# Reachable Sets of Numerical Iteration Schemes

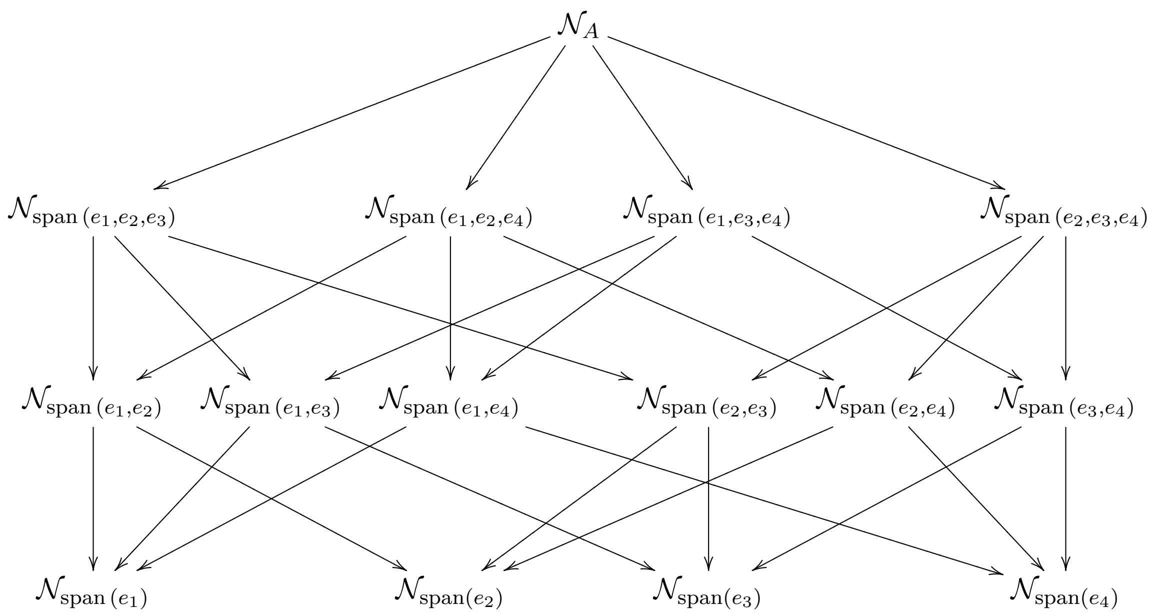
## A System Semigroup Approach

Dissertationsschrift zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilian-Universität Würzburg

vorgelegt von


**Jens Jordan**

aus
Würzburg

# Reachable Sets of Numerical Iteration Schemes

## A System Semigroup Approach

Dissertationsschrift zur Erlangung des naturwissenschaftlichen

Doktorgrades der Bayerischen Julius-Maximilian-Universität Würzburg

vorgelegt von

**Jens Jordan**

aus

**Würzburg**

Eingereicht am: 28.02.2008

bei der Fakultät für Mathematik und Informatik
der Bayerischen Julius-Maximilians-Universität Würzburg

1. Gutachter: Prof. Dr. Uwe Helmke

2. Gutachter: Prof. Dr. Fabian Wirth

Tag der mündlichen Prüfung: 30.07.2008

*"Oh, that was easy," says Man, and for an encore goes on to prove that black is white and gets himself killed on the next zebra crossing.*

*Douglas Adams*

# Contents

# 1 Introduction

Numerical analysis and control theory are two important disciplines in modern mathematics which are closely linked in several aspects. In fact, a large number of numerical techniques have been designed for the treatment of control-theoretical problems. These techniques, algorithms and software packages are necessary tools in engineering applications.

A less traveled road is the converse direction. Many numerical algorithms can be interpreted as dynamical systems and can be analyzed with the corresponding techniques. Interesting examples of such approaches are the works of Ammar and Martin [AM86], Batterson and Smillie, [BS89a, BS89b], Batterson [Bat95] and Shub and Vasquez [SV87], where the dynamics of the QR algorithm and Rayleigh iteration are explored using tools from dynamical systems theory.

Taking one step forward, one can regard the variables of the algorithm – such as shift parameters or step-sizes – as control parameters. Thereby we obtain control systems, which can be studied with the various tools from control-theory. A first step in this direction was established by Gustaffson et al. [GLS88, Gus91, Gus92]. The authors apply simple control-theoretic techniques on step-size selection, such as proportional integral control, to improve the performance of ODE solvers. Other approaches – mainly concerning system solvers, linear and quadratic programming problems and ordinary differential equations – can be found in the recent book of Bhaya and Kaszkurewicz [BK06]. The challenge remains to explore the possibilities that emerge, by applying the full scope of methods from nonlinear control theory.

In this work we investigate iterative numerical algorithms with shifts as nonlinear discrete-time control systems. We emphasize the analysis of reachable sets and their adherence structure. This task is important for three main reasons.

First of all, the design of shift strategies for numerical algorithms often follows heuristic ideas. *The understanding of the algebraic and geometric properties of the reachable sets allows a more systematic way of constructing shift strategies and feedback laws.*

Secondly, the dynamics of algorithms, depends both on the choice of a particular shift strategy as well as on the initial data. Therefore, it is natural to ask if other shift strategies exist, that force the algorithm to converge for generic initial conditions, or if there is a fundamental limitation for the convergence of the algorithm, independent of the choice of shift strategies. Such a *fundamental limitation* might be the following: *The target points are* *not in the topological closure of the initial point.* In such a situation there exists no shift strategy such that the algorithm converges.

Finally, after having understood the reasons why a specific algorithm fails to converge one might be able to *create new algorithms with better convergence behavior.*

In this thesis we will focus mainly on the first two issues, with only few and preliminary results on the third issue.

First attempts to investigate the reachable sets of shifted iterative algorithms are the works of Helmke and Fuhrmann [HF00], Helmke and Wirth [HW01], Chu and Chu [CC06]. All three papers use very different techniques in their analysis. In [HF00], classical inverse iteration with complex shifts are analyzed using polynomial models. The authors show, that there is a bijective correspondence between the topological closures of the reachable sets and the $A$-invariant subspaces. This is not longer the case for classical inverse iteration with real shifts ([HW01]). Here, the authors use the concept of control sets to derive necessary and sufficient conditions for the existence of a dense reachable set. Finally, in [CC06], the authors study the reachable sets of the shifted QR algorithm using matrix decomposition techniques. In particular they show, that that the QR algorithm with shift is neither reflexive nor symmetric.

In this thesis we focus on a different approach that is based on the interpretation of reachable sets as orbits of the system semigroup. The relation between reachable sets and system semigroups has been investigated by several authors, including, e.g., Colonius and Kliemann [CK93, CK00], Mittenhuber [Mit95, Mit01] and Kupka [Kup90] in the continuous-time case and Fliess and Normand-Cyrot [FN81b, FN81a], Mokkabur [Mok89], Agrachev and Gamkrelidze [AG93] and San Martin [San95] for the discrete-time case. Nevertheless, this semigroup approach can run into technical problems. For example, the geometric structure of the system semigroup – viewed as a subset of the diffeomorphism group of $M$ – can be much more complicated than the geometry of the reachable set. Luckily, in the applications in this thesis, the system semigroups are subsemigroups of certain finite dimensional Lie groups. Therefore, we are able to use the underlying differential structure for the investigation of the reachable sets.

Since we are not interested just in reachable sets, but also their boundary points we need to investigate the *adherence structure* of the system, i.e., we analyze if a reachable set is in the topological closure of another reachable set. For this investigation we proceed in three steps.

In the first step we investigate the structure of the system group orbits, i.e., the orbits of the group generated by the system semigroup. Here, we apply a geometrical framework, that has been developed by Jakubzyk, Sontag and others (see [JS90, AS91, AS93]). This expands the well-known Lie-theoretical theory for nonlinear continuous-time systems to a discrete-time setting.

Clearly, the reachable sets are subsets of the corresponding system group orbits. Thus, in a second step of the analysis, we investigate the structure of the reachable sets within a given system group orbit. In this step it is very useful to understand the relation of the system semigroup to the system group. The investigation of this relation will be an important topic in this thesis.

In the case of iterative numerical methods, the target points, such as eigenvectors or solutions of linear equations, are outside of the system group orbit of the initial point, but lie on the boundary of this orbit. Thus, in a third step, we investigate the adherence structure of the system group orbit and the reachable sets. Here, so-called repelling phenomena might occur, i.e., it can happen that the boundary of the orbit and the topological closure of any reachable set of points in this orbit, are disjoint. In this situation there exists no shift strategy such that the controlled sequence converges to the desired solution, regardless how close the initial guess has been. We derive necessary and sufficient conditions for such phenomena.

In Part II of this thesis we apply the semigroup approach to the investigation of the following four numerical iteration schemes.

**Classical inverse iteration** is a method for the calculation of eigenvectors of a given matrix. Given a quadratic matrix $A$ the dynamics of inverse iteration is given by

$$x_{t+1} = (A - u_t I)^{-1} \cdot x_t, \quad x_0 \in \mathbb{RP}^{n-1}. \tag{1}$$

Here $I$ is the identity matrix and $(A - u_t I)^{-1}$ acts canonically on the projective space of lines in $\mathbb{R}^n$. Specific shift strategies yield well-established numerical algorithms, such as inverse power iteration (for constant shifts) or Rayleigh quotient iteration (for Rayleigh shifts). Although, the basic idea of inverse iteration was already introduced by Wielandt in 1944 (see [Wie44]), there is still a lot of active research in this area. For an overview about the history and the state of the art see Ipsen [Ips96, Ips97]. Recent results are e.g. Neymeyer [Ney01], Simoncini and Elden [SE02], Freitag and Spencer [FS07]. It is well known, that inverse iteration with Rayleigh shift converges for almost all symmetric matrices and almost all initial conditions (see Parlett and Kahan [PK69]). In fact, Batterson and Smillie provided a proof based on dynamical system theory, that the set of symmetric matrices for which inverse iteration with Rayleigh shift converges is open and dense (see [BS89a]). On the other hand, Batterson and Smillie also showed that inverse iteration with Rayleigh shift fails for an open set of non-symmetric matrices [BS89b]. It is unknown if there exists a shift strategy such that inverse iteration converges for a generic set of matrices and a generic set of initial conditions. It is actually this lack of theoretical understanding of the inverse iteration method, or closely related, of the QR-iteration, that has motivated this type of research into the geometric analysis of reachable sets. For inverse iteration with complex shifts this structure is now fully understood. More precisely, the reachable sets coincides with the orbits of the centralizer group action (see Helmke and Fuhrmann [HF00]). It contrast, the real case is much more complicated then the complex case and far from being understood. First results for the real case, such as conditions for almost controllability, can be found in Helmke and Wirth [HW01].

Inverse iteration schemes can also be applied to other types of manifolds. For example, inverse iteration on flag manifolds and on Hessenberg varieties are of interest from the numerical point of view, since they are closely related to the QR algorithm (see Ammar and Martin [AM86]). Chu and Chu pointed out, that in general a shifted QR transformation is not invertible by a sequence of shifted QR transformations (see [CC06]). The same phenomenon holds for other generalized inverse iteration system and can easily be explained via the system semigroup approach, since here the reachable sets are smaller then the system group orbits.

**Rational iteration** is an extension from inverse iteration, using a second shift parameter $v_t$. This yields the iteration scheme on projective space

$$x_{t+1} = (A - u_t I)^{-1}(A - v_t I) \cdot x_t, \quad x_0 \in \mathbb{RP}^{n-1} \tag{2}$$

with two control parameters $u_t, v_t$. Rational iteration schemes have been applied in the field of eigenvalue computation as well as for linear equation solvers (see, e.g., Ruhe [Ruh84], Jahrlebring and Voss [JV05], Yong and Vono [YV92]). A one-parameter version of rational iteration is **Cayley iteration**, i.e.,

$$x_{t+1} = (A - u_t I)^{-1}(A + u_t I) \cdot x_t, \quad x_0 \in \mathbb{RP}^{n-1}. \tag{3}$$

Cayley iteration steps have been proposed by several authors (see, e.g., Meerbergen, Spencer and Roose [MSR94], Lehoucq and Meerbergen [LM98]). If $A$ is element of a classical Lie algebra, the Cayley-transform yields an element of the corresponding Lie group, a simple fact that streamlines the Lie group approach to such systems. Nevertheless, to our knowledge, there exists no systematic investigation on the reachable sets for rational iteration schemes. Clearly, the reachable sets of both schemes are always group orbits. We show that for a large set of matrices, but not for all matrices, the reachable sets of rational iteration and Cayley iteration coincide.

Moving from eigenvalue methods to linear equation solvers, we consider **Richardson's method**

$$x_{t+1} = x_t - u_t(Ax_t - b), \quad x_0 \in \mathbb{R}^n. \tag{4}$$

Clearly, a fixed point of this iteration is a solution of the linear equation $Ax = b$. The literature proposed different shift strategies, each of them for certain families of matrices, (see, e.g., Opfer and Schober [OS84], Smorlaski and Saylor [SS88], Golub and Overton [GO88], Calvetti and Reichel [CR96]). In particular, a constant shift strategy $u_t = u$ yields the trivial splitting method, which converges if and only if $\mathrm{Spec}(I - uA)$ lies in the unit disc. Another interesting shift strategy is given by the feedback law $u_t = r_t^\top Ar_t/\|Ar_t\|^2$ with $r_t = b - Ax_t$. This approach yields GMRES(1) which converges if $A + A^\top$ is positive definite. However, a systematic analysis of the reachable sets of Richardson's methods is missing.

A generalization of Richardson's methods are **restarted polynomial iteration of order** $m$

$$x_{t+1} = (I - p_t(A)A)x_t + p_t(A)b, \quad x_0 \in \mathbb{R}^n. \tag{5}$$

Here the controls $p_t$ are polynomials of degree at most $m$. Polynomial restarted iteration can be considered as restarted Krylov methods. See Sorensen [Sor02] for an overview on Krylov methods and polynomial restarting. Note that this setting includes the celebrated GMRES(m) method, which is commonly used in praxis but only partly understood in theory (see Eiermann, Ernst and Schneider [EES00], Joubert [Jou94]). In particular Embree showed some simple examples where GMRES(1) converges while GMRES(2) stagnates ([Emb03]). This phenomena can be extremely sensitive subject to small changes in the initial conditions.

To improve controllability properties we introduce **linear control schemes** as an alternative to the bilinear Richardson's method. Explicitly, we consider

$$x_{t+1} = (I - A)x_t + Bu_t + b, \quad x_0 \in \mathbb{R}^n \tag{6}$$

that has $A^{-1}b$ as an fixed point for the zero control $u_t = 0$. Here, the choice of $B$ can be used to improve the convergence behavior. Linear control systems are well understood (e.g., Kailath [Kai80] and Kučera [Kuc79]). It is known, that (6) is for almost all pairs $(I - A, B)$ controllable. We show that also in many of the uncontrollable cases the topological closure of any reachable set contains the solution of $Ax = b$. For almost all cases a convergent shift strategy $u_t = Kx_t$ can be constructed using linear quadratic controller design, a well-known optimal control technique (see, e.g., Lancester and

Rodman [LR95]). This yields a globally convergent iterative algorithm, called LQRES, for solving linear systems presented by Helmke and Jordan [HJ05].

## 1.1 Main results

The main achievements of this thesis are the following:

- **Development of tools for the systematic analysis of the adherence structure of reachable sets**. We develop a framework merging classical concepts, such as geometric control theory, semigroups and graphs. This framework will be helpful for the analysis of discrete-time control systems.

- **Analysis of the reachable sets of numerical iteration schemes**. We extend the known results about the reachable sets of inverse iteration schemes. Moreover, we investigate the reachable sets of rational iteration schemes, Richardson's methods and linear control schemes.

Now we give a more detailed description. This thesis is divided in two parts. In Part I of this thesis we develop techniques to analyze the structure of reachable sets of invertible discrete-time control systems.

In Chapter 2 we clarify definitions and notations which will be used throughout this manuscript. Moreover, we present some basic observations on discrete-time control systems. We begin with some results on system group orbits in Section 2.1. It is well-known, that the system group orbits of a discrete-time system are immersed submanifolds, provided the system is smoothly invertible (see [JS90]). This fact is a discrete-time version of the well-known *orbit theorem*. We show that system semigroup orbits, i.e., the reachable sets, are not submanifolds in general. Moreover, we show that Mokkadem's *algebraic version of the orbit theorem* (see Theorem 3 in [Mok95]) is wrong and prove a correct version, under the additional assumption, that the system group orbit is semi-algebraic (Theorem 2.7). All systems which appear in Part II share a property, which we termed *right divisibility*. To our knowledge, the concept of right divisible systems is new. In Section 2.1.3 we show some examples and basic properties for such systems. In particular, we prove an equivalent condition for right divisibility which is easier to verify (Theorem 2.15).

The concept of *accessibility* is the topic of Section 2.2. We introduce techniques for checking whether a discrete-time system is accessible or not. First, we briefly recall geometric conditions for accessibility developed by Jakubczyk and Sontag ([JS90]) and then prove an accessibility result for systems where the system group is a Lie group (Theorem 2.23). This result is based on elementary facts on semigroup actions on manifolds, which can be found in [Mit01]. In some cases, accessibility from one point already implies accessibility on the corresponding orbit. This phenomenon is called *Chow property.* In Section 2.2.2 we recall sufficient conditions for Chow property given by Albertini and Sontag ([AS93, AS94]). We prove that any invertible system, where the system group is a Lie group acting continuously on the state space, has the Chow property, provided the corresponding orbit is locally compact (Theorem 2.28).

Section 2.3 deals with the concept of *controllability* and the related notion of *weak reversibility.* We easily see that a system is *weakly reversible* if and only if the system group orbits coincide with the corresponding reachable sets. As a consequence we obtain a condition for controllability analogous to a well-known result of the continuous-time theory (see [Son98]). Afterwards, we list some types of systems, where reachability from one point already implies controllability. This phenomenon is well known for linear systems. We show similar results for abelian systems, weak reversible systems and systems where the system semigroup is "large enough" in a certain topological sense (Theorems 2.39-2.41).

We finish Chapter 2 with some results on *approximatively reachable systems* and *densely reachable systems.* Here we focus on the abelian case. We show that approximatively reachable systems have the property, that for every $y$ in the topological closure of the reachable set of $x$, there exists a control sequence such that the corresponding sequence converges to $y$ (Theorem 2.46). Moreover, we show that – unlike abelian systems which are reachable from one point – abelian systems which are approximatively reachable from every point, do not necessarily have the property that the system semigroup is a group. *Dense reachability* is the property, that a system is approximatively reachable from "almost every" initial state. We show that accessibility from some point together with approximative reachability from one point implies dense reachability (Theorem 2.48).

In Chapter 3 we analyze the relationship between the properties of a given system on state space $M$ and the properties of certain types of related systems, namely *induced systems* and *restricted systems.* Our results are not surprising and probably not entirely unknown. However, to the best of the authors knowledge there exists no systematic investigation for the analysis of *induced systems* or *restricted systems* in terms of system semigroups. Given two systems $\Sigma, \tilde{\Sigma}$ with the same set of control parameters $U$, with state spaces $M$, and respectively, $\tilde{M}$ and with transition maps $f : M \times U \to M$, and respectively, $\tilde{f} : \tilde{M} \times U \to \tilde{M}$, then $\tilde{\Sigma}$ is said to be an *induced system* of $\Sigma$ with respect to $\pi : M \to \tilde{M}$ if $\pi$ is open, continuous and surjective, and $\pi \circ f(\cdot, u) = \tilde{f}(\cdot, u) \circ \pi$ for all $u \in U$. We compare the corresponding system semigroups of the original system and the induced system. Our results imply, that all basic controllability properties of $\Sigma$, such as weak reversibility or dense reachability, are preserved on $\tilde{\Sigma}$ (Theorem 3.4). In Section 3.2 we analyze *restricted systems*, i.e, subsystems restricted to system invariant subsets. We express the system semigroup of the restricted system as a factor semigroup of the system semigroup of the original system (Theorem 3.12). For abelian systems it follows, that controllability of a restricted system on $N \subseteq M$ implies controllability of all systems restricted on orbits in the boundary of $N$ (Theorem 3.13).

In Chapter 4 we discuss the question, how the adherence structure of reachable sets provides limitations for the existence of convergent shift strategies. For that purpose we develop a graph theoretical language which allows us to express the adherence structure of system group orbits and reachable sets graphically. Obviously, a point can not be reached from $x$, if it is outside of the topological closure of the system group orbit of $x$. For that reason we analyze systems restricted on orbits (in Section 4.2) as well as systems restricted on the topological closure of orbits (in Section 4.3). We show that there always exists a sequence of reachable sets such that its union is dense in the orbit, provided the system is right divisible and the orbit is locally compact (Theorem 4.10). Moreover, we prove some conditions for the appearance of repelling phenomena for right divisible systems and abelian systems (Theorems 4.17-4.18).

We finish Part I with the analysis of certain families of systems on Lie groups (Section 5.1) and on homogeneous spaces (Section 5.2). As expected, for systems on Lie groups, we obtain similar results as in the well-known theory on left invariant continuous-time systems by Sussmann and Jurdjevic [JS72, SJ72]. In particular, we show that accessible systems evolving on connected Lie groups are densely reachable if and only if they are controllable (Theorem 5.4). Systems on homogeneous spaces can be regarded as induced systems of a system on a Lie group. Thus, the controllability properties of systems on homogeneous spaces $\tilde{\Sigma}$ are linked to the controllability properties of a certain corresponding system on a Lie group $\Sigma$. We show a condition for weak reversibility of $\tilde{\Sigma}$ in terms of the system semigroup of $\Sigma$ (Theorem 5.8).

In the second part of this thesis we explore the structure of reachable sets of inverse iteration systems, rational iteration systems, Richardson's iteration systems and linear iteration systems.

We start with an investigation of classical inverse iteration systems (1) for cyclic matrices (Chapter 6). First, we analyze the corresponding system group. We show that the system group is an abelian Lie group which acts on the projective space $\mathbb{RP}^{n-1}$ (Theorem 6.3). The isomorphism type depends on the Jordan canonical form of the system matrix $A$. In Section 6.2 we classify all possible isomorphism types in terms of the minimal polynomial of $A$. In the next section we analyze the structure of the system group orbits. We show a one-to-one relation between the adherence structure of the orbits and the lattice structure of the $A$-invariant subspaces (Theorem 6.14). Moreover, we show that there exists one orbit which is open and dense in $\mathbb{RP}^{n-1}$ (Theorem 6.15). In Section 6.4 we focus on the system restricted to the open and dense orbit. In [HW01] it is shown, that the restricted system is only for a certain set of matrices controllable. We extend their results in different aspects. In particular, we show that the restricted system is controllable if and only if the matrix semigroup $S(A)\mathbb{R}^* := \{r\prod_{t=1}^{N}(A - u_t I) \,|\, N \in \mathbb{N}, r \in \mathbb{R}^*, u_t \in \mathbb{R} \setminus \mathrm{Spec}(A)\}$ is equal to the centralizer group $P(A)$ of $A$ (Theorem 6.18). Necessary and sufficient conditions for $S(A)\mathbb{R}^* = P(A)$ are derived in Section 6.5 and Section 6.6. One interesting byproduct is an interpolation result for linear decomposable polynomials (Theorem 6.32). If the restricted system is controllable, the adherence structure of reachable sets is coincides with the adherence structure of the system group orbits. In Section 6.7 we analyze the adherence structure of reachable sets for the cases when the restricted system is not controllable. In particular, we give conditions for the appearance of repelling phenomena (Theorem 6.34). We finish Chapter 6 with a systematic controllability analysis for the cases $n = 2, 3, 4$.

In Chapter 7 we consider generalized inverse iteration systems, i.e., inverse iteration schemes which act on manifolds other than the projected space. In particular we are interested in the cases when the manifold is a complete flag manifold (Section 7.1), a Hessenberg variety (Section 7.2), or a vector space (Section 7.3). In the first case there exist infinitely many system group orbits and all of them have empty interior. This fact was already pointed out by Helmke and Jordan in [HJ02]. We show that the reachable graph and the orbit graph are equivalent if and only if $S(A)\mathbb{R}^* = P(A)$ (Theorem 7.3). The analysis of inverse iteration on Hessenberg varieties is closely related to the QR algorithm on Hessenberg matrices (see [AM86]). We show that there exists a dense reachable set if and only if $S(A)\mathbb{R}^* = P(A)$ (Theorem 7.8). We finish Section 6.8 with an analysis of inverse iteration on $\mathbb{R}^n$. Again, there exists a system group orbit which is open and dense in $\mathbb{R}^n$, provided $A$ is cyclic. We show that the system restricted to this orbit is not controllable for an open and dense set of matrices (Theorem 7.9). Moreover, we present a complete analysis for the case $n = 2$.

In Chapter 8 we explore rational iteration systems (2). Here, the system semigroup is naturally a group isomorphic to the system group of the corresponding generalized inverse iteration system. Thus, the structure of the system group orbits is identically with the structures analyzed in Chapter 6. As a special case we consider Cayley iteration systems. We show an open set of matrices for what the reachable sets of rational iteration and Cayley iteration coincide (Theorem 8.5). In contrast we construct families of matrices, where the reachable sets of Cayley iteration systems (3) are smaller then the reachable sets of inverse iteration systems (Theorem 8.7). We finish Chapter 8 with a complete analysis of Cayley iteration systems in the plane.

In Chapter 9 we explore the reachable sets of Richardson's method (4) and, more generally, polynomial iteration schemes (5) of degree $m$. Here, the system group coincides with $P(A)$. It follows that, if the system semigroup is a group, the solution of $Ax = b$ lies in the topological closure of the reachable set of almost all initial states. We show that the system semigroup is a group if $m > 1$ (Theorem 9.11). However, the situation differs critically for the special case $m = 1$, i.e. for Richardson's systems. On the one hand there exists an open set of matrices, where the system semigroup is a group. For example, this is the case if $A$ has $n$ different real eigenvalues (Theorem 9.6). On the other hand, we construct a family of cyclic matrices where the system semigroup is not a group. In this cases the solution of $Ax = b$ is repelling to a generic subset of $\mathbb{R}^n$ (Theorem 9.7).

In Chapter 10 we investigate linear control schemes (6). Here, the system semigroup is right divisible but not abelian (Theorem 10.2). Moreover, the adherence structure of reachable sets differs fundamentally to the adherence structure of Richardson's systems and polynomial iteration systems. It is well known that generically, linear control systems are controllable. We analyze the adherence structure of reachable sets of the uncontrollable cases. In contrast to Richardson's systems, none of the reachable sets has open interior (Theorem 10.3). However, we show that there exists uncontrollable cases where the topological closure of any reachable set contains the solution of $Ax = b$ (Theorem 10.10). A suitable shift strategy, such that the arising sequence converges to an solution of $Ax = b$, is given by a linear feedback law. The corresponding algorithm (LQRES) is the topic of Section 10.2. LQRES is globally convergent for a generic set of pairs $(A, B)$ (Theorem 10.8). For the special case $B = 0$, LQRES coincides with Richardson's iteration for the constant shift strategy $u \equiv 1$. We show that in some choices of $B$, LQRES converges where Richardson's method fails for all possible shift strategies (Example 10.9). We finish Chapter 10 with some numerical experiments, which point out the influence of the choice of $B$ on the convergence behavior (Examples 10.12-10.13)

## Acknowledgement

I wish to extend my thanks to everybody who contributed to the development of this thesis, be it with advice, guidance and criticism, with discussions, suggestions and corrections or with encouragement, relaxation and patience.

First, I would like to thank my advisor Prof. Uwe Helmke for the chance to work on these interesting problems, for many helpful ideas and suggestions and for valuable insights in particular into the field of geometric control theory. Moreover, I would like to express my gratitude to those, who introduced me to the other mathematical disciplines which are contained in this thesis. In particular I would like to thank Prof. Rodolphe Sepulchre for introducing me to dynamical system theory and and for hosting me at the Université de Liège. I thank Linus Kramer and the Arbeitsgruppe AGF of the Darmstadt University of Technology for fruitful discussions and suggestions on the theory of semigroups. Many thanks also go to Prof. Heinrich Voss for showing me the beauty of numerical analysis and to Prof. Fabian Wirth, in particular for advice in the field of discrete-time control theory.

I wish to thank my colleagues in Würzburg, Darmstadt and Liège who created a pleasant and family like working atmosphere. Especially I want to mention Ingrid Böhm, Oana Curtef, Sarah Drewes, Gunther Dirr, Christophe Germay, Thomas Gregor, Sven Herzberg, Jose Ignacio Iglesias Curto, Annett Keller, Martin Kleinsteuber, Daniela Kraus, Katja Kulas, Indra Kurniawan, Christoph Müller, Oliver Roth, Nils Rosehr, Martin Schröter, Katrin Schumacher, Jörn Steuding, Otto Volk and Wolfgang Weigel for many interesting discussions, not only in cafe breaks and not only about mathematics.

Finally, I thank all my friends, my family and Ina for many things, but in particular for their support and (re)-encouragement throughout the time of my work on this thesis.

This work has been supported by the German Research Foundation Grant DFG HE 1858/10-1 "KONNEW".

# Part I

# Analysis of reachable sets

# 2 Discrete-time control systems

In this chapter we clarify definitions and notations which will be used throughout this manuscript. Moreover, we present some basic observations on discrete-time control systems. We begin with some results on system group orbits in Section 2.1. Then, we introduce the concepts of *accessibility* (Section 2.2), *controllability* (Section 2.3) and *reachability* (Section 2.4).

Iterative algorithms with shift parameters can be regarded as discrete-time control systems. The basic idea is to express every iteration step by a map $f_u := f(\cdot, u)$ which can be manipulated by a shift parameter $u$. This leads to the following definition which is fundamental in this work.

**Definition 2.1 (Discrete-time control systems)** A *discrete-time control system* – or for short a *system* – is a triple $\Sigma = (M, U, f)$ where

- $M$ is a topological space (the *state space*)
- $U$ is a subset of $\mathbb{R}^m$ (the set of *control parameters*)
- $f : M \times U \to M$ is a continuous map (the *transition map*)

A system $\Sigma$ is called

- *abelian* if $f_u \circ f_v = f_v \circ f_u$ for all $u, v \in U$
- *invertible* if $f_u : M \to M$, $x \mapsto f(x, u)$ is a homeomorphism for all fixed $u \in U$
- *smoothly invertible*, if $M$ is a smooth manifold[1] and $f_u : M \to M$ is a diffeomorphism for any $u \in U$.
- *algebraically invertible*, if it is invertible, $M$ is a variety, $U$ is a semi-algebraic set and $f : M \times U \to M$ is a semi-algebraic map [2].

Motivated by the applications on numerical iteration schemes, we focus on invertible systems[3] which are either smoothly invertible, algebraically invertible or both.

[1] As a standard assumption for this thesis, a manifold is always assumed to be smooth and of finite dimension.

[2] In Appendix A we present the definitions and basic properties of *varieties*, *semi-algebraic sets* and *semi-algebraic maps*. Note that $f_u$ and $f_u^{-1}$ are semi-algebraic if $\Sigma$ is algebraically invertible (see Proposition A.1)

[3] see [SW98, Wir98] for the theory of non invertible systems

A discrete-time control system $\Sigma$ describes an iterative method with parameters $u_t \in U$, i.e.,

$$x_{t+1} := f(x_t, u_t), \qquad x_0 \in M, \tag{7}$$

with $t \in \mathbb{N}_0$. In numerical linear algebra, such control parameters or input variables $u_t$ are often called *shifts* and a specific choice of such shifts is called a *shift strategy*. Formally, we define a shift strategy $u$ to be a finite or infinite sequence of control parameters, i.e., $u_0, \dots, u_{T-1} \in U^T$ respectively $u_0, u_1 \cdots \in U^{\mathbb{N}}$. We say *$y$ can be reached from $x$* if there exists $T \in \mathbb{N}$ and $u = (u_0, \dots, u_{T-1}) \in U^T$ such that *$u$ steers $x$ to $y$*, i.e., the recursion $x_{t+1} = f(x_t, u_t)$, $x_0 := x$ yields $x_T = y$. Given a nonempty subset $\mathcal{E} \subseteq M$ (respectively a point $y \in M$) we say that *$x$ converges to $\mathcal{E}$ (respectively to $y$) with respect to $u \in U^{\mathbb{N}_0}$* if the sequence given by the recursion $x_{t+1} = f(x_t, u_t)$, $x_0 = x$ converges to $\mathcal{E}$ (respectively to $y$), i.e., every open subset $\mathcal{V}$ of $M$ such that $\mathcal{E} \subseteq \mathcal{V}$, contains all but finitely many elements of the sequence $(x_t)_{t\in\mathbb{N}}$. We write $x \xrightarrow{u} \mathcal{E}$ (respectively $x \xrightarrow{u} y$). In applications one wants to find an automatic way to obtain suitable shift strategies. If a shift strategy is given by a map $\Phi : M \to U$, $u_t = \Phi(x_t)$, we call $\Phi$ a *feedback law*.

## 2.1 Reachable sets via semigroup orbits

The basic topic of this thesis is the investigation of reachable sets and their adherence structure. They can be described in terms of so-called system semigroups. In the following we will give some definitions and basic properties which are essential in the analysis of abstract discrete-time systems in general, as well as in the analysis of the structure of reachable sets of iterative algorithms.

We will use the following notation. For $T \in \mathbb{N}$ we define $f_T : M \times U^T \to M$ by

$$f_T : (x, u_0, \dots u_{T-1}) \mapsto f_{u_{T-1}} \circ \cdots \circ f_{u_0}(x) \tag{8}$$

with $f_u : M \to M$ given by $f_u := f(\cdot, u)$. In other words, $f_T$ maps an initial point $x$ to the output after $T$ iteration steps with shift parameters $u_0, \dots, u_{T-1}$. For the following definition we stick to the notation in [Son98]. It is analogous to the well known concept of reachable sets in the continuous-time case.

**Definition 2.2 (Reachable sets)** The *reachable set* $\mathcal{R}(x)$ of a point $x$ is the set of all states which can be reached from $x$ in finitely many iterations, using arbitrary controls in each step, i.e.,

$$\mathcal{R}(x) := \{y \in M \mid \exists T \in \mathbb{N}, \exists u \in U^T : y = f_T(x, u)\}. \tag{9}$$

In other words

$$\mathcal{R}(x) = \bigcup_{T=1}^{\infty} \mathcal{R}^T(x) \tag{10}$$

where $\mathcal{R}^T(x)$ is the set of points which can be reached in $T \in \mathbb{N}$ steps, i.e., $\mathcal{R}^T(x) := \{f_T(x,u) \,|\, u \in U^T\}$. We call $x \in M$ a *fixed point* of $\Sigma$ if $\mathcal{R}(x) = \{x\}$.

In this thesis we will extensively use the fact that reachable sets can be interpreted as orbits of certain semigroup actions.

**Definition 2.3 (System semigroup)** The *system semigroup* $S_\Sigma$ of a system $\Sigma = (M, U, f)$ is given by

$$S_\Sigma := \left\{ s : M \to M \,|\, \exists\, T \in \mathbb{N}, \exists\, u \in U^T : s = f_T(\cdot, u) \right\}. \tag{11}$$

Obviously, $S_\Sigma$ is a semigroup with respect to composition of maps, i.e.,

$$s_1 s_2 : x \mapsto s_1(s_2(x)).$$

Note that every element of $S_\Sigma$ is a continuous map $s : M \to M$. It is easy to see, that $\Sigma$ is abelian if and only if $S_\Sigma$ is abelian. Moreover, if $\Sigma$ is invertible, $s \cdot x = s \cdot y$ implies $x = y$ and $s_1 s_2 = \mathrm{id}_M$ implies $s_2 s_1 = \mathrm{id}_M$.

Canonically, the system semigroup acts on the state space via the mapping

$$S_\Sigma \times M \to M, \quad (s, x) \to s \cdot x := s(x). \tag{12}$$

In other words, the reachable set of a discrete-time control system is the orbit of the semigroup action (12), i.e.,

$$\mathcal{R}(x) = \{s(x) \,|\, s \in S_\Sigma\} := S_\Sigma \cdot x. \tag{13}$$

Due to this fact, reachable sets are also called *forward orbits*.

The system semigroup is not a group in general. In particular, $S_\Sigma$ does not always contain the identity homeomorphism $\mathrm{id}_M$. Therefore, we can neither expect that $x$ lies in $\mathcal{R}(x)$ nor that $y \in \mathcal{R}(x)$ implies $x \in \mathcal{R}(y)$. Nevertheless, if the system is invertible, which will be the standard case in this work, the system semigroup generates a group in a canonical way.

**Definition 2.4 (System group)** Let $\Sigma = (M, U, f)$ be an invertible system and $S_\Sigma$ its system semigroup. We call the group

$$G_\Sigma := \langle S_\Sigma \rangle := \left\{ g_N \circ \cdots \circ g_1 \mid N \in \mathbb{N}, g_t \in S_\Sigma \text{ or } g_t^{-1} \in S_\Sigma \right\}$$

the *system group* of $\Sigma$.

Note that $G_\Sigma$ is the smallest group such that $S_\Sigma$ is a subsemigroup of $G_\Sigma$. Every $g \in G_\Sigma$ is a finite composition of continuous maps $g_i \in S_\Sigma \cup S_\Sigma^{-1}$ and therefore continuous. Here $S_\Sigma^{-1} := \{s^{-1} \mid s \in S_\Sigma\}$. It also follows, that $G_\Sigma$ is abelian if and only if $S_\Sigma$ is abelian. The orbits of the group action $G_\Sigma \times M \to M$, $(g, m) \mapsto g(m)$ contain important informations about the structure of the reachable sets due to the trivial but significant observation that

$$\mathcal{R}(x) \subseteq G_\Sigma \cdot x := \{g(x) \mid g \in G_\Sigma\} \tag{14}$$

for all $x \in M$. Nevertheless, in many applications, such as inverse iteration systems (see Section 6), $S_\Sigma$ is a proper subsemigroup of $G_\Sigma$.

### 2.1.1 Orbit theorems

The system group orbits of a system $\Sigma$ are usually better understood than the reachable sets. First of all they form a partition on the state space. Moreover, they have a natural structure of immersed submanifolds in the state space, provided $\Sigma$ is smoothly invertible. This fact is a discrete-time version of the well-known *orbit theorem* of continuous time systems (see Theorem 1, Chapter 2 in [Jur97]).

**Theorem 2.5 (Orbit theorem)** *Let $\Sigma$ be a smoothly invertible system with $U$ open in $\mathbb{R}^m$ and $f : M \times U \to M$ smooth. Then any orbit $G_\Sigma \cdot x$ is an immersed submanifold of $M$ with at most countably many components.*

In other words, $G_\Sigma \cdot x$ can be equipped with a manifold structure, such that the inclusion map $\mathrm{inc} : G_\Sigma \cdot x \to M$ is an immersion. See Theorem 7 in [JS90] and Proposition 8.9 in [Son86] respectively for more details and a proof. Recall that an immersed submanifold is not necessarily a submanifold in the common sense, i.e., the inclusion map is not necessarily an embedding. Now we give an easy example for this phenomenon.

**Example 2.6** Consider $\Sigma = (\mathbb{R}^2, \mathbb{R}, f)$ with

$$f_u : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} \cos \alpha\pi & -\sin \alpha\pi \\ +\sin \alpha\pi & \cos \alpha\pi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Then

$$\begin{aligned} G_\Sigma \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \left\{ \begin{pmatrix} \cos \alpha\pi & -\sin \alpha\pi \\ \sin \alpha\pi & \cos \alpha\pi \end{pmatrix}^z \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \middle| \, z \in \mathbb{Z} \right\} \\ &= \left\{ \begin{pmatrix} \cos z\alpha\pi & -\sin z\alpha\pi \\ \sin z\alpha\pi & \cos z\alpha\pi \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \middle| \, z \in \mathbb{Z} \right\}, \end{aligned}$$

which is a countable dense subset of

$$\|x\|^2 \mathbb{S} := \left\{ (y_1, y_2) \in \mathbb{R}^2 \mid y_1^2 + y_2^2 = \|x\|^2 \right\}$$

and therefore not a submanifold of the state space.

Mokkadem proposes an algebraic version of the orbit theorem (Theorem 3 in [Mok95]). In particular, he claims that $G_\Sigma \cdot x$ is an embedded smooth subvariety, provided $M$ is a smooth variety and $f_u$ is a bijective regular morphism. Note that Example 2.6 is a counterexample to this claim. Nevertheless, assuming that $G_\Sigma \cdot x$ is semi-algebraic[4], we obtain the following version of the orbit theorem.

**Theorem 2.7 (Algebraic orbit theorem)** *Let $\Sigma = (M, U, f)$ be smoothly invertible such that $M$ is a variety in $\mathbb{R}^n$. If $G_\Sigma \cdot x$ is semi-algebraic, then $G_\Sigma \cdot x$ is an embedded smooth submanifold of $M$.*

**Proof.** If $G_\Sigma \cdot x$ is semi-algebraic, it can be written as a finite union of disjoint submanifolds $A_i$, $1 = 1, \dots, l$, such that each $A_i$ is diffeomorphic to $(0,1)^{d_i}$ and that $\dim(G_\Sigma \cdot x) := d := \max\{d_1, \dots, d_l\}$ is uniquely determined (see Theorem A.4).

Moreover, there exists $y \in G_\Sigma \cdot x$ and an open set $U_y \subseteq M$, such that $y \in U_y \cap G_\Sigma \cdot x$ and $U_y \cap G_\Sigma \cdot x$ is diffeomorphic to $(0,1)^d$ (see Lemma A.6).

For all $z \in G_\Sigma \cdot x$ there exists $g \in G_\Sigma$ with $z = g(y)$. Therefore,

$$z \in g(U_y \cap G_\Sigma \cdot x).$$

Since $g$ is bijective and $g(G_\Sigma \cdot x) = gG_\Sigma \cdot x = G_\Sigma \cdot x$, we obtain $g(U_y \cap G_\Sigma \cdot x) = g(U_y) \cap G_\Sigma \cdot x$. Moreover, $g(U_y)$ is open and $g(U_y \cap G_\Sigma \cdot x)$ is diffeomorphic to $(0,1)^d$ since $g : M \to M$ is a diffeomorphism. Hence, $G_\Sigma \cdot x$ is a submanifold of $M$ of dimension $d$. □

In many applications, the system group $G_\Sigma$ carries a canonical Lie group structure. Here, the literature on Lie group actions provides different sufficient conditions for submanifold structure of $G_\Sigma \cdot x$.

**Theorem 2.8** *Let $\Sigma = (M, U, f)$ be a smoothly invertible system. Assume that $G_\Sigma$ carries a Lie group structure such that the group action $\alpha : G_\Sigma \times M \to M$, $(g, x) \mapsto g(x)$ is smooth. Then*

*a) If $G_\Sigma$ is compact then every orbit $G_\Sigma \cdot x$ is a submanifold of $M$.*

*b) If $G_\Sigma$ a semi-algebraic set such that $\alpha$ is semi-algebraic, then every orbit $G_\Sigma \cdot x$ is a submanifold of $M$.*

**Proof.** Statement a) can be found in [GOV97], Theorem 2.3 and statement b) can be found in [HM94], page 353. Moreover, Statement b) is also a consequence of Theorem 2.7, since $G_\Sigma \cdot x$ is the image of the semi-algebraic map $\alpha_x : G_\Sigma \to M$, $g \mapsto g \cdot x$ and therefore semi-algebraic (see Proposition A.1 and Corollary A.3). □

[4] we will show conditions on $\Sigma$ for which $G_\Sigma \cdot x$ is semi-algebraic in Section 2.1.2

In contrast to the system group orbits, system semigroup orbits (the reachable sets) are not necessarily immersed submanifolds of the state space, even if $\Sigma$ is smoothly invertible. An easy example is given by $\Sigma = (\mathbb{R}, \mathbb{R}, f)$ with $f(x,u) = x + u^2$. Here $\mathcal{R}(0) = [0, \infty)$. Another example, which additionally shows that the reachable sets might have locally different dimensions, is the following:

**Example 2.9** Consider $\Sigma = (M, U, f)$ with $M = \mathbb{R}^2$, $U = \mathbb{R}$ and

$$f_u : \mathbb{R}^2 \to \mathbb{R}^2, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} -ux_1 - cx_2 \\ cx_1 - ux_2 \end{pmatrix}.$$

Here $c$ is a real constant with $|c| > 1$. We show that the reachable set of $x = (1,0)^T$ is not an immersed submanifold of $M$ (see Figure 1). Obviously,

$$\mathcal{R}^1(x) = \{f_u(x) |\ u \in \mathbb{R}\} = \left\{(-u, c)^T \,|\, u \in \mathbb{R}\right\}$$

is a one dimensional submanifold of $M$. Moreover, $\mathcal{R}^1(x)$ and the disk

$$C := \left\{ y = (y_1, y_2) \in \mathbb{R}^2 \;\middle|\; \|y\| < c^2 \right\},$$

have nonempty intersection, since $|c| > 1$. On the other hand

$$\mathcal{R}(x) \setminus \mathcal{R}^1(x) = \bigcup_{T=2}^{\infty} \mathcal{R}^T(x)$$

lies outside $C$ as can be shown by induction on $T$. For all $y \in \mathcal{R}^2(x)$ we obtain

$$\begin{aligned} \|y\|_2 &= \|f_{u_0} \circ f_{u_1}(x)\|_2 \\ &= \left\| \begin{pmatrix} u_0 u_1 - c^2 \\ -u_0 c - u_1 c \end{pmatrix} \right\|_2 \\ &= \sqrt{(u_0u_1)^2 + c^4 + (cu_0)^2 + (u_1 c)^2} \\ &\geq c^2. \end{aligned}$$

Now for $T \geq 2$ we assume that $\|y\| \geq c^2$ for all $y \in \mathcal{R}^T(x)$. Recall that

$$\mathcal{R}^{T+1}(x) = \left\{ f_{u_T}\left(\mathcal{R}^T(x)\right) \,|\, u_T \in \mathbb{R} \right\}.$$

In other words, every $z \in \mathcal{R}^{T+1}(x)$ can be written as $z = f_u(y)$ with $y \in \mathcal{R}^T(x)$ and $u \in U$. We obtain

$$\|z\|_2 = \|f_u(y)\|_2 = \sqrt{(u^2 + c^2)(y_1^2 + y_2^2)} \geq |c| \cdot \|y\|_2$$

and therefore $\|z\|_2 \geq c^2$. Hence, $\|z\|_2 \geq c^2$ for all $y \in \mathcal{R}^T(x)$ with $T \geq 2$.

Under the assumption that $\mathcal{R}(x)$ is a manifold, we obtain

$$\dim \mathcal{R}(x) = 1, \tag{15}$$

since $\mathcal{R}^1(x) \cap B_\epsilon = \mathcal{R}(x) \cap B_\epsilon$ for an open ball

$$B_\epsilon := \{y \in \mathbb{R}^2 \,|\, \|y - (0,c)^T\|_2 < \epsilon\}$$

with $\epsilon > 0$ small enough. On the other hand $\mathcal{R}^2(x) = \{f_{u_0} \circ f_{u_1}(x) \,|\, u_0, u_1 \in \mathbb{R}\}$ has open interior, since the Jacobian of the map

$$\mathbb{R}^2 \longrightarrow \mathbb{R}^2, (u_0, u_1) \mapsto f_{u_0} \circ f_{u_1}((1,0)^T)$$

is

$$D = \begin{pmatrix} u_0 & u_1 \\ -c & -c \end{pmatrix}$$

and therefore regular for $u_0 \neq u_1$. Hence, if $\mathcal{R}(x)$ is a manifold, it must have dimension 2 which is a contradiction to (15).

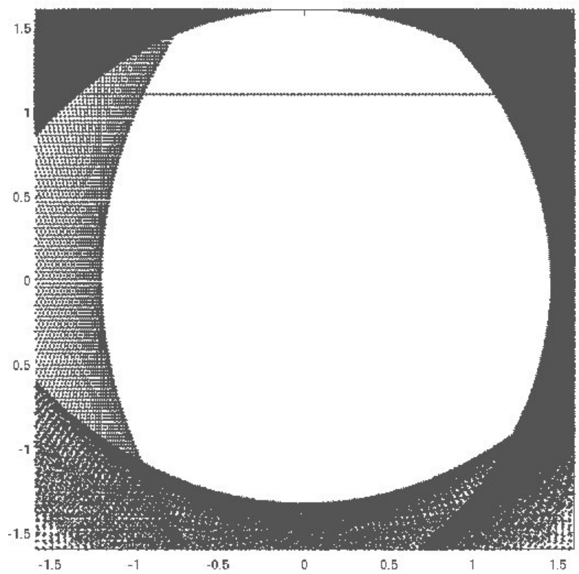



**Figure 1:** *A plot of* $(\mathcal{R}^1(x) \cup \mathcal{R}^2(x) \cup \mathcal{R}^3(x) \cup \mathcal{R}^4(x)) \cap [-1.5, 1.5] \times [-1.5, 1.5]$ *for* $c = 1.1$ *and* $x = (1,0)^T$*. Any point of* $\mathcal{R}^k(x)$ *with* $k > 5$ *is outside of the square* $[-1.5, 1.5] \times [-1.5, 1.5]$*. We see, that* $\mathcal{R}(x)$ *is not a manifold, since the one dimensional line* $\mathcal{R}^1(x)$ *is isolated of* $\mathcal{R}(x) \setminus \mathcal{R}^1(x)$ *close enough to* $(0, 1.1)^T$*. Moreover, the boundary of* $\mathbb{R}^2 \setminus \mathcal{R}(x)$ *is nonsmooth.*

### 2.1.2 Semi-algebraic orbits

If $\Sigma = (M, U, f)$ is algebraically invertible, then for all $T \in \mathbb{N}$ and all $x \in M$ the set $\mathcal{R}^T(x)$ is semi-algebraic, since it is the image of the semi-algebraic set $U^T$ and the semi-algebraic map $(u_0, \dots, u_{T-1}) \mapsto f_T(x, u_0, \dots, u_{T-1})$. Nevertheless, the reachable set $\mathcal{R}(x) = \bigcup_{t=1}^{\infty} \mathcal{R}^t(x)$ or the corresponding system group orbit $G_\Sigma \cdot x$ is not semi-algebraic in general. An easy example is given by $\Sigma = (\mathbb{R}, \mathbb{R}, f)$ with $f(x, u) = x + 1$. Here,

$$\mathcal{R}(x) = \{x + n \,|\, n \in \mathbb{N}\} \quad \text{and} \quad G_\Sigma \cdot x = \{x + z \,|\, n \in \mathbb{Z}\}.$$

In the following we show some sufficient conditions which provide that the reachable sets $\mathcal{R}(x)$ and the system group orbits $G_\Sigma \cdot x$ of an algebraically invertible system are semi-algebraic.

Similarly to the construction of the reachable sets, we define

$$\mathcal{O}^T(x) := \left\{ f_{u_T}^{\epsilon_T} \circ \dots \circ f_{u_1}^{\epsilon_1}(x) \,|\, u_t \in U, \epsilon_t \in \{-1, 1\} \right\}, \quad T \in \mathbb{N}$$

for $x \in M$ and $T \in \mathbb{N}$. Note that $G_\Sigma \cdot x := \bigcup_{t=1}^{\infty} \mathcal{O}^t(x)$. Moreover, we obtain the following lemma:

**Lemma 2.10** *Let $\Sigma = (M, U, f)$ be an invertible system and $T \in \mathbb{N}$. Then*

*a) $\mathcal{R}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^t(x)$ if and only if $\mathcal{R}(x) = \bigcup_{t=1}^{T} \mathcal{R}^t(x)$.*

*b) $\mathcal{O}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{O}^t(x)$ if and only if $G_\Sigma \cdot x = \bigcup_{t=1}^{T} \mathcal{O}^t(x)$.*

*c) If $\Sigma$ is abelian and $\mathcal{R}(x) = \bigcup_{t=1}^{T} \mathcal{R}^t(x)$, then $\mathcal{R}(y) = \bigcup_{t=1}^{T} \mathcal{R}^t(y)$ for all $y \in G_\Sigma \cdot x$.*

*d) If $\Sigma$ is abelian and $\mathcal{R}(x) = \bigcup_{t=1}^{T} \mathcal{R}^t(x)$, then $G_\Sigma \cdot x = \bigcup_{t=1}^{2T} \mathcal{O}^t(x)$.*

**Proof.** a) Obviously, $\mathcal{R}(x) = \bigcup_{t=1}^{T} \mathcal{R}^t(x)$ implies $\mathcal{R}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^t(x)$. Now we assume $\mathcal{R}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^t(x)$. Then

$$\mathcal{R}^{T+2}(x) = \bigcup_{y \in \mathcal{R}^{T+1}(x)} \mathcal{R}^1(y) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^{t+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^t(x),$$

since $y \in \mathcal{R}^{T+1}(x)$ implies $y \in \mathcal{R}^t(x)$ for some $1 \leq t \leq T$ and therefore $\mathcal{R}^1(y) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^{t+1}(x)$. Hence, $\mathcal{R}(x) = \bigcup_{t=1}^{\infty} \mathcal{R}^t(x) = \bigcup_{t=1}^{T} \mathcal{R}^t(x)$.
b) Analogous to a), $G_\Sigma \cdot x = \bigcup_{t=1}^{T} \mathcal{O}^t(x)$ implies $\mathcal{O}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{O}^t(x)$. Moreover, $\mathcal{O}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{O}^t(x)$ implies

$$\mathcal{O}^{T+2}(x) = \bigcup_{y \in \mathcal{O}^{T+1}(x)} \mathcal{O}^1(y) \subseteq \bigcup_{t=1}^{T} \mathcal{O}^{t+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{O}^t(x),$$

and therefore $G_\Sigma \cdot x = \bigcup_{t=1}^{T} \mathcal{O}^t(x)$.
c) Let $y \in G_\Sigma \cdot x$, i.e., $y = g \cdot x = g(x)$ for some $g \in S_\Sigma$. Then, for any $s \in S_\Sigma$

$$s(y) = g \circ s(x) = g \circ f_{u_t} \circ \cdots \circ f_{u_1}(x)$$

for $1 \le t \le T$. Therefore, $s(y) = f_{u_t} \circ \cdots \circ f_{u_1}(y) \in \mathcal{R}^t(y)$. We conclude $\mathcal{R}(y) = S_\Sigma \cdot y = \bigcup_{t=1}^{T} \mathcal{R}^t(y)$.
d) For any $y \in G_\Sigma \cdot x$,

$$y = f_{u_{t_1}}^{-1} \circ \cdots \circ f_{u_1}^{-1} \circ f_{v_{t_2}} \circ \cdots \circ f_{v_1}(x).$$

Since $\mathcal{R}(x) = \bigcup_{t=1}^{T} \mathcal{R}^t(x)$, we can replace $f_{v_{t_2}}, \dots, f_{v_1}$ by a possibly shorter sequence $f_{\tilde{v}_{\tilde{t}_2}}, \dots, f_{\tilde{v}_1}$ such that $\tilde{t}_2 < T$. Moreover, $f_{u_{t_1}} \circ \cdots \circ f_{u_1}(y) = z \in S_\Sigma \cdot y$ with $z := f_{\tilde{v}_{\tilde{t}_2}} \circ \cdots \circ f_{\tilde{v}_1}(x)$. By c) we can replace $f_{u_{t_1}}, \dots, f_{u_1}$ by a shorter sequence $f_{\tilde{u}_{\tilde{t}_1}}, \dots, f_{\tilde{u}_1}$ with $\tilde{t}_1 \le T$. Hence,

$$y = f_{\tilde{u}_{\tilde{t}_1}}^{-1} \circ \cdots \circ f_{\tilde{u}_1}^{-1} \circ f_{\tilde{v}_{\tilde{t}_2}} \circ \cdots \circ f_{\tilde{v}_1}(x) \in \mathcal{O}^{\tilde{t}_1+\tilde{t}_2}(x)$$

with $\tilde{t}_1 + \tilde{t}_2 \le 2T$, and therefore $G_\Sigma \cdot x = \bigcup_{t=1}^{2T} \mathcal{O}^t(x)$. □

From Lemma 2.10 we easily deduce sufficient conditions which provide semi-algebraic orbits respectively semi-algebraic reachable sets.

**Theorem 2.11** *Let $\Sigma = (M, U, f)$ be an algebraically invertible system. Then*

- *a) If $\mathcal{R}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^t(x)$ for one $T \in \mathbb{N}$, then $\mathcal{R}(x)$ is semi-algebraic.*
- *b) If $\mathcal{O}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{O}^t(x)$ for one $T \in \mathbb{N}$, then $G_\Sigma \cdot x$ is semi-algebraic.*
- *c) If $\Sigma$ is abelian and $\mathcal{R}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^t(x)$ for one $T \in \mathbb{N}$, then $G_\Sigma \cdot x$ is semi-algebraic.*

**Proof.** a) and b) For $t \in \mathbb{N}$ and $\epsilon \in \{-1, 1\}^t$ we define

$$F_x^\epsilon : U^t \to M,\ (u_1, \dots, u_t) \mapsto f_{u_t}^{\epsilon_t} \circ \cdots \circ f_{u_1}^{\epsilon_1}(x).$$

Note, that $\mathcal{R}^t(x) = F_x^{(1,\dots,1)}(U^t)$ and

$$\mathcal{O}^t(x) = \bigcup_{\epsilon \in \{-1,1\}^t} F_x^\epsilon(U^t).$$

Now we show, that for all $t \in \mathbb{N}$ and all $\epsilon \in \{-1, 1\}^t$ the set $F_x^\epsilon(U^t)$ is semi-algebraic. Then, under above assumptions, $\mathcal{R}(x)$, respectively $G_\Sigma \cdot x$, are – by Lemma 2.10 – finite unions of semi-algebraic sets and therefore semi-algebraic.

Recall that $f$ is semi-algebraic and $\{x\} \times U$ is semi-algebraic by Proposition A.1. Therefore $F_x^{(1)}(U) = f(\{x\} \times U)$ is semi-algebraic by Corollary A.3. Moreover,

$$\begin{aligned} F_x^{(-1)}(U) &= \{y \in M \mid f_u(y) = x \text{ for some } u \in U\ \} \\ &= \pi_M\left(\{(y,u) \in M \times U \mid f(y,u) = x\}\right) \\ &= \pi_M\left(f^{-1}(\{x\})\right), \end{aligned}$$

where $\pi_M : M \times U \to U$, $(x,u) \mapsto x$. In other words, $F_x^{(-1)}(U)$ is the projection of the semi-algebraic set $f^{-1}(\{x\})$ and therefore semi-algebraic (see Theorem A.2 and Corollary A.3). By induction it follows that $F_x^{\epsilon}(U^t))$ is semi-algebraic, since

$$F_x^{(1,\epsilon)}(U^{t+1}) = f(U \times F_x^{(\epsilon)}(U^t))$$

and

$$F_x^{(-1,\epsilon)}(U^{t+1}) = \pi_M\left(f^{-1}(F_x^{(\epsilon)}(U^t))\right).$$

c) If $\Sigma$ is abelian, $\mathcal{R}^{T+1}(x) \subseteq \bigcup_{t=1}^{T} \mathcal{R}^t(x)$ implies that $G_\Sigma \cdot x$ is the union of finitely many sets $\mathcal{O}^t(x)$, $t \in \mathbb{N}$ (see Lemma 2.10). Therefore, the claim follows from b). □

If $\Sigma$ is abelian, then – by Lemma 2.10 – $\mathcal{R}(x) = \bigcup_{t=1}^{\tilde{T}} \mathcal{R}^t(x)$ implies, that $G_\Sigma \cdot x$ is the union of finitely many sets of the form $\mathcal{O}^t(x)$, $t \in \mathbb{N}$. The following example shows that the converse is false, i.e., that $G_\Sigma \cdot x = \bigcup_{t=1}^{T} \mathcal{O}^t(x)$ does not imply that $\mathcal{R}(x) = \bigcup_{t=1}^{\tilde{T}} \mathcal{R}^t(x)$ for any $\tilde{T} \in \mathbb{N}$.

**Example 2.12** Let $\Sigma = (\mathbb{R}^+, (\frac{1}{2}, \infty), f)$ with $f : (x,u) \mapsto ux$. Here $\mathbb{R}^+$ denotes the set of positive real numbers. Then

$$\mathcal{O}^1(x) = (\frac{1}{2}x, \infty) \cup (0, 2x) = \mathbb{R}^+ = G_\Sigma \cdot x.$$

On the other hand, $\mathcal{R}^T(x) = (\frac{1}{2^T}x, \infty) \neq \mathbb{R}^+$. Therefore,

$$\mathcal{R}(x) = \bigcup_{t=1}^{\infty} \mathcal{R}^t(x) = \mathbb{R}^+ \neq \bigcup_{t=1}^{T} \mathcal{R}^t(x)$$

for any $T \in \mathbb{N}$.

### 2.1.3 Right divisible systems

In many important cases, the system semigroup has additional structure. In fact, most of the systems we are analyzing in Part II have an abelian system semigroup. Nevertheless, in Chapter 10 we deal with linear systems $x_{t+1} = Ax_t + Bu_t$, where the corresponding system semigroup is not abelian, but fulfills weaker conditions, which we named *right divisibility*, and respectively *left divisibility*.

**Definition 2.13 (Right divisible systems)** A subsemigroup $S$ of a group is said to be *right divisible* if $\langle S\rangle = SS^{-1}$, i.e., every $g \in \langle S\rangle$ can be written in the form $g = s_1 s_2^{-1}$ with $s_1, s_2 \in S_\Sigma$. We say that an invertible system $\Sigma = (M, U, f)$ is *right divisible* if its system semigroup $S_\Sigma$ is *right divisible*. Analogously, we say an invertible system is *left divisible* if the system semigroup is *left divisible*, i.e., $\langle S_\Sigma\rangle = S_\Sigma^{-1}S_\Sigma$.

Note that every abelian semigroup is right divisible and left divisible. The following example shows, that the converse is wrong in general.

**Example 2.14** Let $\mathbb{F}$ be a field and $R$ be a subring of $\mathbb{F}$. Assume that for all $f \in \mathbb{F}$ there exists $r \in R$ such that $fr \in R$. Then

$$S := \{(r_{i,j})_{i,j=1,\dots,n} \in \mathrm{GL}_n(\mathbb{F}) \mid r_{i,j} \in R, i, j = 1, \dots, n\}$$

is a right divisible and left divisible semigroup. Note that $S$ is not abelian in general[5]. For any $(f_{i,j})_{i,j=1,\dots,n} \in \mathrm{GL}_n(\mathbb{F})$, we choose $r_{i,j} \in R$, $i, j = 1, \dots, n$ such that $f_{i,j}r_{i,j} \in R$. Then $r := \prod_{i,j=1,\dots,n} r_{i,j} \in R$ has the property $f_{i,j}r \in R$ for all $i, j = 1, \dots, n$. Therefore,

$$(f_{i,j})_{i,j=1,\dots,n} = (rI)^{-1}(f_{i,j}r)_{i,j=1,\dots,n} = (f_{i,j}r)_{i,j=1,\dots,n}(rI)^{-1}.$$

We conclude $\mathrm{GL}_n(\mathbb{F}) = SS^{-1} = S^{-1}S$.

Obviously, a semigroup $S$ is right divisible if the semigroup $S^{-1}$ is left divisible. The following result provides a practical method for checking if a given system is right divisible or not, without knowing $G_\Sigma$ explicitly.

**Theorem 2.15** *An invertible system $\Sigma = (M, U, f)$ is right divisible if and only if the following condition holds:*

$$\textit{for all } s_\alpha, s_\beta \in S_\Sigma \;\; \textit{there exists} \;\; s \in S_\Sigma \;\; \textit{such that} \;\; s_\alpha^{-1}s_\beta s \in S_\Sigma. \tag{16}$$

**Proof.** Assume that $G_\Sigma = S_\Sigma S_\Sigma^{-1}$. Then for any $s_\alpha, s_\beta \in S_\Sigma$ there exists $s_1, s_2 \in S_\Sigma$ such that $s_\alpha^{-1}s_\beta = s_1 s_2^{-1}$. Hence, $s_\alpha^{-1}s_\beta s_2 \in S_\Sigma$. Hence, (16) is

[5] In particular in the special case $\mathbb{F} = \mathbb{R}(x)$ and $r = \mathbb{R}[x]$, $x$ single variable, for $n \geq 2$.

fulfilled.
Conversely, let us assume that (16) is fulfilled. For any $g \in G_\Sigma$ there exists $n \in \mathbb{N}$, $s_1, \dots, s_n \in S_\Sigma$ and $\epsilon_1, \dots, \epsilon_n \in \{-1, 1\}$ such that

$$g = s_1^{\epsilon_1} \dots s_n^{\epsilon_n}.$$

We show that $g \in S_\Sigma S_\Sigma^{-1}$ by induction. For $n = 1$ we have to distinguish between the cases $g \in S_\Sigma$ and $g \in S_\Sigma^{-1}$. In the first case we have $g = gss^{-1} \in S_\Sigma S_\Sigma^{-1}$. In the second case we choose $s_\beta, s \in s_\Sigma$ such that $gs_\beta s =: \tilde{s} \in S_\Sigma$. Then $g = \tilde{s}(s_\beta s)^{-1} \in S_\Sigma S_\Sigma^{-1}$.

Now let $g = g_n \tilde{s}^{\tilde{\epsilon}}$ such that $g_n = s_1^{\epsilon_1} \dots s_n^{\epsilon_n} = \tilde{s}_1 \tilde{s}_2^{-1}$ with $s_1, \dots, s_n, \tilde{s}, \tilde{s}_1, \tilde{s}_2 \in S_\Sigma$ and $\epsilon_1, \dots, \epsilon_n, \tilde{\epsilon} \in \{-1, 1\}$. If $\tilde{\epsilon} = -1$ then

$$g = \tilde{s}_1 \tilde{s}_2^{-1} \tilde{s}^{-1} = \tilde{s}_1 (\tilde{s} \tilde{s}_2)^{-1} \in S_\Sigma S_\Sigma^{-1}$$

and we are done. If $\tilde{\epsilon} = 1$ then $g = \tilde{s}_1 \tilde{s}_2^{-1} \tilde{s}$. Now we choose $s \in S_\Sigma$ such that $\tilde{s}_2^{-1} \tilde{s} s \in S_\Sigma$. Hence,

$$g = s_1 (\tilde{s}_2^{-1} \tilde{s} s) s s^{-1} \in S_\Sigma S_\Sigma^{-1}.$$

□

**Corollary 2.16** *Let $S$ be a subsemigroup of a group $G$ and $N$ a normal subgroup of $G$.*

*a) If $S$ is right divisible, then $NS$ is right divisible.*

*b) If $S$ is left divisible, then $SN$ is left divisible.*

**Proof.** a) For any $n_1 s_1, n_2 s_2 \in NS$ there exists $\tilde{n} \in N$ such that

$$(n_1 s_1)^{-1} n_2 s_2 = (s_1^{-1} n_1^{-1} n_2 s_1) s_1^{-1} s_2 = \tilde{n} s_1^{-1} s_2.$$

If $S$ is right divisible then there exists $s \in S$ such that $\tilde{n} s_1^{-1} s_2 s \in NS$ (see Theorem 2.15). Hence, $NS$ is right divisible. b) If $S$ is left divisible then $S^{-1}$ and $NS^{-1}$ is right divisible. Therefore, $(NS^{-1})^{-1} = SN^{-1} = SN$ is left divisible. □

We finish this section with two examples. In the first example we analyze an explicit system which is right divisible and left divisible but not abelian. In the second example we show a system which is neither right divisible nor left divisible.

**Example 2.17** Let $S_\Sigma$ be the system semigroup of a system on $M = \mathbb{R}^2$ defined by

$$f_u(x) = ux; \quad u \in U := \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \middle| \, a, b, c > 0 \right\}.$$

Obviously, $S_\Sigma$ can be identified with the non abelian matrix semigroup $U$. The following calculation shows, that for every $s_1, s_2 \in S_\Sigma$ there exists $u \in S_\Sigma$ such that $s_1^{-1} s_2 u \in S_\Sigma$. Thus, $S_\Sigma$ is right divisible by Theorem 2.15. Let

$$s_1 = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix}, \; s_2 = \begin{pmatrix} \tilde{a} & \tilde{b} \\ 0 & \tilde{c} \end{pmatrix}$$

with $a, b, c, \tilde{a}, \tilde{b}, \tilde{c} > 0$. Then

$$s_1^{-1} s_2 \begin{pmatrix} 1 & y \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{\tilde{a}}{a} & y\frac{\tilde{a}}{a} + \left(\frac{\tilde{b}}{a} - \frac{b\tilde{c}}{ac}\right) \\ 0 & \frac{\tilde{c}}{c} \end{pmatrix} \in S_\Sigma$$

for $y$ large enough. In particular this shows

$$G_\Sigma = S_\Sigma \cdot S_\Sigma^{-1} = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \middle| \, a, c > 0 \right\}.$$

Hence, $\Sigma$ is right divisible. For any $a_1, b_1, c_1, a_2, b_2, c_2 > 0$ we find $x > 0$ large enough, such that $y := \frac{1}{c_1}(c_1 x + \frac{b_1}{a_1} - \frac{b_2}{a_2})$ is positive. By construction we obtain

$$\begin{pmatrix} \frac{1}{a_1} & x \\ 0 & \frac{1}{c_1} \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ 0 & c_1 \end{pmatrix} = \begin{pmatrix} a_2^{-1} & y \\ 0 & c_2^{-1} \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ 0 & c_2 \end{pmatrix}.$$

In other words, for any $s_1, s_2 \in S_\Sigma$ there exists $\tilde{s}_1, \tilde{s}_2 \in S_\Sigma$ such that $s_1 s_2^{-1} = \tilde{s}_1^{-1} \tilde{s}_2$ and therefore $S_\Sigma S_\Sigma^{-1} \subseteq S_\Sigma^{-1} S_\Sigma \subseteq G_\Sigma$. Hence, $\Sigma$ is left divisible.

**Example 2.18** Consider $\Sigma = (\mathbb{R}^2 \setminus \{0\}, U, f)$ given by

$$U = \left\{ \begin{pmatrix} u_1 & u_2 \\ u_3 & u_4 \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \, \middle| \, u_i > 0, i = 1, \ldots, 4 \right\}$$

and the transition map by $f(x, U) = Ux$. Obviously, the system semigroup can be identified with the matrix set

$$S_\Sigma = \left\{ \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \, \middle| \, u_{11}, u_{12}, u_{21}, u_{22} > 0 \right\}.$$

We show that $S_\Sigma$ is not right divisible using Theorem 2.15. In particular, for

$$s_1 = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix} \in S_\Sigma \;\text{ and }\; s_2 = \begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix} \in S_\Sigma$$

there exists no $u = (u_{i,j})_{i,j=1,2} \in S_\Sigma$ such that $s_1^{-1}s_2u \in S_\Sigma$, since

$$\begin{aligned} s_1^{-1}s_2u &= \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix} \\ &= \begin{pmatrix} 3u_{11} + 2u_{21} & 3u_{12} + 2u_{22} \\ -u_{11} - u_{21} & -u_{12} - u_{22} \end{pmatrix} \notin S_\Sigma, \end{aligned}$$

since $-u_{11} - u_{21} < 0$. Hence, $\Sigma$ is not right divisible, since Condition (16) is not fulfilled.

Now we show that $\Sigma$ is not left divisible. With the notation above we obtain

$$s_1^{-1}s_2s_1^{-1} = \begin{pmatrix} -1 & 5 \\ 0 & -1 \end{pmatrix} \in G_\Sigma.$$

Assuming $s_1^{-1}s_2s_1^{-1} \in (S_\Sigma)^{-1}S_\Sigma$ there exists $s_\alpha, s_\beta \in S_\Sigma$ such that

$$g := s_\alpha \begin{pmatrix} -1 & 5 \\ 0 & -1 \end{pmatrix} = s_\beta.$$

This is not possible, since $g \cdot (6,1)^\top \subseteq \mathbb{R}^- \times \mathbb{R}^-$ but $s_\beta \cdot (6,1)^\top \subseteq \mathbb{R}^+ \times \mathbb{R}^+$. Hence, $\Sigma$ is not left divisible.

Note that in all examples in this thesis the system is either right divisible and left divisible or not right divisible and not left divisible. To our knowledge, it is unknown if right divisibility implies left divisibility.

## 2.2 Accessibility

*Accessibility* is the property that one is able to reach a set of full dimension from a given state. More formally we define:

**Definition 2.19 (Accessibility)** A system $\Sigma = (M, U, f)$ is said to be *accessible from* $x \in M$ if $\mathcal{R}(x)$ has nonempty interior in $M$. We say $\Sigma$ is *accessible* if $\operatorname{int} \mathcal{R}(x) \neq \emptyset$ for any $x \in M$.

In the following subsection we briefly describe two techniques to check whether a system $\Sigma$ is accessible from a certain point. The first one is a geometric framework, developed by Jakubczyk and Sontag. It is similar to the well-known Lie-theoretical approach for continuous-time systems. Afterwards we present technique which uses topological structure of the system group and system semigroup.

In many situations, accessibility from $y \in G_\Sigma \cdot x$ is sufficient for accessibility from all $z \in G_\Sigma \cdot x$. This phenomena is called *Chow property* and will be the topic of Subsection 2.2.2.

### 2.2.1 Conditions for accessibility

First of all, we want to point out a basic necessary condition for accessibility from one point.

