$
$$T(\rho)y = T'(\rho)x - res_{RQI}.$$

In other words, let Δx be an approximate solution of the correction equation in single-vector JD, $res_{JD} = -T'(\rho)x - \Pi_1T(\rho)\Pi_2\Delta x$ be the residual vector, and ν as given in (4.10). Then, an approximate solution $y = \dfrac{x+\Delta x}{\nu}$ of the linear system arising in IRQI can be constructed, and the corresponding residual vector is $res_{RQI} = \dfrac{res_{JD}}{\nu}$. Therefore, define

$$\tau_{JD} = |\nu|\tau_{RQI}\frac{\|T'(\rho)x\|}{\|T(\rho)x\|} \tag{4.11}$$

such that

$$\|res_{JD}\| \le \tau_{JD}\|T(\rho)x\| = \left(|\nu|\tau_{RQI}\frac{\|T'(\rho)x\|}{\|T(\rho)x\|}\right)\|T(\rho)x\| = |\nu|\tau_{RQI}\|T'(\rho)x\|,$$

or equivalently, the residual res_{RQI} corresponding to the approximate solution $y = \dfrac{x+\Delta x}{\nu}$ satisfies

$$\|res_{RQI}\| = \frac{\|res_{JD}\|}{|\nu|} \le \tau_{RQI}\|T'(\rho)x\|.$$

Since $x + \Delta x$ and y differ only by a scaling factor, they represent the same eigenvector approximation. The convergence of single-vector JD can thus be established by applying the convergence of IRQI (Theorem 3.6). Recall that in (4.11), we have $|\nu| \ge C_{11}(\rho - \lambda) = \mathcal{O}(t)$ or $\mathcal{O}(t^2)$, depending on the presence of the local symmetry of $T(\lambda)$ and the choice of p for $\rho_F(x)$, $\|T(\rho)x\| = \mathcal{O}(t)$, and $\|T'(\rho)x\| = \mathcal{O}(1)$. Therefore, we have the following theorem.

THEOREM 4.4. [Local convergence of single-vector JD] *Suppose that the assumptions of Theorem 3.6 hold. If $x^{(0)}$ is sufficiently close to v in direction, and an appropriately small fixed tolerance $\tau^{(i)} = \tau_0$ is used for the inner solve, then Algorithm 4.3 with $\mu_1^{(i)} = \mu_2^{(i)} = \rho_F(x^{(i)})$ converges at least linearly, and it converges at least quadratically if the local symmetry of $T(\lambda)$ is present and $p^{(i)}$ is chosen as in (2.2) for $\rho_F(x^{(i)})$. In addition, if $\tau^{(i)} \le Ct^{(i)} \le \tau_0$ for some appropriately small constant C independent of i, this algorithm converges at least quadratically and at least cubically, respectively, if the local symmetry of $T(\lambda)$ is absent, or if it is present with $p^{(i)}$ chosen as in (2.2) for $\rho_F(x^{(i)})$.*

REMARK 4.5. The order of convergence of single-vector JD shown in Theorem 4.4 is consistent with that given in [23, 39] for linear standard Hermitian problems and in [18] for linear generalized non-Hermitian problems. Our analysis, based on the property of the nonlinear Rayleigh functional and the convergence of IRQI, applies to the general NEP (2.1).

The analyses of the inexact residual inverse iteration and single-vector JD also provide some insight into the local convergence of some iterative subspace projection methods. In fact, two most commonly-used projection methods for (2.1), namely, the nonlinear Arnoldi method and the full version of the Jacobi-Davidson method, respectively, arise from combinations of residual inverse iteration and single-vector JD with subspace projection, where Δx is used to enlarge the subspace for candidate eigenvector approximations; see [30]. Since new eigenvector approximations that are more accurate than $x + \Delta x$ may be extracted from the enlarged subspace, the local convergence rates of the two single-vector algorithms can be used as conservative estimates of their counterparts with subspace projection, which in general converge more rapidly; see, e.g., [20] for this perspective.

We see from Sections 2 to 4 that, though all the algorithms discussed in this paper bear close connections to Newton's method, it is worth noting that only the standard *exact* inverse iteration is mathematically equivalent to Newton's method; that is, given the same initial eigenvector approximation $(\mu^{(0)}, x^{(0)})$, they produce the same sequence of iterates $(\mu^{(i)}, x^{(i)})$. The standard *inexact* inverse iteration is equivalent to a *modified* Newton's method, where the error of the inner solve is equivalently represented as a perturbation of the Jacobian matrix. Rayleigh quotient iteration, residual inverse iteration and the single-vector Jacobi-Davidson method, whether exact or inexact versions, use the nonlinear Rayleigh functional $\rho_F(x)$ as the new eigenvalue approximation, and they cannot be considered as any variants of standard Newton's method that treat the eigenvalue approximation μ as an ordinary entry of the eigenpair iterate $z = \begin{bmatrix} x \\ \mu \end{bmatrix}$. Therefore, the convergence theory of inexact Newton's

method [9] is not applicable to any of the inexact algorithms discussed in this paper. The convergence analyses in Sections 3 and 4 are specifically developed to solve the NEP (2.1).

We have shown that the use of the two-sided Rayleigh functional is critical for IRQI and JD to achieve higher order convergence in the presence of the local symmetry of $T(\lambda)$. In fact, the choice of the normalization vector u for the standard inexact inverse iteration (Algorithm 3.1) directly affects the quality of the new eigenvalue approximation $\mu^{(i+1)}$. An unsuitable choice of u may slow down the convergence of $\mu^{(i+1)}$ toward the desired eigenvalue, and therefore may also delay the convergence of eigenvector approximation in the subsequent iterations [28]. By contrast, the Rayleigh functional $\rho_F(x)$ does not depend on the normalization vector, and it provides a more accurate eigenvalue approximation when the local symmetry of $T(\lambda)$ is present. Therefore, RQI and JD can be interpreted as a special variant of Newton's method with enhancement of the eigenvalue approximation. In Section 5, we see that the standard inexact inverse iteration converges at most quadratically for all test problems, whereas IRQI and single-vector JD show higher order or convergence rates when the local symmetry is present.

5. Numerical Experiments. In this section, we illustrate the convergence results presented in Sections 3 and 4 with numerical examples. We show that with appropriate choices of tolerances for the inner solves, the inexact algorithms discussed in this paper can achieve the same order of convergence rates as the exact methods; in addition, with certain suitable but less stringent tolerances, lower orders of convergence rates may be obtained.

