SIAM J. Numer. Anal.
Vol. 47, No. 2, pp. 1445–1473

© 2009 Society for Industrial and Applied Mathematics

DYNAMICAL SYSTEMS AND NON-HERMITIAN ITERATIVE EIGENSOLVERS[*]

MARK EMBREE[†] AND RICHARD B. LEHOUCQ[‡]

Abstract. Simple preconditioned iterations can provide an efficient alternative to more elaborate eigenvalue algorithms. We observe that these simple methods can be viewed as forward Euler discretizations of well-known autonomous differential equations that enjoy appealing geometric properties. This connection facilitates novel results describing convergence of a class of preconditioned eigensolvers to the leftmost eigenvalue, provides insight into the role of orthogonality and biorthogonality, and suggests the development of new methods and analyses based on more sophisticated discretizations. These results also highlight the effect of preconditioning on the convergence and stability of the continuous-time system and its discretization.

Key words. eigenvalues, dynamical systems, inverse iteration, preconditioned eigensolvers, geometric invariants

AMS subject classifications. 15A18, 37C10, 65F15, 65L20

DOI. 10.1137/07070187X

1. Introduction.

Suppose we seek a small number of eigenvalues (and the associated eigenspace) of the non-Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, having at our disposal a nonsingular matrix $\mathbf{N} \in \mathbb{C}^{n\times n}$ that approximates $\mathbf{A}$. Given a starting vector $\mathbf{p}_0 \in \mathbb{C}^n$, compute

$$\mathbf{p}_{j+1} = \mathbf{p}_j + \mathbf{N}^{-1}(\theta_j - \mathbf{A})\mathbf{p}_j, \tag{1.1}$$

where $\theta_j - \mathbf{A}$ is shorthand for $\mathbf{I}\theta_j - \mathbf{A}$, and

$$\theta_j = \frac{(\mathbf{A}\mathbf{p}_j, \mathbf{p}_j)}{(\mathbf{p}_j, \mathbf{p}_j)}$$

for some inner product $(\cdot,\cdot)$. Knyazev, Neymeyr, and others have studied this iteration for Hermitian positive definite $\mathbf{A}$; see [21, 22] and references therein for convergence analysis and numerical experiments.

Clearly the choice of $\mathbf{N}$ will influence the behavior of this iteration. With $\mathbf{N} = \mathbf{A}$, the method (1.1) reduces to (scaled) inverse iteration:

$$\mathbf{p}_{j+1} = \mathbf{A}^{-1}\mathbf{p}_j\theta_j.$$

We are interested in the case where $\mathbf{N}$ approximates $\mathbf{A}$, yet one can apply $\mathbf{N}^{-1}$ to a vector much more efficiently than $\mathbf{A}^{-1}$ itself. Such a $\mathbf{N}$ acts as a preconditioner for $\mathbf{A}$, and, hence, (1.1) represents a preconditioned iteration.

[*] Received by the editors September 4, 2007; accepted for publication (in revised form) November 7, 2008; published electronically March 13, 2009.
http://www.siam.org/journals/sinum/47-2/70187.html

[†] Department of Computational and Applied Mathematics, Rice University, 6100 Main Street – MS 134, Houston, TX 77005-1892 (embree@rice.edu). This author's research supported by U.S. Department of Energy grant DE-FG03-02ER25531 and National Science Foundation grant DMS-CAREER-0449973.

[‡] Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185–1110 (rblehou@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract DE-AC04-94AL85000.

This method contrasts with a different class of algorithms, based on inverse iteration (or the shift-invert Arnoldi algorithm), that apply a preconditioner to accelerate an "inner iteration" that approximates the solution to a linear system at each step; see, e.g., [24, 13, 16] and [6, Chapter 11]. For numerous practical large-scale non-Hermitian eigenvalue problems, such as those described in [25, 41], these inner iterations can be extremely expensive and highly dependent on the quality of the preconditioner. In contrast, as we shall see, the iteration (1.1) can converge to a leftmost eigenpair even when $\mathbf{N}$ is a suitable multiple of the identity.

This paper provides a rigorous convergence theory that establishes sufficient conditions for (1.1) to converge to the leftmost eigenpair for non-Hermitian $\mathbf{A}$. We obtain these results by viewing this iteration as the forward Euler discretization of the autonomous nonlinear differential equation

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}\left(\mathbf{p}\frac{(\mathbf{Ap},\mathbf{p})}{(\mathbf{p},\mathbf{p})} - \mathbf{Ap}\right) \tag{1.2}$$

with a unit step size. Here $\mathbf{A}$ and $\mathbf{N}$ are fixed but $\mathbf{p}$ depends on a parameter, t; $\dot{\mathbf{p}}$ denotes differentiation with respect to t. In the absence of preconditioning, the differential equation (1.2) has been studied in connection with power iteration [10, 29], as described in more detail below. The nonzero steady-states of this system correspond to (right) eigenvectors of $\mathbf{A}$, and, hence, one might attempt to compute eigenvalues by driving this differential equation to steady-state as swiftly as possible. Properties of the preconditioner determine which of the eigenvectors is an attracting steady-state.

The differential equation (1.2) enjoys a distinguished property, observed, for example, in [10, 29] with $\mathbf{N} = \mathbf{I}$. Suppose that $\mathbf{p}$ solves (1.2), $\theta = (\mathbf{p},\mathbf{p})^{-1}(\mathbf{Ap},\mathbf{p})$, and $\mathbf{N}$ is self-adjoint and invertible ($\mathbf{A}$ may be non-self-adjoint). Then for all t,

$$\begin{aligned}\frac{d}{dt}(\mathbf{p},\mathbf{Np}) &= \big(\mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{Ap}),\mathbf{Np}\big) + \big(\mathbf{p},\mathbf{NN}^{-1}(\mathbf{p}\theta - \mathbf{Ap})\big)\\ &= (\mathbf{p}\theta,\mathbf{p}) - (\mathbf{Ap},\mathbf{p}) + (\mathbf{p},\mathbf{p}\theta) - (\mathbf{p},\mathbf{Ap})\\ &= 0. \end{aligned}\tag{1.3}$$

Thus, $(\mathbf{p},\mathbf{Np})$ is an *invariant* (or *first integral*), as its value is independent of time; see [19, section 1.3] for a discussion of the unpreconditioned case ($\mathbf{N} = \mathbf{I}$), and, e.g., [4, 18] for a general introduction to invariant theory and geometric integration.

The invariant describes a manifold in n-dimensional space, $(\mathbf{p},\mathbf{Np}) = (\mathbf{p}_0,\mathbf{Np}_0)$, on which the solution to the differential equation with $\mathbf{p}(0) = \mathbf{p}_0$ must fall. Simple discretizations, such as Euler's method (1.1), do not typically respect such invariants, giving approximate solutions that drift from the manifold. Invariant-preserving alternatives (see, e.g., [18, 26]) generally require significantly more computation per step (though a tractable method for the unpreconditioned, Hermitian case has been proposed by Nakamura, Kajiwara, and Shiotani [28]). Our goal is to explain the relationship between convergence and stability of the continuous and discrete dynamical systems. In particular, the quadratic invariant is a crucial property of the continuous system, and plays an important role in the convergence theory of the corresponding discretization, even when that iteration does not preserve the invariant.

For a non-Hermitian problem, one naturally wonders how (1.1) can be modified to incorporate estimates of both left and right eigenvectors. In this case, we obtain the coupled iteration (given here without preconditioning)

$$\begin{cases} \dot{\mathbf{p}} = \mathbf{p}\theta - \mathbf{A}\mathbf{p}, \\ \dot{\mathbf{q}} = \mathbf{q}\overline{\theta} - \mathbf{A}^*\mathbf{q}, \end{cases} \qquad \theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})}, \tag{1.4}$$

and a simple derivation reveals that $(\mathbf{p}, \mathbf{q})$ is invariant. Our analysis demonstrates that this two-sided dynamical system often suffers from finite-time blowup; in the discrete scheme this is tantamount to incurable breakdown, a well-known ailment of oblique projection methods (see [5] for a discussion and references to the literature within the context of non-Hermitian Lanczos methods).

A longstanding association exists between eigenvalue iterations and differential equations [1, 2, 3, 10, 11, 15, 19], often involving the observation that iterates of a particular eigenvalue algorithm are *exactly* discrete-time samples of some underlying continuous-time system. Notable examples include Rayleigh quotient gradient flow [10, 27], connections between the QR algorithm for dense eigenproblems and Toda flow [29, 39], and more general "isospectral flows" [42]. For example, Chu notes that the iterates of the standard power method can be obtained as integer-time samples of the solution to the system (1.2) with $\mathbf{N} = \mathbf{I}$ and $\mathbf{A}$ replaced by $\log \mathbf{A}$ [10, eq. (2.7)].

The present study draws upon this body of work, but takes a different perspective: we seek a better understanding of iterations such as (1.1) that provide only *approximate* solutions (with a truncation error due to discretization) to continuous time systems such as (1.2). The distinction is significant: for example, a continuous-time generalization of the power method will converge, with mild caveats, to the largest magnitude eigenvalue, whereas the related systems we study can potentially converge to the leftmost eigenvalue at a shift-independent rate with little more work per iteration than the power method; see Theorems 4.4 and 6.3.

The connection between eigensolvers and continuous-time dynamical systems also arises in applications. For example, the Car–Parrinello method [8] determines the Kohn–Sham eigenstates from a second-order ordinary differential equation, Newton's equations of motion (see [34, p. 1086] for a formulation using (1.2) with no preconditioning). The heavy ball optimization method [35] also formulates the minimum of the Rayleigh quotient via a second order ordinary differential equation. In [7], the ground state solution of Bose–Einstein condensates are determined via a normalized gradient flow discretized by several time integration schemes. (Both the Kohn–Sham eigenstates and Bose–Einstein condensates give rise to self-adjoint nonlinear eigenvalue problems.)

We begin our investigation with a study of various unpreconditioned iterations ($\mathbf{N} = \mathbf{I}$). Section 2 introduces basic differential equations for computation of invariant subspaces of matrix pencils, and then identifies parameter choices that yield invariant-preserving iterations. Near steady states, the solutions to these systems can be viewed as exact invariant subspaces for nearby matrices, as observed in section 3. From this point we focus on single vector iterations for standard eigenvalue problems. Section 4 describes exact solution formulas for two unpreconditioned continuous-time systems, one-sided and two-sided methods. As such exact solutions for the preconditioned case are elusive, we analyze such systems asymptotically using center manifold theory in section 5. These two sections provide the foundation for the main result of section 6, the development of sufficient conditions for convergence of (1.1) for non-Hermitian matrices.

2. Dynamical systems and invariant manifolds.

We first examine properties of the dynamical system (1.2) and various generalizations suitable for computing eigenvalues of non-Hermitian matrix pencils. Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}$ be general matrices with fixed (time-invariant) entries. For the generalized eigenvalue problem $\mathbf{Ax} = \mathbf{Bx}\lambda$ with $\mathbf{N} = \mathbf{I}$, the system (1.2) expands to

$$\dot{\mathbf{p}} = \mathbf{Bp}\theta - \mathbf{Ap}$$

for appropriate $\theta = \theta(t)$. This equation suggests a generalization from a system with the single vector $\mathbf{p} \in \mathbb{C}^n$ to a system that evolves an entire subspace, given by the range of a matrix $\mathbf{P} \in \mathbb{C}^{n\times k}$:

$$\dot{\mathbf{P}} = \mathbf{BPL} - \mathbf{AP},$$

where differentiation is still with respect to the autonomous variable t; we shall address the choice of $\mathbf{L}(t) \in \mathbb{C}^{k\times k}$ momentarily. (Quantities such as $\mathbf{L}$ are t-dependent unless explicitly stated otherwise; we typically suppress the t argument to simplify notation.)

For non-Hermitian problems one might simultaneously evolve an equation for the adjoint to obtain approximations to the left eigenspace, which suggests the system

$$\begin{aligned} \dot{\mathbf{P}} &= \mathbf{BPL} - \mathbf{AP} \\ \dot{\mathbf{Q}} &= \mathbf{B}^*\mathbf{QM}^* - \mathbf{A}^*\mathbf{Q}, \end{aligned} \tag{2.1}$$

with initial conditions $\mathbf{P}(0) = \mathbf{P}_0$ and $\mathbf{Q}(0) = \mathbf{Q}_0$, where $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{n\times k}$, and $\mathbf{L}, \mathbf{M} \in \mathbb{C}^{k\times k}$. The choice we make for the time-dependent $\mathbf{L}, \mathbf{M} \in \mathbb{C}^{k\times k}$ can potentially couple $\mathbf{P}$ and $\mathbf{Q}$ as introduced in (1.4). Here $\cdot^*$ denotes the conjugate transpose and $(\cdot, \cdot)$ the standard Euclidean inner product (though this analysis generalizes readily to arbitrary inner products). If this system is at a steady state, i.e., $\dot{\mathbf{P}} = \dot{\mathbf{Q}} = \mathbf{0}$, then

$$\mathbf{BPL} = \mathbf{AP}, \quad \mathbf{B}^*\mathbf{QM}^* = \mathbf{A}^*\mathbf{Q}, \tag{2.2}$$

and, hence, provided $\mathbf{P}$ and $\mathbf{Q}$ have full column rank, the eigenvalues of $\mathbf{L}$ and $\mathbf{M}$ are included in the spectrum of the pencil $\mathbf{A} - \lambda\mathbf{B}$, while the columns of $\mathbf{P}$ and $\mathbf{Q}$ span right- and left-invariant subspaces of the same pencil. We shall motivate the choice of $\mathbf{L}$ and $\mathbf{M}$ through generalizations of the invariant discussed in the introduction. The following notation facilitates the analysis of these subspace iterations.

Definition 2.1. *Given $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{n\times k}$, define $(\mathbf{P}, \mathbf{Q}) = \mathbf{Q}^*\mathbf{P} \in \mathbb{C}^{k\times k}$; i.e., the (i, j) entry of $(\mathbf{P}, \mathbf{Q})$ satisfies $(\mathbf{P}, \mathbf{Q})_{i,j} := (\mathbf{Pe}_j, \mathbf{Qe}_i)$, where $\mathbf{e}_\ell$ denotes the ℓth column of the $k \times k$ identity matrix.*

In this notation, we have the homogeneity property $(\mathbf{PL}, \mathbf{Q}) = \mathbf{Q}^*\mathbf{PL} = (\mathbf{P}, \mathbf{Q})\mathbf{L}$.

Consider the pairs of (time-dependent) functions

$$(\mathbf{Q}, \mathbf{P}), \quad (\mathbf{P}, \mathbf{Q}) \qquad \text{and} \qquad (\mathbf{P}, \mathbf{P}), \quad (\mathbf{Q}, \mathbf{Q}) \tag{2.3}$$

with derivatives

$$\frac{d}{dt}(\mathbf{Q}, \mathbf{P}) = \left(\dot{\mathbf{Q}}, \mathbf{P}\right) + \left(\mathbf{Q}, \dot{\mathbf{P}}\right), \quad \frac{d}{dt}(\mathbf{P}, \mathbf{Q}) = \left(\dot{\mathbf{P}}, \mathbf{Q}\right) + \left(\mathbf{P}, \dot{\mathbf{Q}}\right),$$

and

$$\frac{d}{dt}(\mathbf{P}, \mathbf{P}) = \left(\dot{\mathbf{P}}, \mathbf{P}\right) + \left(\mathbf{P}, \dot{\mathbf{P}}\right), \quad \frac{d}{dt}(\mathbf{Q}, \mathbf{Q}) = \left(\dot{\mathbf{Q}}, \mathbf{Q}\right) + \left(\mathbf{Q}, \dot{\mathbf{Q}}\right).$$

Inspired by (1.3), we next investigate how best to choose $\mathbf{L}$ and $\mathbf{M}$ to make either pair in (2.3) invariant under the system (2.1).

