Mathematical System Theory

Festschrift in Honor of Uwe Helmke on the Occasion of his Sixtieth Birthday

Knut Hüper and Jochen Trumpf
Editors

Editors

Knut Hüper
Julius-Maximilians-Universität
Würzburg, Germany

Jochen Trumpf
Australian National University
Canberra, Australia

ISBN 978-1470044008
Printed by CreateSpace

Contents

Uwe Helmke

Uwe Helmke was born in 1952 in Bremen, Germany. He studied Mathematics at the University of Bremen where he obtained his Ph.D. (Dr. rer. nat.) in 1983 under the supervision of D. Hinrichsen. The title of his Ph.D. thesis was *Zur Topologie des Raumes linearer Kontrollsysteme* (On the topology of the space of linear control systems). At that time his main interest was focussed on developing algebro-topological methods for linear system theory. A short intermediate period followed, including a research visit to the Division of Applied Sciences, Harvard University to work with C. I. Byrnes.

Uwe's increasing interest in real algebraic geometry lead him to accept a post doctoral position (Akademischer Rat) with the University of Regensburg, Germany. In Regensburg he continued to work on algebraic aspects of mathematical system theory, including such diverse topics as partial realizations, normal forms for linear systems, output feedback stabilization and algebraic invariants for output feedback, the cohomology of moduli spaces for linear systems, and eigenvalue inequalities, to mention just a few. In 1991, Uwe completed his Habilitation with a thesis on *The cohomology of moduli spaces of linear dynamical systems*, and together with obtaining his Venia Legendi (permission to lecture) he was appointed Privatdozent (private lecturer) at the University of Regensburg.

In the following four years Uwe enriched his interests in the direction of more applied mathematics. Between 1991 and 1994 he repeatedly visited The Australian National University in Canberra, Australia, to work with B. D. O. Anderson and J. B. Moore as a visiting fellow. Among other things he became interested in adaptive control and the interplay between numerical aspects of linear algebra and control theory. Uwe's contributions in this area culminated in the monograph *Optimization and Dynamical Systems*, coauthored by J. B. Moore, a book that had major impact across various disciplines in applied mathematics and engineering, popularizing the use of gradient flows and Lie group actions on smooth manifolds in the design of practical optimization algorithms. Within the realm of system theory, the book exploits the intimate relations between gradient flows, completely integrable systems and numerical linear algebra to tackle difficult problems that arise as part of sensitivity analysis. As a byproduct, further connections to isospectral flows, neural networks and balancing were discovered.

Another of Uwe's long standing interests lies in the foundations of algebraic system theory, documented by many joint papers with P. A. Fuhrmann that are covering a wide range of topics in the area, including the parametrization of controlled and conditioned invariant subspaces, observer theory, Bezoutians, and polynomial, rational and tensored models. This line of work culminates in the forthcoming book *Mathematics of Networks of Systems*, that is currently in use as the basis for a masters course at the Julius-Maximilians-Universität Würzburg, Germany, where Uwe has been a full professor and chair holder of the *Lehrstuhl für Mathematik II, Dynamische Systeme und Kontrolltheorie* (Chair for mathematics II, dynamical systems and control theory) since 1995.

Over the years, Uwe continued to reach out across disciplinary boundaries. Together with his coworkers he made important contributions to quantum control, in particular to nuclear magnetic resonance (NMR) applications; to robotics and computer vision; and to Lie-theoretic generalizations of numerical linear algebra algorithms. More recently he became interested in system identification for biological systems and the structure of complex networked systems.

During his time in Würzburg, Uwe has always been active in the self-administration of the faculty. He was one of the driving forces behind the establishment of new study programs such as Mathematical Physics. Between 2010 and 2012 he acted as Dean of the Fakultät für Mathematik und Informatik (Faculty for mathematics and computer science) and more recently, he founded the *Interdisziplinäres Forschungszentrum für Mathematik in Naturwissenschaft und Technik (IFM)* (Center for interdisciplinary research in mathematics, science and technology) at the University of Würzburg. The interdisciplinary center strives to further the collaboration between different scientific faculties and local industry. Uwe currently serves as the inaugural director of IFM.

Counter-acting the proverbial image of the unworldly and unorganized mathematician, Uwe has been very active in the (co-)organization of international workshops, special sessions and conferences across several scientific communities (IEEE, SIAM, MTNS). For many years he has been a member of the MTNS steering commitee.

Uwe's interest in the intersection between mathematical system theory on one side and certain topics in physics, systems biology or electrical engineering on the other side has always been driven by his deep belief that system theory can always contribute something new, possibly better, and maybe even more efficient, if correctly applied. With admiration and affection we honor a colleague, friend and teacher of enormous creativity, energy, mathematical strength and broadness on the occasion of his sixtieth birthday.

The editors would like to thank everyone who contributed to this book. Those who have written chapters did so under tight deadlines and with good grace. The breadth of contributions in this Festschrift reflects Uwe's broad scientific interests and the enormous extent of his international recognition.

K. Hüper and J. Trumpf
February 2013

Ph.D. Students of Uwe Helmke

G. Dirr	*Differentialgleichungen in Frécheträumen*	2001
J. Trumpf	*On the geometry and parametrization of almost invariant subspaces and observer theory*	2003
M. Kleinsteuber	*Jacobi-type methods on semisimple Lie algebras: A Lie algebraic approach to numerical linear algebra*	2006
C. Lageman	*Convergence of gradient-like dynamical systems and optimization algorithms*	2007
M. Baumann	*Newton's method for path-following problems on manifolds*	2008
J. Jordan	*Reachable sets of numerical iteration schemes: A system semigroup approach*	2008
I. Kurniawan	*Controllability aspects of the Lindblad-Kossakowski master equation: A Lie-theoretical approach*	2010
M. Schröter	*Newton methods for image registration*	2012
M. Mauder	*Time-optimal control of the bi-steerable robot: A case study in optimal control of nonholonomic systems*	2013
O. Curtef	*Rayleigh-quotient optimization on tensor products of Grassmannians*	2013
F. Rüppel	in progress	

A differential-geometric look at the Jacobi–Davidson framework

Pierre-Antoine Absil
Université catholique de Louvain
Belgium
`sites.uclouvain.be/absil/`

Michiel E. Hochstenbach
TU Eindhoven
The Netherlands
`www.win.tue.nl/~hochsten/`

Abstract. The problem of computing a p-dimensional invariant subspace of a symmetric positive-definite matrix pencil of dimension n is interpreted as computing a zero of a tangent vector field on the Grassmann manifold of p-planes in $\mathbb{R}^n$. The theory of Newton's method on manifolds is applied to this problem, and the resulting Newton equations are interpreted as block versions of the Jacobi–Davidson correction equation for the generalized eigenvalue problem.

1 Introduction

The Jacobi–Davidson method (JD) [32] is a method to compute certain eigenpairs of standard or generalized eigenvalue problems. JD has been particularly successful for standard eigenproblems where interior eigenvalues are sought, and for generalized types of eigenproblems. JD belongs to the the class of subspace methods, where low-dimensional subspaces are exploited to find approximations to sought eigenvectors. In line with many other subspace methods, JD has two main stages:

(i) The subspace extraction, where approximate eigenpairs are determined from a given search space. This is often done by the Rayleigh–Ritz method, or variants such as the harmonic Rayleigh–Ritz approach (see, e.g., [36]).

(ii) The subspace expansion, where the search space is expanded with an (inexact) solution to the so-called correction equation.

We refer to [20] for a recent overview of several aspects of the JD method; see also [22].

In [33, §6], JD for the standard eigenvalue problem is interpreted as a Newton method. The interpretation is readily extended to the generalized eigenvalue problem as follows. Let (A,B) be a symmetric positive-definite matrix pencil; we refer to Section 3 for the necessary background on the generalized eigenvalue problem. We are interested in an eigenvector y of (A,B). Let $\widetilde{u},w \in \mathbb{R}^n$ be fixed for the time being. In order to remove the scale indeterminacy of eigenvectors, we impose the normalization $\widetilde{u}^\top y = 1$. Consider the function defined for all $u \in \{u : \widetilde{u}^\top u = 1\}$ by

$$F(u) = Au - \theta Bu \quad \text{with} \quad \theta = \theta(u) = \frac{w^\top Au}{w^\top Bu},$$

where we assume that $w^\top Bu \neq 0$. Function F maps the hyperplane $\{u : \widetilde{u}^\top u = 1\}$ to the hyperplane $w^\perp$. Observe that u with $\widetilde{u}^\top u = 1$ is an eigenvector of (A,B) if and only if $F(u) = 0$. If u is the current approximation, then the next Newton iterate for F is $u+s$, where $s \perp \widetilde{u}$ satisfies

$$(\mathrm{D}F(u))s = -F(u). \tag{1}$$

It may be checked that the Jacobian of F acting on $\widetilde{u}^{\perp}$ is given by

$$(\mathrm{D}F(u))s = \left(I - \frac{Buw^{\top}}{w^{\top}Bu}\right)(A - \theta(u)B)s \quad \text{for} \quad s \perp \widetilde{u},$$

hence the Newton equation (1) reads

$$\left(I - \frac{Buw^{\top}}{w^{\top}Bu}\right)(A - \theta(u)B)s = -(A - \theta(u)B)u \quad \text{for} \quad s \perp \widetilde{u}. \tag{2}$$

This Newton process converges locally quadratically to an eigenvector y with $\widetilde{u}^{\top} y = 1$, for fixed $\widetilde{u}$ and w, provided that $w^{\top}By \neq 0$. However, the adaptive choice $\widetilde{u} = u$ and $w = u$ also leads to locally quadratic convergence. For this choice, the Newton correction equation (2) is precisely the correction equation that appears in JD.

In [26], a block Newton method was given for the standard eigenvalue problem, and a connection was made with JD in the appendix. Expected advantages of a block method are better robustness or efficiency in the presence of clustered eigenvalues, as well as exploiting higher-level BLAS; see, e.g., the discussion in [12, §5.1.4], where a block JD for the generalized eigenvalue problem is outlined. In particular, if the desired eigenvalues are multiple or clustered, then difficulties can arise in the (nonblock) Jacobi–Davidson method, because the Jacobi correction equation (1) becomes ill conditioned. Resorting to a block version allows one to "split" the spectrum at a wider eigenvalue gap.

In this paper, we obtain a class of block Jacobi correction equations for the generalized eigenvalue problem. Our approach consists in characterizing the p-dimensional invariant subspaces of (A,B) as the zeros of a tangent vector field on the Grassmann manifold of p-planes in $\mathbb{R}^n$. The Grassmann manifold is described as a quotient of the set $\mathbb{R}_*^{n\times p}$ of all $n \times p$ matrices of full (column) rank, where the equivalences classes gather all matrices that have the same column space. By applying Shub's manifold-based Newton method [30, §3] to the tangent vector field, and by exploiting the leeway offered by the quotient geometry framework, we obtain a whole class of block Jacobi correction equations for the generalized eigenvalue problem.

The paper is organized as follows. Section 2 gives a brief overview of algorithms on manifolds in connection with the eigenvalue problem. The generalized eigenvalue problem is described in Section 3. The tangent vector field on the Grassmann manifold is introduced in Section 4. The geometric Newton method for the vector field is worked out in Section 5. Connections with JD are established in Section 6.

The forthcoming developments will make use of differential-geometric objects on the Grassmann manifold viewed as a quotient of $\mathbb{R}_*^{n\times p}$. However, the necessary differential-geometric concepts are quite limited, and this paper is meant to be accessible without any background in differential geometry.

2 Algorithms on manifolds and eigenvalue problems

The best known field of application of differential geometry is probably relativity theory. More surprisingly perhaps, techniques of differential and Riemannian geometry have found applications in several areas of science and engineering (such as crystallography [25], thermodynamics [10], and information theory [6, 35]), and in particular, they have been utilized to design and analyze eigenvalue algorithms. Numerous papers belong to this line of research, including [1, 3, 7–9, 11, 15–19, 21, 23, 24, 26, 27, 29, 30, 34].

That the Jacobi–Davidson approach, including a block version thereof, is closely related to Newton's method on Grassmann manifolds, was pointed out in [26]. Since then, the area of numerical computations on manifolds has made progress in several directions, some of which will be exploited in this work. Whereas the seminal papers [11, 34] were systematically making use of the Riemannian connection and the Riemannian exponential, more recent papers [1, 4, 5, 28] have relied on the concepts of *retraction* and of *locally smooth family of parameterizations* to relax the Riemannian exponential into a broader class of mappings that offer opportunities to reduce the numerical burden while preserving convergence properties. Likewise, the Newton method on manifolds stated and analyzed in [4, 30] allows for using of any affine connection, instead of restricting to the Riemannian one as in [11, 34]. As a consequence, the Newton method on Riemannian manifolds stated in [34] has turned, without losing its quadratic local convergence, into a class of geometric Newton methods which vary according to the choice of an affine connection and of a retraction.

3 The generalized eigenvalue problem: Notation and assumptions

Given two $n \times n$ matrices A and B, we say that $\lambda \in \mathbb{C}$ is an *eigenvalue*, that $u \in \mathbb{R}^n \setminus \{0\}$ is an *eigenvector*, and that (λ, u) is an *eigenpair* of the *pencil* (A, B) if

$$Ax = \lambda Bx.$$

Finding eigenpairs of a matrix pencil is known as the *generalized eigenvalue problem*. From now on, we assume that A is symmetric and B is symmetric positive-definite (i.e., $x^\top Bx > 0$ for all $x \neq 0$); the pencil (A, B) and the associated generalized eigenvalue problem are then termed *symmetric positive-definite*, abbreviated *S/PD* [36, §4.1]. It follows that the eigenvalues of the pencil are all real and the eigenvectors can be chosen to form a B-orthonormal basis. A subspace $\mathcal{Y}$ is a *(generalized) invariant subspace* (or a *deflating subspace*) [14, §7.7.8] of the S/PD pencil (A, B) if $B^{-1}Ay \in \mathcal{Y}$ for all $y \in \mathcal{Y}$; this can also be written $B^{-1}A\mathcal{Y} \subseteq \mathcal{Y}$ or $A\mathcal{Y} \subseteq B\mathcal{Y}$. It is readily seen that $\mathcal{Y}$ is a one-dimensional invariant subspace of (A, B) if and only if $\mathcal{Y}$ is spanned by an eigenvector of (A, B). More generally, every invariant subspace of an S/PD pencil is spanned by eigenvectors of (A, B). Clearly, the generalized eigenvalue problem reduces to the standard eigenvalue problem when $B = I$.

Given an integer $1 \leq p \leq n$, we let $\mathbb{R}_*^{n \times p}$ denote the set of all $n \times p$ matrices of full (column) rank, and we let $\mathsf{col}(Y)$, termed the *column space* of Y, denote the

p-dimensional subspace of $\mathbb{R}^n$ spanned by the columns of $Y \in \mathbb{R}_*^{n\times p}$, i.e.,

$$\mathsf{col}(Y) = \{Yw : w \in \mathbb{R}^p\}.$$

The set of all matrices $\hat{Y}$ such that $\mathsf{col}(\hat{Y}) = \mathsf{col}(Y)$ is

$$[Y] := Y\,\mathrm{GL}_p := \{YM : M \in \mathrm{GL}_p\}, \tag{3}$$

where

$$\mathrm{GL}_p := \{M \in \mathbb{R}^{p\times p}\}$$

denotes the set of all $p \times p$ invertible matrices. Observe that $\mathcal{Y}$ is a p-dimensional invariant subspace of (A,B) if and only if there is $Y \in \mathbb{R}_*^{n\times p}$ with $\mathcal{Y} = \mathsf{col}(Y)$ such that

$$AY = BYM \tag{4}$$

for some $p \times p$ matrix M.

The *multiplicity* of an eigenvalue λ of (A,B) is its multiplicity as a root of the polynomial $\det(A - \lambda B)$. An invariant subspace $\mathsf{col}(Y)$ of (A,B) is termed *simple* [36] or *spectral* [13] if the multiplicity of the eigenvalues of M is the same as their multiplicity as eigenvalues of (A,B).

We will let $U \mapsto \widetilde{U}_U$ denote any function on $\mathbb{R}_*^{n\times p}$ into $\mathbb{R}_*^{n\times p}$ that satisfies the following two properties. (i) $\mathsf{col}(\widetilde{U}_U)$ only depends on $\mathsf{col}(U)$, i.e., for all $M \in \mathrm{GL}_p$, there is $N \in \mathrm{GL}_p$ such that $\widetilde{U}_{UM} = \widetilde{U}_U N$. For example, the choice $\widetilde{U} = CU$ for a fixed C is adequate. (ii) For all $U \in \mathbb{R}_*^{n\times p}$, $\widetilde{U}_U^\top U$ and $\widetilde{U}_U^\top BU$ are invertible. The motivation for imposing invertibility of $\widetilde{U}_U^\top BU$ will already become clear in Theorem 1. The other assumptions will be instrumental in the differential-geometric approach laid out below.

Finally, we will let

$$P_{E,F} := I - E(F^\top E)^{-1} F^\top \tag{5}$$

denote the projector along the column space of E into the orthogonal complement of the column space of F.

4 Invariant subspaces as zeros of a vector field

We are looking for an iteration function

$$g : \mathbb{R}_*^{n\times p} \to \mathbb{R}_*^{n\times p} \tag{6}$$

such that the sequences of iterates $\mathsf{col}(U_k)$ generated by $U_{k+1} = g(U_k)$ converge locally quadratically to the p-dimensional spectral invariant subspaces of (A,B). The quadratic convergence requirement leads us naturally to Newton-type methods.

Moreover, since we are interested in the sequence of subspaces $\mathsf{col}(U_k)$ rather than in the sequence of p-frames U_k, it makes sense to require that g and col commute; in other words, there must be a function G such that $G(\mathsf{col}(U)) = \mathsf{col}(g(U))$ for all $U \in \mathbb{R}_*^{n\times p}$. The domain and the codomain of G are the set $\mathrm{Grass}(p,n)$ of all p-dimensional subspaces in $\mathbb{R}^n$. This set admits a natural manifold structure, as explained in [18, §C.4], and endowed with this structure, it is called a *Grassmann manifold.* The sought Newton-type method will thus be an iteration on the Grassmann manifold $\mathrm{Grass}(p,n)$.

To this end, we will pursue the following strategy. In this section, we will express the problem of computing a p-dimensional invariant subspace of (A,B) as finding a zero of a particular vector field on $\mathrm{Grass}(p,n)$. Then, in Section 5, we will work out Shub's Newton method for this vector field.

The characterization of the p-dimensional invariant subspaces of (A,B) as the zeros of a vector field on $\mathrm{Grass}(p,n)$ relies on the following result.

Theorem 1. *Let $U \in \mathbb{R}_*^{n\times p}$. Under the assumptions of Section 3, $\mathrm{col}(U)$ is an invariant subspace of (A,B) if and only if*

$$AU - BU(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U AU = 0. \tag{7}$$

Proof. The "if" part is direct in view of (4). For the "only if" part, assume that $\mathrm{col}(U)$ is an invariant subspace of (A,B). Then, in view of (4), there is a matrix M such that $AU = BUM$. Multiplying this equation on the left by $\widetilde{U}_U^\top$ yields that $M = (\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U AU$, hence the claim. □

The rest of this subsection is dedicated to showing that the mapping $U \mapsto AU - BU(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U AU$ that appears in (7) represents a vector field on the set of all p-dimensional subspaces of $\mathbb{R}^n$.

The Grassmann manifold $\mathrm{Grass}(p,n)$ can be viewed as the manifold of rank-p symmetric projection operators of $\mathbb{R}^n$; see [16] for details in the context of Newton's method. It can also be viewed as a homogeneous space for the orthogonal group $\mathrm{O}(n)$; see [11]. In the context of this paper, since elements of $\mathrm{Grass}(p,n)$ are represented as column spaces of elements of $\mathbb{R}_*^{n\times p}$, we find it more convenient to rely on the identification of $\mathrm{Grass}(p,n)$ with the quotient space

$$\mathbb{R}_*^{n\times p}/\mathrm{GL}_p = \{[Y] : Y \in \mathbb{R}_*^{n\times p}\},$$

where $[\cdot]$ is as defined in (3). The one-to-one correspondence between $\mathrm{Grass}(p,n)$ and $\mathbb{R}_*^{n\times p}/\mathrm{GL}_p$ is given by $\mathrm{col}(Y) \leftrightarrow [Y]$. This identification was mentioned in [18, §C.4] and further exploited in [3, 4]. In view of the identification $\mathrm{Grass}(p,n) \simeq \mathbb{R}_*^{n\times p}/\mathrm{GL}_p$, the sought Newton-like iteration G can thus be viewed as an iteration on $\mathbb{R}_*^{n\times p}/\mathrm{GL}_p$.

We now particularize to $\mathbb{R}_*^{n\times p}/\mathrm{GL}_p$ the framework presented, e.g., in [4, §3.5.8] that allows to represent tangent vectors to $\mathbb{R}_*^{n\times p}/\mathrm{GL}_p$ as $n \times p$ matrices by means of so-called horizontal lifts. The difference with the Grassmann-specific developments in [4, Example 3.6.4] is that we depart from the framework of Riemannian submersions, where the horizontal space is constrained to be the orthogonal complement of the vertical space. This additional freedom allows us to obtain a wider class of Newton equations.

For each $U \in \mathbb{R}_*^{n\times p}$, the *vertical space* $\mathcal{V}_U$ is the tangent space to $[U]$ at U. We have

$$\mathcal{V}_U = \{UM : M \in \mathbb{R}^{p\times p}\}.$$

Intuitively, the vertical space $\mathcal{V}_U$ consists of all the elementary variations of U that preserve the column space.

We choose the *horizontal space* $\mathcal{H}_U$ as the set of all $n \times p$ matrices whose columns are orthogonal to the columns of $\widetilde{U}_U$, i.e.

$$\mathcal{H}_U := \{Z \in \mathbb{R}^{n\times p} : \widetilde{U}_U^\top Z = 0\}. \tag{8}$$

The horizontal space is required to be transverse to the vertical space, i.e., $\mathcal{H}_U \cap \mathcal{V}_U = \{0\}$; this is equivalent to the condition that $U^\top \widetilde{U}_U$ be invertible, which is a standing assumption (see Section 3). Moreover, we require the compatibility condition that $\mathcal{H}_{UM} = \mathcal{H}_U$ for all $M \in \mathrm{GL}_p$; this is ensured by the standing assumption that $\mathrm{col}(\widetilde{U}_U)$ only depends on $\mathrm{col}(U)$ (see Section 3).

The purpose of the horizontal space is to provide a unique matrix representation of elementary variations of p-dimensional subspaces of R^n. Given an elementary variation $\xi_{\mathcal{U}}$ of the column space $\mathcal{U}$ of U, there is in $\mathcal{H}_U$ one and only one elementary variation $\overline{\xi}_U$ of U that has the same effect as $\xi_{\mathcal{U}}$, in the sense that, for all real-valued functions f on Grass(p,n), it holds that $\mathrm{D}(f \circ \mathrm{col})(U)[\overline{\xi}_U] = \mathrm{D}f(\mathrm{col}(U))[\xi_{\mathcal{U}}]$. In the parlance of differential geometry, $\xi_{\mathcal{U}}$ is a *tangent vector* to Grass(p,n) at $\mathcal{U}$, and $\overline{\xi}_U$ is the *horizontal lift* of $\xi_{\mathcal{U}}$ at U. The set of all tangent vectors to Grass(p,n) at $\mathcal{U}$ is called the *tangent space* to Grass(p,n) at $\mathcal{U}$ and denoted by $T_{\mathcal{U}}$Grass(p,n). Observe that $\overline{\xi}_U$, in spite of its somewhat unusual notation and its differential geometric origin, is nothing else than an $n \times p$ real matrix. It can be shown (see [4, Proposition 3.6.1]) that the horizontal lifts at different points U and UM of a same equivalence class $[U]$ satisfy the relation

$$\overline{\xi}_{UM} = \overline{\xi}_U M \tag{9}$$

for all $M \in \mathrm{GL}_p$. And any vector field $\mathbb{R}_*^{n\times p} \ni U \mapsto \overline{\xi}_U \in \mathbb{R}^{n\times p}$ that satisfies (9) is a bona-fide horizontal lift.

Now, returning to the generalized eigenvalue problem for the S/PD pencil (A,B), consider

$$\overline{\xi}_U := P_{BU,\widetilde{U}_U} AU, \tag{10}$$

where $P_{BU,\widetilde{U}_U} = I - BU(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U^\top$ in keeping with the notation introduced in (5). Recall the standing assumption that $\widetilde{U}_U^\top BU$ is invertible; hence the right-hand side of (10) is well defined. Moreover, it is readily checked that $\overline{\xi}_U$ of (10) satisfies (9). Thus $\overline{\xi}_U$, being a horizontal lift, defines a vector field ξ on Grass(p,n). In view of Theorem 1, searching for a p-dimensional invariant subspace of $B^{-1}A$ amounts to searching for a zero of the vector field ξ on Grass(p,n) defined by the horizontal lift (10).

5 Geometric Newton for invariant subspace computation

We now work out the geometric Newton equation for the vector field ξ on Grass(p,n) defined in the previous section.

The geometric Newton method for computing a zero of a vector field ξ on a manifold $\mathcal{M}$ requires an *affine connection* ∇ on $\mathcal{M}$, which can be thought of as a generalization of the directional derivative; for details, see, e.g., [4, §5.2]. In the next paragraph, we proceed to describe a class of affine connections on $\mathrm{Grass}(p,n)$.

Let $\widetilde{W}_U$ be an $n\times p$ matrix that depends on U in such a way that $\mathrm{col}(\widetilde{W}_U)$ is constant on the equivalence classes $[U]$, and such that $\widetilde{U}_U^\top \widetilde{W}_U$ is invertible for all U. Define ∇ by

$$\overline{(\nabla_{\eta_{\mathrm{col}(U)}}\xi)}_U = P_{\widetilde{W}_U,\widetilde{U}_U}\mathrm{D}\overline{\xi}(U)[\overline{\eta}_U], \tag{11}$$

where ξ is a vector field on $\mathrm{Grass}(p,n)$ and $\eta_{\mathrm{col}(U)}$ is a tangent vector to $\mathrm{Grass}(p,n)$ at $\mathrm{col}(U)$. Observe that $\nabla_\eta\xi$ is a tangent vector to $\mathrm{Grass}(p,n)$ at $\mathrm{col}(U)$ and that $\overline{(\nabla_{\eta_{\mathrm{col}(U)}}\xi)}_U$ denotes the horizontal lift of that tangent vector. It can be checked that the right-hand side of (11) satisfies the compatibility condition (9) of horizontal lifts, hence (11) is a legitimate definition of a tangent vector $\nabla_{\eta_{\mathrm{col}(U)}}\xi$. It can also be checked that the mapping ∇ thus defined has all the properties of an affine connection.

On an abstract manifold $\mathcal{M}$ equipped with an affine connection ∇, the Newton equation at $x\in\mathcal{M}$ for computing a zero of a vector field ξ reads

$$\nabla_{\eta_x}\xi = -\xi_x. \tag{12}$$

Now let $\mathcal{M}$ be the Grassmann manifold $\mathrm{Grass}(p,n) = \mathbb{R}_*^{n\times p}/\mathrm{GL}_p$, let x be $\mathrm{col}(U)$, and consider the choice (8) for the horizontal space, the choice (11) for the affine connection, and the choice (10) for the vector field ξ. Then, replacing the symbol $\overline{\eta}_U$ by Z for simplicity of notation, the horizontal lift at U of the left-hand side of the Newton equation (12) becomes

$$P_{\widetilde{W}_U,\widetilde{U}_U}\mathrm{D}\overline{\xi}(U)[Z] = P_{\widetilde{W}_U,\widetilde{U}_U}\left(P_{BU,\widetilde{U}_U}AZ - BZ(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U^\top AU - BUE[Z]\right), \tag{13}$$

where $E[Z] := \mathrm{D}(U\mapsto(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U^\top AU)(U)[Z]$.

We choose

$$\widetilde{W}_U := BU \tag{14}$$

to get rid of the $BUE[Z]$ term. The Newton equation (12), in its matrix formulation given by the horizontal lift at U, thus becomes

$$P_{BU,\widetilde{U}_U}(AZ - BZ(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U^\top AU) = -P_{BU,\widetilde{U}_U}AU, \tag{15a}$$

$$\widetilde{U}_U^\top Z = 0. \tag{15b}$$

(Recall that $P_{BU,\widetilde{U}_U} = I - BU(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U^\top$, and that matrix $\widetilde{U}_U$ can be chosen arbitrarily as a function of U under the condition that $\widetilde{U}_U^\top U$ and $\widetilde{U}_U^\top BU$ be invertible and that the column space of $\widetilde{U}_U$ be constant along the equivalence classes $[U]$.)

With the retraction on $\mathrm{Grass}(p,n)$ chosen as in [4, Example 4.1.5], it follows from the convergence theory of the geometric Newton method [4, Algorithm 4] that the iteration on $\mathrm{Grass}(p,n)$ defined by

$$\mathrm{col}(U)\mapsto\mathrm{col}(U+Z_U), \tag{16}$$

where Z_U denotes the solution of (15), converges locally, at least quadratically, to the spectral invariant spaces of $B^{-1}A$.

6 Discussion

The Newton map (16) is the G function announced in the beginning of Section 4. It is an iteration on the quotient space $\mathbb{R}_*^{n\times p}/\mathrm{GL}_p$, or equivalently on the Grassmann manifold since $\mathbb{R}_*^{n\times p}/\mathrm{GL}_p \simeq \mathrm{Grass}(p,n)$. In practice, the iteration is realized numerically by an iteration function g as in (6), such that $\mathrm{col}(U+Z_U) = \mathrm{col}(g(U))$. Any function g such that $g(U) = (U+Z_U)M_U$, where M_U is $p\times p$ and invertible, is suitable. The freedom in M_U can be exploited to keep the iterates (sufficiently close to) orthonormal.

Let W_U be such that $W_U^\top BU$ is invertible. Then, without loss of information, we can multiply (15a) on the left by P_{BU,W_U} to obtain

$$P_{BU,W_U}(AZ - BZ(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U^\top AU) = -P_{BU,W_U}AU. \tag{17}$$

On the other hand, a block generalization of the Jacobi correction equation of [31, Algorithm 3.1], with the hypotheses of [31, Theorem 3.2], would rather be

$$P_{BU,W_U}(AZ - BZ(W_U^\top BU)^{-1}W_U^\top AU) = -P_{BU,W_U}AU. \tag{18}$$

As mentioned in [31, §3.3], when these equations are to be solved with unpreconditioned subspace methods, it is desirable to have $W_U = \widetilde{U}_U$; otherwise the domain space $\mathcal{H}_U$ of the linear map $Z \mapsto P_{BU,W_U}(AZ - BZ(\widetilde{U}_U^\top BU)^{-1}\widetilde{U}_U^\top AU)$ differs from the image space, which implies that powers cannot be formed. In this case where $W_U = \widetilde{U}_U$, (17) and (18) coincide.

If BU in $P_{BU,\widetilde{U}_U}$ appearing in (15) is replaced by U, then some terms that do not go to zero are neglected in the Jacobian, and one can expect that quadratic convergence is lost. This is confirmed in the $p = 1$ case by the experiments reported in [31, §9.1.1].

A full-blown block JD method for the generalized eigenvalue problem would consist in enhancing the Newton equation (15) with a Davidson strategy, where U is selected by a Ritz approximation with respect to a subspace spanned by previous corrections (see [31, Algorithm 4.2]). If the goal is to compute the p-dimensional invariant subspace, $\mathcal{V}$, assumed to be spectral (see Section 3), corresponding to the smallest (resp. largest) eigenvalues of (A,B), and if the p Ritz vectors corresponding to the smallest (resp. largest) Ritz values are used for the next U, then quadratic convergence is preserved. This follows from [2, Proposition 6.1], where the objective function f is the generalized Rayleigh quotient defined by $f(\mathrm{col}(U)) = \mathrm{tr}((U^\top BU)^{-1}U^\top AU)$ (resp. its opposite). To see this, observe that $\mathcal{V}$ is the global minimizer of f (see, e.g., [4, Proposition 2.1.1]), that it is nondegenerate since it is assumed to be spectral (see the discussion in [4, §6.5.1]), and that the Newton iteration (16), since it converges locally quadratically to $\mathcal{V}$, is a descent iteration for f close enough to $\mathcal{V}$.

Bibliography

[1] P.-A. Absil, C. G. Baker, and K. A. Gallivan. Trust-region methods on Riemannian manifolds. *Found. Comput. Math.*, 7(3):303–330, 2007. Cited p. 13.

[2] P.-A. Absil and K. A. Gallivan. Accelerated line-search and trust-region methods. *SIAM J. Numer. Anal.*, 47(2):997–1018, 2009. Cited p. 18.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.*, 80(2):199–220, 2004. Cited pp. 13 and 15.

[4] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. Cited pp. 13, 15, 16, 17, and 18.

[5] R. L. Adler, J.-P. Dedieu, J. Y. Margulies, M. Martens, and M. Shub. Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA J. Numer. Anal.*, 22(3):359–390, 2002. Cited p. 13.

[6] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000. Cited p. 13.

[7] G. Ammar and C. Martin. The geometry of matrix eigenvalue methods. *Acta Appl. Math.*, 5(3):239–278, 1986. Cited p. 13.

[8] M. Baumann and U. Helmke. Riemannian subspace tracking algorithms on Grassmann manifolds. In *Proceedings of the 46th IEEE Conference on Decision and Control*, pages 4731–4736, 2007. Cited p. 13.

[9] M. Baumann and U. Helmke. A time-varying Newton algorithm for adaptive subspace tracking. *Mathematics and Computers in Simulation*, 79(4):1324–1345, 2008. Cited p. 13.

[10] M. Chen. On the geometric structure of thermodynamics. *J. Math. Phys.*, 40(2):830–837, 1999. Cited p. 13.

[11] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1998. Cited pp. 13 and 15.

[12] R. Geus. *The Jacobi-Davidson algorithm for solving large sparse symmetric eigenvalue problems with application to the design of accelerator cavities*. PhD thesis, Swiss Federal Institute of Technology Zürich, 2002. Cited p. 12.

[13] I. Gohberg, P. Lancaster, and L. Rodman. *Invariant Subspaces of Matrices with Applications*, volume 51 of *Classics in Applied Mathematics*. SIAM, 2006. Cited p. 14.

[14] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, third edition, 1996. Cited p. 13.

[15] U. Helmke and P. A. Fuhrmann. Controllability of matrix eigenvalue algorithms: The inverse power method. *Systems and Control Letters*, 41(1):57–66, 2000. Cited p. 13.

[16] U. Helmke, K. Hüper, and J. Trumpf. Newton's method on Grassmann manifolds, 2007. arXiv:0709.2205v2. Cited pp. 13 and 15.

[17] U. Helmke and J. B. Moore. Singular-value decomposition via gradient and self-equivalent flows. *Linear Algebra and Its Applications*, 169:223–248, 1992. Cited p. 13.

[18] U. Helmke and J. B. Moore. *Optimization and Dynamical Systems*. Springer, 1994. Cited pp. 13 and 15.

[19] U. Helmke and F. Wirth. On controllability of the real shifted inverse power iteration. *Systems and Control Letters*, 43(1):9–23, 2001. Cited p. 13.

[20] M. E. Hochstenbach and Y. Notay. The Jacobi–Davidson method. *GAMM Mitteilungen*, 29(2):368–382, 2006. Cited p. 11.

[21] K. Hueper, U. Helmke, and J. B. Moore. Structure and convergence of conventional Jacobi-type methods minimizing the off-norm function. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2124–2129, 1996. Cited p. 13.

[22] The Jacobi–Davidson Gateway.
`http://www.win.tue.nl/casa/research/topics/jd/`.
Cited p. 11.

[23] J. Jordan and U. Helmke. Controllability of the QR-algorithm on Hessenberg flags. In *Proceedings of the Fifteenth International Symposium on Mathematical Theory of Network and Systems*, 2002. No pagination. Cited p. 13.

[24] M. Kleinsteuber, U. Helmke, and K. Hüper. Jacobi's algorithm on compact Lie algebras. *SIAM Journal on Matrix Analysis and Applications*, 26(1):42–69, 2005. Cited p. 13.

[25] E. Kröner. The differential geometry of elementary point and line defects in Bravais crystals. *Internat. J. Theoret. Phys.*, 29(11):1219–1237, 1990. Cited p. 13.

[26] E. Lundström and L. Eldén. Adaptive eigenvalue computations using Newton's method on the Grassmann manifold. *SIAM J. Matrix Anal. Appl.*, 23(3):819–839, 2002. Cited pp. 12 and 13.

[27] R. E. Mahony, U. Helmke, and J. B. Moore. Gradient algorithms for principal component analysis. *J. Austral. Math. Soc. Ser. B*, 37(4):430–450, 1996. Cited p. 13.

[28] J. H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Trans. Signal Process.*, 50(3):635–650, 2002. Cited p. 13.

[29] J. H. Manton, U. Helmke, and I. M. Y. Mareels. A dual purpose principal and minor component flow. *Systems Control Lett.*, 54(8):759–769, 2005. Cited p. 13.

[30] M. Shub. Some remarks on dynamical systems and numerical analysis. In *Proc. VII ELAM.*, pages 69–92. Equinoccio, U. Simón Bolívar, Caracas, 1986. Cited pp. 12 and 13.

[31] G. L. G. Sleijpen, A. G. L. Booten, D. R. Fokkema, and H. A. van der Vorst. Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT*, 36(3):595–633, 1996. Cited p. 18.

[32] G. L. G. Sleijpen and H. A. van der Vorst. A Jacobi–Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17(2):401–425, 1996. Cited p. 11.