We begin with the introduction of the test problems. We selected eight problems from the NLEVP 2.0 toolbox [6], including quadratic (QEP), higher order polynomial (PEP), rational eigenvalue problems (REP) and "genuine" nonlinear problems (NEP) that cannot be equivalently transformed to linear problems. Table 5.1 summarizes some properties of these problems. The identifier describes the structure of the matrix pencil $T(\cdot)$. For example, $T(\cdot)$ is real, symmetric, Hermitian, T-even or T-palindromic, respectively, if $\overline{T(\mu)} = T(\bar{\mu})$, $T(\mu)^T = T(\mu)$, $T(\mu)^H = T(\bar{\mu})$, $T(-\mu)^T = T(\mu)$ or $rev(T(\mu)^T) = T(\mu)$ (see the footnote for the definition of rev) for any $\mu \in \Lambda$.[1] The desired eigenvalue λ and the local symmetry of $T(\lambda)$ are also given. Note that the local symmetry of $T(\lambda)$ is absent for problems **plasma_drift** and **railtrack2**, but it is present for the other six problems, whose names are in *italic*. The parameters are used for the `nlevp` function to generate problems of appropriate size; for example, we use the MATLAB command

$$[\mathtt{T},\mathtt{TP}] = \mathtt{nlevp}('\mathtt{eval}','\mathtt{butterfly}',\mathtt{mu},128);$$

to construct $T = T(\mu)$ and $TP = T'(\mu)$ for *butterfly*. In Tables 5.2 – 5.5, the problem names are abbreviated to two capital letters underlined in Table 5.1. Moreover, the matrix pencils arising in all problems are sparse, allowing for the inner solves to be performed efficiently by Krylov subspace methods.

To verify the order of convergence rates in a conclusive manner, we first comment on a weakness of the standard criterion $\|e^{(i+1)}\|/\|e^{(i)}\|^\ell \le C$ used to establish the ℓth order convergence, and we propose a new criterion that is more descriptive for this purpose. The point is that the errors $e^{(i)}$ satisfying the standard criterion can be observed only in a very small number (e.g., two) of iterations, and therefore it is difficult to determine if the algorithm converges quadratically or cubically; for example, if $\|e^{(0)}\| = 10^{-2}$, $\|e^{(1)}\| = 10^{-5}$, and $\|e^{(2)}\| = 10^{-12}$. This difficulty can be tackled by a methodology similar to the one commonly used to show the order of accuracy of numerical methods for differential equations. Specifically, if an algorithm converges linearly, quadratically or cubically, respectively, a decrease of

[1] For a kth-order matrix polynomial $T(\mu)$, $rev\,(T(\mu)) = \mu^k T(1/\mu)$ is the reversal of $T(\mu)$; see [6].

problem	type	identifier	eigenvalue	local symm	params	size
ButterFly	PEP	real, T-even	$9.3330\,i$	Hermitian	128	16384
FiBer	NEP	real, symm	7.1395×10^{-7}	real symm	–	2400
GuN	NEP	symm	$2.2345 \times 10^{4} + 0.6450\,i$	cplx symm	–	9956
Loaded_String	REP	real, symm	9.6950	real symm	2000, 100, 100	2000
Plasma_Drift	PEP	–	$-0.1087 + 0.3412\,i$	–	512	512
RailTrack2	QEP	T-palindr.	$0.0485 - 0.0010\,i$	–	16	11280
SchröDinger	QEP	real, symm	$2.0033 + 0.1185\,i$	cplx symm	–	1998
SleePer	QEP	real, symm	-16.1974	real symm	4096	4096

TABLE 5.1
Description of the test problems from NLEVP

$\|e^{(0)}\|$ by a factor of 2 should lead to a decrease of $\|e^{(1)}\|$ by a factor of 2, 4 or 8. We have found this approach highly descriptive for verifying the order of convergence rates.

We observe that the convergence analyses in Sections 3 and 4 are developed independent of the specific inner solvers. This independence should be appropriately reflected in the experiments. To this end, we first used IDR(4) [35] as the inner solver with incomplete LU preconditioners generated by MATLAB `ilu` using some drop tolerance τ_{dp}. In general, we found that the estimated order of convergence is insensitive to a wide range of τ_{dp}, unless τ_{dp} is too large or too small. An ILU preconditioner corresponding to an excessively large τ_{dp} is usually too weak or even counterproductive to speed up the inner iteration, leading to very slow or even failure of convergence of the MATLAB `idrs` solver provided by Sonneveld and van Gijzen (Version August 31, 2010). A tiny drop tolerance, on the other hand, leads to an ILU preconditioner that is close to the LU decomposition of the coefficient matrix of the inner linear system. As a result, the inner linear solve converges in a very small number of iterations, and therefore the actual relative residual norm can be significantly smaller than the specified tolerance. This produces considerable noise to our estimation of the order of convergence, which depends on $\|e^{(1)}\|/\|e^{(0)}\|$ as a function of $\tau^{(0)}$. This type of noise is particularly difficult to remove for the problems *fiber* and *loaded_string*, where the matrix pencil $T(\mu)$ is a tridiagonal matrix for all μ, because any drop tolerance is either too large or too small in the above sense. To resolve this difficulty for the purpose of illustration, we also controlled the inner solve errors by solving a linear system with a perturbed right-hand side. Specifically, suppose that $Mz = b$ is the linear system to be solved, and τ is the relative tolerance. To obtain an approximate solution $\hat{z}$ satisfying the tolerance, we generate a random perturbation vector δb such that $\|\delta b\| \leq \tau\|b\|$, and then solve $M\hat{z} = b + \delta b$ by MATLAB backslash operation $M\backslash(b + \delta b)$. Compared to the iterative inner solve strategies, we found for our purposes that this approach leads to negligible difference in the estimated convergence rates for all the problems, except for *fiber* and *loaded_string*, for which we did not obtain reasonable estimated order of convergence by using IDR(4) with ILU preconditioners.

As we discussed in Section 4, the choice of the normalization vector u may affect the convergence of the standard inexact inverse iteration (Algorithm 3.1). Here, a normalized u is chosen such that $\angle(u, v)$ is not too close to $\pi/2$. Specifically, we use $u = \tilde{u}/\|\tilde{u}\|$, where $\tilde{u}^T = [1/n, 2/n, ..., (n-1)/n, 1]^{1/2}$ for *schrödinger*, $\tilde{u}^T = [1, -1, 1, -1, ...(-1)^n]$ for *sleeper*, and $\tilde{u}^T = [1, 1, 1, ..., 1]$ for all other problems.