**Proposition 2.20** *Let $\Sigma = (M, U, f)$ be an invertible system. If $\Sigma$ is accessible from $x \in M$, then the system group orbit $G_\Sigma \cdot x$ is open in $M$.*

**Proof.** Since $\mathcal{R}(x) = S_\Sigma \cdot x$ has nonempty interior, there exists $s \in S_\Sigma$ such that $s \cdot x \in \operatorname{int}_M(S_\Sigma \cdot x)$. For any $y \in G_\Sigma \cdot x$ there exists $g \in G_\Sigma$ such that

$$y = g \cdot x = gs^{-1}(s \cdot x) \subseteq \underbrace{gs^{-1} \left(\operatorname{int}_M \mathcal{R}(x)\right)}_{:=V} \subseteq G_\Sigma \cdot x$$

Since $gs^{-1}$ is a homeomorphism, $V$ is a neighborhood of $y$ in $M$. Hence, $G_\Sigma \cdot x$ is open. □

In particular, knowing the structure of the system group orbits of $\Sigma$, it is enough to check the elements of the open orbits for accessibility.

Now we introduce sufficient conditions for accessibility. Here we assume that $U \subseteq \mathbb{R}^m$ is open and that $f : M \times U \to U$ is smooth. Let $\tilde{U}$ be a subset of $U$ such that every connected component of $U$ has at least one element in $\tilde{U}$. For $u \in U$, $k \in \mathbb{N}_0$, $1 \leq i \leq m$ and $u_1, \ldots, u_k \in \tilde{U}$ we define the *Lie derivative vector field*

$$\operatorname{Ad}_{u_1,\ldots,u_k} f_{u,i} : M \to TM$$

given by

$$x \mapsto \frac{\partial}{\partial v_i}\bigg|_{v=0} (f_{u_k} \circ \cdots \circ f_{u_1})^{-1} \circ f_u^{-1} \circ f_{u+v} \circ (f_{u_k} \circ \cdots \circ f_{u_1})(x). \tag{17}$$

In particular if (i) $\Sigma$ is abelian or if (ii) $U$ is connected and $f_u = \mathrm{id}$ for some $u \in U$, (17) reduces to

$$x \mapsto \frac{\partial}{\partial v_i}\bigg|_{v=0} f_u^{-1} \circ f_{u+v}(x).$$

The above family of vector fields generates a Lie-algebra $\mathcal{L}_\Sigma$, i.e., the smallest Lie algebra which contains all elements $\mathrm{Ad}_{u_1,\dots,u_k} f_{u,i}$, for $u_1, \dots, u_k \in \tilde{U}$, $k \in \mathbb{N}_0$, $u \in U$, $i = 1, \dots, m$. For every $x \in M$ a linear space of tangent vectors at $x$ is given by

$$\mathcal{L}_\Sigma(x) := \{X(x) \mid X \in \mathcal{L}_\Sigma\} \subseteq T_x M. \tag{18}$$

The following theorem gives a necessary and sufficient condition for accessibility in terms of $\mathcal{L}_\Sigma$.

**Theorem 2.21 (Jakubczyk and Sontag [JS90])** *Let $\Sigma = (M, U, f)$ be a smoothly invertible system, such that $U$ is an open subset of $\mathbb{R}^m$ and $f : M \times U \to M$ is smooth. Then $\Sigma$ is accessible if and only if* $\dim \mathcal{L}_\Sigma(x) = n$ *for all $x \in M$.*

A proof of Theorem 2.21 can be found in [JS90] (See Theorem 3 for the case where $U$ is a connected subset of $\mathbb{R}$ and Theorem 9 for the generalization to $U \subseteq \mathbb{R}^m$). Note that $\dim \mathcal{L}_\Sigma(x)$ is independent from the choice of $\tilde{U}$ (see Remark 4.5 in [JS90]).

Now we present another technique for checking accessibility, which is particularly useful for systems $\Sigma$, where $S_\Sigma$ and $G_\Sigma$ have an additional structure. In particular, in our applications in Part 5.2.2 we will deal with systems, given by well-known numerical algorithms, where $G_\Sigma$ turns out to be a subgroup of a Lie group $G$. In this situation we equip $G_\Sigma$ and $S_\Sigma$ with the subspace topology relative to $G$. Obviously, accessibility properties are linked with topological relations between $S_\Sigma$ and $G_\Sigma$.

**Lemma 2.22** *Let $\Sigma = (M, U, f)$ be an invertible system and $G_\Sigma$ equipped with a topology.*

*a) If $h_x : S_\Sigma \to M;\ s \mapsto s \cdot x$ is continuous, then* $\mathrm{int}_M \mathcal{R}(x) \neq \emptyset$ *implies* $\mathrm{int}_{G_\Sigma} S_\Sigma \neq \emptyset$.

*b) If $h_x : S_\Sigma \to M;\ s \mapsto s \cdot x$ is an open map, then* $\mathrm{int}_{G_\Sigma} S_\Sigma \neq \emptyset$ *implies accessibility for all $y \in G_\Sigma \cdot x$.*

**Proof.** a) If $\mathrm{int}_M \mathcal{R}(x)$ is nonempty and the map

$$h_x : S_\Sigma \to M; s \mapsto s \cdot x$$

is continuous, then $h_x^{-1}(\mathrm{int}_M \mathcal{R}(x)) \subseteq G_\Sigma$ is an open subset of $S_\Sigma$.
b) Suppose $y \in G_\Sigma \cdot x$ and $\mathrm{int}_{G_\Sigma} S_\Sigma \neq \emptyset$. Then there exist $g \in G_\Sigma$ such that $y = g \cdot x$. If $h_x$ is open, then also $h_y$ is open, since $h_y = h_x \circ r_{s^{-1}g}$ with

$$r_{s^{-1}g} : G_\Sigma \to G_\Sigma,\ h \to hs^{-1}g.$$

Therefore, $\mathcal{R}(y)$ has open interior, since

$$\mathcal{R}(y) = S_\Sigma \cdot y \supseteq \mathrm{int}_{G_\Sigma} S_\Sigma \cdot y = h_y \left(\mathrm{int}_{G_\Sigma} S_\Sigma\right).$$

□

In applications it is reasonable to choose a topology on $G_\Sigma$ such that $G_\Sigma \times M \to M$ is continuous and therefore $h_x$ is continuous for all $x \in M$. In most important cases we obtain $\mathrm{int}_{G_\Sigma} S_\Sigma \neq \emptyset$ (but not always, see Example 2.24). The assumption that $h_x$ is open is certainly very restrictive. On the other hand, one can always restrict the system to the group orbit $G_\Sigma \cdot x$. This will be the topic of Section 3.2 and Section 4.2. Then, for the restricted system, $G_\Sigma$ acts transitively[6] on $M$ and – under weak assumptions on $G_\Sigma \cdot x$ and $G_\Sigma$ – the map $h_x$ is open (see Theorem B.8). Using techniques from the theory of topological semigroups, we obtain the following sufficient condition for accessibility.

**Theorem 2.23** *Let $\Sigma = (M, U, f)$ be a system on a manifold $M$ and $G_\Sigma$ be equipped with a Lie group structure, such that $G_\Sigma \times M \to M$, $(g, x) \mapsto g(x)$ is transitive and continuous. If*

$$\mathrm{int}_{G_\Sigma} S_\Sigma \cap \mathrm{Stab}_x \neq \emptyset \tag{19}$$

*for $x \in M$, then $\Sigma$ is accessible from $x$ and $x \in \mathrm{int}_M \mathcal{R}(x)$.*

Here $\mathrm{Stab}_x$ denotes the stabilizer subgroup $\mathrm{Stab}_x := \{g \in G_\Sigma \mid g \cdot x = x\}$.

**Proof.** The claim follows from known results on actions of subsemigroups of Lie groups. By Theorem B.8 the map $h_x : G_\Sigma \to M$, $g \mapsto g \cdot x$ is open. Hence, $\Sigma$ is accessible by Lemma 2.22. Moreover, if Condition (19) is fulfilled, then there exists a neighborhood $U$ of $x$ such that $S_\Sigma$ acts transitively on $U$. In other words for all $u_1, u_2 \in U$ there exists $s \in S_\Sigma$ such that $s \cdot u_1 = u_2$ (see Proposition B.7). Hence, $U \subseteq \mathcal{R}(x)$ and $x \in \mathrm{int}_M U \subseteq \mathrm{int}_M \mathcal{R}(x)$. □

[6] In the literature on discrete-time systems (for example [JS90] and [AS93, AS94]), transitivity often means, that the orbit $G_\Sigma \cdot x$ has nonempty interior in $M$ (and is therefore open by Proposition 2.20). Nevertheless, in this work transitivity means that $G_\Sigma \cdot x = M$ for some (and therefore for all) $x \in M$.

In order to apply part b) of Lemma 2.22 and Theorem 2.40 the system semigroup $S_\Sigma$ needs to have nonempty interior with respect to the topology of $G_\Sigma$. Using exotic topologies, such as the indiscrete topology, one can easily construct examples such that the interior of $S_\Sigma$ is empty. In fact the following example shows, that this can also be done if $G_\Sigma$ has a Lie group topology, provided $U$ is sufficiently anomalous.

**Example 2.24** Let us consider the following system $\Sigma = (\mathbb{R}, U, f)$ where $U \subseteq \mathbb{R}$ is given by the following construction. Recall that $\mathbb{R}$ is a topological $\mathbb{Q}$-vectorspace, with respect to the usual topology of $\mathbb{R}$. We choose a basis $\{b_i \mid i \in I\}$ such that $b_1 = 1$ and $b_2 = -\sqrt{2}$ and define

$$U = \left\{ \sum_{i \in \tilde{I}} \lambda_i b_i \;\middle|\; \tilde{I} \subseteq I, |\tilde{I}| < \infty, \lambda_i \in \mathbb{Q}^+ \right\}.$$

Now let $f(x,u) = x + u$. Obviously we can identify the semigroup $S_\Sigma$ with $U$. Moreover $G_\Sigma = \mathbb{R}$, since every $r \in \mathbb{R}$ is a sum of two elements, one from $S_\Sigma$ and one from $-S_\Sigma$. On the other hand we have

$$\operatorname{int}_{G_\Sigma} S_\Sigma = \emptyset,$$

since $-\mathbb{Q}^+$ and $\sqrt{2}\mathbb{Q}^+$ are disjoint to $S_\Sigma$.

### 2.2.2 Chow property

A useful property for analytic continuous-time systems is that every reachable set has nonempty interior in the corresponding system-group orbit. This fact is known as *the positive form of Chow's lemma* ( See [Kre74], Theorem 1 for a proof). In particular, a system is accessible from $x \in M$ if and only if it is accessible from all $y$ contained in the system-group orbit of $x$.

For discrete-time systems – even if the transition map is analytic – it might happen that reachable sets $\mathcal{R}(x)$ have empty interior in $G_\Sigma \cdot x$. An example has been given by Albertini and Sontag in [AS93] (Example 5.1). Using the same example, we show that accessibility from $x \in M$ does not imply accessibility from $y \in G_\Sigma \cdot x$ (see Example 2.29 below).

Nevertheless, in the following we present some sufficient conditions on $\Sigma$ which imply the following useful property.

**Definition 2.25 (Chow property)** A system $\Sigma = (M, U, f)$ has the *Chow Property* if accessibility from $x \in M$ implies accessibility from all $y \in G_\Sigma \cdot x$.

We start with a trivial but important observation on systems with abelian semigroup.

**Theorem 2.26** *Every abelian invertible system has the Chow property.*

**Proof.** If $S_\Sigma$ is abelian, then

$$\mathcal{R}(y) = S_\Sigma \cdot y = S_\Sigma g \cdot x = g(S_\Sigma \cdot x) \supseteq g(\operatorname{int}_M \mathcal{R}(x)).$$

Hence, $\mathcal{R}(y)$ has nonempty interior, since $g$ is a homeomorphism. □

Using the geometric framework developed by Jakubczyk and Sontag (see [JS90], respectively Theorem 2.21), Albertini and Sontag provide some sufficient conditions for the Chow property. Recall that a point $x \in M$ is said to be *positively Poisson stable* if for each neighborhood $V$ of $x$, there exists an integer $T \in N$ and $u_1, \dots, u_T \in U$ such that $f_{u_T} \circ \dots \circ f_{u_1}(x) \in V$.

**Theorem 2.27 (Albertini and Sontag [AS93, AS94])** *Let $\Sigma$ be an invertible system with $U$ open in $\mathbb{R}^m$. We assume that $M$ is an analytic manifold and that $f$ is analytic.*

*a) If $x \in M$ is positively Poisson stable, then accessibility from $x$ implies accessibility from all $y \in G_\Sigma \cdot x$.*

*b) If $G_\Sigma \cdot x$ is compact and $M = G_\Sigma \cdot x$, then $\Sigma$ has the Chow property.*

**Proof.** If $\Sigma$ is accessible from $x \in M$, then $G_\Sigma \cdot x$, and therefore $G_\Sigma \cdot y$ for any $y \in G_\Sigma \cdot x$, is open in $M$ (see Proposition 2.20). Now the claims follow immediately from the results in [AS93] and [AS94]. In particular from the assumptions in *a*) (in *b*)) it follows, that $\operatorname{int}_M G_\Sigma \cdot y \neq \emptyset$ implies accessibility from $y$, by Theorem 1 in [AS94] (by Theorem 4.4 in [AS93]). □

In our applications in Chapters 6-10 the system semigroups carry a Lie group structure and the system semigroups carry the topology induced by $G_\Sigma$. In particular, using Lemma 2.22, we obtain the following sufficient condition for the Chow property.

**Theorem 2.28** *Let $\Sigma = (M, U, f)$ be an invertible system. Assume that $G_\Sigma$ is a Lie group such that the action $G_\Sigma \times M \to M$ is continuous. Let $x \in M$ such that $G_\Sigma \cdot x$ is locally compact*[7]*. Then $\Sigma$ has the Chow property.*

**Proof.** Obviously, the restricted action $G_\Sigma \times G_\Sigma \cdot x \to G_\Sigma \cdot x$ is continuous and transitive. Recall that a Lie group is a locally compact Lindelöf space. Now, by Theorem B.8 it follows that $h_x : S_\Sigma \to M$, $s \mapsto s \cdot x$ is continuous and open. If $\operatorname{int}_M \mathcal{R}(x) \neq \emptyset$, then $\operatorname{int}_{G_\Sigma} S_\Sigma \neq \emptyset$ by part a) of Lemma 2.22. Then $\operatorname{int}_M \mathcal{R}(y) \neq \emptyset$ for all $y \in G_\Sigma \cdot x$ by part b) of Lemma 2.22. Hence, accessibility from $x$ implies accessibility from all $y \in G_\Sigma \cdot x$.

□

[7] In particular, $G_\Sigma \cdot x$ is a submanifold and therefore locally compact, if $\Sigma = (M, U, f)$ is smoothly invertible and $G_\Sigma \cdot x$ is semi-algebraic (see Theorem 2.7). Note that a semi-algebraic set is not locally compact in general, see for example $M = (\mathbb{R}^+ \times \mathbb{R}) \cup \{(0, 0)\}$.

At the end of this section we show — using an example of Albertini and Sontag — that not every analytic system has the Chow property.

**Example 2.29** Let $\Sigma = (M, U, f)$ be given by $M = \mathbb{Z} \times \mathbb{R}$, $U = \mathbb{R}$ and

$$f : M \times U \to M, \left( \begin{pmatrix} x \\ y \end{pmatrix}, u \right) \mapsto \begin{pmatrix} x+1 \\ y + uh(x) \end{pmatrix}.$$

Here, $h : \mathbb{R} \to \mathbb{R}$ is any analytic function with $h(0) = 1$ and $h(x) = 0$ if and only if $x \in \mathbb{N}$. Note that $f$ is analytic. Moreover, $f_u$ is a diffeomorphism with

$$f_u^{-1} : M \to M, \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x-1 \\ y - uh(x-1) \end{pmatrix}.$$

for any $u \in \mathbb{R}$. Now we prove that $\Sigma$ is accessible from $(0,0)^\top$ but not accessible from $(0,1)^\top \in G_\Sigma \cdot (0,0)^\top$. This shows in particular, that $\Sigma$ does not have the Chow property.

(i) First we demonstrate that the system group orbit of $(0,0)^\top \in M$ is

$$G_\Sigma \cdot (0,0)^\top = \mathbb{Z} \times \mathbb{R},$$

i.e., we show that for any $(x,y)^\top \in \mathbb{Z} \times \mathbb{R}$ there exists $g \in G_\Sigma$ such that

$$g \cdot (0,0)^\top = (x,y)^\top. \tag{20}$$

Recall that $h(0) = 1$ and $h(-1) \neq 0$. If $x = 0$, then (20) is fulfilled by the choice $g = f_0 \cdot f_u^{-1}$ with $u = -y/h(-1)$, since

$$f_0 \cdot f_u^{-1} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = f_0 \begin{pmatrix} -1 \\ -uh(-1) \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix}.$$

If $x < 0$, then for $u = y$ we obtain

$$f_u \circ \underbrace{f_0^{-1} \circ \cdots \circ f_0^{-1}}_{-x+1 \text{ times}} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = f_u \begin{pmatrix} -(-x+1) \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ uh(0) \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Hence, (20) is fulfilled by $g = f_u \cdot f_0^{x-1}$.

If $x > 0$, we choose $u = y/h(0)$. Then (20) is fulfilled by $g = f_u$ if $x = 1$, or by $g = f_0^{x-1} \circ f_u$ if $x > 0$, since

$$\underbrace{f_0 \circ \cdots \circ f_0}_{x-1 \text{ times}} \circ f_u \begin{pmatrix} 0 \\ 0 \end{pmatrix} = f_0^{x-1} \begin{pmatrix} 1 \\ uh(0) \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

We conclude $G_\Sigma \cdot (0,0)^\top = M$. Note that in the case $x > 0$, it is possible to find $g \in S_\Sigma$ to fulfill (20).

(ii) Now we show that $\Sigma$ is accessible from $(0,0)^\top$. In (i) we have seen that every element of $\mathbb{N} \times \mathbb{R}$ can be reached from $(0,0)^\top$ by elements of $S_\Sigma$. Hence $\mathcal{R}((0,0)^\top) \supseteq \mathbb{N} \times \mathbb{R}$ and $\Sigma$ is accessible from $(0,0)^\top$.
(iii) In particular, (ii) shows that $(j,0)^\top \in \mathcal{R}((0,0)^\top)$ for $j \in \mathbb{N}$. On the other hand, Albertini and Sontag have pointed out that $\Sigma$ is not accessible from $(j,0)^\top$, $j \in \mathbb{N}$. In fact, we obtain

$$S_\Sigma \cdot \begin{pmatrix} j \\ 0 \end{pmatrix} = \bigcup_{i=j+1}^{\infty} \left\{ \begin{pmatrix} i \\ 0 \end{pmatrix} \right\},$$

since for all $u \in U$, $j \in \mathbb{N}$ we have $f_u \cdot (j,0) = (j+1, 0 + uh(j)) = (j+1,0)$ and therefore $s \cdot (j,0) = f_{u_1} \circ \cdots \circ f_{u_T} \cdot (j,0) = (j+T,0)$.

## 2.3 Controllability

In this section we introduce the concept of *controllability*. First we give the classical definition that can be found in any textbook dealing with discrete-time nonlinear control systems (see for example [Son98], Definition 3.1.6). Afterwards we show necessary as well as sufficient conditions for controllability and other related properties. Here we always emphasize the semigroup orbit structure of the reachable sets.

**Definition 2.30 (Controllability)** A system $\Sigma = (M, U, f)$ is said to be

- *reachable from* $x \in M$ if for any $y \in M$ there exist $T \in \mathbb{N}$ and $u \in U^T$ such that $f_T(x, u) = y$.
- *controllable*[8] if for all $x, y \in M$ there exist $T \in \mathbb{N}$ and $u \in U^T$ such that $f_T(x, u) = y$.
- *controllable on* $N \subseteq M$ if for all $x, y \in N$ there exist $T \in \mathbb{N}$ and $u \in U^T$ such that $f_T(x, u) = y$.

Obviously a system $\Sigma = (M, U, f)$ is reachable from $x \in M$ if and only if $\mathcal{R}(x) = S_\Sigma \cdot x = M$. Moreover, the following proposition shows the basic relation between controllability and reachability.

**Proposition 2.31** *For an invertible system* $\Sigma = (M, U, f)$ *the following statements are equivalent:*

*(i)* $\Sigma$ *is controllable*

*(ii)* $\Sigma$ *is reachable from every point in* $M$

*(iii)* $S_\Sigma \cdot x = G_\Sigma \cdot x = M$ *for all* $x \in M$.

**Proof.** Obviously $(i) \Rightarrow (ii)$ and $(iii) \Rightarrow (ii)$. For any $x \in M$ we have $S_\Sigma \cdot x \subseteq G_\Sigma \cdot x \subseteq M$. Therefore, reachability from every point implies $S_\Sigma \cdot x = G_\Sigma \cdot x = M$ for all $x \in M$. This also implies controllability of $\Sigma$ since $y \in S_\Sigma \cdot x$ for every $x, y \in M$. □

Obviously, $S_\Sigma = G_\Sigma$ implies that $\Sigma$ is controllable, provided $M = G_\Sigma \cdot x$. Nevertheless, the following example shows, that controllability does not necessarily imply $S_\Sigma = G_\Sigma$.

[8] Note that in the literature of linear systems *controllability* often means that every point $x$ can be steered to 0. In the discrete-time case this is not equivalent to Definition 2.30, see Example 2.11 in [AM06]. Nevertheless, in the literature of nonlinear systems Definition 2.30 is common (see Definition 3.1.6 in [Son98] or Definition 9, Chapter 3 in [Jur97]).

**Example 2.32** Let $M := \mathbb{R}$, $U := \mathbb{R}$ and $f(x,u) = x^3 + u$. Note that $\Sigma = (M,U,f)$ is an invertible system, because every $f_u$ is an homeomorphism. Since

$$\mathcal{R}(x) \supseteq \mathcal{R}^1(x) = \{x^3 + u \mid u \in \mathbb{R}\} = \mathbb{R}$$

for all $x \in M$ it follows from Proposition 2.31 that $\Sigma = (M,U,f)$ is controllable. On the other hand every element of $S_\Sigma$ is a non-constant polynomial and therefore has no inverse in $S_\Sigma$. Hence, $S_\Sigma \neq G_\Sigma$.

It is well-known, that a *linear system* $\Sigma = (\mathbb{R}^n, \mathbb{R}^m, f)$, i.e., a system given by $f(x,u) = Ax + Bu$ with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, is controllable if and only if the *Kalman rank condition* holds, i.e.,

$$\operatorname{rank}[B, AB, A^2B, \ldots, A^{n-1}B] = n.$$

(see Theorem 2, [Son98], Chapter 3). The nonlinear case is more complicated and requires more sophisticated techniques such as the concept of *accessibility*. Moreover, we need the notion of *weak reversibility* and *reachability from one point* which will be the topics of the following two subsections.

### 2.3.1 Weak reversibility

Accessibility is a necessary but not a sufficient[9] condition for controllability. On the other hand, it is well-known that for continuous-time systems accessibility implies controllability, provided that the system is *weakly reversible* (see Corollary 4.3.12 in [Son98]). In the following we show a similar result for discrete-time systems.

Analogous to the continuous-time case (see Definition 4.3.9 in [Son98]) we define *weak reversibility* as follows.

**Definition 2.33 (Weak reversibility)** A system $\Sigma = (M,U,f)$ is *weakly reversible* if (i) for every $x \in M$ there exists $y \in M$, such that $x \in \mathcal{R}(y)$ and (ii) for all $x, y \in M$ either $\mathcal{R}(x) = \mathcal{R}(y)$ or $\mathcal{R}(x) \cap \mathcal{R}(y) = \emptyset$.

In other words, $\Sigma$ is weakly reversible, if the reachable sets form a partition on the state space. Due to that, weak reversibility is also called *partition property*. Note that invertible systems always fulfill (i) since $x \in \mathcal{R}(s^{-1} \cdot x)$ for any $s \in S_\Sigma$.

In the classical definition of weak reversibility in the continuous-time case it is additionally assumed that $x \in \mathcal{R}(x)$ for all $x \in M$. The following

[9] An example for an accessible system which is not controllabel are certain Inverse Iteration systems (see Chapter 6)

proposition shows that in the discrete-time case, $x \in \mathcal{R}(x)$ follows from $(i)$ and $(ii)$ of Definition 2.33.

**Proposition 2.34** *If $\Sigma$ is weakly reversible, then $x \in \mathcal{R}(x)$ for all $x \in M$.*

**Proof.** By definition, weakly reversible implies $x \in \mathcal{R}(y)$ for some $y \in M$, i.e., $x = s \cdot y$ with $s \in S_\Sigma$. Therefore,

$$\mathcal{R}(x) = S_\Sigma \cdot x \subseteq S_\Sigma s \cdot y \subseteq S_\Sigma \cdot y = \mathcal{R}(y)$$

Part (ii) of the definition yields $\mathcal{R}(x) = \mathcal{R}(y)$. Hence, $x \in \mathcal{R}(x)$. □

The following result clarifies the term *weakly reversible*, i.e., it shows, that $\Sigma$ is weakly reversible if and only if any iteration step $x \xrightarrow{u} y$, $u \in U$ can be reversed by a finite control sequence.

**Lemma 2.35** *Let $\Sigma = (M, U, f)$ be an invertible system. Then the following statements are equivalent.*

*(i) $\Sigma$ is weakly reversible,*

*(ii) $x \in \mathcal{R}(y)$ implies $y \in \mathcal{R}(x)$ for all $x, y \in M$,*

*(iii) $G_\Sigma \cdot x = \mathcal{R}(x)$ for all $x \in M$.*

**Proof.** Note that $x \in \mathcal{R}(y)$ implies $\mathcal{R}(x) = S_\Sigma \cdot (s \cdot y) \subseteq S_\Sigma \cdot y = \mathcal{R}(y)$. Assuming that $\Sigma$ is weakly reversible, we obtain $\mathcal{R}(x) = \mathcal{R}(y)$ and therefore it follows from Proposition 2.34 that $y \in \mathcal{R}(x)$. Hence, $(i) \Rightarrow (ii)$.

Now we assume $(ii)$ to be fulfilled. In particular we obtain $s^{-1} \cdot x \in \mathcal{R}(x)$ for any $s \in S_\Sigma$, since $x \in \mathcal{R}(s^{-1} \cdot x)$. Moreover, $s \cdot x \in \mathcal{R}(x)$. Recall that $G_\Sigma = \langle S_\Sigma \rangle$. Therefore, for any $y \in G_\Sigma \cdot x$ there exist $g_1, \ldots, g_n \in S_\Sigma \cup S_\Sigma^{-1}$ such that $y = g_1 g_2 \ldots g_n \cdot x$. By induction it follows that $g \cdot x \in \mathcal{R}(x)$. We conclude

$$\mathcal{R}(x) = S_\Sigma \cdot x \subseteq G_\Sigma \cdot x = \bigcup_{g \in G_\Sigma} g \cdot x \subseteq \mathcal{R}(x).$$

Now we assume that (iii) is fulfilled. Then

$$\mathcal{R}(x) = G_\Sigma \cdot x = G_\Sigma \cdot y = \mathcal{R}(y)$$

if $y \in G_\Sigma \cdot x$, or

$$\mathcal{R}(x) \cap \mathcal{R}(y) = \emptyset$$

if $y \notin G_\Sigma \cdot x$. Moreover, $x \in \mathcal{R}(s^{-1} \cdot x)$ for any $s \in S_\Sigma$. Hence, $\Sigma$ is weakly reversible. □

By Lemma 2.35, $\Sigma$ is weakly reversible whenever $S_\Sigma$ is a group. The following example shows that the converse is not true in general.

**Example 2.36** Let $M = \mathbb{R}$, $U = \mathbb{R}^+ := (0, \infty)$ and

$$f(x,u) = \begin{cases} ux & \text{if } x \geq 0, \\ 2ux & \text{if } x < 0. \end{cases} \tag{21}$$

Note that $f_u$ is a homeomorphism for every $u \in U$. Every element of $S_\Sigma$ has the form

$$s(x) = \begin{cases} ux & \text{if } x \leq 0 \\ 2^k ux & \text{if } x < 0 \end{cases} \tag{22}$$

with $u \in U$ and $k \in \mathbb{N}$. In particular, $f_u^{-1} \notin S_\Sigma$ and therefore $S_\Sigma$ is not a group. On the other hand $\Sigma$ is weakly reversible, since $\mathcal{R}(x) = \mathbb{R}^+$ for $x > 0$, $\mathcal{R}(0) = \{0\}$ and $\mathcal{R}(x) = \mathbb{R}^-$ for $x < 0$.

We finish this section with a result analogous to the situation in continuous time (see Corollary 4.3.12 in [Son98]).

**Theorem 2.37** *Let $\Sigma = (M, U, f)$ be an invertible system on a connected manifold $M$. If $\Sigma$ is weakly reversible and accessible such that $x \in \operatorname{int}_M \mathcal{R}(x)$ for all $x \in M$, then $\Sigma$ is controllable.*

**Proof.** For any $y \in \mathcal{R}(x)$ weak reversibility implies $\mathcal{R}(y) = \mathcal{R}(x)$ and therefore $y \in \operatorname{int}_M \mathcal{R}(x)$. Hence, $\mathcal{R}(x)$ is open for all $x \in M$. Again, weak reversibility implies, that the reachable sets form a partition of open sets on the set $M$. Since $M$ is connected, $M = \mathcal{R}(y)$ for any $y \in M$. Hence, $\Sigma$ is controllable by Proposition 2.31. □

### 2.3.2 Reachability from one point

Obviously, controllability implies reachability from one point. We will show that the converse is false in general (see Example 2.42). Nevertheless, for certain types of systems, reachability from one point already implies reachability from every point and therefore controllability. In particular it is well known that linear systems have this property.

**Theorem 2.38** *Let $\Sigma = (M, U, f)$ be an invertible linear system, i.e. $M := \mathbb{R}^n$, $U = \mathbb{R}^m$ and*

$$f(x,u) := Ax + Bu, \quad A \in \mathbb{R}^{n \times n} \textit{ invertible}, \ B \in \mathbb{R}^{n \times m}.$$

*Then $\Sigma$ is controllable if and only if $\Sigma$ is reachable from one point.*

See for example [AM06], Theorem 2.22, for a proof. In the sequel we list other types of systems where reachability from one point implies controllability.

**Theorem 2.39** *Let $\Sigma = (M, U, f)$ be an invertible system.*

a) *Assume that $\Sigma$ is weakly reversible. Then $\Sigma$ is controllable if and only if $\Sigma$ is reachable from one point.*

b) *Assume that $\Sigma$ is abelian. Then the following statements are equivalent.*

  (i) *$S_\Sigma = G_\Sigma$ and $G_\Sigma \cdot x = M$ for some $x \in M$.*

  (i) *$\Sigma$ is controllable*

  (ii) *$\Sigma$ is reachable from one point.*

**Proof.** a) The claim follows immediately from Proposition 2.35. If $\Sigma$ is weakly reversible, $\mathcal{R}(x) = G_\Sigma \cdot x$ for all $x \in M$. Hence, $\mathcal{R}(x) = M$ implies $\mathcal{R}(y) = M$ for all $y \in G_\Sigma \cdot x = M$ and therefore controllability (see Proposition 2.31).

b) Obviously, $(i)$ implies $(ii)$ and $(ii)$ implies $(iii)$. Now we assume that $\mathcal{R}(x) = M$ for one $x \in M$. Then for any $g \in G_\Sigma$ there exists $s \in S_\Sigma$ such that

$$s \cdot x = g \cdot x. \tag{23}$$

Moreover, for any $y \in M$ there exists $s_y \in S_\Sigma$ such that $y = s_y \cdot x$. Therefore, (23) implies $ss_y^{-1} \cdot y = gs_y^{-1} \cdot y$ for all $y \in M$ and thus $s_y^{-1} s \cdot y = s_y^{-1} g \cdot y$ for all $y \in M$. It follows that the maps $s$ and $g$ are identical and in particular $g \in S_\Sigma$. Hence, $S_\Sigma = G_\Sigma$. □

In many applications the system group is equipped with a Lie group structure and is therefore a topological group. In the following two theorems we apply certain results of the theory of topological semigroups to our situation.

**Theorem 2.40** *Let $\Sigma = (M, U, f)$ be an invertible system, where $G_\Sigma$ is a topological group. If $f_u^{-1} \in \overline{S_\Sigma}$ for all $u \in U$ and $\operatorname{int}_{G_\Sigma} S_\Sigma \neq \emptyset$, then*

$$S_\Sigma = G_\Sigma.$$

*In this case, $\Sigma$ is controllable if and only if $\Sigma$ is reachable from one point.*

**Proof.** Let $f_u^{-1} \in \overline{S_\Sigma}$ for all $u \in U$. This implies $(f_{u_1} \circ \cdots \circ f_{u_T})^{-1} \in \overline{S_\Sigma}$ for any $T \in \mathbb{N}$ and $u_1, \ldots, u_T \in U$, since $\overline{S_\Sigma}$ is a semigroup (see Lemma B.2). In other words, $\overline{S_\Sigma} = G_\Sigma$ and therefore, by Lemma B.6, $S_\Sigma = G_\Sigma$. □

In fact, Theorem 2.39,b and Theorem 2.40 provide conditions for the equality $S_\Sigma = G_\Sigma$ and therefore for controllability on orbits. However, we have seen, that controllability does not necessarily imply $S_\Sigma = G_\Sigma$ (see Example 2.32). The following result deals with such situations.

**Theorem 2.41** *Let $\Sigma = (M, U, f)$ be an invertible system on a connected manifold $M$. Assume that $G_\Sigma$ is a Lie group such that the action $(g, x) \mapsto g \cdot x$ is continuous. Moreover, assume that $\Sigma$ is reachable from one point. Then $\Sigma$ is controllable if and only if*

$$\mathrm{int}_{G_\Sigma}(S_\Sigma) \cap \mathrm{Stab}_x \neq \emptyset \quad \textit{for all } x \in M. \tag{24}$$

**Proof.** If $\Sigma$ is reachable from one point $x \in M$, then the inclusion

$$M = S_\Sigma \cdot x \subseteq G_\Sigma \cdot x \subseteq M$$

implies that $G_\Sigma$ acts transitively on $M$. Now the claim follows from a result about semigroup actions on manifolds. If a Lie group $G$ acts transitively on a connected manifold $M$, then a subsemigroup $S \subseteq G$ acts transitively on $M$ if for any $x \in M$ there exists $s \in \mathrm{int}_G S$ such that $s \cdot x = x$. (see Lemma B.7).

Conversely, assuming that $\Sigma$ is controllable, we obtain $\mathrm{int}_{G_\Sigma} S_\Sigma \neq \emptyset$ by Lemma 2.22. Now for any $x \in M$ and $s \in \mathrm{int}_{G_\Sigma} S_\Sigma$ there exists $\tilde{s} \in S_\Sigma$ such that $\tilde{s} \cdot (s \cdot x) = x$. It follows that $\tilde{s}s \in \mathrm{Stab}_x$. Moreover, $\tilde{s}s \in \mathrm{int}_{G_\Sigma} S_\Sigma$, since $S_\Sigma \,\mathrm{int}_{G_\Sigma} S_\Sigma \subseteq \mathrm{int}_{G_\Sigma} S_\Sigma$ (see Lemma B.5). Hence $\mathrm{int}_{G_\Sigma}(S_\Sigma) \cap \mathrm{Stab}_x \neq \emptyset$. □

In this section we have shown sufficient conditions for which reachability from one point implies controllability. The following example illustrates, that in general, reachability from one point is not sufficient for controllability.

**Example 2.42** Let $\Sigma = (M, U, f)$ be given by $M = U = \mathbb{R}$ and

$$f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}; \quad (x, u) \mapsto (2|u| + 1 - u)x + u.$$

Note that $\Sigma$ is invertible and smooth, i.e., $f_u$ is a diffeomorphism for any $u \in U$. We show that $\Sigma$ is reachable from one point, but not controllable.

Obviously, $\mathcal{R}^1(0) = \mathcal{R}(0) = \mathbb{R}$, since $f_u(0) = u$. Hence, $\Sigma$ is reachable from 0. For all $x \geq 1$ we have

$$f_u(x) = x + \underbrace{u + |u|x}_{\geq 0} + \underbrace{(|u| - u)x}_{\geq 0} \geq x.$$

Therefore, $f_{u_0} \circ \cdots \circ f_{u_{T-1}}(1) \geq 1$ for all $u_0, \ldots, u_{T-1} \in U$. It follows that $\mathcal{R}(1) \subseteq [1, \infty)$. Hence, $\Sigma$ is indeed not controllable.

## 2.4 Approximatively reachable systems

In many applications it is impossible to reach desired points in finitely many steps. On the other hand it is the very nature of some algorithms to converge to desired points without reaching them exactly. Therefore, the topological closures of reachable sets are of interest. In particular we are interested if there exists a point $x$, such that every other point can be reached approximatively from $x$.

**Definition 2.43** A system $\Sigma$ is *approximatively reachable from* $x$ if any state $y \in M$ can be reached arbitrarily close from $x$, i.e.,

$$\overline{\mathcal{R}(x)} = M.$$

Whether a desired state can be reached approximatively or not depends on the choice of the initial state, which is often chosen randomly. Therefore one wants to find conditions, under which it is possible to reach any state approximatively from "almost all" initial states. This yields the following definition.

**Definition 2.44** We say a subset $N \subseteq M$ of a topological space $M$ is a *generic subset* of $M$, if $\overline{\operatorname{int}(N)} = M$. A system $\Sigma = (M, U, f)$ is said to be *densely reachable* if there exists a generic subset $N \subseteq M$ such that $\Sigma$ is approximatively reachable from any $x \in N$.

In the following we show properties of abelian invertible systems which are approximatively reachable. Afterwards we show sufficient conditions for dense reachability.

### 2.4.1 Approximative reachability

Let $\Sigma = (M, U, f)$ be a system and $\mathcal{E} \subseteq M$. Obviously, the existence of a shift strategy $u \in U^{\mathbb{N}_0}$ such that $x \xrightarrow{u} \mathcal{E}$ implies

$$\overline{\mathcal{R}(x)} \cap \mathcal{E} \neq \emptyset.$$

The following example shows that the converse is not true in general, i.e., $y \in \overline{\mathcal{R}(x)}$ does not necessarily imply the existence of $u \in \mathbb{R}^{\mathbb{N}}$ (or $u \in \mathbb{R}^N, N \in \mathbb{N}$) such that $x \xrightarrow{u} y$.

**Example 2.45** Let $\Sigma = (\mathbb{R}, U, f)$ be given by $U = \mathbb{R}^+$ and

$$f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R};\ (x, u) = x + u.$$

Note that $\Sigma$ is a smooth invertible system and $\mathcal{R}(x_0) = (x_0, \infty)$ for all $x_0 \in \mathbb{R}$. It follows

$$\overline{\mathcal{R}(x_0)} \cap \{x_0\} \neq \emptyset.$$

Nevertheless, choosing any first control $u_0 \in U$ the reachable set of $x_1 = f(x_0, u_0)$ is $\mathcal{R}(x_1) = (x_0 + u_1, \infty)$. For any further controls we have $x_t \in \mathcal{R}(x_1)$ for all $t \in \mathbb{N} \setminus \{1\}$. Hence, $(x_t)_{t\in\mathbb{N}}$ does not converge to $x_0$.

Nevertheless, the following result shows that approximative reachability from $x \in M$ implies that the sequence $x_{t+1} = f(x_t, u_t)$ can be steered arbitrary close to any $y \in \partial\mathcal{R}(x) := \overline{\mathcal{R}(x)} \setminus \mathcal{R}(x)$, provided $S_\Sigma$ is abelian.

**Theorem 2.46** *Let $\Sigma = (M, U, f)$ be an invertible system with abelian system semigroup. Moreover, let $\Sigma$ be approximatively reachable from $x \in M$. Then*

*a) There exists $N \subseteq M$ with $\overline{N} = M$ such that $\Sigma$ is approximatively reachable from all $y \in N$. In particular, $\Sigma$ is approximatively reachable from $y \in G_\Sigma \cdot x$.*

*b) For any $y \in M \setminus \mathcal{R}(x)$ and any open neighborhood $\mathcal{U}$ of $y$ there exists a control sequence $u_1, \ldots, u_N$, $n \in \mathbb{N}$ such that $x_n \in \mathcal{U}$.*

*c) $\Sigma$ is controllable on $\mathcal{R}(x)$ if and only if $S_\Sigma = G_\Sigma$.*

**Proof.** a) If $\overline{\mathcal{R}(x)} = M$, then $N := G_\Sigma \cdot x \supseteq S_\Sigma \cdot x = \mathcal{R}(x)$ is dense in $M$. Moreover, for any $y = g \cdot x \in G_\Sigma \cdot x$ we obtain

$$M \supseteq \overline{\mathcal{R}(y)} = \overline{S_\Sigma \cdot y} = \overline{S_\Sigma g \cdot x} = \overline{g(S_\Sigma \cdot x)} \supseteq g(\overline{S_\Sigma \cdot x}) = g(M) = M,$$

since $g \in G_\Sigma$ is bijective and continuous.
b) Let $(\mathcal{U}_n)_{n\in\mathbb{N}}$ be a sequence of neighborhoods of $y \in M \setminus \mathcal{R}(x)$ such that $\mathcal{U}_{n+1} \subseteq \mathcal{U}_n$ and $\bigcap_{n=1}^{\infty} \mathcal{U}_n = \{y\}$. Since $\Sigma$ is approximatively reachable from $x$, we can choose $u_1, \ldots, u_{T_1} \in U$ such that $x_{T_1} := f_{u_{T_1}} \circ \cdots \circ f_{u_1}(x)$ lies in $\mathcal{U}_1$. From a) we deduce that $\Sigma$ is approximatively reachable from $x_{T_1}$, since $x_{T_1} \in G_\Sigma \cdot x$. Therefore, we can choose $u_{T_1+1}, \ldots, u_{T_2} \in U$ such that $x_{T_2} := f_{u_{T_1+1}} \circ \cdots \circ f_{u_{T_2}}(x_{T_2}) \in \mathcal{U}_2$. By induction, it follows that for any $\mathcal{U}_n$ there exist controls $u_{T_{n-1}+1}, \ldots, u_{T_n}$ such that

$$\begin{aligned} x_{T_n} &= f_{u_{T_n}} \circ \cdots \circ f_{u_{T_{n-1}+1}}(x_{T_{n-1}}) \\ &= f_{u_{T_n}} \circ \cdots \circ f_{u_{T_{n-1}+1}} \circ f_{u_{T_{n-1}}} \circ \cdots \circ f_{u_1}(x) \\ &\in \mathcal{U}_n. \end{aligned}$$

c) Obviously, $G_\Sigma = S_\Sigma$ implies controllability on $\mathcal{R}(x)$, since for all $y = s \cdot x \in \mathcal{R}(x)$ we obtain

$$x = s^{-1} \cdot y \in G_\Sigma \cdot y = \mathcal{R}(y).$$

Note that this conclusion remains true if $S_\Sigma$ is non-abelian.

Now we assume that $\Sigma$ is controllable on $\mathcal{R}(x)$. Since $G_\Sigma = \langle S_\Sigma \rangle$ is abelian, every element of $G_\Sigma$ can be decomposed in the form $g = s_1^{-1}s_2$ with $s_1, s_2 \in S_\Sigma$. For $s_1 \cdot x, s_2 \cdot x \in \mathcal{R}(x)$ there exists $s \in S$ such that $ss_1 \cdot x = s_2 \cdot x$. It follows that $s \cdot x = g \cdot x$ and $s\tilde{g} \cdot x = g\tilde{g} \cdot x$ for all $\tilde{g} \in G_\Sigma$. Since $G_\Sigma$ acts transitively on $G_\Sigma \cdot x$, we obtain $g \cdot y = s \cdot y$ for all $y \in G_\Sigma \cdot x$. Therefore, the continuous maps $g_{|G_\Sigma \cdot x}$ and $s_{|G_\Sigma \cdot x}$ are identical. Since $\overline{G_\Sigma \cdot x} = M$, we obtain $g = s$.

□

In Theorem 2.39 we have seen that for systems with abelian system semigroup reachability from *one* point already implies $S_\Sigma = G_\Sigma$. The following example illustrates that approximative reachability from *every* point does not imply $S_\Sigma = G_\Sigma$, even if $S_\Sigma$ is abelian.

**Example 2.47** Let $\Sigma = (\mathbb{T}, U, f)$ be a system on the torus $\mathbb{T} := \mathbb{S} \times \mathbb{S}$ given by $U = \mathbb{R}^+$ and

$$f : \mathbb{T} \times U \to \mathbb{T},\ ((x_1, x_2), u) = (e^{iu}x_1, e^{i\sqrt{2}u}x_2).$$

Note that $\mathbb{T}$ is a topological group and therefore

$$\Phi_g : \mathbb{T} \to \mathbb{T}, x \mapsto gx,\ g \in \mathbb{T}$$

is a homeomorphism. We shall show that $S_\Sigma \neq G_\Sigma$ and that $\Sigma$ is approximatively reachable from every point $x \in \mathbb{T}$.

For all $u_1, u_2 \in U$ we have $f_{u_1} \circ f_{u_2} = f_{u_1+u_2}$ and therefore

$$S_\Sigma = \{f_u \,|\, u \in \mathbb{R}^+\}.$$

Moreover, $\mathrm{id}_\mathbb{T} \notin S_\Sigma$ because $f_u = \mathrm{id}_\mathbb{T}$ implies

$$u = 2k_1\pi = \frac{1}{\sqrt{2}}k_2\pi,\ k_1, k_2 \in \mathbb{Z},$$

which contradicts $u \in \mathbb{R}^+$. Hence, $S_\Sigma \neq G_\Sigma$.

Now we show that $\overline{\mathcal{R}(x)} = \mathbb{T}$ for all $x \in \mathbb{T}$. In fact it is sufficient to show that $\overline{\mathcal{R}(x)} = \mathbb{T}$ for one $x \in \mathbb{T}$, since $\overline{\mathcal{R}(x)} = \mathbb{T}$ implies

$$\mathbb{T} = \Phi_{yx^{-1}}\left(\overline{\mathcal{R}(x)}\right) = \overline{\Phi_{yx^{-1}}\left(\mathcal{R}(x)\right)} = \overline{yx^{-1}S_\Sigma \cdot x} = \overline{S_\Sigma \cdot x} = \overline{\mathcal{R}(y)}$$

for all $y, x \in \mathbb{T}$. It is well known that the set

$$G_\Sigma = \{(e^{iu}, e^{i\sqrt{2}u}) \,|\, u \in \mathbb{R}\}$$

is a dense subgroup of the torus (see [HN91], Proposition I.3.13). Since $\mathbb{T}$ is compact, $\overline{S_\Sigma} \subseteq \mathbb{T}$ is compact. Recall, that the closure of a subsemigroup of a topological group is a semigroup (see Lemma B.4) and that a compact subsemigroup of a topological group is a group (see Lemma B.2). It follows, that $\overline{S_\Sigma}$ is a group, and therefore $s^{-1} \in \overline{S_\Sigma}$ for all $s \in S_\Sigma$. We conclude that $G_\Sigma \subseteq \overline{S_\Sigma}$ and

$$\overline{S_\Sigma \cdot e} = \overline{S_\Sigma} = \overline{G_\Sigma} = \mathbb{T}.$$

### 2.4.2 Dense reachability

In general, approximative reachability does not imply dense reachability (see Example 2.50). Nevertheless, for abelian systems we obtain the following:

**Theorem 2.48** *Let $\Sigma = (M, U, f)$ be an aberlian invertible system and $x, y \in M$ such that $\Sigma$ is accessible from $x$ and approximatively reachable from $y$. Assume, that the system group is a Lie group such that $G_\Sigma \times M \to M$ is continuous. Then:*

*a) If $\Sigma$ is abelian then $\Sigma$ is densely reachable.*

*b) $S_\Sigma = G_\Sigma$,*

*c) $\Sigma$ is controllable on $G_\Sigma \cdot x$,*

**Proof.** a) By Proposition 2.20 the orbit $G_\Sigma \cdot x$ is an open subset of $M$. Moreover, for all $y \in G_\Sigma \cdot x$ we have $\overline{\mathcal{R}(y)} = M$ (by Theorem 2.46). Since $\mathcal{R}(y) \subseteq G_\Sigma \cdot x \subseteq M$, it follows $\overline{G_\Sigma \cdot x} = M$. Hence, $G_\Sigma \cdot x$ is a generic subset and $\Sigma$ is densely reachable. b) Since $G_\Sigma \times M \to M$ is continuous, accessibility from $x$ implies

$$\mathrm{int}_{G_\Sigma} S_\Sigma \neq \emptyset$$

by Lemma 2.22. Now we show that $f_u^{-1} \in \overline{S_\Sigma}$ for all $u \in U$ and thus, $G_\Sigma = S_\Sigma$ by Theorem 2.40. $G_\Sigma \cdot x$ is open by Proposition 2.20 and thus locally compact. The Lie group $G_\Sigma$ acts transitively on $G_\Sigma \cdot x$. Following Theorem B.8, the map $h_x : G_\Sigma \to G_\Sigma \cdot x$, $g \mapsto g \cdot x$ is open. It follows that

$$(G_\Sigma \setminus \overline{S_\Sigma}) \cdot x = h_x(G_\Sigma \setminus \overline{S_\Sigma})$$

is open in $G_\Sigma \cdot x$ and, by Proposition 2.20, open in $M$. Recall that $\overline{\mathcal{R}(x)} = M$. Assuming $(G_\Sigma \setminus \overline{S_\Sigma}) \cdot x \neq \emptyset$, we obtain

$$(G_\Sigma \setminus \overline{S_\Sigma}) \cdot x \cap \mathcal{R}(x) \neq \emptyset.$$

Thus, there exists $g \in (G_\Sigma \setminus \overline{S_\Sigma})$ and $s \in S_\Sigma$ such that

$$g \cdot x = s \cdot x.$$

Since $G_\Sigma$ is abelian and acts transitively on $G_\Sigma \cdot x$, we have $g \cdot y = s \cdot x$ for all $y \in G_\Sigma \cdot x$. Therefore, the continuous maps $s|_{G_\Sigma \cdot x}$ and $g|_{G_\Sigma \cdot x}$ are identical. Since $\overline{G_\Sigma \cdot x} = M$, we obtain $s = g$ which is a contradiction to $g \in G_\Sigma \setminus \overline{S_\Sigma}$. Thus, $G_\Sigma \setminus \overline{S_\Sigma} = \emptyset$ and therefore $f_u^{-1} \in \overline{S_\Sigma}$ for any $u \in U$.
c) The claim follows immediately by b) and by Proposition 2.31. □

Theorem 2.48 implies the following characterization of dense reachability. This result will be essential for our analysis of Inverse Iteration systems.

**Corollary 2.49** *Let $\Sigma = (M, U, f)$ be an abelian invertible system and $x \in M$. Assume, that the system group is a Lie group acting continuously and transitively on $M$. Moreover, we assume $\operatorname{int}_{G_\Sigma} S_\Sigma \neq \emptyset$. Then the following statements are equivalent:*

*(i) $\Sigma$ is approximatively reachable from some $x \in M$,*

*(ii) $\Sigma$ is densely reachable,*

*(iii) $S_\Sigma = G_\Sigma$,*

*(iv) $\Sigma$ is controllable on $M$.*

**Proof.** Obviously, $(iv)$ implies $(i)$. Moreover, $(iii)$ implies $(iv)$, since $G_\Sigma$ acts transitively on $M$. By Theorem 2.46, $(i)$ implies $\overline{\mathcal{R}(y)} = M$ for all $y \in G_\Sigma \cdot x = M$ and therefore $(ii)$. Now we assume $(ii)$. Again, $h_x : G_\Sigma \to M$, $g \mapsto g \cdot x$ is open for all $x \in M$ since $G_\Sigma$ acts transitively (see Theorem B.8). Therefore, $\Sigma$ is accessible by Proposition 2.22. Thus, all conditions for Theorem 2.48 are fulfilled. In particular it follows that $S_\Sigma = G_\Sigma$. □

The following example shows that none of the claims of Theorem 2.48 remains true if we drop the assumption that $S_\Sigma$ is abelian.

**Example 2.50** Consider $\Sigma = (M, U, f)$ of example 2.18, i.e., $M = \mathbb{R}^2 \setminus \{0\}$,

$$U = \left\{ \begin{pmatrix} u_1 & u_2 \\ u_3 & u_4 \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \,\middle|\, u_i > 0, i = 1, \ldots, 4 \right\}$$

and $f : M \times U \to M$, $(x, U) \mapsto Ux$. Note that $\Sigma$ is smoothly invertible, i.e., $f_u$ is a diffeomorphism for all $u \in U$. We show that $\Sigma$ is reachable and accessible from $x_0 := (1, -1)^\top$. On the other hand $S_\Sigma \neq G_\Sigma$ and $\Sigma$ is neither densly reachable nor controllable on $G_\Sigma \cdot x_0$.

Obviously, $U$ is closed under matrix multiplication. Therefore, $S_\Sigma$ can be identified with $U$. This already shows $S_\Sigma \neq G_\Sigma$, since $U$ is not a group. Moreover we deduce $\operatorname{int}_{G_\Sigma} S_\Sigma \neq \emptyset$, since $G_\Sigma \subseteq \mathrm{GL}_2(\mathbb{R})$ and $\operatorname{int}_{\mathrm{GL}_2(\mathbb{R})} S_\Sigma \neq \emptyset$. For $x_0 := (1, -1)^\top$ we obtain

$$S_\Sigma \cdot x_0 = \left\{ \begin{pmatrix} u_1 - u_2 \\ u_3 - u_4 \end{pmatrix} \,\middle|\, \begin{array}{c} u_i > 0, i = 1, \ldots, 4, \\ u_1 u_4 \neq u_2 u_3 \end{array} \right\}.$$

For any $(a, b)^\top \in \mathbb{R}^2 \setminus \{0\}$ we choose $\mu, \lambda \in \mathbb{R}$ such that

$$a\mu \neq \lambda b, \quad \lambda > |a| \quad \text{and} \quad \mu > |b|.$$

Then the control parameters

$$u_1 := \lambda + a, \quad u_2 := \lambda, \quad u_3 := \mu + b, \quad u_4 := \mu$$

are all positive and have the property

$$u_1 u_4 - u_2 u_3 = a\mu - \lambda b \neq 0.$$

Moreover,

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} u_1 - u_2 \\ u_3 - u_4 \end{pmatrix} \in S_\Sigma \cdot x_0.$$

This shows that $M = S_\Sigma \cdot x_0 = G_\Sigma \cdot x_0$. Therefore, $\Sigma$ is reachable and accessible from $x_0$.

On the other hand, for any $(a,b)^\top \in M$ with $a \geq 0$, $b \geq 0$ we obtain

$$S_\Sigma \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \left\{ \begin{pmatrix} au_1 + bu_2 \\ au_3 + bu_4 \end{pmatrix} \middle| \, u_1, u_2, u_3, u_4 > 0 \right\} \subseteq \mathbb{R}^+ \times \mathbb{R}^+$$

This shows that $\Sigma$ is neither controllable on $G_\Sigma \cdot x_0 = M$ nor densely reachable.

# 3 Structure theory for subsystems

In many situations, two systems $\Sigma$, $\tilde{\Sigma}$ are related via a map between the state spaces, that preserves crucial parts of the system structure. One important example is the inverse iteration system on flag manifolds and inverse iteration on Hessenberg varieties (see Section 6.8). If the structure of reachable sets of $\Sigma$ is analyzed, one can exploit this information for the analysis of $\tilde{\Sigma}$.

In this chapter we develop a structure theory for such situations. In particular we analyze *induced systems* in Section 3.1 and *restricted systems* in Section 3.2. The results in this chapter are probably not entirely unknown. However, to the best of the authors knowledge, a systematic development of a structure theory for subsystems in terms of system semigroups and system groups is unknown.

## 3.1 Induced systems

**Definition 3.1** Let $\Sigma = (M, U, f)$ and $\tilde{\Sigma} = (\tilde{M}, U, \tilde{f})$ be invertible systems, and $\pi : M \to \tilde{M}$ be a surjective, continuous and open map. We say that $\tilde{\Sigma}$ is an *induced system of* $\Sigma$ *with respect to* $\pi$ if

$$\pi \circ f_u = \tilde{f}_u \circ \pi$$

for all $u \in U$. We say that $\Sigma$ and $\tilde{\Sigma}$ are *isomorphic systems* if $\pi$ is a homeomorphism.

### 3.1.1 Reachable sets of induced systems

The following lemma shows, that the system groups of $\Sigma$ and the system group of $\tilde{\Sigma}$ are closely related.

**Lemma 3.2** *Let $\tilde{\Sigma}$ be an induced system of $\Sigma$ with respect to $\pi$.*

*a) For all $g \in G_\Sigma$, there exists a unique $\tilde{g} \in G_{\tilde{\Sigma}}$ such that*

$$\pi \circ g = \tilde{g} \circ \pi.$$

*b) For all $\tilde{g} \in G_{\tilde{\Sigma}}$ there exists $g \in G_\Sigma$ such that*

$$\pi \circ g = \tilde{g} \circ \pi.$$

**Proof.** Since $G_\Sigma = \langle S_\Sigma \rangle$, every element of $G_\Sigma$ can be written as a product

$$g = f_{u_T}^{\epsilon_T} \cdots f_{u_0}^{\epsilon_0} \tag{25}$$

with $T \in \mathbb{N}$, $\epsilon_k \in \{-1,1\}$ and $u_k \in U$ for $k = 0, \ldots, T$. Analogously, every element in $G_{\tilde{\Sigma}}$ can be written in the form

$$\tilde{g} = \tilde{f}_{\tilde{u}_{\tilde{T}}}^{\tilde{\epsilon}_{\tilde{T}}} \cdots \tilde{f}_{\tilde{u}_0}^{\tilde{\epsilon}_0} \in G_{\tilde{\Sigma}} \tag{26}$$

with $\tilde{T} \in \mathbb{N}$, $\tilde{\epsilon}_k \in \{-1,1\}$ and $\tilde{u}_k \in U$ for $k = 0, \ldots, \tilde{T}$. We show that

$$\pi \circ g = \tilde{g} \circ \pi \tag{27}$$

if $T = \tilde{T}$ and $u_k = \tilde{u}_k$ for $k = 0, \ldots, \tilde{T}$.

By assumption we obtain $\pi \circ g = \tilde{g} \circ \pi$ and therefore $\tilde{g}^{-1} \circ \pi = \pi \circ g^{-1}$ for $g = f_u$ with $u \in U$. Moreover, if $\pi \circ g_i = \tilde{g}_i \circ \pi$ holds for $g_1, g_2 \in G_\Sigma$, then

$$\pi \circ g_1 g_2 = \tilde{g}_1 \circ \pi \circ g_2 = \tilde{g}_1 \tilde{g}_2 \circ \pi.$$

By induction, Equation (27) follows for any product of elements $f_u^\epsilon$, $u \in U$, $\epsilon \in \{-1,1\}$, and therefore for all $g \in G_\Sigma$.

Moreover, $\tilde{g} \circ \pi = \tilde{h} \circ \pi$ for $\tilde{h} \in G_{\tilde{\Sigma}}$ implies $\tilde{g} = \tilde{h}$ since $\pi$ is surjective. Hence, $\tilde{g}$ of statement a) is unique. □

Note that the decompositions in (25) and (26) are not unique in general. Therefore, we cannot expect uniqueness in Part b) of Lemma 3.2, i.e., $\tilde{g} \circ \pi = \pi \circ g_1 = \pi \circ g_2$ does not imply $g_1 = g_2$.

**Lemma 3.3** *Let $\tilde{\Sigma}$ be an induced system of $\Sigma$ with respect to $\pi : M \to \tilde{M}$. For all $x \in M$ we have*

*(i)* $\pi(G_\Sigma \cdot x) = G_{\tilde{\Sigma}} \cdot \pi(x)$

*(ii)* $\pi(\mathcal{R}_\Sigma(x)) = \mathcal{R}_{\tilde{\Sigma}}(\pi(x))$

*(iii)* $\pi(\overline{\mathcal{R}_\Sigma(x)}) \subseteq \overline{\mathcal{R}_{\tilde{\Sigma}}(\pi(x))}$

*(iv)* $\pi(\overline{\mathcal{R}_\Sigma(x)}) = \overline{\mathcal{R}_{\tilde{\Sigma}}(\pi(x))}$, *provided $M$ is compact.*

**Proof.** By Lemma 3.2 it follows

$$\pi(G_\Sigma \cdot x) = \{\pi(g \cdot x) \,|\, g \in G_\Sigma\} = \{\tilde{g} \cdot \pi(x) \,|\, \tilde{g} \in G_{\tilde{\Sigma}}\} = G_{\tilde{\Sigma}} \cdot \pi(x)$$

and

$$\pi(\mathcal{R}_\Sigma(x)) = \{\pi(s \cdot x) \,|\, s \in S_\Sigma\} = \{\tilde{s} \cdot \pi(x) \,|\, \tilde{s} \in S_{\tilde{\Sigma}}\} = S_{\tilde{\Sigma}} \cdot \pi(x) = \mathcal{R}_{\tilde{\Sigma}}(\pi(x)).$$

Moreover, we obtain

$$\pi(\overline{\mathcal{R}_\Sigma(x)}) \subseteq \overline{\pi(\mathcal{R}_\Sigma(x))} = \overline{\mathcal{R}_{\tilde{\Sigma}}(\pi(x))}$$

since $\pi$ is continuous. If $M$ is compact, then $\overline{\mathcal{R}_\Sigma(x)}$ is compact. Therefore, $\overline{\mathcal{R}_{\tilde{\Sigma}}(\pi(x))}$ is closed. It follows $\overline{\mathcal{R}_{\tilde{\Sigma}}(\pi(x))} \subseteq \pi(\overline{\mathcal{R}_\Sigma(x)})$. □

Now we can easily show basic relations between $\Sigma$ and $\tilde{\Sigma}$ concerning controllability, accessibility, weak reversibility, approximative reachability and dense reachability.

**Theorem 3.4** *Let $\tilde{\Sigma}$ be an induced system of $\Sigma$ with respect to $\pi : M \to \tilde{M}$.*

*a) If $\Sigma$ is reachable from $x \in M$, then $\tilde{\Sigma}$ is reachable from $\pi(x) \in \tilde{M}$.*

*b) If $\Sigma$ is controllable, then $\tilde{\Sigma}$ is controllable.*

*c) If $\Sigma$ is accessible from $x \in M$, then $\tilde{\Sigma}$ is accessible from $\pi(x) \in \tilde{M}$.*

*d) If $\Sigma$ is weakly reversible, then $\tilde{\Sigma}$ is weakly reversible.*

*e) If $\Sigma$ is approximatively reachable from $x \in M$, then $\tilde{\Sigma}$ is approximatively reachable from $\pi(x) \in \tilde{M}$.*

*f) If $\Sigma$ is densely reachable, then $\tilde{\Sigma}$ is densely reachable.*

**Proof.** a) If $\Sigma$ is reachable from $x \in M$, i.e. then $\mathcal{R}_\Sigma(x) = M$, then Lemma 3.3 implies

$$\mathcal{R}_{\tilde{\Sigma}}(\pi(x)) = \pi(\mathcal{R}_\Sigma(x)) = \pi(M) = \tilde{M}$$

since $\pi$ is surjective. Hence, $\tilde{\Sigma}$ is reachable from $\pi(x) \in \tilde{M}$.
b) By Proposition 2.31 a system is controllable if and only if it is reachable from all $x \in M$ (from all $\tilde{x} \in \tilde{M}$). Therefore, the claim follows from a).
c) By Lemma 3.3 it is

$$\begin{aligned} \operatorname{int}_{\tilde{M}} \mathcal{R}_{\tilde{\Sigma}}(\pi(x)) &= \operatorname{int}_{\tilde{M}} \pi(\mathcal{R}_\Sigma(x)) \\ &\supseteq \operatorname{int}_{\tilde{M}} \pi(\operatorname{int}_M(\mathcal{R}_\Sigma(x))) \\ &= \pi(\operatorname{int}_M(\mathcal{R}_\Sigma(x))) \end{aligned}$$

since $\pi$ is an open map. Therefore, $\operatorname{int}_M(\mathcal{R}_\Sigma(x)) \neq \emptyset$ implies

$$\operatorname{int}_{\tilde{M}} \mathcal{R}_{\tilde{\Sigma}}(\pi(x)) \neq \emptyset.$$

d) By Lemma 2.35, $\Sigma$ is weakly reversible if and only if $G_\Sigma \cdot x = \mathcal{R}(x)$ for all $x \in M$. This implies $G_{\tilde{\Sigma}} \cdot \pi(x) = \mathcal{R}_{\tilde{\Sigma}}(\pi(x))$, by Lemma 3.3. Hence, $\tilde{\Sigma}$ is weakly reversible.
e) By Lemma 3.3, $\overline{\mathcal{R}_\Sigma(x)} = M$ implies

$$\overline{\mathcal{R}_{\tilde{\Sigma}}(\pi(x))} \supseteq \pi(\overline{\mathcal{R}_\Sigma(x)}) = \pi(M).$$

Therefore, $\tilde{\Sigma}$ is approximatively reachable from $\pi(x) \in \tilde{M}$, since $\pi(M) = \tilde{M}$.
f) If $N$ is generic in $M$, i.e., $\overline{\operatorname{int} N} = M$, then $\pi(N)$ is generic in $\tilde{M}$, since $\pi$ is open, continuous and surjective. Thus,

$$\tilde{M} \supseteq \overline{\operatorname{int} \pi(N)} \supseteq \overline{\pi(\operatorname{int} N)} \supseteq \pi(\overline{\operatorname{int} N}) = \pi(M) = \tilde{M}.$$

From d) we conclude, that $\tilde{\Sigma}$ is approximatively reachable from all $\tilde{x} \in \pi(N)$. Hence, $\tilde{\Sigma}$ is densely reachable. $\square$

### 3.1.2 Relation between $S_\Sigma$ and $S_{\tilde{\Sigma}}$

In the following we analyze the relation between the system semigroup $S_\Sigma$ of system $\Sigma$ and the system semigroup $S_{\tilde{\Sigma}}$ of the induced system $\tilde{\Sigma}$. We define the *core of* $\pi$

$$C_\pi := \{g \in G_\Sigma \,|\, \pi(g \cdot x) = \pi(x),\ \forall x \in M\}. \tag{28}$$

In particular $C_\pi = \{\mathrm{id_M}\}$ if $\Sigma$ and $\tilde{\Sigma}$ are isomorphic, since $g \cdot x = x$ for all $x \in M$ implies $g = \mathrm{id}_M$. In general the core $C_\pi$ has the following useful properties.

**Lemma 3.5** *Let $\tilde{\Sigma}$ be an induced system of $\Sigma$ with respect to $\pi$.*

*a) $C_\pi$ is a normal subgroup of $G_\Sigma$.*

*b) If $G_\Sigma$ is a topological group, such that $h_x : G_\Sigma \to M$, $g \mapsto g \cdot x$ is continuous for all $x \in M$, then $C_\pi$ is a closed subgroup of $G_\Sigma$. In particular, $C_\pi$ is a Lie subgroup of $G_\Sigma$, provided $G_\Sigma$ is a Lie group.*

**Proof.** a) If $f, g \in C_\pi$ then $g^{-1}$ and $fg$ are elements of $C_\pi$, since

$$\pi(x) = \pi(g \cdot (g^{-1} \cdot x)) = \pi(g^{-1} \cdot x)$$

and

$$\pi(fg \cdot x) = \pi(f \cdot (g \cdot x)) = \pi(g \cdot x) = \pi(x).$$

Hence, $C_\pi$ is a subgroup of $G_\Sigma$. Moreover, for all $g \in G_\Sigma$, $c \in C_\pi$, $x \in M$ we obtain

$$\begin{aligned} \pi \circ gcg^{-1}(x) &= \pi \circ g(cg^{-1}(x)) \\ &= \tilde{g} \circ \pi(c(g^{-1} \cdot x)) \\ &= \tilde{g} \circ \pi(g^{-1} \cdot x) \\ &= \pi \circ g(g^{-1} \cdot x) \\ &= \pi(x). \end{aligned}$$

This shows that $gcg^{-1} \in C_\pi$ for all $g \in G_\Sigma$, $c \in C_\pi$. Hence, $C_\pi$ is a normal subgroup of $G_\Sigma$.
b) Let $g_n$ be a sequence in $C_\pi$ with $g_n \to g \in G_\Sigma$. Since $\pi \circ h_x$ is continuous, we obtain

$$\begin{aligned} \pi(g \cdot x) &= \pi(h_x(\lim_{n\to\infty} g_n)) \\ &= \lim_{n\to\infty} \pi(h_x(g_n)) \\ &= \lim_{n\to\infty} \pi(g_n \cdot x) \\ &= \pi(x). \end{aligned}$$

Hence, $C_\pi$ is closed. If $G_\Sigma$ is a Lie group, then every closed subgroup is a Lie subgroup (see Theorem 3.6, Chapter 2 in [GOV97]). Hence, $C_\pi$ is a Lie subgroup of $G_\Sigma$. □

Since $C_\pi$ is a normal subgroup of $G_\Sigma$ we obtain

$$g_1 C_\pi g_2 C_\pi = g_1 g_2 \underbrace{g_2^{-1} C_\pi g_2}_{=C_\pi} C_\pi = g_1 g_2 C_\pi$$

for all $g_1, g_2 \in G_\Sigma$. This allows us to define a group structure (respectively a semigroup structure) on the set of cosets

$$G_\Sigma / C_\pi := \{ g C_\pi \mid g \in G_\Sigma \}$$

(respectively the set of cosets $S_\Sigma / C_\pi := \{ s C_\pi \mid s \in S_\Sigma \}$) via the product

$$g_1 C_\pi g_2 C_\pi := g_1 g_2 C_\pi \tag{29}$$

with $g_1, g_2 \in G_\Sigma$ (respectively $g_1, g_2 \in S_\Sigma$). The following theorem shows the relation between the system group of $\Sigma$ and the system group of $\tilde{\Sigma}$.

**Theorem 3.6** *Let $\tilde{\Sigma}$ be an induced system of $\Sigma$ with respect to $\pi : M \to \tilde{M}$.*

*a) $G_{\tilde{\Sigma}}$ and $G_\Sigma / C_\pi$ are isomorphic as groups.*

*b) $S_{\tilde{\Sigma}}$ and $S_\Sigma / C_\pi$ are isomorphic as semigroups.*

*a) $S_{\tilde{\Sigma}} = G_{\tilde{\Sigma}}$ if and only if $S_\Sigma C_\pi = G_\Sigma$.*

**Proof.** a) Recall that $G_\Sigma = \langle S_\Sigma \rangle$. Therefore, every $g \in G_\Sigma$ can be written in the form $g = f_{u_T}^{\epsilon_T} \dots f_{u_0}^{\epsilon_0}$ with $T \in \mathbb{N}$, $u_k \in U$, $\epsilon_k \in \{-1, 1\}$, $k = 0, \dots, T$. Moreover, $\tilde{g} = \tilde{f}_{u_T}^{\epsilon_T} \dots \tilde{f}_{u_0}^{\epsilon_0}$ is the unique element of $G_{\tilde{\Sigma}}$ such that $\pi \circ g = \tilde{g} \circ \pi$ (see Lemma 3.2). Therefore, the map

$$\Phi : G_\Sigma \to G_{\tilde{\Sigma}},\ f_{u_T}^{\epsilon_T} \dots f_{u_0}^{\epsilon_0} \mapsto \tilde{f}_{u_T}^{\epsilon_T} \dots \tilde{f}_{u_0}^{\epsilon_0} \tag{30}$$

is well defined and surjective. Moreover, $\Phi$ is a group homomorphism, since $\Phi(g_1 g_2) = \tilde{g}_1 \tilde{g}_2 = \Phi(g_1)\Phi(g_2)$ for all $g_1, g_2 \in G_\Sigma$.