[33] G. L. G. Sleijpen and H. A. van der Vorst. The Jacobi–Davidson method for eigenvalue problems and its relation with accelerated inexact Newton schemes. In S. D. Margenov and P. S. Vassilevski, editors, *Iterative Methods in Linear Algebra, II.*, volume 3 of *IMACS Series in Computational and Applied Mathematics*, pages 377–389, 1996. Cited p. 11.

[34] S. T. Smith. Optimization techniques on Riemannian manifolds. In A. Bloch, editor, *Hamiltonian and Gradient Flows, Algorithms and Control*, volume 3 of *Fields Inst. Commun.*, pages 113–136. Amer. Math. Soc., 1994. Cited p. 13.

[35] S. T. Smith. Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Trans. Signal Process.*, 53(5):1610–1630, 2005. Cited p. 13.

[36] G. W. Stewart. *Matrix Algorithms. Vol. II.* SIAM, 2001. Cited pp. 11, 13, and 14.

On partial stabilizability of linear systems

Mustapha Ait Rami
University of Valladolid
Spain
aitrami@autom.uva.es

John B. Moore
Australian National University
Canberra, Australia
john.moore@anu.edu.au

Abstract. A generalization of linear system stability theory is presented. It is shown that the partial stabilizability problem can be cast as a Linear Matrix Inequality (LMI) condition. Also, the set of all initial conditions for which the system is stabilizable by open-loop controls (the stabilizability subspace) is characterized in terms of Semi-Definite Programming (SDP).

1 Introduction

This study brings a novel treatment to the stability and control of linear systems which are not necessary stabilizable. Its contribution lies in the simplicity of the proposed approach. By appropriate problem formulations, all of the results are derived only from first principles. The solutions are simply obtained by a matrix algebra manipulation and/or tackled by convex optimization. Sharing the same point of view of [5], and avoiding high level analysis this work is self containing and provides accessible and complete treatment to the partial stabilizability problem.

Our focus is a fundamental stability issue in linear systems. We investigate the following *partial stabilizability* problem. Consider the following finite dimensional linear time-invariant system

$$\frac{dx(t)}{dt} = Ax(t) + Bu(t), \qquad x(0) = x_0 \in \mathbb{R}^n, \tag{1}$$

where A and B are real matrices of dimension $n \times n$ and $n \times n_u$, respectively; then under which conditions there exists a control law $u(\cdot)$, such that the resulting trajectory $x(\cdot)$ converges asymptotically to zero for a specific initial condition x_0?

Of course the answer and the solution specialize to known results when the system is stabilizable in the classical sense and it can be given by the solution to the classical Riccati equation or by using the well-know LMI techniques [2] (based on the classical Lyapunov equation).

We shall first develop preliminary results for the stability, then these results are applied to the stability synthesis problem. Specifically, it will be shown that the partial stability problem can be expressed in terms of Linear Matrix Inequality (LMI). The set of all initial conditions for which the system is stabilizable by open-loop controls: the stabilizability subspace, is characterized by its projection operator which turn out to be a solution to an adequate (Semi-Definite Programming) SDP problem, (see [6] on LMI and SDP). Also, we shall show that this stabilizability subspace is exactly the set of initial conditions for which the system is stabilizable by static state-feedback control laws.

The remainder of the chapter is organized as follows. Section 2 is devoted to the stability analysis. Section 3 deals with the stabilizability synthesis problem and provides necessary and sufficient conditions for this problem in terms of LMI. Section 4 provides some concluding remarks.

Notation. We make use of the following notation. $\mathbb{R}$ denotes the set of real numbers. $M^\top$ denotes the transpose of the real the matrix M. $M^\dagger$ denotes the Moore-Penrose inverse of the matrix M. **Tr**(M) is the sum of diagonal elements of a square matrix M. For a real matrix M, $M > 0$ (resp. $M \geq 0$) means M is symmetric and positive-definite (resp. positive semidefinite). I denotes the identity matrix, with size determined from the context. $\mathbf{span}(v_1, ..., v_k)$ represents the linear space generated by the vectors $v_1, ..., v_k$.

2 Stability analysis

In what follows we develop a simple stability theory of linear systems based on convex optimization. We stress that neither the modal analysis of linear systems nor the classical Lyapunov Theorem (which actually does not apply for stability with a single fixed initial condition) are used. The new results are derived from first principles.

Definition 1. System (1) is called x_0-stable if the associated trajectory from an initial condition $x(0) = x_0$ and $u(\cdot) = 0$, vanishes at infinity i.e. $\lim_{t\to+\infty} x(t) = 0$.

Theorem 2. *Given an initial condition $x(0) = x_0$ for System (1) with $u = 0$, then the following statements are equivalent.*

(i) The System (1) is x_0-stable.

(ii) There exists a positive semidefinite matrix $P \geq 0$ such that

$$AP + PA^\top + x_0 x_0^\top = 0. \tag{2}$$

Proof. Let $x(t)$ be the trajectory of System (1) associated with $u = 0$, $x(0) = x_0$. Then the implication $(i) \Rightarrow (ii)$ is straightforward by using the integration

$$\int_0^{+\infty} \left. \frac{dx(t)x(t)^\top}{dt} \right|_{t=s} ds$$

and setting $P = \int_0^{+\infty} x(s)x(s)^\top ds$ in (2). We stress that $\int_0^{+\infty} x(s)x(s)^\top ds < +\infty$, since $x(t)$ goes to zeros and the expression of $x(t)$ contains only the products of polynomials and exponents. Next, The implication $(ii) \Rightarrow (i)$ can be shown as follows. Let $P \geq 0$ be a solution to (2) and define $\Phi(t) = e^{tA} P e^{tA^\top}$. Since any matrix commutes with its exponential, a simple calculation gives $\dot{\Phi}(t) = -x(t)x(t)^\top$. Therefore, $\Phi(t)$ is decreasing as time goes to infinity. So the limit $l(P) = \lim_{t\to+\infty} \Phi(t)$ exists since $\Phi(t)$ is necessarily bounded from below by 0 and

$$l(P) - P = -\int_0^{+\infty} x(s)x(s)^\top ds < +\infty.$$

Hence $\lim_{t\to+\infty} x(t) = 0$. □

In the forthcoming results the following set will be used.

Definition 3. The stability subspace $\mathcal{S}_0$ is defined as

$$\mathcal{S}_0 \triangleq \{x_0 \in \mathbb{R}^n \mid \text{ System (1) with } u = 0 \text{ is } x_0\text{-stable}\}. \tag{3}$$

Remark 4. It is trivial that $\mathcal{S}_0$ is a linear subspace of $\mathbb{R}^n$. Also, any free trajectory of System (1) belongs to $\mathcal{S}_0$ whenever its initial condition belongs to $\mathcal{S}_0$.

Along the same line of reasoning, a useful generalization of the previous result can be derived as follows.

Theorem 5. *Given $x_0,\ldots,x_k \in \mathbb{R}^n$, the following statements are equivalent.*

(i) $\mathbf{span}(x_0,\ldots,x_k) \subset \mathcal{S}_0$

(ii) There exists positive semidefinite matrix $P \geq 0$ such that

$$AP + PA^\top + \sum_{i=0}^{k} x_i x_i^\top = 0. \tag{4}$$

The following result can be viewed as a generalization of the classical Lyapunov theorem.

Corollary 6. *Given a matrix C, the following statements are equivalent.*

(i) range$(C) \subset \mathcal{S}_0$

(ii) There exists $P \geq 0$ satisfying

$$AP + PA^\top + CC^\dagger = 0. \tag{5}$$

Moreover, when either any previous item holds, then A is a Hurwitz matrix (all its eigenvalues have strictly negative real part) if and only if $\ker(C) \subset \mathcal{S}_0$*, or equivalently if and only if the following equality is feasible:*

$$A\tilde{P} + \tilde{P}A^\top + I - C^\dagger C = 0, \;\; \tilde{P} \geq 0. \tag{6}$$

Proof. The equivalences between (i),$(ii)$and (iii) are an immediate consequence of Theorem 5. The rest of the proof follows from the fact that $CC^\dagger$ is the projection on the range of C and $I - C^\dagger C$ is the projection on the null space of C. □

Remark 7. Before presenting the next result we would like to stress out that any projection operator M onto a subspace E, is idempotent ($M^2 = M$) and possesses only 0 and 1 as eigenvalues. Such fact is a well-know result. Moreover, M is symmetric $M = M^\top$ and positive semidefinite $M \geq 0$.

Now, consider the projection operator X_0 onto the linear subspace $\mathcal{S}_0$. The following result provides a characterization of X_0 in terms of semidefinite programming (SDP).

Theorem 8. *Consider the following optimization problem*

$$\begin{cases} \min -\mathbf{Tr}(X) \\ \textit{subject to:} \\ AP + PA^\top + X = 0, \;\; I \geq X, \;\; P \geq 0. \end{cases} \tag{7}$$

The minima P,X of (7) are always achievable and unique in the variable X. Moreover, the projection operator X_0 onto the linear subspace $\mathcal{S}_0$, is the only optimal solution for (7) and the optimal index value $\mathbf{Tr}(X_0)$ equals the dimension of $\mathcal{S}_0$.

Proof. We show first that the projection X_0 is feasible solution to (7). It is trivial that the projection X_0 satisfies $X_0 \leq I$. Let k be the dimension of $\mathcal{S}_0$, then $X_0 = \sum x_i x_i^\top$ with $x_1, \ldots, x_k$ an orthonormal basis of $\mathcal{S}_0$. So that Theorem 5 implies that there exist P such that P,X_0 is feasible solution to (7). Next, we prove that X_0 is the only minimum of (7). Let X^* be any minimum, then necessarily it has only 0 and 1 as eigenvalues. Thus X^* is a projection onto a subspace of $\mathcal{S}_0$. Necessarily, $X^* = X_0$, since X_0 is feasible and $k = \mathbf{Tr}(X_0) \leq \mathbf{Tr}(X^*)$. □

3 Stability synthesis

In the sequel, the solution of the x_0-stabilizability problem is shown to be equivalently expressed in terms of LMI. The following definitions will be essential in our development.

Definition 9. The System (1) is called x_0-stablizable if there exists a control law such that the corresponding trajectory with $x(0) = x_0$ vanishes at infinity. In this case, the control law is called x_0-stabilizing.

Definition 10. The stabilizability subspace $\mathcal{S}_u$ is defined as

$$\mathcal{S}_u \triangleq \{x_0 \in \mathbb{R}^n \mid \text{System (1) is } x_0\text{-stabilizable}\}. \tag{8}$$

Remark 11. It is trivial that the stabilizability subspace $\mathcal{S}_u$ is a linear subspace of $\mathbb{R}^n$. Also, any trajectory $x(\cdot)$ of System (1) with $x(0) \in \mathcal{S}_u$, belongs to $\mathcal{S}_u$.

Next, we show that the x_0-stabilizability of System (1) can be expressed in terms of LMI. For this purpose, we need the following key lemmas.

Lemma 12 (Extended Schur's Lemma [1]). *Let matrices $\Gamma = \Gamma^\top$, $\Theta = \Theta^\top$ and Δ be given with appropriate sizes. Then the following conditions are equivalent:*

(i) $\Gamma - \Delta\Theta^\dagger\Delta^\top \geq 0$, $\Theta \geq 0$, *and* $\Delta(I - \Theta\Theta^\dagger) = 0$.

(ii) $\begin{bmatrix} \Gamma & \Delta \\ \Delta^\top & \Theta \end{bmatrix} \geq 0.$

The following lemma can be found in many references but its origin is due to Penrose [3, 4].

Lemma 13. *Let matrices* Γ, Δ *and* Θ *be given with appropriate sizes. Then the following matrix equation*

$$\Gamma X \Delta = \Theta, \tag{9}$$

has a solution X if and only if

$$\Gamma\Gamma^{\dagger}\Theta\Delta^{\dagger}\Delta = \Theta. \tag{10}$$

Moreover, the set of all solutions to (9) *is given by*

$$X = \Gamma^{\dagger}\Theta\Delta^{\dagger} + Y - \Gamma^{\dagger}\Gamma Y \Delta\Delta^{\dagger}, \tag{11}$$

where Y *is an arbitrary matrix of appropriate size.*

Now, we are in position to provide the following result.

Theorem 14. *Given* $v_1, ..., v_k \in \mathbb{R}^n$*, then following conditions are equivalent.*

(i) The System (1) *is* x_0*-stablizable for every* $x_0 \in \mathbf{span}(v_1, ..., v_k)$.

(ii) For all $x_0 \in \mathbf{span}(v_1, ..., v_k)$*, there exists a* x_0*-stabilizing static state-feedback control.*

(iii) There exist $S = S^{\top}$, $T = T^{\top}$ *and* U *such that*

$$\begin{cases} AS + SA^{\top} + BU + U^{\top}B^{\top} + \sum_{0}^{k} v_i v_i^{\top} = 0, \\ \begin{bmatrix} S & U^{\top} \\ U & T \end{bmatrix} \geq 0. \end{cases} \tag{12}$$

Moreover, from (iii) *we have that the state feedback control law*

$$u(t) = [US^{\dagger} + Y(I - SS^{\dagger})]x(t),$$

is x_0*-stabilizing for any arbitrary matrix* Y.

Proof. Assume that System (1) is x_0-stabilizable for given initial conditions $v_1, \ldots,$ v_k. Denote by $u_{v_1}, \ldots, u_{v_k}$ the associated controls and by $x_{v_1}, \ldots, x_{v_k}$ the corresponding trajectories, then

$$S = \sum_{i=0}^{k} \int_{0}^{+\infty} x_{v_i} x_{v_i}^{\top}\, dt, \qquad U = \sum_{i=0}^{k} \int_{0}^{+\infty} x_{v_i} u_{v_i}^{\top}\, dt, \qquad T = \sum_{i=0}^{k} \int_{0}^{+\infty} u_{v_i} u_{v_i}^{\top}\, dt.$$

Since

$$\begin{bmatrix} S & U \\ U^{\top} & T \end{bmatrix} \geq 0,$$

by integrating $\dfrac{d(xx^{\top})}{dt}$ it is easily seen that S, U, T satisfy condition (iii). Now, assume that (iii) holds. Using the Schur Lemma we have $U(I - SS^{\dagger}) = 0$. So that by Lemma 13 the equation $KS = U$ has a solution with $K = US^{\dagger} + Y(I - SS^{\dagger})$ and Y arbitrary. Substituting this expression into (12) we obtain

$$(A+BK)^{\top}S + S(A+BK) + \sum_{0}^{k} v_i v_i^{\top} = 0.$$

Then by Theorem 5 we conclude that the state-feedback control $u = Kx$ is x_0-stabilizing for any x_0 in $\mathbf{span}(v_1,\dots,v_k)$. □

Next, consider the projection operator X_u onto the linear stabilizability subspace $\mathcal{S}_u$, i.e $X_u x = x,\ \forall x \in \mathcal{S}_u$. Then the following result provides a characterization of X_u in terms of SDP.

Theorem 15. *Consider the following optimization problem*

$$\left\{\begin{array}{l} \min -\mathbf{Tr}(X) \\ \textit{subject to:} \\ AS + SA^{\top} + BU + U^{\top}B^{\top} + X = 0, \\ \begin{bmatrix} S & U^{\top} \\ U & T \end{bmatrix} \geq 0,\ I \geq X. \end{array}\right. \tag{13}$$

The minima P, X of (7) are always achievable and unique in the variable X. Moreover, The projection operator $X_u \geq 0$ onto $\mathcal{S}_u$ is the only optimal solution for (13) with optimal index values $\mathbf{Tr}(X_u)$ equal to the dimension of $\mathcal{S}_u$.

Proof. It suffices to apply Theorem 14 to get the first part of the result. The proof of the second part follows the same reasoning as for Theorem 8. □

The main contribution of this work is now stated.

Theorem 16. *Let X_u be the projection onto the stabilizability subspace $\mathcal{S}_u$. Then the following statements are equivalent:*

(i) There exist $S = S^{\top}$, $T = T^{\top}$ and U such that

$$\left\{\begin{array}{l} AS + SA^{\top} + BU + U^{\top}B^{\top} + X_u = 0, \\ \begin{bmatrix} S & U^{\top} \\ U & T \end{bmatrix} \geq 0. \end{array}\right. \tag{14}$$

In this case, for any arbitrary matrix Y the state feedback control law

$$u(t) = (US^{\dagger} + Y(I - SS^{\dagger}))x(t),$$

is x_0-stabilizing $\forall x_0 \in \mathcal{S}_u$.

(ii) There exist $S = S^{\top}$, $T = T^{\top}$ and U such that

$$\left\{\begin{array}{l} AX_uSX_u + X_uSX_uA^{\top} + BUX_u + X_uU^{\top}B^{\top} + X_u = 0, \\ \begin{bmatrix} S & U^{\top} \\ U & T \end{bmatrix} \geq 0. \end{array}\right. \tag{15}$$

In this case, for any arbitrary matrix Y the feedback

$$u(t) = [UX_u(X_uSX_u)^{\dagger} + Y(I - X_uSX_u(X_uSX_u)^{\dagger})]x(t),$$

is x_0-stabilizing $\forall x_0 \in \mathcal{S}_u$.

Proof. Let $X_u = \sum_0^k v_i v_i^{\top}$, then using Theorem 14 the control law $u = US^{\dagger}$ is x_0-stabilizing for any initial condition in $\mathbf{span}(v_1, \ldots, v_k)$. Denote by $x_{v_1}, \ldots, x_{v_k}$ the corresponding trajectories, then

$$S = \sum_0^k \int_{i=0}^{+\infty} x_{v_i} x_{v_i}^{\top}\, dt, \qquad U = \sum_0^k \int_{i=0}^{+\infty} u_{v_i} x_{v_i}^{\top}\, dt, \qquad T = \sum_0^k \int_{i=0}^{+\infty} u_{v_i} u_{v_i}^{\top}\, dt$$

satisfy

$$\begin{cases} AS + SA^{\top} + BU + U^{\top}B^{\top} + X_u = 0, \\ \begin{bmatrix} S & U^{\top} \\ U & T \end{bmatrix} \geq 0. \end{cases}$$

Since the trajectories stay in the stabilizability subspace, we have also $X_uS = S$ and $UX_u = U$. The rest of the proof is straightforward and follows the same line of argument as in Theorem 14. □

4 Conclusion

It has been shown that the partial stabilizability problem can be expressed in terms of an LMI condition. Moreover, the set of all initial conditions for which the system is stabilizable by open-loop controls or static state-feedback controls (the stabilizability subspace) can be characterized via SDP.

Bibliography

[1] A. Albert. Conditions for positive and nonnegative definiteness in terms of pseudoinverses. *SIAM J. Appl. Math.*, 17:434–440, 1969. Cited p. 26.

[2] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory.* SIAM, 1994. Cited p. 23.

[3] R. Penrose. A generalized inverse of matrices. *Proc. Cambridge Philos. Soc.*, 51:406–413, 1955. Cited p. 26.

[4] R. Penrose. On the best approximate solutions of linear matrix equations. *Proc. Cambridge Philos. Soc.*, 52:17–19, 1955. Cited p. 26.

[5] R. E. Skelton. Increased roles of linear algebra in control education. In *Proc. American Cont. Conf.*, pages 393–397, 1994. Cited p. 23.

[6] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996. Cited p. 23.

On the zero properties of tall linear systems with single-rate and multirate outputs

Brian D. O. Anderson
Australian National University
Canberra, Australia
brian.anderson@anu.edu.au

Mohsen Zamani
Australian National University
Canberra, Australia
mohsen.zamani@anu.edu.au

Giulio Bottegal
University of Padova
Padova, Italy
bottegal@dei.unipd.it

Abstract. The zero properties of tall discrete-time multirate linear systems are studied in this paper. In the literature, zero properties of multirate linear systems are defined as those of their corresponding blocked systems, which are time-invariant systems whose behavior is broadly equivalent to that of the generating multirate system. In this paper, we review some recent scattered results of the authors and their colleagues dealing with the zero properties of the blocked systems associated with multirate systems. First, we show that tall linear time-invariant unblocked systems are zero-free when the parameter matrices A,B,C,D assume generic values. Then it is argued that tall blocked systems obtained from blocking of tall linear time-invariant systems with generic parameter matrices A,B,C,D, are also zero-free. Finally, we illustrate that tall blocked systems associated with multirate systems generically have no finite nonzero zeros.

1 Introduction

Our motivation for studying the zero properties of tall transfer function matrices, transfer function matrices which have more outputs than inputs, comes from their potential application in generalized dynamic factor models. Such models arise in a number of fields e.g. econometric modeling, signal processing and systems and control, and the associated transfer functions are almost always tall. Hence the authors of this work have become interested in the zero properties of tall systems due to their application in generalized dynamic factor models, though now consider these properties of interest in their own right. Now as just mentioned, in generalized dynamic factor models, it is very common to have models with a larger number of outputs compared to their number of inputs i.e. tall models; furthermore, when they are used for econometric modeling, it is also very common to have some outputs measured monthly while some other outputs may be obtained quarterly or even annually [10], [15], [14]. Thus, the models are periodic. Moreover, in this context, the latent variables, i.e. the noiseless part of the outputs, or the part remaining after removal of contaminating additive measurement noise, are modeled by systems with unobserved white noise inputs. In a single-rate setting i.e. monthly data only, [8] has shown that model tallness generically implies that the generalized dynamic factor model is zero-free, and then the latent variables can be modeled as a singular autoregressive process whose parameters can be easily identified from covariance data using Yule-Walker equations. A corresponding demonstration is still lacking for the multirate case i.e. with both monthly and quarterly data. The results of this paper are aimed at enabling us to understand better the properties of multirate factor models, and in particular establishing that tall systems again are generically zero-free; this is done with a view to later establishing the utility of the Yule-Walker approach for identifying multirate factor models from their covariance data. Quite apart from this motivation however, the results of this paper suggest in relation to classical control design that, if one adds extra sensors to make a plant have more outputs than inputs, then controller design will be much easier due to the generic absence of plant zeros. Note though that this paper does not focus on the applications problem, but rather on the system theoretical issues involved with the zero properties of multirate systems with tall structure.

Our main goal is to establish generic conditions for when the system matrix associated with (a blocked version of) a tall multirate system will have a property corresponding to the system having no zeros, and also to establishing generic conditions for when all zeros are simple and finite. In order to reach this goal we review some of our recent results regarding the zero properties of multirate systems, and in particular, how the technique of blocking or lifting can allow them to be treated as time-invariant systems.

In the systems and control literature, the technique of blocking or lifting has been used for a long time to transfer a multirate linear system into a linear time-invariant system which is generally referred to as a blocked system (see e.g., [3], [16], [2]). The nomenclature arises because blocked systems can be obtained from stacking the input and output vectors of multirate systems within a period into new larger vectors, see [3], [16] and later in this paper. Moreover, in the literature, the zeros of multirate systems are defined as those of their corresponding blocked systems [3], [7], [5].

The zero properties of blocked linear systems have been studied to some degree in the literature; for instance, [4], [11] have explored the zero properties of blocked systems obtained from blocking of linear periodic systems (a class of systems which includes multi-rate systems). The results show that the blocked system has a finite zero if it is obtained from a time-invariant unblocked system, and the latter has a finite zero, which is a form of sufficiency condition. However, this reference does not provide a necessary condition for a blocked system obtained this way to have a finite zero. This gap has been covered in [19] and [6] where the authors have introduced some additional information about the zero properties of blocked systems obtained from blocking of time-invariant systems. References [19] and [6] have used different approaches but they have obtained largely similar results. The results in those references show that the blocked system is zero-free if and only if its associated linear time-invariant unblocked system is zero-free. In contrast to the case where the unblocked system is time-invariant, very few results indeed however deal with zeros of blocked systems where these systems have been obtained by blocking of a truly multirate system, i.e. one that is not time-invariant. We note that [18] does have some partial results.

In this paper, to achieve the main goal stated in the second paragraph, some of the recent works of the authors and their colleagues are recalled. In particular, we utilize results from [1], [19] and [18]. First in Section 2 one of the results of [1] is modified to study the zero properties of a tall unblocked linear time-invariant system under a generic setting i.e. when parameter matrices A,B,C,D assume generic values. Then in Section 3 the results of [19] are exploited to explore the zero properties of the blocked systems associated with unblocked linear time-invariant systems. Subsequently, these results are used in Section 4 to examine the zero-freeness of tall blocked systems associated with tall multirate linear systems. Finally, Section 5 provides concluding remarks and plans for future works.

It would be inappropriate to close this section without reflecting on the fact that this volume celebrates the achievements of Uwe Helmke. Two of the three authors of this paper count among their happiest professional experiences their interactions with Uwe, a person who is both very talented professionally, and possessing of an affable and generous nature. A special skill has characterized all these interactions: his ability to put himself out of the world of mathematics and into the world of engineering, and to speak the language of engineering where needed, with an infinite supply of patience in answering the engineers' questions.

2 Tall linear time-invariant unblocked systems

In this section we study the zero properties of tall linear time-invariant unblocked systems. Here, one of the results of [1] is modified to show that a tall system with generic parameter matrices A,B,C,D is zero-free i.e. its associated system matrix has full-column rank for all $z \in \mathbb{C} \cup \{\infty\}$.

Consider the following time-invariant unblocked system

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t \\ y_t &= Cx_t + Du_t \end{aligned} \tag{1}$$

where $t \in \mathbb{Z}$, $x_t \in \mathbb{R}^n$, $y_t \in \mathbb{R}^p$ and $u_t \in \mathbb{R}^m$, $p \geq m$. For this system, we assume that y_t is available at every time instant t.

In order to study the zero properties of the system (1), we need to provide a proper definition for the zeros. Here, we first recall the following definition for zeros of the unblocked system (1) from [13] and [12] (page 178).

Definition 1. The finite zeros of the transfer function $W(z) = C(zI - A)^{-1}B + D$ with minimal realization $[A,B,C,D]$ are defined to be the finite values of z for which the rank of the following system matrix falls below its normal rank

$$M(z) = \begin{bmatrix} zI - A & -B \\ C & D \end{bmatrix}. \tag{2}$$

Further, $W(z)$ is said to have an infinite zero when $n + rank(D)$ is less than the normal rank of $M(z)$, or equivalently the rank of D is less than the normal rank of $W(z)$.

The following lemma inspired from results in [9] studies the zero properties of the system (1) when $p = m$ and the parameter matrices A,B,C,D accept generic values. It states that for generic parameter matrices, the rank reduction of the system matrix at any zero is 1. The lemma will help us to prove the main result of this section regarding the zero-freeness of tall linear time-invariant unblocked systems with generic parameter matrices.

Lemma 2. *The set $\mathcal{F} = \{[A,B,C,D] | p = m, rank(D) = m, rank(M(z)) \geq n+m-1, \forall z \in \mathbb{C}\}$ is open and dense in the set $\{[A,B,C,D] | p = m, rank(D) = m\}$.*

Proof. **Dense**: Consider the system matrix $M(z)$ and suppose that there exists a z_0 such that $rank(M(z_0)) = n+m-2$ (note that only the case where rank drops to $n+m-2$ is discussed here and generalization to $n+m-k, k \geq 2$ is straightforward). Therefore, there exist two linearly independent vectors, say x_1 and x_2, which span the kernel of $M(z_0)$. Let $x_i = [x_{i1}^\top \, x_{i2}^\top]^\top$, $i = 1,2$ with $x_{i1} \in \mathbb{R}^n$, then x_{11} and x_{21} must be linearly independent otherwise there would exist nonzero scalars a_1 and a_2 such that $a_1x_1 + a_2x_2 = [0 \; a_1x_{12}^\top + a_2x_{22}^\top]^\top$ with $D[a_1x_{12} + a_2x_{22}] = 0 \Longrightarrow a_1x_{12} + a_2x_{22} = 0$. The latter implies that x_1 and x_2 are linearly dependent which violates the initial assumption. Now it is easy to verify that $[z_0I - A + BD^{-1}C][x_{11} \; x_{21}] = 0$, which implies that $A - BD^{-1}C$ has a repeated eigenvalue. By manipulation of an entry of A by an arbitrarily small amount, we see that the kernel of the new $M(z)$ for any z will have dimension at most 1 since $A - BD^{-1}C$ will have nonrepeated eigenvalues.

Open: Set $\mathcal{F}$ being open is equivalent to its complement, call it $\mathcal{F}^C$, being closed. To obtain a contradiction, suppose $\mathcal{F}^C$ is not closed. Then there must exist a sequence $[A_m,B_m,C_m,D_m]_{m\in\mathbb{N}}$ in $\mathcal{F}^C$ with $[A_m,B_m,C_m,D_m] \to [A_0,B_0,C_0,D_0] \in \mathcal{F}$. Moreover, there exists a $z_m \in \mathbb{C}$ such that $rank(M_m(z_m)) \leq n+m-2$, where $M_m(z)$ denotes the system matrix associated with $[A_m,B_m \, ,C_m,D_m]$. Consequently, $\sigma_1(M_m(z_m)) = \sigma_2(M_m(z_m)) = 0$ where $\sigma_i(F)$ denotes the i-th smallest singular value of F. Now $M_m(z_m) \to M_0(z_0)$ holds as $[A_m,B_m,C_m,D_m] \to [A_0 \, ,B_0,C_0,D_0]$ and $z_m \to z_0$. Hence, $\sigma_2(M_m(z_m)) \to \sigma_2(M_0(z_0))$; however, by assumption $\sigma_2(M_0 \, (z_0)) > 0$ which contradicts the fact that $\sigma_2(M_0(z_0)) \to 0$ and the result follows. □

Theorem 3. *Consider a transfer function matrix $W(z)$ with minimal realization $[A,B,C,D]$ of dimension n in which B,C have m columns and p rows respectively with $p > m$. If the entries of A,B,C,D assume generic values, then $W(z)$ has no finite or infinite zeros.*

Proof. Observe first that the normal rank (which is the rank for almost all z) of a generic $M(z)$ i.e. system matrix $M(z)$ with generic matrices, is $n+m$: to see this, take $A = C = 0$ and D as any full column rank matrix, to get a particular $M(z)$ which for any nonzero z has rank $n+m$. Since the normal rank cannot exceed $n+m$ and this rank is attained for a particular choice of A etc, so $n+m$ must be the normal rank for generic $M(z)$. Observe also that D generically has rank m, and hence the normal rank of M equals $n + \mathrm{rank}\, D$, which shows that generically $W(z)$ has no infinite zero. For the finite zeros, observe that any such zero must be a zero of every minor of dimension $(n+m) \times (n+m)$. Since $M(z)$ has normal rank $n+m$, there must be at least one minor of dimension $(n+m) \times (n+m)$ which is nonzero for almost all values of z. Choose A,B and the first m rows of C,D generically, and consider the associated minor. For each of the finite set of values of z for which the minor is zero, determine the associated kernel which has the dimension at most one based on the result of Lemma 2. Then a generic $(n+m)$-dimensional vector will not be orthogonal to any single one of these kernels, and since there are a finite number of such kernels, a generic $(n+m)$-dimensional vector will not be orthogonal to any of the kernels considered simultaneously. If the next, i.e $(m+1)$-th, row of $[C\ D]$ is set equal to this vector, then any vector in any of the finite set of kernels of the $(n+m)$-dimensional minors formed using the first m rows of $[C\ D]$ will not be orthogonal to the added row of $[C\ D]$, which means that the $(m+n+1)$ row matrix obtained by adjoining the new row of $[C\ D]$ must have an empty kernel for any value of z, i.e. there is no zero. Given that C,D are actually generic and may have more rows again, the result is now evident. □

3 Tall blocked linear systems

It was shown in the previous section that tall time-invariant unblocked systems are generically zero-free. In this section we study the zero properties of their associated blocked systems. The results of this section enable us to study the zero properties of blocked systems resulted from blocking of linear systems with multirate output in the next section. We note that, proofs for some of the theorems are omitted and an interested reader can refer to [19] for detailed proofs.

Now we define for a fixed but arbitrary positive number $N > 1$

$$\begin{aligned} U_t &= \begin{bmatrix} u_t^\top & u_{t+1}^\top & \cdots & u_{t+N-1}^\top \end{bmatrix}^\top, \\ Y_t &= \begin{bmatrix} y_t^\top & y_{t+1}^\top & \cdots & y_{t+N-1}^\top \end{bmatrix}^\top, \end{aligned} \tag{3}$$

where $t = 0, N, 2N, \ldots$.

Then the blocked system is given by

$$\begin{aligned} x_{t+N} &= A_b x_t + B_b U_t \\ Y_t &= C_b x_t + D_b U_t. \end{aligned} \tag{4}$$

The blocked system, mapping the U_t sequence to the Y_t sequence, has a time-invariant state-variable description given by

$$\begin{aligned} A_b &= A^N, \quad B_b = \begin{bmatrix} A^{N-1}B & A^{N-2}B & \ldots & B \end{bmatrix}, \\ C_b &= \begin{bmatrix} C^\top & A^\top C^\top & \ldots & A^{(N-1)^\top} C^\top \end{bmatrix}^\top, \\ D_b &= \begin{bmatrix} D & 0 & \ldots & 0 \\ CB & D & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{N-2}B & CA^{N-3}B & \ldots & D \end{bmatrix}. \end{aligned} \tag{5}$$

An operator Z is defined such that $Zx_t = x_{t+N}$, $ZU_t = U_{t+N}$, $ZY_t = Y_{t+N}$. In the rest of this paper, the symbol Z is also used to denote a complex value. We denote the transfer function associated with (4) as $V(Z) = D_b + C_b(ZI - A_b)^{-1}C_b$ and it is worthwhile remarking that minimality of $[A,B,C]$ is equivalent to minimality of $[A_b, B_b, C_b]$.

3.1 The zero properties of blocked linear systems

Since in this section we are interested in the zero properties of the system (4), we need to first define zeros for that system. Similar to Definition 1 we have the following definition for the zeros of the system (4).

Definition 4. The finite zeros of the transfer function $V(Z) = C_b(ZI - A_b)^{-1}B_b + D_b$ with minimal realization $[A_b, B_b, C_b, D_b]$ are defined to be the finite values of Z for which the rank of the following system matrix falls below its normal rank

$$M_b(Z) = \begin{bmatrix} ZI - A_b & -B_b \\ C_b & D_b \end{bmatrix}. \tag{6}$$

Further, $V(Z)$ is said to have an infinite zero when $n + rank(D_b)$ is less than the normal rank of $M_b(Z)$, or equivalently the rank of D_b is less than the normal rank of $V(Z)$.

According to the above definition the normal rank of the system matrix $M_b(Z)$ plays an important role in the zero properties of its associated blocked system.

Lemma 5. *Suppose that $p \geq m$. Then the normal rank of $M(z)$ is $n+m$ if and only if the normal rank $M_b(Z)$ is $n+Nm$.*

Proof. One can refer to [19] for a complete proof. □

The above lemma establishes a relation between the normal rank of $M(z)$ and the normal rank of $M_b(Z)$. In the following we recall the relation between zeros of these system matrices.

Theorem 6. *Suppose the system matrix of* (1) *has full-column normal rank. Then if* (1) *has a finite zero at $z = z_0 \neq 0$, then the system* (4) *has a finite zero at $Z = Z_0 = z_0^N \neq 0$. Conversely, if the system* (4) *has a finite zero at $Z = Z_0 = z_0^N \neq 0$, then the system* (1) *has a finite zero at one or more of $z = z_0 \neq 0$ or $z = \omega z_0 \neq 0, \ldots, z = \omega^{N-1} z_0 \neq 0$, where $\omega = \exp(\dfrac{2\pi j}{N})$.*

Proof. The proof is omitted and an interested reader can refer to [19] for a complete proof. □

So far we have related nonzero zeros of the system (4) and those of the system (1). Now, we present theorems which establish a relation between zeros of those aforementioned systems at infinity and the origin, see [19] for the proofs, which are straightforward.

Theorem 7. *Suppose the system matrix of* (1) *has full-column normal rank. Then the system* (4) *has a zero at* $Z_0 = \infty$ *if and only if the system* (1) *has a zero at* $z_0 = \infty$.

The above theorem treats the zeros of systems (4) and (1) at the infinity. In the theorem below, we deal with zeros of the blocked system (4) and its associated unblocked system (1) at the origin.

Theorem 8. *Suppose the system matrix of* (1) *has full-column normal rank. Then the system* (4) *has a zero at* $Z_0 = 0$ *if and only if the system* (1) *has a zero at* $z_0 = 0$.

3.2 Zero-free blocked system

The results in the previous subsection established a clear connection between zeros of the blocked system (4) and zeros of the associated unblocked system (1). Furthermore, in the first section, it was shown that the system (1) with $p > m$ and a choice of generic parameter matrices $[A,B,C,D]$, is zero-free. Now, to complete our analysis for the zero properties of tall blocked systems we provide the theorem below which studies the zero properties of the blocked system (4) obtained from blocking of a linear time-invariant system with generic parameter matrices. It is worthwhile remarking that parameter matrices of the blocked system (4) cannot be generic because they are structured.

We first need the lemma below which studies the normal rank of $M_b(Z)$ when matrices A,B,C,D assume generic values.

Lemma 9. *For a generic choice of matrices* $[A,B,C,D]$ *with* $p \geq m$*, the system matrix of* (4) *has normal rank equal to* $n + Nm$.

Proof. In the generic setting and $p \geq m$, matrix D is of full column rank. So, due to the structure of D_b, see (5), one can easily conclude that D_b is of full column rank as well. Then the result easily follows. □

Theorem 10. *Consider the system* (1) *defined by the quadruple* $[A,B,C,D]$*, in which the individual matrices are generic. Then*

1. *If* $p > m$*, the system matrix of the blocked system has full column rank for all* Z.
2. *If* $p = m$*, then the system matrix of the blocked system can only have finite zeros with one-dimensional kernel.*

Proof. Suppose first that $p > m$. Using the results of Lemma 9 and Lemma 5, it can be readily seen that the system matrix of tall unblocked systems generically have full-column normal rank. Furthermore, Theorem 3 shows that tall unblocked systems are generically zero-free. If the blocked system had its system matrix with less than full column rank for a finite $Z_0 \neq 0$, then according to Theorem 6, there would be necessarily a nonzero nullvector of the system matrix of the unblocked system for $z_0 \neq 0$ equal to some $N-th$ root of Z_0, which would be a contradiction. If the blocked system had a zero at $Z_0 = \infty$, then based on Theorem 8 the D matrix of the unblocked system would be less than full column rank which would be a contradiction. Analogously, using the argument in Theorem 6, one can easily conclude that the blocked system has full column rank system matrix at $Z_0 = 0$.