We begin the exposition by reviewing the results for the standard inexact inverse iteration (Algorithm 3.1). Take the problem *butterfly* as an example. The left part of Table 5.2 describes the initial eigenpair approximation error and the initial relative tolerance. The initial

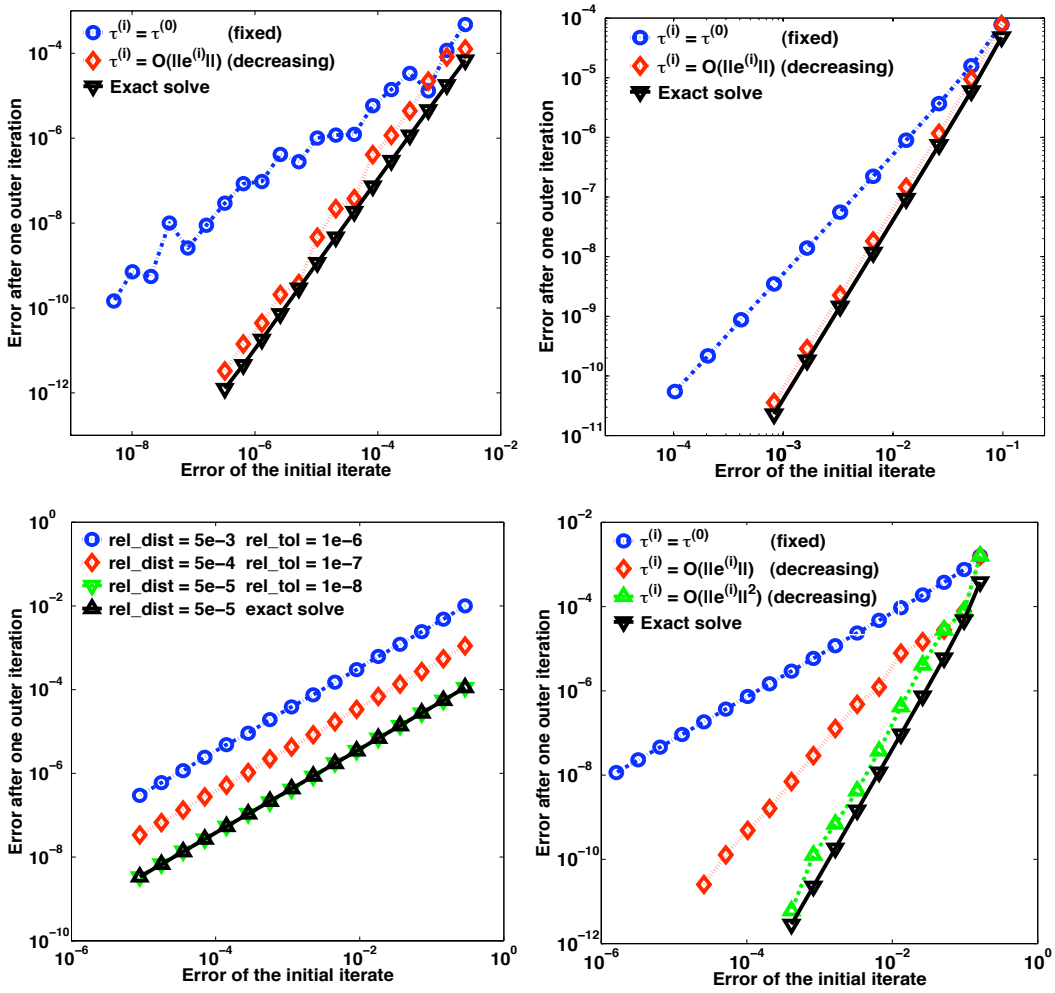

FIG. 5.1. *Log-log plots of the sequence of* $(\|e^{(0)}\|, \|e^{(1)}\|)$ *used to demonstrate the order of convergence for "butterfly" Top left: standard inverse iteration Top right: RQI Bottom left: residual inverse iteration Bottom right: single-vector JD.*

eigenvector approximation $x_1^{(0)}$ is obtained from a linear combination of a given random vector f and the exact eigenvector v, such that $\tan\angle(x_1^{(0)}, v) = 2.5 \times 10^{-3}$; the initial eigenvalue approximation $\mu_1^{(0)}$ is chosen such that $|\mu_1^{(0)} - \lambda|/|\lambda| = 4 \times 10^{-3}$ (see *vec err* and *val err*, respectively). The corresponding relative tolerance is $\tau_1^{(0)} = 10^{-3}$. We then construct a *sequence* of initial eigenpair approximations $\{(\mu_k^{(0)}, x_k^{(0)})\}$ such that $\|e_{k+1}^{(0)}\| = \frac{1}{2}\|e_{k+1}^{(0)}\|$, where $e_k^{(i)} = \begin{bmatrix} x_k^{(i)} - v \\ \mu_k^{(i)} - \lambda \end{bmatrix}$ $(i = 0, 1)$. We then choose either a sequence of fixed tolerance $\{\tau_k^{(0)}\}$ where $\tau_k^{(0)} \equiv \tau_1^{(0)}$, or a decreasing sequence of tolerances $\{\tau_k^{(0)}\}$ where $\tau_{k+1}^{(0)} = \frac{1}{2}\tau_k^{(0)}$. For each initial eigenpair $(\mu_k^{(0)}, x_k^{(0)})$, we run one iteration of Algorithm 3.1 where the inner solve is performed with the relative tolerance $\tau_k^{(0)}$, and we obtain $(\mu_k^{(1)}, x_k^{(1)})$ for which $e_k^{(1)}$ can

be computed. Since an algorithm that has ℓth order convergence satisfies $\frac{\|e^{(1)}\|}{\|e^{(0)}\|^\ell} \leq C$, i.e., $\log \|e^{(1)}\| = \ell \log \|e^{(0)}\| + \log C$, estimates for ℓ and C can be obtained by solving a linear least squares problem for the sequence of pairs $\left\{\left(\log \|e_k^{(0)}\|, \log \|e_k^{(1)}\|\right)\right\}$.

Figure 5.1 plots the curves of $\left\{\left(\log \|e_k^{(0)}\|, \log \|e_k^{(1)}\|\right)\right\}$ for the four inexact algorithms discussed in this paper. The top left plot for the standard inexact inverse iteration includes three curves obtained by applying the following tolerances for the inner solves.

- *Dash-dot curve with* $\bigcirc$ — a small fixed tolerance $\tau = 10^{-3}$.
- *Dotted line with* $\Diamond$ — a decreasing sequence of tolerances $\tau_k^{(0)} = \mathcal{O}(\|e_k^{(0)}\|)$.
- *Solid line with* $\bigtriangledown$ — exact inner solve ($\tau = 0$).

We can see from the figure that the slope of the linear interpolation of the curve with $\bigcirc$ markers is roughly 1, and the slopes of the lines with $\Diamond$ and $\bigtriangledown$ markers are approximately 2. This observation is consistent with the convergence analysis of the standard inexact inverse iteration (Theorem 3.5): the local convergence of Algorithm 3.1 is linear or quadratic, respectively, if an appropriately small fixed tolerance or a decreasing sequence of tolerance is used. The performance of this algorithm for all test problems is summarized in Table 5.2. The right part of the table gives the estimated order of convergence. The integers in parenthesis are the total numbers of the pairs in the sequence $\left\{\left(\log \|e_k^{(0)}\|, \log \|e_k^{(1)}\|\right)\right\}$ used for each solution strategy to estimate the convergence rates. Similar to the results for *butterfly*, all the estimated order of convergence are consistent with Theorem 3.5. In particular, since the standard inexact inverse iteration obtains new eigenvalue approximations from the standard Newton scheme (instead of using the Rayleigh functional), only quadratic convergence can be achieved, no matter if the local symmetry of $T(\lambda)$ is present or not.