THEOREM 2.2. *For the system of ordinary differential equations* (2.1) *with initial conditions* $\mathbf{P}(0) = \mathbf{P}_0 \in \mathbb{C}^{n\times k}$ *and* $\mathbf{Q}(0) = \mathbf{Q}_0 \in \mathbb{C}^{n\times k}$*, the choices*

$$\mathbf{L} = (\mathbf{BP}, \mathbf{Q})^{-1}(\mathbf{AP}, \mathbf{Q}), \quad \mathbf{M}^* = (\mathbf{Q}, \mathbf{BP})^{-1}(\mathbf{Q}, \mathbf{AP}) \tag{2.4}$$

give

$$\frac{d}{dt}(\mathbf{P}, \mathbf{Q}) = \frac{d}{dt}(\mathbf{Q}, \mathbf{P}) = \mathbf{0},$$

and, hence, $(\mathbf{P}, \mathbf{Q}) = (\mathbf{P}_0, \mathbf{Q}_0)$ *and* $(\mathbf{Q}, \mathbf{P}) = (\mathbf{Q}_0, \mathbf{P}_0)$ *hold for all* t*.*

Proof. Note that

$$\begin{aligned}\frac{d}{dt}(\mathbf{P}, \mathbf{Q}) &= \left(\dot{\mathbf{P}}, \mathbf{Q}\right) + \left(\mathbf{P}, \dot{\mathbf{Q}}\right)\\ &= (\mathbf{BP}, \mathbf{Q})\mathbf{L} - (\mathbf{AP}, \mathbf{Q}) + \mathbf{M}(\mathbf{P}, \mathbf{B}^*\mathbf{Q}) - (\mathbf{P}, \mathbf{A}^*\mathbf{Q})\\ \left(\frac{d}{dt}(\mathbf{Q}, \mathbf{P})\right)^* &= \left(\mathbf{P}, \dot{\mathbf{Q}}\right) + \left(\dot{\mathbf{P}}, \mathbf{Q}\right)\\ &= \mathbf{M}(\mathbf{P}, \mathbf{B}^*\mathbf{Q}) - (\mathbf{P}, \mathbf{A}^*\mathbf{Q}) + (\mathbf{BP}, \mathbf{Q})\mathbf{L} - (\mathbf{AP}, \mathbf{Q}),\end{aligned}$$

where we have used (2.1) and the homogeneity property. We can force $(d/dt)(\mathbf{P}, \mathbf{Q})$ to zero by setting $\mathbf{L}$ and $\mathbf{M}$ as in (2.4). □

The next result is a direct analogue of Theorem 2.2 for the second pair in (2.3). We omit the proof, a minor adaptation of the last one.

THEOREM 2.3. *For the system of ordinary differential equations* (2.1) *with initial conditions* $\mathbf{P}(0) = \mathbf{P}_0 \in \mathbb{C}^{n\times k}$ *and* $\mathbf{Q}(0) = \mathbf{Q}_0 \in \mathbb{C}^{n\times k}$*, the choices*

$$\mathbf{L} = (\mathbf{BP}, \mathbf{P})^{-1}(\mathbf{AP}, \mathbf{P}), \quad \mathbf{M}^* = (\mathbf{Q}, \mathbf{BQ})^{-1}(\mathbf{Q}, \mathbf{AQ})$$

give

$$\frac{d}{dt}(\mathbf{P}, \mathbf{P}) = \frac{d}{dt}(\mathbf{Q}, \mathbf{Q}) = \mathbf{0},$$

and, hence, $(\mathbf{P}, \mathbf{P}) = (\mathbf{P}_0, \mathbf{P}_0)$ *and* $(\mathbf{Q}, \mathbf{Q}) = (\mathbf{Q}_0, \mathbf{Q}_0)$ *for all* t*.*

The formulations for $\mathbf{L}$ and $\mathbf{M}$ given in Theorems 2.2 and 2.3 are known as *generalized Rayleigh quotients* [38]. With these values of $\mathbf{L}$ and $\mathbf{M}$, we refer to (2.1) as the *two-sided* and *one-sided* dynamical systems. Theorem 2.2 shows that if $\mathbf{P}_0^*\mathbf{Q}_0 = \mathbf{I}$, then the two-sided solutions will preserve this property (allowing for biorthogonal bases for left and right invariant subspaces), though possibly at the expense of growing $\|\mathbf{P}\|$ or $\|\mathbf{Q}\|$. Theorem 2.3, on the other hand, shows that the one-sided iteration maintains $\|\mathbf{P}\|$ and $\|\mathbf{Q}\|$, though biorthogonality will generally be lost. From the invariants we also see that the system preserves the rank of solutions to both one- and two-sided equations—provided they exist (see section 4). Since $(\mathbf{P}, \mathbf{P})$ is fixed for the one-sided system, so too are all singular values (and, thus, the rank) of $\mathbf{P}$. For the two-sided system, if $(\mathbf{P}_0, \mathbf{Q}_0)$ is full rank, $(\mathbf{P}, \mathbf{Q})$ must always be as well, and, hence, $\mathbf{P}$ and $\mathbf{Q}$ individually have full rank.

We denote the dynamical systems (2.1) given the generalized Rayleigh quotients of Theorems 2.2 and 2.3 as “two-sided” and “one-sided”, respectively. We refer to the ensuing schemes that result from discretizing (2.1) as “two-sided” and “one-sided” iterations.

3. Invariants and backward stability. We saw in (2.2) that, at a steady state, the eigenvalues of $\mathbf{L}$ and $\mathbf{M}$ are exact eigenvalues of the pencil $\mathbf{A} - \lambda\mathbf{B}$. As the system *approaches* a steady state, how well do the eigenvalues of the invariant-preserving choices for $\mathbf{L}$ and $\mathbf{M}$ approximate the eigenvalues of the pencil?

First, consider the one-sided system, with $\mathbf{L}$ as given in Theorem 2.3 and $\mathbf{P}$ full rank. The first part of (2.1) can then be written as

$$\mathbf{0} = \mathbf{BPL} - \left(\mathbf{A} + \dot{\mathbf{P}}(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^*\right)\mathbf{P},$$

from which we see that the eigenvalues of $\mathbf{L}$ form a subset of the spectrum of the perturbed pencil $(\mathbf{A} + \dot{\mathbf{P}}(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^*) - \lambda\mathbf{B}$. How large can such perturbations be? Note that $(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^* = \mathbf{P}^+$ is the pseudoinverse of $\mathbf{P}$, and so

$$\left\|\dot{\mathbf{P}}(\mathbf{P},\mathbf{P})^{-1}\mathbf{P}^*\right\| \le \left\|\dot{\mathbf{P}}\right\| \|\mathbf{P}^+\| = \frac{\left\|\dot{\mathbf{P}}\right\|}{\sigma_k},$$

where σ_k is the smallest singular value of $\mathbf{P} \in \mathbb{C}^{n\times k}$. As discussed at the end of section 2, the choice of $\mathbf{L}$ in Theorem 2.3 that makes $(\mathbf{P},\mathbf{P})$ invariant also makes σ_k invariant. Thus, when $\|\dot{\mathbf{P}}\|$ is small, i.e., near a steady state, we conclude that the eigenvalues of $\mathbf{L}$ are the exact eigenvalues of a nearby pencil, with σ_k^{-1} acting as a condition number does in a backward error bound; that condition number can be set to one simply by taking $(\mathbf{P}_0,\mathbf{P}_0) = \mathbf{I}$. (This is related to an error bound for Rayleigh–Ritz eigenvalue estimates for a Hermitian matrix using a nonorthogonal basis; see [32, Theorem 11.10.1].) This analysis suggests that a departure from orthogonality in a numerical integration of the differential equation is reflected in degrading accuracy of the approximate eigenvalues.

Now consider the two-sided system with $\mathbf{L}$ and $\mathbf{M}$ as given by Theorem 2.2 with nonsingular $(\mathbf{BP},\mathbf{Q})$. We wish to rewrite (2.1) in the form

$$\begin{aligned}\mathbf{0} &= \mathbf{BPL} - (\mathbf{A}+\mathbf{E})\mathbf{P}\\ \mathbf{0} &= \mathbf{B}^*\mathbf{QM}^* - (\mathbf{A}^*+\mathbf{E}^*)\mathbf{Q}\end{aligned}$$

for the same $\mathbf{E}$ in both iterations. Lemma 1 of [20] implies that such a perturbation $\mathbf{E}$ exists if and only if

$$(\mathbf{BP},\mathbf{Q})\mathbf{L} = \mathbf{M}(\mathbf{BP},\mathbf{Q}),$$

which holds for the choice of $\mathbf{L}$ and $\mathbf{M}$ given in Theorem 2.2. The perturbation $\mathbf{E}$ is not unique, but $\mathbf{EP} = \dot{\mathbf{P}}$ and $\mathbf{E}^*\mathbf{Q} = \dot{\mathbf{Q}}$. Moreover, the "main theorem" of [20] gives

$$\min \|\mathbf{E}\|_2 = \max\left\{\left\|\dot{\mathbf{P}}\right\|_2, \left\|\dot{\mathbf{Q}}\right\|_2\right\}$$

if $(\mathbf{P},\mathbf{P}) = \mathbf{I}_k$ and $(\mathbf{Q},\mathbf{Q}) = \mathbf{I}_k$. However, as the authors of [20] explain, a small $\|\mathbf{E}\|_2$ is irrelevant unless $\|(\mathbf{P},\mathbf{Q})^{-1}\|_2$ is also small. In particular, when $\mathbf{P}$ is orthogonal to $\mathbf{Q}$, $\min\|\mathbf{E}\|_2$ is undefined. The discussion following Theorem 4.1 in subsection 4.1 explains that a large (or undefined) $\|(\mathbf{P},\mathbf{Q})^{-1}\|_2$ is equivalent to near breakdown (or serious breakdown) of the two-sided dynamical system.

We caution the reader that backward stability alone does not provide information on forward error, or accuracy, of the steady-states when $\mathbf{A} \ne \mathbf{A}^*$. The relevance of backward stability is that the solution of our one- and two-sided systems are, at all times, steady-states for a related dynamical system. The distance to this related perturbed system depends upon the norm of the residuals.

4. Convergence analysis.

At least for single-vector iterations (i.e., $k = 1$), the analysis of the one- and two-sided dynamical systems follows readily from the remarkable fact that, in many cases, simple formulas give the exact solutions of these nonlinear differential equations. This observation, inspired by a lemma of Nanda [29], informs convergence analysis of the eigeniterations that result from the discretization of these equations. Though expressed for the standard eigenvalue problem, these results can naturally be adapted to the generalized case by replacing $\mathbf{A}$ with $\mathbf{B}^{-1}\mathbf{A}$. We discuss the solution operators for two-sided systems, followed by one-sided systems.

4.1. Two-sided systems.

The following result generalizes a result of Nanda [29, Lemma 1.4] for the two-sided dynamical system.

Theorem 4.1. *Consider the partitioned set of ordinary differential equations*

$$\begin{aligned}\dot{\mathbf{p}} &= \mathbf{p}\theta - \mathbf{A}\mathbf{p}\\ \dot{\mathbf{q}} &= \mathbf{q}\bar{\theta} - \mathbf{A}^*\mathbf{q},\end{aligned} \tag{4.1}$$

with $\mathbf{p}(0) = \mathbf{p}_0$ *and* $\mathbf{q}(0) = \mathbf{q}_0$*, where* $\mathbf{p}, \mathbf{q} \in \mathbb{C}^n$*,* $(\mathbf{p}_0, \mathbf{q}_0) \neq 0$*, and*

$$\theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})}.$$

Then there exists some $t_\mathrm{f} > 0$ *such that for all* $t \in [0, t_\mathrm{f})$*,*

$$\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\pi(t), \quad \mathbf{q}(t) = e^{-\mathbf{A}^*t}\mathbf{q}_0\overline{\pi(t)},$$

where

$$\pi(t) = \sqrt{\frac{(\mathbf{p}_0, \mathbf{q}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}}. \tag{4.2}$$

Proof. We define $\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\pi(t)$ and $\mathbf{q}(t) = e^{-\mathbf{A}^*t}\mathbf{q}_0\overline{\pi(t)}$, and will show that these formulas satisfy the system (4.1). Note that

$$\begin{aligned}\dot{\pi} &= \frac{\pi}{2}\frac{\big((\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0) + (e^{-\mathbf{A}t}\mathbf{p}_0, \mathbf{A}^*e^{-\mathbf{A}^*t}\mathbf{q}_0)\big)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}\\ &= \pi\frac{(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0)}\\ &= \pi\frac{(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0\pi, e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi})}{(e^{-\mathbf{A}t}\mathbf{p}_0\pi, e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi})} \;=\; \pi\frac{(\mathbf{A}\mathbf{p}, \mathbf{q})}{(\mathbf{p}, \mathbf{q})} \;=\; \pi\theta.\end{aligned}$$

Differentiating the formulas for $\mathbf{p}$ and $\mathbf{q}$, thus, gives

$$\begin{aligned}\dot{\mathbf{p}} &= -\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0\pi + e^{-\mathbf{A}t}\mathbf{p}_0\dot{\pi} &&= -\mathbf{A}\mathbf{p} + \theta\mathbf{p}\\ \dot{\mathbf{q}} &= -\mathbf{A}^*e^{-\mathbf{A}^*t}\mathbf{q}_0\bar{\pi} + e^{-\mathbf{A}^*t}\mathbf{q}_0\dot{\bar{\pi}} &&= -\mathbf{A}^*\mathbf{q} + \bar{\theta}\mathbf{q},\end{aligned}$$

as required. The hypothesis that $(\mathbf{p}_0, \mathbf{q}_0) \neq 0$ ensures the existence of the solution at time $t = 0$. The formula will hold for all $t > 0$, until potentially

$$\left(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0\right) = 0. \tag{4.3}$$

We define t_{f} to be the smallest positive t for which (4.3) holds. If no such positive t exists, the solution exists for all $t > 0$ and we can take $t_{\mathrm{f}} = \infty$ in the statement of the theorem. ▯

Theorem 4.1 gives $(\mathbf{p}, \mathbf{q}) = (\mathbf{p}_0, \mathbf{q}_0)$, precisely as Theorem 2.2 indicates. Under the conditions of Theorem 4.1, solutions of the two-sided single-vector equations (4.1) have the same direction as solutions of the simpler linear systems $\dot{\mathbf{x}} = -\mathbf{A}\mathbf{x}$, $\mathbf{x}(0) = \mathbf{p}_0$ and $\dot{\mathbf{y}} = -\mathbf{A}^*\mathbf{y}$, $\mathbf{y}(0) = \mathbf{q}_0$, but the magnitudes of $\mathbf{p}$ and $\mathbf{q}$ vary nonlinearly with (4.2). In particular, the inner product of $\mathbf{p}$ and $\mathbf{q}$ can be zero—even with both $\mathbf{p}$ and $\mathbf{q}$ nonzero—leading to finite time blow-up of (4.1). Note that if

$$\left(\frac{e^{-\mathbf{A}t}\mathbf{p}_0}{\sqrt{(\mathbf{p}_0, \mathbf{q}_0)}}, \frac{e^{-\mathbf{A}^*t}\mathbf{q}_0}{\sqrt{(\mathbf{q}_0, \mathbf{p}_0)}} \right) = 0,$$

then $\pi(t)$ is undefined. Hence, finite time blow-up is analogous to *serious breakdown* [43, p. 389], a problem endemic to oblique projection methods (see, e.g., [5]). This ratio will be nonzero but small in the vicinity of blow-up (or *near-breakdown*), a situation that commonly occurs in discretizations of these equations. The salient issue is that $\mathbf{p}$ and $\mathbf{q}$ are nearly orthogonal and so

$$\frac{(\mathbf{p}, \mathbf{q})}{\|\mathbf{p}\|\,\|\mathbf{q}\|} = \left(\frac{e^{-\mathbf{A}t}\mathbf{p}_0}{\|e^{-\mathbf{A}t}\mathbf{p}_0\|}, \frac{e^{-\mathbf{A}^*t}\mathbf{q}_0}{\|e^{-\mathbf{A}^*t}\mathbf{q}_0\|} \right) \tag{4.4}$$

is a useful quantity to measure. This number is small when the secant of the angle between $\mathbf{p}$ and $\mathbf{q}$ is large. In section 6 we shall see the important consequences of these observations for eigensolvers derived from the discretization of (4.1).