Now we consider the case $p = m$; since D is generic, it has full column rank. Hence, based on the conclusion of Theorem 8, both the unblocked system and the blocked system do not have zeros at infinity. In the second part of this proof we use the conclusion of Theorem 6. Furthermore, one should note that since the matrices A, B, C and D assume generic values it can be easily understood that the quadruple $[A_b, B_b, C_b, D_b]$ is a minimal realization. Now, based on the fact that D_b is nonsingular, one can conclude that the zeros of the blocked system are the eigenvalues of $A_b - B_b D_b^{-1} C_b$. If the eigenvalues of $A_b - B_b D_b^{-1} C_b$ are distinct, then the associated eigenspace for each eigenvalue is one-dimensional; it is equivalent to saying that the associated kernel of $M_b(Z)$ evaluated at the eigenvalue has dimension one. One should note that the unblocked system has distinct zeros due to the genericity assumption. Furthermore, zeros of the unblocked system generically have distinct magnitudes except for complex conjugate pairs. It is obvious that those zeros of the unblocked system with distinct magnitudes produce distinct blocked zeros. Now, we focus on zeros of the unblocked system with the same magnitudes, i.e. complex conjugate pairs. The only case where the generic unblocked system has distinct zeros but its corresponding blocked system has non-distinct zeros happens when the $N-th$ power of the complex conjugate zeros of the unblocked system coincide. We now show by contradiction that this is generically impossible. In order to illustrate a contradiction, suppose that the unblocked system has a complex conjugate pair, say z_{01} and z_{01}^*. If they produce an identical zero for the blocked system, their $N-th$ powers must be the same. The latter condition implies that the angle between z_{01} and z_{01}^* has to be exactly $\frac{2\pi h}{N}$, where h is an integer, which contradicts the genericity assumption for the unblocked system. Hence, the zeros of the blocked system generically have distinct values and consequently the corresponding kernels of system matrix evaluated at the zeros are one-dimensional. □

4 Multirate systems

In Section 2 we consider the system (1), in which y_t exists for all t, and, as a separate matter, can be measured at every time t. The zero properties of this system for a choice of generic parameter matrices were completely discussed in Section 2. Then in Section 3 it was shown that when the system (1) and its associated blocked system (4) are tall, they are generically zero-free. Now, in this section we are also interested in the situation where y_t exists for all t, but not every entry is measured for all t. In particular, we consider a case where y_t has components that are observed at different rates. For simplicity, a case where outputs are provided at two rates which we refer to as the fast rate and the slow rate is assumed. In this section, we use the result of the previous section to specify conditions under which blocked systems associated with multirate systems are generically zero-free.

As mentioned earlier we discuss the case where y_t has components that are measured at two rates so, without loss of generality we decompose y_t as

$$y_t = \begin{bmatrix} y_t^f \\ y_t^s \end{bmatrix}$$

where $y_t^f \in \mathbb{R}^{p_1}$ is observed at all t, the fast part, and $y_t^s \in \mathbb{R}^{p_2}$ is observed at $t = 0, N, 2N, \ldots$, the slow part, also $p_1 > 0, p_2 > 0$ and $p_1 + p_2 = p$. Accordingly, we decompose C and D as

$$C = \begin{bmatrix} C^f \\ C^s \end{bmatrix}, D = \begin{bmatrix} D^f \\ D^s \end{bmatrix}.$$

Thus, the multirate linear system corresponding to what is measured has the following dynamics:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t \quad t = 0, 1, 2, \ldots \\ y_t^f &= C^f x_t + D^f u_t \; t = 0, 1, 2, \ldots \\ y_t^s &= C^s x_t + D^s u_t \;\; t = 0, N, 2N, \ldots \end{aligned} \tag{7}$$

We have actually N distinct alternative ways to block the system, depending on how fast signals are grouped with the slow signals. Even though these N different systems share some common zero properties, their zero properties are not identical in the whole complex plane (see [3], pages 173-179).

For, $\tau \in \{1, 2, \cdots, N\}$, define

$$\begin{aligned} U_t^\tau &\triangleq \begin{bmatrix} u_{t+\tau} \\ u_{t+\tau+1} \\ \vdots \\ u_{t+\tau+N-1} \end{bmatrix}, \\ Y_t^\tau &\triangleq \begin{bmatrix} y_{t+\tau}^f \\ y_{t+\tau+1}^f \\ \vdots \\ y_{t+\tau+N-1}^f \\ y_{t+N}^s \end{bmatrix}, \; t = 0, N, 2N, \ldots \\ x_t^\tau &\triangleq x_{t+\tau}. \end{aligned} \tag{8}$$

Then the blocked system $\sum_\tau$ is defined by

$$\begin{aligned} x_{t+N}^\tau &= A_\tau x_t^\tau + B_\tau U_t^\tau \\ Y_t^\tau &= C_\tau x_t^\tau + D_\tau U_t^\tau, \end{aligned} \tag{9}$$

where,

$$
\begin{aligned}
A_\tau &\triangleq A^N, \\
B_\tau &\triangleq \begin{bmatrix} A^{N-1}B & A^{N-2}B & \dots & AB & B \end{bmatrix}, \\
C_\tau &\triangleq \begin{bmatrix} C^{f\top} & A^\top C^{f\top} & \dots & A^{(N-1)^\top} C^{f\top} & A^{(N-\tau)^\top} C^{s\top} \end{bmatrix}^\top, \\
D_\tau &\triangleq \begin{bmatrix} D^f & 0 & \dots & 0 \\ C^f B & D^f & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C^f A^{N-2} B & C^f A^{N-3} B & \dots & D^f \\ C^s A^{N-\tau-1} B & \dots & D^s & * \end{bmatrix},
\end{aligned}
\tag{10}
$$

where "*" at the very right corner denotes $\tau - 1$ zero matrices of size $p_2 \times m$ and when $N - \tau - 1 < 0$, $C^s A^{-1} B$ is replaced by D^s and rest of the terms in the last row are replaced by zero matrices of size $p_2 \times m$.

Reference [3] defines a zero of (7) at time τ as a zero of its corresponding blocked system $\sum_\tau$ [1]. Hence, in the rest of this section we focus on the zero properties of the blocked system $\sum_\tau$, $\forall \tau \in \{1,2,\dots,N\}$.

Definition 11. The finite zeros of system $\sum_\tau$ are defined to be finite values of Z for which the rank of the following system matrix falls below its normal rank

$$
M_\tau(Z) = \begin{bmatrix} ZI - A_\tau & -B_\tau \\ C_\tau & D_\tau \end{bmatrix}.
$$

Further, $V_\tau(Z) = C_\tau(ZI - A_\tau)^{-1} B_\tau + D_\tau$, $\tau \in \{1,2,\dots,N\}$, is said to have an infinite zero when $n + rank(D_\tau)$, $\tau \in \{1,2,\dots,N\}$, is less than the normal rank of $M_\tau(Z)$, $\tau \in \{1,2,\dots,N\}$, or equivalently the rank of D_τ, $\tau \in \{1,2,\dots,N\}$, is less than the normal rank of $V_\tau(Z)$, $\tau \in \{1,2,\dots,N\}$.

In addition to the above definition the following results from [6] and [7] are useful to the rest of this paper.

Lemma 12. *The pair (A,B) is reachable if and only if the pairs (A_τ, B_τ), $\forall \tau \in \{1,2,\dots,N\}$ are reachable.*

The above lemma studies the reachability property of $\sum_\tau$, $\forall \tau \in \{1,2,\dots,N\}$ and the lemma below explores its transfer function.

Lemma 13. *The transfer function $V_\tau(Z)$ associated with the blocked system* (9) *has the following property*

$$
V_{\tau+1}(Z) = \begin{bmatrix} 0 & I_{p_1(N-1)} & 0 \\ ZI_{p_1} & 0 & 0 \\ 0 & 0 & I_{p_2} \end{bmatrix} V_\tau(Z) \begin{bmatrix} 0 & Z^{-1} I_m \\ I_{m(N-1)} & 0 \end{bmatrix},
$$

where $\tau \in \{1,2\dots,N\}$.

[1] Zeros of the transfer function obtained from (9) and defined following [3] are identical with those defined here, provided the quadruple $\{A_\tau, B_\tau, C_\tau, D_\tau\}$is minimal.

The result of the above lemma is crucial for the study of the zero properties of $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$, for the choice of finite nonzero zeros. The latter is the main focus for the remainder of this section. We treat the zero properties of $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$, under genericity and tallness assumptions. Given that $p_1, p_2 > 0$ and tallness is defined by $Np_1 + p_2 > Nm$, it proves convenient to consider partitioning the set of p_1, p_2 defining tallness into two subsets, as follows.

1. $p_1 > m$.
2. $p_1 \leq m, Np_1 + p_2 > Nm$.

The first case is common, perhaps even overwhelmingly common in econometric modeling but the second case is important from a theoretical point of view, and possibly in other applications. Moreover, our results are able to cover both cases.

4.1 Case $p_1 > m$

According to Definition 11, the normal rank for the system matrix of $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$, plays an important role in the analysis of its zero properties; thus, we make the following observation on the normal rank of $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$ using the conclusion of Lemma 9.

Remark 14. For generic choice of the matrices $[A, B, C^s, C^f, D^f, D^s]$, $p_1 \geq m$, the system matrix of $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$, has normal rank of $n + Nm$.

In the situation where $p_1 > m$, obtaining a result on the absence of finite nonzero zeros is now rather trivial, since the blocked system contains a tall subsystem obtained by deleting some outputs which is provably zero-free.

Theorem 15. *For a generic choice of the matrices $[A, B, C^s, C^f, D^s, D^f]$, $p_1 > m$, the system matrix of $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$, has full column rank for all finite zero Z.*

Proof. With the help of the conclusion of Theorem 10, one can easily conclude that there is a submatrix of $M_{\tau}(Z)$, obtained by deleting rows of $M_{\tau}(Z)$ associated with slow part, which is full-column rank for all finite Z. Then the conclusion of the theorem easily follows. □

4.2 Case $p_1 \leq m$, $Np_1 + p_2 > Nm$

In the previous subsection the case $p_1 > m$ was treated where only considering the fast outputs alone generically leads to a zero-free blocked system, and the zero-free property is not disturbed by the presence of the further slow outputs. A different way in which the blocked system will be tall arises when $p_1 \leq m$ and $Np_1 + p_2 > Nm$. The main result of this subsection is to show that $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$ with $p_1 \leq m$, $Np_1 + p_2 > Nm$ is again generically zero-free. In order to study the latter case we need to review properties of the Kronecker canonical form of a matrix pencil. Since the system matrix of $\sum_{\tau}$, $\forall \tau \in \{1,2,\ldots,N\}$ is actually a matrix pencil, the Kronecker canonical form turns out to be a very useful tool to obtain insight into the zeros of (9) and the structure of the kernels associated with those zeros.

The main theorem on the Kronecker canonical form of the matrix pencil is obtained from [17].

Theorem 16. *[17] Consider a matrix pencil $zR+S$. Then under the equivalence defined using pre- and postmultiplication by nonsingular constant matrices $\widetilde{P}$ and $\widetilde{Q}$, there is a canonical quasidiagonal form:*

$$\widetilde{P}(zR+S)\widetilde{Q} = \operatorname{diag}[L_{\varepsilon_1},\dots,L_{\varepsilon_r},\tilde{L}_{\eta_1},\dots,\tilde{L}_{\eta_s},zN-I,zI-K], \tag{11}$$

where:

1. *L_μ is the $\mu \times (\mu+1)$ bidiagonal pencil*

$$\begin{bmatrix} z & -1 & 0 & \dots & 0 & 0 \\ 0 & z & -1 & \dots & 0 & 0 \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & z & -1 \end{bmatrix}. \tag{12}$$

2. *$\tilde{L}_\mu$ is the $(\mu+1)\times\mu$ transposed bidiagonal pencil*

$$\begin{bmatrix} -1 & 0 & \dots & 0 & 0 \\ z & -1 & \dots & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & z & -1 \\ 0 & 0 & \dots & 0 & z \end{bmatrix}. \tag{13}$$

3. *N is a nilpotent Jordan matrix.*

4. *K is in Jordan canonical form.*

Note the possibility that $\mu = 0$ exists. The associated L_0 is deemed to have a column but not a row and $\tilde{L}_0$ is deemed to have a row but not a column, see [17].

The following corollary can be directly derived easily from the above theorem and provides detail about the vectors in the null space of the Kronecker canonical form. Because the matrices $\tilde{P}$ and $\tilde{Q}$ are nonsingular, it is trivial to translate these properties back to an arbitrary matrix pencil, including a system matrix.

Corollary 17.

1. *For all z except for those which are eigenvalues of K, the kernel of the Kronecker canonical form has dimension equal to the number of matrices L_μ appearing in the form; likewise the co-kernel dimension is determined by the number of matrices $\tilde{L}_\mu$.*

2. *The vector $[1\ z\ z^2 \dots z^\mu]^\top$ is the generator of the kernel of L_μ, a set of vectors $[0\dots 0\ 1\ z\ z^2 \dots z^\mu\ 0\dots 0]^\top$ are generators for the kernel of the whole canonical form which depend continuously on z, provided that z is not an eigenvalue of K; when z is an eigenvalue of K, the vectors form a subset of a set of generators.*

3. *When z equals an eigenvalue of K, the dimension of the kernel jumps by the geometric multiplicity of that eigenvalue, the rank of the pencil drops below the normal rank by that geometric multiplicity, and there is an additional vector or vectors in the kernel apart from those defined in point 2, which are of the form $[0\,0\ldots v^\top]^\top$, where v is an eigenvector of K. Such a vector is orthogonal to all vectors in the kernel which are a linear combination of the generators listed in the previous point.*

4. *When z is an eigenvalue, say z_0 of K, the associated kernel of the matrix pencil can be generated by two types of vectors: those which are the limit of the generators defined by adding extra zeros to vectors such as $[1\; z_0\; z_0^2 \ldots, z_0^\mu]^\top$ (these being the limits of the generators when $z \neq z_0$ but continuously approaches z_0), and those obtained by adjoining zeros to the eigenvector(s) of K with eigenvalue z_0, the latter set being orthogonal to the former set.*

In the rest of this subsection, we explore the zero properties of $M_\tau(Z)$, $\forall\ \tau \in \{1, 2,\ldots,N\}$. To achieve this, we first focus on the particular case of $M_1(Z)$. Later, we introduce the main result for the zero properties of $M_\tau(Z)$, $\forall \tau \in \{1,2,\ldots,N\}$.

First we need to introduce some parameters. To this end, we argue first that the first $n+Np_1$ rows of $M_1(Z)$ are linearly independent. For the submatrix formed by these rows is the system matrix of the blocked system obtained by blocking the fast system defined by $\{A,B,C^f,D^f\}$, and accordingly has full row normal rank, since the unblocked system is generic and square or fat under the condition $p_1 \leq m$. Now define the square submatrix of $M_1(Z)$:

$$N(Z) \triangleq \begin{bmatrix} ZI - A_1 & -B_1 \\ \mathcal{C}_1 & \mathcal{D}_1 \end{bmatrix}, \tag{14}$$

such that normal rank $N(Z)$ =normal rank $M_1(Z)$, by including the first $n+Np_1$ rows of $M_1(Z)$ and followed by appropriate other rows of $M_1(Z)$ to meet the normal rank and squareness requirements. Hence there exists a permutation matrix P such that

$$PM_1(Z) = \begin{bmatrix} N(Z) \\ \mathcal{C}_2\ \ \mathcal{D}_2 \end{bmatrix} \tag{15}$$

where $\mathcal{C}_2$ and $\mathcal{D}_2$ capture those rows of C_1 and D_1 that are not included in $\mathcal{C}_1$ and $\mathcal{D}_1$, respectively.

The zero properties of $N(Z)$ are studied in the following proposition.

Proposition 18. *Let the matrix $N(Z)$ be the submatrix of $M_1(Z)$ formed via the procedure described. Then for generic values of the matrices A,B, etc. with $p_1 \leq m$ and $Np_1 + p_2 > Nm$, for any finite Z_0 for which the matrix $N(Z_0)$ has less rank than its normal rank, its rank is one less than its normal rank.*

Proof. The proof is omitted; an interested reader can refer to [18] for a complete proof. □

The result of the previous proposition, although restricted to $\tau = 1$, enables us to establish the following main result applicable for any τ.

Theorem 19. *Consider the system* $\sum_\tau$*,* $\forall \tau \in \{1, 2, \ldots, N\}$*, with* $p_1 \leq m$*, and* $Np_1 + p_2 > Nm$*. Then for generic values of the defining matrices* $\{A, B, C^f, D^f, C^s, D^s\}$ *the system matrix* $M_\tau(Z)$ *,* $\forall \tau \in \{1, 2, \ldots, N\}$*, has rank equal to its normal rank for all finite nonzero values of* Z_0*, and accordingly* Σ_τ *has no finite nonzero zeros.*

Proof. We first focus on the case $\tau = 1$. Now, apart from the $p_2 - N(m - p_1)$ rows of the C^s, D^s which do not enter the matrix $N(Z)$ defined by (14), choose generic values for the defining matrices, so that the conclusions of the preceding proposition are valid.

Let $Z_a, Z_b, \ldots$ be the finite set of Z for which $N(Z)$ has less rank than its normal rank (the set may have less than n elements, but never has more), and let $w_a, w_b, \ldots$ be vectors which are in the corresponding kernels (*not co-kernels*) and orthogonal to the subspace in the kernel obtained from the limit of the kernel of $N(Z)$ as $Z \to Z_a, Z_b, \ldots$ etc. Now, due to the facts that $M_1(Z)$ and $N(Z)$ have the same normal rank and any existing vector in the kernel of $M_1(Z)$ is in the the kernel of $N(Z)$ one can conclude that the subspace in the kernel obtained from the limit of the kernel of $N(Z)$ as $Z \to Z_a, Z_b, \ldots$ etc, coincides with the subspace in the kernel obtained from the limit of the kernel of $M_1(Z)$ as $Z \to$ zeros of $M_1(Z)$.

Now, to obtain a contradiction, we suppose that the system matrix $M_1(Z)$ is such that, for $Z_0 \neq 0$, $M_1(Z_0)$ has rank less than its normal rank, i.e. the dimension of its kernel increases. Since the kernel of $M_1(Z_0)$ is a subspace of the kernel of $N(Z_0)$, Z_0 must coincide with one of the values of $Z_a, Z_b, \ldots$ and the rank of $M_1(Z_0)$ must be only one less than its normal rank; moreover, there must exist an associated nonzero w_1 unique up to a scalar multiplier, in the kernel of $M_1(Z_0)$ which is orthogonal to the limit of the kernel of $M_1(Z)$ as $Z \to Z_0$. Then w_1 is necessarily in the kernel of $N(Z_0)$, orthogonal to the limit of the kernel of $N(Z)$ as $Z \to Z_0$ and thus w_1 in fact must coincide to within a nonzero multiplier with one of the vectors $w_a, w_b, \ldots$.

Write this w_1 as

$$w_1 = \begin{bmatrix} x_1 \\ u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}, \tag{16}$$

and suppose the input sequence $u(i) = u_i$ is applied for $i = 1, 2 \ldots, N$ to the original system, starting in initial state x_1 at time 1. Let $y^f(1), y^f(2), \ldots$ denote the corresponding fast outputs and $y^s(N)$ the slow output at time N. Break this up into two subvectors, $y^{s1}(N), y^{s2}(N)$, where $y^{s1}(N)$ is associated with those rows of C^s, D^s

which are included in $\mathcal{C}_1$, $\mathcal{D}_1$ and $y^{s2}(N)$ is related with the remaining rows of C^s and D^s . We have

$$N(Z_0)w_1 = \begin{bmatrix} Z_0 I_n - A^N & -A^{N-1}B & -A^{N-2}B & \dots & -B \\ C^f & D^f & 0 & \dots & 0 \\ C^f A & C^f B & D^f & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ C^f A^{N-1} & C^f A^{N-2}B & C^f A^{N-3}B & \dots & D^f \\ C^{s1} A^{N-1} & C^{s1} A^{N-2}B & C^{s1} A^{N-3}B & \dots & D^{s1} \end{bmatrix} w_1 = \begin{bmatrix} Z_0 x_1 - x(N+1) \\ y^f(1) \\ y^f(2) \\ \vdots \\ y^f(N) \\ y^{s1}(N) \end{bmatrix} = 0. \tag{17}$$

Now it must be true that $x_1 \neq 0$. For otherwise, we would have $N(Z)w_1 = 0$ for all Z, which would violate assumptions. Since also $Z_0 \neq 0$, there must hold $x(N+1) \neq 0$. Hence there cannot hold both $x(N) = 0$ and $u(N) = 0$. Consequently, we can always find C^{s2}, D^{s2} such that $y^{s2}(N) = C^{s2}x(N) + D^{s2}u(N) \neq 0$, i.e. the slow output value is necessarily nonzero, no matter whether $w_1 = w_a, w_b$, etc. Equivalently, the equation $[\mathcal{C}_2 \;\; \mathcal{D}_2]w_1 = 0$ cannot hold. Hence, if $M_1(Z)$ defines a system with a finite zero and it is nonzero, this is a nongeneric situation. Hence, $M_1(Z)$ generically has rank equal to its normal rank for all finite nonzero Z. Now, we show that the latter property holds for all $M_\tau(Z)$, $\tau \in \{1,2,\dots,N\}$. First, note that the pair (A,B) is generically reachable so, according to Lemma 12 the pair (A_τ, B_τ), $\forall \tau \in \{1,2,\dots,N\}$, is also reachable. Consider $Z_\zeta \in \mathbb{C} - \{0,\infty\}$, if Z_ζ does not coincide with the eigenvalues of A_τ then

$$rank(M_\tau(Z_\zeta)) = n + rank(V_\tau(Z_\zeta)). \tag{18}$$

Hence, using the result of Lemma 13, it is immediate that $rank(M_\tau(Z_\zeta)) = rank\,(M_{\tau+1}(Z_\zeta))$. If Z_ζ does coincide with an eigenvalue of A_τ then $rank(V_\tau(Z_\zeta))$ is ill-defined. However, since zeros of $M_\tau(Z)$, $\tau \in \{1,2\dots,N\}$, are invariant under state feedback and the pair (A_τ, B_τ) is reachable, one can easily find a state feedback to shift that eigenvalue [20] and then (18) is a well-defined equation and $rank(M_\tau(Z_\zeta))$ $= rank(M_{\tau+1}(Z_\zeta))$. Thus, we can conclude that all $M_\tau(Z)$, $\tau \in \{1,2,\dots,N\}$ generically have no finite nonzero zeros. This ends the proof. □

The above theorem studies the case of finite nonzero zeros. The cases of zeros at the origin and at infinity seem to be more complicated because the structure of the system matrices depend on τ; furthermore, the various $M_\tau(Z)$, for $\tau \in \{1,2,\dots,N\}$, may not share the same zeros at those aforementioned points. Hence, these two points need special treatments. Here, we offer the following conjecture which partly treats the case of zeros at the origin and infinity.

Conjecture 20. Consider the system $\sum_\tau$, $\forall\tau \in \{1,2,\dots,N\}$, with $p_1 < m$ and $Np_1 + p_2 > Nm$. Then for generic values of the defining matrices $[A,B,C^f,D^f,C^s,D^s]$ the system matrix $M_\tau(Z)$, $\tau \in \{1,2,\dots,N\}$ always has zeros at either $Z = 0$ or $Z = \infty$ or at both points.

The above conjecture has been proved for a particular case where the normal rank of the system matrix $M_\tau(Z)$ is equal to the number of columns. Furthermore, it is consistent with numerical examples. Here, we provide the following example which exhibits a very simple scenario and is consistent with the conclusion of the conjecture.

Example 21. Consider a tall multi-rate system with $n = 1$, $m = 3$, $N = 2$, $p_1 = 1$, $p_2 = 5$. Let the parameter matrices for the multi-rate system be $A = a$, $B = [b_1\, b_2\, b_3]$, $C = [{c^f}^\top {C^s}^\top]^\top$, $C^s = [c_1^s\, c_2^s\, c_3^s\, c_4^s\, c_5^s]^\top$, $D = [D^f\, D^s]$, $D^f = [d_1^f\, d_2^f\, d_3^f]$ and

$$D^s = \begin{bmatrix} d_{11}^s & d_{12}^s & d_{13}^s \\ \vdots & \vdots & \vdots \\ d_{51}^s & d_{52}^s & d_{53}^s \end{bmatrix}.$$

First, consider $\tau = 1$ and write the associated system matrix as

$$M_1(Z) = \begin{bmatrix} Z-a^2 & -ab_1 & -ab_2 & -ab_3 & -b_1 & -b_2 & -b_3 \\ c^f & d_1^f & d_2^f & d_3^f & 0 & 0 & 0 \\ c^f a & c^f b_1 & c^f b_2 & c^f b_3 & d_1^f & d_2^f & d_3^f \\ c_1^s a & c_1^s b_1 & c_1^s b_2 & c_1^s b_3 & d_{11}^s & d_{12}^s & d_{13}^s \\ c_2^s a & c_2^s b_1 & c_2^s b_2 & c_2^s b_3 & d_{21}^s & d_{22}^s & d_{23}^s \\ c_3^s a & c_3^s b_1 & c_3^s b_2 & c_3^s b_3 & d_{31}^s & d_{32}^s & d_{33}^s \\ c_4^s a & c_4^s b_1 & c_4^s b_2 & c_4^s b_3 & d_{41}^s & d_{42}^s & d_{43}^s \\ c_5^s a & c_5^s b_1 & c_5^s b_2 & c_5^s b_3 & d_{51}^s & d_{52}^s & d_{53}^s \end{bmatrix}.$$

It is obvious that first two rows are linearly independent. Now, consider the rows 3 to 8; they can be written as

$$\begin{bmatrix} c^f & c^f & c^f & c^f & d_1^f & d_2^f & d_3^f \\ c_1^s & c_1^s & c_1^s & c_1^s & d_{11}^s & d_{12}^s & d_{13}^s \\ c_2^s & c_2^s & c_2^s & c_2^s & d_{21}^s & d_{22}^s & d_{23}^s \\ c_3^s & c_3^s & c_3^s & c_3^s & d_{31}^s & d_{32}^s & d_{33}^s \\ c_4^s & c_4^s & c_4^s & c_4^s & d_{41}^s & d_{42}^s & d_{43}^s \\ c_5^s & c_5^s & c_5^s & c_5^s & d_{51}^s & d_{52}^s & d_{53}^s \end{bmatrix} \operatorname{diag}(a,b_1,b_2,b_3,I_3) = G\operatorname{diag}(a,b_1,b_2,b_3,I_3)$$

The matrix G has rank at most 4; hence, with generic parameter matrices the normal rank of $M(Z)$ equals 6; furthermore, it is easy to observe that the system matrix has a zero at $Z = 0$. However, for $\tau = 2$ we can write the system matrix $M_2(Z)$ as

$$M_2(Z) = \begin{bmatrix} Z - a^2 & -ab_1 & -ab_2 & -ab_3 & -b_1 & -b_2 & -b_3 \\ c^f & d_1^f & d_2^f & d_3^f & 0 & 0 & 0 \\ c^f a & c^f b_1 & c^f b_2 & c^f b_3 & d_1^f & d_2^f & d_3^f \\ c_1^s & d_{11}^s & d_{12}^s & d_{13}^s & 0 & 0 & 0 \\ c_2^s & d_{21}^s & d_{22}^s & d_{23}^s & 0 & 0 & 0 \\ c_3^s & d_{31}^s & d_{32}^s & d_{33}^s & 0 & 0 & 0 \\ c_4^s & d_{41}^s & d_{42}^s & d_{43}^s & 0 & 0 & 0 \\ c_5^s & d_{51}^s & d_{52}^s & d_{53}^s & 0 & 0 & 0 \end{bmatrix}.$$

Observe that the normal rank of the system matrix is still 6 and the matrix D_2 (with its nonzero entries assuming generic values) has rank 4; hence, the only zero of the system matrix is now at infinity.

5 Conclusions and future works

The zero properties of tall discrete-time multirate linear system were addressed in this paper. The zero properties of multirate linear systems were defined as those of their corresponding blocked systems. In this paper several required results from [1], [19] and [18] were reviewed in order to prove the main results about the zero properties of the blocked systems associated with multirate systems. In particular, it was illustrated that tall unblocked linear time-invariant systems are generically zero-free. Then, the zero properties of blocked systems associated with tall unblocked linear time-invariant systems were discussed and it was presented that tall blocked systems are generically zero-free. Finally, it was shown that tall blocked systems associated with multirate systems generically have no finite nonzero zeros. However, the behavior at $Z = 0$ and $Z = \infty$, turns out to be more complicated and we provided a conjecture which specifies a situation where tall blocked systems always have a zero at $z = 0$. As part of our future work, we intended to provide a formal proof for Conjecture 20. Moreover, we intend to generalize the results of this paper in respect of the output rates. More specifically, we are interested in the general case where there are two output streams, one available every ν time instants and the other every $\overline{\nu}$ time instants, with ν and $\overline{\nu}$ coprime integers with neither equal to 1.

Acknowledgements

The authors would like to thank Professor Manfred Deistler from the Department of Mathematical Methods in Economics, Technical University of Vienna for his valuable comments. Furthermore, support by the ARC Discovery Project Grant DP1092571, by the FWF (Austrian Science Fund) under contracts P17378 and P20833/N18 and by the Österreichische Forschungsgemeinschaft is gratefully acknowledged.

Bibliography

[1] B. D. O. Anderson and M. Deistler. Properties of zero-free transfer function matrices. *SICE Journal of Control, Measurement and System Integration*, 82(4):284–292, 2007. Cited pp. 33 and 47.

[2] S. Bittanti. Deterministic and stochastic linear periodic systems. *Lecture Notes in Control and Information Sciences*, 86:141–182, 1986. Cited p. 32.

[3] S. Bittanti and P. Colaneri. *Periodic Systems: Filtering and Control*. Springer, 2009. Cited pp. 32, 39, and 40.

[4] P. Bolzern, P. Colaneri, and R. Scattolini. Zeros of discrete-time linear periodic systems. *IEEE Transactions on Automatic Control*, 31(11):1057–1058, 1986. Cited p. 32.

[5] T. Chen and B. A. Francis. *Optimal Sampled-Data Control Systems*. Springer, 1995. Cited p. 32.

[6] W. Chen, B. D. O. Anderson, M. Deistler, and A. Filler. Properties of blocked linear system. *Proceedings of the International Federation of Automatic Control Conference*, pages 4558–4563, 2011. Cited pp. 32 and 40.

[7] P. Colaneri and S. Longhi. The realization problem for linear periodic systems. *Automatica*, 31(5):775–779, 1995. Cited pp. 32 and 40.

[8] M. Deistler, B. D. O. Anderson, A. Filler, C. Zinner, and W. Chen. Generalized linear dynamic factor models: An approach via singular autoregressions. *European Journal of Control*, 3:211–224, 2010. Cited p. 31.

[9] A. Filler. *Generalized dynamic factor models – structure theory and estimation for single frequency and mixed frequency data*. PhD thesis, Vienna University of Technology, 2010. Cited p. 34.

[10] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82(4):540–554, 2000. Cited p. 31.

[11] O. M. Grasselli and S. Longhi. Zeros and poles of linear periodic multivariable discrete-time systems. *Circuits, Systems, and Signal Processing*, 7:361–380, 1988. Cited p. 32.

[12] J. P. Hespanha. *Linear systems theory*. Princeton University Press, 2009. Cited p. 33.

[13] T. Kailath. *Linear systems*. Prentice-Hall, 1980. Cited p. 33.

[14] A. Raknerud, T. Skjerpen, and A. R. Swensen. Forecasting key macroeconomic variables from a large number of predictors: A state space approach. Discussion Papers 504, Research Department of Statistics Norway, 2007. Cited p. 31.

[15] C. Schumacher and J. Breitung. Real-time forecasting of GDP based on a large factor model with monthly and quarterly data. Discussion Paper Series 1: Economic Studies 33, Deutsche Bundesbank, Research Centre, 2006. Cited p. 31.

[16] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, 1993. Cited p. 32.

[17] P. Van Dooren. The computation of Kronecker's canonical form of a singular pencil. *Linear Algebra and Its Applications*, 27:103–140, 1979. Cited pp. 41 and 42.

[18] M. Zamani and B. D. O. Anderson. On the zeros properties of linear discrete-time systems with multirate outputs. *Proceedings of the American Control Conference*, pages 5182–5187, 2012. Cited pp. 32, 33, 44, and 47.

[19] M. Zamani, W. Chen, B. D. O. Anderson, M. Deistler, and A. Filler. On the zeros of blocked linear systems with single and mixed frequency data. *Proceedings of the IEEE Conference on Decision and Control*, pages 4312–4317, 2011. Cited pp. 32, 33, 35, 36, and 47.

[20] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, 1996. Cited p. 45.

Behavior of responses of monotone and sign-definite systems

David Angeli
Electrical & Electronic Eng.
Imperial College London, U.K.
and Dip. Sistemi e Informatica
Università di Firenze, Italy
d.angeli@imperial.ac.uk

Eduardo D. Sontag
Dept. of Mathematics
Rutgers University, NJ, USA
sontag@math.rutgers.edu

Abstract. This paper study systems with sign-definite interactions between variables, providing a sufficient condition to characterize the possible transitions between intervals of increasing and decreasing behavior.

1 Introduction

We consider systems with inputs and outputs

$$\dot{x} = f(x,u), \;\; y = h(x) \tag{1}$$

for which the entries of the Jacobian of f and h with respect of x and u have a constant sign. For such systems, we provide a graph-theoretical characterization of the possible transitions between intervals of increasing and decreasing behavior of state variables (or output variables). A particular case is that of monotone systems, for which it follows that only monotonic behavior can occur, provided that the input is monotonic and the initial state is a steady state. These results, although very simple to prove, are very useful when invalidating models in situations, such as in systems molecular biology, where signs of interactions are known but precise models are not. We also provide a discussion illustrating how our approach can help identify interactions in models, using information from time series of observations.

1.1 Notations and definitions

We assume in (1) that states $x(t)$ evolve on some subset $X \subseteq \mathbb{R}^n$, and input and output values $u(t)$ and $y(t)$ belong to subsets $U \subseteq \mathbb{R}^m$ and $Y \subseteq \mathbb{R}^p$ respectively. The maps $f : X \times U \to \mathbb{R}^n$ and $h : X \to Y$ are taken to be continuously differentiable, in the sense that they may be extended as $\mathcal{C}^1$ functions to open subsets, and technical conditions on invariance of X are assumed, [1]. (Much less can be assumed for many results, so long as local existence and uniqueness of solutions is guaranteed.) An *input* is a signal $u : [0,\infty) \to U$ which is measurable and bounded on finite intervals (in some of our results, we assume that $u(t)$ is differentiable on t). We write $\varphi(t,x_0,u)$ for the solution of the initial value problem $\dot{x}(t) = f(x(t),u(t))$ with $x(0) = x_0$, or just $x(t)$ if x_0 and u are clear from the context, and $y(t) = h(x(t))$. See [4] for more on i/o systems. For simplicity of exposition, we make the blanket assumption that solutions do not blow-up on finite time, so $x(t)$ (and $y(t)$) are defined for all $t \geq 0$. Given three partial orders on X,U,Y (we use the same symbol $\leq$ for all three orders), a *monotone I/O system (MIOS)*, with respect to these partial orders, is a system (1) such that h is a monotone map (it preserves order) and, for all initial states x_1, x_2 and all all inputs u_1, u_2, the following property holds: if $x_1 \leq x_2$ and $u_1 \leq u_2$ (meaning that $u_1(t) \leq u_2(t)$ for all $t \geq 0$), then $\varphi(t, x_1, u) \leq \varphi(t, x_2, u_2)$ for all $t \geq 0$. Here we consider partial orders induced by closed proper cones $K \subseteq \mathbb{R}^\ell$, in the sense that $x \leq y$ iff $y - x \in K$. The cones K are assumed to have a nonempty interior and are pointed, i.e. $K \bigcap -K = \{0\}$.

The most interesting particular case is that in which K is an *orthant* cone in $\mathbb{R}^n$, i.e. a set S_ε of the form $\{x \in \mathbb{R}^n \,|\, \varepsilon_i x_i \geq 0\}$, where $\varepsilon_i = \pm 1$ for each *i. Cooperative systems* are by definition systems that are monotone with respect to orthant cones. For such cones, there is a useful test for monotonicity, which generalizes Kamke's condition from ordinary differential equations [3] to i/o systems. Let us denote by $\sigma(x)$ the usual sign function: $\sigma(x) = 1, 0, -1$ if $x > 0$, $= 0$, or < 0 respectively. Suppose that

$$\sigma\left(\frac{\partial f_i}{\partial x_j}(x,u)\right) \text{ is constant } \forall\, i \neq j,\ \forall\, x \in X,\ \forall\, u \in U \tag{2}$$

and similarly

$$\sigma\left(\frac{\partial h_i}{\partial x_j}(x)\right) \text{ is constant } \forall\, i, j,\ \forall\, x \in X$$

(subscripts indicate components) We also assume that X is convex. We then associate a directed graph G to the given MIOS, with $n + m + p$ nodes, and edges labeled "+" or "−" (or ±1), whose labels are determined by the signs of the appropriate partial derivatives (ignoring diagonal elements of $\partial f / \partial x$). An undirected loop in G is a sequence of edges transversed in either direction, and the *sign* of an undirected loop is defined by multiplication of signs along the loop. (See e.g. [2] for more details.) Then, it is easy to show that a system is monotone with respect to *some* orthant cones in $\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^p$ if and only if there are no negative undirected loops in G.