The convergence results for IRQI (Algorithm 3.2) are shown in Figure 5.1 and Table 5.3. Since the Rayleigh functional $\rho_F(x)$ depends only on the eigenvector approximation, the initial eigenvalue approximation $\mu^{(0)}$ has no impact on the convergence rates. For this algorithm, we define $e^{(i)} = \tan \angle(x^{(i)}, v)$, which is proportional to the generalized tangent; see (2.18). The three curves of $\left\{\left(\log |e_k^{(0)}|, \log |e_k^{(1)}|\right)\right\}$ in the top right part of Figure 5.1 are obtained by using the same types of tolerances applied to the standard inverse iteration. We see from the figure that the slope of the curve with $\bigcirc$ markers is approximately 2, and the slopes of the lines with $\Diamond$ and $\bigtriangledown$ markers are roughly 3. These slopes, implying quadratic and cubic convergence respectively, are consistent with Theorem 3.6: with a small fixed tolerance for the inner solves, IRQI converges quadratically or linearly, respectively, if the local symmetry of $T(\lambda)$ is present or not; with a decreasing tolerance $\tau = \mathcal{O}(|e|)$, IRQI converges cubically or quadratically, depending on the presence of the local symmetry. The estimated orders of convergence for all problems are given in Table 5.3, and they are consistent with Theorem 3.6. We see that IRQI achieves lower order of convergence for problems **plasma_drift** and **railtrack2**, because the local symmetry of $T(\lambda)$ is absent, and the one-sided Rayleigh functional is computed. In addition, it is worth noting that *exact* RQI converges quartically for the problems *loaded_string* and *fiber*. We found that quartic convergence for these two problems can also be achieved by IRQI with a more stringent sequence of tolerance $\tau_k^{(0)} = \mathcal{O}(|e_k^{(0)}|^2)$.

To study the inexact residual inverse iteration (Algorithm 4.1), recall from Section 4.1 that the error is defined as $\|e^{(i)}\| = \left\| \frac{1}{u^H x^{(i)}} x^{(i)} - v \right\|$. We first refer to the left bottom part of Figure 5.1, where the curves of $\left\{\left(\log \|e_k^{(0)}\|, \log \|e_k^{(1)}\|\right)\right\}$ are obtained by using the following four types of shifts and tolerances.

prob	initial parameters			order of convergence		
	vec err	val err	rel tol	fixed tol	tol $= \mathcal{O}(e)$	exact
BF	$2.5e{-3}$	$4e{-3}$	$5e{-3}$	1.024 (20)	2.031 (14)	1.988 (14)
FB	$5e{-1}$	$4e{-3}$	$8e{-1}$	0.916 (16)	1.971 (9)	2.013 (9)
GN	$2e{-4}$	$2e{-3}$	$5e{-4}$	1.130 (18)	2.095 (12)	2.001 (12)
LS	$1e{-1}$	$1e{-3}$	$1e{-2}$	1.045 (15)	2.189 (9)	2.136 (9)
PD	$1e{-3}$	$1e{-3}$	$5e{-1}$	1.013 (20)	2.081 (10)	2.000 (10)
RT	$1e{-2}$	$5e{-3}$	$1e{-3}$	0.974 (22)	1.940 (15)	2.001 (15)
SD	$4e{-4}$	$1e{-2}$	$5e{-1}$	1.119 (20)	1.981 (13)	2.061 (13)
SP	$5e{-4}$	$1e{-6}$	$1e{-4}$	1.084 (18)	1.904 (12)	1.997 (12)

TABLE 5.2
Initial iterates and orders of local convergence rates of standard inexact inverse iteration (Algorithm 3.1)

prob	initial parameters		order of convergence		
	vec err	rel tol	fixed tol	tol $= \mathcal{O}(e)$	exact
BF	$1e{-3}$	$5e{-1}$	2.040 (11)	3.041 (8)	3.036 (8)
FB	$3e{-4}$	$1e{-1}$	2.042 (11)	3.071 (9)	**4.070** (7)
GN	$5e{-3}$	$1e{-2}$	2.086 (10)	2.940 (8)	2.995 (8)
LS	$4e{-4}$	$1e{-1}$	1.997 (15)	3.011 (12)	**3.937** (10)
PD	$2e{-5}$	$1e{-1}$	1.044 (20)	2.114 (9)	2.118 (9)
RT	$5e{-4}$	$2e{-3}$	1.165 (16)	2.047 (10)	1.993 (10)
SD	$1e{-4}$	$5e{-1}$	2.082 (13)	3.037 (9)	3.001 (9)
SP	$2e{-2}$	$5e{-1}$	2.206 (12)	3.177 (9)	3.011 (9)

TABLE 5.3
Initial iterates and orders of local convergence rates of IRQI (Algorithm 3.2)

- *Dash-dot line with* $\bigcirc$ — fixed σ with $\frac{|\sigma-\lambda|}{|\lambda|} = 5 \times 10^{-3}$; $\tau = 10^{-6}$.
- *Dotted line with* $\Diamond$ — fixed σ with $\frac{|\sigma-\lambda|}{|\lambda|} = 5 \times 10^{-4}$; $\tau = 10^{-7}$.
- *Dashed line with* $\bigtriangledown$ — fixed σ with $\frac{|\sigma-\lambda|}{|\lambda|} = 5 \times 10^{-5}$; $\tau = 10^{-8}$.
- *Solid line with* $\bigtriangleup$ — fixed σ with $\frac{|\sigma-\lambda|}{|\lambda|} = 5 \times 10^{-5}$; exact solve ($\tau = 0$).

Here, we used 16 pairs of $\left(\log \|e_k^{(0)}\|, \log \|e_k^{(1)}\|\right)$ to evaluate the convergence rates; see Table 5.4. It is clear from the figure that the slopes of all the lines are approximately 1, implying linear convergence of this algorithm. In this situation, the convergence factor is a more descriptive criterion to evaluate the performance. We can see from Table 5.4 that each time we decrease both $|\sigma - \lambda|$ and τ by a factor of 10, the convergence factor also decreases roughly by a factor of 10 (from 3.39×10^{-2} to 3.76×10^{-3} and 3.83×10^{-4}). This pattern holds approximately for all the problems, and is consistent with Theorem 4.2. In addition, we see that the inexact inverse iteration converges superlinearly for problems *loaded_string* and *schrödinger*, and it converges *superquadratically* for the problem $fiber$, for which the convergence factor C becomes less important. As we already explained, these results also provide some insight into the local convergence of the nonlinear Arnoldi method, which arises from a combination of the inexact residual inverse iteration and subspace projection.