Since $\pi : M \to \tilde{M}$ is surjective we obtain

$$\begin{aligned}
\mathrm{Ker}(\Phi) &:= \{ g \in G_\Sigma \mid \Phi(g) = \mathrm{id}_{\tilde{M}} \} \\
&= \{ g \in G_\Sigma \mid \tilde{g}(y) = y;\ \forall y \in \tilde{M} \} \\
&= \{ g \in G_\Sigma \mid \tilde{g} \circ \pi(x) = \pi(x);\ \forall x \in M \} \\
&= \{ g \in G_\Sigma \mid \pi(g \cdot x) = \pi(x);\ \forall x \in M \} \\
&= C_\pi.
\end{aligned}$$

By the *homomorphism theorem*,

$$\Psi : G_\Sigma/C_\pi \to G_{\tilde\Sigma}, gC_\pi \mapsto \tilde g \tag{31}$$

is an isomorphism.
b) Every $\tilde s \in S_{\tilde\Sigma}$ has a preimage $sC_\pi \in S_\Sigma/C_\pi$ such that $\Psi(sC_\pi) = \tilde s$. Moreover, $s \in S_\Sigma$ implies $\Psi(sC_\pi) = \tilde s \in S_{\tilde\Sigma}$. Therefore, $\Psi^{-1}(S_{\tilde\Sigma}) = S_\Sigma/C_\pi$. Hence, $\Psi|_{S_\Sigma/C_\pi} : S_\Sigma/C_\pi \to S_{\tilde\Sigma}$ is an isomorphism of semigroups.

c) We have $C_\pi S_\Sigma = G_\Sigma$ if and only if for all $g \in G_\Sigma$ there exists $s \in S_\Sigma$ such that $gC_\pi = sC_\pi$. In other words $G_\Sigma/C_\pi = S_\Sigma/C_\pi$ which is equivalent to $S_{\tilde\Sigma} = G_{\tilde\Sigma}$ by a) and b). □

If $G_\Sigma$ is a Lie group, $C_\pi$ is a closed subgroup of $G_\Sigma$ (see Lemma 3.5). Moreover, the quotient group $G_\Sigma/C_\pi$ carries a Lie group structure (See Theorem 3.2. in [GOV97]). Here, the open sets of $G_\Sigma/C_\pi$ are given by the projection

$$p : G_\Sigma \to G_\Sigma/C_\pi,\ g \mapsto gC_\pi, \tag{32}$$

i.e., a subset of $G_\Sigma/C_\pi$ is open if and only if its preimage is open in $G_\Sigma$. In particular, $p$ is an open map and a homomorphism of Lie groups (see Corollary 1.11.5 in [DK00]). In the following we show, that $G_{\tilde\Sigma}$ carries canonically the Lie group structure of $G_\Sigma/C_\pi$.

**Theorem 3.7** *Let $\tilde\Sigma$ be an induced system of $\Sigma$ with respect to $\pi : M \to \tilde M$. Assume that $\Sigma$ and $\tilde\Sigma$ are smoothly invertible and that $\pi$ is a submersion. Moreover, we assume that $G_\Sigma$ is a Lie group such that the action*

$$\alpha : G_\Sigma \times M \to M,\ (g, x) \mapsto g \cdot x$$

*is smooth.*

*a) $G_{\tilde\Sigma}$ carries a Lie group structure, such that $G_{\tilde\Sigma}$ and $G_\Sigma/C_\pi$ are isomorphic as Lie groups and*

$$\tilde\alpha : G_{\tilde\Sigma} \times \tilde M \to \tilde M,\ (\tilde g, \tilde x) \mapsto \tilde g \cdot \tilde x$$

*is a smooth action.*

*b) There exists a group homeomorphism $\Phi : G_\Sigma \to G_{\tilde\Sigma}$ which is open, continuous and surjective. In particular $\operatorname{int}_{G_{\tilde\Sigma}} S_{\tilde\Sigma} \neq \emptyset$ if and only if $\operatorname{int}_{G_\Sigma} S_\Sigma C_\pi \neq \emptyset$.*

**Proof.** a) Let $p : G_\Sigma \to G_\Sigma/C_\pi$ be the homomorphism of Lie groups defined in (32). By the isomorphism of groups $\Psi : G_\Sigma/C_\pi \to G_{\tilde\Sigma}$ given by

(31) we define a Lie group structure on $G_{\tilde{\Sigma}}$. We have to show, that the action $\tilde{\alpha} : G_{\tilde{\Sigma}} \times \tilde{M} \to \tilde{M}$ is smooth with respect to this Lie group structure.

The diagram

$$\begin{tikzcd} G_\Sigma \times M \arrow[r, "\alpha"] \arrow[d, "p\times \mathrm{id_M}"] & M \arrow[dd, "\pi"] \\ G_\Sigma/C_\pi \times M \arrow[d, "\Psi\times\pi"] & \\ G_{\tilde{\Sigma}} \times \tilde{M} \arrow[r, "\tilde{\alpha}"] & \tilde{M} \end{tikzcd} \tag{33}$$

commutes, since for every $(g, x) \in G_\Sigma \times M$ we have

$$\begin{aligned} \tilde{\alpha} \circ (\Psi \times \pi) \circ (p \times \mathrm{id_M})(g, x) &= \tilde{\alpha} \circ (\Psi \times \pi) \circ (gC_\pi, x) \\ &= \alpha \circ (\tilde{g}, \pi(x)) \\ &= \tilde{g} \circ \pi(x) \\ &= \pi(g \cdot x) \\ &= \pi \circ \alpha(g, x). \end{aligned}$$

Recall that the maps $\pi, \Psi$ and $\mathrm{id}_M$ are submersions. Moreover, every surjective homomorphism of Lie groups is a submersion, since it has constant rank (see Theorem 2.2 in [GOV97]). Therefore, the map

$$\Delta := (\Psi \times \pi) \circ (p \times \mathrm{id}\,|_M) : G_\Sigma \times M \to G_{\tilde{\Sigma}} \times \tilde{M}$$

is a submersion. Since $\tilde{\alpha} \circ \Delta = \alpha \circ \pi$ is smooth and $\Delta$ is a submersion, we conclude that $\tilde{\alpha}$ is smooth (see Theorem 0.5 in [DP82]).

b) Consider $\Phi := \Psi \circ p : G_\Sigma \to G_{\tilde{\Sigma}}$. Note that $\Phi$ coincides with the homomorphism defined in (30). Recall that $\mathrm{Ker}(\Phi) = C_\pi$. Therefore $\Phi(S_\Sigma C_\pi) = \Phi(S_\Sigma) = S_{\tilde{\Sigma}}$. Conversely, $\Phi^{-1}(S_{\tilde{\Sigma}}) = S_\Sigma C_\pi$, since $\Psi \circ p(g) \in S_{\tilde{\Sigma}}$ implies $g \in S_\Sigma C_\pi$. Since $\Psi$ and $p$, and therefore $\Phi$, are open and continuous it follows $\Phi(\mathrm{int}_{G_\Sigma} S_\Sigma C_\pi) = \mathrm{int}_{G_{\tilde{\Sigma}}} S_{\tilde{\Sigma}}$. In particular, $\mathrm{int}_{G_{\tilde{\Sigma}}} S_{\tilde{\Sigma}} \neq \emptyset$ if and only if $\mathrm{int}_{G_\Sigma} S_\Sigma C_\pi \neq \emptyset$. $\square$

## 3.2 Restricted systems

In many applications it is not necessary to understand the dynamic of the system on the entire state-space. Instead, the dynamic can be separated on subsets which are invariant under all elements of the system semigroup.

### 3.2.1 $\Sigma$-invariant subsets

**Definition 3.8 (Restricted systems)** Let $\Sigma = (M, U, f)$ be an invertible system. We say a subset $N \subseteq M$ is $\Sigma$-*invariant*, if $f_u(N) = N$ and for all $u \in U$. The system $\Sigma_{|N} := (N, U, f_{|N\times U})$ is called the *restricted system* with respect to the $\Sigma$-invariant subset $N$. Here, $N$ is equipped with the induced topology with respect to $M$.

Under adequate assumptions on $N$, the topological, algebraic and geometric structure of $\Sigma$ transfers to the restricted system.

**Proposition 3.9** *Let $\Sigma = (M, U, f)$ be an invertible system and $N \subseteq M$ a $\Sigma$-invariant subset. Then*

*a) $\Sigma_{|N}$ is an invertible system,*

*b) if $\Sigma$ is smoothly invertible and $N$ is a submanifold of $M$, then $\Sigma_{|N}$ is smoothly invertible,*

*c) if $\Sigma$ is algebraically invertible and $N$ is a semi-algebraic subset of $M$, then $\Sigma_{|N}$ is algebraically invertible.*

**Proof.** By definition, $f_{u|N}(N) = N$ and $f_{u|N}$ is bijective. The first two claims are obvious, since $f_{u|N}$ and $f_u^{-1}{}_{|N}$ are continuous and $f_{u|N}$ and $f_u^{-1}{}_{|N}$ are smooth, if $f_u$ is a diffeomorphism and $N$ is a submanifold. Now we assume $\Sigma$ to be algebraically invertible and $N$ to be a semi-algebraic subset of $M$. The map $\imath_{N\times U} : N \times U \to M \times U$, $(n, u) \mapsto (n, u)$ is semi-algebraic. By Proposition A.1 also $f_{|N\times U} := f \circ \imath_{N\times U}$ is a semi-algebraic map. Hence, $\Sigma_{|N}$ is algebraically invertible. □

The following observation shows, that every $\Sigma$-invariant subset can be built up by orbits of the system group.

**Proposition 3.10** *Let $\Sigma = (M, U, f)$ be an invertible system.*

*a) A subset $N \subseteq M$ is $\Sigma$-invariant if and only if $N$ is the union of system group orbits, i.e.,*

$$N = \bigcup_{x\in L} G_\Sigma \cdot x$$

*for some subset $L \subseteq N$.*

*b)* $\overline{G_\Sigma \cdot x}$ *is* $\Sigma$*-invariant for all* $x \in M$.

*c)* $\partial(G_\Sigma \cdot x)$ *is* $\Sigma$*-invariant for all* $x \in M$.

**Proof.** a) Obviously, system group orbits are $\Sigma$-invariant, since $f_u(G_\Sigma \cdot x) = G_\Sigma \cdot x$ for any $u \in U, x \in M$. Moreover, unions of $\Sigma$-invariant subsets of $M$ are $\Sigma$-invariant. Now we assume $N$ to be $\Sigma$-invariant. For all $g \in G_\Sigma$ it is $g(N) = N$, since $g = f_1^{\epsilon_1} f_2^{\epsilon_2} \cdots f_n^{\epsilon_n}$ for $n \in \mathbb{N}$, $f_i \in S_\Sigma$ and $\epsilon_i \in \{-1, 1\}$. Therefore, $G_\Sigma \cdot x \subseteq N$ for all $x \in N$ which yields

$$\bigcup_{x \in N} G_\Sigma \cdot x \subseteq N.$$

On the other hand, $\mathrm{id} \in G_\Sigma$ and therefore

$$N \subseteq \bigcup_{x \in N} G_\Sigma \cdot x.$$

b) Obviously

$$\overline{G_\Sigma \cdot x} \subseteq \bigcup_{y \in \overline{G_\Sigma \cdot x}} G_\Sigma \cdot y$$

since $\mathrm{id} \in G_\Sigma$. On the other hand $y \in \overline{G_\Sigma \cdot x}$ implies

$$g \cdot y \subseteq g(\overline{G_\Sigma \cdot x}) \subseteq \overline{g G_\Sigma \cdot x} = \overline{G_\Sigma \cdot x}$$

since $g : M \to M$ is continuous. Hence, the claim follows from a).
c) Since $f_u$ is bijective, a) and b) imply

$$\begin{aligned} f_u\left(\partial(G_\Sigma \cdot x)\right) &= f_u\left(\overline{G_\Sigma \cdot x} \setminus G_\Sigma \cdot x\right) \\ &= f_u\left(\overline{G_\Sigma \cdot x}\right) \setminus f_u\left(G_\Sigma \cdot x\right) \\ &= \overline{G_\Sigma \cdot x} \setminus G_\Sigma \cdot x \\ &= \partial(G_\Sigma \cdot x). \end{aligned}$$

Hence, $\partial(G_\Sigma \cdot x)$ is $\Sigma$ invariant. □

**Corollary 3.11** *Let* $\Sigma = (M, U, f)$ *be an invertible system and* $N \subseteq M$ *such that* $f_u(N) = N$ *for all* $u \in U$.

*a) If* $\Sigma$ *is reachable from any* $x \in M$ *then* $N = M$.

*b) If* $\Sigma$ *is weakly reversible then* $\Sigma_{|N}$ *is weakly reversible.*

*c) If* $\Sigma$ *is accessible from* $x \in N$ *then* $\Sigma_{|N}$ *is accessible from* $x \in N$.

*d) If* $\Sigma$ *is approximatively reachable then* $\Sigma_{|N}$ *is approximatively reachable.*

**Proof.** a) If $\Sigma$ is reachable from $x$ then $G_\Sigma \cdot x = M$. Following Proposition 3.10, $N = G_\Sigma \cdot x = M$. b) If $\Sigma$ is weakly reversible then $\mathcal{R}(x) = G_\Sigma \cdot x$ for all $x \in M$ (see Lemma 2.35). By Proposition 3.10 $N$ is the union of system group orbits. It follows $\mathcal{R}(x) = G_\Sigma \cdot x$ for all $x \in N$. Hence, $\Sigma_{|N}$ is weakly reversible. c) Since $\mathcal{R}(x) \subseteq N$, $\mathrm{int}_M(\mathcal{R}(x)) \neq \emptyset$ clearly implies $\mathrm{int}_N(\mathcal{R}(x)) \neq \emptyset$. d) If $\mathcal{R}(x) \subseteq N$ is dense in $M$ it is also dense in $N \subseteq M$. □

### 3.2.2 System semigroup of $\Sigma_{|N}$

If we restrict a system to a $\Sigma$-invariant subset, the system semigroup $S_{\Sigma_{|N}}$ of $\Sigma_{|N}$ is not necessarily isomorphic to $S_\Sigma$ or to one of its subsemigroups. Nevertheless, it can be expressed as a factor semigroup of $S_\Sigma$. Given a $\Sigma$-invariant subset $N$ of $M$ we define

$$C_N := \{c \in G_\Sigma \,|\, c_{|N} = \mathrm{id}_{|N}\}. \tag{34}$$

The group $C_N$ is a normal subgroup of $G_\Sigma$, since

$$g^{-1} c \underbrace{g(n)}_{\in N} = g^{-1} g(n) = n$$

for all $g \in G_\Sigma$ and for all $c \in C_N$. Analogously to the construction in subsection 3.1.2 we can introduce a group structure and respectively a semigroup structure on the coset space $G_\Sigma/C_N$ and $S_\Sigma/C_N$, respectively.

The following result describes the relation between the system semigroup of a system $\Sigma$ and the system semigroup of a restricted system $\Sigma_{|N}$ corresponding to a $\Sigma$-invariant set $N \subseteq M$.

**Theorem 3.12** *Let $\Sigma = (M, U, f)$ be an invertible system and $N$ a $\Sigma$-invariant subset of $M$.*

*a) The system semigroup $S_{\Sigma_{|N}}$ of the restricted system $\Sigma_{|N} = (N, U, f_{|N\times U})$ is isomorphic to $S_\Sigma/C_N$. In particular, $S_{\Sigma_{|N}}$ is a group if $S_\Sigma$ is a group.*

*b) Let $\Sigma$ be smoothly invertible such that $G_\Sigma$ is a Lie group and $\alpha : G_\Sigma \times M \to M$, $(g, x) \mapsto g(x)$ is a smooth Lie group action. If $N$ is a $\Sigma$-invariant submanifold, then $G_{\Sigma_{|N}}$ carries a Lie group structure, such that*

$$\tilde{\alpha} : G_{\Sigma_{|N}} \times N \to N, (\tilde{g}, x) \mapsto \tilde{g}(x)$$

*is smooth.*

*c) Assume that $N$ is dense in $M$. Then $S_\Sigma$ and $S_{\Sigma_{|N}}$ are isomorphic as semigroups. In particular, $S_\Sigma$ is a group if and only if $S_{\Sigma_{|N}}$ is a group.*

**Proof.** a) Obviously, the map

$$\Phi : G_\Sigma \to G_{\Sigma_{|N}},\ g \mapsto g_{|N}$$

is a surjective group homomorphism. Moreover,

$$\mathrm{Ker}(\Phi) := \{g \in G_\Sigma \,|\, \Phi(g) = \mathrm{id}_N\} = C_N.$$

Therefore, $\Psi : G_\Sigma/C_N \to G_{\Sigma_{|N}}$, $gC_N \mapsto g_{|N}$ is a group isomorphism. Since, $\Psi(S_\Sigma/C_N) = S_{\Sigma_{|N}}$ and $\Psi^{-1}(S_{\Sigma_{|N}}) = S_\Sigma/C_N$ we conclude, that $\Psi_{|S_\Sigma/C_N}$ is an isomorphism of semigroups.
b) Note that $C_N$ is a closed subgroup of $G_\Sigma$, since $h_x : G_\Sigma \to M$, $g \mapsto g(x)$ is continuous, and therefore, $g_n \in C_N$, $n \in \mathbb{N}$ and $\lim_{n\to\infty} g_n \mapsto c$ imply

$$x = \lim_{n\to\infty} g_n(x) = \lim_{n\to\infty} h_x(g_n) = h_x(c) = c(x)$$

for any $x \in N$. Hence, $c \in C_N$. It follows, that $G_\Sigma/C_N$ carries a Lie structure and $p : G_\Sigma \to G_\Sigma/C_\pi$ is a submersion (see Theorem 2.2 in [GOV97]). Via the identification $\Psi$ of part a), we equip $G_{\Sigma_{|N}}$ with a Lie structure.

Note that the diagram

$$\begin{tikzcd} G_\Sigma \times N \arrow[r, "\alpha"] \arrow[d, "(\Psi\circ p)\times\mathrm{id_N}"'] & N \\ G_{\Sigma_{|N}} \times N \arrow[ur, "\tilde\alpha"] & \end{tikzcd} \tag{35}$$

commutes, since for any $(g, x) \in G_\Sigma \times N$

$$\tilde\alpha \circ ((\psi \circ p) \times \mathrm{id}_N)(g, x) = \tilde\alpha(g_{|N}, x) = g(x) = \alpha(g, x).$$

Note that $\alpha_{|G_\Sigma\times N}$ is smooth and $(\psi \circ p) \times \mathrm{id}_N$ is a submersion. Thus, $\alpha_{|G_\Sigma\times N} = \tilde\alpha \circ ((\psi \circ p) \times \mathrm{id}_N)$ implies, that $\tilde\alpha$ is smooth (see Theorem 0.5 in [DP82])
c) Since every $c \in G_\Sigma$ is continuous, $c_{|N} = \mathrm{id}$ implies $c = \mathrm{id}$. Therefore, $C_N = \{\mathrm{id}\}$ and $S_{\Sigma_{|N}} \cong S_\Sigma$. The second claim follows, since $G_\Sigma = \langle S_\Sigma \rangle$ and $G_{\Sigma_{|N}} = \langle S_{\Sigma_{|N}} \rangle$ □

We finish this section with an interesting consequence of Theorem 3.12 for abelian systems.

**Theorem 3.13** *Let $\Sigma = (M, U, f)$ be an abelian invertible system. Assume that $\Sigma_{|G_\Sigma\cdot x}$ is controllable for some $x \in M$. Then $\Sigma_{|G_\Sigma\cdot z}$ is controllable for any $z \in \partial(G_\Sigma \cdot x)$.*

**Proof.** By Theorem 3.12, the restricted systems $\Sigma_{|\overline{G_\Sigma\cdot x}}$, $\Sigma_{|G_\Sigma\cdot x}$ and $\Sigma_{|G_\Sigma\cdot z}$ are abelian. If $\Sigma_{|G_\Sigma\cdot x}$ is controllable then $S_{\Sigma_{|G_\Sigma\cdot x}}$ is a group (see Theorem 2.39). Therefore, $S_{\Sigma_{|\overline{G_\Sigma\cdot x}}}$ and $S_{\Sigma_{|G_\Sigma\cdot z}}$ are groups by Theorem 3.12. Hence, $\Sigma_{|G_\Sigma\cdot z}$ is controllable by Theorem 2.39. □

# 4 Performance limits via reachable sets

Given a control system $\Sigma = (M, f, U)$ we want to design shift sequences such that $x_{t+1} = f(x_t, u_t)$, $x_0 \in M$ converge to a certain set of interesting points – such as eigenvectors or solutions of equations. The adherence structure of reachable sets provides fundamental limitations for the existence of such shift strategies.

Certainly, a necessary condition for the existence of $u \in U^{\mathbb{N}}$ with $x_0 \overset{u}{\to} z$ is,

$$z \in \overline{G_\Sigma \cdot x_0}. \tag{36}$$

Therefore, as a first step, we analyze the adherence structure of the system group orbits. Nevertheless, (36) does not imply that $x \overset{u}{\to} z$ for any $u \in U^{\mathbb{N}}$. A stronger necessary condition[10] is

$$z \in \overline{\mathcal{R}(x_0)}. \tag{37}$$

Therefore, as a second step, one analyzes the adherence structure of the reachable sets within $G_\Sigma \cdot x$ or within $\overline{G_\Sigma \cdot x}$.

Obviously, (37) implies (36). On the other hand, it is easier to check whether or not (36) is fulfilled. This is due to the fact, that group orbits have more pleasant properties than semigroup orbits[11]. Moreover, the cardinality of the set of reachable sets might be larger than the cardinality of the set of system group orbits.

In Section 4.1 we develop a graph-theoretical language which allows us to express the adherence structure of the reachable sets and the system group orbits graphically.

Even if $z \in \overline{G_\Sigma \cdot x}$ is satisfied, it is not clear if $z$ is reachable or approximatively reachable from $x$. Therefore, we focus on the properties of the reachable structure of the restricted system to $G_\Sigma \cdot x$ (in Section 4.2) respectively of the restricted system to $\overline{G_\Sigma \cdot x}$ (in Section 4.3). In the latter case it might happen that $z \in \overline{G_\Sigma \cdot x}$ is not approximatively reachable from any initial state $y \in G_\Sigma \cdot x$. We show some necessary conditions for this so-called *repelling phenomenon*.

## 4.1 Orbit graph and reachable graph

In the following we describe the adherence structure of system group orbits and reachable sets in terms of directed graphs. See the Appendix C for a brief summary of the basic notations concerning directed graphs.

[10] Note that Condition (37) is not sufficient for the existence of $u \in U^{\mathbb{N}}$ such that $x \overset{u}{\to} z$, see Example 2.45.

[11] such as the partition property and – in the case of analytic systems – a differential structure as an immersed submanifold of the state space (see Theorem 2.5).

**Definition 4.1 (Orbit graph and reachable graph)** Let $\Sigma = (M, f, U)$ be an invertible system. For any pair of subsets $N_1, N_2 \subseteq M$ we write $N_1 \longleftarrow N_2$ if $N_1 \subseteq \overline{N_2}$.

- The *orbit graph* $\mathcal{G}_O(\Sigma) = (V_O(\Sigma), \longleftarrow)$ is given by the set of orbits $V_O(\Sigma) := \{G_\Sigma \cdot x \,|\, x \in M\}$ and the relation $\longleftarrow$ restricted to $V_O(\Sigma)$.

- The *reachable graph* $\mathcal{G}_R(\Sigma) = (V_R(\Sigma), \longleftarrow)$ is given by the set of orbits $V_R(\Sigma) := \{S_\Sigma \cdot x \,|\, x \in M\}$ and the relation $\longleftarrow$ restricted to $V_R(\Sigma)$.

The relation $\longleftarrow$ is reflexive and transitive. As described in Appendix C we neglect those redundant edges in figures.

The following example is related to the well-known power iteration. It illustrates the concept of orbit graphs and reachable graphs.

**Example 4.2** Let $M = \mathbb{RP}^{n-1}$, $U = \mathbb{N}$. For a matrix $A \in \mathbb{R}^{n\times n}$ the *power iteration system* $\Sigma = (M, U, f)$ is given by

$$f(x, u) = A^u \cdot x.$$

Here we denote the canonical action on $\mathbb{RP}^{n-1}$ with $A \cdot x$. For simplicity we analyze the case

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Since $f(\cdot, u_1) \circ f(\cdot, u_2) = f(\cdot, u_1 + u_2)$ for all $u_1, u_2 \in U$ we easily obtain

$$S_\Sigma = \{x \mapsto A^u \cdot x, \mid u \in \mathbb{N}\}$$

for the system semigroup and

$$G_\Sigma = \{x \mapsto A^u \cdot x, \mid u \in \mathbb{Z}\}$$

for the system group. For the eigenspaces $e_1 := \operatorname{span}(1\,,\,0)^\top$ and $e_2 := \operatorname{span}(0\,,\,1)^\top$ we obtain

$$\mathcal{R}(e_1) = G_\Sigma \cdot e_1 = \{e_1\} \quad \text{and} \quad \mathcal{R}(e_2) = G_\Sigma \cdot e_2 = \{e_2\}.$$

The following diagram illustrates the reachable graph of $\Sigma$. Here $x, y, z$

are three different initial states in $\mathbb{RP}^1 \setminus \{e_1, e_2\}$.

$$\begin{tikzcd}
 & {} \arrow[d, dotted] & {} \arrow[d, dotted] & {} \arrow[d, dotted] \\
 & \mathcal{R}(A^{-1} \cdot x) \arrow[d] & \mathcal{R}(A^{-1} \cdot y) \arrow[d] & \mathcal{R}(A^{-1} \cdot z) \arrow[d] \\
 & \ldots \mathcal{R}(x) \ldots \arrow[d] & \ldots \mathcal{R}(y) \ldots \arrow[d] & \ldots \mathcal{R}(z) \ldots \arrow[d] \\
 & \mathcal{R}(A \cdot x) \arrow[drr, dotted] & \mathcal{R}(A \cdot y) \arrow[dr, dotted] & \mathcal{R}(A \cdot z) \arrow[d, dotted] \\
\mathcal{R}(e_1) & & & \mathcal{R}(e_2)
\end{tikzcd}$$

Note that $\mathcal{R}(e_1)$ is not contained in the topological closure of any other reachable set. In other words there exists no initial state $x \in \mathbb{RP}^1 \setminus \{e_1\}$ and no choice of shift parameters $u_1, u_2 \cdots \in \mathbb{N}$ such that the sequenece $A^n \cdot x$ converges to $e_1$. On the other hand we have $G_\Sigma \cdot e_1 \longleftarrow G_\Sigma \cdot x$ for all $x \in M \setminus \{e_2\}$. For the orbit graph of $\Sigma$ we obtain

$$\begin{tikzcd}
 & & G_\Sigma \cdot e_1 \\
\ldots G_\Sigma \cdot x \ldots \arrow[urr] \arrow[drr] & \ldots G_\Sigma \cdot y \ldots \arrow[ur] \arrow[dr] & \ldots G_\Sigma \cdot z \ldots \arrow[u] \arrow[d] \\
 & & G_\Sigma \cdot e_2
\end{tikzcd}$$

Now let $N \subseteq M$ be a $\Sigma$-invariant subset and $\Sigma_{|_N} = (N, U, f_{|_{N \times U}})$ be the restricted system with respect to $N$. We denote the reachable graph, and respectively the orbit graph of $\Sigma_{|_N}$ with $\mathcal{G}_R(\Sigma_{|_N}) = (V_R(\Sigma_{|_N}), \longleftarrow_N)$, and respectively with $\mathcal{G}_O(\Sigma_{|_N}) = (V_O(\Sigma_{|_N}), \longleftarrow_N)$. The following result shows the relation between $\mathcal{G}_R(\Sigma_{|_N})$ and $\mathcal{G}_R(\Sigma)$.

**Proposition 4.3** *Let $\Sigma = (M, U, f)$ be an invertible system and $N$ a $\Sigma$-invariant subset of $M$. Then*

*a) $\mathcal{G}_R(\Sigma_{|_N})$ is an induced subgraph of $\mathcal{G}_R(\Sigma)$,*

*b) $\mathcal{G}_O(\Sigma_{|_N})$ is an induced subgraph of $\mathcal{G}_O(\Sigma)$.*

**Proof.** Let $\mathcal{R}_\Sigma(x)$ be the reachable set of $x$ with respect to $\Sigma$ and $\mathcal{R}_{\Sigma_{|_N}}(x)$ be the reachable set of $x$ with respect to $\Sigma_{|_N}$. Since $N$ is $\Sigma$-invariant we have $\mathcal{R}_\Sigma(x) = \mathcal{R}_{\Sigma_{|N}}(x)$ for any $x \in N$. Recall that $N \subseteq M$ is the induced topology with respect to $M$. Thus,

$$\operatorname{closure}_N \mathcal{R}_{\Sigma_{|N}}(x) = \operatorname{closure}_M \mathcal{R}_\Sigma(x) \cap N.$$

Here, $\operatorname{closure}_A B$ denotes the topological closure of $B \subseteq A$ with respect to the topology on $A$. It follows

$$\mathcal{R}_\Sigma(x) \longleftarrow \mathcal{R}_\Sigma(y) \;\Leftrightarrow\; \mathcal{R}_{\Sigma_{|N}}(x) \longleftarrow_N \mathcal{R}_{\Sigma_{|N}}(y).$$

Thus $\mathcal{G}_R(\Sigma_{|N})$ is an induced subgraph of $\mathcal{G}_O(\Sigma)$. The proof for claim b) is completely analogous. □

**Example 4.4** Let $\Sigma = (\mathbb{RP}^1, \mathbb{N}, f)$ be the power iteration system of Example 4.2. By Proposition 3.10 any $\Sigma$-invariant subset of $\mathbb{RP}^1$ is the union of system group orbits. We choose $x := \operatorname{span}(1\,,\,1)^\top$, $e_2 := \operatorname{span}(0\,,\,1)^\top$ and

$$N := G_\Sigma \cdot x \cup G_\Sigma \cdot e_2 = \{\operatorname{span}(1\,,\,2^u)^\top \,|u \in \mathbb{Z}\} \cup \{e_2\}.$$

The orbit graph $\mathcal{G}_O(\Sigma_{|N})$ is given by

$$\begin{tikzcd} G_\Sigma \cdot e_2 & G_\Sigma \cdot x \arrow[l] \end{tikzcd}$$

and he reachable graph $\mathcal{G}_R(\Sigma_{|N})$ is given by

$$\begin{tikzcd} \mathcal{R}(e_2) & \mathcal{R}(A \cdot x) \arrow[l, dotted] & \mathcal{R}(x) \arrow[l] & \mathcal{R}(A^{-1} \cdot x) \arrow[l] & {} \arrow[l, dotted] \end{tikzcd} .$$

Note that the map $V_R(\Sigma) \to V_O(\Sigma)$, $\mathcal{R}(x) \mapsto G_\Sigma \cdot x$ is well defined and surjective. Therefore one might conjecture, that $\mathcal{G}_O(\Sigma)$ is isomorphic to a subgraph of $\mathcal{G}_R(\Sigma)$. In fact Example 4.2 already shows, that this is not true in general.

**Example 4.5** Again, let $\Sigma = (\mathbb{RP}^1, \mathbb{N}, f)$ be the power iteration system of Example 4.2. Obviously,

$$\begin{tikzcd} & G_\Sigma \cdot e_1 & \\ G_\Sigma \cdot y \arrow[ur] \arrow[dr] & & G_\Sigma \cdot z \arrow[ul] \arrow[dl] \\ & G_\Sigma \cdot e_2 & \end{tikzcd}$$

is a subgraph of $\mathcal{G}_O(\Sigma)$ but not isomorphic to any subgraph of $\mathcal{G}_R(\Sigma)$. Hence, by Proposition C.4, $\mathcal{G}_O(\Sigma)$ is not isomorphic to any subgraph of $\mathcal{G}_R(\Sigma)$.

In some applications the system semigroup and the system group coincide. In this case also the orbit graph and the reachable graph coincide. The converse direction is not true in general[12]. Nevertheless, $\mathcal{G}_R(\Sigma) = \mathcal{G}_O(\Sigma)$ always holds, provided $\Sigma$ is weakly reversible.

**Theorem 4.6** *Let $\Sigma = (M, U, f)$ be an invertible system. The orbit graph and the reachable graph coincide if and only if $\Sigma$ is weakly reversible.*

**Proof.** By Definition 4.1, $\mathcal{G}_O(\Sigma) = \mathcal{G}_R(\Sigma)$ if and only if $G_\Sigma \cdot x = S_\Sigma \cdot x$ for all $x \in M$. This is equivalent to weak reversibility by Proposition 2.34. $\square$

[12] In particular, in Example 2.36 we have $\mathcal{R}(x) = G_\Sigma \cdot x$ for all $x \in M$, and therefore $\mathcal{G}_R(\Sigma) = \mathcal{G}_O(\Sigma)$. Nevertheless we have $S_\Sigma \neq G_\Sigma$.

## 4.2 Reachable sets within an orbit

Following Proposition 3.10 we can always restrict a system $\Sigma = (M, U, f)$ to any orbit[13] $G_\Sigma \cdot x$, $x \in M$. In the following we analyze the reachable sets of the restricted system $\Sigma_{|G_\Sigma \cdot x}$. Note that here, $G_\Sigma$ (as well as $G_{\Sigma|G_\Sigma \cdot x}$) acts transitively on $G_\Sigma \cdot x$. In many situations it is useful to state the results in terms of the original system $\Sigma = (M, U, f)$, i.e., in terms of $G_\Sigma$, $S_\Sigma$ instead of $G_{\Sigma|G_\Sigma \cdot x}$ and $S_{\Sigma|G_\Sigma \cdot x}$.

According to Definition 3.8, $G_\Sigma \cdot x$ is equipped with the subspace topology with respect to $M$. In this section we always assume, that $G_\Sigma \cdot x$ is locally compact. Recall that this is the case if $G_\Sigma \cdot x$ is a submanifold of $M$ and in particular if $\Sigma$ is smoothly invertible and $G_\Sigma \cdot x$ is semi-algebraic (see Theorem 2.7). With the tools developed in Chapter 2 we easily obtain the following observation.

**Proposition 4.7** *Let $\Sigma = (M, U, f)$ be an invertible system and $x \in M$ such that $G_\Sigma \cdot x$ is locally compact. Assume that $G_\Sigma$ is a Lie group acting continuously on $M$ and $\operatorname{int}_{G_\Sigma} S_\Sigma \neq \emptyset$. Then*

*a) $\Sigma_{|G_\Sigma \cdot x}$ is accessible.*

*b) For any $y \in G_\Sigma \cdot x$ there exists an open set $\mathcal{O}_y$ in $G_\Sigma \cdot x$ such that $y \in \mathcal{R}(z)$ for all $z \in \mathcal{O}_y$.*

**Proof.** a) The restricted group action

$$G_\Sigma \times G_\Sigma \cdot x \mapsto G_\Sigma \cdot x; \quad (g, h \cdot x) \mapsto gh \cdot x$$

is continuous and transitive. Therefore, the map $h_x : G_\Sigma \to G_\Sigma \cdot x$, $g \mapsto g \cdot x$ is open by Theorem B.8. Now it follows by Lemma 2.22 that $\operatorname{int}_{G_\Sigma \cdot x} \mathcal{R}(y) \neq \emptyset$ for all $y \in G_\Sigma \cdot x$. Hence, $\Sigma_{|G_\Sigma \cdot x}$ is accessible.
b) Obviously, $y \in \mathcal{R}(z)$ with $z \in S_\Sigma^{-1} \cdot y$. Therefore, it is enough to show, that $S_\Sigma^{-1} \cdot y$ has nonempty interior. Since $\imath : g \mapsto g^{-1}$ is a homeomorphism, $\operatorname{int}_{G_\Sigma} S_\Sigma^{-1} = \imath(\operatorname{int}_{G_\Sigma} S_\Sigma)$ is nonempty. Moreover, for $\tilde{g} \in G_\Sigma$ with $y = \tilde{g} \cdot x$, the map $r_{\tilde{g}} : g \mapsto g\tilde{g}$ is a homeomorphism, and therefore $h_y := h_x \circ r_{\tilde{g}}$, $g \mapsto g \cdot y$ is open. Hence,

$$S_\Sigma^{-1} \cdot y = h_y(S_\Sigma^{-1}) \supseteq h_y(\operatorname{int}_{G_\Sigma} S_\Sigma^{-1})$$

has nonempty interior. □

For the remaining part of this section we assume, that $\Sigma$ is right divisible, left divisible or abelian.

[13] but not to a smaller set $N \subsetneqq G_\Sigma \cdot x$.

**Theorem 4.8** *Let $\Sigma = (M, U, f)$ be an invertible system, $x \in M$ and $y, z \in G_\Sigma \cdot x$.*

*a) If $\Sigma$ is right divisible, then there exists $w \in G_\Sigma \cdot x$ such that*

$$\mathcal{R}(w) \supseteq \mathcal{R}(y) \cup \mathcal{R}(z).$$

*b) If $\Sigma$ is left divisible, then there exists $w \in G_\Sigma \cdot x$ such that*

$$\mathcal{R}(w) \subseteq \mathcal{R}(y) \cap \mathcal{R}(z).$$

**Proof.** a) For all $y, z \in G_\Sigma \cdot x$ there exists $g \in G_\Sigma$ such that $y = g \cdot z$. Since $S_\Sigma$ is right divisible, we obtain $w := s_1^{-1} \cdot y = s_2^{-1} \cdot z$ with $s_1, s_2 \in S_\Sigma$. Therefore,

$$\mathcal{R}(w) = S_\Sigma s_1^{-1} \cdot y \supseteq S_\Sigma \cdot y = \mathcal{R}(y).$$

Analogously, we deduce $\mathcal{R}(w) \supseteq \mathcal{R}(z)$.
b) Now we assume $G_\Sigma = (S_\Sigma)^{-1} S_\Sigma$. Then, $g = s_1^{-1} s_2$ for some $s_1, s_2 \in S_\Sigma$. Thus $\mathcal{R}(w) \subseteq \mathcal{R}(y)$ and $\mathcal{R}(w) \subseteq \mathcal{R}(z)$ for $w := s_1 \cdot y = s_2 \cdot z$. □

**Corollary 4.9** *Let $\Sigma = (M, U, f)$ be an invertible system with right divisible system semigroup $S_\Sigma$. Assume that the restricted system $\Sigma_{|G_\Sigma \cdot x}$ has a finite number of reachable sets. Then $\Sigma_{|G_\Sigma \cdot x}$ is reachable from some $y \in G_\Sigma \cdot x$.*

**Proof.** We assume there exists $y_1, \ldots, y_n \in G_\Sigma \cdot x$ such that for any $y \in G_\Sigma \cdot x$ $\mathcal{R}(y) = \mathcal{R}(y_i)$ for some $i = 1, \ldots, n$. In particular we obtain

$$G_\Sigma \cdot x = \bigcup_{k=1,\ldots,n} \mathcal{R}(y_k).$$

By Theorem 4.8 we deduce, that there exists $y_{1,2} \in G_\Sigma \cdot x$ such that $\mathcal{R}(y_{1,2}) \supseteq \mathcal{R}(y_1) \cup \mathcal{R}(y_2)$. Then, by induction, there exists $y \in G_\Sigma \cdot x$ such that

$$\mathcal{R}(y) \supseteq \mathcal{R}(y_1) \cup \cdots \cup \mathcal{R}(y_n) = G_\Sigma \cdot x = G_\Sigma \cdot y.$$

Hence, $\mathcal{R}(y) = G_\Sigma \cdot y$. □

Even if $\Sigma$ restricted on $G_\Sigma \cdot x$ is not reachable from any $y \in G_\Sigma \cdot x$, then there exists a sequence $(x_t)_{t \in \mathbb{N}}$ in $G_\Sigma \cdot x$, such that $\mathcal{R}(x_{t+1}) \supseteq \mathcal{R}(x_t)$ (Theorem 4.8). The following result describes this phenomena in more detail under some reasonable topological assumptions.

**Theorem 4.10** *Let $\Sigma = (M, U, f)$ be an invertible right divisible system evolving on a manifold $M$ and $x \in M$ such that $G_\Sigma \cdot x$ is locally compact. Assume that the system group $G_\Sigma$ is a Lie group acting continuously on $M$. Moreover, we assume that $\operatorname{int}_{G_\Sigma} S_\Sigma \neq \emptyset$. Then* *a) for any* $y \in G_\Sigma \cdot x$*, there exists a sequence* $(y_t)_{t\in\mathbb{N}}$ *in* $G_\Sigma \cdot x$ *such that*

*(i)* $y_1 = y$

*(ii)* $\mathcal{R}(y_{t+1}) \supseteq \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t)$; $\forall t \in \mathbb{N}_0$

*(iii)* $\bigcup_{t=0}^{\infty} \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t)$ *is dense in* $G_\Sigma \cdot x$

*b) Assume that the sequence* $(y_t)_{t\in\mathbb{N}}$ *in a) has a limit point* $\tilde{y} \in G_\Sigma \cdot x$*. Then* $\Sigma_{G_\Sigma\cdot x}$ *is approximatively reachable from some* $\tilde{z} \in G_\Sigma \cdot x$*. In particular,* $\Sigma_{|G_\Sigma\cdot x}$ *is controllable if* $\Sigma$ *is abelian.*

**Proof.** a) Recall, that $\Sigma_{|G_\Sigma\cdot x}$ is accessible by Proposition 4.7. In particular, for any $s \in \operatorname{int}_{G_\Sigma}(S_\Sigma)$ and $y \in G_\Sigma \cdot x$, $s \cdot y$ is an inner point of $\mathcal{R}(y)$ (with respect to $G_\Sigma \cdot x$) since $h_y : G_\Sigma \longrightarrow G_\Sigma \cdot x, g \mapsto g \cdot y$ is open and therefore

$$s \cdot y \in \operatorname{int}_{G_\Sigma} S_\Sigma \cdot y = h_y(\operatorname{int}_{G_\Sigma} S_\Sigma) \subseteq \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y).$$

The manifold $M$, and therefore $G_\Sigma \cdot x \subseteq M$, is separable. In particular, there exists a countable set

$$Q := \{q_1, q_2, \dots\} \subseteq G_\Sigma \cdot x$$

such that $\overline{Q} = G_\Sigma \cdot x$ (with respect to the topology of $G_\Sigma \cdot x$). Note that $s \cdot Q$ is also countable and dense in $G_\Sigma \cdot x$, since $s_{|G_\Sigma\cdot x}$ is continuous and therefore

$$\overline{s \cdot Q} \supseteq s \cdot \overline{Q} = s \cdot G_\Sigma \cdot x = G_\Sigma \cdot x.$$

For an arbitrary $y_0 \in G_\Sigma \cdot x$ we construct a recursive sequence in the following way.
If $\operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t)$ is dense in $G_\Sigma \cdot x$ then the constant sequence $y_{t+s} := y_t$, $s \in \mathbb{N}$ fulfills (ii) and (iii). If $\operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t)$ is not dense in $G_\Sigma \cdot x$ then we choose $i_t$ minimal, such that

$$s \cdot q_{i_t} \notin \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t).$$

This must be possible, because $s \cdot Q \subseteq \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t)$ implies $\overline{\operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t)} = G_\Sigma \cdot x$. Since $G_\Sigma$ acts transitively on $G_\Sigma \cdot x$, there exists $g \in G_\Sigma$ such that $g \cdot y_t = q_{i_t}$. From $G_\Sigma = S_\Sigma S_\Sigma^{-1}$ it follows that $g s_1 = s_2$ for some $s_1, s_2 \in S_\Sigma$. Now we define $y_{t+1} := s_1^{-1} \cdot y_t$. Note that $y_{t+1} = s_2^{-1} g \cdot y_t = s_2^{-1} \cdot q_{i_t}$ and therefore $\mathcal{R}(y_{t+1}) \supseteq \mathcal{R}(y_t)$ and $\mathcal{R}(y_{t+1}) \supseteq \mathcal{R}(q_{i_t})$. It follows

$$\mathcal{R}(y_{t+1}) \supseteq \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t) \cup \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(q_{i_t}).$$

By construction we have $s \cdot q_{i_t} \notin \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t)$, but $s \cdot q_{i_t} \in \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(q_{i_t})$. Hence,

$$\operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_{t+1}) \supsetneqq \operatorname{int}_{G_\Sigma\cdot x} \mathcal{R}(y_t).$$

Since $s \cdot q_1, \ldots, s \cdot q_{t_{i-1}} \in \operatorname{int}_{G_\Sigma \cdot x} \mathcal{R}(y_{t-1})$, we obtain

$$s \cdot Q \subseteq \bigcup_{t=1}^{\infty} \operatorname{int}_{G_\Sigma \cdot x} \mathcal{R}(y_t)$$

Now the claim follows, since $s \cdot Q$ is dense in $G_\Sigma \cdot x$.
b) Without loss of generality we may assume that $y_t$ converges to $\tilde{y} \in G_\Sigma \cdot x$. Then $\tilde{y}$ lies in the open set $\operatorname{int}_{G_\Sigma} S_\Sigma s^{-1} \cdot \tilde{y}$ for any $s \in \operatorname{int}_{G_\Sigma} S_\Sigma$. It follows, that $y_t \in \mathcal{R}(s^{-1} \cdot \tilde{y})$ for $t$ large enough. Therefore, there exists $s_t \in S_\Sigma$ such that $y_t = s_t s^{-1} \cdot \tilde{y}$. We obtain

$$\mathcal{R}(y_t) = \mathcal{R}(s_t s^{-1} \cdot \tilde{y}) = S_\Sigma s_t s^{-1} \cdot \tilde{y} \subseteq \mathcal{R}(s^{-1} \cdot \tilde{y})$$

From (iii) it follows, that $\Sigma_{G_\Sigma \cdot x}$ is approximatively reachable from $\tilde{z} := s^{-1} \cdot z$. By Proposition 4.7 and Theorem 2.48, $\Sigma_{|G_\Sigma \cdot x}$ is controllable provided $\Sigma$ is abelian. □

For the rest of this subsection we deal with abelian systems. Here we observe the following useful properties.

**Theorem 4.11** *Let $\Sigma = (M, U, f)$ be an abelian invertible system. Then*

*a) $\Sigma$ restricted on $G_\Sigma \cdot x$ is either controllable or there exist infinitely many different reachable sets in $G_\Sigma \cdot x$.*

*b) For all $x_1, x_2 \in G_\Sigma \cdot x$ there exist $y_1, y_2 \in G_\Sigma \cdot x$ such that*

$$\mathcal{R}(y_1) \subseteq \mathcal{R}(x_1) \cap \mathcal{R}(x_2) \quad \textit{and} \quad \mathcal{R}(x_1) \cup \mathcal{R}(x_2) \subseteq \mathcal{R}(y_2).$$

**Proof.** a) The statement is an immediate consequence of Corollary 4.9 and Theorem 2.39. Assuming there exists a finite number of reachable sets in $G_\Sigma \cdot x$, then $\Sigma_{|G_\Sigma \cdot x}$ is reachable from one point. This implies controllability of $\Sigma_{|G_\Sigma \cdot x}$, since $S_\Sigma$ is abelian.
b) Recall that abelian system semigroups are right divisible and left divisible. Thus, the claim follows from Theorem 4.8.

□

Recall that $G_\Sigma \cdot x$ is a $\Sigma$-invariant subset and

$$C_{G_\Sigma \cdot x} := \left\{ g \in G_\Sigma \;\middle|\; g_{|G_\Sigma \cdot x} = \operatorname{id}_{|G_\Sigma \cdot x} \right\}$$

is a subgroup of $G_\Sigma$.

**Theorem 4.12** *Let $\Sigma = (M, U, f)$ be an abelian invertible system, and $x \in M$ such that $G_\Sigma \cdot x$ is locally compact. We assume that $G_\Sigma$ is a Lie group acting continuously on $M$. For $y, z \in G_\Sigma \cdot x$ and $g \in G_\Sigma$ such that $g \cdot y = z$ we have*

$$z \in \overline{\mathcal{R}(y)} \quad \text{if and only if} \quad g \in \overline{S_\Sigma C_{G_\Sigma \cdot x}}.$$

**Proof.** If $g \in \overline{S_\Sigma C_{G_\Sigma \cdot x}}$ then $s_n c_n \to g$ for a sequence $(s_n c_n)_{n\in\mathbb{N}}$ in $S_\Sigma C_{G_\Sigma \cdot x}$. Since $h_y : G_\Sigma \to G_\Sigma \cdot x$, $g \mapsto g \cdot y$ is continuous we obtain $s_n c_n \cdot y = s_n \cdot y \to g \cdot y = z$. Hence, $z \in \overline{\mathcal{R}(y)}$.

Conversely, if $z \in \overline{\mathcal{R}(y)}$, then there exists a sequence $(s_n)_{n\in\mathbb{N}}$ in $S_\Sigma$ such that $s_n \cdot y \to z$. Let us assume

$$g \notin \overline{S_\Sigma C_{G_\Sigma \cdot x}}. \tag{38}$$

By Theorem B.8 $h_y$ is an open map. It follows that $z = g \cdot y$ lies in the open set

$$\big(G_\Sigma \setminus \overline{S_\Sigma C_{G_\Sigma \cdot x}}\big) \cdot y = h_y(G_\Sigma \setminus \overline{S_\Sigma C_{G_\Sigma \cdot x}}).$$

Therefore, $s_n \cdot y = \tilde{g} \cdot y$ for $n$ large enough. Since $G_\Sigma$ is abelian we obtain $s_n \hat{g} \cdot y = \tilde{g}\hat{g} \cdot y$ for any $\hat{g} \in G_\Sigma$. In other words

$$s_{n|G_\Sigma \cdot x} = \tilde{g}_{|G_\Sigma \cdot x}.$$

We conclude $s_n^{-1}\tilde{g} \in C_{G_\Sigma \cdot x}$, which is a contradiction to (38). Hence, $g \in \overline{S_\Sigma C_{G_\Sigma \cdot x}}$. □

In some situations $G_\Sigma \cdot x$ is dense in $M$. In particular this will be the case for classical inverse iteration systems (see Section 6). By continuity, $C_{G_\Sigma \cdot x} = \{\mathrm{id}\}$ if $\overline{G_\Sigma \cdot x} = M$. Assuming the conditions of Theorem 4.12 we obtain $z \in \overline{\mathcal{R}(y)}$ if and only if $g \in \overline{S_\Sigma}$ for $g \in G_\Sigma$ with $g \cdot y = z$.

We finish this subsection with two examples. The first one shows that the claims of Theorem 4.11 and of Theorem 4.12 become false if drop the assumption that $\Sigma$ is abelian.

**Example 4.13** Consider $\Sigma = (\mathbb{R} \times \mathbb{R}^+, U, f)$ with

$$U := \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \,\middle|\, a, b, c > 0 \right\}$$

and $f : M \times U \to M$, $(x, U) \mapsto Ux$. Note that $S_\Sigma$ can be identified with $U$ and that $\Sigma$ is right divisible but not abelian (see Example 2.17). Moreover, $G_\Sigma$ acts transitive on $\mathbb{R} \times \mathbb{R}^+$, since

$$\begin{aligned} G_\Sigma \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \begin{pmatrix} \tilde{a} & \tilde{b} \\ 0 & \tilde{c} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \,\middle|\, a, b, c, \tilde{a}, \tilde{b}, \tilde{c} > 0 \right\} \\ &= \left\{ \begin{pmatrix} \frac{a}{\tilde{a}} - \frac{a\tilde{b}}{\tilde{a}\tilde{c}} + \frac{b}{\tilde{c}} \\ \frac{c}{\tilde{c}} \end{pmatrix} \,\middle|\, a, b, c, \tilde{a}, \tilde{b}, \tilde{c} > 0 \right\} \\ &= \mathbb{R} \times \mathbb{R}^+. \end{aligned}$$

Hence, $\Sigma$ can be regarded as the restriction on an orbit of the system in Example 2.17.

Now we show that $\Sigma$ has only two different reachability sets but is not controllable. For $(\alpha, \beta)^\top \in \mathbb{R} \times \mathbb{R}^+$ we obtain

$$\begin{aligned} \mathcal{R}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}\right) = S_\Sigma \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} &= \left\{ \begin{pmatrix} a\alpha + b\beta \\ c\beta \end{pmatrix} \,\middle|\, a, b, c > 0 \right\} \\ &= \begin{cases} \mathbb{R} \times \mathbb{R}^+ & \text{if } \alpha < 0 \\ \mathbb{R}^+ \times \mathbb{R}^+ & \text{if } \alpha \geq 0 \end{cases} \end{aligned} \tag{39}$$

From (39) it follows, that there exist only two different reachable sets and that $\Sigma$ is reachable from some $y \in \mathbb{R} \times \mathbb{R}^+$. Note that the latter is also a consequence of Corollary 4.9. Nevertheless, (39) also shows, that $\Sigma$ is not controllable, since $(-1, 1)^\top \notin \mathcal{R}\left((1,1)^\top\right)$. In particular this shows, that claim a) of Theorem 4.11 is not fulfilled if $\Sigma$ is not abelian.

Now we show that also Theorem 4.12 becomes false if we drop the assumption that $\Sigma$ is abelian. Recall that

$$G_\Sigma = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \,\middle|\, a, c > 0 \right\}$$

(see Example 2.17). In particular, $G_\Sigma$ is a Lie group acting continuously on $\mathbb{R} \times \mathbb{R}^+$.

Let $z := (0,1)^\top$, $y := \overline{(1,1)}^\top$ and $g \in G_\Sigma$ such that $g \cdot y = z$. In (39) we have seen, that $z \in \overline{\mathcal{R}(y)}$. The only linear map$f : \mathbb{R}^2 \to \mathbb{R}^2$ with $f_{|_{\mathbb{R}\times\mathbb{R}^+}} = \mathrm{id}_{|_{\mathbb{R}\times\mathbb{R}^+}}$ is $f : x \mapsto x$. In other words

$$C_{G_\Sigma \cdot x} = \left\{ f \in S_\Sigma S_\Sigma^{-1} \mid f_{|_{\mathbb{R}\times\mathbb{R}^+}} = \mathrm{id}_{|_{\mathbb{R}\times\mathbb{R}^+}} \right\} = \{\mathrm{id}\}.$$

Therefore,

$$\overline{S_\Sigma C_{G_\Sigma \cdot x}} = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \,\middle|\, a, c > 0; b \geq 0 \right\}.$$

On the other hand, $g \cdot y = z$ for $g \in G_\Sigma$ implies

$$g = \begin{pmatrix} a & -a \\ 0 & 1 \end{pmatrix} \qquad \text{with } a \in \mathbb{R}^+.$$

Hence, $g \notin \overline{S_\Sigma C_{G_\Sigma \cdot x}}$ but $z \in \overline{\mathcal{R}(y)}$.

Theorem 4.8 shows, that reachable sets within an orbit have nonempty intersection, provided $\Sigma$ is left divisible. The following example shows, that this is not the case for general systems.

**Example 4.14** Consider $\Sigma = (\mathbb{R}^2 \setminus \{0\}, U, f)$ of Example 2.18, i.e.,

$$U := \left\{ \begin{pmatrix} u_1 & u_2 \\ u_3 & u_4 \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \,\middle|\, u_i > 0, i = 1, \ldots, 4 \right\}$$

and $f(x, U) = Ux$. Recall that $\Sigma$ is not left divisible. For

$$s_1 = \begin{pmatrix} 2 & 3 \\ 1 & 1 \end{pmatrix} \in S_\Sigma \quad \text{and} \quad s_2 = \begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix} \in S_\Sigma.$$

we obtain

$$g := s_1^{-1} s_2 s_1^{-1} = \begin{pmatrix} -1 & 5 \\ 0 & -1 \end{pmatrix} \in G_\Sigma.$$

Therefore, $(6,1)^\top$ and $g \cdot (6,1)^\top = (-1,-1)^\top$ are in the same system group orbit. On the other hand, $\mathcal{R}((6,1)^\top) \subseteq \mathbb{R}^+ \times \mathbb{R}^+$ and $\mathcal{R}((-1,-1)^\top) \subseteq \mathbb{R}^- \times \mathbb{R}^-$.

## 4.3 Systems restricted to $\overline{G_\Sigma \cdot x}$

We have seen, that the topological closure of a system group orbit is $\Sigma$-invariant (see Proposition 3.10). In the following we focus on the analysis on the restricted system $\Sigma_{|\overline{G_\Sigma \cdot x}}$. As in Subsection 4.2 we assume that the system semigroup is right divisible or abelian. It is easy to see that $z \in \overline{G_\Sigma \cdot x}$ does not imply $z \in \overline{\mathcal{R}(x)}$. In fact, it might happen that $z \notin \overline{\mathcal{R}(y)}$ for any $y \in G_\Sigma \cdot x$. This phenomenon motivates the following definition.

**Definition 4.15 (Repelling phenomenon)** Let $\Sigma = (M, U, f)$ be a system and $\mathcal{E}$ a subset of $M$. We say that $\mathcal{E}$ is *repelling* with respect to $G_\Sigma \cdot x$ if $\mathcal{E} \cap \overline{\mathcal{R}(y)} = \emptyset$ for all $y \in G_\Sigma \cdot x$.

An easy example for the repelling phenomenon is the following.

**Example 4.16** Let $\Sigma = (\mathbb{R}, U, f)$ be the system given by $f(x,u) = xu$ with $U = (1, \infty)$ and $\mathcal{E} = \{0\}$. Note that $\mathcal{E} \subseteq \overline{G_\Sigma \cdot x} = \mathbb{R} \setminus \overline{\{0\}}$ for all $x \in \mathbb{R} \setminus \{0\}$. Hovever, $\mathcal{E}$ is repelling to ${}_\Sigma \cdot x$, $x \in \mathbb{R} \setminus \{0\}$ since $\mathcal{E} \cap \overline{\mathcal{R}(x)} = \emptyset$. In particular, no shift strategy will steer any initial state $x \neq 0$ arbitrary close to $\mathcal{E}$, regardless how close the initial state was to the interesting point.

Obviously, a point $z \in \overline{G_\Sigma \cdot x}$ which is repelling to $G_\Sigma \cdot x$ has to be in the boundary of $G_\Sigma \cdot x$, since $z \in \mathcal{R}(s^{-1} \cdot z)$ for all $s \in S_\Sigma$. The next result gives a condition for the existence of an repelling point in $\partial(G_\Sigma \cdot x)$.

**Theorem 4.17** *Let $\Sigma = (M, U, f)$ be an invertible right divisible system and $x \in M$ such that $\partial(G_\Sigma \cdot x) \neq \emptyset$. Then one of the following alternatives is true:*

*(i) There exists $z \in \partial(G_\Sigma \cdot x)$ which is repelling with respect to $G_\Sigma \cdot x$.*

*(ii) For any finite subset $\mathcal{E} \subseteq \partial(G_\Sigma \cdot x)$ there exists $y \in G_\Sigma \cdot x$ such that $\mathcal{E} \subseteq \overline{\mathcal{R}(y)}$.*

**Proof.** Obviously, $(ii)$ implies that $(i)$ is false. Now we assume that statement $(i)$ is false. Then, for any finite set $\mathcal{E} = \{z_1, \dots, z_N\} \subseteq \partial(G_\Sigma \cdot x)$ there exists a set $\{y_1, \dots, y_N\} \subseteq G_\Sigma \cdot x$ such that $z_n \in \overline{\mathcal{R}(y_n)}$, $n = 1 \dots, N$. By Theorem 4.8 there exists $y_T \in G_\Sigma \cdot x$ such that $\{y_1, \dots, y_N\} \subseteq \mathcal{R}(y_T)$. Hence, $\mathcal{E} \subseteq \overline{\mathcal{R}(y_T)}$, since $\overline{\mathcal{R}(y_n)} \subseteq \overline{\mathcal{R}(y_N)}$ for $n = 1, \dots, N$. □

Now we focus on the case where $\Sigma$ is abelian. Here, it is sufficient to analyze $\overline{\mathcal{R}(y)} \cap \mathcal{E}$ for one $y \in G_\Sigma \cdot x$ to decide if a $\Sigma$-invariant subset $\mathcal{E}$ is repelling to $G_\Sigma \cdot x$.

**Theorem 4.18** *Let $\Sigma = (M, U, f)$ be an abelian invertible system and $x \in M$. For any $\Sigma$-invariant subset $\mathcal{E} \subseteq \partial(G_\Sigma \cdot x)$ the following two statements are equivalent.*

*(i) $\mathcal{E}$ is repelling to $G_\Sigma \cdot x$*

*(ii) There exists $y \in G_\Sigma \cdot x$ such that $\overline{\mathcal{R}(y)} \cap \mathcal{E} = \emptyset$*

**Proof.** a) The implication $(i) \Rightarrow (ii)$ is trivial. Now we show that $\overline{\mathcal{R}(y)} \cap \mathcal{E} \neq \emptyset$ implies $\overline{\mathcal{R}(w)} \cap \mathcal{E} \neq \emptyset$ for any $w \in G_\Sigma \cdot x$. Recall, that there exists $g \in G_\Sigma$, such that $g \cdot y = w$. Moreover, $g \cdot \mathcal{E} = \mathcal{E}$ for all $g \in G_\Sigma$ since $f_u(\mathcal{E}) = \mathcal{E}$ for all $u \in U$. If $\overline{\mathcal{R}(y)} \cap \mathcal{E} \neq \emptyset$ then there exists a sequence $(s_n)_{n\in\mathbb{N}}$ in $S_\Sigma$ such that $s_n \cdot y$ converges to $\mathcal{E}$. We conclude, that $\overline{\mathcal{R}(w)} \cap \mathcal{E} \neq \emptyset$ since

$$s_n \cdot w = g(s_n \cdot y) \to g \cdot \mathcal{E} = \mathcal{E}.$$

□

We finish this section with an example which shows, that the claim of Theorem 4.18 is wrong, if we drop the assumption that $\Sigma$ is abelian, even if $\Sigma$ is right divisible.

**Example 4.19** Consider $\Sigma = (\mathbb{R} \times \mathbb{R}_0^+, U, f)$ with

$$U = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R}) \,\middle|\, a, b, c > 0 \right\}$$

and $f : M \times U \to M$, $(x, U) \mapsto Ux$. Note that $\Sigma$ is right divisible and that $G_\Sigma \cdot x = \mathbb{R} \times \mathbb{R}^+$ for $x \in \mathbb{R} \times \mathbb{R}^+$ (see Example 4.13). Therefore, $\Sigma$ can be regarded as the restriction on $\overline{G_\Sigma \cdot x}$ of the system in Example 2.17. We obtain

$$\begin{array}{llll}
G_\Sigma \cdot z_0 = \{z_0\}, & \mathcal{R}(z_0) = \{z_0\} & \text{for} & z_0 = (0,0)^\top, \\
G_\Sigma \cdot z_1 = \mathbb{R}^+ \times \{0\}, & \mathcal{R}(z_1) = \mathbb{R}^+ \times \{0\} & \text{for all} & z_1 \in \mathbb{R}^+ \times \{0\}, \\
G_\Sigma \cdot z_2 = \mathbb{R}^- \times \{0\}, & \mathcal{R}(z_2) = \mathbb{R}^- \times \{0\} & \text{for all} & z_2 \in \mathbb{R}^- \times \{0\}, \\
G_\Sigma \cdot x_1 = \mathbb{R} \times \mathbb{R}^+, & \mathcal{R}(x_1) = \mathbb{R}_0^+ \times \mathbb{R}^+ & \text{for all} & x_1 \in \mathbb{R}_0^+ \times \mathbb{R}^+, \\
G_\Sigma \cdot x_2 = \mathbb{R} \times \mathbb{R}^+, & \mathcal{R}(x_2) = \mathbb{R} \times \mathbb{R}^+ & \text{for all} & x_2 \in \mathbb{R}^- \times \mathbb{R}^+.
\end{array}$$

The orbit graph and the reachable graph are given by

$$\begin{tikzcd}
 & G_\Sigma \cdot z_1 \arrow[dl] & \\
G_\Sigma \cdot z_0 & G_\Sigma \cdot x_1 \arrow[l] \arrow[u] \arrow[d] & \\
 & G_\Sigma \cdot z_2 \arrow[ul] &
\end{tikzcd}
\qquad
\begin{tikzcd}
 & \mathcal{R}(z_1) \arrow[dl] & \\
\mathcal{R}(z_0) & \mathcal{R}(x_1) \arrow[l] \arrow[u] & \mathcal{R}(x_2) \arrow[l] \arrow[ul] \arrow[dl] \\
 & \mathcal{R}(z_2) \arrow[ul] &
\end{tikzcd}$$

In particular we see, that $\Sigma_{|G_\Sigma \cdot z_i}$, $i = 1, 2, 3$ is controllable, but that $G_\Sigma \cdot z_2 = \overline{\mathcal{R}(z_2)}$ is not a subset of $\overline{\mathcal{R}(x_1)}$. Moreover, $\mathcal{E} := G_\Sigma \cdot z_2$ is $\Sigma$-invariant and $\overline{\mathcal{R}(x_1)} \cap \mathcal{E} = \emptyset$. However, $\mathcal{E}$ is not repelling to $G_\Sigma \cdot x_1$ since $x_2 \in G_\Sigma \cdot x_1$ but $\overline{\mathcal{R}(x_2)} \cap \mathcal{E} \neq \emptyset$. This shows, that the claim of Theorem 4.18 does not hold, if $\Sigma$ is not abelian.

# 5 Systems on homogeneous spaces

In the following we apply the results of the previous chapters to systems evolving on Lie groups and homogeneous spaces. Here, the geometric framework developed by Jakubczyk and Sontag (see Section 2.2.1) comes into play. In particular, in Section 5.1, we prove discrete-time versions of results by Jurdjevic and Sussmann on controllability of continuous-time systems on Lie groups (see [JS72] and [SJ72]). Systems on homogeneous spaces can be regarded as induced systems of a system on a Lie group. Thus, the controllability properties of systems on homogeneous spaces $\tilde{\Sigma}$ are linked to the controllability properties of certain related system on a Lie group $\Sigma$. In Section 5.2 we show a condition for weak reversibility of $\tilde{\Sigma}$ in terms of the system semigroup of $\Sigma$. Moreover, we investigate the situation for systems on flag manifolds and projective spaces.

## 5.1 Systems on Lie groups

**Definition 5.1** Let $G$ be a Lie group. A smoothly invertible system $\Sigma = (G, U, f)$ is *evolving* on $G$ if for any $u \in U$ there exists a group element $g \in G$ such that $f_u \cdot x = gx$ for all $x \in G$. We identify $f_u$ with $g \in G$. In particular we write $e := \mathrm{id}_G$.

Note that in this case $G_\Sigma$ is a subgroup of $G$ and that

$$\mathcal{R}(\mathrm{e}) = \left\{ \prod_{t=1}^{T} f_{u_t} \,\middle|\, T \in \mathbb{N}, u_t \in U \right\} = S_\Sigma.$$

In other words, $\Sigma$ is accessible from $e$ if and only if $\mathrm{int}_G\, S_\Sigma \neq \emptyset$. In fact, $\mathrm{int}_G\, S_\Sigma \neq \emptyset$ is equivalent to accessibility from any point.

**Proposition 5.2** *Let $\Sigma = (G, U, f)$ be a system evolving on a Lie group. Then*

*a) $\Sigma$ is accessible if and only if $\Sigma$ is accessible from one point.*

*b) $\Sigma$ is controllable if and only if $S_\Sigma = G$.*

**Proof.** a) Let $\Sigma$ be accessible from $g \in G$. For any $h \in G$ the map

$$r_h : G \to G,\ x \mapsto xh$$

is a homeomorphism. Therefore,

$$\mathcal{R}(\tilde{g}) = S_\Sigma \tilde{g} = S_\Sigma g g^{-1} \tilde{g} = r_{g^{-1}\tilde{g}}(\mathcal{R}(g))$$

has nonempty interior.
b) Obviously, $S_\Sigma = G$ implies controllability. Conversely, if $\Sigma$ is controllable, for every $g \in G$ there exists $s \in S_\Sigma$ such that $sg^{-1} = e$. Therefore, for any $g \in G$ we have $g \in S_\Sigma$. Hence, $S_\Sigma = G$. □

To check if $\Sigma$ is accessible or not, one can apply the geometric framework developed by Jakubczyk and Sontag (see Theorem 2.21). We choose $\tilde{U} \subseteq U$ such that every connected component of $U$ has at least one element in $\tilde{U}$. Following the construction in Section 2.2.1, the *Lie derivative vector fields*

$$\mathrm{Ad}_{\mathrm{u_1,\ldots,u_k}} f_{u,i} : G \mapsto TG, \quad u \in U, k \in N_0, u_1, \ldots, u_k \in \tilde{U}, 1 \leq i \leq m$$

given by

$$g \mapsto \left.\frac{\partial}{\partial v_i}\right|_{v=0} (f_{u_k} \cdots f_{u_1})^{-1} f_u^{-1} f_{u+v} (f_{u_k} \cdots f_{u_1})(g)$$

generate the Lie algebra $\mathcal{L}_\Sigma$. We denote $T_eG$, the *Lie algebra* [14] of $G$, with $\mathfrak{g}$.

**Proposition 5.3** *Let $\Sigma = (G, U, f)$ be a smooth system evolving on a Lie group $G$ with corresponding Lie algebra $\mathfrak{g}$. Moreover we assume, that $U \subseteq \mathbb{R}^m$ is open. Then, $\Sigma$ is accessible if and only if $\mathcal{L}_\Sigma(e) = \mathfrak{g}$.*

**Proof.** Obviously we have $\mathcal{L}_\Sigma(e) \subseteq T_eG = \mathfrak{g}$. The case $\mathcal{L}_\Sigma(e) \neq \mathfrak{g}$ immediately implies $\dim \mathcal{L}_\Sigma(e) < n = \dim(G)$. Therefore, $\Sigma$ is not accessible by Theorem 2.21.
Now we assume $\mathcal{L}_\Sigma(e) = \mathfrak{g}$. For any $g \in G$ we define $l_g : G \to G$, $h \mapsto gh$ and $Tl_g : TG \to TG$ as the corresponding tangent map. For any $X = \mathrm{Ad}_{\mathrm{u_1,\ldots,u_k}} f_{u,i}$ with $u \in U, k \in \mathbb{N}_0, u_1, \ldots, u_k \in \tilde{U}, 1 \leq i \leq m$ we have

$$Tl_g \circ X(e) = Tl_g(e, X(e)) = (g, X(g)) = X \circ l_g(e).$$

In other words, all vector fields $\mathrm{Ad}_{\mathrm{u_1,\ldots,u_k}} f_{u,i}$, and therefore all vector fields $X \in \mathcal{L}_\Sigma$, are left invariant. Moreover, for any $X \in \mathcal{L}_\Sigma$, the isomorphism $T_el_g : T_eG \to T_gG$ maps $X(e)$ on $X(g)$. Therefore, for any $g \in G$ we obtain $\dim \mathcal{L}_\Sigma(e) = \dim \mathcal{L}_\Sigma(g)$, since

$$\mathcal{L}_\Sigma(e) = \{X(e) \mid X \in \mathcal{L}_\Sigma\} = \left\{(T_el_g)^{-1} X(g) \,\middle|\, X \in \mathcal{L}_\Sigma\right\} = (T_el_g)^{-1} \mathcal{L}_\Sigma(g).$$

Hence, $\Sigma$ is accessible by Theorem 2.21, since $\dim \mathcal{L}_\Sigma(g) = n$ for all $g \in G$. □

[14] In fact, $\mathfrak{g} := T_eG$, equipped with the product $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, $(X, Y) \mapsto (\mathrm{ad}\, X)(Y)$ is called Lie algebra. Nevertheless, in the following we do not use the algebra structure of $T_eG$.

In the following we show that for systems evolving on Lie groups, controllability, approximative reachability and dense reachability are equivalent concepts, provided $\Sigma$ is accessible.

**Theorem 5.4** *Let $\Sigma = (G, U, f)$ be a system evolving on a Lie group $G$. Assume that $G^i \cap S_\Sigma \neq \emptyset$ for all connected components $G^i$ of $G$. If $\Sigma$ is accessible, then the following statements are equivalent:*

*(i) $S_\Sigma$ is a group,*

*(ii) $\Sigma$ is controllable,*

*(iii) $\Sigma$ is approximatively reachable from one point $g \in G$,*

*(iv) $\Sigma$ is densely reachable.*

**Proof.** The implications $(ii) \Rightarrow (iv)$ and $(iv) \Rightarrow (iii)$ follow immediately from the definition. Moreover, $(ii) \Rightarrow (i)$ follows from Proposition 5.2.
Now we show $(i) \Rightarrow (ii)$. Recall that $\Sigma$ is accessible if and only $\Sigma$ is accessible from $e$. Assuming that $S_\Sigma$ is a group, i.e., $S_\Sigma = G_\Sigma$, the reachable sets $\mathcal{R}(g)$ are all open in $G$ by Proposition 2.20. In particular, it follows

$$e \in S_\Sigma = \mathrm{int}_G(S_\Sigma).$$

Therefore, $S_\Sigma = G$ by Lemma B.4. In particular, $\Sigma$ is controllable[15].
We finish the proof by showing $(iii) \Rightarrow (i)$. Let $g \in G$ be such that $\overline{\mathcal{R}(g)} = G$. Since the map $r_{g^{-1}} : G \to G$, $x \mapsto xg^{-1}$ is a homeomorphism we obtain

$$G = r_{g^{-1}}(\overline{S_\Sigma g}) = \overline{r_{g^{-1}}(S_\Sigma g)} = \overline{S_\Sigma}.$$

In particular this shows $f_u^{-1} \in \overline{S_\Sigma}$ for all $u \in U$. Moreover, accessibility of $\Sigma$ implies $\mathrm{int}_G\, S_\Sigma = \mathrm{int}_G\, \mathcal{R}(\mathrm{id}_G) \neq \emptyset$. Now $S_\Sigma = G_\Sigma$ follows from Theorem 2.40. □

Similar to the situation for continuous time systems (see Theorem 6.5 in [JS72]), accessibility implies controllability provided $G$ is compact.

**Theorem 5.5** *Let $\Sigma = (G, U, f)$ be a system evolving on a compact Lie group $G$. Assume that $G^i \cap S_\Sigma \neq \emptyset$ for all connected components $G^i$ of $G$. Then $\Sigma$ is controllable if and only if $\Sigma$ is accessible.*

[15] If $S_\Sigma$ is a group, $\Sigma$ is weakly reversible. Moreover, accessibility implies that $G_\Sigma \cdot g$ is open (see Proposition 2.20) and therefore $g \in \mathrm{int}_G\, \mathcal{R}(g)$ for all $g \in G$. Hence, $(i) \Rightarrow (ii)$ also follows immediately from Theorem 2.37 provided $G$ is connected.