1.2 Monotone responses

Suppose now that our system (1) is monotone with respect to an orthant order, and with a scalar input ($U \subseteq \mathbb{R}$ with the usual order). We will prove below that, starting from a steady state, if an external input is a either non-increasing or non-decreasing in time (for example, a step function), then the system has the property that the response of every node is monotonic as well. That is to say, each node must respond as a non-decreasing function, like the one shown in the left panel of Figure 1, or a non-increasing function. A biphasic response like the one shown in the right panel of Figure 1 can never occur, at any of the nodes. In fact, we will show a stronger result, valid for any monotone system and any input that is non-decreasing in time with respect to the order structure in U, $u(t_1) \leq u(t_2)$ for all $t_1 \leq t_2$: states then non-decreasing in time with respect to the order structure in X, $x(t_1) \leq x(t_2)$ for all $t_1 \leq t_2$. For the special case of orthant orders, this means that each coordinate of the state will either satisfy $x_i(t_1) \leq x_i(t_2)$ for all $t_1 \leq t_2$ or $x_i(t_1) \geq x_i(t_2)$ for all $t_1 \leq t_2$ ($i \in \{1, 2, \ldots, n\}$). Analogously, if inputs are non-increasing, that is, $u(t_2) \leq u(t_1)$ for all $t_1 \leq t_2$, then, by reversing the orders in X and U, we obtain a new monotone system in which now $u(t)$ is non-decreasing, and therefore the same conclusions hold (with

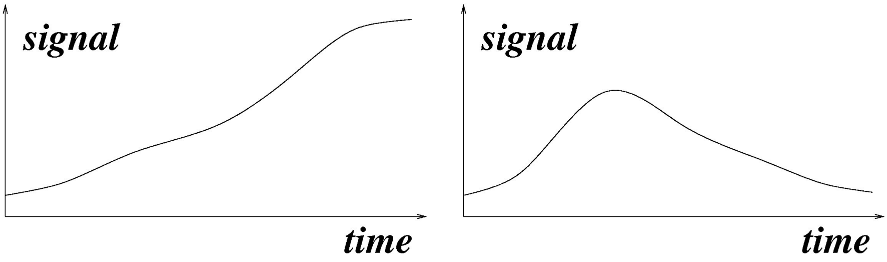


Figure 1: Monotonic and biphasic responses

reversed orders). Let $\varphi(t,x_0,v)$ denote the solution of $\dot{x} = f(x,u)$ at time $t > 0$ with initial condition $x(0) = x_0$ and input signal $v = v(t)$.

Theorem 1. *Suppose that* (1) *is a monotone I/O system. Pick an input v that is nondecreasing in time with respect to the partial order in U, and an initial state x_0 that is a steady state with respect to $v_0 = v(0)$, that is, $f(x_0,v_0) = 0$. Then, $x(t) = \varphi(t,x_0,v)$ is non-decreasing with respect to the partial order in X. Also, the output $y(t) = h(x(t))$ is nondecreasing.*

The proof is given in Section 3.

1.3 Feedback and feedforward architectures

Theorem 1 can be specialized to the study of responses from a single input of interest to a single output. The idea is to let only one input monotonically vary, while other input signals are kept constant at their equilibrium value. This allows to establish monotonicity of I/O responses beyond the case of cooperative systems which is studied in Theorem 1. In order to state the result we need the following graph-theoretic definitions.

Given a directed graph $(\mathcal{V},\mathcal{E} \subset \mathcal{V}\times\mathcal{V})$, we define the *accessible* subgraph from a node $v \in \mathcal{V}$ to be

$$Acc(v) = (\mathcal{V}_v,\mathcal{E}_v)$$

defined as follows:

$$\mathcal{V}_v = \{w \in \mathcal{V} : \exists \text{ directed path from } v \text{ to } w\}$$

while $\mathcal{E}_v = \mathcal{E} \cap \mathcal{V}_v \times \mathcal{V}_v$. We define the *co-accessible* subgraph to a node $z \in \mathcal{V}$ to be:

$$coAcc(z) = (\mathcal{V}_z,\mathcal{E}_z)$$

where:

$$\mathcal{V}_z = \{w \in \mathcal{V} : \exists \text{ directed path from } w \text{ to } z\}$$

and $\mathcal{E}_z = \mathcal{E} \cap \mathcal{V}_z \times \mathcal{V}_z$.

Intuitively, given an input node v_i and an output node v_o in $\mathcal{V}$, in order to investigate monotonicity of the input-output response from the associated input signal to the corresponding output signal, it is enough to consider the graph:

$$\mathcal{G}_{i/o} := (\mathcal{V}_{i/o}, \mathcal{E}_{i/o}) = \mathrm{Acc}(v_i) \cap \mathrm{coAcc}(v_o).$$

The crucial features of this graph that may prevent monotonicity of the response is existence of two or more directed paths from v_i to v_o with inconsistent sign. Such paths can only exist if the graph $\mathcal{G}_{i/o}$ exhibits incoherent feedforward loops (IFFL's) and/or negative directed feedback loops. This condition may be verified for two nodes v_i and v_o even if the overall system is not monotone. For example, Fig. 2 shows a system that (a) is not monotone yet (b) has no IFFL's nor negative feedback loops. However, such a counterexample does not contradict our assertion, since we are interested in knowing how one input (affecting only one node) affects any given particular output node. Indeed, if all we ask is that input/output question, then the following is true:

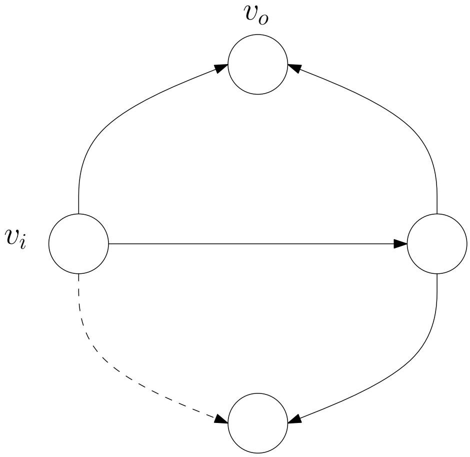


Figure 2: The graph of a non-monotone system fulfilling I/O monotonicity conditions. The dashed edge is negative and all other edges are positive

Theorem 2. *Suppose that* (1) *is a monotone I/O system, with scalar inputs and outputs ($U \subseteq \mathbb{R}$ and $Y \subseteq \mathbb{R}$ with the usual orders), and that the parities of any two directed paths from the input node to the output node are the same. Then, if the system is initially at some equilibrium, the response to a monotonic input is monotonic.*

Observe that "paths" include feedforward loops as well as closed loops in which a cycle occurs. The simple proof is omitted here; it relies upon the pruning all nodes that do not lie in any such path, reducing to the monotone case.

1.4 More general systems with sign-definite Jacobians

In this section, we relax the monotonicity assumptions. We assume that (2) holds. Our goal is to understand, given a certain input with a particular monotone trend, that is such that $\mathrm{sign}(\dot{u}(t))$ is constant in time, what are the possible shapes that solutions $x(t,x_0,u)$ can take, and in particular, what $\mathrm{sign}(\dot{x}(t))$ may look like. Let

$$\mathcal{V} := \{-1,0,1\}^{n+m},$$

which we regard as the set of all possible sign-patterns of vectors $[\dot{x}',\dot{u}']' \in \mathbb{R}^{n+m}$, and define a matrix $J \in \{-1,0,1\}^{n\times(n+m)}$ as follows (σ is applied to each entry):

$$\sigma\left(\begin{bmatrix} 0 & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \cdots & \frac{\partial f_1}{\partial x_n} & \frac{\partial f_1}{\partial u_1} & \cdots & \frac{\partial f_1}{\partial u_m} \\ \frac{\partial f_2}{\partial x_1} & 0 & \frac{\partial f_2}{\partial x_3} & \cdots & \frac{\partial f_2}{\partial x_n} & \frac{\partial f_2}{\partial u_2} & \cdots & \frac{\partial f_2}{\partial u_m} \\ \vdots & & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & & & \ddots & \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_{n-1}} & 0 & \frac{\partial f_n}{\partial u_1} & \cdots & \frac{\partial f_n}{\partial u_m} \end{bmatrix}\right)$$

Let

$$\mathcal{V}_0^2 := \left\{ (v_1,v_2) \in \mathcal{V}^2 \text{ s.t. } \sum_{i=1}^{n} |v_{1i} - v_{2i}| = 1 \right\}$$

(in other words, pairs of elements v_1 and v_2 which differ in exactly one position, located among their first n coordinates, and this difference is between 0 and 1, or between -1 and 0). For such pairs, we denote by $i_{v_1,v_2} \in \{1,2,\ldots,n\}$ the uniquely defined integer for which $v_{1i} \neq v_{2i}$. Regarding $\mathcal{V}$ as a set of vertices in a directed graph, we denote by $\mathcal{E} \subset \mathcal{V}_0^2$ the set of edges for which

$$\exists k \in \{1,\ldots,n+m\} \text{ s.t. } \quad J_{i_{v_1,v_2}k} v_{1k} (v_{2i_{v_1,v_2}} - v_{1i_{v_1,v_2}}) = 1. \tag{3}$$

Intuitively, in equation (3) we allow a directed edge pointing from node v_1 to node v_2 only if the nodes differ by a single entry, the i-th one, and if among the input/states variables that affect $\dot{x}_i$ (with the exception of x_i itself), at least one has an influence on $\dot{x}_i$ which is equal in sign to that of the jump $v_{2i} - v_{1i}$).

In Section 3, we prove the following result:

Theorem 3. *Let $I_1 < I_2$ be disjoint non-empty intervals of the real line such that $I = I_1 \cup I_2$ is also an interval. Let $x(t) : I \to X$ be a solution of (1) corresponding to the $\mathcal{C}^1$ input u of constant sign pattern $\sigma(\dot{u}(t))$. Assume that there exists v_1 and v_2 in $\mathcal{V}$ such that $\sigma([\dot{x}(t)',\dot{u}(t)']) = v_1$ for all $t \in I_1$ and $\sigma([\dot{x}(t)',\dot{u}(t)']) = v_2$ for all $t \in I_2$ and $|v_1 - v_2| = 1$. Then $(v_1,v_2) \in \mathcal{E}$.*

Note that we are allowing either interval to consist of only one point. Theorem 3 can be used to infer the potential shapes of solutions of nonlinear systems with sign-definite Jacobians, subject to piecewise monotone inputs. It generalizes Theorem 1, in the following sense. Suppose that our system is monotone with respect to the standard order, i.e. with respect to the cone $K = S_\varepsilon$, where $\varepsilon = (1,1,\ldots,1)$. Then (Kamke conditions) the sign Jacobian matrix J has all its elements non-negative. In that case, Theorem 3 clearly implies that the two subsets of nodes $\{0,1\}^{n+m}$ and $\{0,-1\}^{n+m}$ are forward-invariant in the graph with edges $\mathcal{E}$. This implies, in particular: (1) if the input is non-decreasing and if we start from a steady state (first n coordinates of edges are zero), then all reachable nodes have non-negative coordinates (that is to say, the solutions of the system are non-decreasing), and (2) if the input is non-increasing, then nodes are non-positive (solutions of the system are non-increasing), thus recovering the conclusions of Theorem 1.

1.5 A toy example

To illustrate the applicability of Theorem 3 we consider the bidimensional nonlinear system:

$$\begin{array}{rcl} \dot{x}_1 & = & ux_1 - k_1x_1x_2 \\ \dot{x}_2 & = & -k_2x_2 + k_3x_1x_2 \end{array} \tag{4}$$

with state space $X = (0,+\infty)^2$ and input taking values in $(0,+\infty)$ and k_1,k_2,k_3 being arbitrary positive coefficients. Notice that this can be interpreted as a model of predator-prey interactions with the reproduction rate of preys being an exogenous input u. Obviously the system is not cooperative due to the presence of a negative feedback loop. The J matrix in this case is given by:

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

Next we build the graph $(\mathcal{V},\mathcal{E})$ with nodes:

$$\mathcal{V} = \{-1,0,1\}^3.$$

Let us focus on increasing inputs. This means we restrict our attention to nodes of the type $\{-1,0,1\}^2 \times \{1\}$ and for the sake of simplicity we may drop the $\dot{u}$ label in Fig. 3. This represents all the edges allowed by Theorem 3. Notice that commutations in the sign of $\dot{x}_2(t)$ (the predators) are only allowed in order to match the sign of $\dot{x}_1(t)$. This restricts the possible sign-patterns of $\dot{x}(t)$ which are compatible with a model of this kind even without assuming any knowledge of the specific values of the k_is (provided their sign is known a priori).

The previous example also suggests the possibility of introducing a reduced graph, which we define by considering a reduced set of nodes and a new set of edges. In particular, we may let: $\mathcal{G}_{red} = (\mathcal{V}_{red},\mathcal{E}_{red})$, where $\mathcal{V}_{red} = \{1,-1\}^{n+m}$, $\mathcal{E}_{red} = \{(v_1,v_2) \in \mathcal{V}_{red}^2 : \exists$ path of length 2 in $\mathcal{G}$ from v_1 to $v_2\}$. This graph represents, for a given and fixed sign pattern of the input variable, the set of all possible transitions between sets $\{x : f(x,u) \in \mathcal{O}\}$, where $\mathcal{O}$ denotes an arbitrary *closed* orthant and edges are only allowed between neighboring orthants (that is orthants sharing a face of maximal dimension). In particular, the orthant $\{x : f(x,u) \in \mathcal{O}\}$ where $\mathcal{O} = \mathrm{diag}(v)[0,+\infty)^n$, and

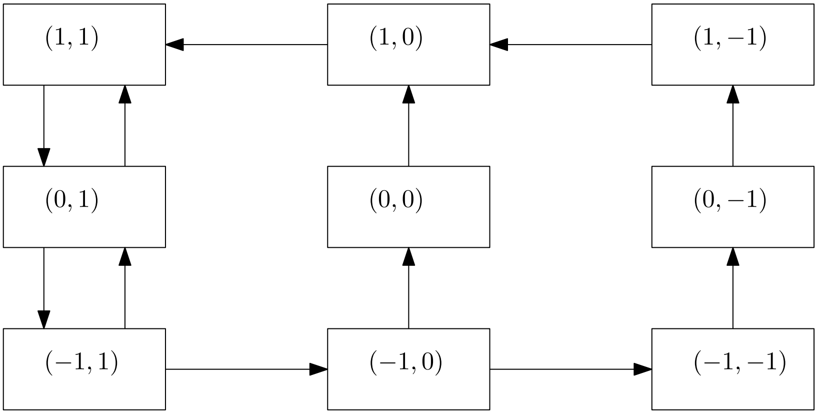


Figure 3: Graph of allowed transitions for increasing inputs

v is an arbitrary element of $\{1,-1\}^n$ is associated to the node v. It is straightforward to see that

$$\mathcal{E}_{red} = \{(v_1,v_2) \in \mathcal{V}_{red}^2 : \exists k \in \{1,\dots,n+m\} \text{ s.t. } \\ J_{i_{v_1,v_2}k} v_{1k}(v_{2i_{v_1,v_2}} - v_{1i_{v_1,v_2}}) = 2\},$$

where with a slight abuse of notation i_{v_1,v_2} denotes the unique index i such that $|v_{1i} - v_{2i}| = 2$.

2 Identification of signed interactions

In the following we exploit the results of previous Sections, and in particular Theorem 3, in order to formulate and discuss an algorithm for identification of signed interactions based on available measured data. This is a systematic tool for hypothesis generation. The method assumes sign definite interactions between variables and allows, under such qualitative constraints, to find the family of minimal signed graphs which are compatible with given measured data. Our discussion in this section will be done very informally. A future paper will provide more precise formulations.

For the sake of simplicity all variables are assumed to be measured continuously so that no issue arises of what has been the intersample behaviour of individual variables and whether or not the adopted sampling time is sufficiently small to unambiguously detect changes of sign in the derivatives of the considered set of variables. Also we assume that at most one variable can switch at any given time (this assumption is reasonable only when there are no conservation laws involving exactly two variables).

The algorithm is particularly flexible as it allows to generate several plausible scenarios compatible with an initial hypothesis $\mathcal{H}_0$ which gathers all the apriori information available, namely all the interactions between variables which have been validated and invalidated by other means. In its basic formulation it assumes that all variables are known and available for measurement.

The following definitions are useful in order to precisely formulate the algorithm. Notice that we will identify a graphical object which is different from the graphs previously described.

Definition 4. A signed graph $\mathcal{G}$ is a triple $\{\mathcal{V},\mathcal{E}_+,\mathcal{E}_-\}$, in which $\mathcal{V}$ is a finite set of nodes (corresponding to the variables of the system), $\mathcal{E}_+ \subset \mathcal{V}\times\mathcal{V}\backslash\{(v,v) : v \in \mathcal{V}\}$ is the set of positive edges, each corresponding to directed excitatory influence of one variable to another, and $\mathcal{E}_- \subset \mathcal{V}\times\mathcal{V}\backslash\{(v,v) : v \in \mathcal{V}\}$ is the set of negative edges, corresponding to directed inhibitory influences.

Notice that variables may be states and inputs. In this respect it is convenient to partition $\mathcal{V}$ as $\mathcal{V}_s \cup \mathcal{V}_i$, with $\mathcal{V}_s \cap \mathcal{V}_i = \varnothing$ denoting the set of nodes corresponding to state variables and input variables respectively. The assumption of signed interactions means that $\mathcal{E}_+ \cap \mathcal{E}_- = \varnothing$. Notice also that we do not consider self-loops in our graphs (and, consequently, no assumption of signed self-interaction is made). We say that a graph is compatible with the observed data if all sign-switches of derivatives in the data are allowed by the sign-pattern of the adjacency matrix of $\mathcal{G}$ according to Theorem 3. Moreover, we say that a signed graph $\tilde{\mathcal{G}} = \{\mathcal{V},\tilde{\mathcal{E}}_+,\tilde{\mathcal{E}}_-\}$ is an edge-subgraph of $\mathcal{G}$ if $\tilde{\mathcal{E}}_+ \subset \mathcal{E}_+$ and $\tilde{\mathcal{E}}_- \subset \mathcal{E}_-$. If at least one inclusion is strict we say that it is a proper edge-subgraph. We also say that $\mathcal{G}$ is an edge-supergraph of $\tilde{\mathcal{G}}$. An apriori hypothesis $\mathcal{H}$ is a signed graph with 2 types of signed edges $\{\mathcal{V},\mathcal{E}^h_+,\mathcal{E}^h_-,\mathcal{F}^h_+,\mathcal{F}^h_-\}$ where $\mathcal{E}^h_+$ and $\mathcal{E}^h_-$ are respectively positive and negative edges which have already been validated (and are therefore known to exist in the graph of the system being identified), while $\mathcal{F}^h_+$ and $\mathcal{F}^h_-$ are forbidden positive and negative edges respectively.

Notice that $\mathcal{E}^h_+ \cap \mathcal{E}^h_- = \varnothing$, while the same is not necessarily true for $\mathcal{F}^h_+$ and $\mathcal{F}^h_-$. For instance, if a certain variable is known to be an input of the system, then all its incoming edges, both positive and negative should be listed as forbidden.

Definition 5. A graph $\mathcal{G}$ is said to be a minimal graph compatible with data and with hypothesis $\mathcal{H}$ if no proper edge-subgraph of $\mathcal{G}$ exists that is both compatible with the data and an edge-supergraph of $\mathcal{H}$ with $\mathcal{F}^h_+ \cap \mathcal{E}_+ = \varnothing$ and $\mathcal{F}^h_- \cap \mathcal{E}_- = \varnothing$.

The first algorithm we discuss below allows to generate all minimal signed graphs compatible with the measured data and the given apriori hypothesis $\mathcal{H}$, (which could be empty, namely $\mathcal{H} = \{\mathcal{V},\varnothing,\varnothing,\varnothing,\varnothing\}$). As more than one such graph may exist, depending on the data available, the algorithm creates a number of plausible scenarios by storing them in a tree, starting from the root node $\mathcal{H}$. The parent of each node is a proper edge-subgraph of all of its children. Measured data is scanned from initial to final time. Each time a sign switch is detected all leaves of the current tree are checked to see whether the switch is compatible with the graphs they represent. If so, nothing is done; otherwise, a single edge is added in order to restore compatibility of data with the graph. If more than one edge may be capable of restoring such compatibility multiple children are created for the considered parent node. If no such edge exists, (namely because the constraint $\mathcal{E}_+ \cap \mathcal{E}_- = \varnothing$ does not allow it), then that node is labeled as *Invalidated.*

In the following we denote by $\mathcal{L}(\mathcal{T})$ the set of leaves of a tree $\mathcal{T}$. Notice that, for the sake of simplicity, we assume that at each time t at most one variable may switch the sign of its derivative.

1. Let $\mathcal{H} = (\mathcal{V}, \mathcal{E}_+^h, \mathcal{E}_-^h)$ be the root of the tree $\mathcal{T}$;
2. Let $t_1, t_2, \dots t_N$ denote the time instants at which sign switches in state variable derivatives are detected;
3. For $i = 1 \dots N$
4. For $\mathcal{H} \in \mathcal{L}(\mathcal{T})$
5. If $\mathcal{H}$ is labeled 'Invalidated' or 'Redundant' do nothing, else:
6. If variable $v \in \mathcal{V}_s$ switches its derivative from positive to negative [from negative to positive] at time t_i then:
    - Check if there exists an edge in $\mathcal{E}_+$ from a node w with negative [positive] derivative (at t_i) to v or if there exists an edge in $\mathcal{E}_-$ from a node w with positive [negative] derivative (at t_i) to v;
    - If the check succeeds then do nothing. If the check fails then for all nodes u with positive derivative, such that (u, v) does not belong to $\mathcal{E}_+ \cup \mathcal{F}_-^h$, add the edge (u, v) to $\mathcal{E}_-$ and attach as a son to $\mathcal{H}$ the newly created graph;
    - Similarly, if the check fails, for all nodes u with negative derivative, such that (u, v) does not belong to $\mathcal{E}_- \cup \mathcal{F}_+^h$, add the edge (u, v) to $\mathcal{E}_+$ and attach as a son to $\mathcal{H}$ the newly created graph;
    - If no such nodes as in the previous two items exist, then label $\mathcal{H}$ as 'Invalidated';
7. End For $\mathcal{H}$;
8. Label all leaves of $\mathcal{T}$ that are proper edge-subgraph of other leaves as 'Redundant';
9. label as 'Redundant' all leaves except one of those which are equal to one another;
10. End For i;

The algorithm terminates with the set of non invalidated and non redundant leaves representing all minimal sign-definite graphs which are compatible with the initial hypothesis.

To illustrate the algorithms we apply it to synthetic data generated by numerically integrating the following differential equation:

$$\begin{aligned} \dot{x}_1 &= -x_1 + x_1 x_2 \\ \dot{x}_2 &= x_2 x_3 - x_1 x_2 \\ \dot{x}_3 &= x_3 - 1.2 x_2 x_3 . \end{aligned} \tag{5}$$

This can be seen as a toy model of an ecosystem comprising 3 interacting species: Predators, Vegetarians and Vegetables, (x_1,x_2 and x_3 respectively). Clearly the algorithm does not assume knowledge of the 'nature' of the variable being measured

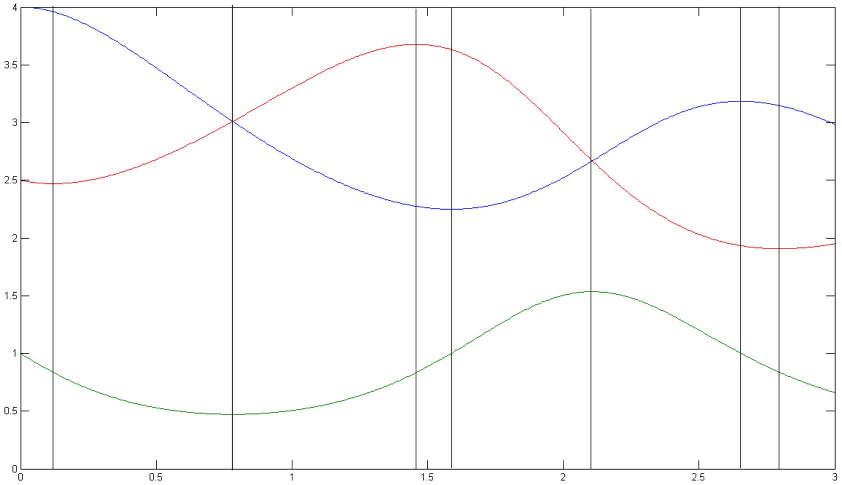


Figure 4: Simulated species data. Blue plot (largest value at $t = 0$) denotes predators, red vegetables, and green (smallest value at $t = 0$) vegetarians.

and in fact the goal of the identification is precisely to find out the sign of interactions between such species, that is the role of each species in the ecosystem. The measured data is shown in Fig. 4, using 3 different colors for the 3 variables.

Notice that 7 sign switches of derivatives are detected in the finite time window considered and these are highlighted by vertical lines in the picture so as to emphasize the order in which variables switch their monotonicity. We start with the empty hypothesis comprising 3 nodes (labeled in the graph given in Figure 4 by colors: blue (bottom left node) = predators, green (right node) = vegetarians, and red (top node) = vegetables), and no validated nor invalidated edges. The execution of the algorithm is shown in Fig. 5 Notice that the algorithm generates two minimal graphs compatible with the measured data. Two edges appear in both graphs and are therefore validated and should be present in any set of differential equations generating such monotonicity patterns. The remaining edge can be picked from any of the two scenarios. In fact the model used to generate the data is a supergraph of both scenarios and is given by their union. This, of course, need not always be the case. Extra data and experiments would be needed in order to refine the model. In fact, the outcome of the algorithm may be used in order to design further experiments targeting specific edges of the graph.

3 Proofs

Proof of Theorem 1

Since $v(t)$ is non-decreasing, we have that $v(t) \geq v(0)$ (coordinate-wise), so that, by comparison with the input that is identically equal to $v(0)$, we know that

$$\varphi(h, x_0, v) \geq \varphi(h, x_0, v_0)$$

where by abuse of notation v_0 is the function that has the constant value v_0. We used the comparison theorem with respect to inputs, with the same initial state.

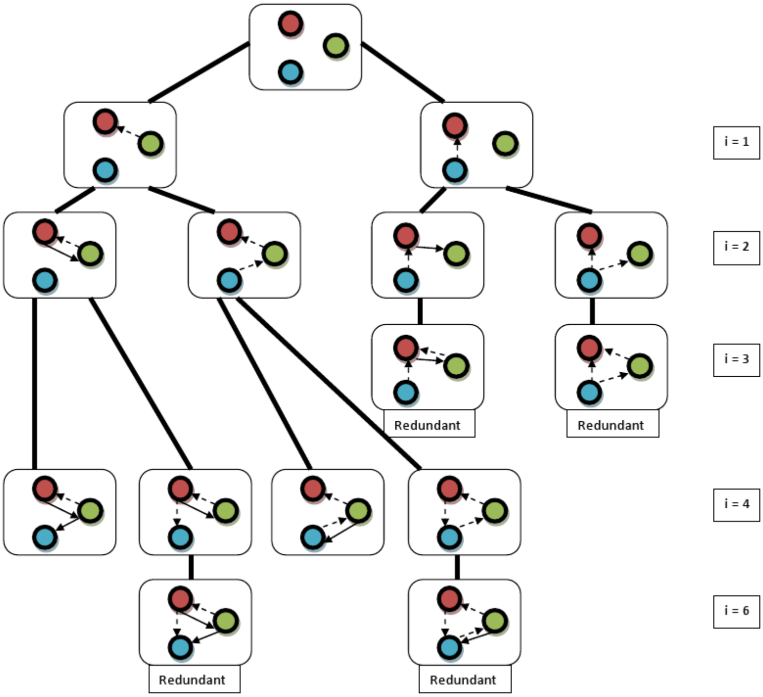


Figure 5: Generation of minimal graphs compatible with available data. Dashed arrows indicate negative edges.

The assumption that the system starts at a steady state gives that $\varphi(h,x_0,v_0) = x_0$. Therefore:

$$x(h) \geq x(0) \qquad \text{for all } h \geq 0\,. \tag{6}$$

Next, we consider any two times $t \leq t+h$. We wish to show that $x(t) \leq x(t+h)$. Using (6) and the comparison theorem with respect to initial states, with the same input, we have that:

$$x(t+h) = \varphi(t,x(h),v_h) \geq \varphi(t,x(0),v_h)\,,$$

where v_h is the "tail" of v, defined by: $v_h(s) = v(s+h)$. On the other hand, since the function v is non-decreasing, it holds that $v_h \geq v$, in the sense that the inputs are ordered: $v_h(t) \geq v(t)$ for all $t \geq 0$. Therefore, using once again the comparison theorem with respect to inputs and with the same initial state, we have that

$$\varphi(t,x(0),v_h) \geq \varphi(t,x(0),v) = x(t)$$

and thus we proved that $x(t+h) \geq x(t)$. So x is a non-decreasing function. The conclusion for outputs $y(t) = h(x(t))$ follows by monotonicity of h. □

Proof of Theorem 3

Consider the function

$$z(t) := \dot{x}_i(t) = f(x(t),u(t)).$$

Differentiating with respect to time we have by the chain rule:

$$\dot{z}(t) = \frac{\partial f}{\partial x}(x(t),u(t))\dot{x}(t) + \frac{\partial f}{\partial u}(x(t),u(t))\dot{u}(t)$$

Looking at the equation for the i-th component of z yields:

$$\begin{aligned}\dot{z}_i(t) &= \sum_j \frac{\partial f_i}{\partial x_j}(x(t),u(t))z_j(t) + \sum_{j=1}^m \frac{\partial f_i}{\partial u_j}(x(t),u(t))\dot{u}_j(t)\\ &= a(t)z_i(t) + b(t)\end{aligned}$$

provided we define:

$$a(t) = \frac{\partial f_i}{\partial x_i}(x(t),u(t))$$

and:

$$b(t) = \sum_{j\neq i} \frac{\partial f_i}{\partial x_j}(x(t),u(t))z_j(t) + \sum_{j=1}^m \frac{\partial f_i}{\partial u_j}(x(t),u(t))\dot{u}_j(t).$$

Let v_1 and v_2 be as in the statement of the theorem, and let $i = i_{v_1,v_2}$. There are four cases to consider:

1. $v_{1i} = 0$ and $v_{2i} = 1$
2. $v_{1i} = 0$ and $v_{2i} = -1$
3. $v_{1i} = -1$ and $v_{2i} = 0$
4. $v_{1i} = 1$ and $v_{2i} = 0$.

Case 1. We have $z_i(t) = 0$ for all $t \in I_1$ and $z_i(t) > 0$ for all $t \in I_2$. It follows that I_2 cannot be a one-point interval. Let $t_2 := \inf I_2$, and note that $z_i(t_2) = 0$. From the variation of parameters formula for the solution of $\dot{z}_i(t) = a(t)z_i(t) + b_i(t)$, it follows that if $z_i(t_2) = 0$ and $z_i(t) > 0$ for an open interval $[0, t_2+\varepsilon)$, then there must exist some $\tau \in I_2$ such that $b(\tau) > 0$. Thus, at least one of the terms in the definition of $b(\tau)$ must be positive, which means that

$$J_{i_{v_1,v_2}k} v_{2k} = 1\,.$$

Note that this k is by definition not equal to i, so $v_{2k} = v_{1k}$ (because v_1 and v_2 differ only on their ith entry). Thus $J_{i_{v_1,v_2}k} v_{1k} = 1$. Moreover, in this case $v_{2i} - v_{1i} = 1 - 0 = 1$, so it follows that $J_{i_{v_1,v_2}k} v_{1k}(v_{2i_{v_1,v_2}} - v_{1i_{v_1,v_2}}) = 1$, as claimed.

Case 2. An analogous argument gives that there is some k such that $J_{i_{v_1,v_2}k} v_{1k} = J_{i_{v_1,v_2}k} v_{2k} = -1$, but now $v_{2i} - v_{1i} = -1 - 0 = -1$, so again $J_{i_{v_1,v_2}k} v_{1k}(v_{2i_{v_1,v_2}} - v_{1i_{v_1,v_2}}) = 1$.

Case 3. Now we argue with the final-time problem $\dot{z}_i(t) = a(t)z_i(t) + b_i(t)$, $z_i(t_1) = 0$, where $t_1 = \sup I_1$. We conclude that there is some k such that $J_{i_{v_1,v_2}k} v_{1k} = 1$, and since $v_{2i} - v_{1i} = 0 - (-1) = 1$, we have $J_{i_{v_1,v_2}k} v_{1k}(v_{2i_{v_1,v_2}} - v_{1i_{v_1,v_2}}) = 1$.

Case 4. Analogously, $J_{i_{v_1,v_2}k} v_{1k} = -1$, $v_{2i} - v_{1i} = 0 - 1 = -1$, so $J_{i_{v_1,v_2}k} v_{1k}(v_{2i_{v_1,v_2}} - v_{1i_{v_1,v_2}}) = 1$. □

Acknowledgements

This research was supported in part by grants NIH 1R01GM086881 and AFOSR FA9550-11-1-0247.

Bibliography

[1] D. Angeli and E. Sontag. Monotone control systems. *IEEE Trans. Automat. Control*, 48(10):1684–1698, 2003. Cited p. 51.

[2] D. Angeli and E. Sontag. Multi-stability in monotone input/output systems. *Systems Control Lett.*, 51(3–4):185–202, 2004. Cited p. 52.

[3] H. Smith. *Monotone dynamical systems: An introduction to the theory of competitive and cooperative systems*, volume 41 of *Mathematical Surveys and Monographs*. AMS, 1995. Cited p. 52.

[4] E. Sontag. *Mathematical Control Theory. Deterministic Finite-Dimensional Systems*, volume 6 of *Texts in Applied Mathematics*. Springer, second edition, 1998. Cited p. 51.

Synchronization without periodicity

Roger Brockett
Harvard University
Cambridge MA, USA

Abstract. In this paper we study a model for synchronization where the solutions are only approximately periodic. More precisely, we present a model involving a small parameter for which conventional averaging theory does not predict the existence of a periodic solution and numerical studies show qualitative synchronization together with small amplitude irregular motion. In spite of this, it seems that the phase difference between the oscillations stays bounded for all time.

1 Introduction

Mathematically speaking, questions about synchronization are usually thought of as questions about the stability properties of a one dimensional manifold. It might be posed as follows. Given a compact manifold $X \subset \mathbb{R}^n$ and a closed curve $\Gamma \subset X$, together with a differential equation $\dot{x} = f(x)$ such that the manifold X is (locally) attracting, determine circumstances under which all solutions will approach Γ. Splitting the problem up this way is helpful because we can then treat it in pieces, posing the problem in terms of a vector field f such that solutions of $\dot{x} = f(x)$ approach X and an interaction term g such that g leaves invariant the manifold X such that solutions of $\dot{x} = f(x) + g(x)$ approach Γ. The existence of feedback controls which stabilize submanifolds has been investigated in the recent work of Mansouri [5] [6].

Since the work of Huygens in the mid 17th century synchronization has been the subject of much speculation and analysis but, for a variety of reasons, there has been an sharp uptick in interest in recent years. One reason stems from the fact that almost all computing and communication systems depend on the synchronization of dispersed signals and clocks; other reasons come from the importance of mode locking in laser physics as well as in more classical areas, as in the work of Kuramoto [4]. There are also various studies in the biological sciences involving collections of quasi independent agents such as fireflies which seem to display some form of synchronization. Along with this, there has also been considerable interest in quasi periodic motions that coexist but do not synchronize over time, especially in connection with the study of invariant tori in KAM theory.

One of the intriguing aspects of the apparent synchronization seen in physical problems is the apparent lack of sensitivity to the size of the coupling terms. This suggests that what one observes is the consequence of the buildup of small effects over time, as in integral control. In this paper we study a synchronization model based on the simplest possible nonintegrable terms. Our objective is to describe a general mechanism that results in a type of frequency synchronization and a form of phase locking. We will give a mathematical description of this mechanism and show that it is robust. Our basic system consists of oscillators that would oscillate autonomously in the absence of any coupling together with an interaction term consisting of nonintegrable terms that provide a measure of the phase differences. The coupled equations involve a small parameter ε and take the form

$$\ddot{x}+\varepsilon f(\dot{x})+x+\varepsilon^2(Q+\dot{x}x^\top-x\dot{x}^\top)x=0\ ;\quad x\in\mathbb{R}^n \tag{1}$$

When $\varepsilon=0$ the equations describe n decoupled oscillators having the same frequency. The symmetric matrix Q provides detunning of the oscillators and the problem is to show that the $x\dot{x}^\top-\dot{x}x^\top$ term restores synchronization. See [1, 2]. Numerical studies suggest that for small ε there is synchronization and mode locking but with a small irregular motion which averages out; the phase difference between the various oscillators is determined by Q but subject to a small jitter. One of our main points is that there is no solution to the averaging equations ordinarly used to establish periodic solutions.

This paper is dedicated to Uwe Helmke who has long been a leader in bringing new mathematical techniques to bear on problems in control.

2 Preliminaries on eigenvalue placement

In this section we establish two facts about linear algebra that are relevant to the later developments. Theorem 1 is, as far as I know, new and provides the essential motivation for the model studied here. Although remarkable in some ways, it is an easy consequence of the well known Schur-Horn theorem. Recall that the Schur-Horn polytope associated with a set of n real numbers $\{\lambda_1,\lambda_2,...,\lambda_n\}$ is defined as the convex hull of the $n!$ vectors, $v_1,v_2,\cdots v_{n!}$ where the v_i are obtained from the λ's by selecting some order for the λ's and treating them as the components of a vector in $\mathbb{R}^n$.