Finally, the performance of the single-vector Jacobi-Davidson method (Algorithm 4.3) is shown in Figure 5.1 and Table 5.5. The results are parallel to those of IRQI. We use the same error $e^{(i)} = \tan \angle(x^{(i)}, v)$ as that for IRQI. In the bottom right part of Figure 5.1, the four curves of $\left\{\left(\log |e_k^{(0)}|, \log |e_k^{(1)}|\right)\right\}$ are obtained by using the following four types of tolerances.

- *Dash-dot curve with* $\bigcirc$ — a small fixed tolerance $\tau = 10^{-2}$.
- *Dotted line with* $\Diamond$ — a decreasing sequence of tolerances $\tau_k^{(0)} = \mathcal{O}(|e_k^{(0)}|)$.

prob	parameters			init_dist		0.1×init_dist		0.01×init_dist	
	init	init	#	tol = 1e−6		tol = 1e−7		tol = 1e−8	
	vec err	dist	pts	order	factor	order	factor	order	factor
BF	$5e{-3}$	$5e{-3}$	16	1.001	$3.39e{-2}$	1.000	$3.76e{-3}$	1.000	$3.83e{-4}$
FB	$8e{-4}$	$5e{-4}$	10	2.597	$8.60e{+2}$	2.665	$1.42e{+2}$	2.701	$1.84e{+1}$
GN	$2e{-4}$	$5e{-3}$	14	1.000	$2.18e{-3}$	1.000	$2.20e{-4}$	1.000	$2.21e{-5}$
LS	$4e{-3}$	$1e{-3}$	11	1.100	$1.22e{-3}$	1.209	$2.46e{-4}$	1.272	$3.69e{-5}$
PD	$5e{-3}$	$2e{-3}$	12	1.056	$2.97e{-3}$	1.058	$2.98e{-4}$	1.062	$3.05e{-5}$
RT	$5e{-3}$	$5e{-3}$	16	0.997	$7.25e{-1}$	0.990	$6.97e{-2}$	0.998	$7.36e{-3}$
SD	$1e{-4}$	$1e{-3}$	9	1.084	$1.46e{-2}$	1.366	$1.33e{-3}$	1.581	$1.08e{-4}$
SP	$5e{-1}$	$5e{-7}$	10	0.967	$4.65e{-3}$	0.957	$6.15e{-4}$	0.993	$7.03e{-5}$

TABLE 5.4
Initial iterates and orders of local convergence rates of residual inverse iteration (Algorithm 4.1)

prob	initial parameters		order of convergence			
	vec err	rel tol	fixed tol	tol = $\mathcal{O}(e)$	tol = $\mathcal{O}(e^2)$	exact
BF	$2e{-3}$	$1e{-2}$	1.010 (18)	1.869 (14)	2.967 (10)	3.080 (10)
FB	$3e{-4}$	$1e{-2}$	1.013 (18)	2.030 (13)	3.013 (10)	**4.070** (7)
GN	$5e{-3}$	$1e{-4}$	0.982 (16)	2.003 (12)	3.124 (9)	2.994 (8)
LS	$4e{-4}$	$1e{-3}$	0.992 (15)	1.987 (12)	2.965 (9)	**3.983** (6)
PD	$2e{-5}$	$5e{-4}$	1.043 (20)	1.971 (9)	–	2.026 (9)
RT	$5e{-4}$	$2e{-3}$	1.000 (16)	2.042 (10)	–	1.993 (10)
SD	$1.5e{-5}$	$1e{-6}$	1.030 (14)	2.068 (9)	2.953 (6)	2.981 (6)
SP	$2.5e{-3}$	$1e{-3}$	1.000 (20)	1.946 (15)	2.967 (10)	3.000 (9)

TABLE 5.5
Initial iterates and orders of local convergence rates of the single-vector JD method (Algorithm 4.3)

- *Dashed line with* $\bigtriangledown$ — a decreasing sequence of tolerances $\tau_k^{(0)} = \mathcal{O}(|e_k^{(0)}|^2)$.
- *Solid line with* $\bigtriangledown$ — exact inner solve ($\tau = 0$).

Clearly, the slopes of the four lines are roughly 1, 2, 3 and 3, respectively, implying linear, quadratic, and cubic convergence. The estimated orders of convergence for all test problems are given in Table 5.5, and they are consistent with the convergence analysis of single-vector JD; see Theorem 4.4. Note that for the two problems **plasma_drift** and **railtrack2** where the local symmetry of $T(\lambda)$ does not exist, single-vector JD achieves only quadratic convergence. In addition, exact JD converges quartically for the problems *loaded_string* and $fiber$, as exact RQI does. We found that with a more stringent decreasing sequence of tolerances $\tau_k^{(0)} = \mathcal{O}(|e_k^{(0)}|^3)$, quartic convergence can also be achieved by single-vector JD.

We complete this section by a comment on the computation of the Rayleigh functional $\rho_F(x)$, which may have considerable impact on the efficiency of all the algorithms discussed except for the standard inverse iteration. In practice, $\rho_F(x)$ needs to be computed by solving a nonlinear equation $p^H T(\rho)x = 0$, typically by Newton's method. This process can be expensive if one uses the iteration scheme $\rho^{(k+1)} = \rho^{(k)} - (p^H T(\rho^{(k)})x)/(p^H T'(\rho^{(k)})x)$ directly. If $T(\mu)$ is of the form $T(\mu) = \sum_{i=1}^m f_i(\mu)A_i$ where $A_i \in \mathbb{C}^{n\times n}$ $(1 \le i \le m)$ are fixed matrices, we can compute the scalars $p^H A_i x$ only once, and apply

$$\rho^{(k+1)} = \rho^{(k)} - \frac{\sum_{i=1}^m (p^H A_i x) f_i(\rho^{(k)})}{\sum_{i=1}^m (p^H A_i x) f_i'(\rho^{(k)})}.$$

The computational cost of this approach is much cheaper for all the test problems. In addition, if $T(\mu)$ is a matrix polynomial of degree less than 5, $\rho_F(x)$ can be computed directly by the formula for the roots of polynomial equations.

6. Conclusions. We presented detailed local convergence analyses of several inexact Newton-type algorithms for the solution of a simple eigenpair of the general nonlinear eigenvalue problem (2.1). We investigated the standard inexact inverse iteration, inexact Rayleigh quotient iteration, inexact residual inverse iteration and the single-vector Jacobi-Davidson method, showing how the errors of inner solves affect the order of local convergence. We show that if an appropriate sequence of tolerances is used for the inner solve, the inexact algorithms can achieve the same order of convergence as the exact methods, and lower order of convergence may be obtained by applying certain suitable and less stringent tolerances. These results are illustrated by numerical experiments.