**Proof.** Obviously, controllability implies accessibility. Now let $\Sigma$ be accessible and $s \in \operatorname{int}_G S_\Sigma$. Since $G$ is compact, $\overline{S_\Sigma}$ is compact and therefore a group by Lemma B.2. Lemma B.5 yields

$$e = s^{-1}s \in \overline{S_\Sigma}\, \operatorname{int}_G \overline{S_\Sigma} \subseteq \operatorname{int}_G \overline{S_\Sigma}.$$

for any $s \in \operatorname{int}_G \overline{S_\Sigma}$.

Since $G^i \cap \overline{S_\Sigma} \neq \emptyset$ for all connected components $G^i$ of $G$, we obtain $\overline{S_\Sigma} = G$ by Lemma B.4. Moreover, $\overline{S_\Sigma} = G$ and $\operatorname{int}_G S_\Sigma \neq \emptyset$ implies $S_\Sigma = G$ by Lemma B.6. Thus, $\Sigma$ is controllable. □

The assumption of accessibility in Theorem 5.4 and Theorem 5.5 cannot be dropped. In fact, the system of Example 2.47 evolves on a compact Lie group. Here, $\Sigma$ is densely reachable but not controllable. Moreover, $S_\Sigma \neq G$.

## 5.2 Homogeneous spaces

Let $\Sigma = (G, U, f)$ be a system evolving on a Lie group $G$ as introduced in subsection 5.1, i.e., for all $u \in U$ there exists $g \in G$ such that $f_u(x) = gx$ for all $x \in G$. Again we identify $f_u$ with $g$. Now let $\alpha : G \times M \to M$ be a transitive smooth group action on a set $M$. We choose a fixed reference element $m \in M$. Moreover, we assume, that $\mathrm{Stab}_m = \{g \in G \,|\, g \cdot m = m\}$ is a closed subgroup of $G$. Then $M$ is a *homogeneous space* with respect to $\alpha$ and it can be equipped with a canonical differential structure (See Appendix F). Here, the projection $\pi_m : G \to M,\ g \mapsto g(m)$ defines the open sets in $M$, i.e., $\mathcal{U} \subseteq M$ is open if and only if $\mathcal{U} = \pi_m(\tilde{\mathcal{U}})$ for an open set in $G$.

For $u \in U$ we define

$$\tilde{f} : M \times U \to M,\ (m, u) \mapsto f_u \cdot m.$$

Note that $\tilde{f}_u : m \mapsto f_u(m)$ is a diffeomorphism for all $u \in U$. The inverse $\tilde{f}_u^{-1}$ is given by $m \mapsto f_u^{-1}(m)$. This defines a smoothly invertible system $\tilde{\Sigma} = (M, U, \tilde{f})$ on the homogeneous space $M$.

**Proposition 5.6** *$\tilde{\Sigma} = (M, U, \tilde{f})$ is an induced system of $\Sigma = (G, U, f)$ with respect to $\pi_m : G \to M,\ g \mapsto g(m)$.*

**Proof.** By construction, $\pi$ is surjective, continuous and open. Moreover, for all $g \in G$ and all $u \in U$ it follows that

$$\tilde{f}_u \circ \pi_m(g) = \tilde{f}_u(g(m)) = f_u g(m) = \pi_m(f_u g) = \pi_m \circ f_u(g).$$

Hence, $\tilde{f}_u \circ \pi_m = \pi_m \circ f_u$ for any $u \in U$. □

Recall that the core $C_M = \bigcap_{m \in M} \mathrm{Stab}_m$ is a normal subgroup of $G$. This implies that $G_\Sigma \cap C_M$ is a normal subgroup of $G_\Sigma$. Analogous to the construction in Section 3.1.2 the product

$$s_1(G_\Sigma \cap C_M) s_2(G_\Sigma \cap C_M) := s_1 s_2(G_\Sigma \cap C_M)$$

defines a semigroup structure on the set of cosets $S_\Sigma/(G_\Sigma \cap C_M)$. The following proposition shows the relation between $C_M$ and the core of $\pi_M$, i.e. $C_{\pi_m} = \{g \in G_\Sigma \,|\, \pi_m(g \cdot x) = \pi_m(x), \forall x \in M\}$.

**Proposition 5.7** *Let $\Sigma$, $\tilde{\Sigma}$ and $\pi_m$ be defined as above. Then*

$$C_{\pi_m} = G_\Sigma \cap C_M.$$

*In particular, $C_{\pi_m}$ is independent of the choice of the reference point $m \in M$. Moreover $S_{\tilde{\Sigma}}$ and $S_\Sigma/(G_\Sigma \cap C_M)$ are isomorphic as semigroups and $G_{\tilde{\Sigma}}$ and $G_\Sigma/(G_\Sigma \cap C_M)$ are isomorphic as groups.*

**Proof.** A straightforward calculation shows

$$\begin{aligned} C_{\pi_m} &= \{g \in G_\Sigma \,|\, \pi_m \circ g = \pi_m\} \\ &= \{g \in G_\Sigma \,|\, \pi_m \circ g(h) = \pi_m(h),\, \forall h \in G\} \\ &= \{g \in G_\Sigma \,|\, gh \cdot m = h \cdot m,\, \forall h \in G\}. \end{aligned}$$

Since $G$ acts transitively on $M$, we conclude

$$\begin{aligned} \{g \in G_\Sigma \,|\, gh \cdot m = h \cdot m,\, \forall h \in G\} &= \{g \in G_\Sigma \,|\, g \cdot \tilde{m} = \tilde{m},\, \forall \tilde{m} \in M\} \\ &= G_\Sigma \cap C_M. \end{aligned}$$

By Theorem 3.6, $S_{\tilde{\Sigma}}$ and $S_\Sigma/(G_\Sigma \cap C_M)$, and respectively $G_{\tilde{\Sigma}}$ and $G_\Sigma/(G_\Sigma \cap C_M)$ are isomorphic. □

In particular, Proposition 5.7 shows, that $C_{\pi_m}$ is independent of the choice of the reference point $m \in M$. Therefore we write $C_\pi := C_{\pi_m}$.

Using the machinery developed in the previous sections, we can easily affirm a reformulated version of Theorem 3.2 in [Jor06], which will be important in our analysis of inverse iteration systems.

**Theorem 5.8** *Let $\Sigma = (G, U, f)$ be a system evolving on a Lie group $G$ which acts transitively on a set $M$. Let $\tilde{\Sigma} = (M, U, \tilde{f})$ be the induced system on the homogeneous space $M$.*

*a) If $G_\Sigma = C_\pi S_\Sigma$ then $\tilde{\Sigma}$ is weakly reversible*

*b) If there exists a reference point in $M$ such that $\mathrm{Stab}_m \cap G_\Sigma \subseteq C_M$. Then $\mathcal{R}_{\tilde{\Sigma}}(m) = G_{\tilde{\Sigma}} \cdot m$ implies $G_\Sigma = C_\pi S_\Sigma$.*

**Proof.** a) Assuming $G_\Sigma = C_\pi S_\Sigma$ then $G_{\tilde{\Sigma}} = S_{\tilde{\Sigma}}$ by Theorem 3.6 and therefore $G_{\tilde{\Sigma}} \cdot m = \mathcal{R}_{\tilde{\Sigma}}(m)$ for all $m \in M$. Hence, $\tilde{\Sigma}$ is weakly reversible by Proposition 2.35.
b) We assume $\mathcal{R}_{\tilde{\Sigma}}(m) = G_{\tilde{\Sigma}} \cdot m$. In other words, for all $g \in G_\Sigma$, there exists $s \in S_\Sigma$ such that $g^{-1}s \in \mathrm{Stab}_m$. Since $g, s \in G_\Sigma$ and $\mathrm{Stab}_m \cap G_\Sigma \subseteq C_M$ we obtain $g^{-1}s \in G_\Sigma \cap C_M = C_\pi$. It follows, $g = cs$ for some $c \in C_\pi$. Therefore $G_\Sigma \subseteq C_\pi S_\Sigma$. Moreover, $C_\pi S_\Sigma \subseteq G_\Sigma$ since $C_\pi$ and $S_\Sigma$ are subsemigroups of $G_\Sigma$. Hence $\mathcal{R}_{\tilde{\Sigma}}(m) = G_{\tilde{\Sigma}} \cdot m$ implies $G_\Sigma = C_\pi S_\Sigma$. □

We finish this section with some observations with two special cases, namely system on *flag manifolds* and systems on *projective spaces*.

### 5.2.1 Systems on flag manifolds

Let $\Sigma = (\mathrm{GL}_n(\mathbb{R}), U, f)$ be a system evolving on $\mathrm{GL}_n(\mathbb{R})$, i.e., $f_u \in \mathrm{GL}_n(\mathbb{R})$ for all $u \in U$. Recall that $\mathrm{GL}_n(\mathbb{R})$ acts transitively on the flag manifold $\mathrm{Flag}(d, \mathbb{R}^n)$ (see Appendix F). We denote the identity element of $\mathrm{GL}_n(\mathbb{R})$ with $I$. Following the construction in Section 5.2 we define a new system on $\tilde{\Sigma} = (\mathrm{Flag}(d, \mathbb{R}^n), U, \tilde{f})$ on $\mathrm{Flag}(d, \mathbb{R}^n)$, with $\tilde{f}$ given by

$$\tilde{f} : ((V_1, \ldots, V_k), u) \mapsto (f_u V_1, \ldots, f_u V_k)) .$$

Here, $f_u V_i$ denotes the image of the $d_i$-dimensional subspace $V_i$ under the linear map $f_u$. The previous results yield:

**Theorem 5.9** *Let $\Sigma$ and $\tilde{\Sigma}$ be systems as above and $\mathcal{V} = (V_1, \ldots, V_k)$ a reference flag in* $\mathrm{Flag}(d, \mathbb{R}^n)$.

*a) System $\tilde{\Sigma}$ is an induced system of $\Sigma$ with respect to*

$$\pi_{\mathcal{V}} : \mathrm{GL}_n(\mathbb{R}) \mapsto \mathrm{Flag}(d, \mathbb{R}^n),\ x \mapsto (xV_1, \ldots, xV_k) .$$

*b) $S_{\tilde{\Sigma}}$ is isomorphic to $S_\Sigma / C_\pi$ and $G_{\tilde{\Sigma}}$ is isomorphic to $G_\Sigma / C_\pi$ Here, $C_\pi = G_\Sigma \cap \mathbb{R}^* I$.*

*c) If $\mathcal{V}$ fulfills* $\mathrm{Stab}(\mathcal{V}) \cap G_\Sigma \subseteq \mathbb{R}^* I$, *then $\mathcal{R}_{\tilde{\Sigma}}(\mathcal{V}) = G_{\tilde{\Sigma}} \cdot \mathcal{V}$ if and only if $\mathcal{R}_{\tilde{\Sigma}}(\tilde{\mathcal{V}}) = G_{\tilde{\Sigma}} \cdot \tilde{\mathcal{V}}$ for all $\tilde{\mathcal{V}} \in$* $\mathrm{Flag}(d, \mathbb{R}^n)$.

**Proof.** The first statement follows immediately from Proposition 5.6. Recall that the core of $\mathrm{Flag}(d, \mathbb{R}^n)$ is $C_{\mathrm{Flag}(d,\mathbb{R}^n)} = \mathbb{R}^* I$ (see Proposition F.2). Therefore, Statement b) follows from Proposition 5.7. Finally, the third statement follows from Theorem 5.8. since $\tilde{\Sigma}$ is weakly reversible if and only if $\mathcal{R}_{\tilde{\Sigma}}(\tilde{\mathcal{V}}) = G_{\tilde{\Sigma}} \cdot \tilde{\mathcal{V}}$ for any $\tilde{\mathcal{V}} \in \mathrm{Flag}(d, \mathbb{R}^n)$ (see Lemma 2.35). □

**Corollary 5.10** *Let $\Sigma$ and $\tilde{\Sigma}$ be systems as above and $\mathcal{V} = (V_1, \ldots, V_k)$ a reference flag in* $\mathrm{Flag}(d, \mathbb{R}^n)$, *such that* $\mathrm{Stab}(\mathcal{V}) \cap G_\Sigma \subseteq \mathbb{R}^* I$. *Then $\tilde{\Sigma}$ is reachable from $\mathcal{V}$ if and only if $\tilde{\Sigma}$ is controllable.*

**Proof.** Clearly, controllability implies reachability from any point. Conversely, if $\tilde{\Sigma}$ is reachable from $\mathcal{V}$, then $\mathcal{R}_{\tilde{\Sigma}}(\mathcal{V}) = \mathrm{Flag}(d, \mathbb{R}^n) = G_{\tilde{\Sigma}} \cdot \mathcal{V}$. By Proposition 5.9 we obtain $\mathcal{R}_{\tilde{\Sigma}}(\tilde{\mathcal{V}}) = G_{\tilde{\Sigma}} \cdot \tilde{\mathcal{V}} = \mathrm{Flag}(d, \mathbb{R}^n)$ for any $\tilde{\mathcal{V}} \in \mathrm{Flag}(d, \mathbb{R}^n)$ and therefore, by Proposition 2.31, controllability. □

### 5.2.2 Systems on projective spaces

We finish Section 5.2 with a remark on the special the case $d = (1)$, i.e., to systems on projective spaces. As described in the previous section, a system evolving on $\mathrm{GL}_n(\mathbb{R})$ induces a system on $\mathbb{RP}^{n-1}$. A more common way to induce systems on $\mathbb{RP}^{n-1}$ is via time-varying linear invertible systems (see [Hom93, Wir95]). We show that both constructions yield the same family of systems.

An invertible system $\hat{\Sigma} = (\mathbb{R}^n, U, \hat{f})$ is *time-varying linear (non-affine)* if $\hat{f}_u : \mathbb{R}^n \to \mathbb{R}^n$ is a linear map for all $u \in U$. Obviously, the set $\{0\} \subseteq \mathbb{R}^n$ is an $\hat{\Sigma}$-invariant subset. Therefore we focus on the restricted system $\hat{\Sigma}|_{\mathbb{R}^n \setminus \{0\}}$. To shorten notations we write $\hat{\Sigma} := \hat{\Sigma}|_{\mathbb{R}^n \setminus \{0\}}$. Consider the map

$$\pi : \mathbb{R}^n \setminus \{0\} \to \mathbb{RP}^{n-1},\ x \mapsto \mathrm{span}(x). \tag{40}$$

For $u \in U$, $x \in \mathbb{RP}^{n-1}$ and $v_1, v_2 \in \pi^{-1}(x)$ it is

$$\pi(\hat{f}_u(v_1)) = \pi(\hat{f}_u(v_2)).$$

In other words, the map

$$\tilde{f} : \mathbb{RP}^{n-1} \times U \to \mathbb{RP}^{n-1},\ (x, u) \mapsto \pi(f_u(v)),\ v \in \pi^{-1}(x)$$

is well defined and $\tilde{f}_u = \tilde{f}(\cdot, u)$ is bijective. This yields a new system $\tilde{\Sigma} = (\mathbb{RP}^{n-1}, U, \tilde{f})$ on $\mathbb{RP}^{n-1}$.

**Proposition 5.11** *$\tilde{\Sigma}$ is an induced system of $\hat{\Sigma}|_{\mathbb{R}^n \setminus \{0\}}$ with respect to $\pi$.*

**Proof.** Obviously, $\pi$ is surjective and for any $v \in \mathbb{R}^n \setminus \{0\}$ we obtain

$$\tilde{f}_u \circ \pi(v) = \mathrm{span}(f_u v) = \pi \circ f_u(v).$$

We show that $\pi$ is continuous and open. Recall, that the topology of $\mathbb{RP}^{n-1}$ is defined by the surjective map $\pi_{\mathcal{V}} : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{RP}^{n-1}$, $g \mapsto g(\mathcal{V})$, for a reference flag $\mathcal{V} \in \mathbb{RP}^{n-1}$. We choose $v \in \mathbb{R}^n \setminus \{0\}$ such that $\mathcal{V} := \mathrm{span}(v)$. Let

$$\pi_v : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{R}^n \setminus \{0\},\ g \mapsto g(v). \tag{41}$$

The diagram

$$\begin{tikzcd} & \mathrm{GL}_n(\mathbb{R}) \arrow[dl, "\pi_v"'] \arrow[dr, "\pi_{\mathcal{V}}"] & \\ \mathbb{R}^n \setminus \{0\} \arrow[rr, "\pi"] & & \mathbb{RP}^{n-1} \end{tikzcd}$$

commutes, since

$$\pi_{\mathcal{V}}(g) = g(\mathrm{span}(v)) = \mathrm{span}(g(v)) = \pi(g(v)) = \pi \circ \pi_v(g).$$

The maps, $\pi_{\mathcal{V}}$ and $\pi_v$ are both surjective, continuous and open. Therefore

$$\pi(\mathcal{O}) = \pi(\pi_v(\pi_v^{-1}(\mathcal{O}))) = \pi_{\mathcal{V}}(\pi_v^{-1}(\mathcal{O}))$$

is for all open subsets $\mathcal{O} \subseteq \mathbb{R}^n \setminus \{0\}$ open. Similarly,

$$\pi^{-1}(\mathcal{U}) = \pi^{-1}(\pi_{\mathcal{V}}\pi_{\mathcal{V}}^{-1}(\mathcal{U})) = \pi^{-1}\pi\pi_v\pi_{\mathcal{V}}^{-1}(\mathcal{U}) = \pi_v(\pi_{\mathcal{V}}^{-1}(\mathcal{U}))$$

is open for all open subsets $\mathcal{U} \subseteq \mathbb{RP}^{n-1}$. $\square$

We have seen, that there exists two canonical approaches to construct systems on $\mathbb{RP}^{n-1}$. Now we show that both approaches yield the same family of systems.

**Proposition 5.12** *Let $\tilde{\Sigma} = (\mathbb{RP}^{n-1}, U, \tilde{f})$ be an invertible system and $\hat{\Sigma} = (\mathbb{R}^n \setminus \{0\}, U, \hat{f})$ a time-varying linear system. For any $u \in U$ we associate $f_u \in \mathrm{GL}_n(\mathbb{R})$ to the linear map $\hat{f}_u : \mathbb{R}^n \to \mathbb{R}^n$. Then the following statements are equivalent.*

*(i) $\tilde{\Sigma}$ is an induced system of $\hat{\Sigma}$ (with respect to $\pi$)*

*(ii) $\tilde{\Sigma}$ is an induced system of the system $\Sigma = (\mathrm{GL}_n(\mathbb{R}), U, f)$ evolving on $\mathrm{GL}_n(\mathbb{R})$ (with respect to $\pi_{\mathcal{V}}$ for some reference flag $\mathcal{V}$)*

**Proof.** For any $v \in \mathbb{R}^n \setminus \{0\}$ we define $\pi_v$ as in (41). Obviously,

$$\hat{f}_u \circ \pi_v(g) = \hat{f}_u(g(v)) = f_u g(v) = \pi_v \circ f_u(g) \tag{42}$$

for any $u \in U$ and any $g \in \mathrm{GL}_n(\mathbb{R})$. In other words, $\hat{\Sigma} = (\mathbb{R}^n \setminus \{0\}, U, \hat{f})$ is an induced system of $\Sigma = (\mathrm{GL}_n(\mathbb{R}), U, f)$ with respect to $\pi_v$. Choose $v \in \mathbb{R}^n \setminus \{0\}$ and set $\mathcal{V} = \mathrm{span}(v)$. As shown before, all maps $\pi, \pi_v$ and $\pi_{\mathcal{V}}$ are surjective, open and continuous. We only have to show, that for any $u \in U$, $(i)$ $\pi \circ \hat{f}_u = \tilde{f}_u \circ \pi$ is equivalent to $(ii)$ $\pi_{\mathcal{V}} \circ f_u = \tilde{f}_u \circ \pi_{\mathcal{V}}$.
$(i) \Rightarrow (ii)$ : Using $\pi \circ \pi_v = \pi_{\mathcal{V}}$ and (42) we obtain

$$\tilde{f}_u \circ \pi_{\mathcal{V}} = \pi \circ \hat{f}_u \circ \pi_v = \pi \circ \pi_v \circ f_u = \pi_{\mathcal{V}} \circ f_u.$$

$(ii) \Rightarrow (i)$: For $x := \pi \circ \hat{f}_u$ we obtain

$$x \circ \pi_v = \pi \circ \pi_v \circ f_u = \pi_{\mathcal{V}} \circ f_u = \tilde{f}_u \circ \pi_{\mathcal{V}}.$$

For any $w \in \mathbb{R}^n \setminus \{0\}$, there exists $g \in \mathrm{GL}_n(\mathbb{R})$ such that $w = g(v)$. Therefore,

$$x(w) = x \circ \pi_v(g) = \tilde{f}_u \circ \pi_{\mathcal{V}}(g) = \tilde{f}_u(g(\mathcal{V})) = \tilde{f}_u(\mathrm{span}(g(v))) = \tilde{f}_u \circ \pi(w).$$

Hence, $x = \tilde{f}_u \circ \pi$. $\square$

# Part II

# Reachable sets of numerical iteration schemes

# 6 Classical inverse iteration

*Inverse iteration* is one of the oldest established methods for calculating eigenvectors of a given matrix. Although its basic idea goes back to the early days of numerics, inverse iteration schemes are still a topic of active research. We refer to Ipsen [Ips96, Ips97] for an overview and the state of the art, respectively Golub and Ye [GY00], Neymeyr [Ney05], Freitag and Spencer [FS07] for examples of recent research. In contrast to the standard literature, which mostly considers convergence performances for certain shift strategies, we analyze the entire structure of reachable sets. This allows us to formulate fundamental limitations on the convergence behavior of possible shift strategies and feedback laws.

Let $A \in \mathbb{R}^{n\times n}$ and denoted by $\operatorname{Spec}(A)$ the spectrum of $A$, i.e., set of eigenvalues in $\mathbb{C}$. The aim of *classical inverse iteration* is to find eigenspaces of $A$. Therefore, the corresponding system evolves on the projective space and fit in the setting of Section 5.2.

**Definition 6.1 (Classical inverse iteration system)** For $A \in \mathbb{R}^{n\times n}$ let $U_A := \mathbb{R} \setminus \operatorname{Spec}(A)$ and

$$f_A : \mathbb{RP}^{n-1} \times U_A \to \mathbb{RP}^{n-1};\ (x, u) \mapsto (I - uA)^{-1} \cdot x.$$

The corresponding system $\Sigma^{II}(A) = (\mathbb{RP}^{n-1}, U_A, f_A)$ is called *classical inverse iteration system* (with respect to the *system matrix* $A \in \mathbb{R}^{n\times n}$). Here, $\mathrm{GL}_n(\mathbb{R}) \times \mathbb{RP}^{n-1} \to \mathbb{RP}^{n-1}$, $(B, x) \mapsto B \cdot x$ denotes the canonical action on $\mathbb{RP}^{n-1}$.

In [HF00] and [HW01] the authors investigated the controllability properties of $\Sigma^{II}(A)$. We extend their work using the following strategy. First, in Sections 6.1, 6.2 and 6.3 we analyze the system groups, and respectively, the system group orbit structure of $\Sigma^{II}(A)$. Then, in Section 6.4, we show certain controllability properties of $\Sigma^{II}(A)$. In particular we give necessary and sufficient conditions for controllability of $\Sigma^{II}(A)$ (restricted on an open system group orbit) in terms of the eigenvalue constellations of $A$ (Sections 6.5 and 6.6). Then, in Section 6.7, we analyze the adherence structure of reachable sets for the cases when the restricted system is not controllable. In particular, we give conditions for the appearance of repelling phenomena. We finish this Chapter with a systematic analysis of the adherence structure of the reachable sets for the cases $n = 2, 3, 4$.

## 6.1 System group

Following Definition 2.4 the system group $G_{\Sigma^{II}(A)}$ of the inverse iteration system $\Sigma^{II}(A)$ is a group of homeomorphisms $g : \mathbb{RP}^{n-1} \to \mathbb{RP}^{n-1}$, generated by the maps $x \mapsto (A - uI)^{-1} \cdot x$, $u \in U_A$. Note that $\Sigma^{II}(A)$ can be seen as the induced system of $\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A) = (\mathrm{GL}_n(\mathbb{R}), U, \hat{f}_A)$ with respect to $\pi_x : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{RP}^{n-1}$, $g \mapsto g \cdot x$ for any reference point $x \in \mathbb{RP}^{n-1}$ (see Theorem 5.9). Here, $\hat{f}_A : (A, u) \mapsto (A - uI)^{-1}$. Obviously, the system semigroup and the system group of $\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)$ is given by

$$S_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = \left\{ \prod_{t=1}^{T} (A - u_t I)^{-1} \,\middle|\, T \in \mathbb{N}, u_t \in U_A, \right\},$$

respectively

$$G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = \left\{ \prod_{t=1}^{T_1} (A - u_t I)^{-1} \prod_{t=1}^{T_2} (A - v_t I) \,\middle|\, T_1, T_2 \in \mathbb{N}, u_t, v_t \in U_A \right\}.$$

Note that $S(A) := S_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)}$ and $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)}$ are[16] abelian subsemigroups of $\mathrm{GL}_n(\mathbb{R})$. More precisely we have:

**Proposition 6.2** *Let $m_A$ be the minimal polynomial of $A \in \mathbb{R}^{n \times n}$. $S(A)$ and $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)}$ are subsemigroups of the abelian Lie group*

$$P(A) := \{p(A) \,|\, p \in \mathbb{R}[x] \text{ coprime to } m_A\} \subseteq \mathrm{GL}_n(\mathbb{R}).$$

*The dimension of $P(A)$ is $\deg(m_A)$. $P(A)$ is a closed subgroup of the centralizer group*

$$Z(A) := \{Z \in \mathrm{GL}_n(\mathbb{R}) \,|\, ZA = AZ\}.$$

*In particular we have $P(A) = Z(A)$ and $\dim P(A) = n$ if and only if $A$ is cyclic.*

**Proof.** Obviously, $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)}$ is an abelian subsemigroup of $Z(A)$. For every $p$ coprime to $m_A$ there exist polynomials $\tilde{p}, k$ such that $1 = p\tilde{p} + km_A$ (theorem of Bezout). From the Cayley-Hamilton theorem it follows, that $p(A)^{-1} = \tilde{p}(A)$. Hence, $p(A)^{-1}$ is an element of $P(A)$. Therefore, $S(A)$ and $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)}$ are subsemigroups of $P(A)$. Moreover, any $p(A) \in P(A)$ can be expressed with $\tilde{p}(A)$ for a unique polynomial $\tilde{p}$ of degree at most $\deg m_A - 1$. It follows, that

$$P(A) = \mathrm{GL}_n(\mathbb{R}) \cap \operatorname{span}(I, A, \dots, A^{\deg(m_A)-1}) \tag{43}$$

[16]The abbreviation $S(A)$ will be very useful for the rest of the thesis. We refrain from abbreviating $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)}$ at this point, since soon we will show $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = P(A)$

is a closed subgroup of $\mathrm{GL}_n(\mathbb{R})$ and therefore a Lie group. Note that the set $P(A)$ is open in $\mathrm{span}(I, A, \ldots, A^{\deg(m_A)-1})$. Hence, $\dim P(A) = \deg m_A$. If $A$ is cyclic, the last claim follows from Proposition D.3, and respectively Proposition D.4. $\square$

Now we show the main result of this subsection.

**Theorem 6.3** *Let $\Sigma^{II}(A)$ be the classical Inverse iteration system with respect to a matrix $A \in \mathbb{R}^{n\times n}$.*

*a) $S_{\Sigma^{II}(A)}$ and $S(A)/\mathbb{R}^*I := \{s\mathbb{R}^* \mid s \in S(A)\}$ are isomorphic as semigroups.*

*b) $G_{\Sigma^{II}(A)}$ is a Lie group of dimension $\deg(m_A)-1$ isomorphic to $P(A)/\mathbb{R}^*I$. Moreover, $G_{\Sigma^{II}(A)} \times \mathbb{RP}^{n-1} \to \mathbb{RP}^{n-1}$, $(g, x) \mapsto g(x)$ is a smooth action.*

**Proof.** We show

$$G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = P(A). \tag{44}$$

Then, a) follows by Theorem 5.9, since $C_\pi = P(A) \cap \mathbb{R}^*I = \mathbb{R}^*I$. Moreover we obtain b) by Theorem 3.7.

To show (44) we analyze the system $\Sigma_{P(A)}(A) := (P(A), U_A^2, \tilde{f}_A)$ given by $U_A^2 = (\mathbb{R} \setminus \mathrm{Spec}(A))^2$ and

$$\tilde{f}_A : P(A) \times U_A^2 \to P(A),\ (B, (u, v)) \mapsto (A - uI)(A - vI)^{-1}B.$$

Note that $\Sigma_{P(A)}(A)$ is a smoothly invertible system evolving on the Lie group $P(A)$. Obviously,

$$S_{\Sigma_{P(A)}(A)} = \left\{ \prod_{t=1}^{T} (A - u_tI)(A - v_tI)^{-1} \;\middle|\; T \in \mathbb{N}, (u_t, v_t) \in U_A^2 \right\} \tag{45}$$

is a group. Moreover we obtain

$$S_{\Sigma_{P(A)}(A)} \subseteq \left\{B_1B_2 \mid B_1 \in S(A), B_2^{-1} \in S(A)\right\} \subseteq G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} \subseteq P(A). \tag{46}$$

Now we show that $\Sigma_{P(A)}(A)$ is controllable. Here, we distinguish between the case when $A$ is cyclic and the case when $A$ is non cyclic. Then, by Proposition 5.2 it follows

$$S_{\Sigma_{P(A)}(A)} = P(A) \tag{47}$$

and thus (44).

**Cyclic case:** Let us assume that $A$ is a cyclic matrix. By Theorem 5.4, $\Sigma_{P(A)}(A)$ is controllable if the following two claims are true:

**Claim 1:** *System $\Sigma_{P(A)}(A)$ is accessible.*

**Claim 2:** *Every connected component $P(A)^i$, $i \in \mathcal{I}$ of $P(A)$ has nonempty intersection with $S_{\Sigma_{P(A)}(A)}$.*

**Proof of Claim 1:** Following Proposition 5.2 it is enough to show that $\mathcal{R}(I) = S_{\Sigma_{P(A)}(A)}$ has nonempty interior in $P(A)$.

Recall that $A$ is cyclic and $P(A)$ is an open subset of the $n$ dimensional vectorspace $\operatorname{span}(I, A, \ldots, A^{n-1})$. For fixed $v_1, \ldots, v_n \in U_A$ we define $p(A) := \prod_{t=1}^{n}(A - v_t I)^{-1}$. Now we consider the map

$$\Psi : U_A^n \to P(A),\ (u_1, \ldots, u_n) \mapsto p(A) \prod_{t=1}^{n} (A - u_t I). \tag{48}$$

By construction, the image of $\Psi$ lies in $S_{\Sigma_{P(A)}(A)}$. Using the inverse function theorem we proof that $\Psi(U_A^n)$ has nonempty interior in $\operatorname{span}(I, A, \ldots, A^{n-1})$ and therefore $\operatorname{int}_{P(A)} \mathcal{R}(I) \neq \emptyset$. In particular we show, that $(u_1, \ldots, u_{n-1})$ is a regular value of $\Psi$ provided that $u_i \neq u_j$ for $i \neq j$.

We express the term $\Psi(u_1, \ldots, u_n) = \prod_{t=1}^{n}(A - u_t I)p(A)$ by *elementary symmetric polynomials* $\sigma_i^n : U_A^n \to \mathbb{R}$, $i = 0, \ldots, n$ (see Definition E.1). In particular, Proposition E.5 yields

$$\Psi(u_1, \ldots, u_n) = \sum_{t=0}^{n} (-1)^t \sigma_t^n(u_1, \ldots, u_n) e_t$$

with $e_t := A^{n-t}p(A)$, $t = 0, 1, \ldots, n$. Recall that $I, A, \ldots, A^{n-1}$ is a basis of $\operatorname{span}(I, A, \ldots, A^{n-1})$. The set $\{e_1, \ldots, e_n\}$ is linearly independent, since

$$0 = \sum_{t=1}^{n} \alpha_t e_t = p(A) \left( \sum_{t=0}^{n} \alpha_t A^{n-t} \right) \Leftrightarrow \alpha_t = 0,\ t = 1, \ldots, n.$$

Moreover, the Cayley-Hamilton theorem yields $P(A)A^k \in \operatorname{span}(I, \ldots, A^{n-1})$ for all $k \in \mathbb{N}$. It follows

$$\operatorname{span}(e_1, \ldots, e_n) = p(A) \operatorname{span}(I, \ldots, A^{n-1}) \subseteq \operatorname{span}(I, \ldots, A^{n-1}).$$

In other words, $\{e_1, \ldots, e_n\}$ is a basis of $\operatorname{span}(I, A, \ldots, A^{n-1})$ and $e_0 = \sum_{t=1}^{n} \alpha_t e_t$ for some $\alpha_t \in \mathbb{R}$, $t = 1, \ldots, n$. With respect to this basis we calculate the Jacobian $D\Psi$ of

$$\Psi(u_1, \ldots, u_n) = \sum_{t=1}^{n} \left( (-1)^t \sigma_t^n(u_1, \ldots, u_n) + \alpha_t \right) e_t$$

in the point $(u_1, \dots, u_n) \in U_A^n$. For the partial derivations we obtain

$$\begin{aligned}
\frac{\partial \Psi_t}{\partial u_k} &= \frac{\partial \left((-1)^t \sigma_t^n(u_1, \dots, u_n) + \alpha_t\right)}{\partial u_k} \\
&= (-1)^t \sum_{\substack{i_1 < \cdots < i_{t-1} \\ i_t \neq k}} u_{i_1} \cdot \dots \cdot u_{i_{t-1}} \\
&= (-1)^t \sigma_{t-1}^n(u_1, \dots, u_{k-1}, 0, u_{k+1}, \dots, u_n).
\end{aligned}$$

Now we show, that the Jacobian $D\Phi$ is invertible, if and only if $u_i \neq u_j$ for $i \neq j$. We define $f : \mathbb{R}^n \to \mathbb{R}$ by $f(u_1, \dots, u_n) = \det(D\Psi(u_1, \dots, u_n))$. Note that $\deg \frac{\partial(\Psi_t)}{\partial u_k}(u_1, \dots, u_k) = t - 1$ and therefore

$$\begin{aligned}
\deg(f) &= \deg\left(\sum_{\pi \in \mathrm{Sym}(n)} (-1)^{\mathrm{sgn}(\pi)} \frac{\partial(\Psi_t)}{\partial u_{\pi(t)}}(u_1, \dots, u_n)\right) \\
&\leq 1 + \dots + n - 1.
\end{aligned} \tag{49}$$

Now let $C_k(u_1, \dots, u_n)$ be the $k$-th column vector of $D\Psi(u_1, \dots, u_n)$, i.e.,

$$C_k(u_1, \dots, u_n) = \left((-1)^t \sigma_{t-1}^n(u_1, \dots, u_{k-1}, 0, u_{k+1}, \dots, u_n)\right)_{t=1,\dots,n}$$

Moreover, for $u = (u_1, \dots, u_{k_1}, \dots, u_{k_2}, \dots, u_n)$ we define

$$\tilde{u} := (u_1, \dots, u_{k_2}, \dots, u_{k_1}, \dots, u_n).$$

Clearly $C_k(u) = C_k(\tilde{u})$ if $k \neq k_1$ and $k \neq k_2$, since all polynomials $\sigma_t^n$ are symmetric. Moreover, for $k = k_1$ respectively $k = k_2$ we obtain

$$\begin{aligned}
C_{k_1}(u) &= \left((-1)^t \sigma_{t-1}^n(\dots, \underbrace{0}_{k=k_1}, \dots, u_{k_2}, \dots)\right)_{t=1,\dots,n} \\
&= \left((-1)^t \sigma_{t-1}^n(\dots, u_{k_1}, \dots, \underbrace{0}_{k=k_2}, \dots)\right)_{t=1,\dots,n} \\
&= C_{k_2}(\tilde{u}).
\end{aligned}$$

It follows,

$$\begin{aligned}
f(u) &= \det((C_1(u), \dots, C_{k_1}(u), \dots, C_{k_2}(u), \dots, C_n(u))) \\
&= -\det((C_1(\tilde{u}), \dots, C_{k_2}(\tilde{u}), \dots, C_{k_1}(\tilde{u}), \dots, C_n(\tilde{u}))) \\
&= -f(\tilde{u}).
\end{aligned}$$

In other words the polynomial $f$ is skew-symmetric. By Proposition E.2 $f$ can be written in the form

$$f(u_1,\dots,u_n)=\prod_{i<j}(u_i-u_j)\cdot g(u_1,\dots,u_n)$$

with a symmetric polynomial $g\in\mathbb{R}[u_1,\dots,u_n]$. Note that $\prod_{i<j}(u_i-u_j)$ has degree $1+2+\cdots+(n-1)$. By (49), $f$ has degree $1+2+\cdots+(n-1)$ and $g$ is constant. This shows, that $D\Psi$ is invertible, if and only if $u_i\neq u_j, i\neq j$. Hence, $D\Psi$ is invertible in exactly those points. By the inverse function theorem, for any $u_1,\dots,u_n\in U_A$ with $u_i\neq u_j$, $i\neq j$ there exists an open neighborhood $\mathcal{O}\subseteq U_A^n$ such that $\Psi:\mathcal{O}\to\Psi(\mathcal{O})$ is a diffeomorphism. Therefore, $\Psi(\mathcal{O})$ is an open subset of $\mathcal{R}(e)=S_{\Sigma_{P(A)}(A)}$ with respect to $P(A)$.

**Proof of Claim 2:** For an arbitrary $B\in P(A)^i$ we construct a continuous path

$$\omega:[0,1]\to P(A)\ \text{ with }\ \omega(0)=B\ \text{ and }\ \omega(1)\in S_{\Sigma_{P(A)}(A)}.$$

For the construction we need the following technical result:

**Lemma 6.4** *Let $A\in\mathbb{R}^{n\times n}$.*

*a) For all $r\in\mathbb{R}^*$ there exists $u\in\mathbb{R}\setminus\operatorname{Spec}(A)$ and a continuous path $\alpha:[0,1]\to P(A)$ such that*

$$\alpha(0)=rI\ \text{ and }\ \alpha(1)=(A-uI).$$

*b) For any normed quadratic polynomial $p\in\mathbb{R}[x]$ without real roots there exists $u\in\mathbb{R}\setminus\operatorname{Spec}(A)$ and a continuous path $\beta:[0,1]\to P(A)$ such that*

$$\beta(0)=p(A)\ \text{ and }\ \beta(1)=(A-uI)^2.$$

*c) For any $u\in\mathbb{R}\setminus\operatorname{Spec}(A)$ there exists $v\in\mathbb{R}\setminus\operatorname{Spec}(A)$ and a continuous path $\gamma:[0,1]\to P(A)$ such that*

$$\gamma(0)=(A-uI)\ \text{ and }\ \gamma(1)=(A-uI)(A-vI)^{-1}.$$

*d) For any $u\in\mathbb{R}\setminus\operatorname{Spec}(A)$ there exists $v\in\mathbb{R}\setminus\operatorname{Spec}(A)$ and a continuous path $\delta:[0,1]\to P(A)$ such that*

$$\delta(0)=(A-uI)^2\ \text{ and }\ \delta(1)=(A-uI)^2(A-vI)^{-2}.$$

**Proof of Lemma 6.4:** a) If $r < 0$ we choose $u \in \mathbb{R}$ such that $u > \lambda$ for all $\lambda \in \mathrm{Spec}(A) \cap \mathbb{R}$. Otherwise we choose $u < \lambda$ for all $\lambda \in \mathrm{Spec}(A) \cap \mathbb{R}$. Now we define $\alpha : [0,1] \to P(A)$ by

$$\alpha(t) = tA + (-ut + (1-t)r)I$$

Note that $\alpha(0) = rI$, $\alpha(1) = A - uI$ and

$$\alpha(t) = t\,(A - (u + r(1 - 1/t))) \in P(A)$$

for $t \in (0,1]$, since $1 - 1/t < 0$ and therefore $u + r(1-1/t) > u$ if $r < 0$, respectively, $u + r(1 - \frac{1}{t}) < u$ if $r > 0$.

b) Let $p(x) = (x - w)(x - \overline{w})$ with $w \in \mathbb{C} \setminus (\mathbb{R} \cup \mathrm{Spec}(A))$ We fix $u \in \mathbb{R} \setminus \mathrm{Spec}(A)$. Note that $\mathbb{C} \setminus \mathrm{Spec}(A)$ is pathwise connected since $\mathrm{Spec}(A)$ is a finite set. Therefore, there exists a continuous a path $\zeta : [0,1] \to \mathbb{C} \setminus \mathrm{Spec}(A)$ such that $\zeta(0) = w$ and $\zeta(1) = u$. For every $t \in [0,1]$ we define the quadratic polynomial

$$p_t : x \mapsto (x - \zeta(t))(x - \overline{\zeta(t)}).$$

Note that $p_t \in \mathbb{R}[x]$ for all $t \in [0,1]$. Now let $\beta : [0,1] \to P(A)$ be the path $t \mapsto p_t(A)$. By construction $\beta(t) \in P(A)$ for all $t \in [0,1]$. Moreover, $\beta(0) = P(A)$ and $\beta(1) = (A - uI)^2$.

c) By a) there exists $\alpha : [0,1] \to P(A)$ such that $\alpha(0) = I$ and $\alpha(1) = A - vI$ for some $v \in \mathrm{Spec}(A) \cap \mathbb{R}$. Therefore, the path $\gamma : [0,1] \to P(A)$, $t \mapsto (A - uI)\alpha(t)^{-1}$ fulfills what is claimed.

d) Let $\gamma$ be a path with $\gamma(0) = (A - uI)$ and $\gamma(1) = (A - uI)(A - vI)^{-1}$. Then $\delta : [0,1] \to P(A)$, $t \mapsto \gamma(t)^2$ fulfills $\delta(0) = (A - uI)^2$ and $\delta(1) = (A - uI)^2(A - vI)^{-2}$.

Now we continue the proof of Theorem 6.3. For any $B \in P(A)^i$ there exists a polynomial $p \in \mathbb{R}[x]$ such that $B = p(A)$. The real polynomial $p$ can be decomposed in the form

$$p(x) = r l_1(x) \dots l_{m_1}(x) p_{m_1+1}(x) \dots p_{m_2}(x)$$

with $r \in \mathbb{R}^*$, linear polynomials $l_j(x) = (x - u_j)$, $j = 1, \dots, m_1$, and quadratic polynomials $p_j : x \mapsto (x - w_j)(x - \overline{w_j})$ with $w_j \in \mathbb{C} \setminus (\mathbb{R} \cup \mathrm{Spec}(A))$, $j = m_1 + 1, \dots, m_2$.

By Lemma 6.4 there exist $u_0, u_j \in \mathbb{R} \setminus \mathrm{Spec}(A)$, $j = m_1 + 1 \dots, m_2$, $v_j \in \mathbb{R} \setminus \mathrm{Spec}(A)$, $j = 0, \dots, m_2$ and continuous paths $\alpha, \beta_j, \gamma_j, \delta_j : [0,1] \to P(A)$ such that

$$\begin{array}{lll}
\alpha(0) = rI, & \alpha(1) = (A - u_0 I); & \\
\beta_j(0) = p_j(A), & \beta_j(1) = (A - u_j I)^2, & j = m_1 + 1, \dots, m_2; \\
\gamma_j(0) = A - u_j I, & \gamma_j(1) = (A - u_j I)(A - v_j I)^{-1}, & j = 0, \dots, m_1; \\
\delta_j(0) = (A - u_j I)^2, & \delta_j(1) = (A - u_j I)^2 (A - v_j I)^{-2}, & j = m_1 + 1, \dots, m_2.
\end{array}$$

Recall that the product of two paths $\alpha, \beta : [0,1] \to P(A)$ with $\beta(0) = \alpha(1)$ is a path $\beta \bullet \alpha : [0,1] \mapsto P(A)$ given by

$$\beta \bullet \alpha : t \mapsto \begin{cases} \alpha(2t) & t \in [0, \frac{1}{2}]; \\ \beta(2t-1) & t \in [\frac{1}{2}, 1]. \end{cases}$$

Then $\omega : [0,1] \to P(A)$, defined by

$$\omega : t \mapsto (\gamma_0 \bullet \alpha)(t) \cdot \gamma_1(t) \dots \gamma_{m_1}(t) \cdot (\delta_{m_1+1} \bullet \beta_{m_1+1})(t)$$

is a continuous path with $\omega(0) = P(A) = B$ and

$$\omega(1) = \prod_{k=0}^{m_1} (A - u_k I)(A - v_k I)^{-1} \prod_{k=m_1+1}^{m_2} (A - u_{m_1+1} I)^2 (A - v_{m_1+1} I)^{-2}.$$

In particular we obtain $\omega(1) \in S_{\Sigma_{P(A)}(A)}$. Hence, every connected component $P(A)^i$ of $P(A)$ has an element of $S_{\Sigma_{P(A)}(A)}$.

**Non cyclic case:** We show that Equation (44) also holds for non cyclic matrices. Obviously,

$$G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(TAT^{-1})} = T G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} T^{-1} \quad \text{and} \quad P(TAT^{-1}) = T P(A) T^{-1}.$$

In particular we can assume, that $A$ is in block diagonal form

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

such that $A_1$ is cyclic and $m_A = m_{A_1}$ (see Appendix D). By Proposition 6.2 it is $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} \subseteq P(A)$. Moreover, $G_{\Sigma^{II}_{\mathrm{GL}_{n_1}(\mathbb{R})}(A_1)} = P(A_1)$ (cyclic case). By Lemma D.5 $\Phi : P(A) \to P(A_1)$, $p(A) \mapsto p(A_1)$ is an isomorphism. Thus,

$$\Phi_{|G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A_1)}} : G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A_1)} \to P(A_1)$$

is a homomorphism and by construction surjective. Therefore, $P(A_1) = G_{\Sigma^{II}_{\mathrm{GL}_{n_1}(\mathbb{R})}(A_1)}$ is isomorphic to a subgroup of $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)}$. Hence, $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = P(A)$. □

In the sequel we point out three interesting byproducts of Theorem 6.3. We start with an observation, which will be essential in the analysis of rational iteration schemes in Section 8.

**Corollary 6.5** *For all $A \in \mathbb{R}^{n \times n}$ we have*

$$P(A) = \left\{ \prod_{t=1}^{T} (A - u_t I)(A - v_t I)^{-1} \;\middle|\; T \in \mathbb{N}, u_t, v_t \in U_A, \right\}.$$

The claim follows immediately from the equations (45), (46) and (47).

For some $A \in \mathbb{R}^{n\times n}$ the system semigroup of $\Sigma^{II}(A)$ is a group, i.e., $S_{\Sigma^{II}(A)} = G_{\Sigma^{II}(A)}$, but in general this is not the case[17]. Nevertheless, from the proof of Theorem 6.3 we can deduce that the system semigroup of $\Sigma^{II}(A)$ is large in a topological sense.

**Corollary 6.6** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic.*

*a) If $u_1, \dots, u_N \in U_A$ with at least $n$ pairwise different values then*

$$\prod_{t=1}^{N} (A - u_t I)^{-1} \in \operatorname{int}_{P(A)} S(A).$$

*b)* $\operatorname{int}_{G_{\Sigma^{II}(A)}} S_{\Sigma^{II}(A)} \neq \emptyset$,

*c)* $e \in \overline{\operatorname{int}_{G_{\Sigma^{II}(A)}} S_{\Sigma^{II}(A)}}$

**Proof.** a) Without loss of generality we assume $u_i \neq u_j$ for all $i \neq j$ with $i, j \leq n$. Let $\Psi : U_A^n \to P(A)$ and $p \in \mathbb{R}[x]$ be defined as in equation (48). Recall that for any $u_1, \dots, u_n \in U_A$ with $u_i \neq u_j$, $i \neq j$ there exists an open neighborhood $\mathcal{O} \subseteq U_A^n$ of $\prod_{t=1}^{n} (A - u_t I)^{-1}$ such that $\Phi(\mathcal{O})$ is open in $P(A)$. The map $\Upsilon : P(A) \to P(A)$, $g \mapsto g^{-1} p(A)$ is a homeomorphism. Therefore $\Upsilon \circ \Phi(V)$ is open in $P(A)$. Now $\operatorname{int}_{P(A)} S(A) \neq \emptyset$ follows, since

$$\begin{aligned} \Upsilon \circ \Phi(\mathcal{O}) &= \left\{ p(A)(\Psi(u_1, \dots, u_n))^{-1} \,\middle|\, (u_1, \dots, u_n) \in \mathcal{O} \right\} \\ &= \left\{ \prod_{t=1}^{n} (A - u_t I)^{-1} \,\middle|\, (u_1, \dots, u_n) \in \mathcal{O} \right\} \\ &\subseteq S(A). \end{aligned}$$

More precisely we have shown, that every $\prod_{t=1}^{n} (A - u_t I)^{-1}$ with $u_i \neq u_j$, $i \neq j$ is an interior point of $S(A)$. Moreover, $\prod_{t=n+1}^{N} (A - u_t I)^{-1} : P(A) \to P(A)$ is a homeomorphism. Therefore

$$\prod_{t=1}^{N} (A - u_t I)^{-1} \in \prod_{t=n+1}^{N} (A - u_t I)^{-1} \left( \Upsilon \circ \Phi(\mathcal{O}) \right) \subseteq S(A).$$

Now we prove c) which immediately implies b). Choose $u_1, \dots u_n \in U_A \setminus \{0\}$ such that $u_i \neq u_j$ for $i \neq j$. For any $r \in \mathbb{R}^*$ large enough[18] we have $B_r := \prod_{t=1}^{n} (A - r u_t I)^{-1} \in \operatorname{int}_{P(A)} S(A)$. Recall that $\Sigma^{II}(A)$ is an induced

[17] We will see examples for both cases in Sections 6.5 and 6.8

[18] $r$ should be large enough such that $r u_i \neq \operatorname{Spec}(A)$ for all $i \in U_A$.

system of $\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)$ with respect to some $\pi : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{RP}^{n-1}$. By Theorem 5.9 and Theorem 6.3 we obtain

$$C_\pi = P(A) \cap \mathbb{R}^* I = P(A) \cap \mathbb{R}^* I = \mathbb{R}^* I.$$

By Theorem 3.7 there exists a continuous group homomorphism $\Phi : P(A) \to G_{\Sigma^{II}(A)}$ such that $\Phi(\mathrm{int}_{P(A)} S(A)C_\pi) = \mathrm{int}_{G_{\Sigma^{II}(A)}} S_{\Sigma^{II}(A)}$. Therefore, $\Phi(B_r) \in \mathrm{int}_{G_{\Sigma^{II}(A)}} S_{\Sigma^{II}(A)}$ for all $r \in \mathbb{R}^+$ large enough. Following the construction of $\Phi$ (see Theorem 3.7) it follows $\Phi(Bc) = \Phi(B)$ for all $B \in P(A)$ and $c \in C_\pi$. It follows

$$\begin{aligned}
\lim_{r\to\infty} \Phi(B_r) &= \lim_{r\to\infty} \Phi\left( r^n \prod_{t=1}^{n} \left( \frac{1}{r} A - u_t I \right)^{-1} \right) \\
&= \Phi\left( \lim_{r\to\infty} \prod_{t=1}^{n} \left( \frac{1}{r} A - u_t I \right)^{-1} \right) \\
&= \Phi\left( \prod_{t=1}^{n} \frac{1}{u_t} I \right) \\
&= e.
\end{aligned}$$

Hence, $e \in \overline{\mathrm{int}_{G_{\Sigma^{II}(A)}} S_{\Sigma^{II}(A)}}$. □

Finally, Theorem 6.3 provides an interesting property of the set of linear decomposable polynomials, i.e., of the set

$$\mathcal{L} := \left\{ r \prod_{t=1}^{T} (x - u_t) \,\middle|\, r \in \mathbb{R}^*, u_t \in \mathbb{R} \right\}.$$

**Corollary 6.7** *For any $p, m \in \mathbb{R}[x]$ such that $p$ and $m$ are coprime, there exist $q_1, q_2 \in \mathcal{L}$ with $\deg q_1 = \deg q_2$ such that*

$$q_1 p = q_2 \mod m$$

**Proof.** Let $A \in \mathbb{R}^{n\times n}$ be a matrix with minimal polynomial $m \in \mathbb{R}[x]$. By Theorem 6.3 and Corollary 6.5 it is

$$P(A) = \left\{ \prod_{t=1}^{T} (A - u_t I)(A - v_t I)^{-1} \,\middle|\, T \in \mathbb{N}, u_t, v_t \in U_A, \right\}.$$

For any $p$ coprime to $m$ it is $p(A) \in P(A)$. Therefore, there exists $T \in \mathbb{N}$ and $u_1, \ldots, u_T, v_1, \ldots, v_T \in \mathbb{R}$ such that

$$p(A) = q_2(A)(q_1(A))^{-1}$$

with $q_1(x) = \prod_{t=1}^{T}(x - u_t)$ and $q_2(x) = \prod_{t=1}^{T}(x - v_t)$. Hence, $q_1 p = q_2 + km$ for some $k \in \mathbb{R}[x]$. □

## 6.2 Lie group types of $G_{\Sigma^{II}(A)}$

From Theorem 6.3 we know, that $G_{\Sigma^{II}(A)}$ is a real abelian Lie group of dimension $m_A - 1$. Therefore it must be isomorphic to $D \times \mathbb{R}^{k_1} \times \mathbb{T}^{k_2}$ with a discrete group $D$ (see [GOV97], Theorem 2.12). Here we denote the additive group of real numbers with $\mathbb{R}$ and the $k$-dimensional torus with

$$\mathbb{T}^k := \underbrace{\mathbb{S} \times \cdots \times \mathbb{S}}_{k\text{-times}}.$$

Note that $\mathbb{R}^* \cong C_2 \times \mathbb{R}$ and $\mathbb{C}^* \cong \mathbb{R} \times \mathbb{S}$ where $C_2$ denotes the group with two elements. In this section we explicitly determine the Lie group type of $G_{\Sigma^{II}(A)}$, in terms of the minimal polynomial $m_A$ of $A$. Note that parts of this results were implicitly used (see [KM83], Theorem 1), but to our knowledge, not explicitly written down and proved.

**Theorem 6.8** *Let $m_A = l_1^{\alpha_1} \dots l_{k_1}^{\alpha_{k_1}} q_1^{\beta_1} \dots q_{k_2}^{\beta_{k_2}}$ be the minimal polynomial of $A \in \mathbb{R}^{n\times n}$ with coprime linear factors $l_1, \dots, l_{k_1}$ and irreducible coprime quadratic factors $q_1, \dots, q_{k_2}$. The group $G_{\Sigma^{II}(A)}$ is isomorphic to*

$$C_2^{k_1} \times \mathbb{R}^{\alpha_1+\cdots+\alpha_{k_1}+2\beta_1+\cdots+2\beta_{k_2}-k_2-1} \times \mathbb{T}^{k_2}.$$

**Proof.** Equivalently we show that $P(A) = G_{\Sigma^{II}(A)}/\mathbb{R}^* I$ is isomorphic to

$$(\mathbb{R}^* \times \mathbb{R}^{\alpha_1-1}) \times \cdots \times (\mathbb{R}^* \times \mathbb{R}^{\alpha_{k_1}-1}) \times (\mathbb{C}^* \times \mathbb{C}^{\beta_1-1}) \times \cdots \times (\mathbb{C}^* \times \mathbb{C}^{\beta_{k_2}-1}).$$

It is sufficient to prove this relation for the cases $m_A = (t-\lambda)^\alpha$ for $\lambda \in \mathbb{R}$ and $m_A = ((t-\lambda)(t-\overline{\lambda}))^\beta$ for $\lambda \in \mathbb{C} \setminus \mathbb{R}$. The minimal polynomial is a product of such polynomials. Thus, the Lie group type of $P(A)$ can be deduced by Lemma D.5.
(i) Let $m_A = l^\alpha$ with a linear polynomial $l(x) = (x - \lambda)$. Without loss of generality we can assume, that

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix} \in \mathbb{R}^{\alpha\times\alpha}$$

since $P(TAT^{-1} - \lambda I) \cong P(A)$and $P(A) \cong P(A_1)$ if $m_A = m_{A_1}$ (see Lemma D.5). Recall that

$$P(A) = \operatorname{span}\left(I, A, \dots, A^{\alpha-1}\right) \cap \mathrm{GL}_\alpha(\mathbb{R}).$$

The matrix $p(A)$ is invertible, if and only if $l(t) = t$ is coprime to $p$, i.e. $p(0) \neq 0$. Therefore, $P(A)$ is the set of those matrices, which can be expressed in the form

$$B = a_0 I + a_1 A + \cdots + a_{\alpha-1} A^{\alpha-1}$$

with $\alpha_0 \in \mathbb{R}^*$ and $\alpha_i \in \mathbb{R}$ for $i = 1, \ldots, \alpha - 1$. Since $A^\alpha = 0$ we get

$$\begin{aligned} & \left(a_0 I + a_1 A + \ldots a_{\alpha-1} A^{\alpha-1}\right)\left(b_0 I + b_1 A + \ldots b_{\alpha-1} A^{\alpha-1}\right) \\ = \; & (a_0 b_0 I + (a_0 b_1 + a_1 b_0) A + \ldots (a_0 b_{\alpha-1} + \cdots + a_{\alpha-1} b_0) A^{\alpha-1}) \end{aligned}$$

for the product of two elements of $P(A)$. Therefore, $P(A)$ can be expressed as the abelian matrix Lie group

$$\left\{ \left( \begin{array}{cccc} a_0 & a_1 & \ldots & a_{\alpha-1} \\ & a_0 & \cdots & \\ & & \ddots & a_1 \\ & & & a_0 \end{array} \right) \middle| \, a_0 \in \mathbb{R}^*, \alpha_i \in \mathbb{R}, \, i = 1, \ldots, \alpha - 1 \right\}.$$

Obviously, $P(A)$ has two connected components and dimension $\alpha$. Therefore, $P(A)$ has to be diffeomorphic to $C_2 \times \mathbb{R}^{\alpha_1} \times \mathbb{T}^{\alpha_2}$ with $\alpha_1 + \alpha_2 = \alpha$. Moreover, booth components are convex subsets in $\mathbb{R}^{n \times n}$ and therefore simply connected. Thus, $\alpha_2$ has to be zero. We conclude

$$P(A) \cong \mathbb{R}^* \times \mathbb{R}^{\alpha-1}.$$

(ii) Now let $m_A = q^\beta$ with a quadratic irreducible polynomial $q$. As in (i) we apply Lemma D.5 to reduce our analysis of a certain type. Without loss of generality we can assume, that $A$ is a block matrix

$$A = \left( \begin{array}{cccc} B & I & & \\ & B & I & \\ & & \ddots & I \\ & & & B \end{array} \right) \in \mathbb{R}^{2\beta \times 2\beta} \text{ with } B = \left( \begin{array}{cc} 0 & 1 \\ -1 & 0 \end{array} \right)$$

since

$$P(J) = P\left(\frac{1}{\operatorname{Im}(\lambda)}(J - \operatorname{Re}(\lambda) I)\right) \text{ for } J = \left( \begin{array}{cc} \operatorname{Re}(\lambda) & \operatorname{Im}(\lambda) \\ -\operatorname{Im}(\lambda) & \operatorname{Re}(\lambda) \end{array} \right).$$

Every polynomial of $A$ is again a matrix of block-type, with blocks $(p(A))_{i,j} \in \mathbb{R}^{2\times 2}$, $i, j = 1, \ldots, \beta$. Obviously, $(p(A))_{i,j} = 0$ for $i < j$. The diagonal blocks are equal, i.e., $(p(A))_{i,i} = (p(A))_{j,j}$. They are invertible, if and only if $p$ is coprime to $m_A$. By induction it can be shown, that for $i > j$ the block

$(A^k)_{i,j}$ is a polynomial of $B$ and equal to $(A^k)_{\tilde{i},\tilde{j}}$ if $i-j=\tilde{i}-\tilde{j}$. It follows, that

$$\begin{aligned} P(A) &= \{p(A) \,|\, p \text{ coprime } x^2+1\} \\ &= \left\{a_0 I + \cdots + a_{2\beta-1}A^{2\beta-1} \,\middle|\, a_0,\ldots,a_{2\beta-1}\in\mathbb{R}\right\} \cap \mathrm{GL}_{2\beta}(\mathbb{R}) \end{aligned}$$

is a subgroup of the abelian matrix group

$$\tilde{P}(A) := \left\{ \left( \begin{array}{cccc} p(B) & p_1(B) & \ldots & p_{\beta-1}(B) \\ & \ddots & \ddots & \\ & & \ddots & p_1(B) \\ & & & p(B) \end{array} \right) \,\middle|\, \begin{array}{l} p(B) \text{ invertible}, \\ p_i \in \mathbb{R}[x] \end{array} \right\}.$$

Now we show, that $\tilde{P}(A)$ is isomorphic to the connected Lie group $\mathbb{S}\times\mathbb{R}^{2\beta-1}$.

We express $\tilde{P}(A)$ with a semidirect product of groups isomorphic to $\mathbb{C}^*$, respectively $\mathbb{C}$. We define the following subgroups of $\mathrm{GL}_{2\beta}(\mathbb{R})$

$$P_1 := \left\{ \left( \begin{array}{ccc} p(B) & & \\ & \ddots & \\ & & p(B) \end{array} \right) \,\middle|\, p \text{ coprime } x^2+1 \right\}$$

and for $k=2,\ldots,\beta$

$$P_k := \left\{ M_{p(B)} := \left( \begin{array}{ccccc} I & 0\cdots 0 & \overbrace{p(B)}^{\text{block } (1,k)} & 0\ldots 0 & 0 \\ & \ddots & & \ddots & \\ & & \ddots & & p(B) \\ & & & \ddots & \\ & & & & I \end{array} \right) \,\middle|\, p_i \in \mathbb{R}[x] \right\}.$$

Recall that the field $\mathbb{C}$ is isomorphic to the field of matrices

$$\mathbb{F} := \left\{ \left( \begin{array}{cc} a & b \\ -b & a \end{array} \right) \,\middle|\, a,b\in\mathbb{R} \right\}.$$

Note that $\{p(B)\,|\,p\in\mathbb{R}[x]\}$ coincides with the set $\mathbb{F}$ such that matrix multiplication in $\mathbb{F}$ and $\{p(B)\,|\,p\in\mathbb{R}[x]\}$ are corresponding. It follows, that the group $P_0$ is isomorphic to $\mathbb{C}^*$. Moreover, matrix multiplication in $P_k, k=2,\ldots,\beta$ corresponds to the addition of two $(1,k)$-block elements, i.e., $M(p_a(B))M(p_b(B)) = M(p_a(B)+p_b(B))$. Therefore, $P_i$ is isomorphic to the additive group $\mathbb{C}$. Every group $P_i$ is a normal subgroup of $\tilde{P}(A)$ since it is abelian. Moreover, from the structure of the elements it is clear that

$\tilde{P}(A) = P_0 P_1 \dots P_{\beta-1}$ and $P_i \cap P_j = \{I\}$ for $i \neq j$. Hence, $\tilde{P}_A$ is a semidirect product of the groups $P_0, \dots, P_{\beta-1}$. We conclude

$$\tilde{P}(A) \cong \mathbb{C}^* \times \mathbb{C}^{\beta-1} \cong \mathbb{S} \times \mathbb{R}^{2\beta-1}.$$

Now we show, that $\tilde{P}(A) = P(A)$. By Proposition 6.2 it becomes clear, that $\dim P(A) = 2\beta = \dim \tilde{P}(A)$. Therefore, the factor group $\tilde{P}(A)/P(A)$ is discrete and must be trivial, since $\tilde{P}(A)$ is connected. Hence $P(A) = \tilde{P}(A)$.
(iii) Now let $m_A = l_1^{\alpha_1} \dots l_{k_1}^{\alpha_{k_1}} q_1^{\beta_1} \dots q_{k_2}^{\beta_{k_2}}$ be the minimal polynomial of $A$, with $l_i, q_i$ as in the statement of Theorem 6.8. The Jordan canonical form is a block matrix with blocks $L_1, \dots, L_{k_1}, Q_1, \dots, Q_{k_2}$ with $L_i \in \mathbb{R}^{\alpha_i \times \alpha_i}$, $i = 1, \dots k_1$ and $Q_j \in \mathbb{R}^{2\beta_j \times 2\beta_j}$, $j = 1, \dots, k_2$. By Lemma D.5 we conclude

$$P(A) \cong P(L_1) \times \dots P(L_{k_1}) \times P(Q_1) \times \dots \times P(Q_{k_2}).$$

Thus, the claim follows from (i) and (ii). $\square$

## 6.3 Structure of orbits

Now we analyze the structure of system group orbits of $\Sigma^{II}(A)$. In particular we show that, similar to the case of complex inverse iteration (see [HF00]), there is always one "large" orbit, which is open and dense in $\mathbb{RP}^{n-1}$, provided $A$ is cyclic.

For inverse iteration systems $\Sigma^{II}(A)$ the state space has a canonical decomposition in $\Sigma$-invariant subspaces. This decomposition is related with the $A$-invariant subspaces. We will use the following notation:

**Definition 6.9** Let $A$ be cyclic. We denote the set of $A$-*invariant subspaces* with $\mathrm{Inv}_A$. For $W \in \mathrm{Inv}_A \setminus \{0\}$ we define $\mathrm{Inv}_A^W := \{V \in \mathrm{Inv}_A \mid V \subseteq W, V \neq W\}$ and

$$N_W := W \setminus \bigcup_{V \in \mathrm{Inv}_A^W} V \subseteq \mathbb{R}^n.$$

Let $\pi : \mathbb{R}^n \setminus \{0\} \to \mathbb{RP}^{n-1}$ the canonical projection, i.e. $\pi(x) = \mathrm{span}(x)$. We define

$$\mathcal{N}_W := \pi(N_W) \subseteq \mathbb{RP}^{n-1}.$$

In the case $W = \mathbb{R}^n$ we write $N_A := N_{\mathbb{R}^n}$, respectively $\mathcal{N}_A := \mathcal{N}_{\mathbb{R}^n}$.

**Proposition 6.10** *The set* $\mathcal{N}_W$ *is* $\Sigma$*-invariant for all* $W \in \mathrm{Inv}_A$.

**Proof.** Recall that $f_u : \mathbb{RP}^{n-1} \to \mathbb{RP}^{n-1}$ is bijective for all $u \in U_A$. Moreover, $f_u(\pi(V)) = \pi\left((A - uI)^{-1}V\right) = \pi(V)$ for all $V \in \mathrm{Inv}_A$. Hence,

$$\begin{aligned}
f_u(\mathcal{N}_W) &= f_u(\pi(N_W)) \\
&= f_u(\pi(W)) \setminus f_u\left(\pi\left(\bigcup_{V \in \mathrm{Inv}_A^W} V\right)\right) \\
&= \pi\left(W \setminus \left(\bigcup_{V \in \mathrm{Inv}_A^W} f_u(V)\right)\right) \\
&= \mathcal{N}_W
\end{aligned}$$

□

Now we show, that the sets $\mathcal{N}_W$, $W \in \mathrm{Inv}_A$ are system group orbits of $\Sigma^{II}(A)$, i.e., $G_{\Sigma^{II}(A)}x = \mathcal{N}_W$ for $x \in \mathcal{N}_W$.

**Lemma 6.11** *Let $A$ be cyclic and* $W \in \mathrm{Inv}_A$.

a) *The map* $P(A) \times N_W \to N_W$, $(B, v) \mapsto Bv$ *is a transitive group action. Moreover,* $\mathrm{Stab}_v = \{B \in P(A) \mid B_{|W} = \mathrm{id}_{|W}\}$.

*b) The map* $G_{\Sigma^{II}(A)} \times \mathcal{N}_W \to \mathcal{N}_W$, $(g, x) \mapsto g \cdot x$ *is a transitive group action. Moreover,* $\mathrm{Stab}_x = \{g \in G_{\Sigma^{II}(A)} \,|\, g_{|W} = \mathrm{id}_{|W}\}$.

**Proof.** Both maps are group actions. In particular, $p(A)v \in V$ for any $p(A) \in P(A)$ and any $A$-invariant subspace $V \subseteq W$. Analogously, $p(A)v \in V$ implies $p(A)^{-1}p(A)v \in V$ and therefore $v \in W \setminus V$ implies $P(A)v \in W \setminus V$. Hence, $p(A)v \in N_W$ for all $v \in N_W$. Thus, $p(A)N_W = N_W$ for all $p(A) \in P(A)$. Moreover, $g \cdot \mathcal{N}_A = \mathcal{N}_A$ follows immediately from Proposition 6.10.

Now we show transitivity of $P(A) \times N_W \to N_W$. Let $v \in N_W$. Since $v \in W$, but $v \notin V \in \mathrm{Inv}_A$ for $V \subsetneqq W$ we have $\mathrm{span}(v, Av, \dots, A^{k-1}v) = W$ ($k := \dim W$). In other words, every $w \in W$ can be written in the form

$$w = \sum_{i=0}^{k-1} w_i A^i v = p(A)v$$

for some $w_0, \dots, w_{k-1} \in \mathbb{R}$ and $p(t) = \sum_{i=0}^{k-1} w_i t^i$. Assume that $w \in N_W$. Then $w, Aw, \dots, A^{k-1}w$ is a basis of $W$. Therefore, $p(A)$ is invertible, since it maps the basis $v, Av, \dots A^{k-1}v$ on the basis $w, Aw, \dots, A^{k-1}w$. Hence, for all $v, w \in N_W$ there exists $p(A) \in P(A)$ such that $p(A)v = w$. Moreover, it follows $p(A)v = v$ if and only if $p(A)_{|W} = \mathrm{id}_{|W}$, since $p(A)$ maps the basis $v, Av, \dots, A^{k-1}v$ on itself.
b) For $x, y \in \mathcal{N}_A$ we choose $v, w \in N_A$ and $B \in P(A)$ such that $Bv = w$ and $\pi(v) = x$ and $\pi(w) = y$. The map $g : \mathbb{RP}^{n-1} \to \mathbb{RP}^{n-1}$, $z \mapsto B \cdot z$ is element of $G_{\Sigma^{II}(A)}$ and we obtain

$$g(x) = B \cdot \pi(v) = \pi(Bv) = \pi(w) = y.$$

Moreover, $g(x) = x$ with $x \in \mathcal{N}_A$ if and only if $g_{|W} = \mathrm{id}_{|W}$. □

Now we show, that the adherence structure of the system group orbits can be described by the lattice structure of the $A$-invariant subspaces.

**Definition 6.12** Let $\mathrm{Inv}_A$ be the set of nontrivial $A$-invariant subspaces. The *subspace graph*[19] $\mathcal{G}_A = (\leftarrow, \mathrm{Inv}_A \setminus \{0\})$ is given by the vertices $\mathrm{Inv}_A$ and the relation

$$U \leftarrow V :\Leftrightarrow U \subseteq V.$$

Note that the subspace graph of $A$ is finite, provided $A$ is cyclic. The following example illustrates this concept.

[19] Note that $\mathrm{Inv}_A$ together with the relation $U \leftarrow V \;:\Leftrightarrow\; U \subseteq V$ forms a lattice structure . The subspace graph is a subgraph of the corresponding Hasse diagram.

**Example 6.13** For

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

we obtain $\mathrm{Inv}_A = (\mathrm{span}\,(e_2), \mathrm{span}\,(e_3), \mathrm{span}\,(e_1, e_2), \mathbb{R}^3)$. The subspace graph is given by

$$\begin{tikzcd}
\mathrm{span}\,(e_3) & \mathbb{R}^3 \arrow[l] \arrow[r] \arrow[d] & \mathrm{span}\,(e_3) \\
& \mathrm{span}\,(e_1, e_2) \arrow[ur] &
\end{tikzcd}$$

**Theorem 6.14** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic, $\Sigma^{II}(A)$ be the inverse iteration system of $A$. The orbit graph $\mathcal{G}_O(\Sigma^{II}(A))$ and the subspace graph $\mathcal{G}_A$ are isomorphic.*

**Proof.** By Lemma 6.11 the sets $\mathcal{N}_W$, $W \in \mathrm{Inv}_A$ coincide with the system group orbits of $\Sigma^{II}(A)$. Therefore, the map

$$\Psi : \mathrm{Inv}_A \to \{G_{\Sigma^{II}(A)} \cdot x \,|\, x \in \mathbb{RP}^{n-1}\},\ W \mapsto \mathcal{N}_W$$

is surjective. Moreover, $\Psi$ is injective, since $V \neq W$ implies $\mathcal{N}_W \neq \mathcal{N}_V$.