Theorem 1. *Given $Q=Q^\top$ with eigenvalues $\{\lambda_1,\lambda_2,...,\lambda_n\}$, and given any vector $[\mu_1,u_2,...,\mu_n]$ in the Schur-Horn polytope defined by the eigenvalues of Q, there exists $Z=-Z^\top$ such that the eigenvalues of $Q-Z$ are $[\mu_1,u_2,...,\mu_n]$. Moreover, if the eigenvalues of $Q+Z$ are real, they must lie in this polytope and if they lie on the boundary of the polytope between two distinct eigenvalues of Q they must be associated with an elementary divisor of degree two or higher.*

Proof. From the Schur-Horn theorem on the possible diagonals of symmetric matrices having a given spectrum, it is known that it is possible to find Θ so that the diagonal elements of $\Theta^\top Q\Theta$ viewed as a vector in $\mathbb{R}^n$, can be any element of the convex hull of the $n!$ vectors having the eigenvalues as components. Given Q choose Θ so that $\Theta^\top Q\Theta$ has the desired eigenvalues on the diagonal. Then choose Z_1 so as to make $\Theta^\top Q\Theta+Z_1$ upper triangular. In this case $Z=\Theta Z_1\Theta^\top$ is the desired Z. To show that the eigenvalues must lie in this polytope, we can apply Schur's argument. Order the eigenvalues of Q as $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_n$. Clearly $x^\top(H+Z)x\leq\lambda_1$ let $M_{(p)}$ denote the p^{th} compound of M and reason that $x^\top(Q+Z)_{(2)}x\leq\lambda_1+\lambda_2$ etc. The fact about the elementary divisors will follow from the next theorem. □

Theorem 2. *Let Q and Z be as above. If Q has no eigenvalue equal to $tr(Q/n)$ and if $Q+Z$ has all its eigenvalues equal to $tr(Q/n)$ then the degree of the minimal polynomial of $Q+Z$ is at least two.*

Proof. The eigenvalues of $Q-Z$ are the same as those of $Q+Z$. Let $(Q-Z)x=\lambda x$. Then $(Q+Z)x=2Zx$ must be nonzero because λ is not an eigenvalue of Q and it is not proportional to x because otherwise Z would have a nonzero real eigenvalue. Thus x and $(Q+Z)x$ are linearly independent and so no linear combination of I and $Q+Z$ can vanish. □

Remark 3. Observe that for single frequency, sinusoidal oscillations the nonlinear term in equation (1) yields a skew-symmetric matrix $x(t)=a\sin\mu t+b\cos\mu t$ then

$$\dot{x}x^{\top}-x\dot{x}^{\top}=\mu(ba^{\top}-ab^{\top}) \tag{2}$$

3 Synchronization on the line

Consider the problem of of arranging for x and y to move along the line approaching unit speed and approaching each other. This behavior is described, for example, by the second order equations

$$\begin{bmatrix}\ddot{x}\\ \ddot{y}\end{bmatrix}+\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix}+\begin{bmatrix}0&0\\1&-1\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}1\\0\end{bmatrix} \tag{3}$$

whose solutions are such that x approaches $t+\alpha$, with α dependent on the initial conditions and $x-y$ approaching zero, independent of the initial conditions. These equations model a situation of the leader-follower type; here x evolves independently of y and y follows x.

Theorem 4. *If affine functions of time are to be stable solutions of $\ddot{x}+\dot{x}+Ax=b$ in the sense that for any initial condition x approaches a solution $x=ct+\alpha d$ with c and d being independent of initial conditions and α a scalar dependent on the initial conditions, then it is necessary and sufficient that $Ac=0$, $c+Ad=b$ and that the polynomial $det(I(s^2+s)+A)$ has all but one of its roots in the left half-plane. If e is the vector with components all one, then for x to be asymptotically synchronized in the sense that x approaches a solution of the form $x(t)=et+\alpha e$ it is necessary that $Ae=0$ and $b=e$.*

If the eigenvalues of A are λ_i then the solutions of $s^2+s+\lambda_i$ are the relevant eigenvalues of the system. If these eigenvalues are to satisfy the stability condition we need $\pm\sqrt{-\lambda_i+1/4}\leq 1/2$, with equality holding for exactly one of the $2n$ possibilities.

Remark 5. If $Q=Q^{\top}\geq 0$, is of rank $Q=n-1$, and $Qe=0$ then all solutions of $\ddot{x}+\dot{x}+Qx=e$ synchronize.

If the system is symmetric in the sense that $\Pi A\Pi^{\top}=A$ for any permutation matrix Π then x and $z=\Pi x$ satisfy the same equation. This is the condition for each element of the group to play the same role in the evolution; i.e., for the group to be without a leader. The equation

$$\begin{bmatrix}\ddot{x}\\ \ddot{y}\end{bmatrix}+\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix}+\begin{bmatrix}-1&1\\1&-1\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix} \tag{4}$$

is an example.

Recasting this in a more geometric language, for the solutions of a system to synchronize it necessary for the vector field to leave invariant a one dimensional manifold and for this manifold to be stable in the above sense. Stated in this way we can include nonlinear versions of the idea. From the point of view of physical systems it is natural to consider equations of the form

$$\ddot{x} + f(x,\dot{x})\dot{x} + g(x,\dot{x}) = e. \tag{5}$$

and to look for conditions such that there is a stable solution of the form

$$x(t) = et + \alpha e + p(t) \tag{6}$$

with p being a zero mean periodic function of time.

4 Stabilizing a manifold

As we have posed it, synchronization is characterized in terms of the asymptotic stability of a one-dimensional submanifold. In the case of synchronization of periodic solutions this manifold will be diffeomorphic to a circle. Before attacking the synchronization problem directly we remark on the stabilization of more general submanifolds.

Definition 6. Let $X \subset \mathbb{R}^n$ be a asymptotically stable submanifold for $\dot{x} = f(x)$. By the *submanifold stabilization problem for $\dot{x} = f(x) + \sum g_i(x)u_i$, we understand the problem of finding a control law $u(x)$ such that a given submanifold $X_1 \subset X$ is attracting.*

Example 7. Consider the system in $\mathbb{R}^4$ defined by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} v & 1 & \sqrt{2} & -u \\ -1 & v & 0 & 0 \\ -\sqrt{2} & 0 & v & 0 \\ u & 0 & 0 & v \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \tag{7}$$

If we let $v = ||x|| - 1$ and $u = 0$ it is clear that this system leaves invariant the submanifold $X \sim S^3$ on which $||x||$ equals one and that this manifold is attracting in the sense that solutions starting near it approach it asymptotically. It is also true that when u is zero a submanifold of the form $X_1 \subset X$ consisting of points of the form $x_4 = 0$ is invariant but not asymptotically stable. However, the control law $u = -x_1$ makes $X_1 \sim S^2$ asymptotically stable.

Theorem 8. *Consider*

$$\dot{x} = f(x) + \sum g_i(x)u_i\,; \quad x \in \mathbb{R}^n \tag{8}$$

Let $X \subset \mathbb{R}^n$ be a compact, invariant submanifold for $\dot{x} = f(x)$ and assume that X is asymptotically stable. Let $X_1 \subset X$ also be invariant under the flow defined by $\dot{x} = f(x)$. Then if $\{g_i\}$ span the normal bundle of X_1 in a tubular neighborhood of X_1 then there exists a control law $u = u(x)$ that makes X_1 asymptotically stable.

Proof. Because X is assumed to be asymptotically stable we can limit our attention to initial conditions in X. For x in a neighborhood of X_1 let $d(x)$ denote the euclidean distance to X_1. Pick the u_i so that $\langle \nabla d, g_i u_i \rangle \leq 0$ Because collectively, the g_i span X_1, we see that along trajectories the distance is monotone decreasing and vanishes only when $d = 0$. □

5 Synchronization of frequency

Returning for a moment to Huygens, clocks can synchronize in the sense that their frequencies can become the same without settling down to a fixed phase relationship. The frequency of a periodic function of time is unambiguously defined as the inverse of the period. However, two periodic signals can have the same period while having very different wave shapes. It is only for sinusoids that the concept of phase is clearly defined.

Example 9. Consider the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & f_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & f_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} u \tag{9}$$

where $f_1 = 1 - x_1^2 - x_2^2$ and $f_2 = 1 - x_3^2 - x_4^2$. When u is zero it is clear that this system leaves invariant the submanifold $X \sim S^1 \times S^1$ on which f_1 and f_2 vanish and that this manifold is attracting in the sense that solutions starting near it approach it asymptotically as time goes to infinity. It is also true that when u is zero a submanifold of the form $X_1 \subset X$ consisting of points of the form

$$\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{10}$$

for a fixed value of θ is invariant. However, this one-dimensional submanifold $X_1 \sim S^1$ is not attracting. By some definitions the synchronization problem consists of finding a control law u which makes X_1 attracting.

More generally, for $x \in \mathbb{R}^n$, let $Q = Q^\top > 0$ be a fixed matrix. Suppose x is governed by the equations

$$\ddot{x} + \varepsilon f(x, \dot{x}) + x + \varepsilon^2 (Q + x\dot{x}^\top - \dot{x}x^\top) x = 0 \tag{11}$$

where f is a column vector which, we fix by setting its i^{th} entry as $(x_i^2 + \dot{x}_i^2 - 1)\dot{x}_i$. If the eigenvalues of Q are not rationally related there will be weakly stable solutions of $\ddot{x} + f(x, \dot{x}) + (I + \varepsilon^2)x = 0$ that are not periodic. What we want to show is that there exists a range of values for ε such that the solution of the full set of equations is nearly synchronous with an amplitude close to the amplitude of the decoupled equations. In short, the effect of the $x\dot{x}^\top - \dot{x}x^\top$ terms is to synchronize the components of the x variables, leaving the $x\dot{x}^\top - \dot{x}x^\top$ variables nearly constant. We will see the need for the scaling implied by the different powers of ε later.

The synchronization of oscillations is analogous to the tracking problem of the previous section but here the role of e, present on the right-hand side of equation (5), is assumed by the nonlinear terms responsible for the autonomous oscillation. However, the relative simplicity that came about because of the vector space structure is no longer available.

Expressed in first order form, we will be dealing with systems in the form of a linear system together with a nonlinear feedback,

$$\dot{x} = Ax + Bu\,; \quad \dot{x} = Ax + Bf(Cx). \tag{12}$$

If the eigenvalues of A are purely imaginary and if $f = 0$ then of course all solutions will oscillate, unless the Jordan normal form A includes one-chains. In any case there is no possibility for the solutions of a linear time invariant system to support fixed phase relationships that are stable in the sense that the effect of a small change in initial conditions will die out in time and the solution will return to the original phase relationships. Synchronization comes about because, under some circumstances, a nonlinear term can provide this kind of stability.

Example 10. It is not hard to show that for small values of ν and μ, small amplitude solutions of the equation

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1-\mu & f_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1-\nu & f_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ (x_2x_3 - x_1x_4)x_1 \\ 0 \\ (x_2x_3 - x_1x_4)x_3 \end{bmatrix}, \tag{13}$$

will synchronize in the sense that solutions x_1 and x_3 that are initially of different frequencies will approach the same frequency and evolve with a specific phase difference as time progresses. This example illustrates a mechanism by which two oscillators with different frequencies can adjust their frequencies so that they become equal.

Example 11. Consider the system

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & f_1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & f_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + (x_1x_4 - x_2x_3)\begin{bmatrix} 0 \\ -x_3 \\ 0 \\ x_1 \end{bmatrix}, \tag{14}$$

Here the system flows on a two-torus and is asymptotically stable to the circle defined by $x_1^2 = x_2^2 = 1, x_3^2 + x_4^2 = 1, x_1 = x_2$. To see that this is the case, observe that $\theta = \tan^{-1}(x_2/x_1)$ and $\phi = \tan^{-1}(x_4/x_3)$. satisfy $\dot{\theta} = -(\theta - \phi)$ and $\dot{\phi} = +(\theta - \phi)$ so the $\theta - \phi$ is exponentially decreasing to zero. This example illustrates a situation in which two oscillators with the same frequency can adjust to achieve a specified phase relationship in a stable way.

Remark 12. In checking these assertions keep in mind that If $x(t) = \cos(\omega t)$ and $y(t) = \cos(\omega t + \phi)$ then $x\dot{y} - \dot{x}y = -\cos t \sin(t + \phi) + \sin t \cos(t + \phi) = -2\sin\phi$.

6 Domains of attraction

In this section we establish some properties of the solutions of a class of second order oscillators with small, but arbitrary, inputs. The main results are summarized in Theorem 13 below.

For $|\varepsilon|$ sufficiently small, but nonzero, the set

$$S_\varepsilon = \{(x,\dot{x}) \mid (\dot{x}^2 + x^2 - 1) + 2\varepsilon^2 x\dot{x}\,\mathrm{sgn}(\dot{x}^2 + x^2 - 1) = \varepsilon\} \tag{15}$$

contains two smooth, closed contours, one lies outside the unit circle in $(x,\dot{x})$-space and one lies inside, both approach the unit circle as ε goes to zero. Let Γ_ε^+ denote the one outside and let Γ_ε^- denote the one inside. Consider the second order equation

$$\ddot{x} + \varepsilon\dot{x}(\dot{x}^2 + x^2 - 1) + x = \varepsilon^2 u \tag{16}$$

We want to show that for ε sufficiently small the solutions of this equation cross the contour Γ_ε^+ in the direction of the unit circle and also cross the contour Γ_ε^- in the direction of the unit circle provided that $|u| \leq \sqrt{x^2 + \dot{x}^2}$.

First we show this for Γ_ε^+. For the sake of brevity, introduce $d = \dot{x}^2 + x^2 - 1$ and note that

$$\dot{d} = -2\varepsilon\dot{x}^2 d + 2\varepsilon^2 u\dot{x} \tag{17}$$

and that

$$\frac{d}{dt}x\dot{x} = \dot{x}^2 - \varepsilon x\dot{x}d - x^2 + \varepsilon^2 xu \tag{18}$$

Thus

$$\frac{d}{dt}\left(d + \varepsilon^2 x\dot{x}\right) = -2\varepsilon\dot{x}^2 d + 2\varepsilon^2 u\dot{x} + \varepsilon^2\dot{x}^2 - \varepsilon^3 x\dot{x}d) - \varepsilon^2 x^2 + \varepsilon^4 xu \tag{19}$$

Note that if $d + 2\varepsilon^2 x\dot{x} = 2\varepsilon$ then to first order in ε we have $(\dot{x}^2 + x^2 - 1) = 2\varepsilon$. Thus, correct to second order in ε, we have on the contour Γ_ε^+

$$\frac{d}{dt}\left(d + \varepsilon^2 x\dot{x}\right) = \varepsilon^2\left(-\dot{x}^2 - x^2 + u\dot{x}\right)$$

and this is less than zero if $|u| < \sqrt{\dot{x}^2 + x^2}$

Now consider Γ_ε^-. The calculations above are only modified by some sign changes that lead to

$$\frac{d}{dt}(d - \varepsilon^2 x\dot{x}) = -2\varepsilon\dot{x}^2 d + 2\varepsilon^2 u\dot{x} - \dot{x}^2 - \varepsilon x\dot{x}d - x^2 + \varepsilon^2 xu \tag{20}$$

which means that on Γ_ε^-, to second order in ε,

$$\dot{d} - \varepsilon^2\frac{d}{dt}x\dot{x} = \varepsilon^2\left(\dot{x}^2 + x^2 + u\dot{x}\right) \tag{21}$$

which is positive if $|u| < \sqrt{\dot{x}^2 + x^2}$. We summarize these calculations with the following theorem.

Theorem 13. *Let* $\Gamma_\varepsilon^\pm$ *be as above. Then there exist* $\varepsilon_0 > 0$ *such that for all* $0 < \varepsilon < \varepsilon_0$ *the solutions of*

$$\ddot{x}+\varepsilon\dot{x}(\dot{x}^2+x^2-1)+x=\varepsilon^2 u \tag{22}$$

which begin in the annulus bounded by Γ_ε^+ *and* Γ_ε^- *remain in this annulus for all time, provided that* $|u| \leq \sqrt{x^2+\dot{x}^2}$.

Consider now the case of two coupled oscillators with the description

$$\begin{aligned}\ddot{x}+\varepsilon\dot{x}(\dot{x}^2+x^2-1)+x+(\varepsilon^2/3)(\alpha x-\varepsilon^2(x\dot{y}-\dot{x}y)y)&=0\\ \ddot{y}+\varepsilon\dot{x}(\dot{y}^2+y^2-1)+y+(\varepsilon^2/3(-\alpha y-\varepsilon^2(x\dot{y}-\dot{x}y)x)&=0\end{aligned} \tag{23}$$

The above results will establish that the solutions stay close to the unit circle in their respective spaces provided $|x-(x\dot{y}+\dot{x}y)y|<3\sqrt{\dot{x}^2+x^2}$ and $|y+(x\dot{y}+\dot{x}y)x|<3\sqrt{\dot{y}^2+y^2}$. In this case when $(x,\dot{x},y,\dot{y})$ is initially in their respective annular regions they will stay there.

We now focus on the term

$$g=\begin{bmatrix}\alpha & a\\ -a & -\alpha\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} \tag{24}$$

where $a=x\dot{y}-\dot{x}y$. Of course a vanishes if x and y are identical. If x and y are sinusoids of the same frequency a provides a measure of the phase difference between them. The following expression for $\dot{a}$ explains the limits on the growth of a.

$$\dot{a}=-(\varepsilon^2/3)(x^2+y^2)a-\varepsilon(x\dot{y}f_2-\varepsilon y\dot{x}f_1) \tag{25}$$

where, as above, $f_1=1-x^2-\dot{x}^2$ and $f_2=1-y^2-\dot{y}^2$. In the annuli, these terms are of order ε so the two expressions on the right are of comparable size.

If a is to approach a constant, and if the two oscillators are to oscillate at the same frequency, then we must arrange matters so that the two eigenvalues of the matrix appearing in equation (24) are the same. This is the case if $a=\pm\alpha$. Although it might see that a sinusoidal solution with x and y being out of phase by a certain amount would meet all the conditions necessary for a sinusoidal oscillation, difficulties arise because the repeated eigenvalues give rise to a secular term.

Finally, with respect to these equations, we note that

$$\frac{d}{dt}\tan^{-1}\frac{\dot{x}}{x}=1+e_1\ ;\quad \frac{d}{dt}\tan^{-1}\frac{\dot{y}}{y}=1+e_2 \tag{26}$$

where e_1 and e_2 are of order ε. Thus the solutions rotate in their respective annuli.

7 Averaging theory

In the previous section we have described conditions under which solutions of coupled equations are somehow close to sinusoids. However, we show here is that the usual approach to small parameter nonlinear oscillations based on averaging theory do not predict the existence of a periodic solution under these conditions, confirming the results of numerical studies. Reference [3] describes what we mean by the averaging equations.

For $x \in \mathbb{R}^n$ let $D(x,\dot{x})$ be a diagonal matrix whose i^{th} diagonal entry is $x_i^2 + \dot{x}_i^2 - 1$.

Theorem 14. *Let Q be a symmetric matrix with zero trace. For the equation describing n coupled oscillators*

$$\ddot{x} + \varepsilon D(x,\dot{x})\dot{x} + x + \varepsilon^2(Q + x\dot{x}^{\top} - \dot{x}x^{\top})x = 0 \tag{27}$$

the averaging equations relating to a possible synchronous solution, which necessarily takes the form $x(t) = a\sin\omega t + b\cos\omega t$, can be expressed in terms of the column vectors a and b as

$$\begin{bmatrix} Q + 2a^{\top}bI & (a^{\top}a + b^{\top}b)I \\ -(a^{\top}a + b^{\top}b)I & Q - 2\omega a^{\top}bI \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \omega^2 \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} \tag{28}$$

and these equations have no real solutions.

Proof. First of all, observe that if a periodic solution is to be sinusoidal and synchronous then the period must be 2π because if the eigenvalues of $I + \varepsilon^2 Q$ must sum to $\mathrm{tr}I + ab^{\top} - ba^{\top} = n.$ Because D vanishes when the individual oscillators are sinusoids of amplitude one and period 2π, we need to solve

$$Q - 2\omega(ab^{\top} - ba^{\top})(a\cos\omega ta - b\sin\omega tb) = 0 \tag{29}$$

which is equivalent to the diagonal form

$$\begin{bmatrix} Q - 2(a^{\top}b - ba^{\top}) & 0 \\ 0 & Q - 2(a^{\top}b - ba^{\top}) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} \tag{30}$$

These have no solution because a and b can not be aligned if they are to make the eigenvalues of $Q + ab^{\top} - ba^{\top}$ all the same but they must be aligned if they are to satisfy, for example, $(Q + ab^{\top} - ba^{\top})a = a.$ Rearranging equation (28) we see that it is equivalent to the one given in the theorem statement. □

8 Conclusions

We have made the case here that at least in some cases, what appears to be synchronization can be accompanied by small amplitude irregular motion, possibly chaotic, which occurs on such a small scale that it is effectively masked by the larger amplitude oscillations.

Acknowledgments

This work was supported in part by US Army Research Office under Pennsylvania State University Prime W911NF-07-1-0376.

Bibliography

[1] R. Brockett. On the rectification of vibratory motion. *Sensors and Actuators*, 20(1):91–96, 1989. Cited p. 66.

[2] R. Brockett. Pattern generation and the control of nonlinear systems. *IEEE Transactions on Automatic Control*, 48(10):1699–1712, 2003. Cited p. 66.

[3] J. K. Hale. *Oscillations in Nonlinear Systems*. McGraw-Hill, 1963. Cited p. 73.

[4] Y. Kuramoto. *Chemical Oscillations, Waves and Turbulence*. Springer. Cited p. 65.

[5] A. Mansouri. Local asymptotic feedback stabilization to a submanifold: Topological conditions. *Systems and Control Letters*, 56:525–528, 2007. Cited p. 65.

[6] A. Mansouri. Topological obstructions to submanifold stabilization. *IEEE Transactions on Automatic Control*, 55(7):1701–1703, 2010. Cited p. 65.

Subspace entropy and controlled invariant subspaces

Fritz Colonius
Institut für Mathematik
Universität Augsburg
Germany
fritz.colonius@math.uni-augsburg.de

1 Introduction

This paper discusses the following problem: Given a controlled invariant subspace V of a linear control system, what is the minimal amount of information per unit time (measured via an entropy notion) that has to be transferred to a controller in order to keep the system in or near V? This problem connects the analysis of control under communication constraints to classical geometric control theory. It was motivated by earlier investigations on invariance entropy (Colonius and Kawan [4]) for a similar problem, concerning controlled invariance of compact subsets with nonvoid interior in the state space where geometric structures did not play a role. The joint paper Colonius and Helmke [3] presented an important insight-the associated entropy for controlled invariant subspaces coincides with the subspace entropy of the linear flow associated with the uncontrolled system. The latter entropy notion was introduced in this paper and several estimates were derived. The present paper extends this line of research by giving a closer analysis of the subspace entropy.

Since the notion of controlled invariant subspaces is a cornerstone of geometric control theory, it is hoped that this will contribute to a closer connection of the theory of control under communication constraints to the more classical parts of state space control theory.

The contents of this paper is as follows: Section 2 collects results on topological entropy of linear differential equations and defines subspace entropy. Section 3 defines entropy for controlled invariant subspaces and explains the equivalence to subspace entropy. Final Section 4 presents the main results of this paper by analyzing the subspace entropy. It is shown that the subspace entropy is bounded above by the topological entropy of an induced system; a sufficient condition for equality is given which leads to a characterization of the subspace entropy (and hence the invariance entropy) by certain positive eigenvalues of the uncontrolled system.

This problem grew out of a discussion with Uwe, when we returned from a meeting of the DFG Priority Research Program 1305 “Control of Digitally Connected Dynamical Systems”. The successful application for funding of this research initiative by Deutsche Forschungsgemeinschaft (DFG) owes a lot to Uwe’s broad knowledge, his many fruitful ideas, and his vigor.

Notation. The distance of a point x in a normed vector space to a closed subset M is defined by $\operatorname{dist}(x,M) := \inf_{y\in M} \|x-y\|$.

2 Topological entropy and subspace entropy

In this section we first recall results on topological entropy of the flow for a linear differential equations. Then the subspace entropy is defined which is a suitable modification of the topological entropy. Later we will use it for the uncontrolled system $\dot{x} = Ax$ and relate it to the entropy of controlled invariant subspaces. It is worth to emphasize that an open loop control system does not define a flow, since the control functions $u(\cdot)$ are time-dependent, and hence it is not covered by this definition.

For a linear map $A : \mathcal{X} \to \mathcal{X}$ on an n-dimensional normed vector space $\mathcal{X}$, let $\Phi : \mathbb{R} \times \mathcal{X} \to \mathcal{X}, \Phi(t,x) := e^{tA}x$, $t \in \mathbb{R}$, $x \in \mathcal{X}$, be the induced flow (actually, throughout this paper, only the semiflow defined for $t \geq 0$ will be relevant.) A set R in $\mathcal{X}$ is called (T,ε)-spanning if for every $x \in K$ there is $y \in R$ such that for all $t \in [0,T]$ one has

$$\|\Phi(t,x) - \Phi(t,y)\| = \left\|e^{tA}(x-y\right\| < \varepsilon.$$

Denote by $r_{\text{top}}(T,\varepsilon,K)$ the minimal cardinality of such a (T,ε,K)-spanning set.

Definition 1. With the notation above, the topological entropy of Φ with respect to K is defined by

$$h_{\text{top}}(\varepsilon,K) := \limsup_{T\to\infty} \frac{1}{T} \log r_{\text{top}}(T,\varepsilon,K),$$
$$h_{\text{top}}(K) := \lim_{\varepsilon\searrow 0} h_{\text{top}}(\varepsilon,K).$$

and the topological entropy with respect to a subspace V of $\mathcal{X}$ is

$$h_{\text{top}}(V) = \sup_{K\subset V} h_{\text{top}}(K),$$

where the supremum is taken over all compact subsets $K \subset V$.

Where appropriate, we also write $h_{\text{top}}(V;\Phi)$, if the considered flow has to be specified. For the topological entropy of linear flows and $V = \mathcal{X}$, a classical result by R. Bowen [2] shows

$$h_{\text{top}}(\mathcal{X}) := \sup_{K} h_{\text{top}}(K) = \sum_{i=1}^{n} \max(0, \operatorname{Re}\lambda_i),$$

where $\lambda_1,\ldots,\lambda_n$ denote the eigenvalues of A; see also Walters [11, Theorem 8.14] and Matveev and Savkin [8, Theorem 2.4.2] for proofs. The supremum is attained for any compact set K with nonvoid interior in $\mathcal{X}$.

Since all norms on $\mathcal{X}$ are equivalent, we may assume that $\mathcal{X}$ is a Hilbert space and we endow $\mathcal{X}$ with the following inner product which is adapted to the decomposition into the Lyapunov spaces $L_j, 1 \leq j \leq l$. Recall that a Lyapunov space L_j is the sum of all generalized (real) eigenspaces corresponding to an eigenvalue of A with real part equal to λ_j. We order these Lyapunov exponents such that

$$\lambda_1 > \ldots > \lambda_l.$$

Take a basis corresponding to the Jordan normal form: for each j, one has a basis $e_1^j, \ldots, e_{n_j}^j$of L_j which is orthonormal with respect to an inner product in L_j. Define

$$\left\langle e_{i_1}^{j_1}, e_{i_2}^{j_2} \right\rangle = \begin{cases} 0 & \text{for} \quad j_1 \neq j_2 \text{ or } i_1 \neq i_2 \\ 1 & \text{for} \quad j_1 = j_2 \text{ and } i_1 = i_2. \end{cases} \tag{1}$$

In order to simplify the notation a bit, we number the basis elements by $1, \ldots, n$ and denote them by x_j. They form an orthonormal basis for an inner product on $\mathcal{X}$. Recall that we can identify the Grassmannian manifold $\mathbb{G}_k\mathcal{X}$ of k-dimensional subspaces with the subset of projective space $\mathbb{P}(\bigwedge^k \mathcal{X})$ obtained from the indecomposable elements in the exterior product $\bigwedge^k \mathcal{X}$. We endow $\mathbb{G}_k\mathcal{X}$ with the corresponding metric.

Following Colonius, San Martin, da Silva [6] we first describe the chain recurrent components in the Grassmannian; see, e.g., Robinson [9] for a discussion of this notion for flows on compact metric spaces.

Theorem 2. *Let $A : \mathcal{X} \to \mathcal{X}$ be a linear map with flow $\Phi_t = e^{tA}$ on $\mathcal{X}$. Let L_i, $i = 1, \ldots, l$, be the Lyapunov spaces of A. For $k \in \{1, \ldots, n\}$ define the index set*

$$I(k) = \{(k_1, \ldots, k_l) \,|\, k_1 + \ldots + k_l = k \text{ and } 0 \le k_i \le n_i = \dim L_i\}. \tag{2}$$

Then the chain recurrent components (also called Morse sets) of the induced flow on the Grassmannian $\mathbb{G}_k\mathcal{X}$ are

$$\mathcal{M}_{k_1,\ldots,k_l}^k = \mathbb{G}_{k_1}L_1 \oplus \ldots \oplus \mathbb{G}_{k_l}L_l, \ (k_1, \ldots, k_l) \in I(k). \tag{3}$$

Here the sum on the right-hand side denotes the set of all k-dimensional subspaces $V^k \subset \mathcal{X}$ with

$$\dim(V^k \cap L_i) = k_i, \ i = 1, \ldots, l.$$

In particular, every ω-limit set for the induced flow on $\mathbb{G}_k\mathcal{X}$ is contained in one of these chain recurrent components.

If A is hyperbolic, i.e., there are no eigenvalues of A on the imaginary axis, then one can decompose $\mathcal{X}$ into the stable and the unstable subspaces, $\mathcal{X} = \mathcal{X}^- \oplus \mathcal{X}^+$. We denote by $\pi^\pm$ the projection of $\mathcal{X}$ to $\mathcal{X}^\pm$ and let $\Phi^\pm$ be the associated restrictions of Φ.

Theorem 3. *Consider a linear flow $\Phi_t = e^{tA}$ and assume that A is hyperbolic. Let V be a k-dimensional subspace. Then the topological entropy with respect to a compact set $K \subset V$ satisfies*

$$h_{\text{top}}(K, \Phi) = h_{\text{top}}(\pi^+ K, \Phi^+).$$

and the topological entropy of V is

$$h_{\text{top}}(V, \Phi) = \sum_{i=1}^{l} k_i \max(0, \lambda_i),$$

where $\mathcal{M}_{k_1,\ldots,k_l}^k \subset \mathbb{G}_k\mathcal{X}$ is the Morse set that contains the omega limit of the point V for the induced flow $\mathbb{G}_k\Phi$ on the k-Grassmannian $\mathbb{G}_k\mathcal{X}$. Furthermore, for every compact subset $K \subset V$ with nonvoid interior, $h_{\text{top}}(K;\Phi)$ equals the volume growth rate of π^+K under the flow Φ^+.

We note that the Morse set containing the omega limit of V can be determined in the following way. Let $v_1,...,v_k$ be a basis of V. Then we can express the v_i using the standard basis of $\mathcal{X}$ as introduced in (1) by

$$v_i = \alpha_{i1}x_1 + \ \dots \ + \alpha_{in}x_n = \sum_{\alpha_{ij}\neq 0} \alpha_{ij}x_j.$$

There is a minimal number of Lyapunov spaces such that

$$V \subset L_{i_1} \oplus ... \oplus L_{i_j},$$

and we number them such that $\lambda_{i_1} > \ \dots \ > \lambda_{i_j}$. Note that generically $\sum \dim L_{i_j} > k$. Then $\omega(V)$ is contained in the Morse set $\mathcal{M}^k_{k_1,...,k_l}$, where the k_i are recursively obtained in the following way: k_1 is the maximal number of base vectors v_i which have a nontrivial component in L_1. Then eliminate these base vectors v_i and let k_2 be the maximal number of the remaining v_i which have a nontrivial component in L_2, etc.

Next we modify the definition of topological entropy in order to define the subspace entropy introduced in Colonius and Helmke [3]. Let V be a linear subspace of $\mathcal{X}$ and consider a linear map $A : \mathcal{X} \to \mathcal{X}$ with flow $\Phi_t = e^{tA}$. For any compact subset $K \subset V$ and for given $T,\varepsilon > 0$ we call $R \subset K$ a (T,ε)-spanning set, if for all $x \in K$ there exists $y \in R$ with

$$\max_{0\leq t\leq T} \operatorname{dist}(e^{tA}(x-y),V) < \varepsilon. \qquad (4)$$

Let $r_{\text{sub}}(T,\varepsilon,K,V)$ denote the minimal cardinality of a such a (T,ε)-spanning set. If no finite (T,ε)-spanning set exists, we set $r_{\text{sub}}(T,\varepsilon,K,V;\Phi) = \infty$. If there exists some (T,ε)-spanning set, then one also finds a finite (T,ε)-spanning set using compactness of K and continuous dependence on the initial value. Note that the points y in R will, in general, not lead to solutions $e^{tA}y$ which remain for all $t \geq 0$ in the ε-neighborhood of V.

Definition 4. Let A be a linear map on $\mathcal{X}$ with associated flow $\Phi_t = e^{tA}$ and consider a subspace V of $\mathcal{X}$. For a compact subset $K \subset V$, we consider the exponential growth rate of $r_{\text{sub}}(T,\varepsilon,K,V)$ and set

$$h_{\text{sub}}(\varepsilon,K,V) := \limsup_{T\to\infty} \frac{1}{T} \log r_{\text{sub}}(T,\varepsilon,K,V),$$
$$h_{\text{sub}}(K,V) := \lim_{\varepsilon\searrow 0} h_{\text{sub}}(\varepsilon,K,V),$$

and define the entropy of V with respect to Φ by

$$h_{\text{sub}}(V) := \sup_K h_{\text{sub}}(K,V),$$

where the supremum is taken over all compact subsets $K \subset V$.

Where appropriate, we write $h_{\text{sub}}(V;\Phi)$ in order to clarify which flow is considered. As usual in the context of topological entropy, one sees that, by monotonicity, the limit for $\varepsilon \searrow 0$ exists (it might be infinite.) Since all norms on a finite dimensional vector space are equivalent, the entropy does not depend on the norm used in (4). For simplicity, we require throughout that $\mathcal{X}$ is a Hilbert space. One easily sees that the subspace entropy $h(V;\Phi)$ is invariant under state space similarity, i.e., $h(SV;S\Phi S^{-1}) = h(V;\Phi)$ for S in the set $GL(\mathcal{X})$ of isomorphisms on $\mathcal{X}$; here $S\Phi_t S^{-1} = Se^{tA}S^{-1} = e^{SAS^{-1}t}, t \geq 0$.

Remark 5. If we choose $V = \{0\}$ condition (4) is trivial, since only $K = \{0\}$ is allowed; furthermore, if we choose $V = \mathcal{X}$, the distance in (4) is always equal to zero. In particular, the subspace entropy does not recover the usual definition of topological entropy for the linear flow $\Phi(t,x) = e^{tA}x$; see Definition 1.

3 Entropy for controlled invariant subspaces

This section briefly summarize some well-known definitions and facts concerning controlled invariant subspaces. Then invariance entropy for controlled invariant subspaces of linear control systems on $\mathcal{X}$ is defined and related to the subspace entropy of linear flows as defined in the previous section.

The notion of controlled invariant subspaces (also called (A,B)–invariant subspaces) was introduced by Basile and Marro [1]; see the monographs Wonham [12] and Trentelman, Stoorvogel and Hautus [10] for expositions of the theory.

Consider linear control systems in state space form

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{5}$$

with linear maps $A : \mathcal{X} \to \mathcal{X}$ and $B : \mathbb{R}^m \to \mathcal{X}$, where $\mathcal{X}$ is an n-dimensional normed vector space. The solutions $\varphi(t,x,u), t \geq 0$, of (5) with initial condition $\varphi(0,x,u) = x$ are given by the variation-of-constants formula

$$\varphi(t,x,u) = e^{tA}x + \int_0^t e^{A(t-s)}Bu(s)ds.$$

Recall that a subspace V is called controlled invariant, if for all $x \in V$ there is $u \in \mathbb{R}^m$ with $Ax + Bu \in V$, i.e., if $AV \subset V + \operatorname{Im} B$. Equivalently, there is a linear map $F : \mathcal{X} \to \mathbb{R}^m$, called a friend of V, such that

$$(A + BF)V \subset V.$$

This also shows that V is controlled invariant iff for every $x \in V$ there is an (open loop) continuous control function $u : [0,\infty) \to \mathbb{R}^m$ with $\varphi(t,x,u) \in V$ for all $t \geq 0$. In fact, differentiating the solution one finds

$$V \ni \frac{d}{dt}\varphi(0,x,u) = Ax + Bu(0).$$

For the converse, define for $x \in V$ a control by $u(t) = Fe^{(A+BF)t}x, t \geq 0$.

We now introduce the central notion of this paper, invariance entropy for controlled invariant subspaces of linear control system (5) and relate it to the subspace entropy defined in the previous section.

In the following, we consider a fixed controlled invariant subspace V of $\mathcal{X}$ with $\dim V = k$. Furthermore, we admit arbitrary controls in the space $C([0,\infty),\mathbb{R}^m)$ of continuous functions $u : [0,\infty) \to \mathbb{R}^m$.