Acknowledgement

We thank Volker Mehrmann for useful comments and pointers to the literature.

Appendix

In this appendix, we discuss the local convergence of flexible inexact inverse iteration (Algorithm 3.2) and single-vector JD (Algorithm 4.3) with a fixed shift $\mu_1^{(i)} = \sigma$ and a variable $\mu_2^{(i)}$. We study the flexible inexact inverse iteration, showing that the method may converge for linear eigenvalue problems, but not for general NEPs, due to a major difference between the two classes of problems. We also show that single-vector JD converges linearly if the local symmetry of $T(\lambda)$ exists and p is chosen as in (2.2) for the Rayleigh functional $\rho_F(x)$.

Part A: Flexible inexact inverse iteration

We first study the flexible inexact inverse iteration with a fixed shift. Consider the linear eigenvalue problem $Av = \lambda Bv$, for which $T(\mu) = \mu B - A$. To simplify the discussion, assume that there exists a complete set of right and left eigenvectors V and W^H, such that $AV = BV\Lambda$ and $W^H A = \Lambda W^H B$. By imposing the normalization condition $W^H BV = I$, we have $B = W^{-H}V^{-1}$ and $A = W^{-H}\Lambda V^{-1}$; see, e.g., [3]. Suppose that (λ_1, v_1) is the desired simple eigenpair. It follows that the resolvent is

$$\begin{aligned} & T(\mu)^{-1} = (\mu B - A)^{-1} \\ = & \left(W^{-H}(\mu I - \Lambda)V^{-1}\right)^{-1} = V(\mu I - \Lambda)^{-1}W^H \\ = & \frac{1}{\mu - \lambda_1} v_1 w_1^H + V \mathrm{diag}\left([0, (\mu - \lambda_2)^{-1}, ..., (\mu - \lambda_n)^{-1}]\right) W^H \\ \equiv & \frac{1}{\mu - \lambda_1} v_1 w_1^H + F(\mu), \end{aligned}$$

where $F(\mu) = V \mathrm{diag}\left([0, (\mu - \lambda_2)^{-1}, ..., (\mu - \lambda_n)^{-1}]\right) W^H$ is analytic in a neighborhood of λ_1.

To begin the analysis, consider the approximate solution y obtained in Step 1 of Algorithm 3.2. To establish the convergence, it is necessary to show that the generalized tangent of $\angle(y, v)$ is bounded above by $\mathcal{O}(t)$; see (3.11). To this end, since we have $\beta = \mathcal{O}(1)$ for a fixed $\mu_1 = \sigma$ from (3.6), it is enough to show that g_y defined in (3.7) satisfies $\|g_y\| = \mathcal{O}(t)$.

The key observation is that $\|g_y\| = \mathcal{O}(t)$ holds only for linear eigenvalue problems. To see this, note that from (3.7) that $\|g_y\|$ depends on the magnitude of $F'(\lambda_1)T'(\mu_2)x$, $F(\mu_1)T''(\lambda)x$ as well as $F(\mu_1)res$, which depends on the tolerance τ; see (3.2) – (3.4). If $\tau = \mathcal{O}(t)$, it is easy to show that γ_4, the generalized norm of $F(\mu_1)res$, satisfies $\gamma_4 = \mathcal{O}(t)$. For linear eigenvalue problems, since $T''(\mu) = 0$ for any μ, we have $F(\mu_1)T''(\lambda)x = 0$. In addition, we can show that the generalized norm of $F'(\lambda_1)T'(\mu_2)x$ is bounded by $\mathcal{O}(t)$. In fact, assume without loss

of generality that $x = cv_1 + sg$ ($\gamma = 1$; see (2.15)), or equivalently, $x = V(ce_1 + se_g)$, where $e_1 = [1, 0, ..., 0]^T$ and $e_g \in span\{e_2, e_3, ..., e_n\}$ such that $g = Ve_g$. Therefore, we have

$$\begin{aligned} & F'(\lambda_1)T'(\mu_2)x \\ &= V\mathrm{diag}\left([0, -(\lambda_1-\lambda_2)^{-2}, ..., -(\lambda_1-\lambda_n)^{-2}]\right) W^H BV(ce_1 + se_g) \\ &= V\mathrm{diag}\left([0, -(\lambda_1-\lambda_2)^{-2}, ..., -(\lambda_1-\lambda_n)^{-2}]\right) (ce_1 + se_g) \quad (W^H BV = I) \\ &= sV\mathrm{diag}\left([0, -(\lambda_1-\lambda_2)^{-2}, ..., -(\lambda_1-\lambda_n)^{-2}]\right) e_g. \end{aligned}$$

In summary, if $\tau = \mathcal{O}(t)$, then $\gamma_4 = \mathcal{O}(t)$; in addition, we have $\gamma_2 = \mathcal{O}(t)$ and $\gamma_3 = 0$ for linear eigenvalue problems. It follows that $\|g_y\| = \mathcal{O}(t)$, and thus the generalized tangent of $\angle(y, v)$ is also bounded above by $\mathcal{O}(t)$; see (3.11), where $\beta = \mathcal{O}(1)$. This is a necessary condition for the linear convergence of Algorithm 3.2 with a fixed shift for linear eigenvalue problems. In fact, with the assumption that $\mu_1 = \sigma$ is close to λ_1, and $\tau \le Ct$ for some small constant C, the linear convergence has been established; see, for example, [5, 14].

For general NEPs, however, it is easy to show that $\gamma_2 = \mathcal{O}(\sigma - \lambda) = \mathcal{O}(1)$ for the fixed shift $\mu_1 = \sigma$, and therefore $\|g_y\| = \mathcal{O}(1)$. Assume that the error of the inner solve is small such that (3.8) is satisfied, then the generalized tangent of $\angle(y, v)$ is bounded *below* by

$$\frac{\|g_y\| - \mathcal{O}\left(\dfrac{|\mu_2 - \lambda|^3}{|\mu_1 - \lambda|}\right)}{\|\beta v\| + \mathcal{O}\left(\dfrac{|\mu_2 - \lambda|^3}{|\mu_1 - \lambda|}\right)} = \mathcal{O}\left(\frac{|\sigma - \lambda|^2}{d + \mathcal{O}(\sigma - \lambda)}\right),$$

which is a small constant independent of t. Therefore, there is no reason to expect Algorithm 3.2 with fixed $\mu_1 = \sigma$ to converge for the general NEP (2.1). In fact, we have observed in experiments consistent failure of convergence.