One can avoid breakdown altogether by using starting vectors $\mathbf{p}_0$ and $\mathbf{q}_0$ that are sufficiently accurate approximations to the right and left eigenvectors of $\mathbf{A}$ associated with the leftmost eigenvalue. Suppose $\mathbf{A}$ is diagonalizable with a simple leftmost eigenvalue λ_1, and all other eigenvalues strictly to the right of λ_1. Thus, there exists invertible $\mathbf{X}$ and diagonal $\mathbf{\Lambda}$ such that

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$$

with $\mathbf{\Lambda}_{1,1} = \lambda_1$. Write $\lambda_j = \mathbf{\Lambda}_{j,j}$, so that $\operatorname{Re}\lambda_j > \operatorname{Re}\lambda_1$ for $j = 2, \ldots, n$. Define $\mathbf{r} = \mathbf{X}^{-1}\mathbf{p}_0$ and $\mathbf{s} = \mathbf{X}^*\mathbf{q}_0$; i.e., $\mathbf{r}$ and $\mathbf{s}$ are the expansions of the starting vectors in biorthogonal bases of right and left eigenvectors of $\mathbf{A}$.

THEOREM 4.2. *Under the setting established in the last paragraph, the condition*

$$|r_1 s_1| > \sum_{j=2}^{n} |r_j s_j|$$

is sufficient to ensure that the dynamical system (4.1) *has a solution for all* $t \geq 0$ *given by Theorem* 4.1*; i.e., no incurable breakdown occurs.*

Proof. First note that

$$\left(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0\right) = \left(\mathbf{X}e^{-\mathbf{\Lambda}t}\mathbf{X}^{-1}\mathbf{p}_0, \mathbf{X}^{-*}e^{-\mathbf{\Lambda}^*t}\mathbf{X}^*\mathbf{q}_0\right) = (e^{-2\mathbf{\Lambda}t}\mathbf{r}, \mathbf{s}) = \sum_{j=1}^{n} r_j \overline{s}_j e^{-2\lambda_j t}.$$

Since $\operatorname{Re}\lambda_1 < \operatorname{Re}\lambda_j$ for $j > 2$, we have $|e^{-2\lambda_1 t}| \geq |e^{-2\lambda_j t}|$ for all $t \geq 0$. The hypothesis involving $\mathbf{r}$ and $\mathbf{s}$, thus, implies, for $t \geq 0$, that

$$\left|r_1 s_1 e^{-2\lambda_1 t}\right| \geq \sum_{j=2}^{n} \left|r_j s_j e^{-2\lambda_j t}\right|.$$

Given this expression, we can twice apply the triangle inequality to conclude

$$\begin{aligned}0 &< \left|r_1\overline{s}_1 e^{-2\lambda_1 t}\right| - \sum_{j=2}^{n}\left|r_j\overline{s}_j e^{-2\lambda_j t}\right| \\ &\le \left|r_1\overline{s}_1 e^{-2\lambda_1 t}\right| - \left|\sum_{j=2}^{n} r_j\overline{s}_j e^{-2\lambda_j t}\right| \le \left|\sum_{j=1}^{n} r_j\overline{s}_j e^{-2\lambda_j t}\right| = \left|\left(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}^*t}\mathbf{q}_0\right)\right|.\end{aligned}$$

Hence, $\pi(t)$ in Theorem 4.1 is finite for all $t \ge 0$, ensuring that the solution to the dynamical system (4.1) does not blow up at finite time. □

Theorem 4.2 implies that finite-time blow-up (or serious breakdown) is not generic for (4.1). However, the sufficient condition provided suggests that excellent initial approximations to the leftmost (left and right) eigenvectors are needed.

4.2. One-sided systems. The single vector one-sided system possesses a similar exact solution, which has been studied in the context of gradient flows associated with Rayleigh quotient iteration. We shall see that finite-time blow-up is never a concern for such systems. The following is a modest restatement of a result of Nanda [29, Lemma 1.4] (who considers the differential equation acting on the unit ball in $\mathbb{R}^n$).

THEOREM 4.3. *Consider the ordinary differential equation*

$$\dot{\mathbf{p}} = \mathbf{p}\theta - \mathbf{A}\mathbf{p}, \tag{4.5}$$

with $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and initial condition* $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$*, where* $\mathbf{p}_0 \ne \mathbf{0}$ *and*

$$\theta = \frac{(\mathbf{A}\mathbf{p}, \mathbf{p})}{(\mathbf{p}, \mathbf{p})}.$$

Then for all $t \ge 0$*,* (4.5) *has the exact solution*

$$\mathbf{p}(t) = e^{-\mathbf{A}t}\mathbf{p}_0\omega(t),$$

where

$$\omega(t) = \sqrt{\frac{(\mathbf{p}_0, \mathbf{p}_0)}{(e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}t}\mathbf{p}_0)}}.$$

We omit the proof of this result, which closely mimics that of Theorem 4.1. Of course, a similar formula can be written for the one-sided equation for $\mathbf{q}(t)$. The restriction to real matrices guarantees that $(\mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0, e^{-\mathbf{A}t}\mathbf{p}_0) = (e^{-\mathbf{A}t}\mathbf{p}_0, \mathbf{A}e^{-\mathbf{A}t}\mathbf{p}_0)$; the result also hold for complex Hermitian $\mathbf{A}$.

As before, $\mathbf{p}$ has the same direction as the solution to the dynamical system $\dot{\mathbf{x}} = -\mathbf{A}\mathbf{x}$ with $\mathbf{x}(0) = \mathbf{p}_0$, but the magnitude is scaled by the nonlinear scalar ω. Provided $\mathbf{p}_0 \ne \mathbf{0}$, the one-sided system (4.5) cannot blow up in finite time, since $(\mathbf{p}, \mathbf{p}) \ne 0$, in stark contrast to the two-sided iteration. This collinearity implies that the $\mathbf{p}$ vectors produced by the one- and two-sided systems provide equally accurate approximations to the desired eigenvector, at least until the latter breaks down.

When $\mathbf{A}$ has a unique simple eigenvalue of smallest real part and the hypotheses of Theorem 4.1 or 4.3 are met, the asymptotic analysis of the associated dynamical system readily follows; cf. [19, section 1.3] for a generic asymptotic linear stability analysis of the one-sided iteration. In fact, one can develop explicit bounds on the sine of the angle between $\mathbf{p}$ and the desired eigenvector $\mathbf{x}_1$, defined as

$$\sin\angle(\mathbf{p},\mathbf{x}_1) := \min_{\alpha\in\mathbb{C}} \frac{\|\alpha\mathbf{p}-\mathbf{x}_1\|}{\|\mathbf{x}_1\|}.$$

Theorem 4.4. *Suppose $\mathbf{A}$ can be diagonalized, $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$, and the eigenvalues of $\mathbf{A}$ can be ordered as*

$$\mathrm{Real}(\lambda_1) < \mathrm{Real}(\lambda_2) \le \cdots \le \mathrm{Real}(\lambda_n).$$

Let $\mathbf{x}_1$ and $\mathbf{y}_1$ denote right and left eigenvectors associated with λ_1, with $\|\mathbf{x}_1\| = 1$ and $\mathbf{y}_1^*\mathbf{x}_1 = 1$. Then the solution $\mathbf{p}(t)$ to both systems (4.1) *and* (4.5) *satisfies*

$$\sin\angle(\mathbf{p}(t),\mathbf{x}_1) \le \|\mathbf{X}\|\,\|\mathbf{X}^{-1}\|\frac{\|\mathbf{p}_0\|}{|\mathbf{y}_1^*\mathbf{p}_0|}e^{\mathrm{Re}(\lambda_1-\lambda_2)t}$$

for all $t \ge 0$ in the case of (4.5), *and for all $t \in [0, t_f)$ in the case of* (4.1).

Proof. Since $\mathbf{x}_1$ is a unit vector, we can write

$$\sin\angle(\mathbf{p}(t),\mathbf{x}_1) = \min_{\alpha\in\mathbb{C}} \|\alpha\mathbf{p}(t)-\mathbf{x}_1\|.$$

In both (4.5) and (4.1), $\mathbf{p}(t)$ is collinear with $e^{-\mathbf{A}t}\mathbf{p}_0$, so we can proceed with

$$\begin{aligned}\sin\angle(\mathbf{p}(t),\mathbf{x}_1) &= \min_{\alpha\in\mathbb{C}} \left\|\alpha\mathbf{X}e^{-\mathbf{\Lambda}t}\mathbf{X}^{-1}\mathbf{p}_0-\mathbf{x}_1\right\| \\ &\le \left\|\frac{e^{\lambda_1 t}}{\mathbf{y}_1^*\mathbf{p}_0}\mathbf{X}e^{-\mathbf{\Lambda}t}\mathbf{X}^{-1}\mathbf{p}_0-\mathbf{x}_1\right\| \;\le\; \|\mathbf{X}\|\;\|\mathbf{X}^{-1}\|\,\frac{\|\mathbf{p}_0\|}{|\mathbf{y}_1^*\mathbf{p}_0|}e^{\mathrm{Re}(\lambda_1-\lambda_2)t}.\end{aligned}$$

The first inequality follows from choosing a (suboptimal) value of α that cancels the terms in the $\mathbf{x}_1$ direction. (For similar analysis of the Arnoldi eigenvalue iteration, see [37, Proposition 2.1].) □

An analogous bound could be developed for the convergence of $\mathbf{q}$ to the left eigenvector $\mathbf{y}_1$. When $\mathbf{A}$ is far from normal, one typically observes a transient stage of convergence that could be better described via analysis that avoids the diagonalization of $\mathbf{A}$; see, e.g., [40, section 28], which includes similar analysis for the power method.

The two-sided iteration converges to left and right eigenvectors of $\mathbf{A}$ associated with the leftmost eigenvalue, *provided the method does not breakdown on the way to this limit.* Several natural questions arise: How common is breakdown? How well do discretizations mimic this dynamical system? Before investigating these issues in section 6, we first address how preconditioning can accelerate—and complicate—the convergence of these continuous-time systems.

5. Preconditioned dynamical systems.

What does it mean to precondition the eigenvalue problem? Several different strategies have been proposed in the literature (see especially the discussion in [21, pp. 109–110]); here we shall investigate analogous approaches for our continuous time dynamical systems, and the implications such modifications have on the convergence behavior described in the last section.

One might first consider applying to the generalized eigenvalue problem

$$\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{p}\lambda,$$

left and right preconditioners $\mathbf{M}$ and $\mathbf{N}$, so as to obtain the equivalent pencil

$$\left(\mathbf{M}^{-1}\mathbf{A}\mathbf{N}\right)\left(\mathbf{N}^{-1}\mathbf{p}\right) = \left(\mathbf{M}^{-1}\mathbf{B}\mathbf{N}\right)\left(\mathbf{N}^{-1}\mathbf{p}\right)\lambda. \tag{5.1}$$

Provided $\mathbf{B}$ is invertible, one could then define

$$\begin{aligned}\widehat{\mathbf{A}} &:= \left(\mathbf{M}^{-1}\mathbf{B}\mathbf{N}\right)^{-1}\left(\mathbf{M}^{-1}\mathbf{A}\mathbf{N}\right) = \mathbf{N}^{-1}\mathbf{B}^{-1}\mathbf{A}\mathbf{N}\\ \widehat{\mathbf{p}} &:= \mathbf{N}^{-1}\mathbf{p},\end{aligned}$$

then apply the concepts from the preceding sections to the standard eigenvalue problem $\widehat{\mathbf{A}}\widehat{\mathbf{p}} = \widehat{\mathbf{p}}\lambda$. For example, we could seek the leftmost eigenpair of $\widehat{\mathbf{A}}$ by evolving the dynamical system

$$\dot{\widehat{\mathbf{p}}} = \widehat{\mathbf{p}}\widehat{\theta} - \widehat{\mathbf{A}}\widehat{\mathbf{p}},$$

with the (preconditioned) Rayleigh quotient

$$\widehat{\theta} = \frac{\left(\widehat{\mathbf{A}}\widehat{\mathbf{p}}, \widehat{\mathbf{p}}\right)}{(\widehat{\mathbf{p}}, \widehat{\mathbf{p}})} = \frac{\left(\mathbf{N}^{-1}\mathbf{B}^{-1}\mathbf{A}\mathbf{p}, \mathbf{N}^{-1}\mathbf{p}\right)}{(\mathbf{N}^{-1}\mathbf{p}, \mathbf{N}^{-1}\mathbf{p})}.$$

Note that $\widehat{\mathbf{A}}$ and $\mathbf{B}^{-1}\mathbf{A}$ share the same spectrum because they are similar, and, hence, the asymptotic rate in Theorem 4.4 is immune to the preconditioner. The application of $\mathbf{N}$ could affect the system's transient behavior, but $\mathbf{M}$ exerts no influence at all.[1]

Several choices for $\mathbf{N}$ are interesting. Taking $\mathbf{N} = \mathbf{A}^{-1}$ gives $\widehat{\mathbf{A}} = \mathbf{A}\mathbf{B}^{-1}$, an alternative to the $\mathbf{B}^{-1}\mathbf{A}$ form suggested by the original problem. Similarity transformations can also be used to *balance* a matrix to improve the conditioning of the eigenvalue problem [31, 33], in which case $\mathbf{N}$ is constructed as a diagonal matrix that reduces the norm of $\widehat{\mathbf{A}}$. Such balancing tends to decrease the departure from normality associated with the largest magnitude eigenvalues. In fact, in the 1960 article that introduced this idea, Osborne refers to this procedure as "pre-conditioning" [31]. A more extreme—if impractical—approach takes $\mathbf{N}$ to be a matrix that diagonalizes $\mathbf{B}^{-1}\mathbf{A}$ (provided such a matrix exists), a choice that minimizes the constant $\|\mathbf{X}\|\|\mathbf{X}^{-1}\|$ that describes the departure from normality in Theorem 4.4.

