Controllability of the shifted inverse power iteration: The case of real shifts*

Uwe Helmke
Mathematisches Institut
Universitaet Wuerzburg
97074 Wuerzburg, Germany
helmke@mathematik.uni-wuerzburg.de

Fabian Wirth
Zentrum für Technomathematik
Universität Bremen
28334 Bremen, Germany
fabian@math.uni-bremen.de

Abstract

Controllability properties of the inverse power method on projective space are investigated. For complex eigenvalue shifts a simple characterization of the reachable sets in terms of invariant subspaces can be obtained. The real case is more complicated and is investigated in this paper. Necessary and sufficient conditions for complete controllability are obtained in terms of the solvability of a matrix equation. Partial results on the solvability of this matrix equation are given.

1 Introduction

Numerical matrix eigenvalue methods such as the QR algorithm or inverse power iterations provide interesting examples of nonlinear discrete dynamical systems defined on Lie groups or homogeneous spaces. A typical approach from numerical linear algebra to improve convergence properties of such algorithms is via suitable shift strategies for the eigenvalues. We refer to [2, 3, 5] for papers studying the shifted inverse iteration. In particular, the case of complex shifts is studied in [7].

Such eigenvalue shifts can be viewed as control variables and the resulting algorithms can therefore be analyzed using tools from nonlinear control theory. So far the analysis and design of shift strategies in numerical eigenvalue algorithms has been more a kind of an art rather than being guided by systematic design principles. The situation here is quite similar to that of control theory in the 50's before the introduction of state space methods. The advance made during the past two decades in nonlinear control theory indicates that the time may now be ripe for a more systematic investigation of control theoretic aspects of numerical linear algebra.

In this paper we investigate the controllability properties of the well known inverse power iterations for finding the dominant eigenvector of a matrix using *real* shifts. Shifted versions of the inverse power method lead to nonlinear control systems on projective space, or more generally on Grassmann manifolds. The complex case has been studied in [4].

The shifted inverse iteration $\mathbb{R}^n$ in its controlled form is given by

$$x(t+1) = \frac{(A - u(t))^{-1} x(t)}{\|(A - u(t))^{-1} x(t)\|}, \quad t \in \mathbb{N}, \tag{1}$$

where $u(t) \notin \sigma(A)$. The trajectory corresponding to an initial condition x_0 and a control sequence $u = (u(0), u(1), \ldots)$ is denoted by $\phi(t, x_0, u)$. Via the choice $u(t) = x^*(t) A x(t)$ we obtain the Rayleigh quotient iteration studied in [2], [3]. The Rayleigh iteration may therefore

*Research partially supported by the European Nonlinear Control Network.

be interpreted as a *feedback strategy* for the shifted inverse iteration. It is known that in some cases this feedback strategy has undesirable properties [3].
This is a shortened version. Due to space constraints full details of the proofs cannot be given here and will appear elsewhere.

2 The shifted inverse power iteration on projective space

Clearly, if the initial condition x_0 for system (1) lies in an invariant subspace of A than the same holds true for the trajectory $\phi(t, x_0, u)$ regardless of the control sequence u. Therefore we may restrict our attention to those points not lying in a nontrivial invariant subspace of A. Using the fact that the interesting dynamics of (1) are on the unit sphere and identifying opposite points which give no further information about the dynamics we define our state space of interest to be

$$M := \mathbb{P}^{n-1} \setminus \bigcup_{AV \subset V, 0 < dim V < n} V \,, \tag{2}$$

where $\mathbb{P}^{n-1}$ denotes the real projective space of dimension $n-1$. Thus M is the projection of the cyclic vectors of A, that is those $x \in \mathbb{R}^{n \times n}$ such that $\{x, Ax, \dots, A^{n-1}x\}$ is a basis of $\mathbb{R}^n$. In the following we assume that A has a cyclic vector, i.e. $M \neq \emptyset$. To keep notation short let us introduce the union of A-invariant subspaces $\mathcal{V}(A) := \bigcup_{AV \subset V, 0 < dim V < n} V$. The system on M is then given by

$$\begin{aligned} \xi(t+1) &= \mathbb{P}(A - u(t))^{-1} \xi(t) \,, \quad t \in \mathbb{N} \\ \xi(0) &= \xi_0 \in M \,, \end{aligned} \tag{3}$$