Part B: Single-vector JD

Fortunately, the single-vector JD method with fixed $\mu_1 = \sigma$ converges linearly when the local symmetry of $T(\lambda)$ is present and p is chosen as in (2.2) for $\rho_F(x)$. Consider Algorithm 4.3 where $\mu_1^{(i)} = \sigma$ and $\mu_2^{(i)} = \rho_F(x^{(i)})$, for which the correction equation is

$$\Pi_1^{(i)} T(\sigma) \Pi_2^{(i)} \Delta x_{JD}^{(i)} = -T(\rho^{(i)})x^{(i)}, \qquad \text{with } \Delta x_{JD}^{(i)} \perp u^{(i)} \ . \tag{.1}$$

An analysis of this algorithm can be performed by exploring a connection between (.1) and the correction equation $T(\sigma)\Delta x_{RII}^{(i)} = -T(\rho^{(i)})x^{(i)}$ arising in the inexact residual inverse iteration. From now on, we drop the superscripts (i) since this does not lead to confusion. Suppose that (.1) is solved approximately, i.e.,

$$\Pi_1 T(\sigma) \Pi_2 \Delta x_{JD} = -T(\rho)x - res_{JD}, \qquad \text{with } \Delta x_{JD} \perp u. \tag{.2}$$

It can be shown that the corresponding approximate solution is

$$\begin{aligned} \Delta x_{JD} = & \frac{u^H T^{-1}(\sigma)\left(T(\rho)x + res_{JD}\right)}{u^H T^{-1}(\sigma) T'(\rho) x} T^{-1}(\sigma)T'(\rho)x \\ & - T^{-1}(\sigma)(T(\rho)x + res_{JD}). \end{aligned} \tag{.3}$$

Note that $T^{-1}(\sigma)$ is bounded since σ is bounded away from λ. Therefore, if a small tolerance $\tau = C_{12}|\sigma - \lambda|$ is applied to the correction equation (.1) such that $\|res_{JD}\| \le C_{12}|\sigma - \lambda|\|T(\rho)x\|$, it follows from (.3) that

$$\|\Delta x_{JD}\| = \mathcal{O}(1 + (\sigma - \lambda))\|T(\rho)x\|. \tag{.4}$$

We can rewrite (.2) as

$$T(\sigma)\Delta x_{JD} = -T(\rho)x - res_{JD} + \frac{p^H T(\sigma)\Delta x_{JD}}{p^H T'(\rho)x} T'(\rho)x.$$

To establish the connection between single-vector JD and residual inverse iteration, define $res_{RII} = res_{JD} - \dfrac{p^H T(\sigma)\Delta x_{JD}}{p^H T'(\sigma)x} T'(\rho)x$. To study the magnitude of this residual, first note that if the local symmetry of $T(\lambda)$ is present, then p chosen as in (2.2) is an approximate left eigenvector of $T(\lambda)$. We have

$$\begin{aligned} (.5) \qquad & \|p^H T(\sigma)\| = \|T(\sigma)x\| \\ & = \|T(\lambda)(cv + sg) + (\sigma - \lambda)T'(\lambda)x\| + \mathcal{O}(|\sigma - \lambda|^2) \\ & = s\|T(\lambda)g\| + |\sigma - \lambda|\|T'(\lambda)x\| + \mathcal{O}(|\sigma - \lambda|^2) \le C_{13}|\sigma - \lambda|, \end{aligned}$$

for some constant C_{13}, which depends on $T(\cdot)$. Therefore, if $\tau_{JD} \le C_{12}|\sigma - \lambda|$, we have from (.4) and (.5) that

$$\begin{aligned} \|res_{RII}\| &= \left\| res_{JD} - \frac{p^H T(\sigma)\Delta x_{JD}}{p^H T'(\sigma)x} T'(\rho)x \right\| \\ &\le \|res_{JD}\| + \frac{\|p^H T(\sigma)\|\|\Delta x_{JD}\|}{|p^H T'(\rho)x|}\|T'(\rho)x\| \\ &\le \|T(\rho)x\| C_{12}|\sigma - \lambda| + \frac{C_{13}|\sigma - \lambda|\mathcal{O}(1 + (\sigma - \lambda))\|T(\rho)x\|}{|p^H T'(\rho)x|}\|T'(\rho)x\| \\ &\le C_{14}|\sigma - \lambda|\|T(\rho)x\|. \end{aligned}$$

In other words, the approximate solution Δx_{JD} to the JD correction equation (.1) is an approximate solution of the residual inverse iteration correction equation $T(\sigma)\Delta x_{RII} = -T(\rho)x$ with the residual vector res_{RII} satisfying $\|res_{RII}\| \le \tau_{RII}\|T(\rho)x\| = C_{14}|\sigma - \lambda|\|T(\rho)x\|$ for some constant C_{14}. If x is sufficiently close to v in direction, and σ is close enough to λ, it then follows from Theorem 4.2 that single-vector JD with fixed $\mu_1^{(i)} = \sigma$ converges linearly with an appropriately small fixed inner solve tolerance $\tau_{JD} = C_{12}|\sigma - \lambda|$, and the convergence factor is proportional to $|\sigma - \lambda|$.

Note that the presence of the local symmetry is fundamental in the proof of convergence of single-vector JD with a fixed shift, but it is not necessary to prove the convergence of the inexact residual inverse iteration (Algorithm 4.1). Therefore, though single-vector JD with a fixed shift avoids the potential stagnation that may arise for residual inverse iteration, the sufficient condition for its linear convergence discovered here is more stringent than that for the latter.

REFERENCES

[1] E.N. Antoniou and S. Vologiannidis, *A new family of companion forms of polynomial matrices*, Electronic Journal of Linear Algebra , Vol. 11 (2004), pp. 78–87.

[2] Z. Bai and Y. Su, *SOAR: a second-order Arnoldi method for the solution of the quadratic eigenvalue problem*, SIAM Journal on Matrix Analysis and Applications, Vol. 26 (2005), pp. 640–659.

[3] J. Berns-Müller and A. Spence, *Inexact inverse iteration with variable shift for nonsymmetric generalized eigenvalue problems*, SIAM Journal on Matrix Analysis and Applications, Vol. 28 (2006), pp. 1069–1082.

[4] J. Berns-Müller, I. G. Graham, and A. Spence, *Inexact inverse iteration for symmetric matrices*, Linear Algebra and its Applications, Vol. 416 (2006), pp. 389–413.