Finally we show, that $\Psi$ preserves the graph structure, i.e. $V \subseteq W$ if and only if $\Psi(V) \leftarrow \Psi(W)$. Let $v \in V$ such that $G_{\Sigma^{II}(A)} \cdot v = \mathcal{N}_V$. Then $v \in \overline{N_W}$ and therefore

$$\pi(v) \in \pi(\overline{N_W}) \subseteq \overline{\pi(N_W)} = \overline{\mathcal{N}_W}.$$

The set $\overline{\mathcal{N}_W}$ is a union of system group orbits (see Proposition 3.10). Therefore,

$$\mathcal{N}_V = G_{\Sigma^{II}(A)} \cdot \pi(v) \subseteq G_{\Sigma^{II}(A)} \cdot \overline{\mathcal{N}_W} = \overline{\mathcal{N}_W}.$$

Hence, $V \subseteq W$ implies $\mathcal{N}_V \subseteq \overline{\mathcal{N}_W}$. Conversely, if $v \in V \setminus W$, then there exists an open set $\mathcal{O} \subseteq \mathbb{R}^n/W$ such $y \in \mathcal{O}$. It follows

$$\pi(v) \in \mathcal{N}_V \cap \pi(\mathcal{O}) \subseteq \mathbb{RP}^{n-1} \setminus \mathcal{N}_W.$$

Hence, $\Psi(V) \not\leftarrow \Psi(W)$. □

In particular, there is always one system group orbit of $\Sigma^{II}(A)$ which corresponds to $\mathbb{R}^n \in \mathrm{Inv}_A$. Now we show, that this orbit is open and dense in the state space.

**Theorem 6.15** *Let $A \in \mathbb{R}^{n\times n}$ and $\Sigma^{II}(A) = (\mathbb{RP}^{n-1}, U_A, f_A)$ be the classical inverse iteration system corresponding to $A$.*

*a) If $A$ is cyclic then there exists one open and dense system group orbit. More precisely, $\mathcal{N}_A$ is open and dense in $\mathbb{RP}^{n-1}$ and $G_{\Sigma^{II}(A)} \cdot x = \mathcal{N}_A$ for all $x \in \mathcal{N}_A$.*

*b) If $A$ is not cyclic then every system group orbit has empty interior in $\mathbb{RP}^{n-1}$.*

**Proof.** a) Since $A$ is cyclic, it has finitely many $A$-invariant subspaces (see D.3). Therefore, $\bigcup_{V \in \mathrm{Inv}_A} V$ is the union of finitely many proper subspaces. Hence, $N_A = \mathbb{R}^n \setminus \bigcup_{V \in \mathrm{Inv}_A} V$ is open and dense in $\mathbb{R}^n$. Moreover, $\mathcal{N}_A = \pi(N_A)$ is open and dense in $\mathbb{RP}^{n-1}$, since $\pi$ is open, continuous and surjective.
b) If $A$ is not cyclic, every $x \in \mathbb{R}^n$ is element of some proper $A$-invariant subspace $W$. Therefore, $\mathcal{R}(x) \subseteq \pi(W)$. The claim follows, since $\dim \pi(W) = \dim W - 1 < \dim \mathbb{RP}^{n-1}$. $\square$

## 6.4 Controllability properties

In this section we discuss controllability properties of $\Sigma^{II}(A)$. If $W$ is a proper $A$-invariant subspace, then $x \in \pi(W)$ implies $\mathcal{R}(x) \subseteq \pi(W)$. Therefore, $\Sigma^{II}(A)$ is not controllable, provided there exists proper $A$-invariant subspaces[20]. On the other hand, in the previous section we have shown, that there exists an open and dense orbit, provided $A$ is cyclic. Following Section 3.8 we can restrict $\Sigma^{II}(A)$ to $\mathcal{N}_A$.

**Definition 6.16 (Restricted inverse iteration system)** Let $A \in \mathbb{R}^{n\times n}$ be cyclic and $\Sigma^{II}(A) = (\mathbb{RP}^{n-1}, U_A, f_A)$ be the corresponding classical inverse iteration system. Then

$$\Sigma^{II}(A)_{|\mathcal{N}_A} = (\mathcal{N}_A, U_A, f_{A|_{\mathcal{N}_A \times U_A}})$$

is the *restricted inverse iteration system (with respect to $\mathcal{N}_A$).*

Now the question arises if $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable. The analogous question for complex arithmetic was solved by Helmke and Fuhrmann (see [HF00]). Here, the restricted system is controllable if and only if $A$ is cyclic. For real arithmetic, Helmke and Wirth already pointed out, that there exists families of cyclic matrices such that $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is not controllable (see [HW01]). Moreover, using a topological approach via controllable sets, Helmke and Wirth showed the following:

**Theorem 6.17 (Helmke, Wirth [HW01])** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic and $m_A$ its minimal polynomial. Then the following statements are equivalent.*

*(i) System $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable.*

*(ii) System $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is approximatively reachable from some $x \in \mathcal{N}_A$.*

*(iii) There exists $r \in \mathbb{R}^*$ and a control sequence $u = (u_0, \dots, u_{T-1})$ such that*

$$\prod_{t=0}^{T-1} (A - u_t I) = rI$$

*and for $\Phi : U_A^T \times \mathcal{N}_A \to \mathcal{N}_A$, $(u_0, \dots, u_{T-1}, x) \mapsto \prod_{t=0}^{T-1}(A - u_t I)\cdot x$ the rank-condition*

$$\operatorname{rank} \frac{\partial \Phi(x,u)}{\partial u} = n-1$$

*holds for all $x \in \mathcal{N}_A$.*

*(iv) There exists a polynomial $k \in \mathbb{R}[x]$ and a constant $\alpha \in \mathbb{R}^*$. such that $\alpha + k m_A \in \mathcal{L}$ and $\alpha + km$ has at least $n-1$ different roots.*

[20] Note that this is always the case if $n \geq 3$.

**Proof.** All proofs are given in [HW01]. See Theorem 3 for the implications $(i) \Leftrightarrow (ii) \Leftrightarrow (iii)$ and Theorem 5 for $(i) \Leftrightarrow (iv)$. □

In [HW01] the authors widely neglected the fact that the reachable sets are semigroup orbits. We are able to extend their results in different aspects. In particular, we show equivalent conditions for controllability of the restricted systems with respect to the properties of the entire system.

**Theorem 6.18** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic. Consider the inverse iteration system $\Sigma^{II}(A)$. Then the following statements are equivalent.*

*(i)* $\Sigma^{II}(A)_{|\mathcal{N}_A}$ *is controllable.*

*(ii)* $S_{\Sigma^{II}(A)} = G_{\Sigma^{II}(A)}$.

*(iii)* $\Sigma^{II}(A)$ *is approximatively reachable from some* $x \in \mathcal{N}_A$.

*(iv) For all* $x \in \mathcal{N}_A$, *all* $y \in \mathbb{RP}^{n-1}$ *and all neighborhoods* $\mathcal{U} \subseteq \mathbb{RP}^{n-1}$ *of* $y$ *there exists a control sequence* $u_0, \dots u_N \in U$ *such that* $x_n \in \mathcal{U}$.

*(v)* $\Sigma^{II}(A)$ *is weakly reversible.*

*(vi)* $\Sigma^{II}(A)$ *is densely reachable.*

*(vii) The reachable structure and the orbit structure of* $\Sigma^{II}(A)$ *coincide.*

*(viii) There exists a finite number of different reachable sets.*

*(ix)* $S(A)\mathbb{R}^* := \{Br \,|\, B \in S(A), r \in \mathbb{R}^*\} = P(A)$.

**Proof.** $(i) \Leftrightarrow (ii)$: Since $G_{\Sigma^{II}(A)}$ acts transitively on $\mathcal{N}_A$, statement $(ii)$ implies controllability of $\Sigma^{II}(A)_{|\mathcal{N}_A}$. Conversely, controllability of $\Sigma^{II}(A)_{|\mathcal{N}_A}$ implies $S_{\Sigma^{II}(A)_{|\mathcal{N}_A}} = G_{\Sigma^{II}(A)_{|\mathcal{N}_A}}$ by Theorem 2.39. Recall that $\mathcal{N}_A$ is dense in $\mathbb{RP}^{n-1}$ (see Theorem 6.15). Therefore, $S_{\Sigma^{II}(A)_{|\mathcal{N}_A}} = G_{\Sigma^{II}(A)_{|\mathcal{N}_A}}$ is equivalent to $S_{\Sigma^{II}(A)} = G_{\Sigma^{II}(A)}$ by Theorem 3.12.
$(ii) \Rightarrow (v)$: Obviously, $(ii)$ implies $\mathcal{R}(x) = G_{\Sigma^{II}(A)} \cdot x$ for all $x \in \mathbb{RP}^{n-1}$. Now, weakly reversibility follows by Lemma 2.35.
$(iv) \Rightarrow (iii)$: Obviously, $(iv)$ implies $\overline{\mathcal{R}(x)} = \mathbb{RP}^{n-1}$ and therefore approximative reachability from $x$.
$(i) \Rightarrow (iv)$: If $y \in \mathcal{N}_A$ there exists a finite control sequence $u_0, \dots, u_N$, $N \in \mathbb{N}$ such that $x_N = y$ for $x_{t+1} = f^{II}(x_t, u_t)$, $x_0 = x$. If $y \in \partial\mathcal{N}_A$, then the existence of a control sequence $u_0, \dots, u_N$ with $x_n \in \mathcal{U}$ is assured by Theorem 2.46,b).
$(v) \Rightarrow (iii)$: If $\Sigma^{II}(A)$ is weakly reversible, then $\mathcal{R}(x) = G_{\Sigma^{II}(A)} \cdot x$ for all $x \in \mathbb{RP}^{n-1}$ (see Lemma 2.35). In particular $S_{\Sigma^{II}(A)} \cdot x = \mathcal{N}_A$ for any

$x \in \mathcal{N}_A$, since $G_{\Sigma^{II}(A)}$ acts transitively on $\mathcal{N}_A$. Hence, $\Sigma^{II}(A)$ is approximatively reachable from $x \in \mathcal{N}_A$, since $\mathcal{N}_A$ is dense in $\mathbb{RP}^{n-1}$.
*(iii)* $\Leftrightarrow$ *(i)* $\Leftrightarrow$ *(vi)*: By Theorem 3.12, $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is abelian and smoothly invertible. Moreover, $G_{\Sigma^{II}(A)_{|\mathcal{N}_A}}$ has a Lie group structure (isomorphic to $P(A)/\mathbb{R}^*I$) such that $G_{\Sigma^{II}(A)_{|\mathcal{N}_A}} \times \mathcal{N}_A \to \mathcal{N}_A$, $(g,x) \mapsto g(x)$ is smooth and, by Lemma 6.11, transitive. By Corollary 6.6, we have $\mathrm{int}_{G_{\Sigma^{II}(A)}} S_{\Sigma^{II}(A)} \neq \emptyset$. By Corollary 2.49 it follows, that *(i)* is equivalent to dense reachability of $\Sigma^{II}(A)_{|\mathcal{N}_A}$.
*(v)* $\Leftrightarrow$ *(vii)*: This equivalence is a immediate consequence of Theorem 4.6.
*(vii)* $\Rightarrow$ *(viii)*: By Theorem 6.14 there exists a bijection between the set of $A$ invariant subspaces and the system group orbits of $\Sigma^{II}(A)$. Since $A$ is cyclic there exists a finite number of $A$-invariant subspaces (see D.3). Assuming *(vii)*, there exist finitely many reachable sets.
*(viii)* $\Rightarrow$ *(iii)*: Recall that $\mathcal{N}_A$ is a system group orbit of $\Sigma^{II}(A)$ (see Proposition 6.11). If $\Sigma^{II}(A)$ has only a finite number of reachable sets, then $\Sigma^{II}(A)_{|\mathcal{N}_A}$ has only a finite number of reachable sets. Thus, from Corollary 4.9 it follows, that $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is reachable from one $x \in \mathcal{N}_A$. Since $\mathcal{N}_A$ is dense in $\mathbb{RP}^{n-1}$, system $\Sigma_{II}(A)$ is approximatively reachable from $x$.
*(ii)* $\Leftrightarrow$ *(ix)* Recall that that $\mathbb{R}^*I \subseteq P(A)$ (see Theorem 6.3). Moreover, $\Sigma^{II}(A)$ is an induced system of $\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)$ with respect to some $\pi : \mathrm{GL}_n(\mathbb{R}) \to \mathbb{RP}^{n-1}$. Here, $C_\pi = P(A) \cap \mathbb{R}^*I = \mathbb{R}^*I$ (see Theorem 5.9). Now the claim follows by Theorem 3.6, i.e., $S_{\Sigma^{II}}(A)$ is a group if and only if $S(A)C_\pi = P(A)$. □

Note that $S(A) = P(A)$ implies $S(A)\mathbb{R}^* = P(A)$. The following example shows, that the converse is wrong in general.

**Example 6.19** Consider $\Sigma^{II}(A)$ for

$$A := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

We show that $S(A)\mathbb{R}^*$ is a group but $S(A)$ is not. Obviously,

$$B := \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^{-1} \in \left\{ \prod_{t=1}^{N} \begin{pmatrix} -u_t & -1 \\ 1 & -u_t \end{pmatrix}^{-1} \middle| N < \infty, u_t \in \mathbb{R} \right\} = S(A).$$

We assume that $B^{-1} \in S(A)$, i.e., there exist shift parameters $u_1, \dots, u_N \in \mathbb{R}$ such that $B^{-1} = \prod_{t=1}^{N} (A - u_t I)^{-1}$. Then

$$\det(B^{-1}) = \det\left( \prod_{t=1}^{N} \begin{pmatrix} -u_t & -1 \\ 1 & -u_t \end{pmatrix}^{-1} \right) = \prod_{t=1}^{N} \frac{1}{u_t^2+1} \leq 1,$$

which is a contradiction to $\det(B) = \frac{1}{2}$. We conclude, $B^{-1} \notin S(A)$. Hence, $S(A)$ is not a group. On the other hand, the inverse of $(A - uI)^{-1} \in S(A)$ is given by

$$A - uI = (u^2 + 1)A^{-1}A^{-1}(A + uI)^{-1} \in S(A)\mathbb{R}^*.$$

Hence, $S(A)\mathbb{R}^*$ is a group.

## 6.5 Conditions for $S(A)\mathbb{R}^* \neq P(A)$

Theorem 6.18 shows, that in order to find out if $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable or not, it is enough to check $S(A)\mathbb{R}^* = P(A)$, which is a property of matrix semigroups[21]. The question if $S(A)\mathbb{R}^* = P(A)$ is given or not only depends on the canonical form of $A$. More precisely we have:

**Lemma 6.20** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic, $T \in \mathrm{GL}_n(\mathbb{R})$, $\mu \in \mathbb{R}$ and $\gamma \in \mathbb{R}^*$. Then $S(A)\mathbb{R}^* = P(A)$ if and only if $S(\gamma TAT^{-1} - \mu I)\mathbb{R}^* = G(\gamma TAT^{-1} - \mu I)$*

**Proof.** Obviously, $S(\gamma TAT^{-1} - \mu I)\mathbb{R}^* = T(S(A - \frac{\mu}{\gamma} I))T^{-1}\mathbb{R}^*$. Moreover,

$$\begin{aligned} S\left(A - \frac{\mu}{\gamma} I\right) &= \left\{ \prod_{t=0}^{N\in\mathbb{N}} ((A - \frac{\mu}{\gamma} I) - v_t I)^{-1} \,\middle|\, N \in \mathbb{N}, v_t \in U_{A - \frac{\mu}{\gamma} I} \right\} \\ &= \left\{ \prod_{t=0}^{N\in\mathbb{N}} (A - u_t I)^{-1} \,\middle|\, N \in \mathbb{N}, u_t \in U_A \right\} \\ &= S(A). \end{aligned}$$

Now the claim follows, since $\mathbb{R}^* I \subseteq G(B)$ for all $B \in \mathrm{GL}_n(\mathbb{R})$ and therefore

$$\begin{aligned} G(\gamma TAT^{-1} - \mu I) &= \left\langle S(\gamma TAT^{-1} - \mu I) \right\rangle \\ &= \left\langle S(\gamma TAT^{-1} - \mu I)\mathbb{R}^* \right\rangle \\ &= \left\langle T(S(A)\mathbb{R}^*)T^{-1} \right\rangle \\ &= T \left\langle S(A) \right\rangle T^{-1}\mathbb{R}^* \\ &= TP(A)T^{-1}. \end{aligned}$$

□

Recall that every matrix $A \in \mathbb{R}^{n\times n}$ is similar to its *real Jordan canonical form*

$$J_A = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{pmatrix}$$

[21] instead of $S_{\Sigma^{II}}(A) = G_{\Sigma^{II}}(A)$ which is a property of a semigroup generated by maps $f : \mathbb{RP}^{n-1} \to \mathbb{RP}^{n-1}$.

such that every block $J_j$ corresponds to the eigenvalue $\lambda_j$ respectively to the pair $(\lambda_j, \overline{\lambda_j})$ and has one of the following types:

$$\begin{array}{lcl}
\text{Type 1:} \quad J_j & = & (\lambda_j) \in \mathbb{R}^{1\times 1} \\
\text{Type 2:} \quad J_j & = & \begin{pmatrix} \lambda_j & 1 & & & \\ 0 & \lambda_j & 1 & & \\ & & \ddots & & \\ & & & \lambda_j & 1 \\ & \dots & & 0 & \lambda_j \end{pmatrix} \in \mathbb{R}^{k_j\times k_j}, k_j \geq 2 \\
\text{Type 3:} \quad J_j & = & \begin{pmatrix} \mathrm{Re}(\lambda_j) & \mathrm{Im}(\lambda_j) \\ -\mathrm{Im}(\lambda_j) & \mathrm{Re}(\lambda_j) \end{pmatrix} \in \mathbb{R}^{2\times 2},\ \mathrm{Im}\,\lambda_j \neq 0 \\
\text{Type 4:} \quad J_j & = & \begin{pmatrix} J & I & & \\ & J & I & \\ & & \ddots & \\ & & & J \end{pmatrix} \in \mathbb{R}^{k_j\times k_j} \text{with } J \text{ of Type 3}
\end{array}$$

**Proposition 6.21** *Let $J \in \mathbb{R}^{n\times n}$ be a matrix of Type $k \in \{1, 2, 3, 4\}$.*

*a) If $J$ is of Type 1 or 3 then $S(J)\mathbb{R}^* = G(J)$.*

*b) If $J$ is of Type 4 then $S(J)\mathbb{R}^* \neq G(J)$.*

*c) If $J$ is of Type 2 then $S(J)\mathbb{R}^* = G(J)$ if and only if $n = 2$.*

**Proof.** (i) assume that $J$ is of Type 1. Since $S(J) = S(J - \lambda I)$ we obtain $S(J) = S(J)\mathbb{R}^* = \mathbb{R}^* = G(J)$.
(ii) Now let $n = 2$ and $J$ of Type 2. Without loss of generality we assume $\lambda = 0$, since $S(J) = S(J - \lambda I)$. Recall that

$$G(J) = P(J) = \mathrm{GL}_2(\mathbb{R}) \cap \mathrm{span}(I, A)$$

(see (43)). Thus

$$\begin{aligned} G(J) &= \{aI + bJ \,|\, a, b \in \mathbb{R}\} \cap \mathrm{GL}_2(\mathbb{R}) \\ &= \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \,\middle|\, a \in \mathbb{R}^*, b \in \mathbb{R} \right\}. \end{aligned}$$

For $b \neq 0$ we obtain

$$\begin{pmatrix} a & b \\ 0 & a \end{pmatrix} = b \begin{pmatrix} 0 - u & 1 \\ 0 & 0 - u \end{pmatrix}$$

with $u := -\frac{a}{b}$. This shows $\overline{S(J)\mathbb{R}^*} = G(J)$. By Corollary 6.6 have $\mathrm{int}_{P(A)}\, S(A) \neq \emptyset$ and therefore $\mathrm{int}_{P(A)}\, S(A)\mathbb{R}^* \neq \emptyset$. Thus we conclude

$S(J)\mathbb{R}^* = G(J)$ by Lemma B.6.

(iii) Now we assume, that $J$ is of Type 3. The characteristic polynomial of $J$ is $\chi_J(t) = t^2 - 2\operatorname{Re}(\lambda)t + |\lambda|^2$. We obtain

$$\begin{aligned} G(J) &= \{aI + bJ \mid a, b \in \mathbb{R}\} \cap \mathrm{GL}_2(\mathbb{R}) \\ &= \left\{ \left( \begin{array}{cc} a & b \\ -b & a \end{array} \right) \middle| \, a^2 + b^2 \neq 0 \right\}. \end{aligned}$$

For $b \neq 0$ we get

$$\left( \begin{array}{cc} a & b \\ -b & a \end{array} \right) = r \left( \begin{array}{cc} \operatorname{Re}(\lambda) - u & \operatorname{Im}(\lambda) \\ -\operatorname{Im}(\lambda) & \operatorname{Re}(\lambda) - u \end{array} \right)$$

with $r = \frac{b}{\operatorname{Im}(\lambda)}$ and $u = \operatorname{Re}(\lambda) - \frac{a}{b}\operatorname{Im}(\lambda)$. Again we conclude $S(J)\mathbb{R}^* = G(J)$ by Corollary 6.6 and Lemma B.6.
(iv) Let $J$ be a matrix of Type 2 with minimal polynomial $(x-\lambda)^n$, $n \geq 3$. Again we may assume that $\lambda = 0$. Assume that $I \in (S(J)\mathbb{R}^*)^{-1}$, i.e., $I = q(J)$ for a linear decomposable polynomial $q$. Then

$$1 = q(x) + k(x)x^n \tag{50}$$

with $k \in \mathbb{R}[x]$. Derivation gives us $q'(x) = -x^{n-1}(nk(x) + xk'(x))$. Since zero is a root of $q'$ with degree at least 2, zero is also a root of $q$. This contradicts (50). Hence, $I \notin (S(J)\mathbb{R}^*)^{-1}$ and therefore $S(J)\mathbb{R}^* \neq P(A)$.
(v) Let $J$ be of Type 4 with characteristic polynomial $p(x)^n$ such that $p$ is quadratic and irreducible. Assume $I \in (S(J)\mathbb{R}^*)^{-1}$,i.e., $I = q(J)$ for a linear decomposable polynomial $q$. Then $1 = q + kp^n$ with $k \in \mathbb{R}[x]$ and therefore $q' = p(k'p^{n-1} + nkp^{n-1}p')$. It follows $q' \notin \mathcal{L}$. This is a contradiction to $q \in \mathcal{L}$ (see Theorem E.4). Hence, $I \notin (S(J)\mathbb{R}^*)^{-1}$ and therefore $S(J)\mathbb{R}^* \neq G(J)$. □

**Lemma 6.22** *Let $A \in \mathbb{R}^{n \times n}$ be a block-diagonal cyclic matrix*

$$A = \left( \begin{array}{cc} A_1 & 0 \\ 0 & A_2 \end{array} \right) \quad \textit{with} \quad A_1 \in \mathbb{R}^{n_1 \times n_1},\ A_2 \in \mathbb{R}^{(n-n_1)\times(n-n_1)}.$$

*If $S(A_1)\mathbb{R}^* \neq P(A_1)$ then $S(A)\mathbb{R}^* \neq P(A)$.*

**Proof.** By Lemma B.5 we have $P(A_1) \setminus \overline{S(A_1)\mathbb{R}^*} \neq \emptyset$. Choose $p \in \mathbb{R}[x]$ such that $p(A_1) \in P(A_1) \setminus \overline{S(A_1)\mathbb{R}^*} \neq \emptyset$. Without loss of generality[22] $p(A_2)$ is invertible. Then $p(A) \in P(A)$. On the other hand, $P(A) \neq S(A)\mathbb{R}^*$, since $p(A) = q(A)$ with $q \in \mathcal{L}$ implies $p(A_1) = q(A_1)$. □

[22] Recall that $P(A_1) = \operatorname{span}(I, A_1, \ldots, A_1^{n_1-1}) \cap \mathrm{GL}_{n_1}(\mathbb{R})$ (see (43)). Therefore, $p(A_1) + \epsilon I \in \mathrm{GL}_{n_1}(\mathbb{R})$ for all except finitely many $\epsilon \in \mathbb{R}$.

In general, the assumptions $S(A_1)\mathbb{R}^* = P(A_1)$ and $S(A_2)\mathbb{R}^* = P(A_2)$ for a block matrix

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

do not imply $S(A)\mathbb{R}^* = P(A)$. An example for this is given by

$$A_1 = (0) \quad \text{and} \quad A_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

This is one of the consequences of the following theorem.

**Theorem 6.23** *Let $A$ be cyclic and spec$(A) \subseteq \mathbb{C}$ the set of eigenvalues of $A \in \mathbb{R}^{n\times n}$.*

*a) If there exists a real eigenvalue $\lambda$ of multiplicity at least three, then $S(A)\mathbb{R}^* \neq P(A)$.*

*b) If there exists a pair of eigenvalues $\lambda, \overline{\lambda} \in \mathbb{C}\setminus\mathbb{R}$ of multiplicity at least two, then $S(A)\mathbb{R}^* \neq P(A)$.*

*c) If there exist eigenvalues $\lambda_1, \lambda_2 \in \operatorname{Spec}(A)$ of multiplicity one with $\operatorname{Re}(\lambda_1) = \operatorname{Re}(\lambda_2)$ but $\operatorname{Im}(\lambda_1) \neq \operatorname{Im}(\lambda_2)$, then $S(A)\mathbb{R}^* \neq P(A)$.*

*d) If there exists eigenvalues $\lambda_1, \lambda_2, \lambda_3 \in \operatorname{Spec}(A)$ of multiplicity one, with $\lambda_1 \in \mathbb{C}\setminus\mathbb{R}$ and $\lambda_2, \lambda_3 \in \mathbb{R}$ such that $\operatorname{Re}\lambda_1 = \frac{\lambda_2+\lambda_3}{2}$ and $\lambda_2 < \operatorname{Re}\lambda_1 + \operatorname{Im}\lambda_1$, then $S(A)\mathbb{R}^* \neq P(A)$.*

*e) If there exists eigenvalues $\lambda_1, \lambda_2, \lambda_3 \in \operatorname{Spec}(A)$ of multiplicity one, with $\operatorname{Re}(\lambda_3) < \operatorname{Re}(\lambda_1) < \operatorname{Re}(\lambda_2)$, $\operatorname{Re}(\lambda_2)+\operatorname{Re}(\lambda_3) = 2\operatorname{Re}\lambda_1$, $\operatorname{Im}(\lambda_2) = \operatorname{Im}(\lambda_3)$ and $\operatorname{Im}(\lambda_2)^2 > (\operatorname{Re}(\lambda_2) - \operatorname{Re}(\lambda_1))^2 + (\operatorname{Im}\lambda_1)^2$, then $S(A)\mathbb{R}^* \neq P(A)$.*

**Proof.** a) and b) is an immediate consequence of Proposition 6.21 and Lemma 6.22. For the proofs of c),d) and e) we show that the assumption $I \in (S(A)\mathbb{R}^*)^{-1}$ implies, that the eigenvalues of $A$ do not form a constellation as assumed. It follows, that $S(A)\mathbb{R}^* \neq P(A)$.
c) We distinguish between the case where $\lambda_1 \in \mathbb{R}$ and where $\lambda_1 \notin \mathbb{R}$.
(i) Let $\lambda_1 \in \mathbb{R}$. Then in the canonical form of $A$ is a block of the type

$$J = \begin{pmatrix} \operatorname{Re}(\lambda_2) & \operatorname{Im}(\lambda_2) & 0 \\ -\operatorname{Im}(\lambda_2) & \operatorname{Re}(\lambda_2) & 0 \\ 0 & 0 & \lambda_1 \end{pmatrix}.$$

By Lemma 6.20 we can assume

$$J = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

since $\lambda_1 = \mathrm{Re}(\lambda_2)$. Assuming $I \in (S(A)\mathbb{R}^*)^{-1}$, there exist controls $u_1, \ldots, u_T \in \mathbb{R} \setminus \{0\}$ such that both of the following equations are fulfilled.

$$\begin{aligned} (I) \quad I_2 &= r\prod_{t=1}^{T}\left(\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - u_t I\right) \\ (II) \quad 1 &= r\prod_{t=1}^{T}(-u_t). \end{aligned}$$

Applying the determinant function $B \mapsto \det(B)$ on Equation $(I)$ we obtain

$$1 = r^2\prod_{t=1}^{T}(1+u_t^2),$$

which is a contradiction to Equation $(II)$, since $r \neq 0$. Hence, $I \notin (S(J)\mathbb{R}^*)^{-1}$ and by Lemma 6.22 we obtain $S(A)\mathbb{R}^* \neq G_A$.
(ii) Now let $\lambda_1 \in \mathbb{C} \setminus \mathbb{R}$. Then in the canonical form of $A$ is a block of the type

$$J = \begin{pmatrix} \mathrm{Re}(\lambda_1) & \mathrm{Im}(\lambda_1) & 0 & 0 \\ -\mathrm{Im}(\lambda_1) & \mathrm{Re}(\lambda_1) & 0 & 0 \\ 0 & 0 & \mathrm{Re}(\lambda_2) & \mathrm{Im}(\lambda_2) \\ 0 & 0 & -\mathrm{Im}(\lambda_2) & \mathrm{Re}(\lambda_2) \end{pmatrix}.$$

By Lemma 6.20 we can assume

$$J = \begin{pmatrix} 0 & 1 & 0 & \\ -1 & 0 & 0 & \\ 0 & 0 & 0 & \beta \\ 0 & 0 & -\beta & 0 \end{pmatrix}$$

with $\beta > 0$ since $\mathrm{Re}(\lambda_1) = \mathrm{Re}(\lambda_2)$. Suppose there exist controls $u_1, \ldots, u_T \in \mathbb{R} \setminus \{0\}$ such that $I = r\prod_{t=1}^{T}(A - u_t I)$ for any $r \in \mathbb{R}^*$, then both of the following equations are fulfilled

$$\begin{aligned} (I) \quad I_2 &= r\prod_{t=1}^{T}\left(\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - u_t I\right) \\ (II) \quad I_2 &= r\prod_{t=1}^{T}\left(\begin{pmatrix} 0 & \beta \\ -\beta & 0 \end{pmatrix} - u_t I\right). \end{aligned}$$

As in (i) we apply the determinant function $B \mapsto \det(B)$ on $(I)$ and $(II)$ and we obtain

$$r^2\prod_{t=1}^{T}(1+u_t^2) = r^2\prod_{t=1}^{T}(\beta^2+u_t^2).$$

Since $r \neq 0$, we obtain $\beta = 1$. But then $J$ is a matrix of Type 4 and Theorem 6.21 implies $S(A)\mathbb{R}^* \neq P(A)$.
d) If $\lambda_1 \in \mathbb{C} \setminus \mathbb{R}$ and $\lambda_2, \lambda_3 \in \mathbb{R}$ with $\operatorname{Re} \lambda_1 = \frac{\lambda_1 + \lambda_2}{2}$ then the canonical form of $A$ has a block $J$ of the form

$$J = \begin{pmatrix} \operatorname{Re}(\lambda_1) & \operatorname{Im}(\lambda_1) & 0 & 0 \\ -\operatorname{Im}(\lambda_1) & \operatorname{Re}(\lambda_1) & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & \lambda_3 \end{pmatrix}.$$

By Lemma 6.20 we can assume

$$J = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & -\alpha \end{pmatrix}$$

with $\alpha = \frac{\lambda_2 - \operatorname{Re}(\lambda_1)}{\operatorname{Im}(\lambda_1)} > 0$. Note that $\alpha < 1$ by assumption. Suppose there exist controls $u_1, \ldots, u_T \in \mathbb{R} \setminus \{0\}$ such that $I = r \prod_{t=1}^{T} (A - u_t I)$ for any $r \in \mathbb{R}^*$, then both of the following equations are fulfilled

$$\begin{aligned} (I) \quad I_2 &= r \prod_{t=1}^{T} \left( \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - u_t I \right) \\ (II) \quad I_2 &= r \prod_{t=1}^{T} \left( \begin{pmatrix} \alpha & 0 \\ 0 & -\alpha \end{pmatrix} - u_t I \right) \end{aligned}$$

The determinant function applied on Equation (I) and Equation (II) yields

$$r^2 \prod_{t=1}^{T} (\alpha - u_t)(-\alpha - u_t) = r^2 \prod_{t=1}^{T} (1 + u_t^2).$$

But this is a contradiction, since $r \neq 0$ and $1 + u_t^2 > |u_t^2 - \alpha^2|$. We conclude $S(A)\mathbb{R}^* \neq P(A)$.
e) (i) First we assume, that $\operatorname{Im} \lambda_1 \neq 0$. Without loss of generality, $A$ has a block of the type

$$J = \begin{pmatrix} \operatorname{Re}(\lambda_1) & \operatorname{Im}(\lambda_1) & 0 & 0 & 0 & 0 \\ -\operatorname{Im}(\lambda_1) & \operatorname{Re}(\lambda_1) & 0 & 0 & 0 & 0 \\ 0 & 0 & \operatorname{Re}(\lambda_2) & \operatorname{Im}(\lambda_2) & 0 & 0 \\ 0 & 0 & -\operatorname{Im}(\lambda_2) & \operatorname{Re}(\lambda_2) & 0 & 0 \\ 0 & 0 & 0 & 0 & \operatorname{Re}(\lambda_3) & \operatorname{Im}(\lambda_3) \\ 0 & 0 & 0 & 0 & -\operatorname{Im}(\lambda_3) & \operatorname{Re}(\lambda_3) \end{pmatrix}$$

such that $\mathrm{Im}(\lambda_1), \mathrm{Im}(\lambda_2), \mathrm{Im}(\lambda_3) > 0$ and $\mathrm{Re}(\lambda_3) < \mathrm{Re}(\lambda_2) < \mathrm{Re}\,\lambda_1$. Using Lemma 6.20 we transform the problem on the matrix

$$J = \begin{pmatrix} 0 & \alpha & 0 & 0 & 0 & 0 \\ -\alpha & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \gamma & 0 & 0 \\ 0 & 0 & -\gamma & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & \gamma \\ 0 & 0 & 0 & 0 & -\gamma & -1 \end{pmatrix}.$$

Here, $\alpha = \frac{2\,\mathrm{Im}(\lambda_1)}{\mathrm{Re}(\lambda_2)-\mathrm{Re}(\lambda_3)}$ and $\gamma = \frac{2\,\mathrm{Im}(\lambda_2)}{\mathrm{Re}(\lambda_2)-\mathrm{Re}(\lambda_3)}$. Suppose there exist controls $u_1, \ldots, u_T \in \mathbb{R} \setminus \{0\}$ such that $I = r\prod_{t=1}^{T}(A - u_t I)$ for any $r \in \mathbb{R}^*$, then the following three equations are fulfilled

$$\begin{aligned} (I) \quad 1 &= r\prod_{t=1}^{T}\left(\begin{pmatrix} 0 & \alpha \\ -\alpha & 0 \end{pmatrix} - u_t I\right), \\ (II) \quad I_2 &= r\prod_{t=1}^{T}\left(\begin{pmatrix} 1 & \gamma \\ -\gamma & 1 \end{pmatrix} - u_t I\right), \\ (III) \quad I_2 &= r\prod_{t=1}^{T}\left(\begin{pmatrix} -1 & \gamma \\ -\gamma & -1 \end{pmatrix} - u_t I\right). \end{aligned}$$

Applying the determinant function on $(I)$, $(II)$ and $(III)$ we obtain

$$r^2\prod_{t=1}^{T}((1-u_t)^2+\gamma^2) = r^2\prod_{t=1}^{T}((1+u_t)^2+\gamma^2) = r^2\prod_{t=1}^{T}(u_t^2+\alpha^2).$$

In particular it holds

$$\prod_{t=1}^{T}\underbrace{(((1-u_t)^2+\gamma^2)((1+u_t)^2+\gamma^2))}_{:=p_\gamma(u_t)} = \prod_{t=1}^{T}\underbrace{(u_t^2+\alpha^2)^2}_{:=p_\alpha(u_t)}. \tag{51}$$

Note that $p_\gamma(u_t) > 0$ and $p_\alpha(u_t) > 0$. Moreover,

$$\begin{aligned} p_\gamma(u_t) - p_\alpha(u_t) &= u_t^4 - 2u_t^2 + 1 + \gamma^2 + \gamma^2 u_t^2 + \gamma^4 - u_t^4 + 2u_t^2 + \alpha^2 + \alpha^4 \\ &= \underbrace{(2\gamma^2 - 2 - 2\alpha^2)}_{C_1}u_t^2 + \underbrace{(1 + 2\gamma^2 + \gamma^4 - \alpha^4)}_{C_2}. \end{aligned}$$

By assumption we have

$$\begin{aligned} & \mathrm{Im}(\lambda_2)^2 > (\mathrm{Re}(\lambda_2) - \mathrm{Re}(\lambda_1))^2 + (\mathrm{Im}\,\lambda_1)^2 \\ \Leftrightarrow\ & (2\,\mathrm{Im}(\lambda_2))^2 - (\mathrm{Re}(\lambda_2) - \mathrm{Re}(\lambda_1))^2 > (2\,\mathrm{Im}(\lambda_1))^2 \\ \Leftrightarrow\ & \gamma^2 - 1 > \alpha^2. \end{aligned}$$

Therefore, $C_1 > 0$ and $C_2 > 0$. It follows $p_\gamma(u_t) > p_\alpha(u_t)$ which contradicts (51). We conclude $I \notin (S(A)\mathbb{R}^*)^{-1}$ and therefore $S(A)\mathbb{R}^* \neq P(A)$.

(ii) Now we assume, $\mathrm{Im}(\lambda_1) = 0$. Without loss of generality, $A$ has a block of the type

$$J = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & \gamma & 0 & 0 \\ 0 & -\gamma & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & \gamma \\ 0 & 0 & 0 & -\gamma & -1 \end{pmatrix}.$$

Now $\gamma = \frac{\mathrm{Im}(\lambda_2)}{\mathrm{Re}(\lambda_2)-\lambda_1}$. By assumption we obtain $\gamma \geq 1$. Suppose there exist controls $u_1, \ldots, u_T \in \mathbb{R} \setminus \{0\}$ such that $I = r \prod_{t=1}^{T}(A - u_t I)$ for any $r \in \mathbb{R}^*$. Then (II), (III) of (ii) are fulfilled. Moreover it is

$$(I^*) \qquad 1 = r \prod_{t=1}^{T}(-u_t)$$

Applying the determinant function on $(I^*)$, $(II)$ and $(III)$ we obtain

$$r^2 \prod_{t=1}^{T}((1-u_t)^2 + \gamma^2) = r^2 \prod_{t=1}^{T}((1+u_t)^2 + \gamma^2) = r^2 \prod_{t=1}^{T} u_t^2.$$

In particular it holds

$$\prod_{t=1}^{T} \underbrace{(((1-u_t)^2 + \gamma^2)((1+u_t)^2 + \gamma^2))}_{:=p_\gamma(u_t)} = \prod_{t=1}^{T} u_t^4$$

which is a contradiction, since $\gamma \geq 1$ implies

$$p_\gamma(u_t) = u_t^4 + (1+\gamma^2)^2 + 2u_t^2(\gamma^2 - 1) > u_t^4$$

We conclude $S(A)\mathbb{R}^* \neq P(A)$. □

Recall that $S(A)\mathbb{R}^* \neq P(A)$ implies, that $\Sigma^{II}(A)$ restricted on $\mathcal{N}_A$ is not controllable. Therefore, Theorem 6.23 verifies the following controllability results of Helmke and Wirth (see [HW01], Proposition 8,i, Corollary 10,i and Corollary 10,i-ii). .

**Theorem 6.24 (Helmke and Wirth [HW01])** *Let $A \in \mathbb{R}^{n\times n}$ be a cyclic matrix. If the eigenvalue constellation coincides with one of the eigenvalue constellations in Theorem 6.23, then $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is not controllable.*

Recall that a matrix $A \in \mathbb{R}^{n\times n}$ is called *skew-symmetric* if $A^\top = -A^\top$, and respectively *Hamiltonian*, if $n$ is even and

$$A^\top J + JA = 0 \ \text{ for } \ J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Now an immediate consequence of Theorem 6.23 is the following.

**Corollary 6.25** *Let $A \in \mathbb{R}^{n\times n}$.*

*a) If $n \geq 3$ and $A$ is skew-symmetric, then $S(A)\mathbb{R}^* \neq P(A)$ and $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is not controllable .*

*b) If $A$ is a cyclic Hamiltonian matrix with eigenvalue $\lambda_1 \in i\mathbb{R}$. If there exists $\lambda_2 \in \mathrm{Spec}(A)$ such that $\mathrm{Im}(\lambda_2)^2 - \mathrm{Re}(\lambda_2)^2 > \mathrm{Im}(\lambda_1)^2$ then $S(A)\mathbb{R}^* \neq P(A)$ and $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is not controllable.*

**Proof.** Eigenvalues of skew-symmetric matrices are of the form $ir$ with $r \in \mathbb{R}$. Therefore, claim a) is a consequence of statement c) in Theorem 6.23. Hamiltonian matrices have the property, that for any eigenvalue $\lambda$ also $-\lambda$ is an eigenvalue. The conditions in claim b) imply, that there exist $\lambda_2, \lambda_3 \in \mathrm{Spec}(A)$ such that $\mathrm{Re}(\lambda_2) + \mathrm{Re}(\lambda_2) = 2\,\mathrm{Re}(\lambda_1) = 0$. and $\mathrm{Im}(\lambda_2)^2 - \mathrm{Re}(\lambda_2)^2 > \mathrm{Im}(\lambda_1)^2$. Hence, $S(A)\mathbb{R}^* \neq P(A)$ by Theorem 6.23,e). $\square$

## 6.6 Conditions for $S(A)\mathbb{R}^* = P(A)$

Using the results of the previous section one easily construct inverse iteration systems $\Sigma^{II}(A)_{|\mathcal{N}_A}$ which are not controllable. On the other hand, there exists a large set of matrices, such that $S(A)\mathbb{R}^* = P(A)$, which implies controllability of $\Sigma^{II}(A)_{|\mathcal{N}_A}$. In the following we present three sufficient conditions for $S(A)\mathbb{R}^* = P(A)$.

If $A$ is in block diagonal form $A = \operatorname{diag}(A_1, \ldots, A_k)$, then $S(A)\mathbb{R}^* = P(A)$ implies $S(A_i)\mathbb{R}^* = P(A_i)$, $i = 1, \ldots, k$ (Lemma 6.22). Note that the converse is wrong in general. An example for that will be given in Section 6.8.2. The following result, provides a strategy, for checking if a block matrix with $S(A_i)\mathbb{R}^* = P(A_i)$, $i = 1, \ldots, k$ fulfills $S(A)\mathbb{R}^* = P(A)$.

**Theorem 6.26** *Let $A$ be a cyclic block-diagonal matrix $A = \operatorname{diag}(A_1, \ldots, A_k)$ with $A_i \in \mathbb{R}^{n_i \times n_i}$ and $n_1 + \cdots + n_k = n$. Assume that for any $i = 1, \ldots, k$ there exists a dense subset $M_i$ of $P(A_i)$ such that for any $p(A_i) \in M_i$ there exists $q \in \mathcal{L}$ such that*

*(i)* $q(A_i) = p(A_i)$,

*(ii)* $q(A_j) = I_{n_j}$ *for* $j \neq i$.

*Then* $S(A)\mathbb{R}^* = P(A)$.

**Proof.** Since $A$ is cyclic, the minimal poynomial of $A$ is the product of the minimal polynomials of $A_1, \ldots, A_n$. Thus, $P(A) \cong P(A_1) \times \cdots \times P(A_k)$ (see Lemma D.5). For any $p(A) \in M_1 \times \cdots \times M_k$ we choose $q_1, \ldots, q_k$ such that $(i)$ and $(ii)$ are fulfilled. Then

$$q_1(A) \ldots q_k(A) = \operatorname{diag}(q_1(A_1), I, \ldots, I) \ldots \operatorname{diag}(I, \ldots, I, q_k(A_k)) = p(A).$$

Recall that $\operatorname{int}_{P(A)} S(A)\mathbb{R}^* \neq \emptyset$ (see Corollary 6.6). Moreover, $M_1 \times \cdots \times M_k$ is dense in $P(A)$. Thus $S(A)\mathbb{R}^* = P(A)$, by Lemma B.6. □

For the next sufficient condition for $S(A)\mathbb{R}^* = P(A)$, we use the fact, that $P(A)$ is a topological group.

**Theorem 6.27** *Let $A \in \mathbb{R}^{n \times n}$ be cyclic. Then $S(A)\mathbb{R}^* = P(A)$ if and only if $I \in \operatorname{int}_{P(A)} S(A)\mathbb{R}^*$.*

**Proof.** We show that $S(A)\mathbb{R}^*$ intersects every connected component $P(A)^i$ of $P(A) = P(A)$. Then, the equivalence follows from Lemma B.4. For any $B \in P(A)^i$ we choose $p \in \mathbb{R}[x]$ such that $B^{-1} = p(A)$. $p$ can be decomposed as $p(x) = q(x)p_1(x) \ldots p_m(x)$ with $q \in \mathcal{L}$ and $p_j$, $j = 1, \ldots, m$ normed quadratic non-irreducible polynomials. By Lemma 6.4 there exists $u_1, \ldots, u_m \in \mathbb{R} \setminus \operatorname{Spec}(A)$ and continuous paths $\beta_j : [0, 1] \to P(A)$, $j =$

$1, \dots, m$ such that $\beta_j(0) = p_j$ and $\beta_j(1) = (A - u_j I)^2$. Therefore, the path $\tilde{\omega} : [0,1] \to P(A)$ defined by $\tilde{\omega} : t \mapsto q(A)\beta_1(t)\dots\beta_m(t)$ fulfills $\tilde{\omega}(0) = (P(A))^{-1} = B$ and

$$\tilde{\omega}(1) = q(A)^{-1}\prod_{t=1}^{m}(A - u_t I)^{-2} \in S(A)\mathbb{R}^*.$$

□

For the remaining part of Section 6.6 we deal with matrices, where every eigenvalue is real. We will use the following technical result.

**Lemma 6.28** *Let $p \in \mathbb{R}[x]$ be a polynomial of degree $k-1$. For every sequence $\lambda_1 \le \dots \le \lambda_k \in \mathbb{R}$ there exists $M \in \mathbb{R}$ such that $f(x) := p(x) - M\prod_{i=1}^{k}(x - \lambda_i)$ is linear decomposable.*

**Proof.** Let $q(x) := \prod_{i=1}^{k}(x - \lambda_i)$. We define $x_0 < x_1 < \dots < x_k$ such that $x_i \notin \{\lambda_1, \dots, \lambda_k\}$. Moreover, we define $y_0 = q(x_0)$ and $y_{i+1} := -\operatorname{sgn} y_i |q(x_{i+1})|$, $i = 0, \dots, k-1$. By construction $y_i \ne 0$ and

$$\operatorname{sgn} y_i = -\operatorname{sgn} y_{i+1} \quad \text{for} \quad i = 0, \dots, k. \tag{52}$$

Now we define

$$C := 1 + \max_{x\in[x_0,x_k]}|p(x)|, \quad D := \min_{i=0,\dots,k}|y_i|, \quad \text{and} \quad M := -\frac{C}{D}.$$

Obviously, the polynomial

$$f := p - Mq \tag{53}$$

has degree $\deg f = \deg q = k$. In the following we show, that $f$ has $k$ different real roots and therefore $f \in \mathcal{L}$.

We have

$$Mq(x_i) = \left(1 + \max_{x\in[x_0,x_k]}|p(x)|\right)\cdot\left(\frac{y_i}{\min_{j=0,\dots,k}|y_j|}\right)$$

and therefore $Mq(x_i) > p(x_i)$ if $y_i > 0$, respectively, $Mq(x_i) < p(x_i)$ if $y_i < 0$. It follows $\operatorname{sgn} f(x_i) = \operatorname{sgn} y_i$ for all $i = 0, \dots, k$. By (52) we obtain

$$\operatorname{sgn}(f(x_i)) = -\operatorname{sgn}(f(x_{i+1}), \quad \text{for} \quad i = 0, \dots, k-1.$$

Now the mean value theorem yields that $f$ has $k$ different real roots. Hence, $f \in \mathcal{L}$. □

**Theorem 6.29** *If all eigenvalues of $A$ are real and have multiplicity at most two, then $S(A)\mathbb{R}^* = P(A)$.*

**Proof.** We show, that for any $p \in \mathbb{R}[x]$ there exists $q \in \mathcal{L}$ such that $p(A) = q(A)$. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of $A$. By Lemma 6.20 we can assume that $A$ is block diagonal

$$A = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{pmatrix}$$

with $A_i = (\lambda_i)$, and respectively $A_i = \begin{pmatrix} \lambda_i & 1 \\ 0 & \lambda_i \end{pmatrix}$. Note that

$$p(A) = \begin{pmatrix} p(A_1) & & \\ & \ddots & \\ & & p(A_k) \end{pmatrix}$$

with $p(A_i) = (p(\lambda_i))$ and respectively $p(A_i) = \begin{pmatrix} p(\lambda_i) & p'(\lambda_i) \\ 0 & p(\lambda_i) \end{pmatrix}$.
By Lemma 6.28 there exists $M \in \mathbb{R}$ such that $q(x) := p(x) - M\prod_{i=1}^{k}(x-\lambda_i)$ is linear decomposable. Note that $p(\lambda_i) = q(\lambda_i)$. Moreover, $p'(\lambda_i) = q'(\lambda_i)$ if $\lambda_i = \lambda_{i+1}$. Hence, $p(A) = q(A)$ and therefore $S(A)\mathbb{R}^* = P(A)$. □

Note that Theorem 6.29 verifies another result of Helmke and Wirth.

**Theorem 6.30 (Helmke and Wirth [HW01], Proposition 8,ii)** *If all eigenvalues of a cyclic matrix are real and have multiplicity at most two, then $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable.*

Note that Theorem 6.30 shows, that $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable for an open set of matrices $A \in \mathbb{R}^{n\times n}$. All conditions we have found implying $S(A)\mathbb{R}^* \neq P(A)$ assume certain symmetries in the constellation of eigenvalues of $A$ and are therefore nongeneric (see Section 6.5). It remains unclear, whether controllability of $\Sigma^{II}(A)_{|\mathcal{N}_A}$ holds for a generic subset of $\mathbb{R}^n$.

We finish this section with two interesting byproducts of Lemma 6.28, which give new insight on the theory of linear decomposable polynomials.

**Corollary 6.31** *Every real polynomial of degree $k-1$ can be written as the sum of two linear decomposable polynomials of degree $k$.*

**Proof.** Let $p \in \mathbb{R}[x]$ of degree $k-1$. Following 6.28 we write $p = f + q$ with $f \in \mathcal{L}$ and $q(x) = M\prod_{i=1}^{k}(x-\lambda_i) \in \mathcal{L}$. Moreover, $\deg f = \deg q = k$. □

For any $\lambda_1 < \cdots < \lambda_k$ and $b_1, \ldots, b_k \in \mathbb{R}$ there exists a unique polynomial $p \in \mathbb{R}[x]$ of degree $k-1$ such that $f(\lambda_i) = b_i$ for $i = 1, \ldots, k$. This fact is known as the *Lagrange interpolation theorem.* In the following we show a similar theorem for linear decomposable polynomials.

**Theorem 6.32 (Interpolation theorem)** *Let $\lambda_1 < \cdots < \lambda_k \in \mathbb{R}$ and $b_1, \ldots, b_k \in \mathbb{R}$. There exists a linear decomposable polynomial $f \in \mathcal{L}$ of degree $k$ such that $f(\lambda_i) = b_i$.*

**Proof.** Following the Lagrangian interpolation theorem, there exists a unique polynomial $p \in \mathbb{R}[x]$ of degree $k-1$ such that $p(\lambda_i) = b_i$, $i = 1, \ldots, k$. From Lemma 6.28 we deduce the existence of $M \in \mathbb{R}$ such that $f(x) := p(x) - M\prod_{i=1}^{k}(x - \lambda_i)$ is linear decomposable. Hence, $\deg(f) = k$ and

$$f(\lambda_i) = p(\lambda_i) - 0 = b_i.$$

$\square$

## 6.7 Structure of reachable sets

In the following we analyze the adherence structure of the reachable sets for classical inverse iteration systems. If $\mathbb{R}^*S(A) = P(A)$ then the adherence structure of the reachable sets is already given by the orbit graph (see Theorem 6.18). Therefore, we focus on the case $\mathbb{R}^*S(A) \neq P(A)$. Nevertheless, our first observation holds in both cases.

**Proposition 6.33** *Let $A \in \mathbb{R}^{n\times n}$ cyclic.*

*a) For any $x, y \in \mathcal{N}_A$ there exists $z, \tilde{z} \in \mathcal{N}_A$ such that $\mathcal{R}(z) \subseteq \mathcal{R}(x) \cap \mathcal{R}(y)$ and $\mathcal{R}(x) \cup \mathcal{R}(y) \subseteq \mathcal{R}(\tilde{z})$.*

*b) For any $x \in \mathcal{N}_A$ we have $x \in \overline{\mathcal{R}(x)}$.*

*c) If $v$ is an eigenvector with respect to a real eigenvalue $\lambda$ with multiplicity $k$, then $\pi(v) \in \overline{\mathcal{R}(x)}$ for any $x \in \mathcal{N}_A$.*

**Proof.** a) Recall that $\Sigma^{II}(A)$ is an abelian system and therefore right divisible as well as left divisible. Thus, claim a) follows from Theorem 4.8.
b) Recall that $\mathcal{N}_A$ is open and dense in $\mathbb{RP}^{n-1}$. Therefore, we obtain

$$C_{\mathcal{N}_A} = \{g \in G_{\Sigma^{II}(A)} \,|\, g_{|\mathcal{N}_A} = \mathrm{id}_{|\mathcal{N}_A}\} = \{e\}.$$

Moreover, we have $e \in \overline{S_{\Sigma^{II}(A)}}$ by Corollary 6.6. Thus, $x \in \overline{\mathcal{R}(x)}$ follows from Theorem 4.12.
c) We choose a basis, such that

$$A = \begin{pmatrix} A_\lambda & 0 \\ 0 & \tilde{A} \end{pmatrix} \quad \text{with} \quad A_\lambda = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} \in \mathbb{R}^{k\times k}$$

and $\tilde{A} \in \mathbb{R}^{(n-k)\times(n-k)}$. Then $v = (1, 0 \ldots, 0)^\top$ and $\lambda$ is not an eigenvalue of $\tilde{A}$. Without loss of generality we assume that $x_0 := \pi\left((1, 1 \ldots, 1)^\top\right) \in \mathcal{N}_A$. By choosing $u_t = \lambda - \frac{1}{t}$ we obtain

$$\lim_{t\to\infty} (A - u_t I)^{-1} \cdot x_0 = \pi \begin{pmatrix} \lim_{t\to\infty}(A_\lambda - u_t I_k)^{-1} e_k \\ (\tilde{A} - \lambda I_{n-k})^{-1} e_{n-k} \end{pmatrix}$$

with $e_k = (1, \ldots, 1)^\top \in \mathbb{R}^k$ and $e_{n-k} = (1, \ldots, 1)^\top \in \mathbb{R}^{n-k}$. Since

$$(A_\lambda - u_t I_k)^{-1} = t \begin{pmatrix} 1 & -t & t^2 & \ldots & (-t)^{k-1} \\ & 1 & -t & \ddots & (-t)^{k-2} \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & \ddots \\ & & & & 1 \end{pmatrix}$$

it follows $\lim_{t\to\infty} (A - u_t I)^{-1} \cdot x_0 = \pi(v)$ Thus, $\pi(v) \in \overline{\mathcal{R}(x)}$. By Theorem 4.18 it follows $\pi(v) \in \overline{\mathcal{R}(x)}$ for all $x \in \mathcal{N}_A$.

□

Now we focus on the case $S(A)\mathbb{R}^* \neq P(A)$. Here the complexity of the reachable graph is much higher then the complexity of the orbit graph. In particular, there exist infinitely many reachable sets (see Theorem 6.18). More precisely, the reachable sets within $\mathcal{N}_A$ have the following structure.

**Theorem 6.34** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic such that $S(A)\mathbb{R}^* \neq P(A)$.*

*a) For any $y \in \mathcal{N}_A$, there exists a sequence $(y_t)_{t\in\mathbb{N}}$ in $\mathcal{N}_A$ such that*

- *(i) $y_1 = y$,*
- *(ii) $\mathcal{R}(y_{t+1}) \supseteq \operatorname{int}_{\mathcal{N}_A} \mathcal{R}(y_t)$; $\forall t \in \mathbb{N}_0$,*
- *(iii) $\bigcup_{t=0}^{\infty} \operatorname{int}_{\mathcal{N}_A} \mathcal{R}(y_t)$ is dense in $\mathcal{N}_A$,*
- *(iv) $\operatorname{int}_{\mathcal{N}_A}(\mathcal{N}_A \setminus \mathcal{R}(y_t)) \neq \emptyset$*
- *(v) $(y_t)_{t\in\mathbb{N}}$ converges to some $z_y \in \partial\mathcal{N}_A$.*

*b) If there exists $z \in \mathbb{RP}^{n-1} \setminus \mathcal{N}_A$ and $x \in \mathcal{N}_A$ such that*

$$G_{\Sigma^{II}(A)} \cdot z \cap \overline{\mathcal{R}(x)} = \emptyset,$$

*then $G_{\Sigma^{II}(A)} \cdot z$ is repelling to $\mathcal{N}_A$.*

**Proof.** a) Recall that $\mathcal{N}_A$ is open and therefore locally compact. Since $G_{\Sigma^{II}(A)}$ acts continuously on $\mathcal{N}_A$ and $\operatorname{int}_{G_{\Sigma^{II}(A)}} S_{\Sigma^{II}(A)} \neq \emptyset$ we can apply Theorem 4.10, a). Thus, for any $y \in \mathcal{N}_A$ there exists a sequence $(y_t)_{t\in\mathbb{N}}$ fulfilling $(i)$, $(ii)$ and $(iii)$. Assuming that $\operatorname{int}_{\mathcal{N}_A}(\mathcal{N}_A \setminus \mathcal{R}(y_t)) = \emptyset$ for one $t \in \mathbb{N}$, then $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is approximatively reachable from $y_t \in \mathcal{N}_A$. But this implies $S(A)\mathbb{R}^* = P(A)$ by Theorem 6.18. Thus $(iv)$ is fulfilled for all $t \in \mathbb{N}$. Since $\mathbb{RP}^{n-1}$ is compact, $(y_t)_{t\in\mathbb{N}}$ has a convergent subsequence. Since any subsequence of $(y_t)_{t\in\mathbb{N}}$ also fulfills $(i)$, $(ii)$, $(iii)$ and $(iv)$ we may assume that $(y_t)_{t\in\mathbb{N}}$ converges to $z_y \in \mathbb{RP}^{n-1}$. The assumption $z_y \in \mathcal{N}_A$ implies that $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable (see Theorem 4.10, b). But then $S_{\Sigma^{II}(A)_{|\mathcal{N}_A}} = G_{\Sigma^{II}(A)_{|\mathcal{N}_A}}$ by Theorem 2.39, $S_{\Sigma^{II}(A)} = G_{\Sigma^{II}(A)}$ by Theorem 3.12 and $S(A)\mathbb{R}^* = P(A)$ by Theorem 6.18. Thus, $z_y \in \partial\mathcal{N}_A$.
b) Recall that $G_{\Sigma^{II}(A)} \cdot z$ is $\Sigma$-invariant. Since $\mathcal{N}_A$ is open and dense in $\mathbb{RP}^{n-1}$ we have $\mathbb{RP}^{n-1} \setminus \mathcal{N}_A = \partial\mathcal{N}_A$. By Theorem 4.18, $G_{\Sigma^{II}(A)} \cdot z \cap \overline{\mathcal{R}(x)} = \emptyset$ implies $G_{\Sigma^{II}(A)} \cdot z \cap \overline{\mathcal{R}(y)} = \emptyset$ for any $y \in \mathcal{N}_A$. Thus $G_{\Sigma^{II}(A)} \cdot z$ is repelling to $\mathcal{N}_A$.

□

In Theorem 6.23 we presented certain eigenvalue constellations where $P(A) \neq S(A)\mathbb{R}^*$. Now Theorem 6.34 implies the appearance of repelling phenomena for these eigenvalue constellations. For a pair of complex eigenvalues $\lambda, \overline{\lambda}$ of $A$ we call

$$\mathcal{E}_\lambda := \pi\left(\{x \in \mathbb{R}^n \setminus \{0\} \,|\, (A^2 - 2\operatorname{Re}(\lambda) + |\lambda|^2 I)x = 0\}\right) \subseteq \mathbb{RP}^{n-1}$$

the *eigenspace corresponding to* $\lambda$. Here, $\pi : \mathbb{R}^n \setminus \{0\} \to \mathbb{RP}^{n-1}$ denotes the canonical projection. Note that $\mathcal{E}_\lambda$ is a $\Sigma$-invariant subspace of $\Sigma_A^{II}$.

**Corollary 6.35** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic and spec$(A) := \{\lambda_1, \dots \lambda_n\}$ the set of eigenvalues of $A$.*

*a) Let $\lambda_1 \in \mathbb{R}$ $\lambda_2 \in \mathbb{C}\setminus\mathbb{R}$ and $\operatorname{Re}(\lambda_1) = \operatorname{Re}(\lambda_2)$, each with multiplicity 1. Then the eigenspace corresponding to $\lambda_2, \overline{\lambda_2}$, is repelling to $\mathcal{N}_A$.*

*b) Let $\lambda_1, \lambda_2 \in \mathbb{C}\setminus\mathbb{R}$ with $\operatorname{Re}(\lambda_1) = \operatorname{Re}(\lambda_2)$ but $|\operatorname{Im}(\lambda_1)| < |\operatorname{Im}(\lambda_2)|$. Then the eigenspace corresponding to $\lambda_2$ is repelling to $\mathcal{N}_A$.*

**Proof.** Without loss of generality we may assume that the matrices are of size $\mathbb{R}^{n\times n}$ with $n = 3$, and respectively $n = 4$. Let $x \in \mathcal{N}_A$. All eigenspaces of $A$ are elements of $\partial(G_\Sigma \cdot x)$, since $\mathcal{N}_A$ is open and dense in $\mathbb{RP}^{n-1}$. By Theorem 6.34 it is sufficient to show that $\overline{\mathcal{R}(x)} \cap \mathcal{E} = \emptyset$ for one $\mathcal{E} \in \mathcal{N}_A$.
a) Let $\lambda_1 \in \mathbb{R}$. Then in the canonical form of $A$ is

$$J = \begin{pmatrix} \operatorname{Re}(\lambda_2) & \operatorname{Im}(\lambda_2) & 0 \\ -\operatorname{Im}(\lambda_2) & \operatorname{Re}(\lambda_2) & 0 \\ 0 & 0 & \lambda_1 \end{pmatrix}.$$

By Lemma 6.20 we can assume

$$J = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

since $\lambda_1 = \operatorname{Re}(\lambda_2)$. Recall that

$$G_A := \left\{ \begin{pmatrix} b & c & 0 \\ -c & b & 0 \\ 0 & 0 & a \end{pmatrix} \,\middle|\, a \neq 0, b^2 + c^2 \neq 0 \right\}.$$

In Theorem 6.35 we have already seen, that not all elements of $G_A$ can be realized with elements in $S(A)\mathbb{R}^*$. In particular, assuming

$$\begin{pmatrix} b & c & 0 \\ -c & b & 0 \\ 0 & 0 & a \end{pmatrix} \in S(A)\mathbb{R}^*$$

implies that there exists $T \in \mathbb{N}$ and $u_t \in U$ such that

$$\begin{aligned}
(I) \quad \begin{pmatrix} b & c \\ -c & b \end{pmatrix} &= r \prod_{t=1}^{T} \left( \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - u_t I \right)^{-1} \\
(II) \quad a &= r \prod_{t=1}^{T} (-u_t)^{-1}.
\end{aligned}$$

Applying the determinant function $B \mapsto \det(B)$ on Equation $(I)$ and $(II)$ we obtain

$$\frac{b^2+c^2}{a^2} = \prod_{t=1}^{T} \frac{u_t^2}{1+u_t^2}.$$

Hence, $b^2 + c^2 < a^2$. Now we show that $\overline{\mathcal{R}(x)} \cap \operatorname{span}(e_1, e_2) = \emptyset$ for $x = (1,1,1)^\top$, $e_1 = (1,0,0)^\top$ and $e_2 = (0,1,0)^\top$. Assume that there exists a sequence $s_n x \to \alpha e_1 + \beta e_2 \neq 0$ with $s_n \in S(A)\mathbb{R}^*$. Then $b_n + c_n \to \alpha$, $b_n - c_n \to \beta$ with $\alpha^2 + \beta^2 \neq 0$ and $a_n \to 0$. But this is impossible since

$$a_n^2 > b_n^2 + c_n^2 = \alpha^2 + \beta^2.$$

b) By Lemma 6.20 we can assume

$$J = \begin{pmatrix} 0 & 1 & 0 & \\ -1 & 0 & 0 & \\ 0 & 0 & 0 & \beta \\ 0 & 0 & -\beta & 0 \end{pmatrix}$$

with $\beta > 1$ since $\operatorname{Re}(\lambda_1) = \operatorname{Re}(\lambda_2)$. Recall that

$$G_A := \left\{ \begin{pmatrix} a & b & 0 & 0 \\ -b & a & 0 & 0 \\ 0 & 0 & c & d \\ 0 & 0 & -d & c \end{pmatrix} \;\middle|\; a^2 + b^2 \neq 0, c^2 + d^2 \neq 0 \right\}.$$

Suppose $g \in G_A$ is an element of $S(A)\mathbb{R}^*$ then there exist controls $u_1, \ldots, u_T \in \mathbb{R} \setminus \{0\}$ such that $I = r \prod_{t=1}^{T} (A - u_t I)$ for any $r \in \mathbb{R}^*$, then both of the following equations are fulfilled

$$\begin{aligned}
(I) \quad \begin{pmatrix} a & b \\ -b & a \end{pmatrix} &= r \prod_{t=1}^{T} \left( \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - u_t I \right) \\
(II) \quad \begin{pmatrix} c & d \\ -d & c \end{pmatrix} &= r \prod_{t=1}^{T} \left( \begin{pmatrix} 0 & \beta \\ -\beta & 0 \end{pmatrix} - u_t I \right).
\end{aligned}$$

Using the determinant function it follows

$$\frac{a^2+b^2}{c^2+d^2} = \prod_{t=1}^{T} \frac{\beta^2+u_t^2}{1+u_t^2} > 1.$$

Similar to (i) we obtain $\overline{\mathcal{R}(x)} \cap \operatorname{span}(e_3, e_4) = \emptyset$ for $x = (1,1,1,1)^\top$, $e_3 = (0,0,1,0)^\top$ and $e_4 = (0,0,0,1)^\top$. Assume, that there exists a sequence $s_n x \to \gamma e_3 + \delta e_4 \neq 0$ with $s_n \in S(A)\mathbb{R}^*$. Then $b_n + c_n \to 0$, $b_n - c_n \to 0$, $c_n + d_n \to \gamma$ and $c_n - d_n \to \delta$ with $\gamma^2 + \delta^2 \neq 0$. But this is impossible since

$$a_n^2 + b_n^2 > c_n^2 + d_n^2 = \gamma^2 + \delta^2.$$

□

## 6.8 Inverse iteration on $\mathbb{RP}^{n-1}$ for small dimensions

We finish our analysis of classical inverse iteration systems with the investigation of reachable sets for matrices $A \in \mathbb{R}^{n\times n}$, $n = 2, 3, 4$. A necessary condition for the existence of large reachable sets, in the sense that they have open interior in $\mathbb{RP}^{n-1}$, is that $A$ is cyclic (see Theorem 6.15). Therefore, we focus on systems $\Sigma^{II}(A)$ with respect to cyclic matrices $A \in \mathbb{R}^{n\times n}$. Recall that the adherence structure of reachable sets of $\Sigma^{II}(A)$ is invariant to similarity transformations. Therefore, we may assume, that $A$ is given in Jordan canonical form. Then, the $A$-invariant subspaces are spanned[23] by the canonical basis vectors $e_1, \ldots, e_n$. The orbit graph is easily obtained, since it is finite and isomorphic to the subspace graph (see Theorem 6.14). The reachable graph is either isomorphic to the orbit graph (if $P(A) = S(A)\mathbb{R}^*$) or infinite (if $P(A) \neq S(A)\mathbb{R}^*$).

### 6.8.1 Inverse iteration on $\mathbb{RP}^1$

Any cyclic matrix $A \in \mathbb{R}^{2\times 2}$ has a Jordan canonical form of the following types:

**Type 1:** $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ with $\lambda_1, \lambda_2 \in \mathbb{R}$ and $\lambda_1 \neq \lambda_2$,

**Type 2:** $\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$ with $\lambda \in \mathbb{R}$,

**Type 3:** $\begin{pmatrix} \operatorname{Re}\lambda & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & \operatorname{Re}\lambda \end{pmatrix}$ with $\operatorname{Im}\lambda \neq 0$.

By Proposition 6.21 and Theorem 6.29 we always have $G_{\Sigma^{II}(A)} = S_{\Sigma^{II}(A)}$. This verifies the known fact, that the restricted system $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is always controllable, provided $n = 2$ (See [HW01], Proposition 12,a). Thus, the reachable graph and the orbit graph coincide. Thus, the reachable graph $\mathcal{G}_{\mathcal{R}}(\Sigma^{II}(A)$ is given by

$$\mathcal{N}_{\operatorname{span}(e_1)} \longleftarrow \mathcal{N}_A \longrightarrow \mathcal{N}_{\operatorname{span}(e_2)}$$

if $A$ is diagonalizable (Type 1),

$$\mathcal{N}_{\operatorname{span}(e_1)} \longleftarrow \mathcal{N}_A$$

if $A$ has an real eigenvalue of multiplicity 2 (Type 2) and trivial otherwise (Type 3).

[23] but not every subspace spanned by canonical basis vectors is an $A$-invariant subspace

### 6.8.2 Inverse iteration on $\mathbb{RP}^2$

For the case $n = 3$ there exist four different types of cyclic Jordan canonical forms. More precisely, $A$ is similar to one of the following matrices

**Type 1:** $$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \quad \text{with } \lambda_1, \lambda_2, \lambda_3 \in \mathbb{R} \text{ and } \lambda_1 < \lambda_2 < \lambda_3,$$

**Type 2:** $$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 1 \\ 0 & 0 & \lambda_2 \end{pmatrix} \quad \text{with } \lambda_1, \lambda_2 \in \mathbb{R} \text{ and } \lambda_1 \neq \lambda_2,$$

**Type 3:** $$\begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \quad \text{with } \lambda \in \mathbb{R},$$

**Type 4:** $$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \operatorname{Re}\lambda_2 & \operatorname{Im}\lambda_2 \\ 0 & -\operatorname{Im}\lambda_2 & \operatorname{Re}\lambda_2 \end{pmatrix} \quad \text{with } \lambda_1 \in \mathbb{R} \text{ and } \operatorname{Im}\lambda_2 \neq 0.$$

For Type 1 and Type 2 we obtain $G_{\Sigma^{II}(A)} = S_{\Sigma^{II}(A)}$ by Theorem 6.29. Thus, the reachable graph and the orbit graph coincide. If $A$ is diagonalizable (Type 1), the reachable graph is given by

$$\begin{tikzcd}
 & & \mathcal{N}_A \arrow[dll] \arrow[d] \arrow[drr] & & \\
\mathcal{N}_{\mathrm{span}(e_1,e_2))} \arrow[dr] \arrow[drr] & & \mathcal{N}_{\mathrm{span}(e_1,e_3)} \arrow[dl] \arrow[dr] & & \mathcal{N}_{\mathrm{span}(e_2,e_3)} \arrow[dll] \arrow[dl] \\
 & \mathcal{N}_{\mathrm{span}(e_1)} & \mathcal{N}_{\mathrm{span}(e_2)} & \mathcal{N}_{\mathrm{span}(e_3)} &
\end{tikzcd}$$

If $A$ has two different real eigenvalues (Type 2), the reachable graph is given by

$$\begin{tikzcd}
 & & \mathcal{N}_A \arrow[dl] \arrow[dr] & \\
 & \mathcal{N}_{\mathrm{span}(e_1,e_2)} \arrow[dl] \arrow[dr] & & \mathcal{N}_{\mathrm{span}(e_2,e_3)} \arrow[dl] \\
\mathcal{N}_{\mathrm{span}\,(e_1)} & & \mathcal{N}_{\mathrm{span}\,(e_2)} &
\end{tikzcd}$$

If $A$ has one real eigenvalue of multiplicity 3 (Type 3), then the orbit graph is given by

$$\mathcal{N}_{\mathrm{span}(e_1)} \longleftarrow \mathcal{N}_{\mathrm{span}(e_1,e_2)} \longleftarrow \mathcal{N}_A$$

By Theorem 6.23 we have $G_{\Sigma^{II}(A)} \neq S_{\Sigma^{II}(A)}$. Thus, there exist infinitely many reachable sets in $\mathcal{N}_A$. For each of this reachable sets $\mathcal{R}(y)$ we have

$$\mathrm{span}(e_1) = \mathcal{N}_{\mathrm{span}(e_1)} \subseteq \overline{\mathcal{R}(y)} \subsetneq \overline{\mathcal{R}(y_2)} \subsetneq \overline{\mathcal{R}(y_3)} \ldots$$

for a sequence $(y_t)_t \in \mathcal{N}_A$ (see Proposition 6.33 and Theorem 6.34). Thus, neither $\mathcal{N}_{\mathrm{span}(e_1)}$ nor $\mathcal{N}_{\mathrm{span}(e_1,e_2)}$ is repelling with respect to $\mathcal{N}_A$.