As useful as such improvements might be, these strategies fail to alter the asymptotic convergence rate described in Theorem 4.4. To potentially improve this rate, one can apply the preconditioner $\mathbf{N}^{-1}$ directly to the residual $\mathbf{p}\theta - \mathbf{A}\mathbf{p}$. Consider the dynamical system

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p}), \tag{5.2}$$

where θ refers to the usual (unpreconditioned) Rayleigh quotient $\theta = (\mathbf{A}\mathbf{p}, \mathbf{p})/(\mathbf{p}, \mathbf{p})$. Discretization of this system results in the familiar preconditioned eigensolver described in (1.1). For this case, a generalization of Theorem 4.3 has proved elusive; we have found no closed form for the exact solution. Indeed, as we shall next see, the choice of preconditioner can even complicate the system's local behavior.

Let $\mathbf{x}_1$ denote a unit eigenvector of $\mathbf{A}$ associated with the eigenvalue λ_1. Note that $\mathbf{x}_1$ is a steady-state of (5.2), linearizing about which gives the Jacobian

$$\mathbf{J} = \mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)(\lambda_1 - \mathbf{A}). \tag{5.3}$$

As $\mathbf{J}\mathbf{x}_1 = \mathbf{0}$, the Jacobian $\mathbf{J}$ always has a zero eigenvalue, adding complexity to conventional linear stability analysis. The challenge can be magnified by a poor

[1] Alternatively, by substituting $(\mathbf{M}^{-1}\mathbf{B}\mathbf{N})^{-1}\widetilde{\mathbf{p}} := \mathbf{N}^{-1}\mathbf{p}$ in (5.1), we obtain a system driven by $\widetilde{\mathbf{A}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{B}^{-1}\mathbf{M}$ that is independent of $\mathbf{N}$.

choice for $\mathbf{N}$. For example, suppose

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{N} = \mathbf{N}^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \lambda_1 = 1,$$

so that

$$\mathbf{J} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix};$$

i.e., the Jacobian is a Jordan block with a double eigenvalue at zero.

To obtain a rough impression of the behavior of the continuous system when θ is in the vicinity of λ_1, consider the constant-coefficient equation $\dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\lambda_1 - \mathbf{A}\mathbf{p})$, whose solution obeys the simple formula

$$\mathbf{p}(t) = e^{\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})t}\mathbf{p}(0).$$

Hence, the asymptotic behavior of $\mathbf{p}$ is controlled by the spectrum of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$. Assuming that $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ has a simple zero eigenvalue, the convergence of this system to the dominant eigenvector depends on the nonzero eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$: if this matrix has any other eigenvalues in the closed right half plane, the system will not generically converge; if all nonzero eigenvalues are in the open left half plane, then the convergence rate will be determined by the rightmost of them.

Specific choices for $\mathbf{N}^{-1}$ will naturally depend significantly on the application problem at hand; in our general setting we seek to characterize basic traits of effective preconditioners. From the perspective of the convergence rate of the continuous dynamical system, we seek a preconditioner $\mathbf{N}^{-1}$ such that the nonzero eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ are as far to the left as possible. While the leftmost eigenvalues of $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ do not much affect the behavior of the continuous system, they can have a significant effect on the stability of the discretized difference equation, i.e., the related eigensolvers. For example, if $\mathbf{N}^{-1}(\lambda_1 - \mathbf{A})$ moves all nonzero eigenvalues into the left half plane, then replacing $\mathbf{N}$ by $\frac{1}{2}\mathbf{N}$ doubles the convergence rate of the continuous system. (We shall see on page 1461 that there is "no free lunch" for practical computations: the improved convergence rate of the continuous system is counter-balanced by the need to use a smaller step size in the discretized system.)

To rigorously analyze the local behavior of the fully nonlinear system when $\mathbf{p}$ approximates the eigenvector $\mathbf{x}_1$, we shall apply the center manifold theorem [9, 17], a tool for studying a dynamical system whose Jacobian has an eigenvalue on the imaginary axis. (Alternatively, we could restrict the system to the unit sphere in $\mathbb{R}^n$.) We assume that $\mathbf{A} \in \mathbb{R}^{n\times n}$. Without loss of generality, assume that $\lambda_1 = 0$, so that the Jacobian at $\mathbf{x}_1$ (5.3) takes the form $\mathbf{J} = -\mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)\mathbf{A}$. Thus, for $\mathbf{p}$ near $\mathbf{x}_1$ we have

$$\dot{\mathbf{p}} = \mathbf{J}\mathbf{p} + \mathbf{F}(\mathbf{p})$$

for the nonlinear function $\mathbf{F}(\mathbf{p}) = \mathbf{N}^{-1}(\theta(\mathbf{p})\mathbf{p} - (\mathbf{A}\mathbf{p}, \mathbf{x}_1)\mathbf{x}_1)$ that, by definition of the Jacobian, satisfies $\|\mathbf{F}(\mathbf{p})\| = o(\|\mathbf{p} - \mathbf{x}_1\|)$.

Suppose that $\mathbf{J}$ has a simple zero eigenvalue, and the rest of its spectrum is in the open left half plane. There exists some invertible (real, if $\mathbf{J}$ is real) matrix $\mathbf{S}$ with first column $\mathbf{x}_1$ and

$$\mathbf{S}^{-1}\mathbf{J}\mathbf{S} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}$$

for some $\mathbf{C} \in \mathbb{R}^{(n-1)\times(n-1)}$ whose spectrum is in the open left half plane.

We now transform coordinates into a form in which the center manifold theorem can most readily be applied. Define

$$\mathbf{r}(t) = \mathbf{S}^{-1}(\mathbf{p}(t) - \mathbf{x}_1),$$

so that

$$\dot{\mathbf{r}} \;=\; \left(\mathbf{S}^{-1}\mathbf{J}\mathbf{S}\right)\mathbf{S}^{-1}(\mathbf{p} - \mathbf{x}_1) + \mathbf{S}^{-1}\mathbf{F}(\mathbf{p}) \;=\; \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}\mathbf{r} + \mathbf{G}(\mathbf{r}),$$

where $\mathbf{G}(\mathbf{r}) := \mathbf{S}^{-1}\mathbf{F}(\mathbf{S}\mathbf{r} + \mathbf{x}_1) = \mathbf{S}^{-1}\mathbf{F}(\mathbf{p})$. By design, $\mathbf{S}^{-1}\mathbf{x}_1 = \mathbf{e}_1$; hence, $\mathbf{G}(\mathbf{r})$ satisfies

$$\mathbf{G}(\mathbf{r}) = \mathbf{S}^{-1}\mathbf{N}^{-1}\mathbf{S}\left(\left(\frac{(\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{S}\mathbf{r}) + (\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{x}_1)}{(\mathbf{S}\mathbf{r}, \mathbf{S}\mathbf{r}) + 2(\mathbf{x}_1, \mathbf{S}\mathbf{r}) + 1}\right)(\mathbf{r} + \mathbf{e}_1) - (\mathbf{A}\mathbf{S}\mathbf{r}, \mathbf{x}_1)\mathbf{e}_1\right). \tag{5.4}$$

Now we are prepared to cast this diagonalized problem into the conventional setting for center manifold theory. We write

$$\mathbf{r} = \begin{bmatrix} \alpha \\ \mathbf{b} \end{bmatrix}$$

for $\alpha \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^{n-1}$. Using MATLAB index notation for convenience, the $\mathbf{r}$ system is simply

$$\begin{bmatrix} \dot{\alpha} \\ \dot{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}\begin{bmatrix} \alpha \\ \mathbf{b} \end{bmatrix} + \begin{bmatrix} \mathbf{G}([\alpha; \mathbf{b}])_1 \\ \mathbf{G}([\alpha; \mathbf{b}])_{2:n} \end{bmatrix},$$

that is,

$$\dot{\alpha} = \mathbf{G}([\alpha; \mathbf{b}])_1, \quad \dot{\mathbf{b}} = \mathbf{C}\mathbf{b} + \mathbf{G}([\alpha; \mathbf{b}])_{2:n}.$$

Notice that the component α only figures in the nonlinear terms; we wish to determine how that contribution affects the magnitude of the $\mathbf{b}$ component—that is, the portion of the solution that we hope decays as $t \to \infty$. Notice that $\mathbf{b} = \mathbf{0}$ corresponds to the case when $\mathbf{p}$ is collinear with $\mathbf{x}_1$. In this case $\mathbf{p}$ may differ from the unit eigenvector $\mathbf{x}_1$, but regardless it is a fixed point of the dynamical system, and provided $\mathbf{p} \neq \mathbf{0}$ we are content. In particular, if $\mathbf{b} = \mathbf{0}$, then $\mathbf{A}\mathbf{S}\mathbf{r} = \mathbf{0}$ too (recall that $\lambda = 0$), and we can see from (5.4) that $\mathbf{G}(\mathbf{r}) = \mathbf{0}$. In this case

$$\dot{\alpha} = \mathbf{G}([\alpha; \mathbf{0}])_1 = 0, \quad \dot{\mathbf{b}} = \mathbf{C}\mathbf{0} + \mathbf{G}([\alpha; \mathbf{0}])_{2:n} = \mathbf{0},$$

so any such $\mathbf{r}$ is a fixed point of the dynamical system. We can put this in grander language: there exists some $\delta > 0$ such that if

$$\mathbf{r}_0 \in \left\{ \begin{bmatrix} \alpha \\ \mathbf{0} \end{bmatrix} : |x| < \delta \right\} =: \mathcal{M},$$

then the dynamical system with $\mathbf{r}(0) = \mathbf{r}_0$ satisfies $\mathbf{r}(t) \in \mathcal{M}$ for all $t > 0$. (In particular, $\mathbf{r}(t) = \mathbf{r}(0) \in \mathcal{M}$.) The set $\mathcal{M}$ is called a *local invariant manifold.* We can define this manifold (locally) by the requirement that

$$\mathbf{b} = \mathbf{g}(\alpha) := \mathbf{0},$$

which trivially satisfies $\mathbf{g}(0) = \mathbf{0}$ and the Jacobian of $\mathbf{g}$ at $\alpha = 0$ is $D\mathbf{g}(0) = \mathbf{0}$; furthermore, $\mathbf{g}$ is arbitrarily smooth near $\alpha = 0$. Together, these properties ensure that $\mathcal{M}$ is a *center manifold* of the dynamical system. (We are fortunate in this case to have an explicit, trivial expression for this manifold.)

All that remains is to apply Theorem 2 from Carr [9, p. 4]. Consider the equation

$$\dot{u} = \mathbf{G}([u; \mathbf{g}(u)])_1 = \mathbf{G}([u; \mathbf{0}])_1 = 0.$$

The solution $u(t) = 0$ is clearly stable—if $u(t) = \varepsilon$, then $|u(t) - 0| = |\varepsilon|$ is bounded for all $t > 0$—and, thus, Theorem 2(a) from [9] implies that the solution $\mathbf{r}(t) = \mathbf{0}$ is a stable solution of the system

$$\dot{\mathbf{r}} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \mathbf{r} + \mathbf{G}(\mathbf{r}).$$

Note that the solution $u(t) = 0$ is not *asymptotically stable*, that is, we do not have $u(t) \to 0$ if $u(0) = \varepsilon$ for small, nonzero ε. Were this the case, then we would be able to conclude that the $\mathbf{r}$ system was asymptotically stable. This would contradict our expectation that the original dynamical system will converge to something in $\operatorname{span}\{\mathbf{x}_1\}$, not necessarily to $\mathbf{x}_1$ itself. In particular, if $\mathbf{N}$ is self-adjoint, then $(\mathbf{Np}, \mathbf{p})$ is an invariant of the system, and so we expect that $\mathbf{p}(t) \to \xi\mathbf{x}_1$ for ξ determined by

$$|\xi|^2 = \frac{(\mathbf{Np}, \mathbf{p})}{(\mathbf{Nx}_1, \mathbf{x}_1)}.$$

We now have stability of the zero state of the $\mathbf{r}$ system, but that only means that solutions sufficiently close to $\mathbf{r} = \mathbf{0}$ do not diverge. To say more—to say that the solutions actually converge to the center manifold—we can apply Theorem 2(b) of [9], which we slightly paraphrase here. Since the zero solution of the $\mathbf{r}$ equation is stable, for $\|[\alpha(0); \mathbf{b}(0)]\|$ sufficiently small, there exists some solution $u(t)$ of the equation $\dot{u}(t) = \mathbf{G}([u; \mathbf{g}(u)])_1 = 0$ and positive constant γ such that

$$\alpha(t) = u(t) + O\left(e^{-\gamma t}\right), \quad \mathbf{b}(t) = \mathbf{g}(u(t)) + O\left(e^{-\gamma t}\right).$$

In particular, in our setting such solutions $u(t)$ will be constant: $u(t) = c$, and so there exist

$$\alpha(t) = c + O\left(e^{-\gamma t}\right), \quad \mathbf{b}(t) = O\left(e^{-\gamma t}\right),$$

and, in particular, $\|\mathbf{b}(t)\| \to 0$ as $t \to \infty$. Thus, for $\|\mathbf{r}_0\|$ sufficiently small,

$$\mathbf{r}(t) = \begin{bmatrix} c \\ \mathbf{0} \end{bmatrix} + O\left(e^{-\gamma t}\right),$$

so that $\mathbf{p}(t) = \mathbf{Sr}(t) + \mathbf{x}_1 = (1 + c)\mathbf{x}_1 + O(e^{-\gamma t})$. The preceding discussion is summarized in the following result.

THEOREM 5.1. *If $\|\mathbf{p}(0) - \mathbf{x}_1\|$ is sufficiently small and $\mathbf{N}^{-1}(\mathbf{I} - \mathbf{x}_1\mathbf{x}_1^*)(\lambda - \mathbf{A})$ has a simple zero eigenvalue with all other eigenvalues in the open left half plane, then there exists $\gamma > 0$ and $\xi \in \mathbb{R}$ such that, as $t \to \infty$,*

$$\|\mathbf{p}(t) - \xi\mathbf{x}_1\| = O\left(e^{-\gamma t}\right).$$

In the case of self-adjoint, invertible $\mathbf{N}$, $|\xi| = |(\mathbf{p}_0, \mathbf{Np}_0)|$.

Note that if $\mathbf{N}$ is Hermitian and invertible but indefinite, then there always exists some unit vector $\mathbf{p}_0$ such that $(\mathbf{p}_0, \mathbf{Np}_0) = 0$. If this starting vector is sufficiently close to the unit eigenvector $\mathbf{x}_1$ of $\mathbf{A}$, then we have not ruled out the possibility that the system converges to the zero vector, rather than a desired eigenvector.