where $u(t) \in U := \mathbb{R} \setminus \sigma(A)$ (the set of admissible control values) and $\mathbb{P}$ denotes the natural projection onto the projective space. We denote the space of finite and infinite admissible control sequences by U^t, resp. $U^{\mathbb{N}}$. The solution of (3) corresponding to the initial value ξ_0 and a control sequence $u \in U^{\mathbb{N}}$ is denoted by $\varphi(t; \xi_0, u)$. The forward orbit of a point ξ given by

$$\mathcal{O}^+(\xi) := \{\eta \in M \mid \exists t \in \mathbb{N}, u \in U^t \text{ such that } \eta = \varphi(t; \xi, u)\} \,.$$

Similarly the set of points reachable exactly in time t is denoted by $\mathcal{O}_t^+(\xi)$. In order to analyze the controllability properties of (3) we introduce the following definition of regions of approximate controllability in $\mathbb{P}^{d-1}$. A *control set* of system (3) is a set $D \subset M$ satisfying

(i) $D \subset \operatorname{cl} \mathcal{O}^+(\xi)$ for all $\xi \in D$.

(ii) For every $\xi \in D$ there exists a $u \in U$ such that $\varphi(1, x, u) \in D$.

(iii) D is a maximal set (with respect to inclusion) satisfying (i).

System (3) is called *forward accessible* [1], if the forward orbit $\mathcal{O}^+(\xi)$ of every point $\xi \in \mathbb{P}^{d-1}$ has nonempty interior and *uniformly forward accessible (in time t)* if there is a $t \in \mathbb{N}$ such that $\operatorname{int} \mathcal{O}_t^+(\xi) \neq \emptyset$ for all $\xi \in M$. Note that $\operatorname{int} \mathcal{O}^+(\xi) \neq \emptyset$ holds iff there is a $t \in \mathbb{N}$ such that $\operatorname{int} \mathcal{O}_t^+(\xi) \neq \emptyset$. For such a t there exists by Sard's theorem a $u \in U^t$ such that

$$\operatorname{rk} \frac{\partial \varphi(t; \xi, u)}{\partial u} = n - 1 \,.$$

A pair $(\xi, u) \in M \times U^t$ is called *regular* if this rank condition holds. $u \in U^t$ is called *universally regular* if (ξ, u) is a regular pair for every $\xi \in M$. By [6, Corollaries 3.2 & 3.3] forward accessibility is equivalent to the fact that the set of universally regular control sequences U^t_{reg} is open and dense in U^t for all t large enough. (For a precise statement we refer to [6].)
The following result clarifies the situation for the systems under consideration here:

Lemma 2.1 *System (3) is uniformly forward accessible in time $n-1$. A control sequence $u \in U^t$ is universally regular if and only if there are $n-1$ pairwise different values in the sequence $u(0), \dots, u(t-1)$.*

An important subset of a control set D is its *core* defined by

$$\operatorname{core}(D) := \{\xi \in D \mid \operatorname{int} \hat{\mathcal{O}}^-(\xi) \cap D \neq \emptyset \text{ and } \operatorname{int} \hat{\mathcal{O}}^+(\xi) \cap D \neq \emptyset\}.$$

Here $\hat{\mathcal{O}}^-(\xi)$ denotes the points $\eta \in \mathbb{P}^{d-1}$ such that there exist $t \in \mathbb{N}$, $u_0 \in \operatorname{int} U^t$ such that $\varphi(t; \eta, u_0) = \xi$ and (η, u_0) is a regular pair.

3 Statement of Results

Theorem 3.1 *Let $A \in \mathbb{R}^{n \times n}$ be cyclic with characteristic polynomial q. Consider the system (3) on M. The following statements are equivalent:*

(i) There exists a control set $D \subset M$ with $\operatorname{int} D \neq \emptyset$.

(ii) M is a control set of system (3).