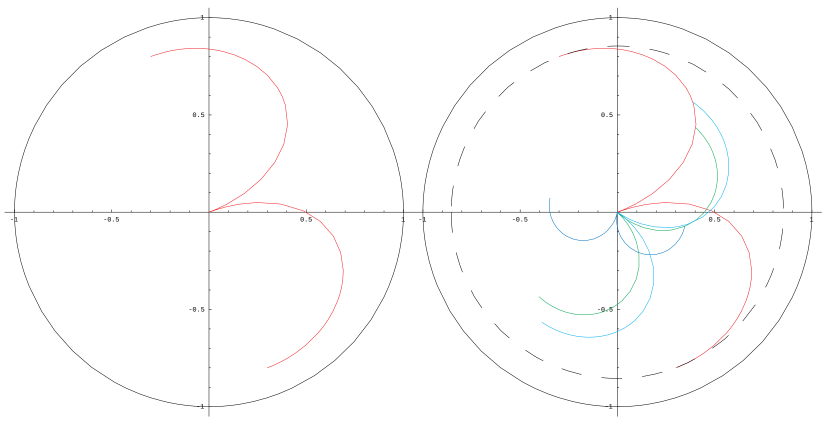



**Figure 2:** *Inverse Iteration for $A \in \mathbb{R}^{3\times 3}$ with $\mathrm{Spec}(A) = \{0, i, -i\}$ (Type 4, Constellation 1). The picture shows possible states of the initial point $x_0 = \mathrm{span}(-0.3, 0.8, 0.854)$ projected on the unit disk. The origin corresponds to $\pi(\mathrm{span}(e_1))$. The boundary of the disk corresponds to $\pi(\mathrm{span}(e_2, e_3))$. On the left hand side we see $\mathcal{R}^1(x_0)$, i.e., the set of all states which can be reached using only one control. On the right hand side we see more possible states after using more then one controls. It is not possible to steer $x_0$ for any sequence of shifts closer than a certain distance (depending on $x_0$) to $\pi(\mathrm{span}(e_2, e_3))$ (see Corollary 6.35).*

If $A$ is of Type 4, then the orbit graph is given by

$$\mathcal{N}_{\mathrm{span}(e_1)} \longleftarrow \mathcal{N}_A \longrightarrow \mathcal{N}_{\mathrm{span}(e_2,e_3)}$$

The adherence structure of reachable sets depends on the constellations of the eigenvalues. We have to distinguish between two cases.

**Constellation 1:** If $\lambda_1 = \operatorname{Re}\lambda_2$ then $P(A) \neq \mathbb{R}^* S(A)$ by Theorem 6.23. Thus, the reachable graph has infinitely many vertices. However, for any $x \in \mathcal{N}_A$ we obtain $\mathcal{N}_{\operatorname{span}(e_1)} \subseteq \overline{\mathcal{R}(x)}$ (see Proposition 6.33). In fact, we can steer any initial state in $\mathcal{N}_A$ arbitrary close to $\operatorname{span}(e_1)$ with only one control (see Figure 2). On the other hand $\mathcal{N}_{\operatorname{span}(e_2,e_3)}$ is repelling with respect to $\mathcal{N}_A$, i.e., $\overline{\mathcal{R}(x)} \cap \mathcal{N}_{\operatorname{span}(e_2,e_3)} = \emptyset$ for all $x \in \mathcal{N}_A$ (see Corollary 6.35).
**Constellation 2:** Now let $\lambda_1 \neq \operatorname{Re}\lambda_2$. We show that $S_{\Sigma^{II}(A)} = G_{\Sigma^{II}(A)}$ and therefore $\mathcal{G}_{\mathcal{R}}(\Sigma^{II}(A)) \cong \mathcal{G}_O(\Sigma^{II}(A))$. By Lemma 6.20 we assume

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \quad \text{with} \quad A_1 = (1) \quad \text{and } A_2 = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}$$

for some $\omega > 0$. Recall that

$$\begin{aligned} P(A) &= \left\{\alpha I + \beta A + \gamma A^2 \,\middle|\, \alpha, \beta, \delta \in \mathbb{R}\right\} \cap \mathrm{GL}_3(\mathbb{R}) \\ &= \left\{ \begin{pmatrix} a & 0 & 0 \\ 0 & b & c \\ 0 & -c & b \end{pmatrix} \,\middle|\, a \neq 0, b^2 + c^2 \neq 0 \right\}. \end{aligned}$$

We define $M_1 = \mathbb{R}^*$ and

$$M_2 = \left\{ \begin{pmatrix} b & c \\ -c & b \end{pmatrix} \,\middle|\, b \neq -\frac{1}{\omega}, c \neq 0, b^2 + c^2 \neq 0 \right\}.$$

Note that $M_2$ is dense in $P(A_2)$.

Now we apply Theorem 6.26, i.e., we show:
**Statement (i)** For any $p(A_1) \in M_1$ there exists $q \in \mathcal{L}$ such that $q(A_1) = p(A_1)$ and $q(A_2) = I_2$.
**Statement (ii)** For any $p(A_2) \in M_2$ there exists $q \in \mathcal{L}$ such that $q(A_2) = p(A_2)$ and $q(A_1) = 1$.
(i) Let

$$q_u(t) = \frac{1}{-(u^2 + \omega^2)}(t - u)(t + u) \quad \text{with} \quad u \in U_A.$$

Then $q_u(A_2) = I_2$ and $q_u(A_1) = \frac{u^2-1}{u^2+\omega^2}$. Note that the image of the map $u \mapsto q_u(A_1)$ is $[-\frac{1}{\omega^2}, 0) \cup (0, 1)$. Thus, in the case $\omega < 1$ there exists $u \in U_A$ such that $q_u^2(A_1) = 1$ and we are done.

In the case $\omega \geq 1$ it is much more complicated to find an adequate $q \in \mathcal{L}$ with the desired requirements. The following construction is similar to the arguments in the proof for Proposition 12 in [HW01]. We define

$$V := \left\{(u_1, u_2) \in (U_A)^2 \,|\, \alpha_{u_1,u_2} := \arg\left((-u_1 + i\omega)(-u_2 + i\omega)\right) \in \pi\mathbb{Q}\right\}.$$

Note that $V$ is dense in $\mathbb{R}^2$. Moreover, for any $(u_1, u_2) \in V$ we have

$$\begin{aligned} ((A_2 - u_1 I_2)(A_2 - u_2 I_2))^m &= T \begin{pmatrix} (r_{u_1,u_2} e^{\alpha_{u_1,u_2}})^{2m} & 0 \\ 0 & (r_{u_1,u_2} e^{\alpha_{u_1,u_2}})^{2m} \end{pmatrix} T^{-1} \\ &= r_{u_1,u_2}^{2m} I_2 \end{aligned}$$

for $T \in \mathrm{GL}_2(\mathbb{C})$, $r^2_{u_1,u_2} = |(-u_1 + i\omega)(-u_2 + i\omega)|^2$ and some $m \in \mathbb{N}$.

Now we show the existence of a pair $(u_1, u_2) \in V$ such that

$$|(1-u_1)(1-u_2)|^2 > r^2_{u_1,u_2}. \tag{54}$$

(54) is equivalent to

$$u_1 u_2 > \frac{1}{2}\frac{\omega^4 + (\omega^2-1)(u_1+u_2)^2 + 2(u_1+u_2) - 1}{1-(u_1+u_2)+\omega^2}. \tag{55}$$

Clearly, the set of solutions of (55) is nonempty. In particular the choice $u_1 = u_2$ yields

$$0 > \frac{1}{2}(\omega^4 + (\omega^2-1)(2u)^2 + 4u - 1 - u^2(\omega^2 + 1 - 2u)$$

which has solutions for any $\omega \in \mathbb{R}^+$. Thus, there exists $(u_1, u_2) \in V$ such that (54) is fulfilled.

Then, $q_{u_1,u_2}(t) := ((t-u_1)(t-u_2)) \in \mathcal{L}$ fulfills

$$\frac{1}{r^{2m}_{u_1,u_2}} q^m_{u_1,u_2}(A_2) = I_2 \quad \text{and} \quad \frac{1}{r^{2m}_{u_1,u_2}} q^m_{u_1,u_2}(A_1) > 1.$$

This proves that for any $r > 0$ (and in particular for $r = 1$) there exists $k_r \in \mathbb{N}$ and $u \in U_A$ such that

$$\left(\frac{1}{r^{2m}_{u_1,u_2}} q^m_{u_1,u_2}\right)^{k_r} q_u(A_1) = r \quad \text{and} \quad \left(\frac{1}{r^{2m}_{u_1,u_2}} q^m_{u_1,u_2}\right)^{k_r} q_u(A_2) = I_2.$$

(ii) For any $p(A_2) = \begin{pmatrix} b & c \\ -c & b \end{pmatrix} \in M_2$ we choose $q(t) = \frac{c}{\omega}(t - b\omega)$. Clearly $q(A_2) = p(A_2)$ and $q(A_1) = \frac{c}{\omega}(1 + b\omega)$. From (i) we know, that there exists $\tilde{q} \in \mathcal{L}$ such that $\tilde{q}(A_1) = \frac{1}{\frac{c}{\omega}(1+b\omega)}$ and $\tilde{q}(A_2) = I_2$. Thus, $q\tilde{q}$ fulfills $q\tilde{q}(A_2) = p(A_2)$ and $q\tilde{q}(A_1) = 1$.
From $(i)$ and $(ii)$ we conclude $S(A)\mathbb{R}^* = P(A)$ by Theorem 6.26 and therefore $S_{\Sigma^{II}(A)} = G_{\Sigma^{II}(A)}$. In particular this shows that there exists a control sequence which steers any initial state $x_0 \in \mathcal{N}_A$ arbitrary close to $\pi(\mathrm{span}(e_2, e_3))$. However, from the proof it is not clear if the number of steps is limited. In Figure 3 we see a possible trajectory for such a steering.

Recall that $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable if and only if $S(A)\mathbb{R}^* = P(A)$. Thus, the results in Section 6.8.2 verify the controllability results of Helmke and Wirth in [HW01] (Proposition 12,b). Moreover, we have characterized the cases where repelling phenomena occur. The following Theorem summerizes the results of this subsection.

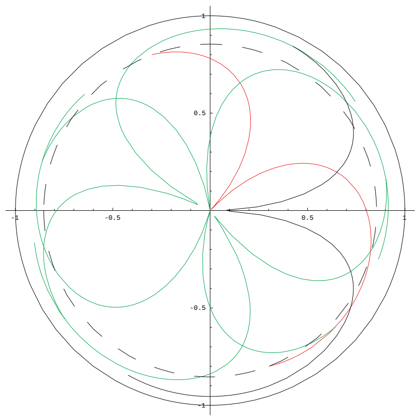



**Figure 3:** *Inverse Iteration for $A$ with eigenvalues $\lambda_1 = 0.1$, $\lambda_2 = i$ and $\lambda_3 = -i$. Again we see possible states of the initial point $x_0 = \pi(-0.3, 0.8, 0.854)$ projected on the unit disk. Here, there exists a sequence of controls such that the sequence of states converges to $\pi(\mathrm{span}(e_2, e_3))$.*

**Theorem 6.36** *Consider classical inverse iteration for a cyclic matrix $A \in \mathbb{R}^{3\times 3}$.*

*a) The restricted system $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable if and only if $A$ is of Type 1, of Type 2 or of Type 4 with $\lambda_1 \neq \mathrm{Re}\,\lambda_2$.*

*b) The repelling phenomenon occurs if and only if $A$ is of Type 4 with $\lambda_1 \neq \mathrm{Re}\,\lambda_2$.*

### 6.8.3 Inverse iteration on $\mathbb{RP}^3$

In the case $n = 4$ any cyclic matrix is similar to one of the following types:

**Type 1:**
$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix} \qquad \begin{array}{l} \lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \mathbb{R}, \\ \lambda_i \neq \lambda_j \text{ for } i \neq j; \end{array}$$

**Type 2:**
$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 1 \\ 0 & 0 & 0 & \lambda_3 \end{pmatrix} \qquad \begin{array}{l} \lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}, \\ \lambda_i \neq \lambda_j \text{ for } i \neq j; \end{array}$$

**Type 3:**
$$\begin{pmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & \lambda_2 \end{pmatrix} \qquad \begin{array}{l} \lambda_1, \lambda_2 \in \mathbb{R}, \\ \lambda_1 \neq \lambda_2; \end{array}$$

**Type 4:**
$$\begin{pmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{pmatrix} \qquad \begin{array}{l} \lambda_1, \lambda_2 \in \mathbb{R}, \\ \lambda_1 \neq \lambda_2; \end{array}$$

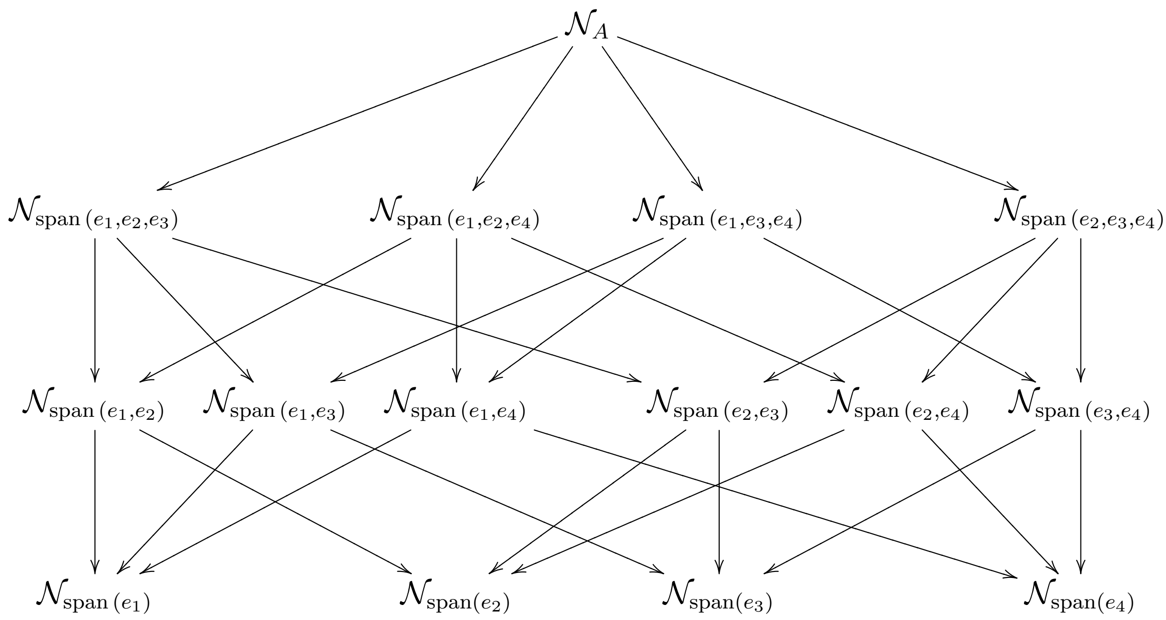
**Type 5:**
$$\begin{pmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix} \qquad \text{with } \lambda \in \mathbb{R};$$

**Type 6:**
$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & & \operatorname{Re}\lambda_3 & \operatorname{Im}\lambda_3 \\ 0 & & -\operatorname{Im}\lambda_3 & \operatorname{Re}\lambda_3 \end{pmatrix} \qquad \begin{array}{l} \lambda_1, \lambda_2 \in \mathbb{R}, \\ \lambda_1 \neq \lambda_2, \\ \operatorname{Im}\lambda_3 \neq 0; \end{array}$$

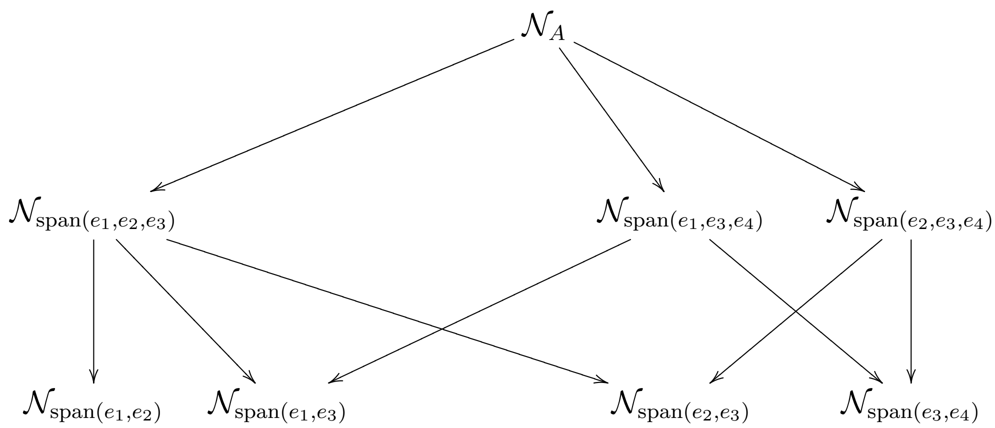
**Type 7:**
$$\begin{pmatrix} \operatorname{Re}\lambda_1 & \operatorname{Im}\lambda_1 & 0 & 0 \\ -\operatorname{Im}\lambda_1 & \operatorname{Re}\lambda_1 & 0 & 0 \\ 0 & & \operatorname{Re}\lambda_2 & \operatorname{Im}\lambda_2 \\ 0 & & -\operatorname{Im}\lambda_2 & \operatorname{Re}\lambda_2 \end{pmatrix} \qquad \begin{array}{l} \operatorname{Im}\lambda_1 \neq 0, \\ \operatorname{Im}\lambda_2 \neq 0, \\ \lambda_1 \neq \lambda_2, \\ \lambda_1 \neq \overline{\lambda_2}; \end{array}$$

**Type 8:**
$$\begin{pmatrix} \operatorname{Re}\lambda & \operatorname{Im}\lambda & 0 & 0 \\ -\operatorname{Im}\lambda & \operatorname{Re}\lambda & 1 & 0 \\ 0 & & \operatorname{Re}\lambda & \operatorname{Im}\lambda \\ 0 & & -\operatorname{Im}\lambda & \operatorname{Re}\lambda \end{pmatrix} \qquad \text{with } \operatorname{Im}\lambda \neq 0.$$

If $A$ is of Type 1, Type 2 or Type 3, then $P(A) = S(A)\mathbb{R}^*$ by Theorem 6.29. Thus, the reachable sets and their adherence structure are completely described by the orbit graph. If $A$ is of Type 1, the orbit graph is given by




If $A$ is of Type 2, the orbit graph is given by



If $A$ is of Type 3, the orbit graph is given by

$$\begin{tikzcd}
 & & \mathcal{N}_A \arrow[dl] \arrow[dr] & & \\
 & \mathcal{N}_{\mathrm{span}(e_1,e_2,e_3)} \arrow[dl] \arrow[dr] & & \mathcal{N}_{\mathrm{span}(e_1,e_3,e_4)} \arrow[dl] \arrow[dr] & \\
\mathcal{N}_{\mathrm{span}(e_1,e_2)} \arrow[dr] & & \mathcal{N}_{\mathrm{span}(e_1,e_3)} \arrow[dl] \arrow[dr] & & \mathcal{N}_{\mathrm{span}(e_3,e_4)} \arrow[dl] \\
 & \mathcal{N}_{\mathrm{span}(e_1)} & & \mathcal{N}_{\mathrm{span}(e_3)} &
\end{tikzcd}$$

If $A$ is of Type 4, Type 5 or Type 8, then we have $P(A) \neq S(A)\mathbb{R}^*$ by Theorem 6.23. Thus the reachable graph is infinite. Nevertheless, the corresponding orbit graphs are easy to deduce by Theorem 6.14. If $A$ is of Type 4, the orbit graph is given by

$$\begin{tikzcd}
 & & \mathcal{N}_A \arrow[dl] \arrow[dr] & & \\
 & \mathcal{N}_{\mathrm{span}(e_1,e_2,e_3)} \arrow[dl] & & \mathcal{N}_{\mathrm{span}(e_1,e_2,e_4)} \arrow[dlll] \arrow[dr] & \\
\mathcal{N}_{\mathrm{span}((e_1,e_2)} \arrow[dr] & & & & \mathcal{N}_{\mathrm{span}((e_1,e_4)} \arrow[dl] \\
 & \mathcal{N}_{\mathrm{span}(e_1)} & & \mathcal{N}_{\mathrm{span}(e_4)} &
\end{tikzcd}$$

If $A$ is of Type 5, the orbit graph is given by

$$\mathcal{N}_{\mathrm{span}(e_1)} \longleftarrow \mathcal{N}_{\mathrm{span}(e_1,e_2)} \longleftarrow \mathcal{N}_{\mathrm{span}(e_1,e_2,e_3)} \longleftarrow \mathcal{N}_A$$

If $A$ is of Type 8, the orbit graph is given by

$$\begin{tikzcd} \mathcal{N}_{\mathrm{span}(e_1,e_2)} & \mathcal{N}_A \arrow[l] \end{tikzcd}$$

If $A$ is of Type 6, the orbit graph is given by

$$\begin{tikzcd}
 & \mathcal{N}_A \arrow[ddl] \arrow[dr] \arrow[drr] & & \\
 & & \mathcal{N}_{\mathrm{span}(e_2,e_3,e_4)} \arrow[dr] \arrow[ddl] & \mathcal{N}_{\mathrm{span}(e_1,e_3,e_4)} \arrow[d] \arrow[ddlll] \\
\mathcal{N}_{\mathrm{span}(e_1,e_2)} \arrow[d] \arrow[dr] & & & \mathcal{N}_{\mathrm{span}(e_3,e_4)} \\
\mathcal{N}_{\mathrm{span}(e_1)} & \mathcal{N}_{\mathrm{span}(e_2)} & &
\end{tikzcd}$$

The answer to the question if $S(A)\mathbb{R}^* = P(A)$ or not, depends on the constellations of the eigenvalues. By Theorem 6.23 we have $S(A)\mathbb{R}^* \neq P(A)$ if $\operatorname{Re}\lambda_3 = \frac{\lambda_1+\lambda_2}{2}$ and $\lambda_1 < \operatorname{Re}\lambda_3 + \operatorname{Im}\lambda_3$. However, it is unknown if $S(A)\mathbb{R}^* = P(A)$ holds for any other constellation (of Type 6).

Now we assume that $A$ is of Type 7. The orbit graph is given by

$$\begin{tikzcd}
 & \mathcal{N}_A \arrow[dl] \arrow[dr] & \\
\mathcal{N}_{\mathrm{span}(e_1,e_2)} & & \mathcal{N}_{\mathrm{span}(e_3,e_4)}
\end{tikzcd}$$

If $\operatorname{Re}\lambda_1 = \operatorname{Re}\lambda_2$, then $S(A)\mathbb{R}^* \neq P(A)$ by Theorem 6.23. It is unknown, if equation $S(A)\mathbb{R}^* = P(A)$ holds for any other eigenvalue constellation (of Type 7).

The following theorem summarizes Section 6.8.3 with respect to the controllability properties of the restricted system $\Sigma^{II}(A)_{|\mathcal{N}_A}$.

**Theorem 6.37** *Consider classical inverse iteration on $\mathbb{RP}^3$ for a cyclic matrix $A \in \mathbb{R}^{4\times 4}$.*

*a) If $A$ is of Type 1, Type 2 or Type 3 then $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable.*

*b) If $A$ is of Type 4, Type 5 or Type 8 then $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is not controllable.*

*c) If $A$ is of Type 6 with $2\operatorname{Re}\lambda_3 = \lambda_1 + \lambda_2$ and $\lambda_1 < \operatorname{Re}\lambda_3 + \operatorname{Im}\lambda_3$ or of Type 7 with $\operatorname{Re}\lambda_1 = \operatorname{Re}\lambda_2$ then $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is not controllable.*

In the remaining cases it is unclear, if $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable. In particular it is unclear if the set of all cyclic matrices $A \in \mathbb{R}^{4\times 4}$, where $\Sigma^{II}(A)_{|\mathcal{N}_A}$ is controllable is generic in $\mathbb{R}^{4\times 4}$.

# 7 Generalized inverse iteration systems

Classical inverse iteration schemes are mainly designed for eigenvector computation. Therefore, their dynamic naturally evolves on the projective space. Nevertheless, different generalizations appear in several situations, such as in the dynamics of the QR algorithm. In the following we investigate inverse iteration systems on flag manifolds (Section 7.1), on Hessenberg varieties (Section 7.2)and on real vector spaces (Section 7.3). The following setting generalizes classical inverse iteration together with all these cases.

**Definition 7.1 (Generalized inverse iteration system)** Let $M$ be a topological space and $\alpha : \mathrm{GL}_n(\mathbb{R}) \times M \to M$ be a transitive group action. For a given matrix $A \in \mathbb{R}^{n\times n}$, we define $U_A := \mathbb{R} \setminus \mathrm{Spec}(A)$ and

$$f_A^{II} : (x, u) \mapsto (A - uI)^{-1} \cdot x.$$

We call the corresponding system $\Sigma^{II}(A) := (M, U_A, f_A^{II})$ the *inverse iteration system of $A$ on $M$* (with respect to $\alpha$).

In particular the case $M = \mathbb{RP}^{n-1}$ with the canonical action yields classical inverse iteration. Clearly, the system group of $\Sigma^{II}(A)$ is related to the matrix semigroup

$$S(A) := \left\{ \prod_{t=1}^{T} (A - u_t I)^{-1} \,|\, T \in \mathbb{N}, u_t \in U_A \right\}.$$

More precisely we obtain:

**Proposition 7.2** *Consider the generalized inverse iterations system $\Sigma^{II}(A)$ $:= (M, U_A, f_A^{II})$ with respect to a group action $\alpha$. The system group $G_{\Sigma^{II}(A)}$ of an inverse iteration system of $A$ on $M$ with respect to $\alpha$ is isomorphic to the group $P(A)/(P(A) \cap C_M)$ where $C_M := \bigcap_{x\in M} \mathrm{Stab}_x$.*

**Proof.** Recall that $\langle S(A)\rangle = P(A)$ (see Theorem 6.3). Two matrices $B, \tilde{B} \in P(A)$ induce the same maps $x \mapsto B \cdot x$, respectively $x \mapsto \tilde{B} \cdot x$ if and only if $B\tilde{B}^{-1}$ is an element of $\mathrm{Stab}_x$ for all $x \in M$. Therefore, the kernel of the group homomorphism $\Phi : P(A) \to G_\Sigma$, $\Phi(B) : x \mapsto B \cdot x$ is $P(A) \cap C_M$. □

## 7.1 Inverse iteration on flag manifolds

In this section we consider inverse iteration systems on flag manifolds, i.e., $\Sigma^{II}(A) = (\mathrm{Flag}(d, \mathbb{R}^n), U_A, f_A^{II})$ with respect to the canonical group action

$$\mathrm{GL}_n(\mathbb{R}) \times \mathrm{Flag}(d, \mathbb{R}^n) \to \mathrm{Flag}(d, \mathbb{R}^n), \quad g \cdot \mathcal{V} = (g(V_1), \dots, g(V_k)))$$

for $\mathcal{V} = (V_1, \ldots, V_k)$. See Appendix F for an introduction on flag manifolds and Section 5.2.1 for general results on systems on flag manifolds. In particular, inverse iteration systems on complete flag manifolds, i.e., $\mathrm{Flag}(\mathbb{R}^n) = \mathrm{Flag}(d, \mathbb{R}^n)$ with $d = (1, 2 \ldots, n-1)$, are of interest. In this situation, $\Sigma^{II}(A)$ is closely related to the shifted $QR$ algorithms. More precisely, a $QR$-step applied on an operator $A - uI \in \mathrm{GL}_n(\mathbb{R})$ with respect to a basis $e_1, \ldots, e_n$ of $\mathbb{R}^n$ is equivalent to one power iteration step $x_{t+1} = (A - uI)x_t$. See [AM86, Amm86, Wat82] for a more detailed description.

The structure of reachable sets for inverse iteration on $\mathrm{Flag}(\mathbb{R}^n)$ is much more complicated as in the classical case. The main reason lies in the fact that the orbit graph is infinite, even if $A$ is cyclic.

**Theorem 7.3** *Consider the inverse iteration system $\Sigma^{II}(A)$ on $\mathrm{Flag}(d, \mathbb{R}^n)$.*

*a) If $n \geq 3$ and $d \notin \{(1), (n-1)\}$ then $\mathrm{Flag}(d, \mathbb{R}^n)$ is a partition of infinitely many different systemgroup orbits.*

*b) If $A$ is cyclic and $d_1 = 1$ then the following statements are equivalent.*

*(i) $S(A)\mathbb{R}^* = P(A)$.*

*(ii) The reachable structure $\mathcal{G}_R(\Sigma^{II}(A))$ coincides with the orbit structure $\mathcal{G}_O(\Sigma^{II}(A))$.*

*(iii) There exists $\mathcal{V} = (V_1, \ldots, V_k) \in \mathrm{Flag}(d, \mathbb{R}^n)$ with $V_1 \in \mathcal{N}_A$ such that $G_{\Sigma^{II}(A)} \cdot \mathcal{V} = \mathcal{R}_{\Sigma^{II}(A)}(\mathcal{V})$.*

**Proof.** Both statements can be deduced from the results of Section 3 and Section 5.2.1. Consider

$$\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A) = (\mathrm{GL}_n(\mathbb{R}), U_A, \hat{f}^{II}) \quad \text{with} \quad \hat{f}^{II}(g, u) = (A - uI)^{-1}g.$$

Recall that here $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = S(A)$ and $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = P(A)$. We choose a reference flag $\mathcal{V} = (V_1, \ldots, V_k)$. Then, $\Sigma^{II}(A)$ is an induced system of $\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)$ with respect to $\pi_{\mathcal{V}} : \mathrm{GL}_n(\mathbb{R}) \to \mathrm{Flag}(d, \mathbb{R}^n)$, $x \mapsto g \cdot \mathcal{V}$ (see Theorem 5.9) and thus, $C_{\pi_{\mathcal{V}}} = \mathbb{R}^* I$. Moreover, $G_{\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)} = P(A)$ and thus, by Theorem 5.9, $C_\pi = \mathbb{R}^* I$.
a) Recall that $P(A)$ is a Lie group of dimension $m - 1$ where $m$ is the degree of the minimal polynomial of $A$. Moreover, $G_{\Sigma^{II}(A)}$ carries a Lie group structure such that $G_{\Sigma^{II}(A)}$ is isomorphic to $P(A)/C_{\pi_{\mathcal{V}}}$ and therefore, $\dim G_{\Sigma^{II}(A)} < n - 1$. Thus, $\dim G_{\Sigma^{II}(A)} \cdot \mathcal{V}$, which is an immersed submanifold by Theorem 2.5, is smaller then $n - 1$. Now the claim follows, since $\dim \mathrm{Flag}(d, \mathbb{R}^n) \geq n$ (see Appendix F).
b) $(i) \Leftrightarrow (ii)$: Recall that $\mathcal{G}_O(\Sigma^{II}(A))$ coincides with $\mathcal{G}_R(\Sigma^{II}(A))$ if and only if $\Sigma^{II}(A)$ is weakly reversible (see Theorem 4.6). Thus, $(i) \Leftrightarrow (ii)$ follows from Theorem 5.8.

$(i) \Rightarrow (iii)$: Assuming, $S(A)\mathbb{R}^* = P(A)$ we have

$$\mathcal{R}_{\Sigma^{II}(A)}(\mathcal{V}) = \mathcal{R}_{\Sigma^{II}(A)}(\pi_{\mathcal{V}}(I)) = \pi_{\mathcal{V}}(S(A)\mathbb{R}^* I) = G_{\Sigma^{II}(A)} \cdot \pi_{\mathcal{V}}(I) = G_{\Sigma^{II}(A)} \cdot \mathcal{V}$$

(see Lemma 3.3). Thus, $(i)$ implies $(iii)$.
$(iii) \Rightarrow (ii)$: If $g \in \mathrm{Stab}_{\mathcal{V}}$, then $g(V_1) = V_1$. Thus, $g \in \mathbb{R}^* I$ by Lemma 6.11. It follows that $\mathrm{Stab}_{\mathcal{V}} \cap P(A) \subseteq \mathbb{R}^* I$. This implies, that $G_{\Sigma^{II}(A)} \cdot \mathcal{W} = \mathcal{R}_{\Sigma^{II}(A)}(\mathcal{W})$ for all $\mathcal{W} \in \mathrm{Flag}(d, \mathbb{R}^n)$ (see Theorem 5.9). Therefore, $\Sigma^{II}(A)$ is weakly reversible by Lemma 2.35. Hence, $\mathcal{G}_O(\Sigma^{II}(A))$ and $\mathcal{G}_R(\Sigma^{II}(A))$ coincide. □

Chu and Chu pointed out, that in general a shifted QR transformation, and therefore inverse iteration on $\mathrm{Flag}(\mathbb{R}^n)$, is not necessarily invertible by a sequence of shifted QR transformations (see ([CC06]). The system semigroup approach explains this phenomenon. In Section 6.5 we have seen various cases, where $S(A)\mathbb{R}^* \neq P(A)$. In this case, not every iteration step is invertible, i.e., there exists $u \in U_A$ such that

$$\prod_{t=1}^{N}(A - u_t I)^{-1} \cdot ((A - uI)^{-1} \cdot \mathcal{V}) \neq \mathcal{V}$$

for any finite control sequence $u_1, \ldots, u_N \in U_A$.

Since there exist infinitely many system group orbits, it is useful to merge related reachable sets to larger classes. The following definition provides a coarser partition of flag manifolds by unions of reachable sets. In the following we focus on the case of complete flag manifolds $\mathrm{Flag}(\mathbb{R}^n)$.

**Definition 7.4** For $A \in \mathbb{R}^{n \times n}$ we denote the set of $A$-invariant subspaces by $\mathrm{Inv}_A$. Two flags $\mathcal{V} = (V_1, \ldots, V_{n-1}) \in \mathrm{Flag}(\mathbb{R}^n)$ and $\mathcal{U} = (U_1, \ldots, U_{n-1}) \in \mathrm{Flag}(\mathbb{R}^n)$ are called equivalent if

$$\dim(U_j \cap W) = \dim(V_j \cap W)$$

for all $W \in \mathrm{Inv}_A$ and all $j = 1, \ldots, n-1$. We denote the set of all flags equivalent to $\mathcal{V}$ by $[\mathcal{V}]$. Moreover, we define a directed graph $\mathcal{G}_{[\,]}(\Sigma^{II}(A)) = (V_{[\,]}, \longleftarrow)$ by the set of equivalence classes $V_{[\,]} := \{[\mathcal{V}] \mid \mathcal{V} \in \mathrm{Flag}(\mathbb{R}^n)\}$ and the relation

$$[\mathcal{U}] \longleftarrow [\mathcal{V}] :\Leftrightarrow [\mathcal{U}] \subseteq \overline{[\mathcal{V}]}.$$

**Theorem 7.5** *Consider the inverse iteration system $\Sigma^{II}(A)$ on $\mathrm{Flag}(\mathbb{R}^n)$.*

*a) Every class $[\mathcal{V}]$ is the disjoint union of system group orbits.*

*b) Let $A$ be cyclic. There exists one class $[\mathcal{V}]$ such that $[\mathcal{U}] \longleftarrow [\mathcal{V}]$ for any $\mathcal{U} \in \mathrm{Flag}(\mathbb{R}^n)$*

*c) If* $[\mathcal{U}] \longleftarrow [\mathcal{V}]$*, then* $\dim(U_j \cap W) \geq \dim(V_j \cap W)$ *for all* $W \in \mathrm{Inv}_A, j = 1, \ldots, n-1$.

**Proof.** a) Let $\mathcal{V} \in \mathrm{Flag}(\mathbb{R}^n)$ and $W \in \mathrm{Inv}_A$. Recall that $P(A)$ acts transitively on $N_W := W \setminus \bigcup_{\mathrm{Imv}_A^W} V$ (see Lemma 6.11). It follows

$$\dim(V_j \cap W) = \dim(p(A)V_j \cap W)$$

for any $p(A) \in P(A)$ and any $j = 1, \ldots, n-1$ and therefore $G_{\Sigma^{II}(A)} \cdot \mathcal{V} \subseteq [\mathcal{V}]$. Thus, $[\mathcal{V}]$ is the union of all system group orbits $G_{\Sigma^{II}(A)} \cdot \mathcal{U}$ with $\mathcal{U} \in [\mathcal{V}]$.
b) If $A$ is cyclic then $\mathrm{Inv}_A$ is finite and $\bigcup_{W \in \mathrm{Inv}_A \setminus \{\mathbb{R}^n\}} W$ is nowhere dense in $\mathbb{R}^n$. Therefore, for any $\mathcal{U} \in [\mathcal{U}]$, we find a sequence $(\mathcal{V}_t)_{t \in \mathbb{N}}$ such that $\mathcal{V}_t \to \mathcal{U}$ and

$$\dim(W \cap V_k^t) = \min\{0, \dim V + \dim W - n\} \leq \dim(W \cap U_j)$$

with $\mathcal{V}_t = (V_1^t, \ldots, V_{n-1}^t)$. Thus, $[\mathcal{U}] \longleftarrow [\mathcal{V}]$.
c) The projection

$$\pi_j : \mathrm{Flag}(\mathbb{R}^n) \to \mathrm{Grass}_j(\mathbb{R}^n),\ \mathcal{V} \mapsto V_i$$

is continuous, and the map

$$F_W : \mathrm{Grass}_i \to \mathbb{N}_0,\ V \mapsto \dim(W \cap V)$$

is upper semicontinuous, i.e., $V_k \to V$ implies $F_W(V_k) \leq F_W(V)$ for $k$ large enough. Therefore, the map $F_{W,j} : \mathrm{Flag}(\mathbb{R}^n) \to \mathbb{N}_0$, $F_{W,j} := F_W \circ \pi_j$ is upper semicontinuous, for all $W \in \mathrm{Inv}_A$ and all $j = 1 \ldots, n-1$. If $[\mathcal{U}] \subseteq \overline{[\mathcal{V}]}$ then every $\mathcal{U}$ in $[\mathcal{U}]$ can be approached with a sequence $(\mathcal{V}_k)_{k \in \mathbb{N}}$ in $[\mathcal{V}]$. I.e., $\mathcal{V}_k \to \mathcal{U}$. Thus,

$$F_{W,j}(\mathcal{V}_k) = \dim(W \cap (V_j)_k) \leq \dim(W \cap U_j)$$

for all $W \in \mathrm{Inv}_A$ and $j = 1, \ldots, n-1$. □

Theorem 7.5 allows us to present information about the adherence structure of reachable sets as a finite graph.

**Example 7.6** We consider the inverse iteration system $\Sigma^{II}(A)$ on $\mathrm{Flag}(d, \mathbb{R}^n)$ with $d = (1, 2)$ and with respect to

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Recall that $S(A)\mathbb{R}^* = P(A)$ (see Section 6.8.2). Thus the system group orbits and the reachable sets coincide. By Theorem 7.5 every class $[\mathcal{V}]$ is the disjoint union of reachable sets. There exist two nontrivial $A$-invariant subspaces $E_1 := \mathrm{span}(e_1)$ and $E_2 := \mathrm{span}(e_2, e_3)$. We identify the equivalence classes $[\,]$ with the values of the map $\Phi : \mathrm{Flag}(\mathbb{R}^3) \to \{0, 1, 2\}^4$ defined by

$$(V_1, V_2) \mapsto (\dim(V_1 \cap E_1), \dim(V_2 \cap E_1), \dim(V_1 \cap E_2), \dim(V_2 \cap E_2))\,.$$

Note that $\Phi$ is not surjective. Clearly, $\dim(V_1 \cap E_1) \le 1$, $\dim(V_2 \cap E_1) \le 1$ and $\dim(V_1 \cap E_2) \le 1$. Moreover, easy linear algebra arguments show further restrictions. In fact, six classes exist. With the notation $N_A := \mathbb{R}^3 \backslash \{E_1 \cup E_2\}$ we obtain

$$\begin{aligned}
\Phi^{-1}(0,0,0,1) &= \{(\mathrm{span}(x), \mathrm{span}(x,y)) \in \mathrm{Flag}(\mathbb{R}^3) \,|\, x \in N_A, y \in N_A \cup E_2\},\\
\Phi^{-1}(0,1,0,1) &= \{(\mathrm{span}(x), \mathrm{span}(x,y)) \in \mathrm{Flag}(\mathbb{R}^3) \,|\, x \in N_A, y \in E_1\},\\
\Phi^{-1}(1,1,0,1) &= \{(\mathrm{span}(x), \mathrm{span}(x,y)) \in \mathrm{Flag}(\mathbb{R}^3) \,|\, x \in E_1, y \in N_A \cup E_2\},\\
\Phi^{-1}(0,0,1,1) &= \{(\mathrm{span}(x), \mathrm{span}(x,y)) \in \mathrm{Flag}(\mathbb{R}^3) \,|\, x \in E_2, y \in N_A\},\\
\Phi^{-1}(0,1,1,1) &= \{(\mathrm{span}(x), \mathrm{span}(x,y)) \in \mathrm{Flag}(\mathbb{R}^3) \,|\, x \in E_2, y \in E_1\},\\
\Phi^{-1}(0,0,1,2) &= \{(\mathrm{span}(x), \mathrm{span}(x,y)) \in \mathrm{Flag}(\mathbb{R}^3) \,|\, x \in E_2, y \in E_2\}.
\end{aligned}$$

By Theorem 7.5, the graph $\mathcal{G}_{[\,]}(\Sigma^{II}(A))$ is given by

$$\begin{tikzcd}
 & (0,0,0,1) \arrow[dl] \arrow[d] & \\
(0,1,0,1) \arrow[d] \arrow[dr] & (0,0,1,1) \arrow[d] \arrow[dr] & \\
(1,1,0,1) & (0,1,1,1) & (0,0,1,2)
\end{tikzcd}$$

## 7.2 Inverse iteration on Hessenberg varieties

In numerical computations one often transforms a matrix $A$ first into Hessenberg form and then applies the QR algorithm to this condensed form. Since the QR algorithm preserves the Hessenberg structure it restricts to a control system on the set of Hessenberg flags. This system can be interpreted as an inverse iteration system on a certain subset of $\mathrm{Flag}(\mathbb{R}^n)$, the Hessenberg variety. See [AM86, Amm87, DS88] for more details. In the following we analyze the structure of reachable sets of inverse iteration systems on Hessenberg varieties.

For a given matrix $A$, the Hessenberg variety is defined as the set

$$\mathrm{Hess}_A := \{\mathcal{V} \in \mathrm{Flag}(\mathbb{R}^n) \mid AV_j \subseteq V_{j+1},\ j = 1, \dots, n-1\}.$$

Here $\mathrm{Flag}(\mathbb{R}^n)$ denotes the complete flag manifold (see Appendix F).

**Proposition 7.7** *Let $A \in \mathbb{R}^{n\times n}$ be invertible. The Hessenberg variety is a $\Sigma$-invariant subset of the inverse iteration system $\Sigma^{II}(A)$ on* $\mathrm{Flag}(\mathbb{R}^n)$*.*

**Proof.** Obviously, $AV_j \subseteq V_{j+1}$ implies $A(A-uI)V_j \subseteq (A-uI)V_{j+1}$ as well as $A(A-uI)^{-1}V_j \subseteq (A-uI)^{-1}V_{j+1}$. Therefore, $f_u(\mathrm{Hess}_A) = \mathrm{Hess}_A$ for all $u \in U$. □

By Proposition 3.10, $\mathrm{Hess}_A$ must be the union of system group orbits. I.e.,

$$\mathrm{Hess}_A := \bigcup_{i\in I} G_{\Sigma^{II}(A)} \cdot \mathcal{V}$$

for some $\mathcal{V}_i \in \mathrm{Flag}(\mathbb{R}^n)$, $i$ in an index set $I$. Moreover, we can restrict $\Sigma^{II}(A)$ to $\mathrm{Hess}_A$. We define the inverse iteration on $\mathrm{Hess}_A$ by

$$\Sigma^{Hess}(A) := \Sigma^{II}(A)|_{\mathrm{Hess}_A}.$$

Following Proposition 7.2 we obtain $G_{\Sigma^{Hess}(A)} \sim P(A)/\mathbb{R}^* I$, since $C_{\Sigma^{Hess}(A)} = \{\mathbb{R}^* \cdot I\}$. Therefore, $G_{\Sigma^{Hess}(A)} = S_{\Sigma^{Hess}(A)}$ if and only if $S(A)\mathbb{R}^* = P(A)$.

We have already seen, that none of the reachable sets of $\Sigma^{II}(A)$ on $\mathrm{Flag}(\mathbb{R}^n)$ is open or dense in $\mathrm{Flag}(\mathbb{R}^n)$, provided $n > 2$. The reason for that was, that the dimension of $\mathrm{Flag}(\mathbb{R}^n)$ is much larger then the dimension of possible group orbits. Using the system semigroup approach we show that there exist reachable sets of $\Sigma^{Hess}(A)$, which have open interior in $\Sigma^{Hess}(A)$. Moreover, $\Sigma^{Hess}(A)$ is densely reachable, provided $P(A) = S(A)\mathbb{R}^*$.

**Theorem 7.8** *Let $A \in \mathbb{R}^{n\times n}$ cyclic and invertible. Consider the inverse iteration system on* $\mathrm{Hess}_A$*.*

*a) There exists a system group orbit $\mathcal{N}_A^{Hess}$ which is open and dense in* $\mathrm{Hess}_A$*.*

*b) For all $x \in \mathcal{N}_A^{Hess}$ the reachable set of $x$ has nonempty interior in $\mathcal{N}_A^{Hess}$.*

*c) The following statements are equivalent.*

*(i)* $S(A)\mathbb{R}^* = P(A)$

*(ii) Orbit graph and reachable graph of $\Sigma^{Hess}(A)$ coincide.*

*(iii) $\Sigma^{Hess}(A)$ is approximatively reachable for some $x \in \mathrm{Hess}_A$.*

*(iv) $\Sigma^{Hess}(A)$ is densely reachable.*

**Proof.** a) Let $\mathcal{N}_A$ be defined as in Definition 6.9, i.e., the set of one dimensional spaces which are not included in any $A$-invariant subspace. Recall that $P(A) \cdot x = \mathcal{N}_A$ for all $x \in \mathcal{N}_A$. The projection

$$\pi : \mathrm{Flag}(\mathbb{R}^n) \to \mathbb{RP}^{n-1}, \ (U_1, U_2, \ldots, U_{n-1}) \mapsto U_1$$

is open and continuous. Thus, $\pi^{-1}(\mathcal{N}_A) \cap \mathrm{Hess}_A$ is open[24] in $\mathrm{Hess}_A$. Recall that all vectors $v \in \mathbb{R}^n$ with $\mathrm{span}(v) \in \mathcal{N}_A$ are cyclic. Thus,

$$\mathcal{K}_v := \big(\mathrm{span}(v), \mathrm{span}(v, Av), \ldots, \mathrm{span}(v, Av, \ldots, A^{n-1}v)\big) \in \mathrm{Hess}_A$$

for all $\mathrm{span}(v) \in \mathcal{N}_A$. We define $\mathcal{N}_A^{Hess} := \{\mathcal{K}_v \in \mathrm{Hess}_A \mid \mathrm{span}(v) \in \mathcal{N}_A\}$. Note that

$$\mathrm{Hess}_A = \mathcal{N}_A^{Hess} \cup \left( \bigcup_{W \in \mathrm{Inv}_A \setminus \{\mathbb{R}^n\}} \mathcal{N}^W \right)$$

with $\mathcal{N}_W := \{(U_1, \ldots, U_{n-1}) \in \mathrm{Hess}_A \mid U_j \subseteq W \text{ for some } j = 1, \ldots, n-1\}$. Clearly, $\dim \mathcal{N}_W < \dim \mathrm{Hess}_A$. Thus, since $\mathrm{Inv}_A$ is finite, $\mathcal{N}_A^{Hess}$ has open interior. Moreover, $\mathcal{N}_A^{Hess}$ is dense in $\mathrm{Hess}_A$ since $\pi$ is open and $\pi(\mathcal{N}_A^{Hess}) = \mathcal{N}_A$ is dense in $\mathbb{RP}^{n-1}$. Recall that $P(A)$ acts transitively on $\mathcal{N}_A$. Therefore, the group action

$$P(A) \times \mathcal{N}_A^{Hess} \to \mathcal{N}_A^{Hess}, (P(A), \mathcal{K}_v) \mapsto \mathcal{K}_{P(A)v}$$

is transitive. Thus

$$G_{\Sigma^{II}(A)} \cdot x = \{P(A) \cdot x \mid x \in \mathcal{N}_A^{Hess}\} = \mathcal{N}_A^{Hess}.$$

By Proposition 2.20, $\mathcal{N}_A^{Hess}$ is open. Hence, $\mathcal{N}_A^{Hess}$ is an open and dense group orbit in $\mathrm{Hess}_A$.
b) Recall that $S(A)\mathbb{R}^*$ has nonempty interior in $P(A)$ (see Corollary 6.6). Therefore, $\mathcal{R}(x) = S(A)\mathbb{R}^* \cdot x$ has nonempty interior in $P(A) \cdot x = \mathcal{N}_A^{Hess}$. Thus, $\mathrm{int}_{\mathrm{Hess}_A} \mathcal{R}(x) \neq \emptyset$.

[24] for the induced topology with respect to $\mathrm{Flag}(\mathbb{R}^n)$.

c) Clearly, $(i) \Rightarrow (ii)$, $(i) \Rightarrow (iv)$ and $(iv) \Rightarrow (iii)$. Assuming that $S(A)\mathbb{R}^* \neq P(A)$, we have $\mathrm{int}_{P(A)}(P(A) \setminus S(A)\mathbb{R}^*) \neq \emptyset$ (see Lemma B.6) and therefore $\mathrm{int}_{\mathrm{Hess}_A}(\mathrm{Hess}_A \setminus \mathcal{R}(x)) \neq \emptyset$. It follows, $\overline{\mathcal{R}(x)} \neq \mathrm{Hess}_A$ for all $x \in \mathrm{Hess}_A$. Thus (iii) implies (i). Moreover, $S(A)\mathbb{R}^* \neq P(A)$ implies that $\Sigma^{Hess}(A)$ is not weakly reversible. Thus, (ii) implies (i) by Theorem 4.6. □

In particular, Theorem 7.8 shows, that $\Sigma^{Hess}(A)$ has reachable sets which are dense in $\mathrm{Hess}_A$ if and only if the corresponding classical inverse iteration system $\Sigma^{II}(A)$ on $\mathbb{RP}^{n-1}$ has reachable sets which are open and dense in $\mathbb{RP}^{n-1}$. This fact has been pointed out earlier by Helmke and Jordan (see Theorem 5.1 in [HJ02]).

## 7.3 Inverse iteration on $\mathbb{R}^n$

We finish Section 7 with an analysis of inverse iteration systems on $M = \mathbb{R}^n$, i.e., $\Sigma^{II}(A) = (\mathbb{R}^n, U_A, f_A^{II})$ with respect to the canonical group action $\mathrm{GL}_n(\mathbb{R}) \times \mathbb{R}^n \to \mathbb{R}^n$. Note that here, $S_{\Sigma^{II}(A)} = S(A)$. Again we assume that $A$ is cyclic. Similar to classical inverse iteration systems there exists an open and dense $\Sigma$-invariant subset

$$N_A := \mathbb{R}^n \setminus \bigcup_{V \in \mathrm{Inv}_A} V.$$

(See Definition 6.9 and Proposition 6.10). The following result shows that $N_A$ is a system group orbit of $\Sigma^{II}(A)$ for all cyclic matrices. On the other hand it shows, that for an open set of matrices, $N_A$ is not a reachable set.

**Theorem 7.9** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic and $\Sigma^{II}(A) = (\mathbb{R}^n, U_A, f^{II})$ be the inverse iteration system on $\mathbb{R}^n \setminus \{0\}$ with respect to $A$.*

*a) $\Sigma^{II}(A)_{|N_A}$ is controllable if and only if $S(A) = P(A)$.*

*b) Let $n \geq 2$. There exists an open set of matrices $A \in \mathbb{R}^{n\times n}$, such that $S(A) \neq P(A)$. In particular this is the case if $A$ has a complex eigenvalue $\lambda$ with $\operatorname{Im}\lambda > 1$.*

**Proof.** a) Obviously, we have $S_{\Sigma^{II}(A)} = S(A)$. Recall that $G_{\Sigma^{II}(A)} := \langle S_{\Sigma^{II}(A)} \rangle = P(A)$ (see Theorem 6.3) and that $P(A)$ acts transitively on $N_A$ (see Lemma 6.11). Thus, $S(A) = P(A)$ implies controllability of $\Sigma^{II}(A)_{|N_A}$. Recall that $\mathrm{Stab}_x = \{I\}$ for all $x \in N_A$. Thus, $Bx = Cx$ with $B, C \in P(A)$ implies $B = C$. Hence, $S(A) \neq P(A)$ yields $\mathcal{R}(x) \subsetneq P(A)x$ for any $x \in N_A$.
b) We show, that for any $A \in \mathbb{R}^{n\times n}$ with complex eigenvalue $\lambda$, $\operatorname{Im}\lambda > 1$ the system semigroup $S(A)$ is not a group. Since $S(TAT^{-1} - vI) = TS(A)T^{-1}$ for $T \in \mathrm{GL}_n(\mathbb{R})$ and $v \in \mathbb{R}$ we may assume, that

$$A = \begin{pmatrix} A_1 & * \\ 0 & * \end{pmatrix} \text{ with } A_1 := \begin{pmatrix} 0 & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & 0 \end{pmatrix}.$$

If $S(A)$ is a group we have $\prod_{t=1}^N (A_1 - u_t I_2) = I_2$ for some $T \in \mathbb{N}$ and $u_1, \ldots, u_T \in U_A$. But this is a contradiction to $\operatorname{Im}\lambda > 1$, since

$$\det\left(\prod_{t=1}^N (A_1 - u_t I_2)\right) = \prod_{t=1}^N (u_t^2 + (\operatorname{Im}\lambda)^2) > 1 = \det I_2.$$

Thus, $S(A) \neq P(A)$. □

### 7.3.1 Inverse iteration in the plane

We finish this section with a complete analysis of system semigroups for inverse iteration systems on $\mathbb{R}^2$. We obtain the following semigroup types for $S(A)$.

**Theorem 7.10** *Let $A \in \mathbb{R}^{2\times 2}$ be cyclic.*

*a) If $A$ has two different real eigenvalues, then $S(A) = P(A) \cong (\mathbb{R}^*)^2$.*

*b) If $A$ has one real eigenvalue with multiplicity 2, then $S(A) = P(A) \cong \mathbb{R} \times \mathbb{R}^*$.*

*c) Assume, that $A$ has a pair of complex eigenvectors $\lambda, \overline{\lambda}$ such that $\operatorname{Im}\lambda \neq 0$.*

*(i) If $|\operatorname{Im}(\lambda)| < 1$, then $S(A) = P(A) \cong \mathbb{C}^*$.*

*(ii) If $|\operatorname{Im}(\lambda)| \geq 1$, then $S(A)$ is not a group.*

*(iii) If $|\operatorname{Im}(\lambda)| = 1$, then $S(A)$ is isomorphic to $\mathbb{D} \cup \{1, i, -1, -i\}$. Here, $\mathbb{D}$ denotes the open unit disc without zero in $\mathbb{C}^*$.*

**Proof.** Recall that $S(TAT^{-1} - vI) = TS(A)T^{-1}$ for all $T \in \mathrm{GL}_2(\mathbb{R})$ and

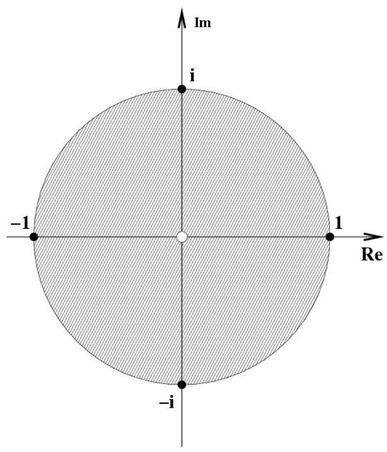



**Figure 4:** *The semigroup $S(A) \subseteq \mathbb{C}^*$ for the case $|\operatorname{Im}(\lambda)| = 1$*

$v \in \mathbb{R}$. Therefore, we can restrict our analysis on the cases

$$a)\ A = \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix}, \quad b)\ A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \text{ and } c)\ A = \begin{pmatrix} 0 & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & 0 \end{pmatrix}$$

with $\lambda \neq 0$ in case a), and respectively $\operatorname{Im}\lambda \neq 0$ in case c).

a) According to Theorem 6.8 we have $P(A) \cong (\mathbb{R}^*)^2$. Recall that $S(A)\mathbb{R}^* = P(A)$ (see Theorem 6.29). Moreover, for any $r \in (-\infty, 0)$ we have $rI \in S(A)$, since

$$(A - \alpha I)^{-1}(A - \beta I)^{-1} = rI$$

with $\alpha = \frac{1}{2}(\lambda + \sqrt{1-4r})$ and $\beta = \frac{x}{\alpha}$. Clearly $rI = (-\sqrt{r})I(-\sqrt{r}I) \in S(A)$ for $r > 0$. Thus, $\mathbb{R}^* I \subseteq S(A)$ and we conclude $S(A) = P(A)$.

b) Here $P(A) \cong \mathbb{R}^* \times \mathbb{R}$ (see Theorem 6.8). Again we have $S(A)\mathbb{R}^* = P(A)$ by Theorem 6.29. Moreover, for any $r \in (-\infty, 0)$ we have $rI \in S(A)$, since

$$\left(A - \frac{1}{\sqrt{-r}} I\right)^{-1} \left(A + \frac{1}{\sqrt{-r}} I\right)^{-1} = rI.$$

Clearly, $rI = (-\sqrt{r}I)(-\sqrt{r}I) \in S(A)$ for $r > 0$. Thus, $\mathbb{R}^* I \subseteq S(A)$ and we conclude $S(A) = P(A)$.

c) $S(A)$ is not a group, if $|\operatorname{Im}(\lambda)| > 1$ (see Theorem 7.9). Thus, we only have to show Claim (i) and Claim (iii). We can identify $A - uI$ with the complex number $-u + \beta i$ with $\beta := \operatorname{Im}\lambda$. Note that the multiplication of matrices $A - u_1 I$, $A - u_2 I$ coincides with the multiplication in $\mathbb{C}^*$. In other words, $S(A)$ can be regarded as a subsemigroup in $\mathbb{C}^*$. Using polar coordinates every element $\prod_{t=0}^{N}(-u_t + i\beta) \in (S(A))^{-1}$ can be written in the form

$$x = \prod_{t=0}^{N} \left( \frac{\beta e^{i\alpha_t}}{\sin \alpha_t} \right), \quad \text{with } \tan \alpha_t = \frac{\beta}{-u_t}, \ \alpha_t \in (0, \pi).$$

For every $N \in \mathbb{N}$ we define $I_N := [\frac{\pi}{2} - \frac{\pi}{2N+2}, \frac{\pi}{2} + \frac{\pi}{2N+2}]$ and

$$\gamma_N : I_N \to (S(A))^{-1}, \quad \alpha \mapsto \prod_{t=0}^{4+4N} (\frac{\beta e^{i\alpha}}{\sin \alpha}).$$

$\gamma_N$ is a closed curve in $\mathbb{C}^*$ which is symmetric with respect to the real axis (i.e. $\gamma_N(\pi/2 - \alpha) = \gamma_N(\pi/2 + \alpha)$). For all $x \in \gamma_N(I_N)$ it is

$$\beta^{4+4N} \leq x \leq \beta^{4+4N} \frac{1}{\sin(\frac{\pi}{2} - \frac{\pi}{2N+2})^{4+4N}}.$$

Moreover,

$$\begin{aligned}
\sin(\frac{\pi}{2} - \frac{\pi}{2N+2}) &= \sin(\frac{\pi}{2}) \cos(\frac{\pi}{2N+2}) - \cos(\frac{\pi}{2}) \sin(\frac{\pi}{2N+2}) \\
&= \cos(\frac{\pi}{2+2N}).
\end{aligned}$$

Thus, the sequence of closed curves $\gamma_N(I_N) \subseteq (S(A))^{-1}$ converges uniformly to $\mathbb{S}$ in the case $\beta = 1$ respectively to $\{0\}$ in the case $\beta < 1$.

Now let $x = ae^{i\alpha} \in \mathbb{C}^*$ such that $\varepsilon e^{i(\alpha+\pi)} \in (S(A))^{-1}$ for some $\varepsilon < a/\beta^2$. This is possible for all $x \in \mathbb{C}^*$ if $0 < \beta < 1$ and for all $x \in \mathbb{C} \setminus \overline{\mathbb{D}}$ if $\beta = 1$. We show that $x \in (S(A))^{-1}$. Choose $\alpha_1 = -\alpha_2$ such that $\varepsilon\beta^2/\sin(\alpha_1)^2 = a$. Then it is

$$\underbrace{\varepsilon e^{i(\alpha+\pi)}}_{(S(A))^{-1}} \underbrace{\frac{\beta e^{i\alpha_1}}{\sin\alpha_1}}_{(S(A))^{-1}} \underbrace{\frac{\beta e^{i\alpha_2}}{\sin\alpha_2}}_{(S(A))^{-1}} = \frac{\varepsilon\beta^2 e^{i(\alpha+\pi)}}{-\sin^2\alpha_1} = \frac{\varepsilon\beta^2 e^{i\alpha}}{\sin^2\alpha_1} = x.$$

This implies Claim (i). Moreover, we can conclude $(S(A))^{-1} \subseteq \mathbb{C} \setminus \overline{\mathbb{D}}$ for $\beta = 1$.

For any $\beta \geq 1$ we can estimate the norm of an arbitrary element $x \in (S(A))^{-1}$ by

$$|x| = \left| \prod_{t=0}^{N} \frac{\beta e^{i\alpha_t}}{\sin\alpha_t} \right| \geq |\beta^N|.$$

For $\beta = 1$ it follows $|x| \geq 1$. Moreover, it is $|x| = 1$ if and only if $x = \prod_{t=0}^{N} i$. We deduce, $(S(A))^{-1} = (\mathbb{C} \setminus \overline{\mathbb{D}}) \cup \{i, -1, -i, 1\}$ which yields Claim (iii). □

In the proof of Theorem 7.10 we have shown a technical result for subsemigroups of $\mathbb{C}^*$ which will be important in Section 9.

**Corollary 7.11** *Let $M_\beta := \{i\beta - u \,|\, u \in \mathbb{R}\}$. The set of finite products of elements of $M_\beta$ is $\mathbb{C}^*$ if $0 < \beta < 1$ and $(\mathbb{C}^* \setminus \overline{\mathbb{D}}) \cup \{1, i, -1, -i\}$ if $\beta = 1$.*

We finish this section with a remark on the case $\operatorname{Im}\lambda > 1$. Note that here $(S(A))^{-1}$ corresponds to the system semigroup of Example 2.9. In this case, $S(A)$ is neither isomorphic to $\mathbb{C}^*$ nor to $\mathbb{D} \cup \{1, i, -1, -i\}$.

**Proposition 7.12** *Let $A \in \mathbb{R}^{2\times 2}$ with a pair of complex eigenvectors $\lambda, \overline{\lambda}$ such that $\operatorname{Im}\lambda \neq 0$. If $|\operatorname{Im}\lambda| > 1$ then $P(A) \setminus S(A)^{-1}$ has at least two connected components.*

**Proof.** We construct a closed loop in $(S(A))^{-1}$ which separates two subsets of $P(A) \setminus (S(A))^{-1}$. Since the inversion map $\mathbb{C}^* \to \mathbb{C}^*$, $z \mapsto z^{-1}$ is a homeomorphism, $P(A) \setminus S(A)$ has at least two components.

Recall that $P(A) \cong \mathbb{C}^*$. The line $l(u) := -u + i\beta$, $u \in \mathbb{R}$ describes the set of points in $S(A)$ which are generated by one factor. Every element generated by more the one factor has a norm larger or equal to $\beta^2$. We construct a connected curve $\gamma : \mathbb{R} \to (S(A))^{-1}$ which intersects the line $l$

on the left and on the right half plane, but has the property $|\gamma(u)| > \beta^3$ for all $u \in \mathbb{R}$.

Consider, $\gamma : u \mapsto (-u - \beta i)^3$. On the one hand,

$$\gamma(u)^3 = \frac{\beta^3 e^{i3\alpha}}{\sin^3 \alpha},\ \alpha \in (0, \pi)$$

shows, that $|\gamma(u)| \geq \beta^3$ and $\mathrm{Im}(\gamma(u)) > \beta$ for $\tan \alpha = \frac{\beta}{-u}$. On the other hand

$$\mathrm{Im}(\gamma(u)) = \mathrm{Im}\left((-u - \beta i)^3\right) = -u^2\beta - \beta^3 + 2\beta u$$

shows, that $\mathrm{Im}(\lambda) < \beta$ for $|u|$ large enough. We conclude, that $l$ and $\gamma$ intersect in the left and in the right complex plane. In particular, they separate the sets

$$M_1 := \{z \in \mathbb{C}^* \mid z \in i\mathbb{R}, \beta < \mathrm{Im}(z) < \beta^2\}$$

and

$$M_2 := \{z \in \mathbb{C}^* \mid |z| < \beta^2, \mathrm{Im}(z) < \beta\}.$$

Thus, $P(A) \setminus (S(A))^{-1}$ has at least two connected components. $\square$

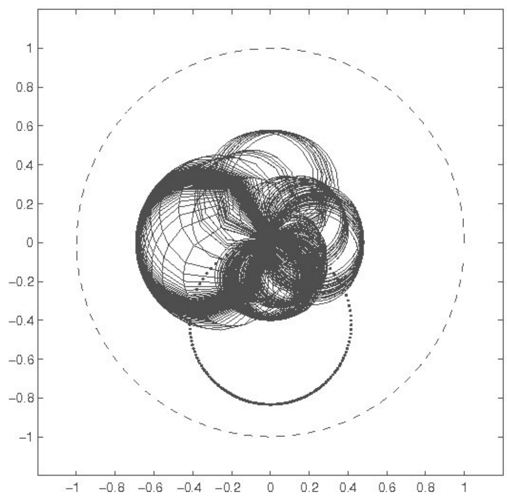


**Figure 5:** *$S(A) \subseteq \mathbb{C}$ for $\beta = 1.2$. In fact the plot shows products of order 1,2,3 and 4 with elements in $\{(-u + i\beta)^{-1} \mid u \in \mathbb{R}\}$. Every element of $S(A)$ lies inside the circle $\{z \in \mathbb{C} \mid |z| = 1\}$. Moreover, $\mathbb{C} \setminus S(A)$ has at least two connected components.*

# 8 Rational iteration

In the previous sections we have seen, that the system semigroup of inverse iteration is not necessarily a group. This situation yields undesired constraints on the convergence behavior of possible shift strategies. To avoid this phenomenon it is advisable to create alternative schemes, such that the reachable sets become easier to investigate. *Rational iteration* is an extension from inverse iteration, using a second shift parameter. Here the system semigroups are always groups. Rational iteration schemes have been applied in the field of eigenvalue computation as well as linear equation solving (see [Ros94, JV05], and respectively, [YV92]). To the authors knowledge, there exists no systematic investigation on the adherence structure of reachable sets of rational iteration systems. This will be the topic of the following section. First we analyze the general setting of rational iteration systems on manifolds (Section 8). Then, in Section 8.2, we consider a one-parameter version of rational iteration called *Cayley iteration*.

## 8.1 Rational iteration systems

**Definition 8.1 (Rational iteration system)** Let $\mathrm{GL}_n(\mathbb{R}) \times M \to M$ be a transitive group action on a manifold $M$. Given $A \in \mathbb{R}^{n\times n}$, we define

$$U_A^{RI} := (\mathbb{R} \setminus \mathrm{Spec}(A))^2 \quad \text{and} \quad f_A^{RI}(x,(u,v)) := (A-uI)^{-1}(A-vI)\cdot x.$$

We call the corresponding system $\Sigma^{RI}(A) := (M, U_A^{RI}, f_A^{RI})$ the *Rational iteration system* of $A$ with respect of the group action $\mathrm{GL}_n(\mathbb{R}) \times M \to M$.

Note that the corresponding system semigroup $S_{\Sigma^{RI}(A)}$ is a group for any matrix $A \in \mathbb{R}^{n\times n}$. More precisely we obtain:

**Proposition 8.2** *Let $A \in \mathbb{R}^{n\times n}$, $m_A$ be the minimal polynomial of $A$ and $C_M := \bigcap_{x\in M} \mathrm{Stab}_x$. The system semigroup of $\Sigma^{RI}(A)$ is a group isomorphic to $P(A)/(P(A)\cap C_M)$.*

**Proof.** $S_{\Sigma^{RI}(A)}$ is a group, since the inverse of

$$s : x \mapsto \prod_{t=1}^{T}(A-u_tI)(A-v_tI)^{-1}\cdot x$$

is given by $x \mapsto \prod_{t=1}^{T}(A-v_tI)(A-u_tI)^{-1}\cdot x$ and therefore an element of $S_{\Sigma^{RI}(A)}$. Recall that

$$\left\{ \prod_{t=1}^{T}(A-u_tI)(A-v_tI)^{-1} \,\middle|\, T\in\mathbb{N}, (u_t,v_t)\in U_A \right\} = P(A)$$

(see Corollary 6.5). Two matrices $B, \tilde{B} \in P(A)$ induce the same maps $x \mapsto B \cdot x$, respectively $x \mapsto \tilde{B} \cdot x$ if and only if $B\tilde{B}^{-1}$ is an element of $\mathrm{Stab}_x$ for all $x \in M$. Therefore, the kernel of the group homomorphism $\Phi : P(A) \to S_{\Sigma^{RI}(A)}$, $\Phi(B) : x \mapsto B \cdot x$ is $P(A) \cap C_M$. □

In particular we are interested in the case when $M_1 = \mathbb{R}^n$, $M_2 = \mathbb{RP}^{n-1}$, $M_3 = \mathrm{Hess}_A(\mathbb{R}^n)$ and $M_4 = \mathrm{Flag}(\mathbb{R}^n)$, each case with respect to the corresponding canonical group action $\alpha_i : \mathrm{GL}_n(\mathbb{R}) \times M_i \to M_i$, $i = 1, 2, 3, 4$. From our analysis of inverse iteration systems we easily deduce the following results:

**Theorem 8.3** *Let $M$ be a topological space, $\alpha : \mathrm{GL}_n(\mathbb{R}) \times M \to M$ be a transitive group action and $\Sigma^{RI}(A) = (M, U_A^{RI}, f_A)$ be the rational iteration system of $A \in \mathbb{R}^{n\times n}$ with respect to $\alpha$.*

*a) The orbit graph $\mathcal{G}_O(\Sigma^{RI}(A))$ and the reachable graph $\mathcal{G}_R(\Sigma^{RI}(A))$ coincide. In particular, $\Sigma^{RI}(A)$ is weakly reversible.*

*b) Let $\alpha_i : \mathrm{GL}_n(\mathbb{R}) \times M_i \to M_i$, $i = 1, 2, 3$ be the canonical group action on $M_i$ with $M_1 = \mathbb{R}^n$, $M_2 = \mathbb{RP}^{n-1}$, $M_3 = \mathrm{Hess}_A(\mathbb{R}^n)$ and $\Sigma_i^{RI}(A) = (M_i, U_A^{RI}, f_A^{RI})$ the rational iteration system of $A \in \mathbb{R}^{n\times n}$ on $M_i$.*

*(i) If $A$ is cyclic, then $N_i$ with $N_1 = N_A$, $N_2 = \mathcal{N}_A$, $N_3 = \mathcal{N}_A^{Hess}$ coincides with one reachable set, which is open and dense in $M_i$. Here $N_A$ and $\mathcal{N}_A$ are defined as in Definition 6.9 and $\mathcal{N}_A^{Hess}$ is defined as in Section 7.2. Moreover, the restricted system $\Sigma^{RI}(A)_{|N_i}$ is controllable.*

*(ii) If $A$ is not cyclic, then none of the reachable sets has open interior in $N_i$.*

*c) Let $\alpha_4 : \mathrm{GL}_n(\mathbb{R}) \times \mathrm{Flag}(\mathbb{R}^n) \to \mathrm{Flag}(\mathbb{R}^n)$, be the canonical group action on $\mathrm{Flag}(\mathbb{R}^n)$. Then any class $[\mathcal{V}]$, $\mathcal{V} \in \mathrm{Flag}(d, \mathbb{R}^n)$ (as defined in Definition 7.4) is the disjoint union of reachable sets.*

**Proof.** a) Since $S_{\Sigma^{RI}(A)}$ is a group, $\Sigma^{RI}(A)$ is weakly reversible by Lemma 2.35. Thus, the claim follows by Theorem 4.6.
b) and c) The reachable sets of $\Sigma^{RI}(A)$ coincide to the system group orbits of the corresponding inverse iteration system. Thus, all claims in b) are immediate consequences of Lemma 6.11 and Theorem 7.8. Moreover, claim c) follows from Theorem 7.5. □

## 8.2 Cayley iteration

As a special case of rational iteration we consider systems generated by Cayley transformations, $x \mapsto (A - uI)(A + uI)^{-1} \cdot x$. Cayley iteration steps have been proposed by several authors (see for example [MSR94] and [LM98]). If $A$ is element of a classical Lie algebra, all states of Cayley iteration remain in the corresponding Lie-group. This fact yields interesting relations for the eigenvalue computation for specific matrices.