6. Discrete dynamical systems. The previous sections have addressed the quadratic invariant and convergence behavior of the continuous-time, one- and two-sided dynamical systems. For purposes of computation, one naturally wonders how closely such properties are mimicked by the solutions to discretizations of these systems. The present section considers the convergence and preservation of the quadratic invariant by the discrete flow under a forward Euler time integration. We focus on this canonical integrator for three reasons: (1) this discretization leads to the algorithm (1.1) proposed in the literature; (2) analysis for forward Euler serves as a first step toward understanding more sophisticated algorithms; (3) more elaborate methods are not always practical. For example, the implicit midpoint rule will preserve the quadratic invariant $(\mathbf{p}, \mathbf{N}\mathbf{p})$ [18, IV.2.1] of the one-sided system (1.2), but since this method takes the form

$$\begin{aligned}\mathbf{p}_{j+1} &= \mathbf{p}_j + h\mathbf{N}^{-1}\left(\theta_{j+1}\left(\frac{\mathbf{p}_j+\mathbf{p}_{j+1}}{2}\right) - \mathbf{A}\left(\frac{\mathbf{p}_j+\mathbf{p}_{j+1}}{2}\right)\right)\\ \theta_{j+1} &= \frac{(\mathbf{p}_j+\mathbf{p}_{j+1})^T\mathbf{A}(\mathbf{p}_j+\mathbf{p}_{j+1})}{(\mathbf{p}_j+\mathbf{p}_{j+1})^T(\mathbf{p}_j+\mathbf{p}_{j+1})},\end{aligned}$$

its implementation requires the solution of a (nonlinear) system of equations at each step: a far more expensive proposition (per step) than the humble forward Euler method. (For a more sophisticated discretization in the unpreconditioned Hermitian case, along with a cautionary note about use of large step-size in the forward Euler method, see [28].)

6.1. Departure from the manifold. Given $\mathbf{A} \in \mathbb{R}^{n\times n}$, for notational convenience we rewrite the two-sided system in the form

$$\begin{aligned}\dot{\mathbf{p}} &= \mathbf{p}\theta - \mathbf{A}\mathbf{p} =: \mathbf{f}(\mathbf{p},\mathbf{q})\\ \dot{\mathbf{q}} &= \mathbf{q}\theta - \mathbf{A}^T\mathbf{q} =: \mathbf{g}(\mathbf{p},\mathbf{q}),\end{aligned} \tag{6.1}$$

with $\theta = (\mathbf{q}^T\mathbf{p})^{-1}\mathbf{q}^T\mathbf{A}\mathbf{p} = \theta^T$ and initial conditions $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$ and $\mathbf{q}(0) = \mathbf{q}_0 \in \mathbb{R}^n$. Similarly, the one-sided system (now including preconditioning) is

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}(\mathbf{p}\theta - \mathbf{A}\mathbf{p}) =: \mathbf{N}^{-1}\mathbf{f}(\mathbf{p},\mathbf{p}), \tag{6.2}$$

with $\theta = (\mathbf{p}^T\mathbf{p})^{-1}\mathbf{p}^T\mathbf{A}\mathbf{p} = \theta^T$ and $\mathbf{p}(0) = \mathbf{p}_0 \in \mathbb{R}^n$.

In section 2 we showed that this system preserves the quadratic invariant $\mathbf{q}^T\mathbf{p}$. To what extent do discretizations respect such conservation, and what are the implications of any drift from this manifold? To understand the role of discrete quadratic invariants, we consider the error when using a forward Euler time integrator.

We begin with the two-sided iteration. The finite-time blow-up established in Theorem 4.1 is a strike against this method. Before abandoning it altogether, we wish to investigate the consequences of the blow-up on the discrete two-sided eigensolver. The forward Euler applied to (6.1) leads to the iteration

$$\mathbf{p}_{j+1} = \mathbf{p}_j + h\mathbf{f}_j \tag{6.3}$$

$$\mathbf{q}_{j+1} = \mathbf{q}_j + h\mathbf{g}_j, \tag{6.4}$$

where $\mathbf{f}_j := \mathbf{f}(\mathbf{p}_j, \mathbf{q}_j)$ and $\mathbf{g}_j := \mathbf{g}(\mathbf{p}_j, \mathbf{q}_j)$. With the mild caveat that $\mathbf{q}_j^T\mathbf{p}_j \neq 0$, the form of the Rayleigh quotient gives

$$\mathbf{q}_j^T\mathbf{f}_j = 0 = \mathbf{p}_j^T\mathbf{g}_j.$$

This simple observation is critical to understanding the drift of the forward Euler iterates from the invariant manifold. It implies, for example, that the first iteration of (6.3)–(6.4) produces a iterate that is quadratically close to the manifold:

$$\mathbf{q}_1^T\mathbf{p}_1 = \mathbf{q}_0^T\mathbf{p}_0 + h^2\left(\mathbf{g}_0^T\mathbf{f}_0\right),$$

which is perhaps surprising given the forward Euler method's $O(h)$ accuracy. Writing the departure from the manifold as

$$d_j = \mathbf{q}_j^T\mathbf{p}_j - \mathbf{q}_0^T\mathbf{p}_0,$$

we, thus, have $d_1 = h^2(\mathbf{g}_0^T\mathbf{f}_0)$. From this we can compute

$$d_2 = \left(\mathbf{q}_2^T\mathbf{p}_2 - \mathbf{q}_1^T\mathbf{p}_1\right) + d_1 = h^2\left(\mathbf{g}_1^T\mathbf{f}_1 + \mathbf{g}_0^T\mathbf{f}_0\right)$$

and, in general, $d_{j+1} = h^2\sum_{k=0}^{j}\mathbf{g}_k^T\mathbf{f}_k$. (This result is a special case of one derived in [18] for partitioned Runge–Kutta systems.) Thus, we can bound the relative drift from the manifold as

$$\frac{\left|\mathbf{q}_{j+1}^T\mathbf{p}_{j+1} - \mathbf{q}_0^T\mathbf{p}_0\right|}{\left|\mathbf{q}_0^T\mathbf{p}_0\right|} \le h^2\sum_{k=0}^{j}\frac{\|\mathbf{f}_k\|\,\|\mathbf{g}_k\|}{\left|\mathbf{q}_0^T\mathbf{p}_0\right|}. \tag{6.5}$$

The definitions of $\mathbf{f}(\mathbf{p},\mathbf{q})$ and $\mathbf{g}(\mathbf{p},\mathbf{q})$ imply

$$\|\mathbf{f}_k\| \le (|\theta_k| + \|\mathbf{A}\|)\,\|\mathbf{p}_k\| \le \left(1 + \frac{\|\mathbf{q}_k\|\,\|\mathbf{p}_k\|}{|\mathbf{q}_k^T\mathbf{p}_k|}\right)\|\mathbf{A}\|\,\|\mathbf{p}_k\|$$

$$\|\mathbf{g}_k\| \le (|\theta_k| + \|\mathbf{A}\|)\,\|\mathbf{q}_k\| \le \left(1 + \frac{\|\mathbf{p}_k\|\,\|\mathbf{q}_k\|}{|\mathbf{p}_k^T\mathbf{q}_k|}\right)\|\mathbf{A}\|\,\|\mathbf{q}_k\|.$$

Substituting these formulas into (6.5), we arrive at the following result.

THEOREM 6.1. *The forward Euler iterates* (6.3)–(6.4) *for the two-sided dynamical system* (6.1) *satisfy*

$$\frac{\left|\mathbf{q}_{j+1}^T\mathbf{p}_{j+1} - \mathbf{q}_0^T\mathbf{p}_0\right|}{\left|\mathbf{q}_0^T\mathbf{p}_0\right|} \le h^2\frac{\|\mathbf{A}\|^2}{\left|\mathbf{q}_0^T\mathbf{p}_0\right|}\sum_{k=0}^{j}\left(1 + \frac{\|\mathbf{q}_k\|\,\|\mathbf{p}_k\|}{|\mathbf{q}_k^T\mathbf{p}_k|}\right)^2\|\mathbf{q}_k\|\,\|\mathbf{p}_k\|. \tag{6.6}$$

This bound implies that the departure from the manifold is proportional to the square of the step size, and involves the secants of the angles formed by $\mathbf{q}_k$ and $\mathbf{p}_k$, $k = 0,\ldots,j$, as well as the norms of $\mathbf{q}_k$ and $\mathbf{p}_k$. Moreover, unless the cosines of the angles between $\mathbf{q}_k$ and $\mathbf{p}_k$ are bounded away from zero, there does not exist a step size h such that all iterates remain near the quadratic manifold. The proof of the theorem demonstrates that the secant of the angle is at least as large as the normalized residuals. Numerical experiments indicate that these bounds are descriptive; see the first example in section 6.3. A conclusion is that serious breakdown (as discussed after Theorem 4.1) leads to *incurable breakdown* of the two-sided iteration because forward Euler mimics the continuous solution and cannot "step-over" the point of blow-up.

Given the shortcomings of the two-sided iteration, we shall, henceforth, focus on the one-sided dynamical system, and also include preconditioning (6.2). The associated forward Euler discretization takes the form

$$\mathbf{p}_{j+1} = \mathbf{p}_j + h\mathbf{N}^{-1}\mathbf{f}_j, \tag{6.7}$$

where now $\mathbf{f}_j = \mathbf{f}(\mathbf{p}_j, \mathbf{p}_j)$. (Here we see that the time-step h directly multiplies the preconditioner $\mathbf{N}$, so that the effect of scaling $\mathbf{N}$ to improve the convergence rate of the continuous-time system, as discussed on page 1456, is equivalent to choosing a smaller time-step in the discrete setting.)

The following analysis will play a useful role in our main convergence result, Theorem 6.3. For the rest of the paper we assume that $\mathbf{N}$ is symmetric and invertible, which, as seen in the Introduction, ensures that solutions of the continuous system reside on an invariant manifold $\mathbf{p}^T\mathbf{N}\mathbf{p} =$ constant. At each time step, the discrete iteration incurs a local departure from that manifold of

$$e_{j+1} := \mathbf{p}_{j+1}^T\mathbf{N}\mathbf{p}_{j+1} - \mathbf{p}_j^T\mathbf{N}\mathbf{p}_j = h^2\mathbf{f}_j^T\mathbf{N}^{-1}\mathbf{f}_j.$$

Hence, if $\mathbf{N}^{-1}$ is additionally positive definite (e.g., $\mathbf{N}^{-1} = \mathbf{I}$), the drift is monotone increasing—an important property for the forthcoming convergence theory.

When $\mathbf{N}$ is positive definite, we can define vector norms

$$\|\mathbf{z}\|_{\mathbf{N}^{-1}}^2 := \mathbf{z}^T\mathbf{N}^{-1}\mathbf{z}, \qquad \|\mathbf{z}\|_{\mathbf{N}}^2 := \mathbf{z}^T\mathbf{N}\mathbf{z}$$

(which in turn induce matrix norms), with $\|\mathbf{z}\|_{\mathbf{N}^{-1}} \le \|\mathbf{N}^{-1}\|\|\mathbf{z}\|_{\mathbf{N}}$. Thus, we write

$$e_{j+1} = h^2\|\mathbf{f}_j\|_{\mathbf{N}^{-1}}^2 \le h^2\left\|\mathbf{N}^{-1}\right\|^2\|\mathbf{f}_j\|_{\mathbf{N}}^2 = h^2\left\|\mathbf{N}^{-1}\right\|^2\|\mathbf{r}_j\|_{\mathbf{N}}^2\|\mathbf{p}_j\|_{\mathbf{N}}^2,$$

where we use the normalized residual $\mathbf{r}_j := \mathbf{f}_j/\|\mathbf{p}_j\|_{\mathbf{N}} = (\theta_j - \mathbf{A})\mathbf{p}_j/\|\mathbf{p}_j\|_{\mathbf{N}}$. Now consider the aggregate, global drift from the manifold:

$$\begin{aligned} d_{j+1} &:= \mathbf{p}_{j+1}^T\mathbf{N}\mathbf{p}_{j+1} - \mathbf{p}_0^T\mathbf{N}\mathbf{p}_0 \\ &= \sum_{k=1}^{j+1} e_k \le h^2\left\|\mathbf{N}^{-1}\right\|^2\sum_{k=0}^{j}\|\mathbf{r}_k\|_{\mathbf{N}}^2\left(d_k + \|\mathbf{p}_0\|_{\mathbf{N}}^2\right). \end{aligned}$$

In particular, d_{j+1} is determined by the step size, the residual norms, and the growth in the norm of the iterates. For further simplification, choose some $M > 0$ such that $\|\mathbf{r}_k\|_{\mathbf{N}}^2 \le M$ for all $k = 0, \ldots, j$. One coarse (but j-independent) possibility is

$$M := \inf_{s\in\mathbb{R}} 4\|\mathbf{A} - s\|_{\mathbf{N}}^2 \ge \inf_{s\in\mathbb{R}}\|(\mathbf{A} - s) - (\theta_k - s)\|_{\mathbf{N}}^2 \ge \|\mathbf{r}_k\|_{\mathbf{N}}^2, \tag{6.8}$$

which is invariant to shifts in $\mathbf{A}$. (In terms of the Euclidean norm, we, thus, have $M \le 4\kappa(\mathbf{N})\inf_{s\in\mathbb{R}}\|\mathbf{A} - s\|^2$, where $\kappa(\mathbf{N}) = \|\mathbf{N}\|\|\mathbf{N}^{-1}\|$.) Hence,

$$d_{j+1} \le h^2M\left\|\mathbf{N}^{-1}\right\|^2\sum_{k=0}^{j}(d_k + \|\mathbf{p}_0\|_{\mathbf{N}})^2 = h^2M\left\|\mathbf{N}^{-1}\right\|^2\left((j+1)\|\mathbf{p}_0\|_{\mathbf{N}}^2 + \sum_{k=1}^{j} d_k\right)$$

(since $d_0 = 0$). Thus, if we define the sequence $\{\widehat{d}_k\}$ by

$$\widehat{d}_{j+1} = h^2M\left\|\mathbf{N}^{-1}\right\|^2\left((j+1) + \sum_{k=1}^{j}\widehat{d}_k\right), \tag{6.9}$$

then the departure from the manifold obeys $d_{j+1} \le \widehat{d}_{j+1}\|\mathbf{p}_0\|_{\mathbf{N}}^2$. Equation (6.9) is a binomial recurrence whose solution can be written explicitly:

$$\widehat{d}_{j+1} = \sum_{k=1}^{j+1}\binom{j+1}{k}\left(h^2M\|\mathbf{N}^{-1}\|^2\right)^k = \left(1 + h^2M\|\mathbf{N}^{-1}\|^2\right)^{j+1} - 1.$$

Theorem 6.2. *Let* $\mathbf{N} \in \mathbb{R}^{n\times n}$ *be symmetric and positive definite, and define* M *by* (6.8). *Then the forward Euler iterates* (6.7) *for the preconditioned one-sided dynamical system* (6.2) *satisfy*

$$0 \le \frac{\mathbf{p}_{j+1}^T\mathbf{N}\mathbf{p}_{j+1} - \mathbf{p}_0^T\mathbf{N}\mathbf{p}_0}{\mathbf{p}_0^T\mathbf{N}\mathbf{p}_0} \le \left(1 + h^2M\|\mathbf{N}^{-1}\|^2\right)^{j+1} - 1, \tag{6.10}$$

the upper bound being asymptotic to $(j+1)h^2\|\mathbf{N}^{-1}\|^2M$ *as* $h \to 0$.

Note that a small eigenvalue of $\mathbf{N}$ results in a small time-step h. The bound also provides an estimate of a critical time-step

$$h\sqrt{j+1} \lessapprox \frac{1}{\|\mathbf{N}^{-1}\|\sqrt{M}}$$

for forward Euler, limiting the departure from the quadratic manifold. Highly nonnormal problems for which $\|\mathbf{A} - s\| \gg \max_k |\lambda_k - s|$ also result in tiny time-steps.