Definition 6. For a compact subset $K \subset V$ and for given $T,\varepsilon > 0$ we call a set $\mathcal{R} \subset C([0,\infty),\mathbb{R}^m)$ of control functions (T,ε)-*spanning* if for all $x_0 \in K$ there is $u \in \mathcal{R}$ with

$$\operatorname{dist}(\varphi(t,x_0,u),V) < \varepsilon \text{ for all } t \in [0,T]. \tag{6}$$

By $r_{\text{inv}}(T,\varepsilon,K,V)$ we denote the minimal cardinality of such a (T,ε)-spanning set. If no finite (T,ε)-spanning set exists, we set $r_{\text{inv}}(T,\varepsilon,K,V) = \infty$.

In other words, we require for a (T,ε)-spanning set $\mathcal{R}$ that, for every initial value in K, there is a control in $\mathcal{R}$ such that up to time T the trajectory remains in the ε-neighborhood of V. Note that, in contrast to the definitions of topological entropy and subspace entropy for flows, Definition 4, here a number of control functions is counted, not a number of initial values. Hence this notion is intrinsic for control systems.

The following elementary observation shows that one cannot require that there are finitely many control functions u such that instead of (6) one has $\varphi(t,x_0,u) \in V$ for all $t \in [0,T]$. Hence the invariance condition has to be relaxed as indicated above using $\varepsilon > 0$.

Proposition 7. *Let V be a controlled invariant subspace. Furthermore, consider a neighborhood K of the origin in V, let $T > 0$, and suppose that there is $v \in V$ with $e^{AT}v \notin V$. Then there is no finite set $\mathcal{R}$ of controls such that for every $x_0 \in K$ there is $u \in \mathcal{R}$ with $\varphi(t,x_0,u) \in V$ for all $t \in [0,T]$.*

Proof. We may assume that $\gamma v \in K$ for all $\gamma \in (0,1)$. The proof is by contradiction. Suppose that $\mathcal{R} = \{u_1,\dots,u_r\}$ is a finite set of controls such that for every $x_0 \in V$ there is a control u_j in $\mathcal{R}$ with $\varphi(T,x_0,u_j) \in V$. There is a control in $\mathcal{R}$, say u_1, with

$$\varphi(T,v,u_1) = e^{TA}v + \int_0^T e^{(T-s)A}Bu_1(s)ds \in V.$$

Since $e^{TA}v \notin V$, it follows that

$$\varphi(T,0,u_1) = \int_0^T e^{(T-s)A}Bu_1(s)ds \notin V.$$

We find for $\gamma \in (0,1)$

$$\begin{aligned}\varphi(T,\gamma v,u_1) &= \gamma\left[e^{TA}v + \int_0^T e^{(T-s)A}Bu_1(s)ds\right] + (1-\gamma)\int_0^T e^{(T-s)A}Bu_1(s)ds \\ &= \gamma\varphi(T,v,u_1) + (1-\gamma)\varphi(T,0,u_1).\end{aligned}$$

This implies $\varphi(T,\gamma v,u_1) \notin V$ for all $\gamma \in (0,1)$. Choose $\gamma_1 \in (0,1)$ and let $v_1 := \gamma_1 v$. There is a control in $\mathcal{R}$, say $u_2 \neq u_1$, such that $\varphi(T,v_1,u_2) \in V$. Iterating the arguments above one arrives at a contradiction. □

On the other hand, there are always finite (T,ε)-spanning sets of controls as shown by the following remark.

Remark 8. Let $K \subset V$ be compact and $\varepsilon, T > 0$. By controlled invariance of V there is for every $x \in K \subset V$ a control function u with $\varphi(t,x,u) \in V$ for all $t \geq 0$. Hence, using continuous dependence on initial values and compactness of K, one finds finitely many controls $u_1,...,u_r$ such that for every $x \in K$ there is u_j with $\mathrm{dist}(\varphi(t,x,u_j),V) < \varepsilon$ for all $t \in [0,T]$. Hence $r_{\mathrm{inv}}(T,\varepsilon,K,V) < \infty$.

Now we consider the exponential growth rate of $r_{\mathrm{inv}}(T,\varepsilon,K,V)$ as in Definition 6 for $T \to \infty$ and let $\varepsilon \to 0$. The resulting invariance entropy is the main subject of the present paper.

Definition 9. Let V be a controlled invariant subspace for a control system of the form (5). Then, for a compact subset $K \subset V$, the *invariance entropy* $h_{\mathrm{inv}}(K,V)$ is defined by

$$h_{\mathrm{inv}}(\varepsilon,K,V) := \limsup_{T\to\infty} \frac{1}{T} \log r_{\mathrm{inv}}(T,\varepsilon,K,V),$$
$$h_{\mathrm{inv}}(K,V) := \lim_{\varepsilon \searrow 0} h_{\mathrm{inv}}(\varepsilon,K,V).$$

Finally, the invariance entropy of V is defined by

$$h_{\mathrm{inv}}(V;A,B) := \sup_K h_{\mathrm{inv}}(K,V),$$

where the supremum is taken over all compact subsets $K \subset V$.

In the sequel, we will use the shorthand notation $h_{\mathrm{inv}}(V)$ for $h_{\mathrm{inv}}(V;A,B)$, when it is clear which control system is considered. Note that $h_{\mathrm{inv}}(\varepsilon_1,K,V) \leq h_{\mathrm{inv}}(\varepsilon_2,K,V)$ for $\varepsilon_2 \leq \varepsilon_1$. Hence the limit for $\varepsilon \to 0$ exists (it might be infinite.) Since all norms on finite dimensional vector spaces are equivalent, the invariance entropy of V is independent of the chosen norm. We will show later that every controlled invariant subspace has finite invariance entropy. It is clear by inspection, that, as the subspace entropy $h_{\mathrm{sub}}(V)$, also the invariance entropy $h_{\mathrm{inv}}(V)$ is invariant under state space similarity; i.e. $h_{\mathrm{inv}}(SV;SAS^{-1},SB) = h_{\mathrm{inv}}(V;A,B)$ for $S \in GL(\mathcal{X})$.

We are interested in the problem to keep the system in the subspace V for all $t \geq 0$. Then the exponential growth rate of the required number of control functions will give information on the difficulty of this task. A motivation to consider open-loop controls in this context comes, in particular, from model predictive control (see, e.g., Grüne and Pannek [7]), where optimal open-loop controls are computed and applied on short time intervals.

The following theorem (taken from Colonius and Helmke [3]) shows that the entropy of a controlled invariant subspace V can be characterized by the entropy of V for the corresponding uncontrolled system $\dot{x} = Ax$. This result will be useful in order to compute entropy bounds.

Theorem 10. *Let V be a controlled invariant subspace for system (5) and consider the invariance entropy $h_{\mathrm{inv}}(V)$ of control system (5) and the subspace entropy $h_{\mathrm{sub}}(V)$ of V of the uncontrolled system $\Phi_t = e^{tA}$. Then*

$$h_{\mathrm{inv}}(V) = h_{\mathrm{sub}}(V;\Phi).$$

Proof. (i) Let $K \subset V$ be compact, and fix $T, \varepsilon > 0$. Consider a (T, ε, K, V)-spanning set $\mathcal{R} = \{u_1, ..., u_r\}$ of controls with minimal cardinality $r = r_{\text{inv}}(T, \varepsilon, K, V)$. This means that for every $x \in K$ there is u_j with

$$\text{dist}(\varphi(t,x,u_j), V) < \varepsilon \text{ for all } t \in [0,T].$$

By minimality, we can for every u_j pick $x_j \in K$ with $\text{dist}(\varphi(t,x_j,u_j),V) < \varepsilon$ for all $t \in [0,T]$. Then, using linearity, one finds for all $x \in K$ a control u_j and a point $x_j \in K$ such that for all $t \in [0,T]$

$$\text{dist}(e^{tA}x - e^{tA}x_j, V) = \text{dist}(\varphi(t,x,u_j) - \varphi(t,x_j,u_j), V) < 2\varepsilon.$$

This shows that the points x_j form a $(T, 2\varepsilon)$-spanning set for the subspace entropy, and hence

$$r_{\text{inv}}(T, \varepsilon, K, V) \geq r_{\text{sub}}(T, 2\varepsilon, K, V).$$

Letting T tend to infinity, then $\varepsilon \to 0$ and, finally, taking the supremum over all compact subsets $K \subset V$, one obtains $h_{\text{inv}}(V) \geq h_{\text{sub}}(V)$.

(ii) For the converse inequality, let K be a compact subset of V and $T, \varepsilon > 0$. Let $E = \{x_1, \ldots, x_r\} \subset K$ be a minimal (T, ε)-spanning set for the subspace entropy which means that for all $x \in K$ there is $j \in \{1, \ldots, r\}, r = r_{\text{sub}}(T, \varepsilon, K, V)$, such that for all $t \in [0,T]$

$$\text{dist}(e^{tA}x - e^{tA}x_j, V) = \inf_{z \in V} \left\| e^{tA}x - e^{tA}x_j - z \right\| < \varepsilon.$$

Since V is controlled invariant, we can assign to each $x_j, j \in \{1, \ldots, r\}$, a control function $u_j \in C([0,\infty), \mathbb{R}^m)$ such that $\varphi(t, x_j, u_j) \in V$ for all $t \geq 0$. Let $\mathcal{R} := \{u_1, \ldots, u_r\}$. Using linearity we obtain that for every $x \in K$ there is j such that for all $t \in [0,T]$

$$\text{dist}(\varphi(t,x,u_j) - \varphi(t,x_j,u_j), V) = \text{dist}(e^{tA}x - e^{tA}x_j, V) < \varepsilon.$$

Since $\varphi(t, x_j, u_j) \in V$ for $t \in [0,T]$, it follows that

$$\begin{aligned}\text{dist}(\varphi(t,x,u_j), V) &= \inf_{z \in V} \left\| \varphi(t,x,u_j) - z \right\| \\ &\leq \left\| \varphi(t,x,u_j) - \varphi(t,x_j,u_j) \right\| < \varepsilon.\end{aligned}$$

Thus for every $x \in K$ there is $u_j \in \mathcal{R}$ such that for all $t \in [0,T]$ one has $\text{dist}(\varphi(t,x,u_j), V) < \varepsilon$. Hence $\mathcal{R}$ is (T, ε)-spanning for the invariance entropy and it follows that

$$r_{\text{inv}}(T, \varepsilon, K, V) \leq r_{\text{sub}}(T, \varepsilon, K, V) \text{ for all } T, \varepsilon > 0,$$

and consequently $h_{\text{inv}}(K, V) \leq h_{\text{sub}}(V; \Phi)$. □

In view of this theorem, we will look more closely at the subspace entropy.

4 Analysis of the subspace entropy

This section presents an analysis of the subspace entropy. The main result is Theorem 20 which shows that the subspace entropy is bounded above by the topological entropy of an induced system; a sufficient condition for equality is given which leads to a characterization of the subspace entropy (and hence the invariance entropy) by certain positive eigenvalues of the uncontrolled system.

First we describe the behavior of the subspace entropy under a semiconjugacy to the induced flow on a quotient space.

Proposition 11. *Let W be an A-invariant subspace for a linear map A on $\mathcal{X}$. Then, for a subspace V of $\mathcal{X}$ the subspace entropies of the flow $\Phi_t = e^{tA}$ on $\mathcal{X}$ and the induced flow $\bar{\Phi}_t$ on the quotient space $\mathcal{X}/W$, respectively, satisfy*

$$h_{\text{sub}}(V,\Phi) \geq h_{\text{sub}}(V/W,\bar{\Phi}).$$

Proof. Let $K \subset V$ be compact and for $T,\varepsilon > 0$ consider a $(T,\varepsilon,K,V;\Phi)$-spanning set $R \subset K$. Denote the projection of $\mathcal{X}$ to $\mathcal{X}/W$ by π, hence $\pi V = V/W$. Then the set πR is (T,ε)-spanning for $\pi K \subset \pi V$ with respect to the flow $\bar{\Phi}$. In fact, let $R = \{x_1,\dots,x_r\}$ and consider $\pi x \in \pi K$ for some element $x \in K$. Then there exists $x_j \in R$ with

$$\max_{0\leq t\leq T} \operatorname{dist}(e^{tA}(x-x_j),V) < \varepsilon.$$

Denoting the map induced by A on $\mathcal{X}/W$ by $\bar{A}$ one finds for all $t \in [0,T]$

$$\begin{aligned}
\operatorname{dist}(e^{t\bar{A}}(\pi x-\pi x_j),\pi V) &= \inf_{z\in V}\left\| e^{t\bar{A}}(\pi x-\pi x_j)-\pi z\right\| \\
&= \inf_{z\in V,w\in W}\left\| e^{tA}(x-x_j)-z-w\right\| \\
&\leq \operatorname{dist}(e^{tA}(x-x_j),V) < \varepsilon.
\end{aligned}$$

It follows that the minimal cardinality $r_{\text{sub}}(T,\varepsilon,K,V)$ for Φ is greater than or equal to the minimal cardinality $r_{\text{sub}}(T,\varepsilon,\pi K,\pi V)$ for $\bar{\Phi}$. Then take the limit superior for $T \to \infty$ and let ε tend to 0. Finally, observe that for every compact set $K_1 \subset V/W$ there is a compact set $K \subset V$ with $\pi K = K_1$. Hence taking the supremum over all compact $K_1 \subset V/W$ one obtains the assertion. □

Note that the map A does not induce a map on the quotient space $\mathcal{X}/V$, since we are interested in the case where V is not invariant. Nevertheless, condition (4) determines a distance in $\mathcal{X}/V$.

Next we show that we may assume that all eigenvalues of A have positive real part. Decompose $\mathcal{X}$ into the center-stable and the unstable subspaces, $\mathcal{X} = \mathcal{X}^{-,0} \oplus \mathcal{X}^{+}$. Thus $\mathcal{X}^{-,0}$ is the sum of all real generalized eigenspaces corresponding to eigenvalues with nonpositive real part and $\mathcal{X}^{+}$ is the sum of all real generalized eigenspaces corresponding to eigenvalues with positive real part. We denote the corresponding projections of $\mathcal{X}$ by $\pi^{-,0}$ and π^{+}, respectively, and let $\Phi^{0,-}$ and Φ^{+} be the associated restrictions of Φ.

Proposition 12. *Let V be a subspace of $\mathcal{X}$. Then the subspace entropy Φ with respect to V and of Φ^{+} with respect to $\pi^{+}V$ coincide.*

Proof. Decompose Φ into $\Phi^{0,-}$ and Φ^{+}. The restriction $\Phi^{0,-}$ to the center-stable subspace has the property, that for a polynomial $p(t)$

$$\left\|\Phi_t^{0,-}(x-y)\right\| \leq p(t)\left\|x-y\right\|,$$

hence the subspace entropy here vanishes. Furthermore, the product of spanning sets for the stable and the unstable part yields spanning sets for the total system, hence

$$h_{\text{sub}}(V,\Phi) \leq h_{\text{sub}}(V,\Phi^{+}) + h_{\text{sub}}(V,\Phi^{0,-}) = h_{\text{sub}}(V,\Phi^{+}).$$

and clearly, $h_{\text{sub}}(V,\Phi^{+}) \leq h_{\text{sub}}(V,\Phi)$. □

Next we show that the subspace entropy is bounded above by the topological entropy of V.

Proposition 13. *Let V be a subspace of $\mathcal{X}$. Then the topological entropy of V and the subspace entropy of V satisfy $h_{\text{sub}}(V) \leq h_{\text{top}}(V)$.*

Proof. Let $K \subset V$ be compact and for $T,\varepsilon > 0$ consider a (T,ε)-spanning set $R = \{x_1,\ldots,x_r\} \subset K$ with minimal cardinality $r = r_{\text{top}}(T,\varepsilon,K)$. For every $x \in K$ there exists $x_j \in R$ such that for all $t \in [0,T]$

$$\left\|e^{tA}(x-x_j)\right\| < \varepsilon.$$

Then one finds for all $t \in [0,T]$

$$\text{dist}(e^{tA}(x-x_j),V) = \inf_{v\in V}\left\|e^{tA}(x-x_j)-v\right\| \leq \left\|e^{tA}(x-x_j)\right\| < \varepsilon.$$

It follows that the minimal cardinality $r_{\text{top}}(T,\varepsilon,K)$ for the topological entropy is greater than or equal to the minimal cardinality $r_{\text{sub}}(T,\varepsilon,K,V)$ for the subspace entropy. Then take the limit superior for $T \to \infty$ and let ε tend to 0. Finally, take the supremum over all compact sets $K \subset V$. □

The next proposition shows that only part of the state space $\mathcal{X}$ is relevant for the subspace entropy.

Proposition 14. *Let $V \subset \mathcal{X}$ be a subspace. Then the subspace entropies of V as a subspace of $\mathcal{X}$ and of the smallest A-invariant subspace $\langle A|V\rangle$ containing V coincide.*

Proof. Let $K \subset V$ be compact and consider $T,\varepsilon > 0$. Let $R \subset K$ in $\mathcal{X}$ be a (T,ε)-spanning set for the system in $\mathcal{X}$. Thus for every $x \in K$ there exists $y \in R$ with

$$\max_{0\leq t\leq T}\text{dist}(e^{tA}(x-y),V) < \varepsilon.$$

Since $e^{tA}(x-y) \in \langle A|V\rangle$, it follows that R is also a (T,ε)-spanning set for the system in $\langle A|V\rangle$. The converse is obvious. □

Hence it suffices to consider the system in $\langle A|V\rangle$. Next we show that we can also neglect the largest A-invariant subspace of V, denoted by $\ker(A;V)$. We denote the projection by

$$\pi : \langle A|V\rangle \to \langle A|V\rangle / \ker(A;V)$$

and hence $V/\ker(A;V) = \pi V$. The linear map A induces a linear map $\bar{A}$ on the quotient space $\langle A|V\rangle/\ker(A;V)$ and we let $\pi\Phi(t,\bar{x}) := e^{\bar{A}t}\bar{x}, t \in \mathbb{R}, \bar{x} \in \pi V$. Note that for all subspaces $W \subset V$ with $AW \subset W$ it follows that $W \subset \ker(A;V)$. Hence for a subspace $\pi W \subset \pi V$ with $\bar{A}(\pi W) = AW + \ker(A;V) \subset \pi W = W + \ker(A;V)$ the subspace $W + \ker(A;V) \subset V + \ker(A;V) \subset V$ is an A-invariant subspace of V and hence contained in $\ker(A;V)$.

Proposition 15. *The subspace entropies of* Φ *with respect to* V *and of* $\pi\Phi$ *with respect to* $V/\ker(A;V)$ *coincide,*

$$h_{\text{sub}}(V;\Phi) = h_{\text{sub}}(\pi V;\pi\Phi).$$

Proof. By Proposition 11, the inequality $h_{\text{sub}}(V;\Phi) \geq h_{\text{sub}}(\pi V;\pi\Phi)$ follows. For the converse, let K be a compact subset of V. Then, for the projection π of $\langle A|V\rangle$ to the quotient space $\langle A|V\rangle/\ker(A;V)$, the set πK is compact and $\pi V = V + \ker(A;V)$. Let $T, \varepsilon > 0$ be given and denote by $E \subset \pi K$ a minimal $(T,\varepsilon,\pi K,\pi V;\pi\Phi)$-spanning set with respect to the flow $\pi\Phi$ on $\langle A|V\rangle/\ker(A;V)$, say $E = \{\pi x_1,\ldots,\pi x_r\}$ with $x_j \in K$ and $r = r_{\text{sub}}(T,\varepsilon,\pi K,\pi V)$. Note that $V + \ker(A;V) = V$. Hence it follows that for all $x \in K$ there is $j \in \{1,\ldots,r\}$ such that for all $t \in [0,T]$

$$\begin{aligned}\inf_{z \in V}\left\| e^{tA}x - e^{tA}x_j - z\right\| &= \operatorname{dist}(e^{tA}x - e^{tA}x_j, V + \ker(A;V)) \\ &= \operatorname{dist}(e^{\bar{A}t}\pi x - e^{\bar{A}t}\pi x_j, \pi V) < \varepsilon.\end{aligned}$$

We have shown that the set $\{x_1,\ldots,x_r\} \subset K$ is (T,ε)-spanning for Φ and hence the minimal cardinality $r_{\text{sub}}(T,\varepsilon,K,V)$ of such a set is equal to or less than $r_{\text{sub}}(T,\varepsilon,$ $\pi K,\pi V)$. Thus the assertion follows. □

This result shows that we have to project things to πV for every time t. Observe that $\dim e^{\bar{A}t}(\pi V) = \dim(\pi V)$. However, the projection of $e^{\bar{A}t}(\pi V)$ to $\langle A|V\rangle/\pi V$ need not have constant dimension. Slightly more generally, we have the following situation: Consider a linear map A on $\mathcal{X}$ and a subspace V of $\mathcal{X}$ which is not invariant under A. Due to Proposition 13 we know that the topological entropy is an upper bound. The following examples show that the subspace entropy may be equal to the topological entropy or less than the topological entropy.

Example 16. Consider a complex conjugate pair of eigenvalues and a one-dimensional subspace V of the real eigenspace. Let K be a compact neighborhood of the origin in V. This can be a controlled invariant subspace: Consider $V = \mathbb{R} \times \{0\}$ and with $\lambda > 0$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \left(\lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\right)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t),$$

i.e.

$$\begin{aligned}\dot{x}_1 &= \lambda x_1 - x_2\\ \dot{x}_2 &= x_1 + \lambda x_2 + u(t)\end{aligned}$$

If we choose $u = -x_1 - \lambda x_2$, then every initial point with $x_2 = 0$ remains in this subspace.

For $u = 0$, the solution is

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = e^{\lambda t} \begin{bmatrix} x_1^0 \cos t - x_2^0 \sin t \\ x_1^0 \sin t + x_2^0 \cos t \end{bmatrix}.$$

Initial values $(x_1^0, 0) \in V$ have as second component

$$x_2(t) = e^{\lambda t}[x_1^0 \sin t + x_2^0 \cos t] = e^{\lambda t} x_1^0 \sin t.$$

Hence the projection of the solutions to $\mathbb{R}^2/V$, identified with the second component, gives for $K \subset V$

$$x_2(t) = e^{\lambda t} \sin t \cdot x_1^0,\ x_1^0 \in K.$$

The solutions $x_2(t)$ move apart with $e^{\lambda t}$, if we consider the limit superior: choose $t = (2n+1)\frac{\pi}{2}$. Hence the subspace entropy is $h_{\mathrm{sub}}(V) = \lambda$. Observe that the image of the projection depends, naturally, on t. In $\mathbb{R}^2/V$ it is one-dimensional, except for $t = n\pi, n \geq 0$, where it drops to 0. In this example, the Lyapunov exponent in L_j determines the subspace entropy.

Example 17. Consider with $\lambda > 0$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t).$$

The eigenspace is $\mathbb{R} \times \{0\}$. The subspace $V = \{0\} \times \mathbb{R}$ is controlled invariant, since we may choose $u = -\lambda x_2$. One has

$$e^{tA} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = e^{\lambda t} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = e^{\lambda t} \begin{bmatrix} t \\ 1 \end{bmatrix}.$$

Thus $e^{tA}V \to \mathbb{R} \times \{0\}$ in projective space for $t \to \infty$. The solution in $\mathbb{R}^2/V$ identified with $\mathbb{R} \times \{0\}$ is given by

$$x_1(t) = te^{\lambda t} x_2^0.$$

The solutions $x_1(t)$ move apart with $e^{\lambda t}$, hence the subspace entropy is given by $h_{\mathrm{sub}}(V) = \lambda$. Again, the Lyapunov exponent in L_j determines the subspace entropy. Note that here $e^{tA}V$ converges to the orthogonal complement of V.

Example 18. Consider with $\lambda > 0$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}.$$

The eigenspace of A is $\mathbb{R}\times\{0\}\times\{0\}\times\{0\}$. The subspace $V=\{0\}\times\mathbb{R}^2\times\{0\}$ only contains the trivial A-invariant subspace and V is controlled invariant, since we may choose $u_1=-\lambda x_2-x_3, u_2=-\lambda x_3-x_4$. One has

$$e^{tA}\begin{bmatrix}0\\1\\0\\0\end{bmatrix}=e^{\lambda t}\begin{bmatrix}1 & t & \frac{t^2}{2} & \frac{t^3}{3!}\\ 0 & 1 & t & \frac{t^2}{2}\\ 0 & 0 & 1 & t\\ 0 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}0\\1\\0\\0\end{bmatrix}=e^{\lambda t}\begin{bmatrix}t\\1\\0\\0\end{bmatrix}, e^{tA}\begin{bmatrix}0\\0\\1\\0\end{bmatrix}=e^{\lambda t}\begin{bmatrix}\frac{t^2}{2}\\t\\1\\0\end{bmatrix}.$$

Thus $e^{tA}V\to\mathbb{R}^2\times\{0\}\times\{0\}$ in the Grassmannian $\mathbb{G}_2$ for $t\to\infty$. The solution in $\mathbb{R}^4/V$ identified with $\mathbb{R}\times\{0\}\times\{0\}\times\mathbb{R}$ is given by

$$\begin{bmatrix}x_1(t)\\x_4(t)\end{bmatrix}=\begin{bmatrix}e^{\lambda t}\frac{t^2}{2}\\0\end{bmatrix}.$$

The solutions in $\mathbb{R}^4/V$ move apart with $e^{\lambda t}\frac{t^2}{2}$, hence the subspace entropy is given by $h_{\text{sub}}(V)=\lambda$. One the other hand, the topological entropy of V in $\mathbb{R}^4$ is 2λ. Note that here $\dim V=2=\dim\mathbb{R}^4/V$.

We impose the following assumption: Let $v_1,\ldots,v_k$ be an orthonormal basis of V. Then there is $\gamma>0$ such that for a sequence $t_i\to\infty$ the absolute value of the volume of the parallelepiped spanned by $\pi(e^{t_iA}v_1),\ldots,\pi(e^{t_iA}v_k)$ is bounded below by a positive constant times the absolute value of the volume of the parallelepiped spanned by $e^{t_iA}v_1,\ldots,e^{t_iA}v_k$. More formally, we require:

There are an orthonormal basis $v_1,\ldots,v_k$ of V and $\gamma>0$ such that for a sequence $t_i\to\infty$

$$\left\|\pi(e^{t_iA}v_1)\wedge\cdots\wedge\pi\left(e^{t_iA}v_k\right)\right\|\geq\gamma\left\|e^{t_iA}v_1\wedge\cdots\wedge e^{t_iA}v_k\right\|. \tag{7}$$

Note that this assumption can only hold, if $n-k=\dim\mathcal{X}/V\geq k=\dim V$.

Proposition 19. *Let V be a subspace of $\mathcal{X}$ and suppose that condition (7) holds . Then for $A:\mathcal{X}\to\mathcal{X}$ the subspace entropy is given by*

$$h_{\text{sub}}(V)=h_{\text{top}}(V).$$

Proof. In view of Proposition 13 it only remains to show that $h_{\text{sub}}(V)\geq h_{\text{top}}(V)$. A consequence of (7) is that for all i

$$\log\left\|\pi(e^{t_iA}v_1)\wedge\cdots\wedge\pi\left(e^{t_iA}v_k\right)\right\|\geq\log\gamma+\log\left\|e^{t_iA}v_1\wedge\cdots\wedge e^{t_iA}v_k\right\|,$$

and hence

$$\limsup_{t\to\infty}\frac{1}{t}\log\left\|\pi(e^{tA}v_1)\wedge\cdots\wedge\pi\left(e^{tA}v_k\right)\right\|\geq\limsup_{t\to\infty}\frac{1}{t}\log\left\|e^{tA}v_1\wedge\cdots\wedge e^{tA}v_k\right\|. \tag{8}$$

Let K be a neighborhood of the origin in V. Then K contains a parallelepiped and we may assume that K contains the parallelepiped P spanned by $v_1,\ldots,v_k$. Then the set $e^{tA}K$ is a neighborhood of the origin in the k-dimensional subspace $e^{tA}V$ and it contains the parallelepiped spanned by

$$e^{tA}v_1, \cdots, e^{tA}v_k.$$

The projected set $\pi(e^{tA}K)$ is a neighborhood of the origin in $\pi(e^{tA}V)$ and, for $t = t_i$, it contains the parallelepiped $\pi(e^{t_iA}P)$ spanned by $\pi(e^{tA_k}v_1)$, $\cdots, \pi\left(e^{tA_k}v_\ell\right)$. By Colonius and Kliemann [5, Theorem 5.2.5] one finds

$$\lim_{t\to\infty}\frac{1}{t}\log\left\|e^{tA}v_1\wedge\cdots\wedge e^{tA}v_\ell\right\| = \sum_{i=1}^{l}k_i\lambda_i,$$

where $(k_1,...,k_l)$ is an element of the index set $I(k)$ given by (2).

It remains to relate the volume growth to the subspace entropy. We argue as in Colonius, San Martin, da Silva [6, Proposition 4.1]:

For $t > 0$ the k-dimensional volume of $\pi(e^{tA}P)$ satisfies

$$\mathrm{vol}^k(\pi(e^{tA}K)) \geq \mathrm{vol}^k(\pi(e^{tA}P)) = \left\|\pi(e^{tA}v_1)\wedge\cdots\wedge\pi\left(e^{tA}v_k\right)\right\|.$$

Let $\varepsilon > 0, T > 0$, and consider a (T,ε)-spanning set $R = \{x_1,...,x_r\} \subset P$ of minimal cardinality $r = r_{\mathrm{sub}}(T,\varepsilon,P,V)$ for the subspace entropy. Then (by the definition of spanning sets) the set $\pi(e^{TA}P)$ is contained in the union of r balls $B(\pi(e^{TA}x_j);\varepsilon)$ of radius ε in $\mathcal{X}/V$,

$$B(\pi(e^{TA}x_j);\varepsilon) = \{z \in \mathcal{X}/V \mid \left\|z - \pi(e^{TA}x_j)\right\| < \varepsilon\}.$$

Each such ball has volume bounded by $c(2\varepsilon)^{n-k}$, where $c > 0$ is a constant. Thus

$$\mathrm{vol}^k(\pi(e^{TA}P)) \leq r\cdot c(2\varepsilon)^{n-k}.$$

This yields

$$\begin{aligned}\log r_{\mathrm{sub}}(T,\varepsilon,P,V) &\geq \log\mathrm{vol}^k(\pi(e^{TA}P)) - \log\left[c(2\varepsilon)^d\right]\\ &= \log\left\|\pi(e^{TA}v_1)\wedge\cdots\wedge\pi\left(e^{TA}v_k\right)\right\| - \log\left[c(2\varepsilon)^{n-k}\right],\end{aligned}$$

and hence

$$\begin{aligned}&\limsup_{T\to\infty}\frac{1}{T}\log r_{\mathrm{sub}}(T,\varepsilon,P,V)\\ &\geq \limsup_{T\to\infty}\frac{1}{T}\log\left\|\pi(e^{TA}v_1)\wedge\cdots\wedge\pi\left(e^{TA}v_k\right)\right\|.\end{aligned}$$

Together with (8) one obtains the assertion for $\varepsilon \to 0$. □

As a consequence of the discussion above, we obtain the following characterization of the subspace entropy. It presents a stepwise reduction of the problem.

Theorem 20. *Let $A : \mathcal{X} \to \mathcal{X}$ be a linear map on a finite dimensional normed vector space X and consider a subspace V. Decompose the associated flow $\Phi_t := e^{tA}$ into the center-stable and the unstable parts $\Phi^{-,0}$ and Φ^+, respectively.*

(i) Then the subspace entropy satisfies

$$h_{\mathrm{sub}}(V,\Phi) = h_{\mathrm{sub}}(V,\Phi^+).$$

(ii) Let $\langle A|V\rangle$ and $\ker(A;V)$ denote the smallest A-invariant subspace containing V and the largest A-invariant subspace contained in V, respectively. Then the reduced flow $\Phi_t^{\mathrm{red}} = e^{tA^{\mathrm{red}}}$ which is induced on $\langle A|V\rangle/\ker(A;V)$ satisfies

$$h_{\mathrm{sub}}(V,\Phi^+) = h_{\mathrm{sub}}(V/\ker(A;V),\Phi^{\mathrm{red}}).$$

(iii) The topological entropy of the subspace $V/\ker(A;V)$ for the reduced flow Φ^{red} is an upper bound of the subspace entropy $h_{\mathrm{sub}}(V/\ker(A;V),\Phi^{\mathrm{red}})$,

$$h_{\mathrm{sub}}(V/\ker(A;V),\Phi^{\mathrm{red}}) \le h_{\mathrm{top}}(V/\ker(A;V),\Phi^{\mathrm{red}}). \tag{9}$$

(iv) If the reduced flow Φ^{red} on $\langle A|V\rangle/\ker(A;V)$ and the subspace $V/\ker(A;V)$ satisfy assumption (7), then equality holds in (9).

(v) The topological entropy of the subspace $V/\ker(A;V)$ for the reduced flow Φ^{red} is determined by certain eigenvalues of A: Let $k := \dim V/\ker(A;V)$. Then

$$h_{\mathrm{top}}(V/\ker(A;V),\Phi^{\mathrm{red}}) = \sum_i k_i \max(0,\lambda_i), \tag{10}$$

where λ_i are the real parts of the eigenvalues of A^{red}, and the k_i are given by the chain recurrent component $\mathcal{M}^k_{k_1,\dots,k_l}$ of Φ^{red} in the k-Grassmannian $\mathbb{G}_k(\langle A|V\rangle/\ker(A;V))$ containing the ω-limit set $\omega(V/\ker(A;V))$.

Proof. Assertion (i) follows from Proposition 12, (ii) is a consequence of Proposition 15 and (iii) follows from Proposition 13. Assertion (iv) holds by Proposition 19 and (v) follows by Proposition 19 applied to the reduced flow Φ^{red}. Finally, (v) is a consequence of the characterization of topological entropy in Theorem 3 applied to the reduced flow. □

In particular, Theorem 20 characterizes the invariance entropy $h_{\mathrm{inv}}(V)$ of a controlled invariant subspace V of a linear control system of the form (5). By Theorem 10 it coincides with the subspace entropy of $\Phi_t = e^{tA}$. One obtains the following corollary to Theorem 20.

Corollary 21. *The invariance entropy of a controlled invariant subspace V of a linear control system of the form (5) is bounded above by the topological entropy of the flow Φ^{red} induced by A on $\langle A|V\rangle/\ker(A;V)$, where $\langle A|V\rangle$ and $\ker(A;V)$ denote the smallest A-invariant subspace containing V and the largest A-invariant subspace contained in V, respectively. Hence*

$$h_{\mathrm{inv}}(V) \le h_{\mathrm{top}}(V/\ker(A;V),\Phi^{\mathrm{red}}) = \sum_i k_i \max(0,\lambda_i),$$

where the sum is over the eigenvalues λ_i of A as in (10). Equality holds, if the subspace $V/\ker(A;V)$ satisfies assumption (7) for Φ^{red}.

Acknowledgments

The author was supported by DFG grant Co 124/17-2 within DFG Priority Program 1305 “Control of Digitally Connected Dynamical Systems”.

Bibliography

[1] G. Basile and G. Marro. Controlled and conditioned invariant subspaces in linear system theory. *J. Optim. Theory Appl.*, 3:306–315, 1969. Cited p. 79.

[2] R. Bowen. Entropy for group endomorphisms and homogeneous spaces. *Trans. Amer. Math. Soc.*, 153:401–414, 1971. Erratum, 181:509–510, 1973. Cited p. 76.

[3] F. Colonius and U. Helmke. Entropy of controlled invariant subspaces. *Zeitschrift für Angewandte Mathematik und Mechanik*, submitted, 2011. Cited pp. 75, 78, and 81.

[4] F. Colonius and C. Kawan. Invariance entropy for control systems. *SIAM J. Control Optim.*, 48:1701–1721, 2009. Cited p. 75.

[5] F. Colonius and W. Kliemann. Dynamics and Linear Algebra. Book manuscript, 2011. Cited p. 88.

[6] F. Colonius, L. A. B. San Martin, and A. J. da Silva. Topological fiber entropy for linear flows on vector bundles. submitted, 2011. Cited pp. 77 and 88.

[7] L. Grüne and J. Pannek. *Nonlinear Model Predictive Control. Theory and Algorithms*. Springer, 2011. Cited p. 81.

[8] A. S. Matveev and A. V. Savkin. *Estimation and Control over Communication Networks*. Birkhäuser, 2009. Cited p. 76.

[9] C. Robinson. *Dynamical Systems. Stability, Symbolic Dynamics, and Chaos*. CRC Press, 1999. Cited p. 77.

[10] H. L. Trentelmann, A. A. Stoorvogel, and M. Hautus. *Control Theory for Linear Systems*. Springer, 2001. Cited p. 79.

[11] P. Walters. *An Introduction to Ergodic Theory*. Springer, 1982. Cited p. 76.

[12] W. M. Wonham. *Linear Multivariable Control: A Geometric Approach*. Springer, 1985. Cited p. 79.

Euclidean norm optimal realization revisited

Tobias Damm
University of Bayreuth
Bayreuth, Germany
tobias.damm@uni-bayreuth.de

Abstract. We consider Euclidean norm optimal realizations of linear control systems and suggest an alternative constructive approach to results obtained by Uwe Helmke. In particular, we avoid the use of methods from invariant theory.

1 Introduction

In the early nineties – or perhaps more appropriately, in *his* early forties, Uwe Helmke (in joint work with several coauthors) produced quite a number of results on balancing, sensitivity minimization, and optimal realization of linear control systems, see [9, 10, 12, 19]. Among his favourite tools at that time were gradient flows blended with methods from algebraic geometry and invariant theory that is, rather intricate methods from my simple point of view.

At a central point, however, many of the problems boiled down to rational matrix equations (see e.g. [19]), which I felt more comfortable with. Didi Hinrichsen and I were happy to take up these equations to apply our results developed for similar matrix equations arising in stochastic control. For the problem of L^2-sensitivity minimization considered in [10, 11] this has been worked out in [4, 6].