**Definition 8.4 (Cayley iteration system)** Let $\alpha : \mathrm{GL}_n(\mathbb{R}) \times M \to M$ be a transitive group action on a manifold $M$. Given a matrix $A \in \mathbb{R}^{n\times n}$, we define

$$U_A := \mathbb{R} \setminus \pm\operatorname{Spec}(A) \quad \text{and} \quad f^{CI}(x,u) := (A - uI)(A + uI)^{-1} \cdot x.$$

We call the corresponding system $\Sigma^{CI}(A) := (M, U_A, f^{CI})$ the *Cayley iteration system* of $A$ with respect of $\alpha$.

Again, the system semigroup is a group. Therefore, $\Sigma^{CI}(A) := (M, U_A^{CI}, f^{CI})$ is always weakly reversible. Cayley iteration systems can be considered as rational iteration with a restriction on the allowed shift strategies, i.e., $v_t = -u_t$. Therefore, the system semigroup $S_{\Sigma^{CI}(A)}$ is a subgroup of $S_{\Sigma^{RI}(A)}$ (see Proposition 8.2).

### 8.2.1 Conditions for $S_{\Sigma^{CI}(A)} = P(A)$

We restrict our analysis to the case where $P(A) \cap C_M$ is trivial [25]. In this situation we have $S_{\Sigma^{CI}(A)} \subseteq S_{\Sigma^{RI}(A)} = P(A)$. In fact, for some but not for all matrices $A \in \mathbb{R}^{n\times n}$, it holds that $S_{\Sigma^{CI}(A)} = P(A)$. In the following we show a condition on $A$ for the property $S_{\Sigma^{CI}(A)} = P(A)$.

**Theorem 8.5** *Let $A \in \mathbb{R}^{n\times n}$ be invertible with $n$ different real eigenvalues $\lambda_1, \dots, \lambda_n$ such that $|\lambda_i| \neq |\lambda_j|$ for $i \neq j$. Then, $S_{\Sigma^{CI}(A)} = P(A)$.*

**Proof.** Recall that the topological closure[26] of $S_{\Sigma^{CI}(A)}$ is a closed subgroup of the Lie group $P(A) = \{\operatorname{diag}(a_1, \dots, a_n) \mid a_k \in \mathbb{R}^*\}$ (See Theorem 6.8) and therefore a Lie group. We show the following two claims:

**Claim 1:** $e \in \operatorname{int}_{P(A)} S_{\Sigma^{CI}(A)}$;

**Claim 2:** $S_{\Sigma^{CI}(A)}$ has nonempty intersection with any connected component of $P(A)$.

Then, by Theorem 5.4 it follows $S_{\Sigma^{CI}(A)} = P(A)$.

[25] In particular this is the case if $M = \mathbb{R}^n$ or if $M = \mathrm{GL}_n(\mathbb{R})$ (and $\alpha$ the corresponding canonical group action on $M$).

[26] with respect to $P(A)$

**Proof of Claim 1:** Without loss of generality we assume, that $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. We show that the map

$$\Phi : (U_A^{CI})^n \to S_{\Sigma^{CI}(A)},\ u \mapsto \operatorname{diag}(f_1(u), \ldots, f_n(u)) \subseteq P(A)$$

with $f_k(u) = f_k((u_1, \ldots, u_n)) = \prod_{j=1}^n \frac{\lambda_k - u_j}{\lambda_k + u_j}$ is locally invertible, if and only if $u_i \neq u_j$ for $i \neq j$. For the Jacobian $D\Phi$ of $\Phi$ we obtain

$$\begin{aligned} D\Phi(u) &= \left( f_k(u) \frac{-2\lambda_k}{\lambda_k^2 - u_j^2} \right)_{k,j=1,\ldots,n} \\ &= \operatorname{diag}(-2\lambda_1 f_1(u), \ldots, -2\lambda_n f_n(u)) \left( \frac{1}{\lambda_k^2 - u_j^2} \right)_{k,j=1,\ldots,n} \end{aligned}$$

The *Cauchy determinant rule* (see [Fuh96], Section 3.4) yields

$$\det \left( \left( \frac{1}{\lambda_k^2 - u_j^2} \right)_{k,j=1,\ldots,n} \right) = \frac{\prod_{k>j} (\lambda_k^2 - \lambda_j^2)(u_k^2 - u_j^2)}{\prod_{k,j} (\lambda_k^2 + u_j^2)}$$

This shows, that

$$\det(D\Phi(u_1, \ldots, u_n)) = (-2)^n \prod_{k=1}^n (\lambda_k f_k(u)) \det \left( \left( \frac{1}{\lambda_k^2 - u_j^2} \right)_{k,j=1,\ldots,n} \right) \neq 0$$

provided $u_i \neq u_j$. From the inverse function theorem it follows, that $\Phi$ is locally invertible. Hence, $\operatorname{int}_{P(A)} S_{\Sigma^{CI}(A)} \neq \emptyset$. Moreover, for any $s \in \operatorname{int}_{P(A)} S_{\Sigma^{CI}(A)}$ we have $s^{-1}s \in \operatorname{int}_{P(A)} S_{\Sigma^{CI}(A)}$ (see Lemma B.5). We conclude

$$e \subseteq \operatorname{int}_{P(A)} S_{\Sigma^{CI}(A)}.$$

**Proof of Claim 2:** Without loss of generality we assume, that $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $0 < |\lambda_1| < \ldots, |\lambda_n|$. Obviously,

$$P(A) = \{\operatorname{diag}(a_1, \ldots, a_n) \mid a_k \in \mathbb{R}^*\}$$

has $2^n$ connected components, which can be identified with the sign vectors $(\operatorname{sign}(a_1), \ldots, \operatorname{sign}(a_n)) \in \{-1, 1\}^n$. We show, that for any sign vector $(\epsilon_1, \ldots, \epsilon_n) \in \{-1, 1\}^n$ there exists $\operatorname{diag}(b_1, \ldots, b_n) \in S_{\Sigma^{CI}(A)}$ such that $\operatorname{sign}(b_k) = e_k$ for any $k = 1, \ldots, n$. Note that

$$\frac{\lambda_k - u}{\lambda_k + u} = \frac{1 - \frac{u}{\lambda_k}}{1 - \frac{u}{\lambda_k}} > 0.$$

if and only if $u < |\lambda_k|$. Therefore, for $u \in [|\lambda_k|, |\lambda_{k+1}|]$ we obtain

$$(A - uI)(A + uI)^{-1} = \operatorname{diag}(\underbrace{b_1, \ldots, b_k}_{<0}, \underbrace{b_{k+1}, \ldots, b_n}_{>0}).$$

Those matrices already generate matrices $\operatorname{diag}(b, \ldots, b_n)$ for any combination $\operatorname{sign} b_k \in \{-1, 1\}$, $k = 1, \ldots, n$. $\square$

In particular, Theorem 8.5 shows, that there exists an open set in $\mathbb{R}^{n\times n}$ such that $S_{\Sigma^{CI}(A)} = P(A)$. Now we show some conditions on $A \in \mathbb{R}^{n\times n}$ such that $S_{\Sigma^{CI}(A)} \neq P(A)$. We will use the following fact:

**Lemma 8.6** *Let $Z \in \mathbb{R}^{n\times n}$. If $A$ is an element of*

$$\mathfrak{g}_Z := \{B \in \mathbb{R}^{n\times n} \,|\, B^\top Z + ZB = 0\}$$

*then $S_{\Sigma^{CI}(A)}$ is an abelian subgroup of the group*

$$\mathrm{G}_Z := \{B \in \mathrm{GL}_n(\mathbb{R}) \,|\, B^\top ZB = Z\}.$$

**Proof.** Let $A$ be en element of $\mathfrak{g}_Z$, i.e. $A^\top Z = -ZA$. Straightforward calculation yields

$$\begin{aligned} \left((A-u_tI)(A+u_tI)^{-1}\right)^\top Z\left((A-u_tI)(A+u_tI)^{-1}\right) &= \\ (A+uI)^{-\top}\left(Z(u^2I-A^2)\right)(A+uI)^{-1} &= \\ (A+uI)^{-\top}Z(uI-A) &= \\ (A+uI)^{-\top}(uI+A^\top)Z &= Z. \end{aligned}$$

Therefore, $(A-u_tI)^{-1}(A+u_tI) \in \mathrm{G}_Z$ for every $u \in U_A$. The claim follows, since every $B \in S_{\Sigma^{CI}(A)}$ is a product of matrices of type $(A-u_tI)^{-1}(A+u_tI)$. □

Note that $\mathrm{G}_Z$ is a Lie group and $\mathfrak{g}_Z$ is the Lie algebra of $\mathrm{G}_Z$. In particular, the choice $Z = I$ yields the *orthogonal group* $\mathrm{O}_n(\mathbb{R})$ and the algebra of *skew-symmetric matrices* $\mathfrak{so}_n(\mathbb{R})$. Moreover, if $n$ is even, the choice $Z = J$ with

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

yields the *symplectic group* $\mathrm{Sp}_n(\mathbb{R})$ and the algebra of *Hamiltonian matrices* $\mathfrak{sp}_n(\mathbb{R})$.

**Theorem 8.7** *Let $A \in \mathbb{R}^{n\times n}$ and $\Sigma^{CI}(A) := (M, U_A, f^{CI})$ be the corresponding Cayley iteration system.*

*a) If $0 \in \mathrm{Spec}(A)$, then $S_{\Sigma^{CI}(A)} \neq P(A)$.*

*b) If $\lambda, -\lambda \in \mathrm{Spec}(A) \cap \mathbb{R}$, then $S_{\Sigma^{CI}(A)} \neq P(A)$.*

*c) If $A$ is skew-symmetric, then $S_{\Sigma^{CI}(A)} \neq P(A)$.*

*d) If $n$ is even and $A$ is Hamiltonian, then $S_{\Sigma^{CI}(A)} \neq P(A)$.*

**Proof.** a) Obviously, there exists $B \in P(A)$ such that neither 1 nor $-1$ is an eigenvalue of $B$. We show that every element of $S_{\Sigma^{CI}(A)}$ has eigenvalue 1 or $-1$. For any $B = \prod_{t=1}^{T}(A - u_t I)(A + u_t I)^{-1} \in S_{\Sigma^{CI}(A)}$ we obtain

$$\begin{aligned} B + (-1)^T I &= \left(\prod_{t=1}^{T}(A - u_t I) + (-1)^T \prod_{t=1}^{T}(A + u_t I)\right) \prod_{t=1}^{T}(A + u_t I)^{-1} \\ &= \prod_{t=1}^{T}(A + u_t I)^{-1} A p(A) \end{aligned}$$

for some $p \in \mathbb{R}[x]$. Since $\det(A) = 0$ it follows $\det(B + (-1)^T I) = 0$. Hence, $B$ has eigenvalue 1 or $-1$.
b) Without loss of generality we may assume

$$A = \begin{pmatrix} A_1 & * \\ 0 & * \end{pmatrix}, \quad \text{with} \quad A_1 = \begin{pmatrix} \lambda & 0 \\ 0 & -\lambda \end{pmatrix}.$$

For any $B = \prod_{t=1}^{T}(A - u_t I)(A + u_t I)^{-1} \in S_{\Sigma^{CI}(A)}$ we obtain

$$B = \begin{pmatrix} B_1 & * \\ 0 & * \end{pmatrix} \quad \text{with} \quad B_1 = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$$

such that $\alpha = \prod_{t=1}^{T}(\lambda - u_t I)(\lambda + u_t I)^{-1}$ and $\beta = \prod_{t=1}^{T}(-\lambda - u_t I)(\lambda + u_t I)^{-1}$. Thus $\frac{\alpha}{\beta} = (-1)^T$.

On the other hand, by the Lagrangian interpolation theorem, for any $\alpha, \beta \in \mathbb{R}^*$ there exists $p(A) \in P(A)$ such that

$$p(A) = \begin{pmatrix} p(A_1) & * \\ 0 & * \end{pmatrix} \quad \text{with} \quad p(A_1) = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}.$$

We conclude $S_{\Sigma^{CI}(A)} \neq P(A)$.
c) If $n = 1$ then $S_{\Sigma^{CI}(A)} \neq P(A)$ by a). Recall that $P(A)$ is an unbounded subset of $\mathrm{GL}_n(\mathbb{R})$ (see Theorem 6.8). If $A$ is skew-symmetric, $S_{\Sigma^{CI}(A)}$ is a subgroup of the compact group $\mathrm{O}_n(\mathbb{R})$ (see Lemma 8.6). In particular $S_{\Sigma^{CI}(A)}$ is bounded. Hence, $S_{\Sigma^{CI}(A)} \neq P(A)$.
d) If $A$ is Hamiltonian, then $S_{\Sigma^{CI}(A)} \subseteq \mathrm{Sp}_n(\mathbb{R})$ by Lemma 8.6. In particular, the determinant of any element in $S_{\Sigma^{CI}(A)}$ is 1. Hence, $S_{\Sigma^{CI}(A)} \neq P(A)$. □

### 8.2.2 Cayley iteration on the plane

Now we focus on Cayley iteration systems on $\mathbb{R}^n$ with respect to the canonical action on $\mathbb{R}^n$. Note that $N_A = \mathbb{R}^n \setminus \bigcup_{V \in \mathrm{Inv}_A^{\mathbb{R}^n}} V$ is a $\Sigma$-invariant subset of $\mathbb{R}^n$. Recall that $P(A)$ acts transitively on $N_A$ and that $\mathrm{Stab}_x = \{I\}$ for any $x \in N_A$ (see Lemma 6.11). Thus, any subgroup $G$ of $P(A)$ acts transitively on $N_A$ if and only if $G = P(A)$. Hence, $\Sigma^{CI}(A)_{|\mathcal{N}_A}$ is controllable if and only if $S_{\Sigma^{CI}(A)} = P(A)$. In the following we classify all cyclic matrices $A \in \mathbb{R}^{2\times 2}$ with $S_{\Sigma^{CI}(A)} = P(A)$.

**Theorem 8.8** *Let $A \in \mathbb{R}^{2\times 2}$ be cyclic.*

*a) Assume that $A$ is real diagonalizable with eigenvalues $\lambda_1, \lambda_2 \in \mathbb{R}$. Then $S_{\Sigma^{CI}(A)} = P(A)$ if and only if $\lambda_1, \lambda_2 \neq 0$ and $|\lambda_1| \neq |\lambda_2|$.*

*b) Assume that $A$ has a real eigenvalue $\lambda$ with multiplicity two. Then $S_{\Sigma^{CI}(A)} = P(A)$ if and only if $\lambda \neq 0$.*

*c) Assume that $A$ has a pair of complex eigenvalues $\lambda, \overline{\lambda}$ ($\operatorname{Im}\lambda \neq 0$). Then $S_{\Sigma^{CI}(A)} = P(A)$ if and only if $\operatorname{Re}\lambda \neq 0$.*

**Proof.** Recall that $S_{\Sigma^{CI}(TAT^{-1})} = TS_{\Sigma^{CI}(A)}T^{-1}$ for $T \in \mathrm{GL}_n(\mathbb{R})$. Thus we can assume, that $A$ is in Jordan canonical form.

a) (i) If

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \text{with} \quad \lambda_1 \neq 0,\ \lambda_2 \neq 0,\ |\lambda_1| \neq |\lambda_2|,$$

then $S_{\Sigma^{CI}(A)} = P(A) = \{\operatorname{diag}(a_1, a_2) \,|\, a_1, a_2 \in \mathbb{R}^*\}$ by Theorem 8.5.
(ii) If

$$A = \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix} \quad \text{with} \quad \lambda \neq 0,$$

then $S_{\Sigma^{CI}(A)} \subsetneq P(A)$ by Theorem 8.7. More precisely we have

$$S_{\Sigma^{CI}(A)} = \left\{ \begin{pmatrix} \epsilon & 0 \\ 0 & a \end{pmatrix} \,\middle|\, \epsilon \in \{-1, 1\}, a \in \mathbb{R}^* \right\}.$$

If

$$A = \begin{pmatrix} \lambda & 0 \\ 0 & -\lambda \end{pmatrix} \quad \text{with} \quad \lambda \neq 0,$$

then $S_{\Sigma^{CI}(A)} \subsetneq P(A)$ by Theorem 8.7 and Lemma 8.6. We obtain

$$S_{\Sigma^{CI}(A)} = \left\{ \begin{pmatrix} a & 0 \\ 0 & \epsilon a \end{pmatrix} \,\middle|\, \epsilon \in \{-1, 1\}, a \in \mathbb{R}^*, \right\}.$$

b) If $A$ has a real eigenvalue of multiplicity two, $P(A)$ is given by

$$P(A) := \left\{ \begin{pmatrix} a & b \\ 0 & a \end{pmatrix} \,\middle|\, a \in \mathbb{R}^*, b \in \mathbb{R} \right\}.$$

(see Section 6.2). In particular, $P(A)$ is an abelian Lie group with two connected components.
(i) Assume that

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

By Theorem 8.7 it holds that $S_{\Sigma^{CI}(A)} \subsetneq P(A)$. More precisely we obtain

$$(A - uI)(A + uI)^{-1} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

for any $u \in U_A^{CI}$. Thus, $S_{\Sigma^{CI}(A)} = \{-I, I\}$.

(ii) Now we assume that

$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \quad \text{with} \quad \lambda \neq 0.$$

Here the group $S_{\Sigma^{CI}(A)}$ is generated by the matrices

$$A_u := (A - uI)(A + uI)^{-1} = \frac{\lambda - u}{\lambda + u} \begin{pmatrix} 1 & \frac{2u}{(\lambda+u)(\lambda-u)} \\ 0 & 1 \end{pmatrix}$$

with $u \in \mathbb{R} \setminus \{-\lambda, \lambda\}$. Clearly, the dimension of $S_{\Sigma^{CI}(A)}$ is larger then two. Moreover, $S_{\Sigma^{CI}(A)}$ has nonempty intersection with both components of $P(A)$. By $S_{\Sigma^{CI}(A)} = P(A)$.
c) If $A$ has a real eigenvalue of multiplicity two, $P(A)$ is given by

$$P(A) := \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \,\middle|\, a^2 + b^2 \neq 0 \right\}.$$

(see Section 6.2).
(i) Assume

$$A = \begin{pmatrix} 0 & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & 0 \end{pmatrix} \quad \text{with} \quad \operatorname{Im}\lambda \neq 0.$$

By Theorem 8.7 and Lemma 8.6 we have $S_{\Sigma^{CI}(A)} \subseteq \mathrm{O}_n(\mathbb{R}) \subsetneq P(A)$ and therefore $S_{\Sigma^{CI}(A)} \neq P(A)$.

(ii) Now we assume that $\operatorname{Re}\lambda \neq 0$. The dimension of $S_{\Sigma^{CI}(A)}$ is larger then 2. Thus, $S_{\Sigma^{CI}(A)}$ coincides with the connected Lie group $P(A)$. □

Recall, that any Cayley iteration system is weakly reversible (even if $S_{\Sigma^{CI}(A)} \neq P(A)$). Thus, the reachable sets always form a partition on $\mathbb{R}^2$. As an immediate consequence of the previous proof, we obtain the adherence structure of the reachable sets.

**Corollary 8.9** *Let $A \in \mathbb{R}^{2\times 2}$ be cyclic.*

*a) Assume that* $A = \operatorname{diag}(\lambda_1, \lambda_2)$.

*(i) If* $\lambda_1, \lambda_2 \neq 0$ *and* $|\lambda_1| \neq |\lambda_2|$ *then the reachable graph is given by*

$$\begin{tikzcd}
& \mathcal{R}((1,1)^\top) \arrow[dl] \arrow[dr] & \\
\mathcal{R}((1,0)^\top) \arrow[dr] & & \mathcal{R}((0,1))^\top \arrow[dl] \\
& \mathcal{R}((0,0))^\top &
\end{tikzcd}$$

*(ii) If* $\lambda_1 = 0$ *and* $\lambda_2 \neq 0$ *then we have infinitely many reachable sets. In particular we obtain*

$$\begin{array}{rcl}
\mathcal{R}((0,0)^\top) & = & (0,0)^\top, \\
\mathcal{R}((x,0)^\top) & = & \{(-x,0)^\top, (x,0)^\top\}, \\
\mathcal{R}((0,y)^\top) & = & \{(0,r)^\top, |r \in \mathbb{R}^*\} \;\; \textit{for } y \in \mathbb{R}^*, \\
\mathcal{R}((x,y)^\top) & = & \{(\epsilon x, r)^\top, |\epsilon \in \{-1,1\}, r \in \mathbb{R}^*\} \;\; \textit{for } (x,y) \in (\mathbb{R}^*)^2.
\end{array}$$

*(iii) If* $\lambda_1 \neq 0$ *and* $\lambda_2 = -\lambda_1$*, then we have infinitely many reachable sets. In particular we obtain*

$$\begin{array}{rcl}
\mathcal{R}((0,0)^\top) & = & (0,0)^\top, \\
\mathcal{R}((x,0)^\top) & = & \{(r,0)^\top, |r \in \mathbb{R}^*\} \;\; \textit{for } x \in \mathbb{R}^*, \\
\mathcal{R}((0,y)^\top) & = & \{(0,r)^\top, |r \in \mathbb{R}^*\} \;\; \textit{for } y \in \mathbb{R}^*, \\
\mathcal{R}((x,y)^\top) & = & \{(rx, \epsilon r y)^\top, |\epsilon \in \{-1,1\}, r \in \mathbb{R}^*\} \;\; \textit{for } (x,y) \in (\mathbb{R}^*)^2.
\end{array}$$

*b) Assume that A has an eigenvalue of multiplicity two.*

*(i) If* $\lambda = 0$*, then* $\mathcal{R}((x,y)^\top) = \{(-x,-y),(x,y)\}$ *for all* $(x,y) \in \mathbb{R}^2$.

*(ii) If* $\lambda \neq 0$*, then there exist only three reachable sets. The reachable graph is given by*

$$\begin{tikzcd}
\mathcal{R}((0,0)^\top) & \mathcal{R}((1,0)^\top) \arrow[l] & \mathcal{R}((1,1)^\top) \arrow[l]
\end{tikzcd}$$

*c) Assume*

$$A = \begin{pmatrix} \operatorname{Re}\lambda & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & \operatorname{Re}\lambda \end{pmatrix} \quad \textit{with} \quad \operatorname{Im}\lambda \neq 0.$$

(i) *If* $\operatorname{Re}\lambda = 0$*, then*

$$\mathcal{R}((x,y)^\top) = \{(a,b) \in \mathbb{R}^2 \mid a^2 + b^2 = x^2 + y^2\}$$

*for all* $(x,y) \in \mathbb{R}^2$*.*

(ii) *If* $\operatorname{Re}\lambda = 0$*, then* $\mathcal{R}((0,0)^\top) = (0,0)^\top$ *and* $\mathcal{R}((x,y)^\top) = \mathbb{R}^2 \setminus (0,0)^\top$ *for any* $(x,y) \in \mathbb{R}^2 \setminus (0,0)^\top$*.*

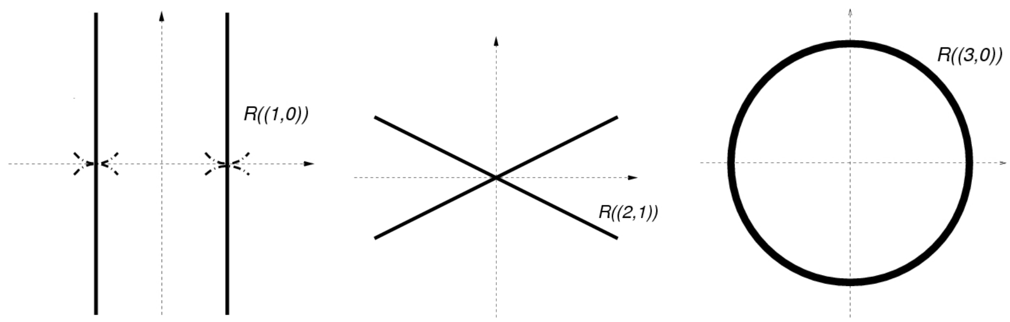



**Figure 6: Left:** *example for case a,ii). Here* $A = \operatorname{diag}(0,1)$*. The reachable set of* $(x,y)^\top$ *with* $y, x \neq 0$ *has four connected components. Moreover, the orbit* $\{(-x,0)^\top, (x,0)^\top\}$ *lies in the topological closure of* $\mathcal{R}((x,y)^\top)$*.* **Middle:** *Example for case a,iii). Let* $A = \operatorname{diag}(-1,1)$*. Again, the reachable set of* $(x,y)^\top$ *with* $y, x \neq 0$ *for* $\Sigma^{CI}(A)$ *has four connected components. The orbit* $\{(0,0)^\top\}$ *lies in the topological closure of* $\mathcal{R}((x,y)^\top)$*.* **Right:** *Example for case c,i) with* $\operatorname{Re}\lambda = 0$ *and* $\operatorname{Im}\lambda = 1$*. Here, none of the reachable sets is in the topological closure of another reachable set.*

# 9 Richardson's method

One of the most important tasks in numerical linear algebra is to solve systems of linear equations $Ax = b$ with $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. Iteration schemes of the form

$$x_{t+1} = x_t + u_t(b - Ax_t), x_0 \in \mathbb{R}^n$$

with $u_t \in \mathbb{R}$ are called *Richardson methods.* In this context, it is also common to call the shift parameters $u_t \in \mathbb{R}$ *relaxation parameters.*

The literature provides different shift strategies, each of them for certain families of matrices, see [OS84, SS88, GO88] and [CR96]. In particular, a constant shift strategy $u_t = u$ yields the so-called *trivial splitting method*, i.e.,

$$x_{t+1} = (I - uA)x_t + ub.$$

It is easy to verify, that a trivial splitting method converges if and only if $\operatorname{Spec}(I - uA) \subseteq \mathbb{D}$ (see [Gre97], Theorem 2.1.1). Another interesting shift strategy is given by the feedback law $u_t = \frac{r_t^\top Ar_t}{\|Ar_t\|^2}$ with $r_t = b - Ax_t$. This approach yields $GMRES(1)$, i.e.,

$$x_{t+1} = \arg\min_{x \in x_t + \operatorname{span}(b - Ax_t)} \|b - Ax\|.$$

It is known, that $GMRES(1)$ converges if $A + A^\top$ is positive definite (see [Mei99], Theorem 4.78). Nevertheless, the convergence properties of $GMRES(1)$ for general matrices is far from being understood (see [Emb03] for some notes on this topic).

The sequence $(x_t)_{t \in \mathbb{N}}$ converges to $A^{-1}b$ if and only if the sequence of residuals $r_t := b - Ax_t$ converges to zero. Thus, the dynamic of the iteration can be equivalently described in terms of the residual vectors, i.e.,

$$r_{t+1} = b - Ax_{t+1} = b - A((I - u_tA)x_t + u_tb) = (I - u_tA)(b - Ax_t) = (I - u_tA)r_t.$$

This motivates the following setting.

**Definition 9.1 (Richardson system)** Let $A \in \mathbb{R}^{n \times n}$ be invertible and $U_{A^{-1}} = \mathbb{R} \setminus \operatorname{Spec}(A^{-1})$. The system $\Sigma^{RS}(A) = (\mathbb{R}^n, U_{A^{-1}}, f_A^{RS})$ given by the transmission map $f_A^{RS} : (r, u) \mapsto (I - uA)r$ is called *Richardson system* (with respect to $A$).

Clearly, the existence of a shift strategy $u = (u_t)_{t \in \mathbb{N}}$ such that $x_t \xrightarrow{u} A^{-1}b$ implies that

$$0 \in \overline{\mathcal{R}(r_0)} \quad \text{for} \quad r_0 = b - Ax_0. \tag{56}$$

In the following we show, sufficient as well as necessary conditions for (56).

## 9.1 Richardson system semigroups

For the system semigroup of $\Sigma^{RS}(A) = (\mathbb{R}^n, U_{A^{-1}}, f_A^{RS})$ we obtain

$$S_{\Sigma^{RS}(A)} = \left\{ \prod_{t=1}^{T} (I - u_t A) \,|\, T \in \mathbb{N}, u_t \in U_{A^{-1}} \right\} \subseteq \mathrm{GL}_n(\mathbb{R}).$$

Obviously, we have $S_{\Sigma^{RS}(\alpha TAT^{-1})} = TS_{\Sigma^{RS}(\alpha A)}T^{-1}$ for any $T \in \mathrm{GL}_n(\mathbb{R})$ and $\alpha \in \mathbb{R} \setminus \{0\}$. $S_{\Sigma^{RS}(A)}$ and $G_{\Sigma^{RS}(A)}$ are closely related to the corresponding objects of inverse iteration systems. In fact, the following Proposition shows, that the system groups of inverse iteration (with respect to $A$) and the system group of Richardson systems (with respect to $A$) coincide.

**Proposition 9.2** *Let $A \in \mathbb{R}^{n\times n}$ be invertible. Then*

$$G_{\Sigma^{RS}(A)} = P(A).$$

**Proof.** Recall that $A^{-1} \in P(A)$ and $A \in P(A^{-1})$. Therefore, it follows $P(A) = P(A^{-1})$. Moreover, every element $B$ of $G_{\Sigma^{RS}(A)} := \langle S_{\Sigma^{RS}(A)} \rangle$ can be written as

$$B = \underbrace{\prod_{t=1}^{T} (I - u_t A)}_{\in P(A)} \underbrace{\prod_{t=1}^{\tilde{T}} (I - \tilde{u}_t A)^{-1}}_{\in P(A)}$$

for some $T, \tilde{T} \in \mathbb{N}$ and $u_t, \tilde{u}_t \in U_{A^{-1}}$. Thus, $G_{\Sigma^{RS}(A)} \subseteq P(A)$. With Corollary 6.5 we obtain

$$\begin{aligned} P(A^{-1}) &= \left\{ \prod_{t=1}^{T} (A^{-1} - u_t I) \prod_{t=1}^{T} (A^{-1} - \tilde{u}_t I)^{-1} \,\middle|\, T \in \mathbb{N}, u_t, \tilde{u}_t \in U_{A^{-1}} \right\} \\ &= \left\{ \prod_{t=1}^{T} (I - u_t A) \prod_{t=1}^{T} (I - \tilde{u}_t A)^{-1} \,\middle|\, T \in \mathbb{N}, u_t, \tilde{u}_t \in U_{A^{-1}} \right\} \\ &\subseteq G_{\Sigma^{RS}(A)}. \end{aligned}$$

Hence, $G_{\Sigma^{RS}(A)} = P(A)$. □

In particular, Proposition 9.2 shows, that similar to the situation for inverse iteration systems on $\mathbb{R}^n$ there exists an open and dense system group orbit $N_A = G_{\Sigma^{RS}(A)} \cdot x$, $x \in N_A$. Again, $N_A$ is defined as

$$N_A := \mathbb{R}^n \setminus \bigcup_{V \in \mathrm{Inv}_A} V \subseteq \mathbb{R}^n$$

where, $\mathrm{Inv}_A$ denotes the proper $A$-invariant subspaces of $A$. Using the techniques developed in Section 4.3 and Section 6 we easily obtain the following result.

**Theorem 9.3** *Let $A$ be cyclic and invertible. Assume that $r_0 \in N_A$.*

*a)* $0 \in \overline{\mathcal{R}(r_0)}$ *if and only if* $0 \in \overline{\mathcal{R}(\tilde{r}_0)}$ *for any* $\tilde{r}_0 \in N_A$.

*b) If* $S_{\Sigma^{RS}(A)} = P(A)$, *then* $0 \in \overline{\mathcal{R}(r_0)}$ *for all* $r_0 \in N_A$.

**Proof.** a) Recall that $P(A)$ acts transitively on $N_A$ (see Lemma 6.11). Moreover, $\{0\}$ is a $\Sigma$-invariant subset with $\{0\} \subseteq \overline{N_A}$. Thus, by Theorem 4.18, $\{0\} \cap \overline{\mathcal{R}(r_0)} = \emptyset$ if and only if $\{0\}$ is repelling to $N_A$.
b) $S_{\Sigma^{RS}(A)} = P(A)$ implies $0 \in \overline{N_A} = \overline{\mathcal{R}(r_0)}$ since

$$\mathcal{R}(r_0) = S_{\Sigma^{RS}(A)} \cdot r_0 = P(A) \cdot r_0 = N_A.$$

□

## 9.2 Conditions for $S_{\Sigma^{RS}(A)} = P(A)$

Similar to inverse iteration systems, the system semigroup is not always a group (see Theorem 9.8). Nevertheless, the following proposition shows, that $S_{\Sigma^{RS}(A)}$ is a large subset of $P(A)$ in a topological sense.

**Proposition 9.4** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic and invertible. Then*

$$\operatorname{int}_{P(A)} S_{\Sigma^{RS}(A)} \neq \emptyset.$$

**Proof.** We have

$$\begin{aligned} S_{\Sigma^{RS}(A)} &= \left\{ A^T \prod_{t=1}^{T} (A^{-1} - u_t I) \,\middle|\, T \in \mathbb{N}, u_t \in U_{A^{-1}} \right\} \\ &\supseteq A^n \left\{ \prod_{t=1}^{n} (A^{-1} - u_t I) \,\middle|\, u_t \in U_{A^{-1}} \right\}. \end{aligned}$$

Recall that $A$ is cyclic if and only if $A^{-1}$ is cyclic. By Corollary 6.6, the set $\{\prod_{t=1}^{n}(A^{-1} - u_t I) \,|\, u_t \in U_{A^{-1}}\}$ has open interior with respect to $P(A)$. Thus, $\operatorname{int}_{P(A)} S_{\Sigma^{RS}(A)} \neq \emptyset$. □

In Section 6, and respectively Section 7.3, we have proved a series of sufficient and necessary conditions, such that $S(A)\mathbb{R}^* = P(A)$, and respectively, $S(A) = P(A)$. It turns out, that neither $S(A) = P(A)$ nor $S(A)\mathbb{R}^* = P(A)$ implies that $S_{\Sigma^{RS}(A)} = P(A)$. Examples for that phenomenon will be given in Section 9.3. Nevertheless, we obtain the following useful fact.

**Lemma 9.5** *Let $A \in \mathbb{R}^{n\times n}$ be invertible.*

*a) If $S_{\Sigma^{RS}(A)}$ is a group, then $S(A)\mathbb{R}^*$ is a group.*

*b) If $\mathbb{R}^* I \subseteq S_{\Sigma^{RS}(A)}$, then $S(A)\mathbb{R}^* = P(A)$ implies $S_{\Sigma^{RS}(A)} = P(A)$.*

**Proof.** a) If $S_{\Sigma^{RS}(A)}$ is a group, then $S_{\Sigma^{RS}(A)} = P(A)$ by Proposition 9.2. Hence, for all $p(A) \in P(A)$ there exist $N \in \mathbb{N}$, $u_1, \dots, u_N \in U_{A^{-1}}$ such that

$$p(A) = \prod_{t=1}^{N} (I - u_t A) = \prod_{t=1}^{N} (-u_t) \prod_{t=1}^{N} \left( A - \frac{1}{u_t} I \right) \in \mathbb{R}^* (S(A))^{-1}.$$

Thus, $p(A) \in \mathbb{R}^*(S(A))^{-1}$. It follows that, $\mathbb{R}^*(S(A))^{-1}$ is a group and therefore $\mathbb{R}^*(S(A))^{-1} = S(A)\mathbb{R}^* = P(A)$.
b) Obviously, $S_{\Sigma^{RS}(A)} \subseteq P(A)$. Moreover, we have $\operatorname{int}_{P(A)} S_{\Sigma^{RS}(A)} \neq \emptyset$ (see Proposition 9.4). Thus, it is enough to show $S(A)\mathbb{R}^* \subseteq S_{\Sigma^{RS}(A)}$.

Let $B := r\prod_{t=1}^{T}(A - u_t I) \in S(A)\mathbb{R}^*$, i.e., $T \in \mathbb{N}$, $u_t \in U_A$ and $r \in \mathbb{R}^*$. If $u_t \neq 0$ for all $t = 1, \ldots, T$, then

$$B = \underbrace{(-1)^T r u_1 \cdot \cdots \cdot u_T}_{\in \mathbb{R}^* I \subseteq S_{\Sigma^{RS}(A)}} \underbrace{\prod_{t=1}^{T}(I - \frac{1}{u_t} I)}_{\in S_{\Sigma^{RS}(A)}}.$$

Note that $\{r\prod_{t=1}^{T}(A - u_t I) \in S(A)\mathbb{R}^* \mid u_t \neq 0\}$ is a dense subset of $S(A)\mathbb{R}^*$ and therefore, $S_{\Sigma^{RS}(A)}$ is a dense subset of $S(A)\mathbb{R}^* = P(A)$. By Lemma B.6 we conclude $S_{\Sigma^{RS}(A)} = P(A)$. □

**Theorem 9.6** *For any $n \in \mathbb{N}$ there exists an open set of invertible matrices, such that $S_{\Sigma^{RS}(A)} = P(A)$. In particular $S_{\Sigma^{RS}(A)} = P(A)$ if $A$ has $n$ different real eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}^n \setminus \{0\}$.*

**Proof.** Without loss of generality we assume that $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Since $S(A)\mathbb{R}^* = P(A)$ (see Theorem 6.29), it is sufficient to show that $\mathbb{R}^* I \subseteq S_{\Sigma^{RS}(A)}$. For any $r \in \mathbb{R}^*$ there exist shifts $v_1, \ldots, v_{n+1} \in U_{A^{-1}}$ such that

$$\prod_{t=1}^{n+1}(I - v_t A) = rI.$$

Define $\lambda_{n+1} = 0$. Let $p$ be the unique polynomial of degree $n$ with $p(\lambda_i) = r$ for $i = 1, \ldots, n$ and $p(\lambda_{n+1}) = 1$. By Lemma 6.28 there exists $M \in \mathbb{R}$ and $f \in \mathcal{L}$ such that

$$p(x) = f(x) - M\prod_{t=1}^{n+1}(x - \lambda_t).$$

Recall that $\deg p = k$ and therefore $f(x) = M\prod_{t=1}^{n+1}(x - u_t)$ for some $u_t \in \mathbb{R}$. Since $\lambda_{n+1} = 0$ we obtain

$$1 = p(0) = f(0) - 0 = M\prod_{t=1}^{n+1}(-u_t).$$

Moreover, $f(\lambda_i) = p(\lambda_i) - 0 = r$ for $i = 1, \ldots, n$. Note that $u_t \neq 0$, since $p(0) \neq 0$. Therefore, $v_t := \frac{1}{u_t}$ yields

$$f(x) = M(-1)^{n+1}u_1 \ldots u_{n+1}\prod_{t=1}^{n+1}(1 - v_t x) = \prod_{t=1}^{n+1}(1 - v_t x).$$

We conclude

$$\prod_{t=1}^{n+1}(I - v_t A) = f(A) = p(A) - M\prod_{t=1}^{n+1}(A - \lambda_t I) = p(A) = rI.$$

□

We finish this section with a result which shows, that $S_{\Sigma^{RS}(A)}$ is not a group in general.

**Theorem 9.7** *Let $A \in \mathbb{R}^n$ be an invertible cyclic matrix with $\lambda, \overline{\lambda} \in \operatorname{Spec}(A)$such that $\operatorname{Im}\lambda \neq 0$ and $\operatorname{Re}\lambda = 0$. Then*

*a) $S_{\Sigma^{RS}(A)} \neq P(A)$.*

*b) $\{0\}$ is repelling to $N_A$, i.e., $\{0\} \cap \overline{\mathcal{R}(r_0)} = \emptyset$ for any $r_0 \in N_A$. In particular, there exists no shift strategy $u = (u_t)_{t\in\mathbb{N}}$ such that $x_t \xrightarrow{u} A^{-1}b$.*

**Proof.** a) Without loss of generality we assume

$$A = \begin{pmatrix} A_1 & * \\ 0 & * \end{pmatrix} \quad \text{with} \quad A_1 = \begin{pmatrix} 0 & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & 0 \end{pmatrix}.$$

Assume, that $S_{\Sigma^{RS}(A)}$ is a group, i.e., $S_{\Sigma^{RS}(A)} = P(A)$. In particular, $rA \in P(A)$ with $r > \frac{1}{\operatorname{Im}\lambda}$ has an inverse in $S_{\Sigma^{RS}(A)}$. Thus, there exist $N \in \mathbb{N}$, $u_1, \ldots, u_N \in U_{A^{-1}}$ such that

$$I_2 = rA_1 \prod_{t=1}^{N} (I - u_t A_1).$$

But this is a contradiction to

$$\det\left( rA_1 \prod_{t=1}^{N} (I - u_t A_1) \right) = r^2 (\operatorname{Im}\lambda)^2 \prod_{t=1}^{N} (1 + u_t^2 (\operatorname{Im}\lambda)^2) > 1. \tag{57}$$

Hence, $S_{\Sigma^{RS}(A)}$ is not a group.
b) By Theorem 9.3 we may assume that $r_0 = (1, 1, 1 \ldots, 1)^\top$. Assuming that $\{0\} \cap \overline{\mathcal{R}(r_0)} \neq \emptyset$. Then there exists a sequence

$$s_n := \begin{pmatrix} B_n & * \\ 0 & * \end{pmatrix} \in S_{\Sigma^{RS}(A)},$$

with $B_n \in \mathbb{R}^{2\times 2}$ such that $B_n (1,1)^\top \to 0$ for $n \to \infty$. Since

$$B_n \subseteq P(A_1) := \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \middle| \; a^2 + b^2 \neq 0 \right\}$$

and $\det B_n = \det \prod_{t=1}^{N} (I - u_t A_1) \geq 1$ we obtain

$$\|B_n (1,1)^\top\|_2 = \sqrt{(a+b)^2 + (a-b)^2} = \sqrt{2\det(B_n)} \geq \sqrt{2}.$$

Thus $\{0\} \cap \overline{\mathcal{R}(r_0)} = \emptyset$. □

## 9.3 Richardson's method on the plane

In this section we classify the semigroup types of $S_{\Sigma^{RS}(A)}$ for invertible cyclic matrices $A \in \mathbb{R}^{2\times 2}$.

**Theorem 9.8** *Let $A \in \mathbb{R}^{2\times 2}$ be cyclic and invertible.*

*a) If $A$ has two different real eigenvalues, then $S_{\Sigma^{RS}(A)} = P(A) \cong (\mathbb{R}^*)^2$.*

*b) If $A$ has one real eigenvalue with multiplicity 2, then $S_{\Sigma^{RS}(A)} = P(A) \cong \mathbb{R} \times \mathbb{R}^*$.*

*c) Assume, that $A$ has a pair of complex eigenvectors $\lambda, \overline{\lambda}$ such that $\operatorname{Im}\lambda \neq 0$.*

*(i) If $\operatorname{Re}(\lambda) \neq 0$ then $S_{\Sigma^{RS}(A)} = P(A) \cong \mathbb{C}^*$.*

*(ii) If $\operatorname{Re}(\lambda) = 0$ then $S_{\Sigma^{RS}(A)}$ is not a group. More precisely,*

$$S_{\Sigma^{RS}(A)} \cong (\mathbb{C} \setminus \overline{\mathbb{D}}) \cup \{1\}.$$

**Proof.** a) The first claim follows immediately from Theorem 9.6 and Theorem 6.8.
b) We show that $\mathbb{R}^* I \subseteq S_{\Sigma^{RS}(A)}$. Then, the claim follows from Lemma 9.5 since $S(A)\mathbb{R}^* = P(A)$ (see Theorem 6.29). If $r \in \mathbb{R} \setminus [0,1]$ then the choice

$$v := \frac{1}{\lambda}\left(1 - r + \sqrt{r(r-1)}\right); \quad u := \frac{1-r}{v\lambda^2}$$

yields

$$(I - uA)(I - vA) = I - (u+v)A + uvA^2 = rI,$$

since $uv = \frac{1-r}{\lambda^2}$ and $u+v = \frac{2(1-r)}{\lambda}$. Any $r = (-1)(-r) \in [0,1]$ is the product of elements of $\mathbb{R} \setminus [0,1]$. Thus, $\mathbb{R}^* I \subseteq S_{\Sigma^{RS}(A)}$.
c) Without loss of generality we assume $\operatorname{Im}\lambda = 1$. We identify the matrix $I - uA$ with the complex number $z(u) := (1 - u\operatorname{Re}\lambda) - iu$. Thus,

$$S_{\Sigma^{RS}(A)} = \left\{ \prod_{t=1}^{T} z(u_t) \,\middle|\, t \in \mathbb{N}, u_t \in \mathbb{R} \right\}.$$

(i) We show that $M_\beta := \{i\beta + u \,|\, u \in \mathbb{R}\} \subseteq S_{\Sigma^{RS}(A)}$ for one $0 < 1 < \beta$. Then, the claim follows, since the set of finite products of elements of $M_\beta$ is $\mathbb{C}^*$ (see Corollary 7.11). There exists an open set in $U \subseteq \mathbb{R}$ with $0 \in \overline{U}$ such that

$$|z(u)| = \sqrt{1 - 2u\operatorname{Re}\lambda + u^2|\lambda^2|} < 1 \text{ for } u \in U.$$

More precisely, $|z(u)| < 1$ for $0 < u < \frac{2\operatorname{Re}\lambda}{|\lambda|^2}$ if $\operatorname{Re}\lambda > 0$, and respectively $|z(u)| < 1$ for $\frac{2\operatorname{Re}\lambda}{|\lambda|^2} < u < 0$ if $\operatorname{Re}\lambda < 0$. Therefore, we can choose $u \in \mathbb{R}$ such that $|z(u)| < 1$ and $\arg z(u) = \frac{\pi}{2n}$ for $n \in \mathbb{N}$ large enough. Then

$$z(u)^n = \beta i \in S_{\Sigma^{RS}(A)} \quad \text{with} \quad \beta = |z(u)|^n.$$

Since $M_\beta = \{\beta i(1 - ui) \,|\, u \in \mathbb{R}\}$ we obtain $M_\beta \subseteq S_{\Sigma^{RS}(A)}$ and thus $S_{\Sigma^{RS}(A)} = \mathbb{C}^*$.

(ii) $S_{\Sigma^{RS}(A)}$ is not a group by Theorem 9.7. Again we identify the matrices $I - uA$, $u \in \mathbb{R}$ with complex numbers. Here, $z(u) := 1 - iu \operatorname{Im}\lambda$. For any $z \in S_{\Sigma^{RS}(A)}$ we have

$$|z| = \prod_{t=1}^{T} \underbrace{|1 - iu_t|}_{\geq 1}.$$

It follows that $|z| \geq 1$ and $|z| = 1$ if and only if $z = 1$. Thus,

$$S_{\Sigma^{RS}(A)} \subseteq (\mathbb{C} \setminus \overline{\mathbb{D}}) \cup \{1\}.$$

Now we show that $M_1 := \{i + u \,|\, u \in \mathbb{R}\} \subseteq S_{\Sigma^{RS}(A)} \cup \{1\}$. Then, the claim follows, since $\mathbb{C}^* \setminus \overline{\mathbb{D}}$ lies in the set of finite products of elements of $M_1$ (see Corollary 7.11).

For $u \in \mathbb{R} \setminus \{0\}$ we construct $u_1, \ldots, u_T$ such that $z(u_1) \cdot \cdots \cdot z(u_T) = i + u$. Let $u_n = \tan\frac{\pi}{2n}$. Then $z(u_n)^n = |z(u_n)|^n i$. Moreover, $|z_n|^n - 1$ is arbitrary small (for $n$ sufficiently large). Now we choose $n$ such that $u^2 > 4|z_n|^n(|z_n|^n - 1)$. Then for

$$v := \frac{1}{2\beta}(u + \sqrt{u^2 - 4|z_n|^n(|z_n|^n - 1)}), \quad r := \frac{|z(u_n)|^n - 1}{v}$$

we have

$$\begin{aligned}
z(u_n)^n z\left(\frac{r}{|z(u_n)|^n}\right) z(v) &= i|z(u_n)|^n \left(1 - i\frac{r}{|z(u_n)|^n}\right)(1 - iv) \\
&= (i|z(u_n)|^n + r)(1 - iv) \\
&= i(|z(u_n)|^n - vr) + r + v|z(u_n)|^n \\
&= i + u.
\end{aligned}$$

Thus, $S_{\Sigma^{RS}(A)} = (\mathbb{C} \setminus \overline{\mathbb{D}}) \cup \{1\}$. □

**Corollary 9.9** *Let $A \in \mathbb{R}^{2\times}$ by cyclic and invertible. 0 is repelling to $N_A$ if and only if $A$ has a pair of complex eigenvalues $\lambda, \overline{\lambda}$ with $\operatorname{Re}\lambda = 0$. In this case*[27] *$\mathcal{R}(z) = |z|(\mathbb{C} \setminus \overline{\mathbb{D}}) \cup \{z\}$ for all $z \neq 0$.*

[27] with respect to the identification $\mathbb{R}^2 \cong \mathbb{C}$ and $I - uA \cong (1 - u\operatorname{Re}\lambda) - iu)$.

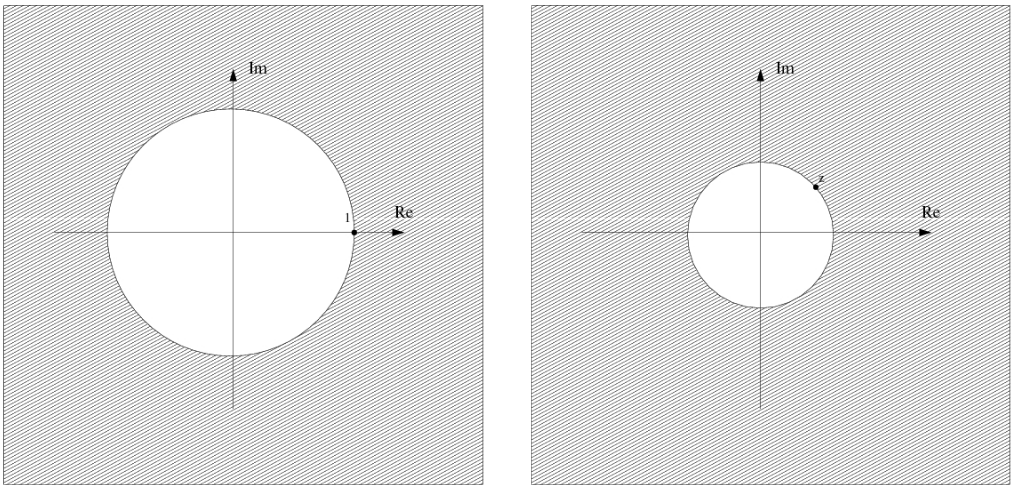



**Figure 7: Left:** *System semigroup of Richardson systems embedded in* $\mathbb{C}^*$ *for* $A \in \mathbb{R}^{2\times 2}$ *with* $\operatorname{Spec} A = \{i, -i\}$. *Here,* $S_{\Sigma^{RS}(A)} \cong (\mathbb{C} \setminus \overline{\mathbb{D}}) \cup \{1\}$. *We obtain* $\mathcal{R}(z) = \{z\tilde{z} \mid \tilde{z} \in S_{\Sigma^{RS}(A)}\} = |z|(\mathbb{C} \setminus \overline{\mathbb{D}}) \cup \{z\}$. **Right:** *Reachable set for* $z = \frac{1}{2}e^{\frac{\pi}{4}i}$.

## 9.4 Restarted polynomial iteration

Given an initial guess $x_0$ for the solution of a linear equation $Ax = b$, $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, a *restarted polynomial iteration of degree* $m$ is an iteration scheme of the form

$$x_{t+1} = x_t - p_t(A)(b - Ax_t) \tag{58}$$

where $p_t \in \mathbb{R}[x]$ with $\deg p_t < m$. Restarted polynomial methods are also called *restarted Krylov methods*, since

$$x_{n+1} \in x_t + \mathcal{K}_m(A, r_t)$$

where $\mathcal{K}_m(A, r_t)$ denotes the *Krylov space* with respect to $A$ and $r_t := b - Ax_t$, i.e., $\mathcal{K}_m(A, r_t) := \operatorname{span}(r_t, Ar_t \ldots, A^{m-1}r_t)$ . Similar to Richardson's method, the dynamics of the iteration can be equivalently described by the dynamics of the residual sequence $(r_t)_{t\in\mathbb{N}}$. We obtain

$$r_{t+1} = b - A(x_t - p_t(A)(b - Ax_t)) = (I - Ap_t(A))r_t.$$

This motivates the following setting.

**Definition 9.10 (Polynomial iteration system)** Let $A \in \mathbb{R}^{n\times n}$ be invertible and

$$U_A^{PI} := \{p \in \mathbb{R}[x] \mid \deg(p) < m + 1, I - Ap(A) \,\text{invertible}\}.$$

The system $\Sigma^{PI}(A) = (\mathbb{R}^n, U_A^{PI}, f_A^{PI})$ given by the transmission map $f_A^{PI}$ : $(r, p) \mapsto (I - Ap(A))r$ is called *Polynomial iteration system* (with respect to $A$).

Note that Richardson's method and restarted polynomial iteration coincide for $m = 1$. We have seen that the Richardson system semigroups are not necessarily groups (see Theorem 9.7). In the following we show, that the system semigroup of polynomial iteration system is a group, provided $m \geq 2$.

**Theorem 9.11** *Let $A \in \mathbb{R}^{n\times n}$ be cyclic and $\Sigma^{PI}(A) = (\mathbb{R}^n, U_A^{PI}, f_A^{PI})$ be a polynomial iteration system of degree $m \geq 2$. Then*

*a) $S_{\Sigma^{PI}(A)}(A) = P(A)$.*

*b) $\Sigma^{PRS}(A)$ is weakly reversible.*

*c) $\Sigma^{PRS}(A)_{|N_A}$ is controllable.*

**Proof.** a) Obviously, $S_{\Sigma^{RS}(A)} \subseteq S_{\Sigma^{PI}(A)} \subseteq P(A)$. Moreover, the system semigroup for polynomial iterations systems for polynomials of degree $m$ is included in the system semigroup for polynomial iterations systems for polynomials of degree $m+1$. Therefore, it is sufficient to show the claim for $m=2$. Recall that $\langle S_{\Sigma^{RS}(A)} \rangle = P(A)$ by Proposition 9.2. Thus, we only have to show that $S_{\Sigma^{PI}(A)}$ is a group, i.e., we show that for any $p \in U_A^{PI}$ there exists $k \in \mathbb{N}$ and $p_1, \dots, p_k \in U_A^{PI}$ such that

$$f_p \circ f_{p_1} \circ \cdots \circ f_{p_k} = I.$$

By the Cayley Hamilton theorem there exists a polynomial $\tilde{p}$ of degree at most $n$ such that

$$(I - p(A)A)^{-1} = \tilde{p}(A). \tag{59}$$

We decompose $\tilde{p}$ in linear or quadratic polynomials, i.e.,

$$\tilde{p}(t) = (\alpha_1 + tr_1(t)) \dots (\alpha_k + tr_k(t)) \quad \text{with} \quad \deg r_j \leq 1, j = 1, \dots, k.$$

Since $\tilde{p}(A)$ is invertible we have $\alpha_j \neq 0, j = 1, \dots, k$. Moreover, (59) implies

$$(1 - p(t)t)\tilde{p}(t) = (1 - p(t)t)(\alpha_1 + tr_1(t)) \dots (\alpha_k + tr_k(t)) = 1 + k(t)m_A(t)$$

for some $k \in \mathbb{R}[t]$. Since $\deg(p) = m = 2$, $\deg \tilde{p} \leq n$ and $\deg m_A(t) = n$ we obtain $\alpha_1 \dots \alpha_k = 1$. Thus,

$$I = (I - Ap(A)(I - Ap_1(A)) \dots (I - Ap_k(A))$$

with $p_j := \frac{-1}{\alpha_j} r_j$. This proves claim a).

b) and c) Clearly, $\Sigma^{PRS}(A)$ is weakly reversible if $S_{\Sigma^{PI}(A)}(A)$ is a group. Moreover, $P(A)$ acts transitively on $N_A$ (see Lemma 6.11). Thus, statements b) and c) are immediate consequences of statement a). □

# 10 Linear control schemes

Another approach to design iterative methods for solving linear equations $Ax = b$ is via *linear control schemes*, i.e., given $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, we want to find a shift sequence $U = (u_1, u_2, \dots)$, $u_t \in \mathbb{R}^m$ with $\lim_{t\to\infty} u_t = 0$ and $B \in \mathbb{R}^{n\times m}$ such that the sequence

$$x_{t+1} = (I - A)x_t + Bu_t + b \tag{60}$$

converges. Then, the limit of $(x_t)_{t\in\mathbb{N}}$ is a solution of the equation $Ax = b$. Without loss of generality we assume that $b$ lies in the *image space* of $B$, i.e., $b \in \operatorname{Image} B := \{By \mid y \in \mathbb{R}^m\}$. Otherwise we set $\tilde{B} := [b, B] \in \mathbb{R}^{n\times(m+1)}$. Assuming that $A$ is invertible, we have

$$x = A^{-1}b = \sum_{j=0}^{n-1} \alpha_j (I - A)^j b$$

for some $\alpha_j \in \mathbb{R}$, $j = 0, \dots, n-1$. Thus, $x \in \operatorname{Image} \mathbf{R}(I - A, B)$ where $\mathbf{R}(I - A, B)$ is the *Kalman matrix* of the pair $(I - A, B)$, i.e.,

$$\mathbf{R}(I - A, B) := [B, (I - A)B, \dots, (I - A)^{n-1}B].$$

This approach yields the following definition.

**Definition 10.1 (linear control system)** Let $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n\times m}$ such that $b \in \operatorname{Image} B$. System $\Sigma^B(A) = (\mathbb{R}^n, \mathbb{R}^m, f^B)$ with

$$f^B : U \times \mathbb{R}^n \to \mathbb{R}^n;\ f^B(u, x) = (I - A)x + Bu + b$$

is called *linear control system* (of equation $Ax = b$ with respect to $B$).

In the following we analyze the system semigroup and the reachable structure of linear control systems (Section 10.1). Our results imply the *Kalman rank condition for controllability*, a well known fact from the theory of linear control systems. Moreover, we present a feedback law such that (60) converges globally to a solution of $Ax = b$ (Section 10.2).

## 10.1 Linear control system semigroup

Obviously, every composition of maps $f^B_{u_1}, \dots, f^B_{u_T}$ is an affine map. Therefore, the system semigroup $S_{\Sigma^B(A)}$ of the linear system $\Sigma^B(A)$ is a subsemigroup of the affine group $\operatorname{Aff}_n(\mathbb{R})$, provided $I - A$ is invertible. By induction we obtain

$$f_{u_1}^B \circ \cdots \circ f_{u_N}^B(x) = (I-A)^N x + \sum_{t=1}^{N} (I-A)^{t-1}(Bu_t + b). \tag{61}$$

In particular this shows, that the identity is not element of the semigroup $S_{\Sigma^B(A)}$ for almost all $A \in \mathbb{R}^{n\times n}$. Thus, $S_{\Sigma^B(A)}$ is in general not a group.

In contrast to the system semigroups we have analyzed in the previous chapters, $S_\Sigma^B(A)$ is not abelian. In fact equation (61) shows

$$f_{u_1}^B \circ f_{u_2}^B(x) = (I-A)^2 x + B(u_1+u_2) + (2I-A)b - ABu_2$$

and therefore

$$f_{u_1}^B \circ f_{u_2}^B(x) - f_{u_2}^B \circ f_{u_1}^B(x) = AB(u_1-u_2).$$

In other words, $f_{u_1}^B$ and $f_{u_2}^B$ commute if and only if $f_{u_1}^B = f_{u_2}^B$, provided $\operatorname{rank}(AB) = m$. Nevertheless, it turns out, that $S_{\Sigma^B(A)}$ is right divisible.

**Theorem 10.2** *Let $A \in \mathbb{R}^{n\times n}$ such that $I-A$ is invertible and $B \in \mathbb{R}^{n\times m}$ such that $b \in \operatorname{Image} B$. Then $\Sigma^B(A)$ is right divisible and left divisible. The system group is given by*

$$G_{\Sigma^B(A)} = \left\{g : x \mapsto (I-A)^Z x + v \,|\, Z \in \mathbb{Z}\ v \in \operatorname{Image} \mathbf{R}(I-A, B)\right\}.$$

**Proof.** Without loss of generality we assume $b = 0$ (otherwise we set $\tilde{u}_t$ such that $Bu_t + b = B\tilde{u}_t$). By equation (61) we easily deduce

$$(f_{u_N}^B)^{-1} \circ \cdots \circ (f_{u_1}^B)^{-1}(x) = x \mapsto (I-A)^{-N}\left(x - \sum_{t=1}^{N}(I-A)^{t-1}Bu_t\right).$$

Obviously, every finite product

$$f_{u_1}^B \circ \cdots \circ f_{u_{N_1}}^B \circ (f_{w_1}^B)^{-1} \circ \cdots \circ (f_{w_{N_2}}^B)^{-1}$$

with $u_1, \ldots, u_{N_1}, w_1, \ldots, w_{N_2} \in \mathbb{R}^m$ is an affine map and therefore contained in the group $\{x \mapsto (I-A)^Z x + v \,|\, Z \in \mathbb{Z}, v \in \mathbb{R}^n\} \subseteq \operatorname{Aff}_n(\mathbb{R})$. More precisely, $v = \sum_{t=Z_1}^{Z_2} \alpha_t (I-A)^t Bu_t$ for some $Z_1, Z_2 \in \mathbb{Z}$ and $\alpha_t \in \mathbb{R}$. Thus,

$$\begin{aligned} S_{\Sigma^B(A)}(S_{\Sigma^B(A)})^{-1} &\subseteq G_{\Sigma^B(A)} \\ &\subseteq \underbrace{\{x \mapsto (I-A)^Z x + v \,|\, Z \in \mathbb{Z}, v \in \operatorname{Image} \mathbf{R}(I-A, B)\}}_{=:G}. \end{aligned}$$

Now we show that for every $Z \in \mathbb{Z}$ and every $v \in \operatorname{Image} \mathbf{R}(I-A, B)$ there exists $N_1, N_2 \in \mathbb{N}$ and $u_1 \ldots, u_{N_1}, w_1, \ldots, w_{N_2} \in \mathbb{R}^m$ such that

$$s_1 s_2^{-1}(x) = (I-A)^Z x + v$$

for $s_1 := f_{u_1} \circ \cdots \circ f_{u_{N_1}}$ and $s_2 := f_{w_1} \circ \cdots \circ f_{w_{N_2}}$. Then $G$ is a subset of $S_{\Sigma^B(A)}(S_{\Sigma^B(A)})^{-1}$ and thus $G_{\Sigma^B(A)} = G$.

**Case I:** We assume that $Z \geq 0$. We choose, $N_1 = Z + n$, $N_2 = n$ and $u_1, \ldots, u_{N_1} = 0$. Since $v \in \operatorname{Image} \mathbf{R}(I - A, B)$ and $\operatorname{Image} \mathbf{R}(I - A, B)$ is $(I - A)$ invariant, there exists $w_1, \ldots, w_n \in \mathbb{R}^m$ such that

$$-(I-A)^{-Z} v = \sum_{t=1}^{n} (I-A)^{t-1} B w_t.$$

Therefore,

$$\begin{aligned} s_1 s_2^{-1}(x) &= (I-A)^{Z+n}(I-A)^{-n}\left(x - \sum_{t=1}^{n}(I-A)^{t-1}Bw_t\right) \\ &= (I-A)^Z x - (I-A)^Z \sum_{t=1}^{n}(I-A)^{t-1}Bw_t \\ &= (I-A)^Z x + v \end{aligned}$$

**Case II:** Now we assume $Z < 0$. We choose $\tilde{w}_1, \ldots, \tilde{w}_n \in \mathbb{R}^m$ such that

$$v = \sum_{t=1}^{n} (I-A)^{t-1} B \tilde{w}_t.$$

From case I we deduce

$$s_1 \tilde{s}_2^{-1}(x) = (I-A)^{\tilde{Z}} x - (I-A)^{\tilde{Z}} v$$

for $\tilde{Z} = -Z$ and $\tilde{s}_2 = f^B_{\tilde{w}_1} \circ \cdots \circ f^B_{\tilde{w}_{N_2}}$. Therefore,

$$\begin{aligned} \tilde{s}_2 s_1^{-1}(x) &= \left(s_1 \tilde{s}_2^{-1}\right)^{-1}(x) \\ &= (I-A)^{-\tilde{Z}}(x + (I-A)^{\tilde{Z}} v) \\ &= (I-A)^Z x + v. \end{aligned}$$

Thus, in both cases $x \mapsto (I-A)^Z x + v$ is an element of $S_{\Sigma^B(A)}\left(S_{\Sigma^B(A)}\right)^{-1}$. We conclude

$$G_{\Sigma^B(A)} = S_{\Sigma^B(A)}\left(S_{\Sigma^B(A)}\right)^{-1} = G.$$

Hence, $\Sigma^B(A)$ is right divisible. Analogously, we can show that any element of $G$ can be written as a product $s_1^{-1}s_2$ with $s_1, s_2 \in S_{\Sigma^B(A)}$. Thus, $\Sigma^B(A)$ is also left divisible. □

Knowing the explicit types of the system group we easily obtain the following result on the adherence structure of the reachable sets. In particular, we deduce the well known *Kalman rank condition* for controllability.

**Theorem 10.3** *Let $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n\times m}$ such that $I - A$ is invertible and $b \in \operatorname{Image} B$. Consider $\Sigma^B(A) := (\mathbb{R}^n, \mathbb{R}^m, f^B)$.*

*a) Every reachable set is the countable union of affine subspaces of $\mathbb{R}^n$ with dimension at most $\operatorname{rank} \mathbf{R}(I - A, B)$.*

*b) Every system group orbit is the countable union of affine subspaces of $\mathbb{R}^n$ with dimension $\operatorname{rank} \mathbf{R}(I - A, B)$.*

*c) $\Sigma^B(A)$ restricted to $G_{\Sigma^B(A)} \cdot 0$ is controllable.*

*d) $\Sigma^B(A)$ is controllable if and only if $\operatorname{rank} \mathbf{R}(I - A, B) = n$.*

**Proof.** a) From (61) and $b \in \operatorname{Image} B$ it follows

$$\begin{aligned} \mathcal{R}(x) &= \left\{ (I-A)^N x + \sum_{t=1}^{N} (I-A)^{t-1} B v_t \,\middle|\, N \in \mathbb{N}, v_t \in \mathbb{R}^m \right\} \quad (62) \\ &= \bigcup_{t\in\mathbb{N}} \big((I-A)^t x + K_t\big) \end{aligned}$$

with $K_t := \{\sum_{j=0}^{t-1} (I-A)^j B w_j \,|\, w_j \in \mathbb{R}^m, j = 0, \ldots, t-1\}$. Note that, $K_n = \operatorname{Image} \mathbf{R}(I - A, B)$ and $\dim K_t \leq \dim K_n$ for any $t \in \mathbb{N}$.
b) By Theorem 10.2 we have

$$G_{\Sigma^B(A)} \cdot x = \bigcup_{t\in\mathbb{Z}} \big((I-A)^t x + K_n(I-A,B)\big) . \quad (63)$$

c) From (62) and (63) it follows $\mathcal{R}(x) = G_{\Sigma^B(A)} \cdot 0$ for all $x \in \operatorname{Image} \mathbf{R}(I - A, B)$. Thus, $\Sigma^B(A)$ restricted to $G_{\Sigma^B(A)} \cdot 0$ is controllable by Proposition 2.31.
d) Obviously, $\operatorname{rank} \mathbf{R}(I-A, B) = n$ implies controllability by c). Conversely, $\operatorname{rank} \mathbf{R}(I - A, B) < n$ implies

$$G_{\Sigma^B(A)} \cdot x = \bigcup_{Z\in\mathbb{Z}} (I-A)^Z x + \operatorname{Image} \mathbf{R}(I-A, B) \neq \mathbb{R}^m .$$

Thus, $\mathcal{R}(x) \neq \mathbb{R}^n$. □

In the following we assume that also $A$ is invertible. Theorem 10.3 shows, that $A^{-1}b \in \mathcal{R}(x)$ if and only if $x \in \operatorname{Image} \mathbf{R}(I - A, B)$. Note that the set of pairs $(I - A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ which satisfy $\operatorname{rank} \mathbf{R}(I - A, B) = n$, is open and dense[28] in $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ (see Proposition 3.3.12 in [Son98]). If the Kalman rank condition does not hold, i.e., $\operatorname{rank} \mathbf{R}(I - A, B) < n$, every

[28] The generaliy of this statement is not restricted by the general assumption $b \in \operatorname{Image} B$, since $\operatorname{rank} \mathbf{R}(I - A, B) = n$ implies $\operatorname{rank} \mathbf{R}(I - A, [b, B]) = n$.

reachable set is the countable union of affine spaces with dimension at most rank $\mathbf{R}(I-A,B) < n$ and therefore of measure zero. Nevertheless, in some (but not all) situations we have $A^{-1}b \subseteq \overline{\mathcal{R}(x)}$ for some $x \in \mathbb{R}^n \setminus \mathcal{R}(0)$ (see Example 10.5). A sufficient condition for this phenomenon will be presented in Section 10.2 (see Theorem 10.10). The following result shows, that $A^{-1}b \subseteq \overline{\mathcal{R}(x)}$ is a property of the entire orbit $G_{\Sigma^B(A)} \cdot x$.

**Theorem 10.4** *Consider $\Sigma^B(A) := (\mathbb{R}^n, \mathbb{R}^m, f^B)$ with $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n\times m}$ such that $A$ and $I-A$ is invertible and $b \in \operatorname{Image} B$. Let $y, z \in G_{\Sigma^B(A)} \cdot x$ for some $x \in \mathbb{R}^n$. Then*

$$A^{-1}b \in \overline{\mathcal{R}(y)} \quad \textit{if and only if} \quad A^{-1}b \in \overline{\mathcal{R}(z)}.$$

**Proof.** Recall that $G_{\Sigma^B(A)}$ is right divisible. Therefore, there exists $w \in G_{\Sigma^B(A)} \cdot x$ such that $\mathcal{R}(y) \cup \mathcal{R}(z) \subseteq \mathcal{R}(w)$ (see Theorem 4.8). It follows, that $z = (I-A)^N w + v$ for some $N \in \mathbb{N}$ and $v \in \operatorname{Image} \mathbf{R}(I-A,B)$. By (63) it follows

$$\mathcal{R}(w) \setminus \mathcal{R}(z) \subseteq \bigcup_{t=1}^{N-1} \left((I-A)^t z + \operatorname{Image} \mathbf{R}(I-A,B)\right).$$

Since $(I-A)^t z \notin \operatorname{Image} \mathbf{R}(I-A,B) = \mathcal{R}(0)$ for any $t = 1, \ldots, N-1$ it follows

$$\mathcal{R}(0) \cap \overline{\mathcal{R}(w) \setminus \mathcal{R}(z)} = \emptyset. \tag{64}$$

Now we assume that $A^{-1}b \in \overline{\mathcal{R}(y)}$. Then $A^{-1}b \in \overline{\mathcal{R}(w)}$, since $\mathcal{R}(y) \cup \mathcal{R}(z) \subseteq \mathcal{R}(w)$. By (64) it follows $A^{-1}b \in \overline{\mathcal{R}(z)}$. The converse direction follows analogous. □

We finish this section with an example which shows, that there exist linear $B$-systems with $A^{-1}b \notin \overline{\mathcal{R}(x)}$ for some $x \in \mathbb{R}^n$ as well as systems with $A^{-1}b \in \overline{\mathcal{R}(x)}$ for some $x \in \mathbb{R}^n \setminus \mathcal{R}(0)$.

**Example 10.5** Consider $\Sigma^B(A_a) = (\mathbb{R}^2, \mathbb{R}, f^B)$ with

$$I - A_a = \begin{pmatrix} 2 & 0 \\ 0 & a \end{pmatrix} \quad \text{and} \quad B = b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Clearly we have $\operatorname{Image} \mathbf{R}(I-A,B) = \{(y,0)^\top \mid y \in \mathbb{R}\}$ and $A_a^{-1}b \in \operatorname{Image} \mathbf{R}(I-A,B)$. For any $(x_1,x_2)^\top$ we have

$$G_{\Sigma^B(A_a)} \cdot x = \left\{ \begin{pmatrix} 2^Z x_1 \\ a^Z x_2 \end{pmatrix} + v, \;\middle|\; Z \in \mathbb{Z}, v \in \operatorname{Image} \mathbf{R}(I-A,B) \right\}$$

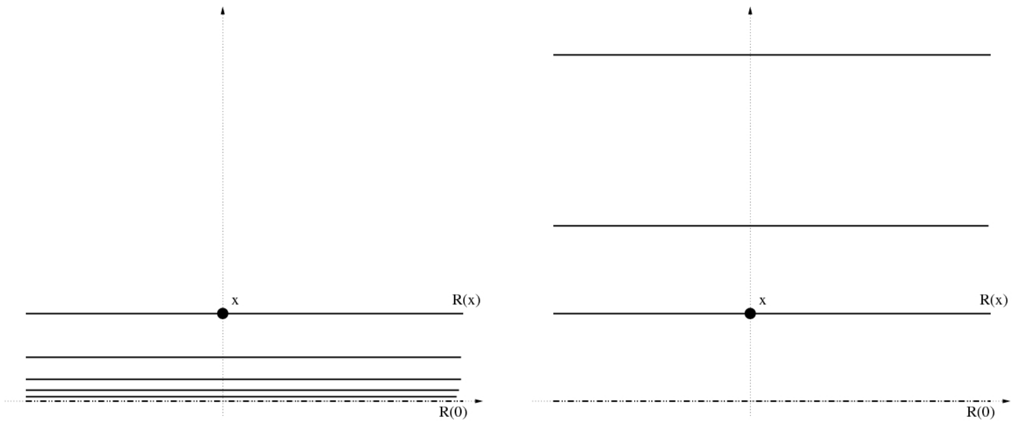



**Figure 8:** *Illustration to Example 10.5. The reachable sets are countable unions of affine subspaces.* **Left:** *$\mathcal{R}(x)$ for the case $0 < a < 1$. Here, $\mathcal{R}(0)$ lies in the topological closure of $\mathcal{R}(x)$.* **Right:** *$\mathcal{R}(x)$ for the case $|a| > 1$. Here, $\mathcal{R}(0) \cap \overline{\mathcal{R}(x)} = \emptyset$.*

and

$$\mathcal{R}(x) = \left\{ \begin{pmatrix} 2^N x_1 \\ a^N x_2 \end{pmatrix} + v, \;\middle|\; N \in \mathbb{N}, v \in \operatorname{Image} \mathbf{R}(I - A, B) \right\}.$$

Thus, for $x = (0,1)^\top$, we have $\mathcal{R}(0) \subseteq \overline{\mathcal{R}(x)}$ if $|a| < 1$ and $\mathcal{R}(0) \cap \overline{\mathcal{R}(x)} = \emptyset$ if $|a| \geq 1$.