[5] J. Berns-Müller, *Inexact Inverse Iteration Using Galerkin Krylov Solvers*, Ph.D Thesis, Department of Mathematics, University of Bath, United Kingdom, 2003.
[6] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, and F. Tisseur, *NLEVP: A Collection of Nonlinear Eigenvalue Problems*, MIMS EPrint 2010.98, School of Mathematics, University of Manchester, November 2010.
[7] T. Betcke and H. Voss, *A Jacobi-Davidson type projection method for nonlinear eigenvalue problems*, Future Generation Computer Systems, Vol. 20 (2004), pp. 363–372.
[8] D. Day and T. Walsh, *Quadratic eigenvalue problems*, Sandia Report, SAND2007-2072, Sandia National Lab., April 2007.
[9] R. S. Dembo, S. C. Eisenstat and T. Steihaug, *Inexact Newton methods*, SIAM Journal on Numerical Analysis, Vol. 19 (1982), pp. 400–408.
[10] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Vol. 16 of Classics in Applied Mathematics, SIAM, Philadelphia, PA, 1996.
[11] M.A. Freitag and A. Spence, *Convergence theory for inexact inverse iteration applied to the generalised nonsymmetric eigenproblem*, Electronic Transactions on Numerical Analysis, Vol. 28 (2007), pp. 40–64.
[12] M.A. Freitag and A. Spence, *Convergence rates for inexact inverse iteration with application to preconditioned iterative solves*, BIT Numerical Mathematics, Vol. 47 (2007), pp. 27–44.
[13] M.A. Freitag and A. Spence, *A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems*, IMA Journal on Numerical Analysis, Vol. 28 (2008), pp. 522–551.
[14] M.A. Freitag, *Inner-Outer Iterative Methods for Eigenvalue Problems - Convergence and Preconditioning*, Ph.D Thesis, Department of Mathematics, University of Bath, United Kingdom, 2007.
[15] I. Gohberg, P. Lancaster and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[16] C.-H. Guo, N. Higham and F. Tisseur, *Detecting and solving hyperbolic quadratic eigenvalue problems*, SIAM Journal on Matrix Analysis and Applications, Vol. (2008), pp. 1593–1613.
[17] C.-H. Guo and P. Lancaster, *Algorithms for hyperbolic quadratic eigenvalue problems*, Mathematics of Computation, Vol. 74 (2005), pp. 1777–1791.
[18] G. Hechme, *Convergence analysis of the Jacobi-Davidson method applied to a generalized eigenproblem*, Comptes rendus Mathématique, Académie des Sciences. Paris, Vol. 345 (2007), pp. 293–296.
[19] N. Higham, S. Mackey, F. Tisseur, and S. Garvey, *Scaling, sensitivity and stability in the numerical solution of quadratic eigenvalue problems*, International Journal for Numerical Methods in Engineering, Vol. 73 (2008), pp. 344–360.
[20] M.E. Hochstenbach, Y. Notay, *Controlling inner iterations in the Jacobi-Davidson method*, SIAM Journal on Matrix Analysis and Applications, Vol. 31(2009), pp. 460–477.
[21] T.-M. Hwang, W.-W. Lin and V. Mehrmann, *Numerical solution of quadratic eigenvalue problems with structure-preserving methods*, SIAM Journal on Scientific Computing, Vol. 24 (2003), pp. 1283–1302.
[22] E. Jarlebring and W. Michiels, *Analyzing the convergence factor of residual inverse iteration*, Preprint, Department of Computer Science, KU Leuven, 2011. To appear in BIT Numerical Mathematics.
[23] Z. Jia and W. Zeng, *A convergence analysis of the inexact Rayleigh quotient iteration and simplified Jacobi-Davidson method for the large Hermitian matrix eigenproblem*, Science in China Series A: Mathematics, Vol. 51 (2009), pp. 2205–2216.
[24] H. B. Keller, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in Applications of Bifurcation Theory, P. H. Rabinowitz, ed., Academic Press, New York, 1977, pp. 359–384.
[25] V. Kozlov and V. Maz'ya *Differential Equations With Operator Coefficients: With Applications to Boundary Value Problems for Partial Differential Equations*, Springer Monographs in Mathematics, Springer - Verlag, Berlin - Heidelberg, 1999.
[26] S. D. Mackey, N. Mackey, C. Mehl and V. Mehrmann, *Vector spaces of linearizations for matrix polynomials*, SIAM Journal on Matrix Analysis and Applications, Vol. 28 (2006), pp. 971–1004.
[27] K. Meerbergen, *The quadratic Arnoldi method for the solution of the quadratic eigenvalue problem*, SIAM Journal on Matrix Analysis and Applications, Vol. 30 (2008), pp. 1463–1482.
[28] V. Mehrmann, private communication, Householder Symposium XVIII, Tahoe City, California, June 2011.
[29] V. Mehrmann and C. Schröder, *Nonlinear eigenvalue and frequency response problems in industrial practice*, DFG Research Preprint, Mathematics for Key Technologies, Berlin, February 2011.
[30] V. Mehrmann and H. Voss, *Nonlinear eigenvalue problems: a challenge for modern eigenvalue methods*, Mitteilungen der Gesellschaft für Angewandte Mathematik und Mechanik, Vol. 27 (2004), pp. 121–152.
[31] A. Neumaier, *Residual inverse iteration for the nonlinear eigenvalue problem*, SIAM Journal on Numerical Analysis, Vol. 22 (1985), pp. 914–923.
[32] A. Ruhe, *A Rational Krylov algorithm for nonlinear matrix eigenvalue problems*, Rossiĭskaya Akademiya

Nauk. Sankt-Peterburgskoe Otdelenie. Matematicheskiĭ Institut im. V. A. Steklova. Zapiski Nauchnykh Seminarov (POMI), Vol. 268 (2000), pp. 176–180, translation in Journal of Mathematical Sciences (New York), Vol. 114 (2003), pp. 1854–1856.

[33] K. Schreiber, *Nonlinear Eigenvalue Problems: Newton-type Methods and Nonlinear Rayleigh Functionals*, Ph.D thesis, Department of Mathematics, TU Berlin, 2008.

[34] H. Schwetlicka and K. Schreiber, *Nonlinear Rayleigh functionals*, Linear Algebra and Its Applications, in press, doi:10.1016/j.laa.2010.06.048, available online 26 August 2010.

[35] P. Sonneveld and M. B. van Gijzen, *IDR(s): a family of simple and fast algorithms for solving large nonsymmetric linear systems*, SIAM Journal on Scientific Computing, Vol. 31 (2008), pp. 1035–1062.

[36] Y. Su and Z. Bai, *Solving rational eigenvalue problems via linearization*, SIAM Journal on Matrix Analysis and Applications, Vol. 32 (2011), pp. 201–216.

[37] D. B. Szyld, *Criteria for combining inverse and Rayleigh quotient iteration*, SIAM Journal on Numerical Analysis, Vol. 25 (1988), pp. 1369–1375.

[38] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Review, Vol. 43 (2001), pp. 235–286.

[39] J. van den Eshof, *The convergence of Jacobi-Davidson iterations for Hermitian eigenproblems*, Numerical Linear Algebra with Applications, Vol. 9 (2002), pp. 163–179.

[40] J. van den Eshof, *Nested Iteration Methods for Nonlinear Matrix Problems*, Ph.D Thesis, Department of Mathematics, Utrecht University, The Netherlands, 2003.

[41] F. Xue, *Numerical Solution of Eigenvalue Problems with Spectral Transformations*, Ph.D Thesis, Applied Mathematics, Statistics, and Scientific Computing, Department of Mathematics, University of Maryland, College Park, 2009.