For the Euclidean norm balancing problem (see e.g. [9, 10, 16]) however, it is more difficult to verify all the assumptions needed in our approach. A major obstacle is the lack of certain definiteness properties that are present in the L^2-minimization problem and help essentially to find a stabilizing initial guess for our iteration scheme. Thus, up to now it has been unclear, whether the existence of a solution of the Euclidean norm balancing problem can also be shown without recurring to the Kempf-Ness theorem or similar tools.

It is the object of the present note to fill this gap and to give an alternative approach to Euclidean norm balanced systems and Euclidean norm optimal realizations. The proof is constructive and immediately amounts to an efficient algorithm, which is much faster than e.g. applying Runge-Kutta methods to the gradient flow as in [8].

In Section 2 we first give a brief account of Euclidean norm balancing and optimal Euclidean norm realization. Then in Section 3 we recall a non-local convergence result for Newton's method. These two issues are brought together in Section 4, in a new derivation of the main result of Section 2 based on the main result in Section 3. The last two sections contain algorithmic formulations and numerical examples.

It is my pleasure to dedicate this paper to Uwe on the occasion of his 60th birthday.

2 Euclidean norm balancing

Consider a strictly proper rational matrix $G \in \mathbb{R}^{p\times m}(s)$ of McMillan degree n and a minimal realization $(A,B,C) \in L_{n,m,p}(\mathbb{R}) := \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n}$ of

$$G(s) = C(sI - A)^{-1}B .$$

The set of all minimal realizations of $G(s)$ is given as the orbit of (A,B,C) under the similarity action $(S,(A,B,C)) \mapsto (SAS^{-1}, SB, CS^{-1})$ of $\mathrm{Gl}_n(\mathbb{R})$ on $L_{n,m,p}(\mathbb{R})$.

On $L_{n,m,p}(\mathbb{R})$ we consider the Euclidean norm (compare [18])

$$\|(A,B,C)\|^2 = \mathrm{tr}(AA^\top) + \mathrm{tr}(BB^\top) + \mathrm{tr}(CC^\top) = \mathrm{tr}(AA^\top + BB^\top + C^\top C) .$$

This norm is *orthogonally invariant* i.e. for all orthogonal matrices S we have

$$\|(A,B,C)\| = \|(SAS^{-1}, SB, CS^{-1})\| .$$

A realization (A,B,C) is called *Euclidean norm minimal*, if for all nonsingular S it satisfies

$$\|(A,B,C)\| \le \|(SAS^{-1}, SB, CS^{-1})\| .$$

If $X > 0$ is given with an arbitrary factorization $X = S^\top S$ then

$$\begin{aligned}\|(SAS^{-1}, SB, CS^{-1})\|^2 &= \mathrm{tr}(SAS^{-1}S^{-\top}A^\top S^\top + SBB^\top S^\top + S^{-\top}C^\top CS^{-1}) \\ &= \mathrm{tr}(AX^{-1}A^\top X + BB^\top X + X^{-1}C^\top C) =: f(X) . \qquad (1)\end{aligned}$$

Thus, to determine a Euclidean norm minimal realization in the similarity orbit of (A,B,C), it suffices to minimize f in (1) over all positive definite matrices X. In first-order approximation we have

$$\begin{aligned}f(X+\Delta) - f(X) &\approx \mathrm{tr}\left(-AX^{-1}\Delta X^{-1}A^\top X + AX^{-1}A^\top \Delta + BB^\top \Delta - X^{-1}\Delta X^{-1}C^\top C\right) \\ &= \mathrm{tr}\left(\left(-X^{-1}A^\top XAX^{-1} + AX^{-1}A^\top + BB^\top - X^{-1}C^\top CX^{-1}\right)\Delta\right) .\end{aligned}$$

Hence, $X_+ > 0$ is a critical point for f, if and only if the rational matrix equation

$$0 = -AX_+^{-1}A^\top - BB^\top + X_+^{-1}\left(A^\top X_+ A + C^\top C\right)X_+^{-1} \qquad (2)$$

is satisfied. This is a necessary condition for X_+ to be a local minimizer of f. If the critical point is unique, then it must be a minimizer, since f is radially unbounded, i.e. for all $X > 0$ we have $f(\alpha X) \to \infty$ for $\alpha \to \infty$.

Criteria for the existence and uniqueness of X_+ are, however, not so obvious. This question and generalizations of it have been answered by Uwe Helmke and his coauthors e.g. in [9, 10, 12, 19]. Using tools from invariant theory like the Kempf-Ness theorem [14] and the Azad-Loeb theorem [2] they derive the following central result.

Theorem 1. *Let $(A,B,C) \in L(n,m,p)$ be controllable and observable. Then the rational matrix equation* (2) *has a unique positive definite solution* $X_+ > 0$.

Remark 2. If X_+ is the unique positive definite solution of (2) and $S^\top S = X_+$, then the realization $(\tilde{A},\tilde{B},\tilde{C}) = (SAS^{-1}, SB, CS^{-1})$ satisfies

$$\begin{aligned} 0 &= S\left(-AX_+^{-1}A^\top - BB^\top + X_+^{-1}\left(A^\top X_+ A + C^\top C\right)X_+^{-1}\right)S^\top \\ &= -\tilde{A}\tilde{A}^\top - \tilde{B}\tilde{B}^\top + \tilde{A}^\top\tilde{A} + \tilde{C}^\top\tilde{C}\,. \end{aligned}$$

Definition 3. A realization (A,B,C) with the property $AA^\top + BB^\top = A^\top A + C^\top C$ is called *Euclidean norm balanced.*

In the following we suggest a different proof of Theorem 1 which is based on a non-local convergence result for Newton's method developed in [4, 6].

3 Newton's method

Let $H^n \subset \mathbb{K}^{n\times n}$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) denote the real space of real or complex $n\times n$ Hermitian matrices, endowed with the Frobenius inner product $\langle X\,,Y\rangle = \operatorname{tr}(XY)$ and the corresponding (Frobenius) norm $\|\cdot\|$. By $H_+^n = \{X \in H^n \mid X \geq 0\}$ we denote the closed convex cone of nonnegative definite matrices and by int (H_+^n) its interior, i.e. the open cone of positive definite matrices. The cone H_+^n induces a partial ordering on H^n. We write $X \geq Y$ if $X - Y \in H_+^n$, and $X > Y$ if $X - Y \in \operatorname{int} H_+^n$.

Following the presentation in [3], we recall three notions for operators on H^n, namely *resolvent positivity*, *concavity*, and *stabilizability* (see also [4, 6] and the references therein). The set-up is simplified slightly.

Definition 4. A linear operator $T : H^n \to H^n$ is called

- *positive* ($T \geq 0$) if it maps H_+^n to H_+^n.
- *inverse positive* if it is invertible and $T^{-1} \geq 0$.
- *resolvent positive*, if $(\alpha I - T)^{-1} \geq 0$ for all sufficiently large $\alpha > 0$.

Further, we denote the adjoint operator by T^* and write $\sigma(T)$ for the spectrum,

$\beta(T) = \max\{\Re(\lambda);\lambda \in \sigma(T)\}$ for the spectral abscissa, and
$\rho(T) = \max\{|\lambda|;\lambda \in \sigma(T)\}$ for the spectral radius of T.

Example 5. (i) Let $A \in \mathbb{K}^{n\times n}$. Then the operator $\Pi_A : H^n \to H^n$ defined by $\Pi_A(X) = A^*XA$ is positive. Its adjoint is $\Pi_A^* = \Pi_{A^*}$.

(ii) All positive operators $\Pi : H^n \to H^n$ are resolvent positive, since for $\alpha > \rho(\Pi)$ the resolvent $(\alpha I - \Pi)^{-1} = \sum_{k=0}^\infty \alpha^{-(k+1)}\Pi^k$ is positive.

(iii) Given $A \in \mathbb{K}^{n\times n}$, the associated *Lyapunov operator*

$$L_A : H^n \to H^n, \quad L_A(X) = A^*X + XA, \tag{3}$$

is resolvent positive but, in general, not positive. Its adjoint is $L_A^* = L_{A^*}$.

We will need the following version of the Perron-Frobenius theorem (see e.g. [15]).

Theorem 6. *Let $T : H^n \to H^n$ be a linear mapping.*

(i) *T is* positive $\Rightarrow \exists V \geq 0$, $V \neq 0$: $T^*(V) = \rho(T)V$.

(ii) *T is* resolvent positive $\Rightarrow \exists V \geq 0$, $V \neq 0$: $T^*(V) = \beta(T)V$.

As a consequence we have a generalization of Lyapunov's matrix theorem.

Theorem 7. [17] *Let $L : H^n \to H^n$ be resolvent positive and $\Pi : H^n \to H^n$ be positive. Then $L + \Pi$ is resolvent positive, and the following are equivalent:*

(i) *$L + \Pi$ is stable, i.e.* $\sigma(L+\Pi) \subset \mathbb{C}_-$.

(ii) $-(L+\Pi)$ *is inverse positive.*

(iii) $\sigma(L) \subset \mathbb{C}_-$ *and* $\rho\left(L^{-1}\Pi\right) < 1$.

(iv) $\exists X < 0 : \quad (L+\Pi)(X) > 0$.

Let R be a Fréchet-differentiable operator from some open domain $\operatorname{dom} R \subset H^n$ to H^n with $H^n_+ \subset \operatorname{dom} R$. By $R'_X(H)$ we denote the derivative of R at X in direction H.

Definition 8. The operator R is said to be *H^n_+-concave* on $\operatorname{dom} R$ if

$$R(Y) - R(Z) + R'_Y(Z-Y) \geq 0.$$

for all $Y \in \operatorname{dom} R$ and all $Z \in H^n_+$.

Definition 9. The operator R is said to be *stabilizable* if there exists a matrix $X \in \operatorname{dom} R$, such that $\sigma(R'_X) \subset \mathbb{C}_-$. The matrix X is then called *stabilizing* (for R).

Now we can state the non-local convergence result for Newton's method.

Theorem 10 ([6], [4])**.** *Let $R : H^n \to H^n$ have the following properties.*

(a) *The derivative R'_X is resolvent positive for all $X \in \operatorname{dom} R$.*

(b) *R is H^n_+-concave on* $\operatorname{dom} R$.

(c) *There exists $X_0 \in \operatorname{dom} R$ with* $R'_{X_0} \subset \mathbb{C}_-$.

(d) *There exists $\hat{X} > 0$ with* $R(\hat{X}) \geq 0$.

Then the iteration scheme

$$X_{k+1} = X_k - (R'_{X_k})^{-1}(R(X_k))$$

defines a sequence (X_k) in $\operatorname{dom} R$ *with the following properties:*

(i) $\forall k = 1, 2, \ldots : X_k \in H^n_+$, $X_k \geq X_{k+1} \geq \hat{X}$, *and* $\sigma(R'_{X_k}) \subset \mathbb{C}_-$.

(ii) *(X_k) converges to a limit matrix $X_+ > 0$ that satisfies $R(X_+) = 0$ and is the largest solution of $R(X) \geq 0$.*

(iii) $\exists X \in H^n_+ : R(X) > 0 \iff \sigma(R'_{X_+}) \subset \mathbb{C}_-$.
In this case the sequence (X_k) converges quadratically, and X_+ is the unique solution of $R(X) = 0$ which is stabilizing for R.

4 Proof of Theorem 1

We define the matrix operator

$$R(X) = -XAX^{-1}A^\top X - XBB^\top X + A^\top XA + C^\top C \tag{4}$$

and its dual

$$\tilde{R}(\tilde{X}) = A\tilde{X}A^\top + BB^\top - \tilde{X}A^\top\tilde{X}^{-1}A\tilde{X} - \tilde{X}C^\top C\tilde{X}\,. \tag{5}$$

Obviously, $X_+ > 0$ satisfies (2), if and only if $R(X_+) = 0$ which again is equivalent to $\tilde{R}(\tilde{X}_+) = 0$ for $\tilde{X}_+ = X_+^{-1}$. In the following, we will verify that R, defined on

$$\operatorname{dom} R = \{X \in H^n \,\big|\, \det X \neq 0\}$$

satisfies the conditions of Theorem 10. The same then holds for $\tilde{R}$ by duality.

By straightforward calculations we obtain the explicit form of the Fréchet derivative of R and check concavity. Analogous arguments can be found in [4, 7] for the problem of L^2-sensitivity minimization.

Lemma 11. *With the notation* (3), *the Fréchet derivative* $R'_X(\Delta)$ *of* R *is given by*

$$R'_X(\Delta) = -L_{(AX^{-1}A^\top + BB^\top)X}(\Delta) + XAX^{-1}\Delta X^{-1}A^\top X + A^\top \Delta A\,. \tag{6}$$

As the sum of a Lyapunov operator and positive operators, R'_X *is resolvent positive.*

Lemma 12. *The operator* R *is* H^n_+*-concave on* $\operatorname{dom} R$.

Proof. Both $X \mapsto A^\top XA$ and the quadratic mapping $X \mapsto -XBB^\top X$ are obviously H^n_+-concave on $\operatorname{dom} R$. It remains to analyze the operator

$$X \mapsto F(X) := -XAX^{-1}A^\top X\,.$$

For nonsingular Y and positive definite Z, we have

$$\begin{aligned} F(Y) - F(Z) + F'_Y(Z-Y) &= -YAY^{-1}A^\top Y + ZAZ^{-1}A^\top Z - ZAY^{-1}A^\top Y \\ &\quad + YAY^{-1}A^\top Y - YAY^{-1}A^\top Z + YAY^{-1}A^\top Y \\ &\quad + YAY^{-1}ZY^{-1}A^\top Y - YAY^{-1}A^\top Y \\ &= (ZAZ^{-1} - YAY^{-1})Z(ZA^\top Z^{-1} - YA^\top Y^{-1}) \geq 0\,. \end{aligned}$$

Thus F is H^n_+-concave on $\operatorname{dom} R$, which completes the proof. □

Now, we show that R and $\tilde{R}$ are stabilizable. This is the most difficult issue.

Lemma 13. *Consider* R *and* $\tilde{R}$ *defined in* (4) *and* (5).

(a) If (A,B) *is controllable, then there exists* $X_0 > 0$ *such that* $\sigma(R'_{X_0}) \subset \mathbb{C}_-$.

(b) If (A,C) *is observable, then there exists* $\tilde{X}_0 > 0$ *such that* $\sigma(\tilde{R}'_{\tilde{X}_0}) \subset \mathbb{C}_-$.

Proof. It suffices to prove (a), since again (b) is just the dual result.

To emphasize the dependence of $R'_X(\Delta)$ on A and B, let us write $R'_X(\Delta) = R'_X(\Delta, A, B)$ for the moment. It then follows from (6) that $R'_{\alpha^2 X}(\Delta, \alpha A, B) = \alpha^2 R'_X(\Delta, A, B)$. In particular $\sigma(R'_X(\cdot, A, B)) \subset \mathbb{C}_-$ if and only if $\sigma(R'_{\alpha^2 X}(\cdot, \alpha A, B)) \subset \mathbb{C}_-$ for some $\alpha > 0$. Therefore, without loss of generality, we can assume that $\rho(A) < 1$.

Under this assumption, we can define $X_0 > 0$ as the inverse of the discrete-time controllability Gramian, satisfying $AX_0^{-1}A^\top + BB^\top = X_0^{-1}$. Then

$$L_{(AX_0^{-1}A^\top + BB^\top)X_0}(\Delta) = L_{X_0^{-1}X_0}(\Delta) = L_I(\Delta) = 2\Delta\,, \tag{7}$$

and

$$\begin{aligned} R'_{X_0}(X_0) &= -2X_0 + X_0 A X_0^{-1} A^\top X_0 + A^\top X_0 A \\ &= -X_0 - X_0 BB^\top X_0 + A^\top X_0 A \le -X_0 BB^\top X_0 \le 0\,. \end{aligned} \tag{8}$$

The first inequality in (8) holds by the following argument. Consider

$$M = \begin{bmatrix} X_0 & A^\top \\ A & X_0^{-1} \end{bmatrix}.$$

For the Schur-complement with respect to the upper left block, we have

$$X_0^{-1} - AX_0^{-1}A^\top = BB^\top \ge 0\,,$$

which implies $M \ge 0$. Hence also the Schur-complement with respect to the lower right block is nonnegative, $X_0 - A^\top X_0 A \ge 0$, proving (8).

Inequality (8) implies that $\sigma(R'_{X_0}) \subset \overline{\mathbb{C}_-}$. To see this, let β be the spectral abscissa of R'_{X_0} and $(R'_{X_0})^*(V) = \beta V$ with $V \ge 0$, $V \ne 0$, according to Theorem 6.
Then we find $\beta \le 0$, since $\langle X, V\rangle > 0$ and

$$\beta \langle X, V\rangle = \left\langle X, (R'_{X_0})^*(V)\right\rangle = \left\langle R'_{X_0}(X), V\right\rangle \le -\left\langle X_0 BB^\top X_0, V\right\rangle \le 0\,. \tag{9}$$

In fact, we even have $\sigma(R'_{X_0}) \subset \mathbb{C}_-$. Otherwise, by Theorem 6, $\beta = 0 \in \sigma(R'_{X_0})$. Inequality (9) then implies $B^\top X_0 V = 0$. Moreover, using (7), we have

$$(R'_{X_0})^*(V) = -2V + X_0^{-1} A^\top X_0 V X_0 A X_0^{-1} + AVA^\top = 0\,.$$

Multiplying with $B^\top X_0$ from the left and with $X_0 B$ from the right we have

$$B^\top X_0 (R'_{X_0})^*(V) X_0 B = B^\top A^\top X_0 V X_0 A B + B^\top X_0 A V A^\top X_0 B = 0\,,$$

whence also $B^\top A^\top X_0 V = 0$. Exploiting this, we get

$$B^\top A^\top X_0 (R'_{X_0})^*(V) X_0 A B = B^\top (A^\top)^2 X_0 V X_0 A^2 B + B^\top A^\top X_0 A V A^\top X_0 A B = 0\,,$$

yielding $B^\top (A^\top)^2 X_0 V = 0$. Inductively, we find $B^\top (A^\top)^k X_0 V = 0$ for all $k \in \mathbb{N}$, contradicting controllability of (A, B). Thus $\sigma(R'_{X_0}) \subset \mathbb{C}_-$. □

Note that the result of this lemma is constructive. The matrix X_0 which plays the rôle of an initial guess in Theorem 10 is obtained from the controllability Gramian of $(\alpha A, B)$ after an arbitrary scaling of A with $\alpha > \rho(A)$.

To complete our derivation of Theorem 1 based on Theorem 10, we need to show that there exists an $\hat{X} > 0$ so that $R(\hat{X}) \geq 0$.

Lemma 14. *Consider R and $\tilde{R}$ defined in* (4) *and* (5)*. If (A,B) is controllable and (A,C) is observable, then there exists $\hat{X} > 0$ such that $R(\hat{X}) > 0$.*

Proof. Choose $\gamma > 0$ so that

$$\tilde{R}(I) + \gamma I = AA^\top + BB^\top - A^\top A - C^\top C + \gamma I \geq 0\,.$$

By the previous lemmas the operator $\tilde{X} \mapsto \tilde{R}(\tilde{X}) + \gamma I$ satisfies the conditions of Theorem 10. Hence there exists a matrix $\tilde{X}_\gamma > I$ with $\tilde{R}(\tilde{X}_\gamma) + \gamma I = 0$. Thus

$$\begin{aligned} 0 &= -\tilde{X}_\gamma^{-1}\left(\tilde{R}(\tilde{X}_\gamma) + \gamma I\right)\tilde{X}_\gamma^{-1} \\ &= -\tilde{X}_\gamma^{-1} A \tilde{X}_\gamma A^\top \tilde{X}_\gamma^{-1} - \tilde{X}_\gamma^{-1} BB^\top \tilde{X}_\gamma^{-1} + A^\top \tilde{X}_\gamma^{-1} A + C^\top C - \gamma X_\gamma^{-2} \\ &= R(\tilde{X}_\gamma^{-1}) - \gamma X_\gamma^{-2}\,, \end{aligned}$$

so that for $\hat{X} = \tilde{X}_\gamma^{-1} > 0$ we have $R(\hat{X}) = \gamma \hat{X}^2 > 0$. □

Altogether, the Lemmata 11 – 14 establish the conditions (a) – (d) of Theorem 10, and thus the existence of a unique stabilizing solution $X_+ > 0$ of $R(X) = 0$. Finally, we show that every positive definite solution necessarily is stabilizing, which proves that X_+ is the only positive definite solution.

Lemma 15. *Consider R defined in* (4) *and assume that (A,C) is observable. If $X > 0$ satisfies $R(X) \leq 0$, then $\sigma(R_X') \subset \mathbb{C}_-$.*

Proof. By our assumptions and Lemma 11, we have

$$R_X'(X) = -XAX^{-1}A^\top X - XBB^\top X + A^\top XA = R(X) - C^\top C \leq -C^\top C \leq 0\,.$$

We argue similarly as in the proof of Lemma 13. Let β be the spectral abscissa of R_X' and $(R_X')^*(V) = \beta V$ for some nonzero $V \geq 0$. Then

$$\beta \langle X, V \rangle = \langle R_X'(X), V \rangle \leq \langle -C^\top C, V \rangle \leq 0\,.$$

Then either $\beta < 0$ or $\beta = 0$. In the latter case, we have $CV = 0$ and thus

$$0 = C(R_X')^*(V)C^\top \geq CAVA^\top C\,,$$

which yields $CAV = 0$. By induction, we find $CA^kV = 0$ for all $k \in \mathbb{N}$ contradicting observability of (A,C). Hence $\beta < 0$, i.e. $\sigma(R_X') \subset \mathbb{C}_-$. □

5 Algorithmic aspects

We reformulate the results of the previous section in algorithmic form.

Algorithm 1: *Euclidean norm balanced realization*

```
 1: inputs
      A system $(A,B,C)$ in minimal realization
 2: outputs
      A Euclidean norm balanced system $(A,B,C)$
 3: Set $\alpha = 2\|A\|_\infty$
 4: Solve the Lyapunov equation $\frac{1}{\alpha^2}AXA^\top - X = -BB^\top$ for $X$
 5: Update $X \leftarrow \alpha^2 X^{-1}$
 6: repeat
 7:    Solve $L_{(AX^{-1}A^\top + BB^\top)X}(\Delta) - XAX^{-1}\Delta X^{-1}A^\top X - A^\top \Delta A = R(X)$ for $\Delta$
 8:    Update $X \leftarrow X + \Delta$
 9: until $\|\Delta\| < \mathrm{tol}$
10: Factorize $X = S^\top S$
11: Update $(A,B,C) \leftarrow (SAS^{-1}, SB, CS^{-1})$
```

Remark 16. (a) The choice $\alpha = 2\|A\|_\infty$ in 3 is made just for convenience. Other upper estimates of $\rho(A)$ might be used as well.

(b) The most expensive operation in the algorithm is the repeated solution of the linear equation in line 7. A naive direct solution has complexity $O(n^6)$. An iterative method of lower complexity has been described in [5].

(c) Another critical issue is the conditioning of the initial matrix X. In particular for high-order single-input single-output systems, the discrete-time controllability Gramian X computed in line 4 is known to be very ill-conditioned. This may cause numerical problems or even destroy the convergence.

To overcome the problem mentioned in (c) we may first solve the modified equation

$$-XAX^{-1}A^\top X - X(BB^\top + I)X + A^\top XA + C^\top C = 0 \tag{10}$$

by applying the steps 4–9 of Algorithm 1. In this case, the initial guess obtained from the equation

$$\frac{1}{\alpha^2}AXA^\top - X = -BB^\top - I$$

typically is well-conditioned. Multiplying equation (10) from both sides with X^{-1} and setting $\tilde{X} = X^{-1}$, we get

$$-A\tilde{X}A^\top - BB^\top - I + \tilde{X}A^\top\tilde{X}^{-1}A\tilde{X} + \tilde{X}C^\top C\tilde{X} = 0\,,$$

that is, after a change of sign,

$$\tilde{R}(\tilde{X}) = A\tilde{X}A^\top + BB^\top - \tilde{X}A^\top\tilde{X}^{-1}A\tilde{X} - \tilde{X}C^\top C\tilde{X} < 0\,.$$

According to Lemma 15, we have $\sigma(\tilde{R}'_{\tilde{X}}) \subset \mathbb{C}_-$, so that we have found an initial guess to solve the dual equation by the loop 6 of Algorithm 1.
Therefore, we suggest the following extended algorithm, which has the additional feature that it computes a *diagonal* balanced realization, where

$$AA^\top + BB^\top = A^\top A + C^\top C = \operatorname{diag}(\sigma_1, \dots, \sigma_n)\,, \quad \text{with } \sigma_1 \geq \dots \geq \sigma_n\,. \tag{11}$$

Algorithm 2: *Euclidean norm diagonal balanced realization*

```
1:  inputs
        A system $(A,B,C)$ in minimal realization
2:  outputs
        The diagonal Euclidean norm balanced system $(A,B,C)$
3:  Set $\alpha = 2\|A\|_\infty$
4:  Solve the Lyapunov equation $\frac{1}{\alpha^2}AXA^\top - X = -BB^\top - I$ for $X$
5:  Update $X \leftarrow \alpha^2 X^{-1}$
6:  repeat
7:      Solve $L_{(AX^{-1}A^\top+BB^\top+I)X}(\Delta) - XAX^{-1}\Delta X^{-1}A^\top X - A^\top \Delta A = R(X) + X^2$
8:      Update $X \leftarrow X + \Delta$
9:  until $\|\Delta\| < \text{tol}$
10: Update $X \leftarrow X^{-1}$
11: repeat
12:     Solve $L_{(A^\top X^{-1}A+C^\top C)X}(\Delta) - XA^\top X^{-1}\Delta X^{-1}AX - A\Delta A^\top = \tilde{R}(X)$
13:     Update $X \leftarrow X + \Delta$
14: until $\|\Delta\| < \text{tol}$
15: Factorize $X = SS^\top$
16: Update $(A,B,C) \leftarrow (S^{-1}AS, S^{-1}B, CS)$.
17: Compute orthogonal $U$ with $U(AA^\top + BB^\top)U^\top = \operatorname{diag}(\sigma_1,\dots,\sigma_n)$, sorted
18: Update $(A,B,C) \leftarrow (UAU^\top, UB, CU^\top)$
```

Remark 17. Again, there is some freedom in the choice of parameters. In line 4 any right-hand side $-BB^\top - P$ with a positive definite matrix P will do. Then also line 7 has to be adapted accordingly. Moreover, in line 9 a different stopping criterion could be used, since not the accurate solution of the matrix equation is required here, but just a stabilizing initial guess for the second iteration.

6 Numerical examples

We implemented Algorithm 2 with a Lyapunov preconditioned Krylov subspace solver for the generalized Lyapunov equations in lines 4 and 12 (cf. [5]). On a laptop with 1.4 GHz and 4 GB RAM it took about one second to balance a randomly generated system with $n = 10$, $m = p = 1$, about a minute for $n = 100$, $m = p = 10$ and about an hour for an example with $n = 500$, $m = p = 50$, all with tolerance 10^{-10}. These numbers are not meant to be significant, they just indicate roughly, what can be expected.

Example 18. As a simple reproducable example, let us consider the system (A,B,C), where A is the $N^2 \times N^2$ Poisson matrix in two dimensions, corresponding to N discretization points in each direction, $B = [I,0]^\top \in \mathbb{R}^{N^2 \times N}$ and $C = [0,I] \in \mathbb{R}^{N \times N^2}$.

The balanced system (A_b, B_b, C_b) for $N = 2$ has the form

$$A_b = \begin{bmatrix} 6.0200 & 0 & -0.2008 & 0 \\ 0 & 4.0200 & 0 & -0.2008 \\ 0.2008 & 0 & 3.9800 & 0 \\ 0 & 0.2008 & 0 & 1.9800 \end{bmatrix}, \; B_b = \begin{bmatrix} 0.4525 & -0.4525 \\ 0.4525 & 0.4525 \\ -0.4525 & 0.4525 \\ -0.4525 & -0.4525 \end{bmatrix}$$

$$C_b = \begin{bmatrix} -0.4525 & -0.4525 & -0.4525 & -0.4525 \\ 0.4525 & -0.4525 & 0.4525 & -0.4525 \end{bmatrix}.$$

For $N = 10$, the sparsity pattern of A_b is shown in Fig. 1. The matrices B_b and C_b are dense. Actually, the norm reduction here is marginal, where $\frac{\|(A_b, B_b, C_b)\|}{\|(A,B,C)\|} = 0.9979$. This is not very surprising, since the original system was almost symmetric.

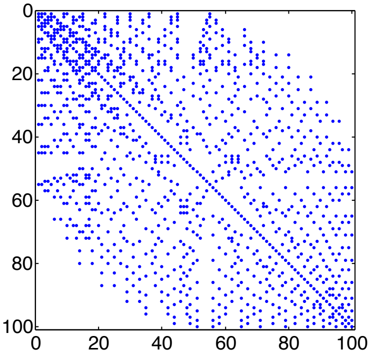


Figure 1: Sparsity pattern of A_b in Example 18 for $n = N^2 = 100$

Remark 19. In Gramian-based balancing, the typically fast decay of Hankel singular values is of special interest, since it indicates how well the system may be approximated by one of lower order (e.g. [1]). For Euclidean norm balancing, the decay of the corresponding values σ_k in (11) seems to be less rapid as can be seen in Fig. 2.

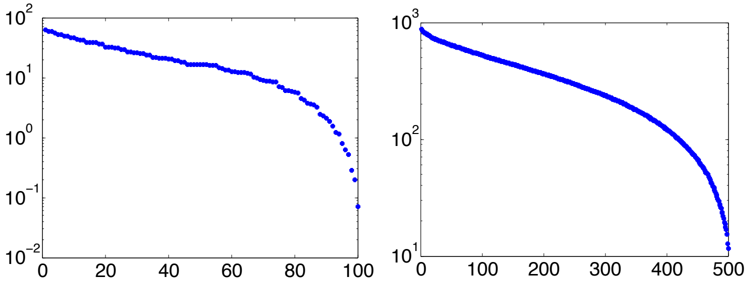


Figure 2: Decay of the σ_j defined in (11) for the system from Example 18 with $n = N^2 = 100$ (left) and a random system with $n = 500$, $m = p = 50$ (right).

Example 20. Euclidean norm balancing might be beneficial for badly conditioned systems. For instance, consider the system

$$A = \begin{bmatrix} -1/2 & 1000 & 0 \\ 0 & -1 & 1000 \\ 0 & 0 & -3/2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},$$

and its Euclidean norm balanced realization

$$A_b = \begin{bmatrix} -23.3635 & -15.7984 & 0.0312 \\ 15.7984 & -1.0000 & -15.8205 \\ 0.0312 & 15.8205 & 21.3635 \end{bmatrix}, \quad B_b = \begin{bmatrix} 15.8224 \\ 22.3607 \\ 15.8003 \end{bmatrix},$$

$$C_b = \begin{bmatrix} 15.8224 & -22.3607 & 15.8003 \end{bmatrix}.$$

Let further V and V_b denote the matrices containing the eigenvectors of A and A_b. It is well-known that the condition number of V and V_b describes the sensitivity of the eigenvalues of A and A_b, respectively, with respect to additive perturbations (see e.g. [13]). In the following table, we compare the condition numbers κ_2 of these matrices.

$\kappa_2(A)$	$\kappa_2(A_b)$	$\kappa_2(V)$	$\kappa_2(V_b)$
$1.3\cdot 10^9$	$4.3\cdot 10^4$	$8.5\cdot 10^6$	$8.5\cdot 10^3$

As can be seen in this example, the balanced system is much better conditioned than the original one. This effect can be observed in many examples (in particular with random data), although the sensitivity is not guaranteed to improve. In Example 18, actually, we have the opposite effect.

Bibliography

[1] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM, 2005. Cited p. 100.

[2] H. Azad and J. Loeb. On a theorem of Kempf and Ness. *Indiana Univ. J.*, 39:61–65, 1990. Cited p. 92.

[3] T. Damm. State-feedback H^∞-type control of linear systems with time-varying parameter uncertainty. *Linear Algebra Appl.*, 351–352:185–210, 2002. Cited p. 93.

[4] T. Damm. *Rational Matrix Equations in Stochastic Control*. Number 297 in Lecture Notes in Control and Information Sciences. Springer, 2004. Cited pp. 91, 93, 94, and 95.

[5] T. Damm. Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations. *Numer. Lin. Alg. Appl.*, 15(9):853–871, 2008. Cited pp. 98 and 99.

[6] T. Damm and D. Hinrichsen. Newton's method for a rational matrix equation occuring in stochastic control. *Linear Algebra Appl.*, 332–334:81–109, 2001. Cited pp. 91, 93, and 94.

[7] T. Damm and D. Hinrichsen. Newton's method for concave operators with resolvent positive derivatives in ordered Banach spaces. *Linear Algebra Appl.*, 363:43–64, 2003. Cited p. 95.

[8] N. Del Buono, L. Lopez, and C. Mastroserio. Runge Kutta type methods for isodynamical matrix flows: Applications to balanced realizations. *Computing*, 68:255–274, 2002. Cited p. 91.

[9] U. Helmke. Balanced realizations for linear systems: A variational approach. *SIAM J. Control Optim.*, 31(1):1–15, 1993. Cited pp. 91 and 92.

[10] U. Helmke and J. B. Moore. *Optimization and Dynamical Systems*. Springer, 1994. Cited pp. 91 and 92.

[11] U. Helmke and J. B. Moore. L^2 sensitivity minimization of linear system representations via gradient flows. *J. Math. Syst. Estim. Control*, 5(1):79–98, 1995. Cited p. 91.

[12] U. Helmke, J. B. Moore, and J. E. Perkins. Dynamical systems that compute balanced realizations and the singular value decomposition. *SIAM J. Matrix. Anal. & Appl.*, 15(3):733–754, 1994. Cited pp. 91 and 92.

[13] D. Hinrichsen and A. J. Pritchard. *Mathematical Systems Theory I. Modelling, State Space Analysis, Stability and Robustness*. Springer, 2005. Cited p. 101.

[14] G. Kempf and L. Ness. The length of vectors in representation spaces. In *Algebraic Geometry*, volume 732 of *Lecture Notes in Math.*, pages 233–244. Springer, 1979. Cited p. 92.

[15] M. A. Krasnosel'skij, J. A. Lifshits, and A. V. Sobolev. *Positive Linear Systems – The Method of Positive Operators*. Heldermann, 1989. Cited p. 93.

[16] J. E. Perkins, U. Helmke, and J. B. Moore. Balanced realizations via gradient flow techniques. *Syst. Control Lett.*, 14:369–380, 1990. Cited p. 91.

[17] H. Schneider. Positive operators and an inertia theorem. *Numer. Math.*, 7:11–17, 1965. Cited p. 94.

[18] E. Verriest. Minimum sensitivity implementation for multi-mode systems. In *Proc. 27th IEEE Conf. Decis. Control*, pages 2165–2170, 1988. Cited p. 92.

[19] W.-Y. Yan, J. B. Moore, and U. Helmke. Recursive algorithms for solving a class of nonlinear matrix equations with applications to certain sensitivity optimization problems. *SIAM J. Control Optim.*, 32:1559–1576, 1994. Cited pp. 91 and 92.

A note on generic accessibility and controllability of bilinear systems

Gunther Dirr
University of Würzburg
Würzburg, Germany
dirr@mathematik.uni-wuerzburg.de

Jens Jordan
University of Würzburg
Würzburg, Germany
jordan@mathematik.uni-wuerzburg.de

Indra Kurniawan
ITK Engineering AG
Stuttgart, Germany
indra.kurniawan@itk-engineering.de

Abstract. This short note deals with the issue of generic accessibility of bilinear control systems. We investigate (right-)invariant control systems evolving on a matrix Lie group G with Lie algebra $\mathfrak{g}$. Thereby, both the drift term and the control terms may vary in possibly different analytic subsets of $\mathfrak{g}$. Based on standard arguments on analytic functions, we derive a necessary and sufficient condition for generic accessibility within this class of bilinear systems. In combination with previous results in the literature, we obtain a particular simple genericity criterion if $\mathfrak{g}$ is semisimple. As an application, we demonstrate that almost all finite dimensional open quantum control systems (modelled by a Lindblad-Kossakowski master equation with controls entering only its Hamiltonian part) are accessible.

1 Introduction

Bilinear control systems constitute a class of nonlinear control systems which find numerous applications in many different areas such as physics, engineering, ecology and medicine [8, 18]. In most of these applications, the underlying dynamical models depend on partially unknown parameters. Therefore, one is interested in control properties which are valid not only for a particular bilinear control system but for all or at least a large subclasses of systems.

Probably, accessibility and controllability are the most fundamental properties of control system. Since the work of Lobry, Stefan and Sussmann (see [22] and the references therein) it is known that both properties are robust against small perturbations and accessibility is even a generic property for non-linear control systems (with respect to the fine C^k-topology). Furthermore, for linear systems a classical result says that also controllability in generic [21].

If it comes to bilinear systems less is known. There are only a few results mainly concerned with semisimple Lie groups. One result by Jurdjevic and Kupka [11, 12] is essentially that the set of all pairs which generate the whole Lie algebra $\mathfrak{sl}_n(\mathbb{R})$ is open and dense and therefore bilinear control systems on $\mathfrak{sl}_n(\mathbb{R})$ are generically accessible. The aim of this note is to extend this result in two directions: First, we derive a necessary and sufficient condition for generic accessibility (controllability) which is applicable to any bilinear system. Secondly, we focus on "structured" bilinear systems on semisimple Lie groups. Here, structured means that the drift term and the control terms are not allowed to vary in the entire Lie algebra but only in a prespecified "thin" subset. Such scenarios often arise in systems whose dynamics is related to some underlying weighted graph structure, where the weights may vary but not the graph structure itself.