Theorem 6.2 leads to an interesting observation—despite the fact that the forward Euler method generally incurs an $O(h)$ truncation error and the global error grows exponentially in j for fixed h (see (6.12) and, e.g., [14, section 1.3]), for a one-sided iteration the drift from the quadratic manifold is $O(h^2)$ and both linear and nondecreasing in j for all starting vectors, under mild restrictions. This monotone departure from the manifold is exploited in the discrete convergence analysis to follow. So, although explicit Runge–Kutta methods (such as forward Euler) do not preserve quadratic invariants (see [18, Chapter IV]), the forward Euler iterates for the one-sided systems remain nearby. The reader is referred to [18, Chapter IV] for further information and references, including the use of projection to remain on the quadratic manifold.

6.2. Discrete convergence theory. Just as the local drift from the manifold at each iteration contributes to the global drift, so local truncation errors committed by each step of an ODE solver aggregate into a global error. How does this accumulated error affect convergence of the discrete method as we compute $\mathbf{p}_j$ with $j \to \infty$?

In this section, we seek conditions that will ensure that the *discrete preconditioned one-sided iteration* (6.7) converges to the same eigenvector as the continuous system.

First, we establish the setting that will be used through this rest of this section. Suppose $\mathbf{A} \in \mathbb{R}^{n\times n}$ has a simple eigenvalue λ_1 strictly to the left of all other eigenvalues (and, hence, real). Without loss of generality (via a unitary similarity transformation) we can assume that $\mathbf{A}$ takes the form

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & \mathbf{d}^T \\ \mathbf{0} & \mathbf{C} \end{bmatrix}. \tag{6.11}$$

Let $\mathbf{x}_1$ and $\mathbf{y}_1$ denote unit-length right and left eigenvectors associated with λ_1; in these coordinates we can take $\mathbf{x}_1 = [1, 0, \ldots, 0]^T$. Theorems 4.3, 4.4, and 5.1 provide conditions under which the solution $\mathbf{p}(t)$ of the continuous system converges in angle to the eigenvector $\mathbf{x}_1$ (e.g., if $\mathbf{N} = \mathbf{I}$ and $\mathbf{y}_1^T\mathbf{p}_0 \neq 0$).

Before beginning the convergence analysis, one should appreciate that the conditions established in the last paragraph are not sufficient to guarantee convergence of the discrete iteration. Consider the following example. When $\mathbf{N} = \mathbf{I}$, the forward Euler iterate of the one-sided system at step k can be written as

$$\mathbf{p}_k = \prod_{j=0}^{k-1} \varphi_j(\mathbf{A})\mathbf{p}_0$$

for linear factors $\varphi_j(z) = 1+h(\theta_j - z)$. If any of these factors has λ_1 as a root, then $\mathbf{p}_k$ will have no component in the direction of the eigenvector $\mathbf{x}_1$, and so λ_1 and $\mathbf{x}_1$ will not influence the iteration: convergence of $\mathbf{p}_k$ to $\mathbf{x}_1$ is impossible. Concrete matrices that exhibit such behavior are simple to construct. For any *fixed* $h > 0$, set

$$\mathbf{A} = \begin{bmatrix} 0 & -1-2/h \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{p}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Theorem 4.3 guarantees that the continuous one-sided system will converge for this $\mathbf{A}$ and $\mathbf{p}_0$. At the first step of the forward Euler method $\theta_0 = -1/h$, so that $\varphi_0(0) = 0$ and $\mathbf{p}_1 = [h+2, -h]^T$ is an eigenvector for $\lambda_2 = 1$, and $\mathbf{p}_k$ will never have a component in the $\mathbf{x}_1$ direction for any $k \geq 1$. (Note that $\varphi_j(\lambda_1) = 1 + h(\theta_j - \lambda_1) = 0$ implies that $\theta_j - \lambda_1 = -1/h < 0$, and this is impossible if $\mathbf{A}$ is normal. As h is reduced, complete deflation requires an increasing departure from normality.) The more sophisticated restarted Arnoldi algorithm exhibits a similar phenomenon; see [12].

Under what circumstances can we guarantee convergence? To answer this question, we first review the conventional global error analysis for the forward Euler method; for details, see, e.g., [14, section 1.3]. The first step begins with the exact solution at time $t = 0$: $\mathbf{p}_0 = \mathbf{p}(0)$. Each subsequent step introduces a local truncation error, while also magnifying the global error aggregated at previous steps. Suppose we wish to integrate for $t \in [0, \tau]$ with $\tau = kh$ for some integer k. With the local truncation error at each step bounded by

$$T_h := \max_{0 \leq t \leq \tau} \tfrac{1}{2} h \|\ddot{\mathbf{p}}(t)\|,$$

one can show that

$$\|\mathbf{p}_k - \mathbf{p}(\tau)\| \leq \frac{T_h}{L}\left(e^{\tau L} - 1\right), \tag{6.12}$$

where L is a Lipschitz constant for our differential equation; in Appendix A we show that $L = 10\|\mathbf{N}^{-1}\|\|\mathbf{A}\|$ will suffice. This expression for the global error captures an essential feature: for fixed τ, the fact that $T_h = O(h)$ implies that we can always select $h > 0$ sufficiently small as to make the difference between the forward Euler iterate $\mathbf{p}_{\tau/h}$ and the exact solution $\mathbf{p}(\tau)$ arbitrarily small. However, if we increase k with $h > 0$ *fixed*, the bound indicates an *exponential* growth in the error. To show that $\mathbf{p}_k$ converges (in angle) to an eigenvector as $k \to \infty$, further work is required. In this effort, the preservation of the quadratic invariant characterized in Theorem 6.2 plays an essential role.

Preconditioning significantly complicates the convergence theory. For simplicity, our analysis imposes the stringent requirement that, in the coordinates in which $\mathbf{A}$ takes the form (6.11), we have

$$\mathbf{N}^{-1} = \begin{bmatrix} \eta & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix} \tag{6.13}$$

in addition to the requirement that $\mathbf{N}^{-1}$ be symmetric and positive definite. The trivial off-diagonal blocks prevent the preconditioner from using the growing component of $\mathbf{p}_k$ in $\mathbf{x}_1$ to enlarge the component in the unwanted eigenspace.

A crucial ingredient in our convergence analysis is the constant

$$\gamma := \|\mathbf{\Pi}_1(\mathbf{I} + h\mathbf{N}(\lambda_1 - \mathbf{A}))\| = \|\mathbf{I} + h\mathbf{M}(\lambda_1 - \mathbf{C})\|,$$

where $\mathbf{\Pi}_1 := \mathbf{I} - \mathbf{x}_1\mathbf{x}_1^T$ is a projector onto the complement of the desired invariant subspace. This constant γ, a function of h, measures the potency of the preconditioner: the smaller, the better. For example, in the ideal case that $\mathbf{M} = (\mathbf{C} - \lambda_1)^{-1}$, we have $\gamma = |1 - h|$, giving $\gamma = 0$ for the large step size $h = 1$, and that $\gamma \to 1$ as $h \to 0$.

With γ in hand, we are prepared to state our convergence result. Here, $\kappa(\mathbf{N}) = \|\mathbf{N}\|\|\mathbf{N}^{-1}\|$ denotes the condition number of the preconditioner.

THEOREM 6.3. *Given* (6.11), (6.13), *and assumptions on* λ_1, $\mathbf{x}_1$, *and* $\mathbf{N}$ *established in the previous paragraphs, suppose that* $\mathbf{p}_0$ *is chosen so that the continuous dynamical system converges in angle to an eigenvector associated with the distinct, simple leftmost eigenvalue* λ_1 *(e.g.,* $\mathbf{y}_1^T\mathbf{p}_0 \neq 0$ *suffices if* $\mathbf{N} = \mathbf{I}$*). Furthermore, suppose there exists* $h > 0$ *for which*

$$\gamma \in [0, 1/\sqrt{\kappa(\mathbf{N})}). \tag{6.14}$$

Then after preliminary iteration with a sufficiently small time-step h_0*, the forward Euler method with time-step* h *will converge (in angle) to the desired eigenvector:*

$$\sin(\angle(\mathbf{p}_k, \mathbf{x}_1)) = O\left(\gamma^k\right). \tag{6.15}$$

Asymptotically, the Rayleigh quotient converges to λ *at the same rate:*

$$|\theta_k - \lambda| = O\left(\gamma^k\right), \tag{6.16}$$

which in the case $\mathbf{d} = \mathbf{0}$ *improves to* $|\theta_k - \lambda| = O(\gamma^{2k})$.

Proof. Denote the kth iterate by

$$\mathbf{p}_k = \begin{bmatrix} \alpha_k \\ \mathbf{b}_k \end{bmatrix}.$$

• *Convergence of the forward Euler method to the continuous solution, and convergence of the continuous solution to the eigenvector, together ensure that preliminary forward Euler steps will get close to the eigenvector.* To show that $\sin(\angle(\mathbf{p}_k, \mathbf{x}_1)) \to 0$ as $k \to \infty$, we will show that $\|\mathbf{b}_k\| \to 0$ while $|\alpha_k|$ is bounded away from zero. The convergence of the forward Euler method at a fixed time $\tau \geq 0$ (see (6.12)), with the assumption that the continuous system converges for the given $\mathbf{p}_0$ (as described in sections 4–5), ensures that we can run the forward Euler iteration with a sufficiently small time-step that, after $k \geq 0$ iterations, $\|\mathbf{b}_k\|$ is sufficiently small that

$$\frac{\|\mathbf{b}_k\|^2\|\lambda_1 - \mathbf{C}\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{b}_k\|\|\mathbf{d}\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \leq \frac{\varepsilon}{h\|\mathbf{M}\|} \tag{6.17}$$

for some $\varepsilon \in [0, 1/\sqrt{\kappa(\mathbf{N})} - \gamma)$; here $\gamma \in [0, 1/\sqrt{\kappa(\mathbf{N})})$ and $h > 0$ are as in the statement of the theorem. Note that the left-hand side of (6.17) will get small when $\|\mathbf{b}_k\|$ is small, since $|\alpha_k|$ is bounded away from zero. This follows from Theorem 6.2 (monotonic drift of the invariant) and the fact that $\mathbf{N}$ is symmetric positive definite, which imply that for any j,

$$\|\mathbf{p}_j\|^2 \geq \frac{1}{\|\mathbf{N}\|}\mathbf{p}_j^T\mathbf{N}\mathbf{p}_j \geq \frac{1}{\|\mathbf{N}\|}\mathbf{p}_{j-1}^T\mathbf{N}\mathbf{p}_{j-1} \geq \frac{1}{\kappa(\mathbf{N})}\|\mathbf{p}_{j-1}\|^2. \tag{6.18}$$

• *Condition* (6.17) *ensures that* θ_k *is close to* λ_1. Since

$$\theta_k = \frac{\lambda_1\alpha_k^2 + \alpha_k\mathbf{d}^T\mathbf{b}_k + \mathbf{b}_k^T\mathbf{C}\mathbf{b}_k}{\alpha_k^2 + \|\mathbf{b}_k\|^2},$$

we have

$$
\begin{aligned}
|\theta_k - \lambda_1| &= \frac{\left|\lambda_1 \alpha_k^2 + \alpha_k \mathbf{d}^T \mathbf{b}_k + \mathbf{b}_k^T \mathbf{C} \mathbf{b}_k - \lambda_1 \left(\alpha_k^2 + \mathbf{b}_k^T \mathbf{b}_k\right)\right|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} \\
&\leq \frac{\left|\mathbf{b}_k^T (\mathbf{C} - \lambda_1) \mathbf{b}_k\right|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{|\alpha_k| \|\mathbf{b}_k\| \|\mathbf{d}\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} \\
&\leq \frac{\|\mathbf{b}_k\|^2 \, \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{b}_k\| \|\mathbf{d}\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}}, \qquad (6.19)
\end{aligned}
$$

where the last inequality uses the fact that $|\alpha_k| \leq \sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}$. Now condition (6.17) implies that the Rayleigh quotient θ_k is sufficiently close to the eigenvalue λ_1:

$$
|\theta_k - \lambda_1| \leq \frac{\varepsilon}{h \|\mathbf{M}\|}. \qquad (6.20)
$$

The next step of the iteration, with time-step $h > 0$ specified in the statement of the theorem, produces

$$
\begin{bmatrix} \alpha_{k+1} \\ \mathbf{b}_{k+1} \end{bmatrix} = \mathbf{p}_{k+1} = \mathbf{p}_k + h \mathbf{N}^{-1} (\theta_k - \mathbf{A}) \mathbf{p}_k = \begin{bmatrix} \alpha_k + \eta h \left((\theta_k - \lambda_1) \alpha_k - \mathbf{d}^T \mathbf{b}_k\right) \\ (\mathbf{I} + h \mathbf{M} (\theta_k - \mathbf{C})) \mathbf{b}_k \end{bmatrix}.
$$

Adding zero in a convenient way gives

$$
\begin{aligned}
\|\mathbf{b}_{k+1}\| &= \|(\mathbf{I} + h \mathbf{M} (\lambda_1 - \mathbf{C})) \mathbf{b}_k + h (\theta_k - \lambda_1) \mathbf{M} \mathbf{b}_k\| \\
&\leq \|\mathbf{I} + h \mathbf{M} (\lambda_1 - \mathbf{C})\| \|\mathbf{b}_k\| + h |\lambda_1 - \theta_k| \|\mathbf{M}\| \|\mathbf{b}_k\| \\
&\leq (\gamma + \varepsilon) \|\mathbf{b}_k\|. \qquad (6.21)
\end{aligned}
$$

In particular, since $0 \leq \gamma + \varepsilon < 1/\kappa(\mathbf{N}) \leq 1$, this guarantees a fixed reduction in the component of the forward Euler iterate in the unwanted eigenspace. (The second inequality follows from condition (6.14) and bound (6.20).) After checking a few details, we shall see that this condition is the key to convergence.

• *Subsequent Rayleigh quotients must also remain close to* λ_1. We now show that the new Rayleigh quotient, θ_{k+1}, automatically satisfies the requirement (6.20) with the same $\varepsilon > 0$ and time-step. Repeating the calculation that culminated in (6.19), we obtain

$$
|\theta_{k+1} - \lambda_1| \leq \frac{\|\mathbf{b}_{k+1}\|^2 \, \|\mathbf{C} - \lambda_1\|}{\alpha_{k+1}^2 + \|\mathbf{b}_{k+1}\|^2} + \frac{\|\mathbf{d}\| \|\mathbf{b}_{k+1}\|}{\sqrt{\alpha_{k+1}^2 + \|\mathbf{b}_{k+1}\|^2}}.
$$

Now we use (6.18), a consequence of the monotonic drift from the invariant manifold, to deduce that

$$
\begin{aligned}
|\theta_{k+1} - \lambda_1| &\leq \frac{\kappa(\mathbf{N}) (\gamma + \varepsilon)^2 \|\mathbf{b}_k\|^2 \, \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\sqrt{\kappa(\mathbf{N})} (\gamma + \varepsilon) \|\mathbf{d}\| \|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}} \\
&\leq \frac{\|\mathbf{b}_k\|^2 \, \|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{d}\| \|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}},
\end{aligned}
$$

since $\gamma + \varepsilon < 1/\sqrt{\kappa(\mathbf{N})}$. The condition (6.17) then implies that

$$
|\theta_{k+1} - \lambda_1| \leq \frac{\varepsilon}{h \|\mathbf{M}\|},
$$

which guarantees that the Rayleigh quotient cannot wander too far from λ_1.