(iii) There exists a universally regular control sequence $u \in U^t$ such that

$$\prod_{s=0}^{t-1} (A - u(s))^{-1} \in R^* I. \tag{4}$$

(iv) For every $B \in \Gamma_A := \{p(A) \mid p \in \mathbb{R}[z], p \wedge q = 1\}$ there exist $t \in \mathbb{N}$, $u \in U^t_{reg}$, $\alpha \in \mathbb{R}^*$ such that

$$B = \alpha \prod_{s=0}^{t-1} (A - u(s)),$$

i.e. $\Gamma_A = \Gamma_A^{\mathbb{R}} := \{p(A) \mid p = \prod_{s=0}^{t-1} (z - u(s)), u(s) \in \mathbb{R}, p \wedge q = 1\}$.

(v) There exists a polynomial f with only real roots and $n-1$ pairwise different roots, $\alpha \in \mathbb{R}^*$ and $r(z) \in \mathbb{R}[z]$ such that

$$f(z) = \alpha + r(z)q(z). \tag{5}$$

Outline of proof: Statement (iii) is equivalent to (v) by the Caley-Hamilton theorem and the implications (ii) $\Rightarrow$ (i) and (iv) $\Rightarrow$ (iii) are obvious.
"(iii) $\Rightarrow$ (ii)" By assumption every $x \in M$ is an element in the core of a control set, [8]. Thus by [8] every connected component of M is contained in the core of a control set. It thus remains to show that it is possible to steer from any connected components of M into any other. To be precise we have to show for every two connected components $Z_1, Z_2 \subset M$ that there exist

$x_i \in Z_i, i = 1,2$ and a control $u \in U^t$ such that $x_2 = \varphi(t, x_1, u)$. Note that different connected components of M are separated by the $n-1$-dimensional A-invariant subspaces. Let the (real) Jordan form of A be given by

$$TAT^{-1} = \begin{pmatrix} \mathrm{diag}(J(\lambda_1), \ldots, J(\lambda_k)) & 0 \\ 0 & B \end{pmatrix},$$

where the $J(\lambda_i)$ are Jordan blocks to real eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_k$ and B has only complex eigenvalues. For $i = 1, \ldots, k$ let x_i denote the unit Jordan vector to $J(\lambda_i)$ of highest index, then the $n-1$-dimensional A-invariant subspaces are given exactly by the sets $T^{-1}\{x \in \mathbb{R}^n \mid x_i = 0\}, i = 1, \ldots, k$. The connected components of M can then be described by index sets I, where $I \subset \{1, \ldots, k\}$ and the connected components of M are of the form

$$Z = \mathbb{P}\{x \in \mathbb{R}^n \setminus \mathcal{V}(A) \mid x_j > 0\,,\, j \in I, x_j < 0\,,\, j \notin I\}\,.$$

(Note that by multiplication by -1 there are now for each connected component two representations of the above kind. But this is irrelevant for our purposes.) It is now clear how it is possible to steer from one connected component into another. Namely, if one chooses $\lambda_j > u > \lambda_j + 1$ then the connected component given by the index set I is mapped onto the connected component given by the index sets $\tilde{I}$, where $\tilde{I} = \{i \in I \mid i \geq j\} \cup \{i \notin I \mid i < j\}$. Using this representation it is easy to see that indeed it is possible to steer from any connected component into any other.
"(i) $\Rightarrow$ (iii)" Let $D \subset M$ be a control set. By Theorem 15 in [8] for every open set $W \subset \mathrm{core}(D)$ there exists an $\xi \in W$ and $u \in U^t_{reg}$ such that $\xi = \varphi(t, \xi, u)$. Choose W such that for all $\eta \in W$ the representation $\eta = \mathbb{P}\sum_{i=1}^{n} \alpha_i x_i$ in terms of a basis given by (generalized) eigenvectors of A implies $\alpha_i \neq 0, i = 1, \ldots, n$. For such an η we now have a representation

$$\eta = \mathbb{P}\prod_{s=0}^{t}(A - u(s))^{-1}\eta\,.$$

That is in the basis $\{x_1, \ldots, x_n\}$ (with an associated change of basis T) we have

$$\alpha = c\prod_{s=0}^{t}(TAT^{-1} - u(s))^{-1}\alpha\,,$$

for a suitable constant $c \neq 0$ and $\alpha = (\alpha_1, \ldots, \alpha_n)$. In particular, for each Jordan block $J(\lambda)$ of A corresponding to the indices $i_{\lambda,1}, \ldots, i_{\lambda,k(\lambda)}$ it holds that

$$\begin{bmatrix} \alpha_{i_{\lambda,1}} \\ \vdots \\ \alpha_{i_{\lambda,k(\lambda)}} \end{bmatrix} = c\prod_{s=0}^{t}(J(\lambda) - u(s))^{-1} \begin{bmatrix} \alpha_{i_{\lambda,1}} \\ \vdots \\ \alpha_{i_{\lambda,k(\lambda)}} \end{bmatrix},$$