## 10.2 Shift strategies via quadratic controller design

In this following we introduce a method, for constructing an explicit shift sequence such that (60) converges globally to a solution of $Ax = b$. We use the following classic result by Kalman [Kal60]. A proof can be found in [LR95], Theorem 16.6.4.

**Theorem 10.6** *Consider the linear control system* $\Sigma = (\mathbb{R}^n, \mathbb{R}^m, L)$, *given by*

$$r_{t+1} = L(r_t, u_t) = \tilde{A} r_t + \tilde{B} u_t$$

*and the cost functional*

$$J_{r_0}(u_0, u_1, \dots) = \sum_{t=0}^{\infty} (\|r_t\|^2 + \|u_t\|^2). \tag{65}$$

*Assume that* $(\tilde{A}, \tilde{B})$ *is discrete-time stabilizable, i.e.,* $\operatorname{rank}[\lambda I - \tilde{A}, \tilde{B}] = n$ *for any* $\lambda \in \mathbb{C}$ *with* $|\lambda| \geq 1$.

a) *The algebraic Riccati equation*

$$P = I_n + \tilde{A}^\top P \tilde{A} + (\tilde{B}^\top P \tilde{A})^\top (I_m + \tilde{B}^\top P \tilde{B})^{-1} \tilde{B}^\top P \tilde{A} \tag{66}$$

*has an unique symmetric positive definite solution* $P \in \mathbb{R}^{n \times n}$.

b) *There exists a unique control sequence* $u = (u_0, u_1, \dots)$ *such that* $J_{r_0}(u_0, u_1, \dots)$ *is minimal. This optimal control sequence is given by the feedback law* $u_t = -K r_t$ *with*

$$K = (I_m + \tilde{B}^\top P \tilde{B})^{-1} \tilde{B}^\top P \tilde{A}. \tag{67}$$

*Moreover,* $J_{r_0}(u_0, u_1, \dots) = r_0^\top P r_0$.

Now we apply Theorem 10.6 to $\Sigma^B(A)$. The dynamics of the residuals $r_t := b - Ax_t$ is given by the linear system

$$r_{t+1} = b - Ax_{t+1} = b - A((I - A)x_t + Bu_t + b) = (I - A)r_t - ABu_t.$$

Assume that $(I - A), -AB)$ is discrete-time stabilizable. By Theorem 10.6, $r_t$ converges to zero if we apply the feedback law $u_t = -Kr_t$ with

$$K = (I_m + (AB)^\top P (AB))^{-1} (-AB)^\top P (I - A).$$

Here $P$ is the unique solution of (66) with $\tilde{A} = I - A$ and $\tilde{B} = -AB$. This yields the following algorithm proposed by Helmke, Jordan and Lanzon ([HJ05, HJL06]).

**Algorithm 10.7 (LQRES)**

(i) Choose $B$ such that $(I - A, -AB)$ is stabilizable

(ii) Calculate the unique positive definite solution of the Riccati Equation (66) for $\tilde{A} = I - A$ and $\tilde{B} = -AB$.

(iii) Calculate $K$ as in Equation (67) for $\tilde{A} = I - A$ and $\tilde{B} = -AB$.

(iv) Iterate the closed loop system

$$x_{t+1} = (I - A)x_t + BK(b - Ax_t) + b. \tag{68}$$

By Theorem 10.6 we immediately obtain the following convergence result for LQRES.

**Theorem 10.8** *If $(I - A, -AB)$ is stabilizable then (68) converges to a solution of $Ax = b$.*

Note that a solution to step (i) may not exist for arbitrary choices of $A$. However, for generic choices of $A$ step (i) is always solvable. Moreover, the freedom in choosing $B$ can be exploited to improve convergence speed (see Example 10.12 and Example 10.13). If the eigenvalues $\lambda$ of $A$ satisfy $|1 - \lambda| < 1$, then one can choose $B = 0$. Then LQRES coincides with the Richardson's method $x_{t+1} = (I - uA)x_t + ub$ with constant shift strategy $u \equiv 1$. The following example shows, that LQRES converges in cases, where Richardson's iteration fails for all possible shift strategies.

**Example 10.9** Consider $Ax = b$ with

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

By Theorem 9.7 $Ax = b$ is not solvable for any Richardson's method. However, $(I - A, -AB)$ is stabilizable for the choice $B := b$. Thus LQRES converges.

Provided the dimension of $U = \mathbb{R}^m$ is relatively small, step (iii) does not cause numerical problems. However, the expensive preconditioning process by solving the algebraic Riccati equation (66) in step (ii) is a serious numerical problem. In fact, any known method is more expensive then solving the origin equation $Ax = b$. Nevertheless, we believe variations of LQRES, using suboptimal techniques for solving equation (66), yield attractive alternatives to the common numerical algorithms.

Theorem 10.8 provides an interesting result on the adherence structure of reachable sets.

**Theorem 10.10** *Let $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n \times m}$ such that $I - A$ and $A$ are invertible, $(I - A, -AB)$ is stabilizable and $b \in \operatorname{Image} B$. Consider $\Sigma^B(A) = (\mathbb{R}^n, \mathbb{R}^m, f^B)$. Then $\mathcal{R}(0) \subseteq \overline{\mathcal{R}(x)}$ for any $x \in \mathbb{R}^n$. In particular $A^{-1}b \in \overline{\mathcal{R}(x)}$ for any $x \in \mathbb{R}^n$.*

**Proof.** By Theorem 10.8 there exists a sequence $(x_t)_{t\in\mathbb{N}}$ with $x_0 = x$ in $\mathcal{R}(x)$ which converges to $A^{-1}b \in \mathcal{R}(0) = \operatorname{Image} \mathbf{R}(I - A, B)$. Thus, $A^{-1}b \in \overline{\mathcal{R}(x)}$. For any $v \in \operatorname{Image} \mathbf{R}(I - A, B)$ the sequence $x_t + (v - A^{-1}b)$ lies in $\mathcal{R}(x)$ since $v - A^{-1}b \in \operatorname{Image} \mathbf{R}(I - A, B)$ and $\mathcal{R}(x) = \bigcup_{t\in\mathbb{N}} ((I - A)^t x + K_t)$. Thus, $\mathcal{R}(0) \subseteq \overline{\mathcal{R}(x)}$. □

Clearly, the statement of Theorem 10.10 is trivial if the Kalman rank condition holds. The following example shows, that the assumptions of Theorem 10.10 do not imply $\operatorname{rank} \mathbf{R}(I - A, B) = n$.

**Example 10.11** Consider $\Sigma^B(A_a) = (\mathbb{R}^2, \mathbb{R}, f^B)$ of Example 10.5 with $a \in (0, 1)$. Recall that $A_a$ and $I - A_a$ are invertible. Moreover,

$$\operatorname{rank}[\lambda I - (I - A_a), -A_a B] = \operatorname{rank} \begin{pmatrix} \lambda - 2 & 0 & -2 \\ 0 & \lambda - a & 0 \end{pmatrix} = 2$$

for all $|\lambda| > 1$. Thus, $(I - A_a, -A_a B)$ is stabilizable. By Theorem 10.10 it follows, that $A^{-1}b \in \overline{\mathcal{R}(x)}$ for any $x \in \mathbb{R}^n$. However,

$$\operatorname{rank} \mathbf{R}(I - A_a, B) = \operatorname{rank} \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix} < 2.$$

We finish this section with some numerical experiments which demonstrate the dependence of the convergence properties of LQRES on the choice of the parameter $B$.

**Example 10.12** Consider $Ax = b$ for

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{and } b = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}.$$

We choose $x_0 = 0$ as an initial guess. This example is known to produce extreme behavior for restarted GMRES algorithms. In particular GMRES(2) fails to converge while GMRES(1) converges (see [Emb03]). We choose

$$B1 = \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}, \quad B2 = \begin{pmatrix} 3 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad B3 = \begin{pmatrix} 3 & -1 \\ 1 & -2 \\ 1 & -3 \end{pmatrix}$$

The convergence behavior of LQRES is shown in Figure 9.

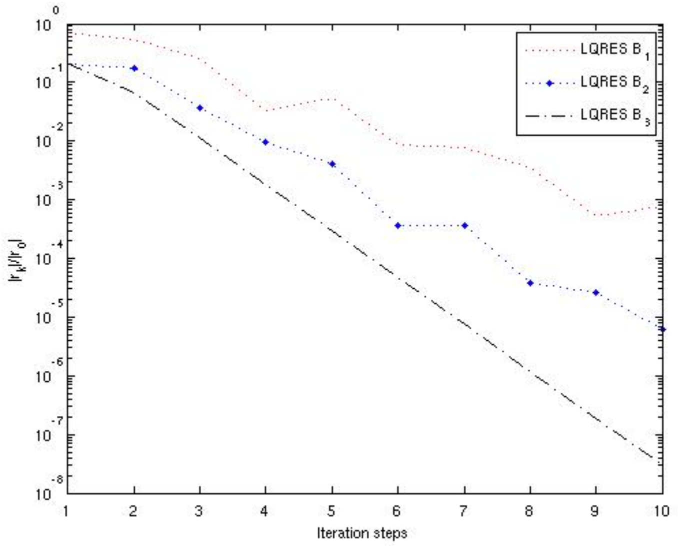



**Figure 9:** *LQRES in Example 10.12. We compare the relative residuals after $n$ outer iteration steps. The algorithm converges for all parameters $B1$, $B2$, $B3$. However, the speed of convergence depends on the choice of $B$.*

**Example 10.13** Now we consider $Ax = b$ where $b = (1, 0, 0, 0, 0)^\top$ and $A$ is the Hilbert matrix of order 5. The elements of the Hilbert matrices are given by $a_{i,j} = \frac{1}{i+j-1}$. It is known that this matrix is poorly conditioned (see [FM67], Chapter 19). We choose

$$B1 = b, \quad B2 = \begin{pmatrix} 1 & 1 \\ 0 & -1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad B3 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & -1 \end{pmatrix}$$

The convergence behavior of LQRES with respect to $B1$, $B2$ and $B3$ is shown in Figure 10.

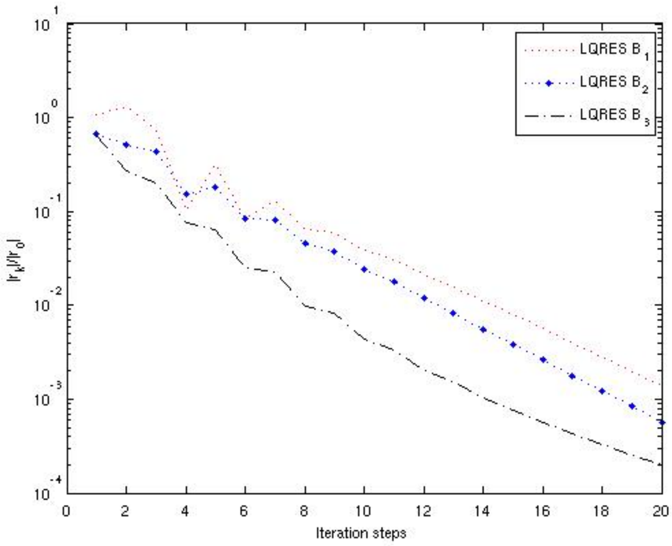



**Figure 10:** *LQRES applied on a Hilbert matrix of dimension 5 (Example 10.13). We compare the relative residuals after n outer iteration steps. We observe that the speed of convergence increases when the number of columns of B gets larger.*

# A Semi-algebraic sets

In Part II of this thesis, we analyze systems where the state space is a real algebraic set and the transition map is a rational homomorphism. To take advantage of this situation we shall use some basic concepts from algebraic geometry. Here, we briefly recall some basic notations and properties of semi-algebraic sets which will be important for our analysis. See [BCR98, CLO91] for a more detailed overview on real algebraic geometry.

We call a set $A \subseteq \mathbb{R}^N$ a *variety* or a real *algebraic set* if there exists a set of polynomials $P \subseteq \mathbb{R}[x_1, \ldots, x_N]$ such that

$$A = \{x \in \mathbb{R}^N \mid p(x) = 0, \forall p \in P\}.$$

A variety $A$ is called *irreducible* if $A = A_1 \cup A_2$ with varieties $A_1$, $A_2$ implies $A = A_1$ or $A = A_2$. A set $A \subseteq \mathbb{R}^N$ is called *semi-algebraic* if it can be written as the finite union of sets of the form

$$\{x \in \mathbb{R}^N \mid f_1(x) = \cdots = f_l(x) = 0, g_1(x) > 0, \ldots, g_m(x) > 0\},$$

where $f_1, \ldots, f_l, g_1, \ldots, g_m \in \mathbb{R}[x_1, \ldots, x_N]$. A map $f : A \to B$ between semi-algebraic sets $A \subseteq \mathbb{R}^M$ and $B \subseteq \mathbb{R}^N$ is called *semi-algebraic* if

$$\mathrm{graph}(f) := \{(a, f(a)) \,|\, a \in A\}$$

is semi-algebraic in $\mathbb{R}^{M+N}$. In particular, every *regular morphism* is semi-algebraic, i.e., every map $f = (f_1, \ldots, f_M) : A \to B$ with rational components $f_k = p_k/q_k$, $k = 1, \ldots, M$ such that $p_k, q_k \in \mathbb{R}[x]$ and $q_k(x) \neq 0$ for all $x \in A$, is semi-algebraic.

One easily obtains the following:

**Proposition A.1** *a) If $A, B \subseteq \mathbb{R}^N$ are semi-algebraic, then $A \cap B$, $A \cup B$ and $A \setminus B$ are semi-algebraic.*

*b) If $A \subseteq \mathbb{R}^M$ and $B \subseteq \mathbb{R}^N$ are semi-algebraic sets, then $A \times B$ is semi-algebraic in $\mathbb{R}^{M+N}$.*

*c) If $f : A \to B$ is a semi-algebraic bijective map, then $f^{-1} : B \to A$ is semi-algebraic.*

*d) The composition $g \circ f$ of semi-algebraic maps $f : A \to B$ and $g : B \to C$ is semi-algebraic.*

*e) If $M$ and $U$ are semi-algebraic sets and $f : M \times U \to M$ is semi-algebraic, then $f_u : M \to M$, $m \mapsto f(m, u)$ is semi-algebraic.*

**Proof.** The proofs of claim a) and b) can be found in [BCR98], Chapter 2.1. Moreover, graph$(f)$ is semi-algebraic if and only if

$$\begin{aligned} \mathrm{graph}(f^{-1}) &= \{(a, f^{-1}(a)) \,|\, a \in A\} \\ &= \{(f(b), b) \,|\, b := f(a) \in B\} \end{aligned}$$

is semi-algebraic. This shows c). Claim d) is proven in [BCR98], Proposition 2.2.6. Claim e) follows from d), since $f_u = f \circ \pi_u$ where $\pi_u : M \to M \times U$, $m \mapsto (m, u)$. □

The following fact is also known as the *Tarski-Seidenberg theorem* .

**Theorem A.2** *Let $A$ be a semi-algebraic subset of $\mathbb{R}^{N+K}$ and $\pi : \mathbb{R}^{N+K} \to \mathbb{R}^N$ the projection on the first $N$ coordinates. Then $\pi(A)$ is a semi-algebraic subset of $\mathbb{R}^N$.*

For a proof we refer to [BCR98] (see Theorem 2.2.1).

Assume that $X \subseteq \mathbb{R}^M$ and $Y \subseteq \mathbb{R}^N$ are semi-algebraic sets and $A \subseteq X$ as well as $B \subseteq Y$ are semi-algebraic subsets. If $f : X \to Y$ is a semi-algebraic map, then $f(A)$ is the image of $(A \times Y) \cap \mathrm{graph}(f)$ under the projection $X \times Y \to Y$ and $f^{-1}(B)$ is the image of $(X \times B) \cap \mathrm{graph}(f)$ under the projection $X \times Y \to X$. By Proposition A.1 and Theorem A.2 we obtain:

**Corollary A.3** *Let $X \subseteq \mathbb{R}^M$ and $Y \subseteq \mathbb{R}^N$ be semi-algebraic sets and let $f : X \to N$ be a semi-algebraic map. Then for all semi-algebraic sets $A \subseteq X$ and $B \subseteq Y$ the sets $f(A)$ and $f^{-1}(B)$ are semi-algebraic.*

Another important property of semi-algebraic sets is that they can be decomposed in manifolds.

**Theorem A.4** *Every semi-algebraic subset $A \subseteq \mathbb{R}^N$ is the disjoint union of a finite number of semi-algebraic submanifolds $A_i \subseteq \mathbb{R}^N$, $i = 1, \dots, l$, such that each $A_i$ is diffeomorphic to $(0,1)^{d_i}$. Here $(0,1)^0$ is a point by convention. Moreover, $d := \max\{d_1, \dots, d_l\}$ is unique.*

See Proposition 2.9.10 in [BCR98] for the decomposition property. The fact that $d$ is unique follows from Corollary 2.8.9. in [BCR98]. We say $d =: \dim_s(A)$ is the *semi-algebraic dimension* of $A$. Note that $\dim_s(A) = \dim(A)$ if $A$ is a manifold (see Proposition 2.8.14 in [BCR98]).

**Lemma A.5** *Let $A \subseteq \mathbb{R}^N$ be a semi-algebraic set with $\dim_s(A) = d$. Then:*

*a) $\overline{A}$ is a semi-algebraic subset of $\mathbb{R}^N$ and $\dim_s(\overline{A}) = \dim_s(A)$.*

*b) $\dim_s(\overline{A} \setminus A) < \dim_s(A)$.*

*c) If $A$ is the finite union of semi-algebraic sets $A_1, \dots, A_k$ with dimensions $d_1, \dots, d_k$, then $d = \max\{d_1, \dots, d_k\}$.*

All claims are well known and can be found in [BCR98] (see Proposition 2.2.2 and Proposition 2.8.2 for claim *a*), Proposition 2.8.13 for claim *b*) and Proposition 2.8.5 for claim *c*).

As a consequence of Theorem A.4 and Lemma A.5 we obtain the following observation which is important in the proof of Theorem 2.7 (algebraic orbit theorem).

**Lemma A.6** *Let $A \subseteq \mathbb{R}^N$ be a semi-algebraic set with $\dim_s(A) = d$. Then there exists $x \in A$ and a neighborhood $U$ of $x$ in $\mathbb{R}^N$ such that $U \cap A$ is diffeomorphic to $(0,1)^d$.*

**Proof.** By Theorem A.4 we can write

$$A = A_1 \cup \dots \cup A_{k_1} \cup A_{k_1+1} \cup \dots \cup A_k$$

such that for all $i = 1, \dots, k_1$, $A_i$ is a submanifold of $\mathbb{R}^N$ diffeomorphic to $(0,1)^d$ and for all $i = k_1+1, \dots, k$, $A_i$ is diffeomorphic to $(0,1)^{\tilde{d}_i}$ with $\tilde{d}_i < d$.

By Lemma A.5, the set

$$\hat{A} := A_1 \setminus \left( \underbrace{\left( \bigcup_{1<i\le k_1} \overline{A_i} \right)}_{=:A_\alpha} \cup \underbrace{\left( \bigcup_{k_1<i\le k} \overline{A_i} \right)}_{=:A_\beta} \right) \tag{69}$$

is semi-algebraic. We shall show that $\dim_s(\hat{A}) = d$.

Since $\dim_s(A_\beta) = \max\{\dim_s(A_{k_1+1}), \ldots, \dim(A_k)\} < d$ we have

$$\dim((A_1 \setminus A_\beta) \cup A_\beta) = \dim_s(A_1 \setminus A_\beta) = \dim(A_1) = d.$$

Recall that $A_i$, $i = 1, \ldots, k$ are disjoint. It follows

$$\hat{A} = (A_1 \setminus A_\beta) \setminus A_\alpha = (A_1 \setminus A_\beta) \setminus \left( \bigcup_{1<i\le k_1} \left( \overline{A_i} \setminus A_i \right) \right).$$

By Lemma A.5 we have $\dim_s \left( \bigcup_{1<i\le k_1} \left( \overline{A_i} \setminus A_i \right) \right) < \dim_s(A_i) = d$. Therefore,

$$\begin{aligned} \dim_s(A_1 \setminus A_\beta) &= \dim_s \left( \left( (A_1 \setminus A_\beta) \setminus \bigcup_{1<i\le k_1} \left( \overline{A_i} \setminus A_i \right) \right) \cup \bigcup_{1<i\le k_1} \left( \overline{A_i} \setminus A_i \right) \right) \\ &= \dim_s \left( \left( (A_1 \setminus A_\beta) \setminus \bigcup_{1<i\le k_1} \left( \overline{A_i} \setminus A_i \right) \right) \right) \\ &= \dim_s(\hat{A}). \end{aligned}$$

This shows that $\dim_s(\hat{A}) = d$ and in particular $\hat{A} \neq \emptyset$. Thus, for all $x \in \hat{A}$ we can find $U \subseteq \mathbb{R}^N$ such that

$$U \cap A = U \cap \hat{A} = U \cap A_1.$$

Since $A_1$ is diffeomorphic to $(0,1)^d$, we can choose $U$ such that $U \cap A$ is diffeomorphic to $(0,1)^d$. $\square$

# B Topological semigroups

System groups are often equipped with a canonical topology such that $G_\Sigma$ is a topological group acting continuously on the state space. Therefore, some basic theory on topological groups and their subsemigroups turns out to be very helpful for the analysis of the reachable set structure of systems and algorithms. In the following we collect some useful properties on this subject which can be found in [Hus66, HN93, HHL89, Mit01] and [SBG+95].

**Definition B.1** A topological space $G$ that is also a group is called a *topological group* if the mappings $G \times G \to G$, $(g_1, g_2) \mapsto g_1 g_2$ and $G \to G$, $g \mapsto g^{-1}$ are both continuous. Analogously, a topological space $S$ that is also a semigroup is called a *topological semigroup* if the mapping $S \times S \to S$, $(s_1, s_2) \mapsto s_1 s_2$ is continuous.

Obviously, every subsemigroup of a topological group is a topological semigroup. Moreover, we observe the following.

**Lemma B.2** *Let $G$ be a topological group and $S$ a nonempty subsemigroup of $G$. Then*

*a) The topological closure of $S$ is a subsemigroup of $G$.*

*b) If $S$ is compact, then $S$ is a group.*

**Proof.** a) For any $s, \tilde{s} \in \overline{S}$ there exist sequences $(s_n)_{n\in\mathbb{N}}$ and $(\tilde{s}_n)_{n\in\mathbb{N}}$ in $S$ such that $s_n \to s$ and $\tilde{s}_n \to \tilde{s}$. Since the product in the topological group $G$ is a continuous map, we obtain $s\tilde{s} \in \overline{S}$. Therefore, $\overline{S}$ is a closed subsemigroup of $G$.
b) Since $S$ is compact, the sequence $s^n$ has a convergent subsequence $s^{n_k}$. Since $\lim_{k\to\infty} s^{n_k} = \lim_{k\to\infty} s^{n_{k+2}}$ it is

$$\lim_{k\to\infty} s^{n_{k+2}} s^{-n_k} s^{-1} = s^{-1}.$$

From $n_{k+2} - n_k > 1$ we deduce $s^{n_{k+2}} s^{-n_k} s^{-1} \in S$ and therefore $s^{-1} \in \overline{S} = S$. Hence, $S$ is a group. □

In the following we denote the neutral element of a topological group $G$ with $e$.

**Theorem B.3** *Let $G$ be a connected topological group. Then for any neighborhood $V$ of $e$ we have*

$$G = \bigcup_{n\in\mathbb{N}} V^n.$$

Here $V^n$ denotes the set of products of $n$ elements $v_i \in V$, i.e. $V^n := \{\prod_{i=1}^n v_i \,|v_i \in V\}$. For a proof we refer to [Hus66], Theorem 23.6.

**Lemma B.4** *Let $G$ be a topological group and $S$ a subsemigroup of $G$. If $e \in \operatorname{int}_G S$ and $S \cap G^i \neq \emptyset$ for every path-connected component $G^i$ of $G$, then $S = G$.*

**Proof.** (i) First we show the claim under the assumption that $G$ is path-connected. Let $V$ be an open set in $\operatorname{int}_G S$ such that $e \in V$. Since $G$ is a topological group,

$$\bigcup_{n\in\mathbb{N}} V^n = G \tag{70}$$

by Theorem B.3.

Since $S$ is a semigroup, it follows $V^n \subseteq S$ for all $n \in \mathbb{N}$ and therefore

$$S \subseteq G = \bigcup_{n\in\mathbb{N}} V^n \subseteq S.$$

(ii) Now we assume that $G$ has different path-connected components $G^i$, all of them having nonempty intersection with $S$. We show that $G^i \subseteq S$ and therefore $S = G$.

Let $G_e$ be the component of $e$ and $g_i$ an element of $G^i \cap S$. We define $r_{g_i} : G_e \to G_i$, $h \mapsto hg_i$. Note that $r_{g_i}$ is a homeomorphism with inverse $r_{g_i}^{-1} = r_{g_i^{-1}}$. Since $g_i$ is an element of the semigroup $S$, we obtain

$$r_{g_i}(S_e) = S_e g_i \subseteq S. \tag{71}$$

Here $S_e$ is the identity component of $S$. We show that $S_e$ is a semigroup. For any $a, b \in S_e$ there exists a path $s_a : [0,1] \to S_e$ with $s_a(0) = e$ and $s_a(1) = a$ and a path $s_b : [0,1] \to S_e$ with $s_b(0) = e$ and $s_b(1) = b$. Therefore the path $s : [0,1] \to S_e^2$, given by $s(t) := s_a(t)s_b(t)$, connects $s(0) = e$ and $s(1) = ab$. Hence, $S_e$ is a semigroup. By (i) it follows that $S_e = G_e$, and we conclude

$$G^i = r_{g_i}(G_e) = r_{g_i}(S_e) = S_e g_i \subseteq S, \tag{72}$$

for all $i \in I$. □

The following useful fact can be found in [HN93] (see Lemma 3.7).

**Lemma B.5** *Let $S$ be a subsemigroup of a connected topological group $G$. Then the following statements hold:*

*a)* $\operatorname{int}_G(S)$ *is a semigroup ideal, i.e.,*

$$\operatorname{int}_G(S)S \subseteq \operatorname{int}_G(S) \qquad \text{and} \qquad S\operatorname{int}_G(S) \subseteq \operatorname{int}_G(S).$$

*b) If* $e \in \overline{\operatorname{int}_G(S)}$*, then*

$$S \subseteq \overline{\operatorname{int}_G(S)} \qquad \text{and} \qquad \operatorname{int}_G(S) = \operatorname{int}_G(\overline{S}).$$

*c) If* $\operatorname{int}_G(S) \neq \emptyset$ *and* $\overline{S} = G$*, then* $S = G$

Typically, we have to deal with system groups with more than one connected component. Nevertheless, statement *c)* of Lemma B.5 also holds if $G$ is not connected.

**Lemma B.6** *Let $S$ be a subsemigroup of a topological group $G$. Assume that* $\operatorname{int}_G(S) \neq \emptyset$ *and* $\overline{S} = G$*. Then* $S = G$*.*

**Proof.** (i) By assumption it follows that $(\operatorname{int}_G S)^{-1} \subseteq \overline{S}$ and since $G \to G$, $g \mapsto g^{-1}$ is an open map,

$$(\operatorname{int}_G S)^{-1} \cap S \neq \emptyset.$$

In other words, there exists $s \in S$ such that $s^{-1} \in \operatorname{int}_G S$. We obtain

$$e = ss^{-1} \subseteq S \operatorname{int}_G S \subseteq \operatorname{int}_G S,$$

since $\operatorname{int}_G S$ is an ideal of $S$ (see Lemma B.5). Hence $e$ is an interior point of $S$, i.e.,

$$e \in \operatorname{int}_G(S).$$

(ii) Since $\overline{S} = G$, there exists for any $g \in G$ a sequence $(s_n)_{n\in\mathbb{N}}$ in $S$ with $s_n \to g$. In other words, the sequence $s_n^{-1}g$ converges to $e \in \operatorname{int}_G(S)$. Thus, $g \in s_n \operatorname{int}_G(S) \subseteq S$ for almost all $n \in \mathbb{N}$. Hence, $S = G$. □

Recall that $\mathrm{Stab}_x := \{g \in G \,|\, g \cdot x = x\}$ is a subgroup of $G$, the so called *stabilizer subgroup*. Reachable sets are orbits of semigroup actions. We say a semigroup $S_\Sigma$ acts *transitively* on $M$ if for $m_1, m_2 \in M$ there exists $s \in S_\Sigma$ such that $s \cdot m_1 = m_2$. If $S$ is a subsemigroup of a Lie group[29] $G$, the following condition for transitivity applies.

**Proposition B.7** *Let $G$ be a Lie group and $S$ a subsemigroup of $G$. We assume that $G$ acts continuous and transitively on a manifold $M$. Then:*

*a) If* $\operatorname{int}_G S \cap \mathrm{Stab}_x \neq \emptyset$*, then there exists a neighborhood of $x$ such that $S_\Sigma$ acts transitively on $U$.*

*b) If $M$ is connected and* $\operatorname{int}_G S \cap \mathrm{Stab}_x \neq \emptyset$ *for all $x \in M$, then $S$ acts transitively on $M$.*

[29] A Lie group is a differential manifold with topological group structure such that product and inversion are smooth maps.

For a proof we refer to Mittenhuber [Mit01](see Proposition 3.3.4 for statement $a$), respectively Proposition 3.3.5 for statement $b$).

We finish this section with an important fact known as *Effros theorem* . Recall that a topological space is *locally compact* if each point is contained in a compact neighborhood. In particular manifolds are locally compact[30]. A *Lindelöf space* is a topological space in which every open cover has a countable subcover. In particular, $G$ is a Lindelöf space if $G$ is a Lie group.

**Theorem B.8** *Let $G$ be a locally compact topological group and $M$ a locally compact topological space. Assume that $G$ is a Lindelöf space. If $G$ acts transitively and continuous on $M$, then the map $h_x : G \to M$, $g \mapsto g \cdot x$ is open.*

The proof of Theorem B.8 is based on *Baire's category theorem.* For more details we refer to [SBG$^+$95] (see Theorem 96.8).

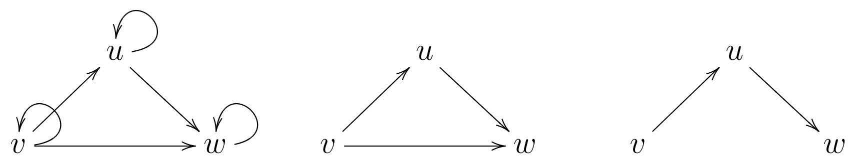
[30] Note that semi-algebraic sets are not locally compact in general.

# C Directed graphs

In this work we will describe the adherence structure of the reachable sets using a graph theoretical language. In the following we give a brief summery on the necessary notations and properties. The following definitions are standard and can be found in the books of Bollobás [Bol98] or Diestel [Die00].

**Definition C.1 (Directed graph)** A *directed graph* $\mathcal{G}$ is a pair $(V, \longleftarrow)$ containing of a set $V$, the set of *vertices* and a relation $\longleftarrow$ on $V$. A pair $(v_1, v_2) \in V \times V$ is called an *edge* from $v_1$ to $v_2$ if $v_2 \longleftarrow v_1$. We say that $\mathcal{G}$ is *infinite* if $V$ has infinitely many elements.

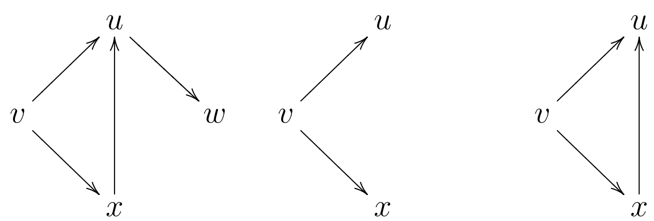
In this work, we only consider graphs $\mathcal{G} = (V, \longleftarrow)$ where the relation $\longleftarrow$ is *reflexive* and *transitive*, i.e., where $v \longleftarrow v$ for all $v \in V$ and $v_2 \longleftarrow v_1$ and $v_3 \longleftarrow v_2$ implies $v_3 \longleftarrow v_1$. Therefore, in figures we neglect *trivial edges*, i.e., edges from $v \in V$ to itself. Moreover we reduce the graph by those edges which are already clear by transitivity. The following diagram illustrates this reduction.




original graph   neglect trivial edges   reduced graph

**Definition C.2 (Subgraph)** Let $\mathcal{G}_2 = (V_2, \longleftarrow_2)$ and $\mathcal{G}_1 = (V_1, \longleftarrow_1)$ be directed graphs such that $V_2 \subseteq V_1$. We say $\mathcal{G}_2$ is a *subgraph* of $\mathcal{G}_1$ if $v \longleftarrow_2 w$ for $v, w \in V_2$ implies $v \longleftarrow_1 w$. We say $\mathcal{G}_2$ is an *induced subgraph* of $\mathcal{G}_1$ if for all $v, w \in V_2$: $v \longleftarrow_1 w$ is equivalent to $v \longleftarrow_2 w$.

Let $\mathcal{G}_1 = (V_1, \longleftarrow_1)$ be a directed graph and $V_2 \subseteq V_1$ a subset of vertices. In general, there exists more then one subgraph but a unique induced subgraph with vertex set $V_1$. In particular, the following graphs show, that not every subgraph is an induced subgraph.




original graph   subgraph   induced subgraph

**Definition C.3 (Graph isomorphism)** Let $\mathcal{G}_1 = (V_1, \longleftarrow_1)$ and $\mathcal{G}_2 = (V_2, \longleftarrow_2)$ be directed graphs. A map $\Phi : V_1 \to V_2$ is called *graph isomorphism* if it is bijective and $v_1 \longleftarrow_1 v_2$ for $v_1, v_2 \in V_1$ is equivalent to $\Phi(v_1) \longleftarrow_2 \Phi(v_2)$.

If $\Phi$ is a graph isomorphism between $\mathcal{G}_1 = (V_1, \longleftarrow_1)$ and $\mathcal{G}_2 = (V_2, \longleftarrow_2)$ and $\tilde{\mathcal{G}} = (\tilde{V}, \longleftarrow_{\tilde{1}})$ is a subgraph of $\mathcal{G}_1$, then we write $\Phi(\tilde{\mathcal{G}})$ for the graph $(\Phi(\tilde{V}), \longleftarrow_{\tilde{2}})$ defined by

$$\Phi(v_1) \longleftarrow_{\tilde{2}} \Phi(v_2) \quad \text{if and only if} \quad v_1 \longleftarrow_{\tilde{1}} v_2.$$

Note that $\Phi(\tilde{\mathcal{G}})$ is a subgraph of $\mathcal{G}_2$. This yields the following useful proposition:

**Proposition C.4** *$\mathcal{G}_1$ is isomorphic to a subgraph of $\mathcal{G}$ if and only if any subgraph $\tilde{\mathcal{G}}$ of $\mathcal{G}_1$ is isomorphic to a subgraph of $\mathcal{G}$.*

# D Cyclic matrices

**Definition D.1** A matrix $A \in \mathbb{R}^{n\times n}$ is called *cyclic*, if there exists $x \in \mathbb{R}^n$ such that the vectors $x, Ax, \ldots, A^{n-1}x$ form a basis of $\mathbb{R}^n$. Such a vector $x$ is also called a *cyclic vector*.

First of all we want to point out, that cyclicity is a *generic property*.

**Proposition D.2** *The set of cyclic matrices is open and dense in* $\mathbb{R}^{n\times n}$.

**Proof.** Consider the polynomial

$$P : \mathbb{R}^n \times \mathbb{R}^{n\times n} \to \mathbb{R}\ (x, A) \mapsto \det(x, Ax, \ldots, A^{n-1}x). \tag{73}$$

A matrix $A \in \mathbb{R}^{n\times n}$ is not cyclic if and only if $P(x, A) = 0$ for all $x \in \mathbb{R}^n$. Therefore, if $A$ is cyclic, there exist an $x \in \mathbb{R}^n$ such that $|P(x, A)| = |c| > 0$. It follows, that also $|P(x, B)| > 0$ for $\|A - B\|$ small enough, since $P$ is continuous. Hence, the set of cyclic matrices is open.

Now we show, that the map of cyclic matrices is dense in $\mathbb{R}^{n\times n}$. If $A$ is not cyclic, then for all $x \in \mathbb{R}^n$ we have $P(x, A) = 0$. Suppose there exists a neighborhood $\mathcal{O}$ of $A$ such that $P(x, B) = 0$ for all $B \in \mathcal{O}$. and all $x \in \mathbb{R}^n$. Then polynomial $P$ has to be constant zero, which is a contradiction to the definition. □

In the following we collect some characterizing properties of cyclic matrices, which will be important in our analysis.

**Proposition D.3** *The following statements are equivalent:*

*(i) A is cyclic*

*(ii) For the characteristic polynomial* $\chi_A(t) = \det(A - tI)$ *and the minimal polynomial* $m_A$ *it is* $\chi_A = (-1)^n m_A$.

*(iii) The matrix A has finitely many A-invariant subspaces.*

**Proof.** The equivalences of $(i)$ and $(ii)$ are shown in [Fuh96], Proposition 6.3.2.
$(ii) \Rightarrow (iii)$: If $\chi_A = (-1)^n m_A$ then for every real eigenvalue, respectively pair of complex eigenvalues, there exists exactly one block in the canonical form. Every block corresponds with exactly one $A$-invariant subspace. The set of invariant subspaces of $A$ consists of all possible sums of this subspaces and is therefore finite.
$(iii) \Rightarrow (ii)$: If $A$ has finite many proper $A$-invariant subspaces then the union of this subspaces is strictly smaller then $\mathbb{R}^n$. For any $x \in \mathbb{R}^n \setminus \{0\}$ which does not belong to one of this invariant subspaces it is $\sum_{i=1}^n \lambda_i A^i x = 0$ if and only if $\lambda_i = 0$ for all $i = 1, \ldots, n$. □

An immediate consequence of Proposition D.3 is the fact that $A$ is cyclic if and only if $T(A-uI)T^{-1}$ with $T \in \mathrm{GL}_n(\mathbb{R})$ and $u \in \mathbb{R}$ is cyclic. Moreover, in the case that $A$ is invertible, $A$ is cyclic if and only if $A^{-1}$ is cyclic.

Recall that the *centralizer* of a matrix $A \in \mathbb{R}^{n\times n}$ is defined as

$$Z(A) := \{Z \in \mathrm{GL}_n(\mathbb{R}) \,|\, ZA = AZ\}.$$

Note that $Z(A)$ is a closed subgroup of $\mathrm{GL}_n(\mathbb{R})$ and therefore a Lie group. Obviously, every element of

$$P(A) := \{p(A) \,|\, p \in \mathbb{R}[x]\,\text{coprime to}\; m_A\}$$

lies in $Z(A)$. The following statement and a proof can be found in [Fuh96] (see Proposition 6.1.2).

**Proposition D.4** *A matrix $A \in \mathbb{R}^{n\times n}$ is cyclic if and only if $Z(A) = P(A)$.*

Note that every matrix is similar to a block matrix

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \qquad A_1 \in \mathbb{R}^{n_1\times n_1}, A_2 \in \mathbb{R}^{(n-n_1)\times(n-n_1)}$$

such that $A_1$ is cyclic and $m_A = m_{A_1}$.

**Lemma D.5** *Let $A \in \mathbb{R}^{n\times n}$ (not necessarily cyclic) and $m_A$ the minimal polynomial of $A$.*

*a) If $A$ is a block matrix*

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \qquad A_1 \in \mathbb{R}^{n_1\times n_1},\ A_2 \in \mathbb{R}^{(n-n_1)\times(n-n_1)},$$

*then $P(A)$ is isomorphic to $P(A_1) \times P(A_2)$ if and only if the minimal polynomial of $A_1$ is coprime to the minimal polynomial of $A_2$.*

*b) If $A$ is a block matrix*

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \qquad A_1 \in \mathbb{R}^{n_1\times n_1},\ A_2 \in \mathbb{R}^{(n-n_1)\times(n-n_1)},$$

*such that the minimal polynomial of $A$ is equal to the minimal polynomial of $A_1$, then $P(A)$ and $P(A_1)$ are isomorphic.*

**Proof.** a) Let $m_1$, $m_2$ respectively $m_A$ be the minimal polynomials of $A_1$, $A_2$ respectively $A$. Obviously, $m_1$ and $m_2$ are divisors of $m_A$. For every $B \in P(A)$ there exists a unique $p \in \mathbb{R}[x]$ with $\deg p < \deg(m_A)$, such that

$B = p(A)$. (i) If $m_1$ and $m_2$ are not coprime, then the degree of $m_A$ is strictly smaller then $\deg m_1 + \deg m_2$. From Proposition 6.2 we deduce

$$\dim P(A) = \deg m_A < \deg m_1 + \deg m_2 = \dim(P(A_1) \times P(A_2))$$

Therefore $P(A) \ncong P(A_1) \times P(A_2)$.
(ii) Now let $m_1$ and $m_2$ be coprime. This is equivalent to $m_A = m_{A_1} m_{A_2}$. We show that

$$\Phi : P(A) \to P(A_1) \times P(A_2),\ p(A) \mapsto (p(A_1), p(A_2))$$

is a group isomorphism.
Obviously, $\Phi$ is well defined and injective, since

$$\begin{aligned} p_1(A) = p_2(A) \quad &\Leftrightarrow \quad p_1 - p_2 \equiv 0 \mod m_A \\ &\Leftrightarrow \quad p_1 - p_2 \equiv 0 \mod m_{A_1} \\ &\qquad \text{and } p_1 - p_2 \equiv 0 \mod m_{A_2} \\ &\Leftrightarrow \quad p_1(A_1) = p_2(A_1) \\ &\qquad \text{and } p_1(A_2) = p_2(A_2). \end{aligned}$$

Moreover, $\Phi$ is a group homomorphism, since

$$\Phi(p_1(A)p_2(A)) = (p_1p_2(A_1), p_1p_2(A_2)) = \Phi(p_1(A))\Phi(p_2(A)).$$

We show that $\Phi$ is surjective, if $m_1$ and $m_2$ are coprime. From Bezouts theorem we know, that there exist $\tilde{k}_1, \tilde{k}_2 \in \mathbb{R}[x]$ such that $1 = \tilde{k}_1 m_1 + \tilde{k}_2 m_2$. For any pair of polynomials $p_1, p_2$ such that $p_1$ is coprime to $m_1$ and $p_2$ is coprime to $p_2$, we define $k_1 := (p_1 - p_2)m_2\tilde{k}_1$ and $k_2 := (p_1 - p_2)m_1\tilde{k}_2$. Note that $p_1 - p_2 = k_1 m_1 - k_2 m_2$. Now we define $p := p_1 - k_1 m_1 = p_2 - k_2 m_2$. Since $p_1$ is coprime to $m_1$ and $p_2$ is coprime to $m_2$ it follows, that $p$ is coprime to $m_A = m_1 m_2$, i.e. $p(A) \in P(A)$. We conclude

$$\begin{aligned} p(A) &= \begin{pmatrix} p(A_1) & 0 \\ 0 & p(A_2) \end{pmatrix} \\ &= \begin{pmatrix} p_1(A_1) & 0 \\ 0 & p_2(A_2) \end{pmatrix}. \end{aligned}$$

b) The map $\Phi : P(A) \to P(A_1), p(A) \mapsto p(A_1)$ is a well defined and injective group isomorphism, since

$$\begin{aligned} p_1(A) = p_2(A) \quad &\Leftrightarrow \quad p_1 - p_2 \equiv \quad \mod m_A \\ &\Leftrightarrow \quad p_1(A_1) = p_2(A_2). \end{aligned}$$

Moreover, $\Phi$ is surjective. This follows from the fact, that $p(A_1)$ is invertible if and only if $p \not\equiv m_A$ and therefore if and only if $p(A)$ is invertible. Hence, $p(A)$ is the preimage of $p(A_1)$. $\square$

# E Real polynomials

In our analysis of inverse iteration schemes and Richardson's methods we use certain families of real polynomials to represent the corresponding system groups and system semigroups. In particular, we have to deal with *symmetric polynomials* and *linear decomposable polynomials.*

**Definition E.1 (Symmetric polynomials)** A polynomial $f \in \mathbb{R}[u_1, \ldots, u_m]$ is called *symmetric* if, for any permutation $\pi$, we have

$$f(u_{\pi(1)}, \ldots, u_{\pi(m)}) = f(u_1, \ldots, u_m).$$

The *elementary symmetric polynomials* $\sigma_i^m : \mathbb{R}^m \to \mathbb{R}$, $i = 0, \ldots, m$ are defined by

$$\begin{aligned} \sigma_0^m(u_1, \ldots, u_m) &= 1, \\ \sigma_1^m(u_1, \ldots, u_m) &= \sum_{i=1}^m u_i, \\ \sigma_k^m(u_1, \ldots, u_m) &= \sum_{i_1 < \cdots < i_k} u_{i_1} \ldots u_{i_k}. \end{aligned}$$

Note that every symmetric polynomial $f(u_1, \ldots, u_m)$ can be expressed as a polynomial of elementary symmetric polynomials. More precisely,

$$f(u_1, \ldots, u_m) = g\left(\sigma_1^m(u_1, \ldots, u_m), \ldots, \sigma_m^m(u_1, \ldots, u_m)\right)$$

for some $g \in \mathbb{R}[u_1, \ldots, u_m]$. Here, $g$ is unique (see [Pra01], Theorem 3.1.1.). A polynomial $f \in \mathbb{R}[u_1, \ldots, u_m]$ is called *skew-symmetric* if

$$f(\ldots, u_i, \ldots, u_j, \ldots) = -f(\ldots, u_j, \ldots, u_i, \ldots), \quad 1 \leq i < j \leq m$$

Skew symmetric polynomials can be expressed by symmetric polynomials in the following way.

**Theorem E.2** *Every skew symmetric polynomial $f(u_1, \ldots, u_m)$ can be represented in the form*

$$\prod_{i<j}(u_i - u_j) g(u_1, \ldots, u_m)$$

*where $g$ is a symmetric polynomial.*

A proof for Theorem E.2 can be found in [Pra01], Theorem 3.1.2.

Now we introduce a type of real polynomials, which will be of particular interest in in Chapter 6 and Chapter 9.

**Definition E.3 (Linear decomposable polynomials)** A polynomial $q \in \mathbb{R}[x]$ for which every irreducible factor is linear, is called *linear decomposable*. We denote the set of all linear decomposable polynomials with $\mathcal{L}$.

The following useful observations can be found in [Dör55], Chapter 36.

**Theorem E.4** *Let $f$ be linear decomposable of degree $n$.*

*(i)* $f' \in \mathcal{L}$.

*(ii)* *For any $c \in \mathbb{R}$ the polynomial $p_c : t \mapsto cf(t) + tf'(t)$ is linear decomposable.*

Note that every linear decomposable polynomial $q$ can be written in the form

$$q(x) = r \prod_{t=1}^{T} (x - u_t)$$

Every $q \in \mathcal{L}$ is a symmetric polynomial in $u_1, \ldots, u_t$ (for fixed $x$) and can be expressed as follows:

**Proposition E.5** *For all $m \in \mathbb{N}$ and $u_t \in \mathbb{R}$ we have*

$$\prod_{t=1}^{m} (x - u_t) = \sum_{t=0}^{m} (-1)^t \sigma_t^m (u_1, \ldots, u_m) x^{m-t}.$$

Proposition E.5 can be shown by straightforward calculation (see [CLO91], Chapter 7.1).

# F Flag manifolds

Now we introduce some facts about *flag manifolds* which will be important in our analysis of generalized inverse iteration systems. For a more detailed overview we refer to [BC64, HM94] and [Tay92].

Let $H$ be a closed subgroup of a Lie group $G$. Recall that the map

$$\pi : G \to G/H \; g \mapsto gH$$

equips the coset space $G/H := \{gH \,|\, g \in G\}$ with a manifold structure. The map $\pi$ is a surjective submersion and therefore open and continuous. Now let $m \in M$ and $G$ be a Lie group acting transitively on a set $M$ such that

$$\operatorname{Stab}_m := \{g \in G \,|\, g \cdot m = m\}$$

is a closed subgroup [31] of $G$. Then,

$$\Phi_m : G/\operatorname{Stab}_m \to M\, ; g \operatorname{Stab}_m \mapsto g \cdot m$$

is a bijective map. Therefore, we can identify $M$ with the coset space $G/\operatorname{Stab}(m)$. This identification provides a smooth structure on $M$. Such a space $M$ is called *homogeneous space.*

A flag $\mathcal{V}$ is an increasing sequence of $\mathbb{R}$-linear subspaces

$$\{0\} \subsetneqq V_1 \subsetneqq V_2 \subsetneqq \ldots \subsetneqq V_k \subseteq \mathbb{R}^n.$$

The *type* of the flag $\mathcal{V} = (V_1, \ldots, V_k)$ is defined by the $k$-tuple $d := (d_1, \ldots, d_k)$ of dimensions $d_i = \dim V_i, i = 1, \ldots, k$. For any such sequence of integers $d = (d_1, \ldots, d_k)$ with $1 \leq d_1 < \cdots < d_k \leq n$, we denote the set of all flags of type $d$ with $\operatorname{Flag}(d, \mathbb{R}^n)$.

The general linear group $\mathrm{GL}_n(\mathbb{R})$ acts on $\operatorname{Flag}(d, \mathbb{R}^n)$ via

$$\pi_{\mathcal{V}} : (g, \mathcal{V}) \mapsto g \cdot \mathcal{V} := (gV_1, \ldots, gV_k) \tag{74}$$

where $gV_i$ is the image of the space $V_i$ under the transformation $g \in \mathrm{GL}_n(\mathbb{R})$. Here, the stabilizer subgroup for $\mathcal{V} \in \operatorname{Flag}(d, \mathbb{R}^n)$ is

$$\operatorname{Stab}(\mathcal{V}) = \{g \in \mathrm{GL}_n(\mathbb{R}) \,|\, g \cdot \mathcal{V} = \mathcal{V}\}.$$

Now we apply the construction above. For that purpose we need the following fact.

**Lemma F.1** *The group action (74) is transitive.*

[31] Note that $\operatorname{Stab}_{\tilde{m}} = \tilde{g}^{-1} \operatorname{Stab}_m \tilde{g}$ for $\tilde{g} \cdot \tilde{m} = m$. Therefore, $\operatorname{Stab}_m$ is a closed subgroup of $G$ for one $m \in M$ if and only if it is a closed subgroup of $G$ for any $m \in M$.

A proof can be found in [Tay92] (Page 28). In particular, Lemma F.1 yields, that for a fixed flag $\mathcal{V} = (V_1, \dots, V_k)$ the map

$$\Phi_{\mathcal{V}} : \mathrm{GL}_n(\mathbb{R})/\operatorname{Stab}(\mathcal{V}) \to \operatorname{Flag}(d, \mathbb{R}^n),\ g\operatorname{Stab}(\mathcal{V}) \mapsto (gV_1, \dots, gV_k).$$

is bijective and provides a smooth structure on $\operatorname{Flag}(d, \mathbb{R}^n)$. We denote $\operatorname{Flag}(d, \mathbb{R}^n)$ as the *flag manifold of type d*. It is well-known, that $\operatorname{Flag}(d, \mathbb{R}^n)$ is a compact and connected manifold of dimension $d_1(n-d_1)+\sum_{i=1}^{k-1}(d_{i+1}-d_i)(n-d_{i+1})$ (see [BC64], Chapter 7.4.13). One important case is $d_c = (1, 2, \dots, n-1)$. The corresponding manifold $\operatorname{Flag}(\mathbb{R}^n) := \operatorname{Flag}(d_c, \mathbb{R}^n)$ is the so called *complete flag manifold*. Another special case is $d = (k)$ yielding the *Grassmann manifold*, and in particular $\operatorname{Flag}((1), \mathbb{R}^n) = \mathbb{RP}^{n-1}$, the *projective space*.

Recall that the *core* of an homogeneous space is defined as

$$C_M := \bigcap_{m \in M} \operatorname{Stab}_m = \{g \in G \,|\, g \cdot m = m,\ \forall m \in M\}. \tag{75}$$

In the case $M = \operatorname{Flag}(d, \mathbb{R}^n)$ we obtain:

**Proposition F.2** *The core of* $\operatorname{Flag}(d, \mathbb{R}^n)$ *is* $C_{\operatorname{Flag}(d,\mathbb{R}^n)} = \mathbb{R}^* I$. *Here, $I$ is the identity matrix $I \in \mathrm{GL}_n(\mathbb{R})$.*

**Proof.** Obviously, $g \cdot \mathcal{V} = \mathcal{V}$ for all $g \in \mathbb{R}^* I$. Conversely, if $g \notin \mathbb{R}^* I$, then there exists $w \in \mathbb{R}^n$ such that $g(w) \notin \operatorname{span}(w)$. We can always choose $\mathcal{V} \in \operatorname{Flag}(d, \mathbb{R}^n)$ such that $w \in V_1$ but $g(w) \notin V_1$. Hence $g \cdot \mathcal{V} = \mathcal{V}$ is not fulfilled and therefore $g \notin \operatorname{Stab}_{\mathcal{V}} \subseteq C_{\operatorname{Flag}(d,\mathbb{R}^n)}$. □

# Index

# Notation

### Control Systems

| | | |
|---|---|---|
| $\Sigma = (M, U, f)$ | discrete-time control system | Definition 2.1 |
| $\tilde{\Sigma} = (\tilde{M}, U, \tilde{f})$ | induced system with respect to $\pi : M \to \tilde{M}$ | Definition 3.1 |
| $\Sigma_{\mid N} = (N, U, f_{\mid N \times U})$ | restricted system on $N \subseteq M$ | Definition 3.8 |
| $M$ | state space | Definition 2.1 |
| $U$ | set of control parameters | Definition 2.1 |
| $U_A$ | set of control parameters for $\Sigma^{II}(A)$ i.e., $U_A = \mathbb{R} \setminus \mathrm{Spec}(A)$ | Definition 2.1 |
| $f$ | transition map $f : M \times U \to M$ | Definition 2.1 |
| $f_u$ | $f_u := f(\cdot, u)$ | Definition 2.1 |
| $\Sigma^{II}(A)$ | inverse iteration system | Definition 7.1 |
| $\Sigma^{RI}(A)$ | rational iteration system | Definition 8.1 |
| $\Sigma^{CI}(A)$ | Cayley iteration system | Definition 8.4 |
| $\Sigma^{RS}(A)$ | Richardson iteration system | Definition 9.1 |
| $\Sigma^{PI}(A)$ | polynomial iteration system | Definition 9.10 |
| $\Sigma^{B}(A)$ | linear control system with respect to $B$ | Definition 10.1 |

### Semigroups

| | | |
|---|---|---|
| $S_\Sigma$ | system semigroup of $\Sigma$ | Definition2.3 |
| $S(A)$ | system semigroup of $\Sigma^{II}_{\mathrm{GL}_n(\mathbb{R})}(A)$ | Page 84 |

### Groups

| | | |
|---|---|---|
| $\langle S \rangle$ | soubgroup of $G$ generated by $S \subseteq G$ | |
| $G_\Sigma$ | system group of $\Sigma$ | Definition 2.4 |
| $P(A)$ | group of polynomials in $A$ such that $p(A)$ is invertible | Proposition $D.4$ |
| $\mathrm{Stab}_x$ | stabilizer subgroup of a group action $G \times M \to M$ | Page 29 |
| $\mathrm{GL}_n(\mathbb{R})$ | general linear group | |
| $\mathrm{O}_n(\mathbb{R})$ | orthogonal group | Page 151 |
| $\mathrm{Sp}_{2n}(\mathbb{R})$ | symplectic group | Page 151 |
| $\mathrm{G}_M$ | M-orthogonal | Page 151 |
| $C_\pi$ | core of $\pi : M \to \tilde{M}$ | Equation (28) |
| $C_M$ | core of the homogenepous space $M$ | Equation (75) |
| $C_N$ | core of the restricted system $\Sigma_{\mid N}$ | Equation (34) |

# References


[AG93] A.A. Agrachev and R.V. Gamkrelidze. Local controllability for families of diffeomorphisms. *Systems & Control Letters*, 20:67–76, 1993.

[AM86] G.S. Ammar and C.F. Martin. The geometry of matrix eigenvalue methods. *Acta Applicandae Mathematicae*, 5:239–278, 1986.

[AM06] P.J. Antsaklis and A.N. Michel. *Linear Systems.* Birkhäuser Boston, 2006.

[Amm86] G. S. Ammar. Inverse eigenvalue methods and flag manifolds. In C.I. Byrnes and A. Lindquist, editors, *Computational and Combinatorial Methods in Systems Theory*, pages 177 – 184. Elsevier Science Publisher (North-Holland), 1986.

[Amm87] G. S. Ammar. Geometric aspects of Hessenberg matrices. *Contemporary Mathematics*, 68:1–21, 1987.

[AS91] F. Albertini and E. Sontag. Some connections between chaotic dynamical systems and control systems. volume 1, pages 158–163, Grenoble, 1991. European Control Conference.

[AS93] F. Albertini and E. Sontag. Discrete-time transitivity and accessibility: analytic systems. *SIAM J. Control & Opt.*, 31:1599–1622, 1993.

[AS94] F. Albertini and E. Sontag. Further results on controllability properies of discrete-time nonlinear systems. *Dynamics and Control*, (4):235–253, 1994.

[Bat95] S. Batterson. Dynamical analysis of numerical systems. *Numer. Linear Algebra Appl.*, 2:297–310, 1995.

[BC64] R. L. Bishop and R. J. Crittenden. *Geometry of Manifolds.* Pure and Applied Mathematics. Academic Press Inc., London, 1964.

[BCR98] J. Bochnak, M. Coste, and M.-F. Roy. *Real Algebraic Geometry*, volume 36 of *A Series of Modern Surveys in Mathematics.* Springe Verlag, Berlin Heidelberg New York, 1998.

[BK06] A. Bhaya and E. Kaszkurewicz. *Control Perspectives on Numerical Algorithms and Matrix Problems*, volume 10 of *Advances in Design and Control.* SIAM, 2006.

[Bol98] B. Bollobás. *Modern Graph Theory.* Number 184 in Graduate Texts in Mathematics. Springer-Verlag New-York Berlin Heidelberg, 1998.

[BS89a] S. Batterson and J. Smillie. The dynamics of Rayleigh quotient iteration. *SIAM J. Numer. Anal.*, (26):624–636, 1989.

[BS89b] S. Batterson and J. Smillie. Rayleigh quotient iteration fails for nonsymmetric matrices. *Appl. Math. Lett.*, 2:19–20, 1989.

[CC06] D. Chu and M. Chu. Reachable matrices by the QR iteration with shift. *SIAM J. Applied Dynamical Systems*, 5(1):91–107, 2006.

[CK93] F. Colonius and W. Kliemann. Linear control semigroups acting on projective space. *J. Dynamics Diff. Equations*, 5:495–528, 1993.

[CK00] F. Colonius and W. Kliemann. *The Dynamics of Control.* Systems & Control: Foundations & Applications. Birkhaeuser, Boston, Basel, Berlin, 2000.

[CLO91] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms.* Springer Science + Business Media Inc., 1991.

[CR96] D. Calvetti and L. Reichel. An adaptive Richardson iteration method for indefinite linear systems. *Numerical Algorithms*, (12):125 – 149, 1996.

[Die00] R. Diestel. *Graph Theory.* Springer, 2000.

[DK00] J.J. Duistermaat and J.A.C. Kolk. *Lie Groups.* Springer-Verlag Berlin Heidelberg, 2000.

[DP82] W. De Melo and J. Palis. *Geometric Theory of Dynamical Systems.* Springer-Verlag New York Heidelberg Berlin, 1982.

[DS88] F. De Mari and M. A. Shayman. Generalized Eulerian numbers and the topology of the Hessenberg variety of a matrix. *Acta Applicandae Math.*, 12:213–235, 1988.

[Dör55] H. Dörrie. *Praktische Algebra.* Oldenbourg-Verlag, München, 1955.

[EES00] M. Eiermann, O. G. Ernst, and O. Schneider. Analysis of acceleration strategies for restarted minimal residual methods. *J. Comp. Appl. Math.*, 123:261 – 292, 2000.

[Emb03] M. Embree. The tortoise and the hare restart GMRES. *SIAM Review*, pages 1074–1109, 2003.

[FM67] G. E. Forsythe and C. M. Moler. *Computer Solution of Linear Algebraic Systems*, volume 2 of *Series in Automatic Computation.* Prentice-Hall, 1967.

[FN81a] M. Fliess and D. Normand-Cyrot. A group-theoretic approach to discrete-time non-linear controllability. IEEE, December 1981.

[FN81b] M. Fliess and D. Normand-Cyrot. A Lie-theoretic approach to nonlinear discrete-time controllability via Ritt's formal differential groups. *Systems and control letters*, 1(3):179–183, 1981.

[FS07] M.A. Freitag and A. Spencer. Convergence of inexact inverse iteration with application to preconditioned iterative solvers. *BIT Numerical Mathematics*, 47:27–44, 2007.

[Fuh96] P. A. Fuhrmann. *A Polynomial Approach to Linear Algebra.* Springer Publ., New York, 1996.

[GLS88] K. Gustafsson, M. Lundh, and G.S. Söderlind. A PI stepsize control for the numerical solution of ordinary differential equations. *BIT*, (28):270–287, 1988.

[GO88] G.H. Golub and M.L. Overton. The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems. *Mumer. Math.*, 53:571 – 593, 1988.

[GOV97] V.V. Gorbatsevich, A.L. Onishchik, and E.B. Vinberg. *Foundations of Lie Theory and Lie Transformation Groups.* Springer, 1997.

[Gre97] A. Greenbaum. *Iterative Methods for Solving Linear Systems*, volume 17 of *Frontiers in Applied Mathematics.* SIAM, 1997.

[Gus91] K. Gustafsson. Control theoretic techniques for stepsize selection in explicit Runge-Kutta methods. *ACM Trans. Math. Softw.*, 17(4):533–554, 1991.

[Gus92] K. Gustafsson. *Control of error and convergence in ODE solvers.* PhD thesis, Lund Institute of Technology, 1992.

[GY00] G.H. Golub and Q. Ye. Inexact inverse iteration for generalized eigenvalue problems. *BIT*, 40(4):671 – 684, 2000.

[HF00] U. Helmke and P. A. Fuhrmann. Controllability of matrix eigenvalue algorithms: the inverse power method. *Systems and Control Letters*, 41:57–66, 2000.

[HHL89] K.H. Hofmann, J. Hilgert, and J.D. Lawson. *Lie Groups, Convex Cones and Semigroups.* Oxford Mathematical Monographs. Clarendon Press, Oxford, 1989.

[HJ02] U. Helmke and J. Jordan. Numerics versus control. In *Mathematical Systems Theory in Biology, Communications, Computations and Finance*, volume 134 of *IMA Conference Volume*, pages 223–236. Springer-Verlag New York, Inc, 2002.

[HJ05] U. Helmke and J. Jordan. Optimal control of iterative solution methods for linear systems of equations. In *Proc. Appl. Math. Mech.*, volume 5 of *PAMM*, pages 163–164, 2005.

[HJL06] U. Helmke, J. Jordan, and A. Lanzon. A control theory approach to linear equation solvers. Kyoto, Japan, 2006. Seventeenth International Symposium on Mathematical Theory of Network and Systems. Proceeding CD-ROM.

[HM94] U. Helmke and J.B. Moore. *Optimization and Dynamical Systems.* Communication and Control Engineering Series. Springer-Verlag, Berlin, 1994.

[HN91] J. Hilgert and K.-H. Neeb. *Lie-Gruppen und Lie-Algebren.* Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1991.

[HN93] J. Hilgert and J. Neeb. *Lie Semigroups and Their Applications.* lecture Notes in Mathematics. Springer-Verlag, Berlin, 1993.

[Hom93] P. Homblé. Ergodic conditions for nonlinear discrete time stochastic dynamical systems with Markovian noise. *Stochastic Analysis and applications*, 11(5):513 – 568, 1993.

[Hus66] T. Husain. *Introduction to topological groups.* Saunders mathematics books, 1966.

[HW01] U. Helmke and F. Wirth. On controllability of the real shifted inverse power iteration. *Systems and Control Letters*, 43:9–23, 2001.

[Ips96] I. C. F. Ipsen. A history of inverse iteration. In B. Huppert and H. Schneider, editors, *Helmut Wieland, Mathematische Werke,* *Mathematical Works*, volume 2, pages 464 – 472. Walter der Gruyter, Berlin, 1996.

[Ips97] I. C. F. Ipsen. Computing an eigenvector with inverse iteration. *SIAM Reviews*, 39:254–291, 1997.

[Jor06] J. Jordan. Discrete-time control systems on homogeneous spaces: Partition property. *Control and Cybernetics*, 35(4):863–871, 2006.

[Jou94] W. Joubert. On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems. *Numerical Linear Algebra with Apllications*, 1(5):427 – 447, 1994.

[JS72] V. Jurdjevic and H.J. Sussmann. Control systems on Lie groups. *J. Diff. Eqs.*, 12:313–329, 1972.

[JS90] B. Jakubczyk and E. D. Sontag. Controllability of nonlinear discrete-time systems: A Lie-algebraic approach. *SIAM J. Control and Opt.*, 28:1–33, 1990.

[Jur97] V. Jurdjevic. *Geometric Control Theory.* Number 51 in Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1997.

[JV05] E. Jarlebring and H. Voss. Rational Krylov for nonlinear eigenproblems, an iterative projection method. *Appl. Math.*, 50:543–554, 2005.

[Kai80] T. Kailath. *Linear Systems.* Prentice-Hall Information and Systems Science Series. Englewood Cliffs, 1980.

[Kal60] R.E. Kalman. Contributions to the theory of optimal control. *Bull. Soc. Math. Mex.*, 5:102 – 119, 1960.

[KM83] P.A. Krishnaprasad and C.F. Martin. On families of systems and deformations. *Int. J. Control*, 38(5):1055 – 1079, 1983.

[Kre74] A.J. Krener. A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems. *SIAM J. Control*, 12(1), 1974.

[Kuc79] V. Kucera. *Discrete Linear Control.* Academia Prague, 1979.

[Kup90] I. Kupka. Applications of semigroups to geometric control theory. In Pym Hofmann, Lawson, editor, *The analytical and Topological Theory of Semigroups.* De Gruyter, 1990.

[LM98] R.B. Lehoucq and K. Meerbergen. Using generalized Cayley transformations within an inexact rational krylov sequence method. *Siam J. Matrix Anal. Appl.*, 20(1):131–148, 1998.

[LR95] P. Lancester and L. Rodman. *Algebraic Riccati Equation.* Oxford Science Publications, 1995.

[Mei99] A. Meister. *Numerik linearer Gleichungssysteme.* Vieweg, 1999.

[Mit95] D. Mittenhuber. Applications of the maximum principle to problems in lie semigroups. In Vinberg Hofmann, Lawson, editor, *Semigroups in Algebra, Geometry and Analysis*, pages 313–337. De Gruyter, 1995.

[Mit01] D. Mittenhuber. Transitive semigroup actions and controllability of systems on Lie groups: A solvable and a semisimple problem. *Habilitationsschrift*, 2001.

[Mok89] A. Mokkadem. Orbites de semi-groupes de morphismes réguliers et systèmes non linéaires en temps discret. *Forum Math.*, 1:359–376, 1989.

[Mok95] A. Mokkadem. Orbit theorems for semigroup of regular morphisms and nonlinear discrete time systems. *Bull. Soc. math. France*, 123:477–491, 1995.

[MSR94] K. Meerbergen, A. Spence, and D. Roose. Shift-invert and Cayley transforms for the detection of eigenvalues wi th largest real part of nonsymmetric matrices. *BIT Numerical Mathematics*, 34(3):409 – 423, 1994.

[Ney01] K. Neymeyr. A geometric theory for preconditioned inverse iteration I: Extrema of the Rayleigh quotient. *Linear Algebra and its Applications*, 322:61–85, 2001.

[Ney05] K. Neymeyr. A note on inverse iteration. *Numerical Linear Algebra with Applications*, 12(1):1 – 8, 2005.

[OS84] G. Opfer and G. Schober. Richardson's iteration for nonsymmetric matrices. *Linear Algebra Appl.*, 58(343-361), 1984.

[PK69] B.N. Parlett and W. Kahan. On the convergence of a practical QR algorithm. *Information Processing 68*, pages 114–118, 1969. Amsterdam, North Holland.

[Pra01] V.V. Prasolov. *Polynomials.* Springer-Verlag Berlin Heidelberg New York, 2001.

[Ros94] J. Rosenthal. On dynamic feedback compensation and compactifications of systems. *SIAM J. Control Opt.*, 32(1):279–296, 1994.

[Ruh84] A. Ruhe. Rational Krylov sequence methods for eigenvalue computaion. *Linear Algebra Appl.*, 58:391–405, 1984.

[San95] L.A.B. San Martin. On global controllability of discrete-time control systems. *Mathematics of Control*, 8:279 – 297, 1995.

[SBG$^+$95] M. Salzmann, D. Betten, T. Grundhöfer, R. Löwen H. Hähl, and M. Stroppel. *Compact projective planes*. Number 21 in De Gruyter Expositions in Mathematics. Walter de Gruyter, Berlin - New York, 1995.

[SE02] V. Simoncini and L. Eldén. Inexact Rayleigh quotient-type methods for eigenvalue computations. *BIT Numerical Mathematics*, 42(1):159–182, 2002.

[SJ72] H.J. Sussmann and V. Jurdjevic. Controllability of nonlinear systems. *Journal of Differential Equations*, 12:95–116, 1972.

[Son86] E. D. Sontag. Orbit theorems and sampling. In M. Fliess and M. Hazewinkel, editors, *Algebraic and Geometric Methods in Nonlinear Control Theory*, pages 441–486. Dordrecht, 1986.

[Son98] E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Number 6 in Texts in Applied mathematics. Springer-Verlag New York, second edition, 1998.

[Sor02] D.C. Sorensen. Numerical methods for large eigenvalue problems. *Acta Numerica*, pages 519–584, 2002.

[SS88] D.C. Smolarski and P.E. Saylor. An optimum iterative method for solving any linear systems with a square matrix. *BIT*, 28:163 – 178, 1988.

[SV87] M. Shub and T. Vasquez. Some linearly induced Morse-Smale systems, the QR algorithm and the Toda lattice. In L. Keen, editor, *The Legacy of Sonya Kovalevskaya*, volume 64 of *Contemporary Mathematics*, pages 181–194. A.M.S., 1987.

[SW98] E. D. Sontag and F. R. Wirth. Remarks on universal nonsingular controls for discrete-time systems. *Systems and Control Letters*, 33:81–88, 1998.

[Tay92] D. E. Taylor. *The Geometry of the Classical Groups*, volume 9 of *Sigma Series in Pure Mathematics*. Heldermann Verlag, 1992.

[Wat82] D. Watkins. Understanding the QR algorithm. *SIAM Reviews*, 24:427–440, 1982.

[Wie44] H. Wielandt. Beiträge zur Behandlung komplexer Eigenwertprobleme Teil V: Bestimmung höhere Eigenwerte durch gebrochene Iteration. Technical Report Bericht B44/J/37, Aerodynamische Versuchsanstalt Göttingen, 1944.

[Wir95] F. Wirth. *Robust Stability of Discrete-Time Systems under Time-Varying Pertubations.* PhD thesis, Univ. Bremen., 1995.

[Wir98] F. Wirth. Dynamics and controllability of nonlinear discrete-time control systems. pages 269–275, Enschede, The Netherlands, 1998. 4th IFAC Nonlinear Control Systems Design Symposium (NOLCOS'98).

[YV92] D. M. Young and B. R. Vono. On the use of rational iterative methods for solving large sparse linear systems. *Appl. Numer. Math.*, 10(3-4), 1992.