• *Subsequent iterates and Rayleigh quotients must eventually converge.* The bound on $|\theta_{k+1} - \lambda_1|$ just established allows us to repeat the argument resulting in (6.21) at future steps, giving

$$\|\mathbf{b}_{k+m}\| \leq (\gamma + \varepsilon)^m \|\mathbf{b}_k\|$$

along with, via a slight modification of (6.18),

$$(6.22)\ |\theta_{k+m} - \lambda_1| \leq \frac{\kappa(\mathbf{N})(\gamma+\varepsilon)^{2m}\|\mathbf{b}_k\|^2\,\|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\sqrt{\kappa(\mathbf{N})}(\gamma+\varepsilon)^m\|\mathbf{d}\|\|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}}$$
$$\leq \frac{\|\mathbf{b}_k\|^2\,\|\mathbf{C} - \lambda_1\|}{\alpha_k^2 + \|\mathbf{b}_k\|^2} + \frac{\|\mathbf{d}\|\|\mathbf{b}_k\|}{\sqrt{\alpha_k^2 + \|\mathbf{b}_k\|^2}}.$$

Thus, $|\theta_{k+m} - \lambda_1| \leq \varepsilon/(h\|\mathbf{M}\|)$ for all $m \geq 1$. As $\|\mathbf{b}_{k+m}\| \to 0$, the component in the desired eigenvector does not vanish, as again a generalization of (6.18) gives

$$\|\mathbf{p}_{k+m}\| \geq \frac{1}{\sqrt{\kappa(\mathbf{N})}}\|\mathbf{p}_0\|.$$

Thus, with $\mathbf{x}_1 = \mathbf{e}_1$, we have

$$\sin\angle(\mathbf{p}_{k+m}, \mathbf{x}_1) = \min_{\xi} \frac{\|\xi\mathbf{p}_{k+m} - \mathbf{x}_1\|}{\|\mathbf{x}_1\|} = \min_{\xi} \left\| \begin{bmatrix} \xi\alpha_{k+m} - 1 \\ \xi\mathbf{b}_{k+m} \end{bmatrix} \right\|$$
$$\leq \frac{\|\mathbf{b}_{k+m}\|}{|\alpha_{k+m}|} \leq (\gamma+\varepsilon)^m \frac{\|\mathbf{b}_k\|}{|\alpha_{k+m}|},$$

where we have taken $\xi = \alpha_{k+m}^{-1}$ for the first inequality. As $|\alpha_{k+m}|$ is bounded away from zero, we have $\sin\angle(\mathbf{p}_{k+m}, \mathbf{x}_1) = O((\gamma+\varepsilon)^m)$ as $m \to \infty$. Since $\|\mathbf{b}_{k+m}\| \to 0$ as $m \to \infty$, we can take the ε used in (6.19) to be arbitrarily small as the iterations progress, giving the asymptotic rate given in (6.15). Similarly, from (6.22) we observe that the Rayleigh quotient converges as in (6.16). The $O(\gamma^m)$ term in that bound falls out if $\mathbf{d} = \mathbf{0}$. □

We now make several remarks concerning Theorem 6.14 and its proof. (1) As $\mathbf{N}$ becomes increasingly ill-conditioned, the hypothesis (6.14) in the theorem becomes more and more difficult to satisfy. We can only guarantee convergence for an ill-conditioned preconditioner if that preconditioner gives a small value of γ, i.e., if it gives a rapid convergence rate. (2) A curiosity of condition (6.17) is that the requirement is more strict when convergence is slower, i.e., when γ is near $\kappa(\mathbf{N})^{-1/2}$. (3) One does not in general know whether θ_k falls to the left or right of λ_1. If $\mathbf{A}$ is normal, then as θ_k must fall the convex hull of its spectrum, and so $\theta_k \geq \lambda_1$; for nonnormal $\mathbf{A}$, it is possible that $\theta_k < \lambda_1$. (4) The proof of the theorem exploits the monotonic drift from the manifold described by Theorem 6.2. This drift is easily monitored, so providing a useful (and cheap) check on convergence of the iteration during computation. If this drift reaches a point where it is not small, projection to the quadratic manifold is easily undertaken; see [18, Chapter IV] for further information.

Theorem 6.3 considers the general case of nonsymmetric $\mathbf{A}$ and a somewhat stringent notion of preconditioning. For the important special case of symmetric positive definite $\mathbf{A}$, Knyazev and Neymeyr [23] provide convergence estimates (and review much literature) for the one-sided forward Euler discretization (6.3). They provide rates of convergence given a symmetric positive definite preconditioner $\mathbf{N}$ for $\mathbf{A}$. However, a connection with dynamical systems is not made and instead optimization is applied to the Rayleigh quotient.

If $\mathbf{M} = \mathbf{I}$, and $\mathbf{C}$ is normal (which is possible even if $\mathbf{A}$ itself is not normal due to $\mathbf{d} \neq \mathbf{0}$) with spectrum given by $\sigma(\mathbf{C}) = \{\lambda_2, \dots, \lambda_n\}$, we can estimate an optimal time-step as follows. We wish to minimize

$$\gamma = \max_{i=2,\dots,n} |1 + h(\lambda_1 - \lambda_i)|,$$

a simple minimax approximation problem on a discrete set; see, e.g., [36, section 8.5]. In particular, if all the eigenvalues are real (i.e., $\mathbf{C}$ is symmetric) and $\lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_n$, then the best h must give

$$1 + h(\lambda_1 - \lambda_2) = -1 - h(\lambda_1 - \lambda_n).$$

This can be solved to obtain $h = 2/(\lambda_2 + \lambda_n - 2\lambda_1)$, from which we compute

$$\gamma = \frac{\lambda_n - \lambda_2}{\lambda_n + \lambda_2 - 2\lambda_1}.$$

Notice that this agrees with the convergence rate of the power method applied to $\mathbf{A} - \sigma\mathbf{I}$ for the optimal shift $\sigma = \frac{1}{2}(\lambda_2 + \lambda_n)$ to the leftmost eigenvector $\mathbf{x}_1$; see, e.g., [43, p. 572]. With the optimal choice of h, the forward Euler method recovers the convergence rate of an optimally shifted power method to $\mathbf{x}_1$.

Again, suppose that $\mathbf{M} = \mathbf{I}$, so that $\gamma = \gamma(h) \to 1$ as $h \to 0$. However, this limit need not be approached from below; that is, for some matrices $\mathbf{C}$ we will have $\gamma(h) > 1$ for all h sufficiently small.[2] The behavior of γ in this limit bears a close connection to the *logarithmic norm* of $\lambda_1 - \mathbf{C}$, which is defined as

$$\beta(\lambda_1 - \mathbf{C}) := \lim_{h \downarrow 0} \frac{\|\mathbf{I} + h(\lambda_1 - \mathbf{C})\| - 1}{h};$$

see, e.g., [30], [40, Chapter 17]. In particular, $\gamma(h) < 1$ for all sufficiently small $h > 0$ provided $\beta(\lambda_1 - \mathbf{C}) < 0$. One can show that the logarithmic norm of a matrix coincides with the numerical abscissa, that is, the real part of the rightmost point in the numerical range:

$$\begin{aligned}\beta(\lambda_1 - \mathbf{C}) &= \max_{\mathbf{v} \in \mathbb{C}^{n-1}, \|\mathbf{v}\|=1} \operatorname{Re} \mathbf{v}^*(\lambda_1 - \mathbf{C})\mathbf{v} \\ &= \max\left\{\eta : \eta \in \sigma(\tfrac{1}{2}\left((\lambda_1 - \mathbf{C}) + \left(\lambda_1 - \mathbf{C}^T\right)\right)\right\};\end{aligned}$$

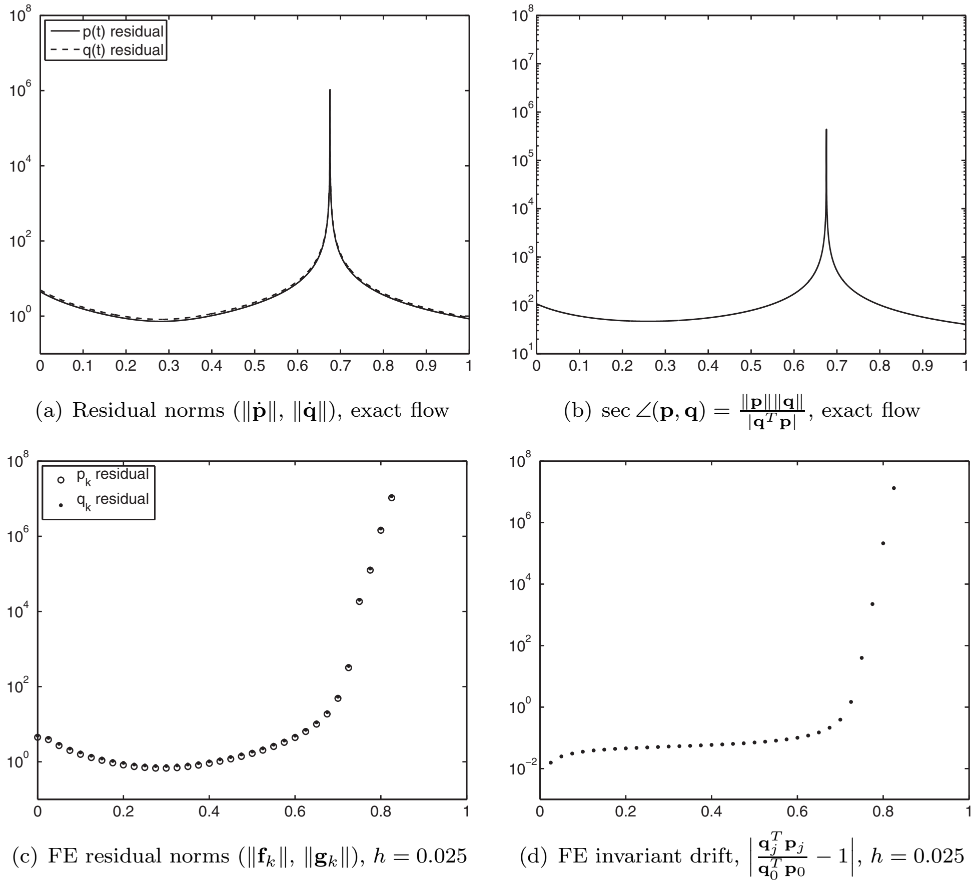
see, e.g., [40, Theorem 17.4]. When is $\gamma(h) > 1$? That is, for what matrices can we not apply our convergence theory by taking h arbitrarily small? We can answer this question by finding requirements on $\mathbf{C}$ that ensure $\beta(\lambda_1 - \mathbf{C}) < 0$. From the above analysis we see that

$$\beta(\lambda_1 - \mathbf{C}) = \lambda_1 - \min_{\mathbf{v} \in \mathbb{C}^{n-1}, \|\mathbf{v}=1\|} \operatorname{Re} \mathbf{v}^*\mathbf{C}\mathbf{v}.$$

Since $\mathbf{C}$ is essentially the restriction $\mathbf{A}|_{\mathbf{x}_1^\perp}$ of $\mathbf{A}$ to the orthogonal complement of the eigenvector $\mathbf{x}_1$, we can summarize as follows.

LEMMA 6.4. *Suppose* $\mathbf{N} = \mathbf{I}$. *Then* $\gamma < 1$ *for all* h *sufficiently small if and only if* λ_1 *is not in the numerical range of* $\mathbf{A}|_{\mathbf{x}_1^\perp}$ *(equivalently,* $\mathbf{C}$*).*

[2] In this case the matrix $\mathbf{A}$ does not satisfy the hypotheses of the theorem; convergence is still possible. Experiments with a small example gave convergence after a bit of initial irregularity.



(a) Residual norms ($\|\dot{\mathbf{p}}\|$, $\|\dot{\mathbf{q}}\|$), exact flow

(b) $\sec\angle(\mathbf{p},\mathbf{q}) = \frac{\|\mathbf{p}\|\|\mathbf{q}\|}{|\mathbf{q}^T\mathbf{p}|}$, exact flow

(c) FE residual norms ($\|\mathbf{f}_k\|$, $\|\mathbf{g}_k\|$), $h = 0.025$

(d) FE invariant drift, $\left|\frac{\mathbf{q}_j^T\mathbf{p}_j}{\mathbf{q}_0^T\mathbf{p}_0} - 1\right|$, $h = 0.025$

FIG. 6.1. *Sampled flow and forward Euler (FE) approximations for the two-sided system with* $\mathbf{T}_\rho^{100}$ *and* $\rho = 1/(20 \cdot 101)$. *The horizontal axis denotes time. Note the blow-up of the exact solution near* $t = 0.675$, *and the consequences of this behavior for the discretized method.*

6.3. Numerical experiments. In this section we investigate Theorems 4.1, 6.1, and 6.3 through several computational examples. Our first experiment applies to the tridiagonal matrix

$$\mathbf{T}_\rho^n \equiv \begin{bmatrix} 2 & -1+\rho & & 0 \\ -1-\rho & 2 & \ddots & \\ & \ddots & \ddots & -1+\rho \\ 0 & & -1-\rho & 2 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

where $n = 100$ and $\rho = 1/(20(n+1))$. The eigenvalues are all real and the condition number of the matrix of eigenvectors is modest. All computations in Figure 6.1 use the same starting vectors $\mathbf{p}_0$ and $\mathbf{q}_0$, which are taken to be (different) random vectors. (Results vary with the other choices for these vectors.)

Figures 6.1(a) and 6.1(b) show the exact solution to the two-sided unpreconditioned system, as given by Theorem 4.1. The residuals $\| \cdot \mathbf{p}\| = \|\mathbf{p}\theta - \mathbf{A}\mathbf{p}\|$ and $\| \cdot \mathbf{q}\| = \|\mathbf{q}\overline{\theta} - \mathbf{A}^*\mathbf{q}\|$ begin to decrease, but then rise as t approaches a critical point

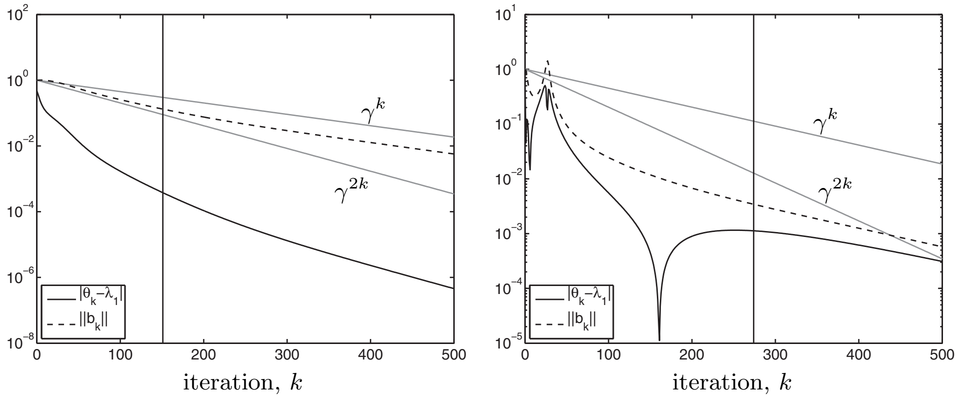


FIG. 6.2. *Computational confirmation of Theorem* 6.3 *for a normal matrix (left) and a nonnormal matrix (right), both with* $\mathbf{N} = \mathbf{I}$. *In the normal case, the residual* $|\theta_k - \lambda|$ *converges like* γ^{2k}, *while in the nonnormal case* $|\theta_k - \lambda|$ *only converges like* γ^k. *The vertical lines denote the point at which the hypotheses of the convergence theorem hold.*

near $t = 0.675$, where cusps develop, indicating that a pole as given by $\pi(t)$ of Theorem 4.1 is encountered by the flow. The same behavior is seen in a plot of the secant of the angle between $\mathbf{p}$ and $\mathbf{q}$. Figures 6.1(c) and 6.1(d) display the discrete flow associated with a forward Euler time integrator with a time step of $h = 0.025$. As expected, when the iterates depart from the quadratic manifold, the residuals explode in size, as in the exact solution. One can also show that the secant of the angle between $\mathbf{p}_j$ and $\mathbf{q}_j$, and the norms of $\mathbf{p}_j$ and $\mathbf{q}_j$, also begin to grow near $t \approx .675$, consistent with Theorem 6.1.