An easy calculation shows that this implies $c\prod_{s=0}^{t}(J(\lambda) - u(s))^{-1} = I$ for every $\lambda \in \sigma(A)$.
(ii) $\Rightarrow$ (iv): The proof of this statement is omitted. It follows the ideas for the complex case in [4]. □

The unusual fact about the system we are studying is thus that by the universally regular representation of one element of the system's semigroup we can immediately conclude that the system is completely controllable.

For brevity we will call a cyclic matrix A *II-controllable* (for inverse iteration controllable), if A satisfies any of the equivalent conditions of Theorem 3.1.
An immediate consequence of Theorem 3.1 lies in a complete characterization of the reachable sets of the complete projected inverse power iteration given by

$$\xi(t+1) = \mathbb{P}(A - u(t))^{-1}\xi(t)\,, \quad t \in \mathbb{N}\,, \qquad \xi(0) = \xi_0 \in \mathbb{P}^{n-1}\,, \tag{6}$$

Corollary 3.2 *Let A be II-controllable, then*
(i) for each $\xi \in \mathbb{P}^{n-1}$ we have

$$\begin{aligned} \mathrm{cl}\,\mathcal{O}^+(\xi) &= \mathbb{P} \bigcap_{x\in V, AV\subset V} V = \mathbb{P}\,\mathrm{span}\{x, Ax, A^2x, \ldots, A^{n-1}x\}\,, \\ \mathcal{O}^+(\xi) &= \mathbb{P} \bigcap_{x\in V, AV\subset V} V \setminus \bigcup_{x\notin V, AV\subset V} V\,. \end{aligned}$$

(ii) There is a one-to-one correspondence between
- *The orbits of system (6).*
- *The closures of the orbits of system (6).*
- *The A-invariant subspaces.*

4 Conditions for complete controllability

The result of the previous section raises the question for which cyclic matrices A admit a representation of the form (4) or equivalently when (5) is possible. With respect to this question we have the following preliminary results.

Proposition 4.1 *Let $A \in \mathbb{R}^{n\times n}$ be cyclic.*

(i) A is not II-controllable, if it satisfies one of the following conditions

(i) A has a nonreal eigenvalue of multiplicity greater than one.

(ii) A has a real eigenvalue of multiplicity greater than two.

(iii) A is II-controllable, if $\sigma(A) \subset \mathbb{R}$ and no eigenvalue has multiplicity greater than two.

We state the following trivial lemma, which for all its triviality will provide a way to construct cases in which representations of the form (5) do not exist.

Lemma 4.2 *(i) If for two eigenvalues $\lambda_1, \lambda_2 \in \sigma(A)$ and all $u \in \mathbb{R}$ we have*

$$|\lambda_1 - u| < |\lambda_2 - u|\,,$$

then A is not II-controllable.

(ii) If the spectrum $\sigma(A)$ is symmetric with respect to the imaginary axis, i.e. $\sigma(A) = -\sigma(A)$, then the a representations of the form (5) exists iff there exists a universally regular control sequence $u \in U^t$ such that

$$\prod_{s=0}^{t-1}(A^2 - u^2(s)) \in R^* I\,. \tag{7}$$

In this case A is not II-controllable if for two eigenvalues $\lambda_1, \lambda_2 \in \sigma(A)$ and for all $u \in [0, \infty)$ but one we have
$$|\lambda_1^2 - u| < |\lambda_2^2 - u| .$$

Corollary 4.3 *If one of the following cases is satisfied, the A is not II-controllable.*

(i) There exist $\lambda_1, \lambda_2 \in \sigma(A)$ with $\lambda_1 \neq \lambda_2$ and $\operatorname{Re}\lambda_1 = \operatorname{Re}\lambda_2$.