Decreasing the time-step h does not avoid the blow-up—in fact, the time at which the explosive growth occurs is largely independent of the time-step because of the onset of incurable breakdown associated with the continuous dynamical system. In contrast to the latter, the discrete dynamical system cannot simply step over the pole associated with continuous dynamical system. Aside from special cases such as the one described by Theorem 4.2, these results appear to be common and do not significantly depend on specially engineered starting vectors (though breakdown will occur at different points in time, of course). We also implemented the symplectic Euler method (that preserves quadratic invariants) for this class of matrices and observed behavior consistent with the forward Euler method combined with a projection. In contrast, the one-sided discretized forward Euler iterations converge to the left eigenvalue and associated eigenvector.

Next, we investigate the convergence analysis described in Theorem 6.3 for a simple example with $\mathbf{N} = \mathbf{I}$. Let $\mathbf{A}$ be the matrix with $a_{j,j} = (j-1)/(N-1)$ for $j = 1, \dots, N$, and all other entries equal to zero except perhaps for the vector $\mathbf{d}^T$ in entries 2 through N of the first row; cf. (6.11). The plots in Figure 6.2 use $N = 64$, comparing $\mathbf{d}^T = \mathbf{0}$ (left) and $\mathbf{d}^T = [1, \dots, 1]$ (right). In both cases we take $h = 1/2$, for which (6.14) gives $\gamma = 0.992\ldots \in [0, 1)$ as required. We take $\mathbf{p}_0$ to be the same randomly generated unit vector in both cases. This initial vector does not satisfy (6.17), but this condition is eventually met after a number of iterations, denoted by the vertical line in each plot. For the normal case in the left plot, $\|\mathbf{b}_k\|$ converges like γ^k, while the error in the Rayleigh quotient $|\theta_k - \lambda_1|$ converges like γ^{2k} as predicted. The nonnormality induced by the $\mathbf{d}$ vector spoils this convergence for the Rayleigh quotient, as seen in the right plot; now both $\|\mathbf{b}_k\|$ and $|\theta_k - \lambda_1|$ converge like γ^k, consistent with Theorem 6.3. The spikes in the latter plot correspond to points where the Rayleigh quotient θ_k crossed over the desired eigenvalue λ_1, something only possible for nonnormal iterations.

7. Summary.

This paper demonstrates the fruitful relationship between several nonlinear dynamical systems and certain simple preconditioned eigensolvers for nonsymmetric eigenvalue problems. Properties of the continuous-time systems, such as system invariants and the asymptotic behavior of the exact solution, can inform the convergence theory for practical algorithms derived from discretizations, as we illustrate with Theorem 6.1 for the forward Euler discretization. Generalizations to more sophisticated discretizations, along with relaxation of the stringent requirements on the preconditioner in Theorem 6.1, are natural avenues for future research.

Appendix A. Lipschitz constant for Euler's method.

To apply the standard convergence theory for the forward Euler method applied to the system

$$\dot{\mathbf{p}} = \mathbf{N}^{-1}(\theta(\mathbf{p})\mathbf{p} - \mathbf{A}\mathbf{p}),$$

we seek a constant $L > 0$ such that

$$\big\|\mathbf{N}^{-1}(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - \mathbf{N}^{-1}(\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\big\| \le L\,\|\mathbf{u} - \mathbf{v}\|$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. First we note that

$$\|(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - (\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \le \|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| + \|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|.$$

We focus attention on the first term on the right:

$$\begin{aligned}\|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| &\le \|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{u} + \theta(\mathbf{v})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| \\ &\le |\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\| + |\theta(\mathbf{v})|\|\mathbf{u} - \mathbf{v}\| \\ &\le |\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\| + \|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|. \qquad \text{(A.1)}\end{aligned}$$

(In this last inequality and others that follow, we neglect the opportunity to take tighter bounds that would lead to smaller constants but greater analytical complexity.)

Next, we need to bound $|\theta(\mathbf{u}) - \theta(\mathbf{v})|\|\mathbf{u}\|$ in terms of $\|\mathbf{u} - \mathbf{v}\|$. For convenience (assuming neither $\mathbf{u}$ nor $\mathbf{v}$ is zero), define the unit vectors $\widehat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|$ and $\widehat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$, with $\boldsymbol{\varepsilon} = \widehat{\mathbf{v}} - \widehat{\mathbf{u}}$, so that

$$\begin{aligned}|\theta(\mathbf{u}) - \theta(\mathbf{v})| &= \big|\widehat{\mathbf{u}}^T\mathbf{A}\widehat{\mathbf{u}} - \widehat{\mathbf{v}}^T\mathbf{A}\widehat{\mathbf{v}}\big| \\ &= \big|\widehat{\mathbf{u}}^T\mathbf{A}\widehat{\mathbf{u}} - \widehat{\mathbf{u}}^T\mathbf{A}\widehat{\mathbf{u}} - \boldsymbol{\varepsilon}^T\mathbf{A}\widehat{\mathbf{u}} - \widehat{\mathbf{u}}^T\mathbf{A}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\mathbf{A}\boldsymbol{\varepsilon}\big| \\ &\le 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + \|\boldsymbol{\varepsilon}\|^2\|\mathbf{A}\|. \qquad \text{(A.2)}\end{aligned}$$

Now note that

$$\|\boldsymbol{\varepsilon}\| = \|\widehat{\mathbf{v}} - \widehat{\mathbf{u}}\| = \frac{\|\|\mathbf{u}\|\mathbf{v} - \|\mathbf{v}\|\mathbf{v} + \|\mathbf{v}\|\mathbf{v} - \|\mathbf{v}\|\mathbf{u}\|}{\|\mathbf{u}\|\,\|\mathbf{v}\|} \le \frac{|\|\mathbf{u}\| - \|\mathbf{v}\||}{\|\mathbf{u}\|} + \frac{\|\mathbf{u} - \mathbf{v}\|}{\|\mathbf{u}\|}.$$

Apply the triangle inequality to obtain $|\|\mathbf{u}\| - \|\mathbf{v}\|| \le \|\mathbf{u} - \mathbf{v}\|$, from which we conclude

$$\|\boldsymbol{\varepsilon}\| \le \frac{2}{\|\mathbf{u}\|}\|\mathbf{u} - \mathbf{v}\|. \qquad \text{(A.3)}$$

Since $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ are unit vectors, we alternatively have the coarse bound $\|\boldsymbol{\varepsilon}\| = \|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\| \leq 2$, which we can apply to (A.2) to obtain

$$\begin{aligned}|\theta(\mathbf{u}) - \theta(\mathbf{v})| &\leq 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + \|\boldsymbol{\varepsilon}\|^2\|\mathbf{A}\| \\ &\leq 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| + 2\|\boldsymbol{\varepsilon}\|\|\mathbf{A}\| \ = \ 4\|\mathbf{A}\|\|\boldsymbol{\varepsilon}\|.\end{aligned}$$

Now using (A.3), the bound first bound on $\|\boldsymbol{\varepsilon}\|$,

$$|\theta(\mathbf{u}) - \theta(\mathbf{v})| \leq 8\frac{\|\mathbf{A}\|}{\|\mathbf{u}\|}\|\mathbf{u} - \mathbf{v}\|.$$

Substituting this bound into (A.1) gives

$$\|\theta(\mathbf{u})\mathbf{u} - \theta(\mathbf{v})\mathbf{v}\| \leq 9\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|,$$

and, finally, we arrive at the Lipschitz constant

$$\left\|\mathbf{N}^{-1}(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - \mathbf{N}^{-1}(\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\right\| \leq 10\left\|\mathbf{N}^{-1}\right\|\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|.$$

Thus, we define

$$L = 10\left\|\mathbf{N}^{-1}\right\|\|\mathbf{A}\|. \tag{A.4}$$

The Rayleigh quotient $\theta(\mathbf{p})$ is undefined in the case that $\mathbf{p} = \mathbf{0}$. However, as $\|\mathbf{p}\| \to 0$, we have that $\|\theta(\mathbf{p})\mathbf{p} - \mathbf{A}\mathbf{p}\| \to 0$, and this motivates the definition that $\theta(\mathbf{p})\mathbf{p} - \mathbf{A}\mathbf{p} = \mathbf{0}$ if $\mathbf{p} = \mathbf{0}$.

The above analysis excludes the case that $\mathbf{u} = \mathbf{0}$ and/or $\mathbf{v} = \mathbf{0}$, but with our definition of this singular case we have, e.g., if $\mathbf{u} = \mathbf{0}$, that

$$\|\,(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - (\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| = \|(\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| \leq 2\|\mathbf{A}\|\|\mathbf{v}\| \leq 10\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|,$$

and obviously if $\mathbf{u} = \mathbf{v} = \mathbf{0}$, we have

$$\|(\theta(\mathbf{u})\mathbf{u} - \mathbf{A}\mathbf{u}) - (\theta(\mathbf{v})\mathbf{v} - \mathbf{A}\mathbf{v})\| = 0 = 10\|\mathbf{A}\|\|\mathbf{u} - \mathbf{v}\|.$$

Hence, the Lipschitz constant (A.4) holds for all $\mathbf{u}$ and $\mathbf{v}$.

Acknowledgments. We thank Pierre-Antoine Absil, Moody Chu, Kyle Gallivan, Anthony Kellems, Christian Lubich, and Qiang Ye, and anonymous referees for their numerous helpful suggestions concerning this work and its presentation.

REFERENCES

[1] P.-A. Absil, *Continuous-time systems that solve computational problems*, Int. J. Uncov. Comput., 2 (2006), pp. 291–304.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008.

[3] P.-A. Absil, R. Sepulchre, and R. Mahony, *Continuous-time subspace flows related to the symmetric eigenproblem*, Pacific J. Optim., 4 (2008), pp. 179–194.

[4] V. I. Arnold, *Ordinary Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1992.

[5] Z. Bai, D. Day, and Q. Ye, *ABLE: An adaptive block Lanczos method for non-Hermitian eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1060–1082.

[6] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000.

[7] W. Bao and Q. Du, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comp., 25 (2004), pp. 1674–1697.

[8] R. Car and M. Parrinello, *Unified approach for molecular dynamics and density functional theory*, Phys. Rev. Lett., 55 (1985), pp. 2471–2474.

[9] J. Carr, *Applications of Centre Manifold Theory*, Springer-Verlag, New York, 1981.

[10] M. T. Chu, *Curves on s^{n-1} that lead to eigenvalues or their means of a matrix*, SIAM J. Alg. Disc. Math., 7 (1986), pp. 425–432.

[11] M. T. Chu, *On the continuous realization of iterative processes*, SIAM Rev., 30 (1988), pp. 375–387.

[12] M. Embree, *The Arnoldi eigenvalue iteration with exact shifts can fail*, SIAM J. Matrix Anal. Appl., to appear.

[13] M. A. Freitag and A. Spence, *Convergence theory for inexact inverse iteration applied to the generalised nonsymmetric eigenvalue problem*, Electron. Trans. Numer. Anal., 28 (2007), pp. 40–64.

[14] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[15] G. H. Golub and L.-Z. Liao, *Continuous methods for extreme and interior eigenvalue problems*, Linear Algebra Appl., 415 (2006), pp. 31–51.

[16] G. H. Golub and Q. Ye, *Inexact inverse iteration for generalized eigenvalue problems*, BIT, (2000), pp. 671–684.

[17] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[18] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd ed., Springer-Verlag, Berlin, 2006.

[19] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer, London, 1994.

[20] W. Kahan, B. N. Parlett, and E. Jiang, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Num. Anal., 19 (1982), pp. 470–484.

[21] A. V. Knyazev, *Preconditioned eigensolvers—an oxymoron?*, Elec. Trans. Numer. Anal., 7 (1998), pp. 104–123.

[22] A. V. Knyazev and K. Neymeyr, *Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method*, Elec. Trans. Numer. Anal., 7 (2003), pp. 38–55.

[23] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.

[24] Y.-L. Lai, K.-Y. Lin, and W.-W. Lin, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 1 (1997), pp. 1–13.

[25] R. B. Lehoucq and A. J. Salinger, *Large-scale eigenvalue calculations for stability analysis of steady flows on massively parallel computers*, Internat. J. Numer. Methods Fluids, 36 (2001), pp. 309–327.

[26] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, Cambridge, 2005.

[27] R. Mahony and P.-A. Absil, *The continuous time Rayleigh quotient flow on the sphere*, Linear Algebra Appl., 368 (2003), pp. 343–357.

[28] Y. Nakamura, K. Kajiwara, and H. Shiotani, *On an integrable discretization of the Rayleigh quotient gradient system and the power method with a shift*, J. Comput. Appl. Math., 96 (1998), pp. 77–90.

[29] T. Nanda, *Differential equations and the QR algorithm*, SIAM J. Numer. Anal., 22 (1985), pp. 310–321.

[30] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.

[31] E. E. Osborne, *On pre-conditioning of matrices*, J. ACM, 7 (1960), pp. 338–345.

[32] B. N. Parlett, *The Symmetric Eigenvalue Problem*, no. 20 in Classics in Applied Mathematics, SIAM, Philadelphia, 1998. Amended reprint of 1980 Prentice-Hall edition.

[33] B. N. Parlett and C. Reinsch, *Balancing a matrix for calculation of eigenvalues and eigenvectors*, Numer. Math., 13 (1969), pp. 293–304.

[34] M. C. Payne, M. P. Teeter, D. C. Allan, T. Arias, and J. Joannopoulos, *Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients*, Rev. Mod. Phys, 64 (1992), pp. 1045–1097.

[35] B. T. Polyak, *Introduction to Optimization*, Translation Series in Mathematics and Engineering, Optimization Software, Inc., New York, 1987.

[36] M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, 1981.

[37] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.

[38] G. W. Stewart and J.-g. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.

[39] W. W. Symes, *The QR algorithm and scattering for the finite nonperiodic Toda lattice*, Physica D, 4 (1982), pp. 275–280.

[40] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.

[41] J. S. Warsa, T. A. Wareing, J. E. Morel, J. M. McGhee, and R. B. Lehoucq, *Krylov subspace iterations for deterministic k-eigenvalue calculations*, Nuc. Sci. Engrg., 147 (2004), pp. 26–42.

[42] D. S. Watkins, *Isospectral flows*, SIAM Rev., 26 (1984), pp. 379–391.

[43] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.