(ii) There exist $\lambda_i, i = 1, \dots, 4 \in \sigma(A)$ with $\lambda_1 \neq \lambda_2 \in \mathbb{R}$, $\lambda_3 = \overline{\lambda_4}$, $\operatorname{Re}\lambda_3 = \frac{1}{2}(\lambda_1 + \lambda_2)$ and $\operatorname{Im}\lambda_3 \geq \frac{1}{2}|\lambda_1 - \lambda_2|$.

(iii) There exist $\lambda_i, i = 1, \dots, 5 \in \sigma(A)$ with $\lambda_1 \in \mathbb{R}$, $\lambda_2 = \overline{\lambda_3}, \lambda_4 = \overline{\lambda_5}$, $\lambda_1 = \frac{1}{4}\sum_{j=2}^{5}\lambda_j$, $\operatorname{Im}\lambda_2 = \operatorname{Im}\lambda_4 \geq |\operatorname{Re}\lambda_1 - \operatorname{Re}\lambda_2| > 0$.

(iv) There exist $\lambda_i, i = 1, \dots, 6 \in \sigma(A)$ with $\lambda_1 \neq \lambda_2 \in \mathbb{R}$, $\lambda_3 = \overline{\lambda_4}$, $\lambda_5 = \overline{\lambda_6}$, $\frac{1}{2}(\lambda_1 + \lambda_2) = \frac{1}{4}\sum_{j=3}^{6}\lambda_j$ with the properties $|\lambda_j - \frac{1}{2}(\lambda_1 + \lambda_2)| \geq \frac{1}{2}|\lambda_1 - \lambda_2|, j = 3, \dots, 6$ and $\operatorname{Re}(\lambda_j - \frac{1}{2}(\lambda_1 + \lambda_2))^2 \leq \frac{1}{2}|\lambda_1 - \lambda_2|$ for $j = 3, \dots, 6$.

(v) There exist $\lambda_i, i = 1, \dots, 6 \in \sigma(A)$ with $\lambda_1 = \overline{\lambda_2} \in \mathbb{C}$, $\lambda_3 = \overline{\lambda_4}$, $\lambda_5 = \overline{\lambda_6}$, $\operatorname{Re}\lambda_1 = \frac{1}{4}\sum_{j=3}^{6}\lambda_j$ with the property $\operatorname{Re}\lambda_j^2 < -(\operatorname{Im}\lambda_1)^2$.

Finally, for low dimensions the following complete result can be given.

Proposition 4.4 *Let $A \in \mathbb{R}^{n \times n}$ be cyclic.*

(i) IF $n = 1, 2$ then A is II-controllable.

(ii) If $n = 3$ then A is II-controllable if and only if the eigenvalues of A do not have a common real part.

References

[1] F. Albertini and E. Sontag, *Discrete-time transitivity and accessibility: analytic systems*, SIAM J. Contr. & Opt., 31 (1993), pp. 1599–1622.

[2] S. Batterson, Dynamical analysis of numerical systems. Numer. Linear Algebra Appl. 2, No.3, 297-310 (1995).

[3] S. Batterson and J. Smillie, Rayleigh quotient iteration for nonsymmetric matrices. Math. Comput. 55, No.191, 169-178 (1990).

[4] U. Helmke and P.A. Fuhrmann, On controllability of matrix eigenvalue algorithms: The inverse power method, preprint 1998.

[5] I.C.F. Ipsen. Computing an eigenvector with inverse iteration. *SIAM Rev.* 39(2):254-291, 1997.

[6] E. D. Sontag and F. R. Wirth, Remarks on universal nonsingular controls for discrete-time systems, *Syst. Cont. Lett.*, 33(2):81-88, 1998.

[7] T. Suzuki, Inverse iteration method with a complex parameter. *Proc. Japan Acad., Ser. A* 68(3):68-73, 1992.

[8] F. Wirth. Dynamics and controllability of nonlinear discrete-time control systems. In *4th IFAC Nonlinear Control Systems Design Symposium (NOLCOS'98)* Enschede, The Netherlands, 1998.