For deriving the first result, we slightly modify the standard proof of generic controllability from linear systems theory. More precisely, the well-known fact that the set

$$\{(A,b) \in \mathfrak{gl}_n(\mathbb{R}) \times \mathbb{R}^n \mid \operatorname{span}\{b, Ab, \ldots, A^{n-1}b\} = \mathbb{R}^n\}$$

is open and dense in $\mathfrak{gl}_n(\mathbb{R}) \times \mathbb{R}^n$ is usually based on a simple argument about the zero set of polynomials. The same idea leads in the bilinear case to an if-and-only-if statement on generic accessibility (controllability). The second result, similar to the work by Jurdjevic and Kupka [11, 12] exploits heavily the structure theory of semisimple Lie algebras.

Finally, in the last section, we present an application of our results to quantum control. Most quantum processes (which satisfy the assumption of Markovian dynamics) can be modelled as bilinear control systems, e.g. [5, 6]. The controlled Lindblad-Kossakowski master equation [10, 17], which describes an open quantum system, i.e. a non-isolated quantum systems interacting with the environment, constitutes for instance a bilinear control system on the space of all density operators. It is known that the Lindblad-Kossakowski master equation with controls entering only its Hamiltonian part is never controllable [2, 6]. Nevertheless, accessibility, which guarantees that the reachable sets have at least non-empty interior, may apply. Our goal is to prove that accessibility is actually a generic property of the Lindblad-Kossakowski master equation even in the single control case. Similar statements dealing with the generic accessibility of open quantum systems also appeared in the work by C. Altafini [3].

The paper is organized as follows. Section 2 provides the basic facts on accessibility and controllability of bilinear control systems on Lie groups. Section 3 contains the main results: the general case is treated in Subsection 3.1; the real semisimple one in Subsection 3.2. In Section 4, we give an application of our results to open quantum systems. Most proofs are only sketched, more comprehensive details will be provided in a forthcoming full paper.

...and now for something completely different: HAPPY BIRTHDAY, UWE!

2 Preliminaries

To fix notation, let $\mathfrak{gl}_n(\mathbb{R})$ and $\mathfrak{gl}_n(\mathbb{C})$ be the Lie algebra of all real and, respectively, complex $n \times n$ matrices. Moreover, let $\mathfrak{so}_n(\mathbb{R}) \subset \mathfrak{gl}_n(\mathbb{R})$ and $\mathfrak{su}(n) \subset \mathfrak{gl}_n(\mathbb{C})$ denote the Lie subalgebras of all skew-symmetric and, respectively, all skew-Hermitian matrices with trace zero. For arbitrary $n \times n$ matrices, the trace and the commutator are given by $\operatorname{Tr}(A) := \sum_{k=1}^{n} a_{kk}$ and $[A,B] := AB - BA$, respectively. The identity matrix of size n is denoted by I_n or plainly by I, whenever the size is clear from the context.

Now, let $\mathfrak{g}$ be a Lie subalgebra of $\mathfrak{gl}_n(\mathbb{R})$, i.e. $\mathfrak{g}$ is a subspace of $\mathfrak{gl}_n(\mathbb{R})$ which is closed under taking commutators. Then there exists a unique Lie subgroup G of $GL_n(\mathbb{R})$ which corresponds to $\mathfrak{g}$ in the sense that the tangent space of G at the identity coincides with $\mathfrak{g}$. A bilinear or, equivalently, a (right)-invariant control systems on G is given by

$$(\Sigma) \qquad \dot{X} = \Big(A_0 + \sum_{k=1}^{m} u_k(t)A_k\Big)X, \quad X(0) = X_0 \in G, \tag{1}$$

where $A_0, A_1, \ldots A_m \in \mathfrak{g}$ and $u(t) := \big(u_1(t), \ldots, u_m(t)\big) \in U \subset \mathbb{R}^m$ is an admissible real-valued control input. For our purposes, the class of piecewise constant controls $u(\cdot)$ assuming values in $\mathbb{R}^m$ (i.e. $U = \mathbb{R}^m$) is convenient. However, in many cases the assumption $U = \mathbb{R}^m$ can be considerably relaxed by requiring that only the convex hull of the control set U contains the origin as an interior point [11].

Next, we define the terms *accessibility* and *controllability* for (bilinear) control systems. To this end, we need the concept of *reachability*. Let $\mathcal{R}_T(X_0)$ be the set of all $X \in G$ which can be reached from X_0 in time $T \geq 0$, i.e.

$$\mathcal{R}_T(X_0) := \{X_u(T) \mid u : [0,T] \to \mathbb{R}^m \text{ admissible control}\}, \tag{2}$$

where $X_u(\cdot)$ denotes the corresponding solution of Σ. Thus the entire *reachable set* of X_0 and Σ is given by

$$\mathcal{R}(X_0) := \bigcup_{T \geq 0} \mathcal{R}_T(X_0). \tag{3}$$

Then, Σ is called *accessible* if for all $X_0 \in G$ the reachable set $\mathcal{R}(X_0)$ has non-empty interior in G, and *controllable* if for all $X_0 \in G$ the reachable set $\mathcal{R}(X_0)$ is equal to G. As Σ is right-invariant one has $\mathcal{R}(X_0) = \mathcal{R}(I)X_0$ and therefore accessibility and controllability of Σ is equivalent to accessibility and, respectively, controllability at the identity I. Moreover, the so-called Lie algebra rank condition (LARC) which is in general only sufficient for accessibility yields the following necessary and sufficient accessibility criterion for right-invariant control systems.

Proposition 1. *Let Σ be defined as in* (1). *Then Σ is accessible if and only if Σ satisfies the LARC-condition at the identity, i.e. $\langle A_0, A_1, \ldots, A_n \rangle_L = \mathfrak{g}$.*

Here and henceforth, $\langle A_0, A_1, \ldots, A_n \rangle_L$ or, more generally, $\langle \mathcal{A} \rangle_L$ denotes the Lie subalgebra generated by $\mathcal{A} \subset \mathfrak{gl}_n(\mathbb{R})$, i.e. $\langle \mathcal{A} \rangle_L$ is the smallest Lie subalgebra of $\mathfrak{gl}_n(\mathbb{R})$ which contains $\mathcal{A}$ or, equivalently, the smallest subspace of $\mathfrak{gl}_n(\mathbb{R})$ which contains $\mathcal{A}$ and all iterated commutators of the form

$$[A_1,A_2], [A_1,[A_2,A_3]], [[A_1,A_2],A_3]], \Big[A_1,\big[A_2,[A_3,A_4]\big]\Big], \Big[\big[A_1,A_2\big],\big[A_3,A_4\big]\Big], \ldots$$

with $A_k \in \mathcal{A}$.

For controllability there is in general no such simple condition as for accessibility. Yet, for some special cases one has the following results.

Proposition 2. *Let Σ be defined as in* (1). *Then one has:*

(a) If Σ is additionally driftless (i.e. $A_0 = 0$) then controllability of Σ is equivalent to the Lie algebra condition $\langle A_1, \ldots, A_n \rangle_L = \mathfrak{g}$.

(b) *If G is compact then controllability of Σ is equivalent to the Lie algebra condition $\langle A_0, A_1, \ldots, A_n \rangle_L = \mathfrak{g}$.*

(c) *For $U = \mathbb{R}^m$, controllability of Σ is guaranteed by the Lie algebra condition $\langle A_1, \ldots, A_n \rangle_L = \mathfrak{g}$.*

Note that in the compact case accessibility and controllability of Σ are equivalent. A proof of both propositions can be found e.g. in [8, 11, 13]

Now, assume that the drift A_0 and the control terms $A_1, \ldots, A_m$ may vary in some non-empty subsets $D, C_1, \ldots, C_m \subset \mathfrak{g}$, respectively. Then $\Sigma(D; C_1, \ldots, C_m)$ denotes the family of all bilinear systems which can be obtained while $(A_0, A_1, \ldots, A_m)$ runs through $D \times C_1 \times \cdots \times C_m$. For $\Sigma(D; C, \ldots, C)$ we also write $\Sigma(D; C^m)$. To specify a particular system in $\Sigma(D; C_1, \ldots, C_m)$ we use the notation $\Sigma(A_0; A_1, \ldots, A_m)$. Thus we are prepared to state precisely what generic accessibility, controllability or, more general, genericity of any property of the family $\Sigma(D; C_1, \ldots, C_m)$ means.

- A property P is called (topologically) *generic* for $\Sigma(D; C_1, \ldots, C_m)$ if the set

  $$\left\{ (A_0, A_1, \ldots, A_m) \in D \times C_1 \times \cdots \times C_m \,\middle|\, \Sigma(A_0; A_1, \ldots, A_m) \text{ satisfies } P \right\} \tag{4}$$

  contains an open and dense subset of $D \times C_1 \times \cdots \times C_m$, where $D \times C_1 \times \cdots \times C_m$ is equipped with the topology induced by $\mathfrak{gl}_n(\mathbb{R})^{m+1}$.

- If D and $C_1, \ldots, C_m$ are smooth submanifolds of $\mathfrak{gl}_n(\mathbb{R})$, then P is called *generic* (with respect to the Lebesgue measure) for $\Sigma(D; C_1, \ldots, C_m)$ if the complement of the set defined by (4) has measure zero in $D \times C_1 \times \cdots \times C_m$ (cf. Remark 3 below).

Remark 3.

(a) If $D \times C_1 \times \cdots \times C_m$ is a Baire space, for instance, if $D, C_1, \ldots, C_m$ are smooth submanifolds, then topological genericity implies that the set defined by (4) is of second category.

(b) In general, topological genericity does not imply genericity with respect to the Lebesgue measure nor vice versa. Counterexamples can be obtained by Cantor-like sets.

(c) Sets of measure zero in $D \times C_1 \times \cdots \times C_m$ can simply be defined locally in coordinate charts and "globalized" via the partition of the unity.

General assumption and convention:

(a) From now on we assume $U = \mathbb{R}^n$ and that D and $C_1, \ldots, C_m$ are real analytic connected submanifolds of $\mathfrak{gl}_n(\mathbb{R})$.

(b) Whenever we do not specify the type of genericity (topological or with respect to the Lebesgue measure), the corresponding result holds for both types.

The trivial, but useful observation that $\Sigma(D, C_1, \ldots, C_{m'})$ is generically accessible (controllable) for all $m' \geq m$ if $\Sigma(D; C_1, \ldots, C_m)$ is generically accessible (controllable) allows us to put emphasis on the case $m = 1$.

We complete this preliminary section with an auxiliary results that is well known in Lie theory but maybe not in control theory. It yields an upper bound for the maximal Lie word length which has to be considered in "constructing" $\langle A_0, \ldots, A_m \rangle_L$. To this end, we define recursively the following sets. Let $\mathcal{A}$ be an arbitrary subset of $\mathfrak{g}$. Then,

$$L_1(\mathcal{A}) := \mathcal{A}, \quad L_n(\mathcal{A}) := \bigcup_{k=1}^{n-1} \big[L_k(\mathcal{A}), L_{n-k}(\mathcal{A})\big] \quad \text{for } n \geq 2 \tag{5}$$

and

$$L'_1(\mathcal{A}) := \mathcal{A}, \quad L'_n(\mathcal{A}) := \big[L'_1(\mathcal{A}), L'_{n-1}(\mathcal{A})\big] \quad \text{for } n \geq 2\,. \tag{6}$$

Clearly, while L_n includes all Lie words (over the alphabet $\mathcal{A}$) of length n, the set L'_n contains only Lie words of length n of the particular type

$$\Big[A_n, \big[A_{n-1}, [\ldots [A_1, A_0]]\big]\Big] \text{ with } A_k \in \mathcal{A}. \tag{7}$$

Lemma 4. *For $\mathcal{A} \subset \mathfrak{g}$ let $L_n(\mathcal{A})$ and $L'_n(\mathcal{A})$ be defined as above. Then*

(a) $\operatorname{span} L_n(\mathcal{A}) = \operatorname{span} L'_n(\mathcal{A})$ *for all $n \in \mathbb{N}$.*

(b) If $\operatorname{span} L'_{n_*+1}(\mathcal{A}) \subset \sum_{k=1}^{n_*} \operatorname{span} L'_k(\mathcal{A})$ *for some $n_* \in \mathbb{N}$ then* $\operatorname{span} L'_{n'}(\mathcal{A}) \subset \sum_{k=1}^{n_*} \operatorname{span} L'_k(\mathcal{A})$ *for all $n' \geq n_*$ and thus*

$$\langle \mathcal{A} \rangle_L = \sum_{k=1}^{n_*} \operatorname{span} L'_k(\mathcal{A})\,. \tag{8}$$

Proof. Part (a) follows by induction and the Jacobi identity; part (b) is a straightforward consequence of (a). A complete proof can be found in [4]. □

3 Main results

Our goal is to obtain necessary and sufficient conditions for generic accessibility and controllability of bilinear systems $\Sigma(D, C_1, \ldots, C_m)$. Our first result, Theorem 5, is in the spirit of the well-known generic controllability result for linear systems, see e.g. [21]. The proof is based on the same standard technique as in the linear case. It uses the fact that the zero set of a real analytic function is closed and nowhere dense.

Our second result, Theorem 7, heavily exploits the structure theory of real semisimple Lie algebras. It extends a well-known result by Jurdjevic and Kupka [12] on generic accessibility of bilinear systems on semisimple Lie groups.

3.1 General case

Theorem 5. *Let D and $C_1, \ldots, C_m$ be real analytic connected submanifolds of $\mathfrak{g} \subset \mathfrak{gl}_n(\mathbb{R})$. Then $\Sigma(D; C_1, \ldots, C_m)$ is generically accessible if and only if there exists at least one system $\Sigma(A_0; A_1, \ldots, A_m)$ such that $\langle A_0, A_1, \ldots, A_m \rangle_L = \mathfrak{g}$.*

Proof. According to Proposition 1 it suffices to show that the set

$$P := \{(A_0, A_1, \ldots, A_n) \in D \times C_1 \times \cdots \times C_m \mid \langle A_0, A_1, \ldots, A_m \rangle_L = \mathfrak{g}\} \tag{9}$$

contains an open and dense subset of $D \times C_1 \times \cdots \times C_m$. To this end, let $N := \dim \mathfrak{g}$ and let $L'_N(A_0, A_1, \ldots, A_m)$ be defined as in (6). Then Lemma 4 guarantees that

$$V_N := \sum_{k=1}^{N} \operatorname{span} L'_k(A_0, A_1, \ldots, A_m) \tag{10}$$

coincides with $\langle A_0, A_1, \ldots, A_m \rangle_L$ and hence $\langle A_0, A_1, \ldots, A_m \rangle_L = \mathfrak{g}$ is equivalent to $\dim V_N = N$. Now, let W denote the matrix that collects all Lie words of type (7) up to length N over the alphabet $A_0, A_1, \ldots, A_m$ (as column vectors in some coordinate representation). Since the condition $\dim V_N = N$ can be expressed as a rank condition on the matrix W, which is clearly a polynomial and thus real analytic condition, we can conclude that the set P is open dense in $D \times C_1 \times \cdots \times C_m$ if there is at least one $A_0, A_1, \ldots, A_m$ such that the matrix W has full rank. This, however, is guaranteed by the assumption that there exists at least one system such that $\langle A_0, A_1, \ldots, A_m \rangle_L = \mathfrak{g}$. Since the condition on W is polynomial the complement of P has also Lebesgue measure zero in $D \times C_1 \times \cdots \times C_m$. □

Proposition 2 immediately leads to the following controllability result.

Corollary 6. *If $\Sigma(D; C_1, \ldots, C_m)$ is a family of driftless systems, i.e. $D = \{0\}$, or if G is compact then the condition of Theorem 5 is equivalent to generic controllability of $\Sigma(D; C_1, \ldots, C_m)$.*

3.2 Real semisimple case

To follow and adapt the ideas by Jurdjevic and Kupka [11, 12] we first collect some basic facts on *strongly regular* elements of *real semisimple Lie-algebra*. For more details on semisimple Lie algebras we recommend in addition [14].

Let $\mathfrak{g} \subset \mathfrak{gl}_n(\mathbb{R})$ be a real semisimple Lie algebra and $\mathfrak{g}^{\mathbb{C}} := \mathfrak{g} \oplus i\mathfrak{g} \subset \mathfrak{gl}_n(\mathbb{C})$ be its complexification. Consider the corresponding adjoint representations

$$\operatorname{ad} : \mathfrak{g} \to \operatorname{End}(\mathfrak{g}) \quad \text{and} \quad \operatorname{ad}^{\mathbb{C}} : \mathfrak{g}^{\mathbb{C}} \to \operatorname{End}(\mathfrak{g}^{\mathbb{C}}).$$

For $A \in \mathfrak{g}$, define

$$\operatorname{Sp}(A) := \{\lambda \in \mathbb{C} \setminus \{0\} \mid \ker(\operatorname{ad}_A^{\mathbb{C}} - \lambda I_n) \neq \{0\}\} \tag{11}$$

and $E_\lambda^{\mathbb{C}}(A) := \ker(\operatorname{ad}_A^{\mathbb{C}} - \lambda I_n)$ for $\lambda \in \operatorname{Sp}(A)$ as the λ-eigenspace of $\operatorname{ad}_A^{\mathbb{C}}$. Then, an element $A \in \mathfrak{g}^{\mathbb{C}}$ is called *strongly regular* if it satisfies the following conditions:

- All nonzero eigenvalues of $\operatorname{ad}_A^{\mathbb{C}}$ are simple, i.e. the algebraic multiplicity of each $\lambda \in \operatorname{Sp}(A)$ is equal to one.
- The generalized eigenspace $E_0^{\mathbb{C}}(A) := \bigcup_{n \geq 1} \ker(\operatorname{ad}_A^{\mathbb{C}})^n$ does not contain any non-trivial ideal of $\mathfrak{g}^{\mathbb{C}}$.

It is known that strong regularity is a generic property in $\mathfrak{g}^{\mathbb{C}}$ and $\mathfrak{g}$ as well. More precisely, the set of all strongly regular elements is open and dense in $\mathfrak{g}^{\mathbb{C}}$ and its intersection with $\mathfrak{g}$ is again open and dense in $\mathfrak{g}$. In both cases, the complement has Lebesgue measure zero. Furthermore, the following facts about strongly regular elements are well-known and can be found e.g. in [12] or [14].

(1) With respect to a strongly regular element $A \in \mathfrak{g}$, the complex Lie algebra $\mathfrak{g}^{\mathbb{C}}$ decomposes as a direct sum

$$\mathfrak{g}^{\mathbb{C}} = E_0^{\mathbb{C}}(A) \oplus \bigoplus_{\lambda \in \mathrm{Sp}(A)} E_\lambda^{\mathbb{C}}(A). \tag{12}$$

(2) For every $\lambda \in \mathrm{Sp}(A)$, the set $\left[E_\lambda^{\mathbb{C}}(A), E_{-\lambda}^{\mathbb{C}}(A)\right]$ is a one-dimensional vector space contained in $E_0^{\mathbb{C}}(A)$. The sum of all $\left[E_\lambda^{\mathbb{C}}(A), E_{-\lambda}^{\mathbb{C}}(A)\right]$, $\lambda \in \mathrm{Sp}(A)$, equals $E_0^{\mathbb{C}}(A)$, i.e.

$$\sum_{\lambda \in \mathrm{Sp}(A)} \left[E_\lambda^{\mathbb{C}}(A), E_{-\lambda}^{\mathbb{C}}(A)\right] = E_0^{\mathbb{C}}(A). \tag{13}$$

Note that the above sum is in general *not* a direct sum.

(3) For $\lambda, \mu \in \mathrm{Sp}(A) \cup \{0\}$ one has

$$\left[E_\lambda^{\mathbb{C}}(A), E_\mu^{\mathbb{C}}(A)\right] = \begin{cases} E_{\lambda+\mu}^{\mathbb{C}}(A) & \text{for } \lambda + \mu \in \mathrm{Sp}(A) \cup \{0\}, \\ \{0\} & \text{for } \lambda + \mu \notin \mathrm{Sp}(A) \cup \{0\}. \end{cases} \tag{14}$$

(4) It turns out that $E_0^{\mathbb{C}}(A) = \ker(\mathrm{ad}_A^{\mathbb{C}})$. Moreover, $E_0^{\mathbb{C}}(A)$ is a Cartan subalgebra of $\mathfrak{g}^{\mathbb{C}}$, i.e. a maximal abelian subalgebra of $\mathfrak{g}$ whose ad-action on $\mathfrak{g}$ is simultaneously diagonalizable. For more details on Cartan subalgebras we refer to [14].

(5) With respect to a strongly regular element $A \in \mathfrak{g}$, the real Lie algebra $\mathfrak{g}$ decomposes as a direct sum

$$\mathfrak{g} = E_0(A) \oplus \bigoplus_{\lambda \in \mathrm{Sp}(A),\ \mathrm{Im}(\lambda) \geq 0} E_\lambda(A), \tag{15}$$

with $E_0(A) := E_0^{\mathbb{C}}(A) \cap \mathfrak{g}$ and $E_\lambda(A) := \left(E_\lambda^{\mathbb{C}}(A) + E_{\bar\lambda}^{\mathbb{C}}(A)\right) \cap \mathfrak{g}$, where $\bar\lambda$ denotes the complex conjugate of λ. Note that $E_\lambda(A) = E_{\bar\lambda}(A)$. Thus, depending on whether λ is real or not, $E_\lambda(A)$ is the real counterpart either to the eigenspace $E_\lambda^{\mathbb{C}}(A)$ or to the pair of eigenspaces $E_\lambda^{\mathbb{C}}(A)$ and $E_{\bar\lambda}^{\mathbb{C}}(A)$. Therefore, any $B \in \mathfrak{g}$ has a unique decomposition

$$B = B_0 + \sum_{\lambda \in \mathrm{Sp}(A),\ \mathrm{Im}(\lambda) \geq 0} B_\lambda, \tag{16}$$

where $B_0 \in E_0(A)$ and $B_\lambda \in E_\lambda(A)$ for $\lambda \in \mathrm{Sp}(A)$ with $\mathrm{Im}(\lambda) \geq 0$.

Now, for any strongly regular element $A \in \mathfrak{g}^{\mathbb{C}}$ we define

$$\Gamma(A) := \{B \in \mathfrak{g}^{\mathbb{C}} \mid B_\lambda \neq 0 \text{ for all } \lambda \in \mathrm{Sp}(A),\ \mathrm{Im}(\lambda) \geq 0\}. \tag{17}$$

Thus, we are prepared to state our main result on generic accessibility in the semi-simple case.

Theorem 7. *Let D and C be real analytic connected submanifolds of a real semisimple Lie algebra $\mathfrak{g} \subset \mathfrak{gl}_n(\mathbb{R})$. If D contains a strongly regular element (in the above sense) and if $C \cap \Gamma(A) \neq \varnothing$ then $\Sigma(D;C)$ is generically accessible.*

Before proceeding with the proof of Theorem 7, a few comments may be helpful.

Remark 8.

(a) Concerning accessibility, the role of D and C in Theorem 7 is completely interchangeable.

(b) The condition $C \cap \Gamma(A) \neq \varnothing$ can be refined by the results of Gauthier and Bornard [9], Silva-Leite and Crouch [19, 20] or Jurdjevic and Kupka [12]. However, these improvements are more of interest for analysing the accessibility and controllability of an individual system. For a simple genericity test the above condition $C \cap \Gamma(A) \neq \varnothing$ is usually sufficient.

For the proof of Theorem 7 we need two auxiliary results.

Lemma 9. *Let $A \in \mathfrak{g}$ be a strongly regular element. Then one has*

$$E_0(A) \subseteq \sum_{\lambda \in \mathrm{Sp}(A),\ \mathrm{Im}(\lambda) \geq 0} [E_\lambda(A), E_{-\lambda}(A)]. \tag{18}$$

A proof of Lemma 9 which follows straightforwardly from property (2) can be found in [15].

Lemma 10. *Let A be a strongly regular element in a real semi-simple Lie algebra $\mathfrak{g}$ and let $B \in \Gamma(A)$ Then one has $\langle A, B\rangle_L = \mathfrak{g}$.*

Proof. The inclusion $\langle A, B\rangle_L \subset \mathfrak{g}$ is trivial. Conversely, the span of $\mathrm{ad}_A B, \ldots, \mathrm{ad}_A^k B$ is an invariant subspace of ad_A for k sufficiently large. Therefore, we have

$$\mathrm{span}\{\mathrm{ad}_A B, \ldots, \mathrm{ad}_A^k B\} = \bigoplus_{\lambda \in \mathrm{Sp}(A),\ \mathrm{Im}(\lambda) \geq 0} E_\lambda(A) \subset \mathfrak{g}, \tag{19}$$

since all $E_\lambda(A)$ are irreducible subspaces of ad_A and, by assumption, $B_\lambda \neq 0$ for all $\lambda \in \mathrm{Sp}(A)$. By Lemma 9, summing $[E_\lambda(A), E_{-\lambda}(A)]$ for all $\lambda \in \mathrm{Sp}(A)$ with $\mathrm{Im}(\lambda) \geq 0$, we eventually generate $E_0(A)$. Hence, we obtain

$$\langle A, B\rangle_L \supset E_0(A) \oplus \bigoplus_{\lambda \in \mathrm{Sp}(A),\ \mathrm{Im}(\lambda) \geq 0} E_\lambda(A) = \mathfrak{g}, \tag{20}$$

and the result follows. □

Proof of Theorem 7. Due to Theorem 5 it suffices to show that there is at least one pair $A,B \in D \times C$ such that $\langle A,B \rangle_{LA} = \mathfrak{g}$. This, however, is guaranteed by Lemma 10 and the assumption that D contains a strongly regular A such that $C \cap \Gamma(A) \neq \varnothing$. □

Corollary 11. *(a) If G is compact then the conditions of Theorem* 7 *are sufficient for generic controllability of* $\Sigma(D;C)$.

(b) If $\Sigma(D;C_1,C_2)$ *is a family of systems with two controls then the conditions of Theorem* 7 *with* D,C *replaced by* C_1,C_2 *are sufficient for generic controllability of* $\Sigma(D;C_1,C_2)$.

The proof follows immediately from Proposition 2.

Based on Theorem 7, we can improve a result by Jurdjevic and Kupka in the sense that generic accessibility can be guaranteed for semisimple Lie algebras once one of the two sets D or C is sufficiently large.

Corollary 12. *Let* $C \neq \{0\}$ *be a real analytic connected submanifold of a real semisimple Lie algebra* $\mathfrak{g} \subset \mathfrak{gl}_n(\mathbb{R})$ *and let* $D = \mathfrak{g}$. *Then* $\Sigma(D;C)$ *is generically accessible.*

Proof sketch. Choose any strongly regular element $A \in \mathfrak{g}$ and any non-trivial $B \in C$. Now, consider the G-orbit of B, i.e. $\mathcal{O}_G(B) := \{XBX^{-1} \mid X \in G\}$, where G is the unique connected matrix Lie group with Lie algebra $\mathfrak{g}$. If $\Gamma(A) \cap \mathcal{O}_G(B) \neq \varnothing$ we are done, because $XBX^{-1} \in \Gamma(A)$ implies $B \in \Gamma(X^{-1}AX)$ and thus we can apply Theorem 7 to the strongly regular element $X^{-1}AX$. Therefore, we focus on the condition $\Gamma(A) \cap \mathcal{O}_G(B) \neq \varnothing$. To this end, choose any $\lambda \in \mathrm{Sp}(A)$. All we have to show is that the map $X \mapsto (XBX^{-1})_\lambda$ does not vanish identically. The fact that a holomorphic function vanishes identically on $\mathbb{C}^k$ if and only if its restriction to $\mathbb{R}^k$ vanishes identically allows us to pass form G to $G^{\mathbb{C}}$, the unique matrix Lie group which corresponds to $\mathfrak{g}^{\mathbb{C}}$. Now, we can exploit familiar properties of the Cartan subalgebra $E_0^{\mathbb{C}}(A)$, in particular, the transitive action of the associated Weyl group of the root spaces, cf. [14]. □

The following final result in this section turns out to be quite useful for reductive Lie algebras, i.e. if $\mathfrak{g} = \mathfrak{g}_0 \oplus \mathfrak{z}_0$ decomposes into a direct sum of a semisimple Lie algebra $\mathfrak{g}_0$ and a center $\mathfrak{z}_0$. In many cases, Proposition 13 allows to extend Theorem 7 to the reductive case.

Proposition 13. *Let* $\mathfrak{g} = \mathfrak{g}_0 \oplus \mathfrak{z}_0 \subset \mathfrak{gl}_n(\mathbb{R})$ *be a real reductive Lie algebra with semisimple component* $\mathfrak{g}_0$ *and center* $\mathfrak{z}_0$. *Moreover, let* $A = A_0 + Z$ *and* $B = B_0 + Z'$ *with* $A_0,B_0 \in \mathfrak{g}_0$ *and* $Z,Z' \in \mathfrak{z}_0$. *Then,* $\langle A,B \rangle_L = \mathfrak{g}_0 \oplus \mathrm{span}\{Z,Z'\}$ *if and only if* $\langle A_0,B_0 \rangle_L = \mathfrak{g}_0$.

Proof. Clearly, $\langle A,B \rangle_L$ is a subset of $\langle A_0,B_0 \rangle_L \oplus \mathrm{span}\{Z,Z'\}$. Therefore, $\langle A,B \rangle_L = \mathfrak{g}_0 \oplus \mathrm{span}\{Z,Z'\}$ implies $\langle A_0,B_0 \rangle_{LA} = \mathfrak{g}_0$. Conversely, if $\langle A_0,B_0 \rangle_L = \mathfrak{g}_0$, then by the semisimplicity of $\mathfrak{g}_0$ it follows $[\mathfrak{g}_0,\mathfrak{g}_0] = \mathfrak{g}_0$ and thus A_0 and B_0 are contained in the commutator algebra of $\langle A_0,B_0 \rangle_L$. Since the two commutator algebras of $\langle A_0,B_0 \rangle_L$ and $\langle A,B \rangle_L$ obviously coincide, one has $A_0,B_0 \in \langle A,B \rangle_L$ and hence the identity $\langle A,B \rangle_L = \mathfrak{g}_0 \oplus \mathrm{span}\{Z,Z'\}$. □

4 An application to open quantum systems

Here, we present a typical application of the previous results to bilinear control systems arising in open quantum dynamics. Let $\mathfrak{her}_0(n)$ be the set of all hermitian $n \times n$-matrices with trace zero and consider the following bilinear control system on $GL\big(\mathfrak{her}_0(n)\big)$:

$$\dot{X} = \Big(A_0 - \mathrm{i}\sum_{k=1}^{m} u_k(t)\mathrm{ad}_{H_k}\Big)\cdot X\,, \quad X(0) := I_{\mathfrak{her}_0(n)}\,, \tag{21}$$

where all H_k are traceless hermitian $n \times n$-matrices and A_0 can be for now an arbitrary linear operator acting on $\mathfrak{her}_0(n)$. Moreover, let $\mathrm{ad}_{\mathfrak{su}(n)}$ denote the adjoint action of $\mathfrak{su}(n)$ of $\mathfrak{her}_0(n)$, i.e.

$$\mathrm{ad}_{\mathfrak{su}(n)} := \{\mathrm{iad}_H := [\mathrm{i}H,\cdot] : \mathfrak{her}_0(n) \to \mathfrak{her}_0(n) \mid \mathrm{i}H \in \mathfrak{su}(n)\}\,. \tag{22}$$

Then, Corollary 12 and Proposition 13 imply the following preliminary result.

Theorem 14. *Let* $D := \mathfrak{gl}\big(\mathfrak{her}_0(n)\big)$ *and* $C := \mathrm{ad}_{\mathfrak{su}(n)} \subset \mathfrak{gl}\big(\mathfrak{her}_0(n)\big)$. *Then* (21) *is generically accessible.*

The controlled unital Lindblad-Kossakowski master equation

The state of a finite dimensional n-level quantum system is completely described by its density operator ρ. Thus the entire state space is given by the compact convex set

$$\mathcal{D} := \big\{\rho \in \mathbb{C}^{n\times n} \,|\, \rho = \rho^\dagger \geq 0\,,\ \mathrm{Tr}(\rho) = 1\big\} \tag{23}$$

of all positive semidefinite operators with trace one acting on the Hilbert space $\mathcal{H} := \mathbb{C}^n$. In what follows, we consider only open quantum control systems described by the *Lindblad-Kossakowski* master equation [10, 17] with *coherent* control inputs, i.e. the control inputs enter only the Hamiltonian part of the systems. Precisely, we have

$$\dot{\rho} = L(\rho) = -\mathrm{i}\Big[H_0 + \sum_{k=1}^{m} u_k(t)H_k, \rho\Big] + L_D(\rho)\,, \quad \rho(0) = \rho_0 \in \mathcal{D} \tag{24}$$

where $H_0 \in \mathfrak{her}_0(n)$ and $H_1,\dots,H_m \in \mathfrak{her}_0(n)$ denote the internal *drift Hamiltonian* and external *control Hamiltonians*, respectively. As before, $u(t) := \big(u_1(t),\dots,u_m(t)\big)$ are admissible time-dependent control signals with values in $U := \mathbb{R}^m$. The *dissipative drift* term L_D, which models various interactions with the environment, can be expressed as a linear operator of the following form [10, 17]

$$L_D(\rho) = \frac{1}{2}\sum_{j,k=1}^{n^2-1} a_{jk}\big([B_j,\rho B_k] + [B_j\rho, B_k]\big)\,. \tag{25}$$

Here, without loss of generality, we take $(B_1,\dots,B_{n^2-1})$ to be any orthonormal basis of $\mathfrak{her}_0(n)$. Moreover, $A := (a_{jk})_{j,k=1,\dots,n^2-1}$ has to be positive semidefinite to guarantee *complete positivity* of the semi-flow $\big(e^{tL}\big)_{t\geq 0}$.

For the definition of complete positivity and issues related to its physical interpretations in open quantum systems, we recommend to consult e.g. [1]. For further issues on completely positive maps and their relations to Lie semigroups, Lie wedges and reachable sets of open quantum systems, see also [7, 15, 16] and the references therein. The Lindblad-Kossakowski master equation (24) is called *unital* if its flow leaves the density matrix $\rho = I_n/n$ invariant, i.e. if $L(I_n) = 0$. Otherwise, when $L(I_n) \neq 0$, it is called *non-unital*. Now, we are ready to state and prove the announced genericity result for the unital Lindblad-Kossakowski master equation.

Theorem 15. *The unital n-level Lindblad-Kossakowski master equation with a single coherent control is generically accessible. More precisely, let* $C := \mathrm{ad}_{\mathfrak{su}(n)}$ *and let* D *denote the set of all operators acting on* $\mathfrak{her}(n)$ *of the form* $-\mathrm{ad}_{\mathrm{i}H_0} + L_D$*, where* $\mathrm{i}H_0$ *is in* $\mathfrak{su}(n)$ *and* L_D *is unital and given by* (25)*. Then, the family of bilinear control systems described by* (24) *is generically accessible with respect to* $D \times C$*.*

Proof. Instead of $\rho \in \mathcal{D}$ consider the reduced density matrix $\hat{\rho} := \rho - I_n/n \in \mathfrak{her}_0(n)$. Since (24) is assumed to be unital, the time evolution of $\hat{\rho}$ follows again (24). Moreover, if we can show that the group lift of (24) to $GL\big(\mathfrak{her}_0(n)\big)$ given by

$$\dot{X} = \Big(L_D - \mathrm{i}\mathrm{ad}_{H_0} - \mathrm{i}\sum_{k=1}^{m} u_k(t)\mathrm{ad}_{H_k}\Big)\cdot X\,, \quad X(0) := I_{\mathfrak{her}_0(n)}\,, \tag{26}$$

is generically accessible then the same holds for (24) as $GL\big(\mathfrak{her}_0(n)\big)$ acts clearly transitively on $\mathfrak{her}_0(n)$. Now, by Theorem 14 we know that generic accessibility holds with respect to $\mathfrak{gl}\big(\mathfrak{her}_0(n)\big) \times \mathrm{ad}_{\mathfrak{su}(n)}$. Since it is known [15] that the set D is a closed convex cone of $\mathfrak{gl}\big(\mathfrak{her}_0(n)\big)$ with non-empty interior the result follows. □

A similar result holds for the non-unital case, the interested reader is referred to [15].

Acknowledgments

The work by I. Kurniawan was fully supported by the Elite Network of Bavaria (ENB) from the Bavarian State Ministry of Science, Research and Arts, under the framework of the International Doctorate Program in Engineering and Computer Science: *Identification, Optimization and Control with Applications in Modern Technologies*.

Bibliography

[1] R. Alicki and K. Lendi. *Quantum dynamical semigroups and applications*, volume 717 of *Lecture Notes in Physics*. Springer, second edition, 2007. Cited p. 113.

[2] C. Altafini. Controllability properties for finite dimensional quantum Markovian master equations. *J. Math. Phys.*, 44(6):2357–2372, 2003. Cited p. 104.

[3] C. Altafini. Coherent control of open quantum dynamical systems. *Phys. Rev. A*, 70:062321, 2004. Cited p. 104.

[4] N. Bourbaki. *Lie Groups and Lie Algebras, Chapters 1-3*. Springer, second edition, 1989. Cited p. 107.

[5] D. D'Alessandro. *Introduction to Quantum Control and Dynamics*. Chapman & Hall/CRC, 2008. Cited p. 104.

[6] G. Dirr and U. Helmke. Lie theory for quantum control. *GAMM-Mitteilungen*, 31(1):59–93, 2008. Cited p. 104.

[7] G. Dirr, U. Helmke, I. Kurniawan, and T. Schulte-Herbrüggen. Lie-semigroup structures for reachability and control of open quantum systems: Kossakowski-Lindblad generators form Lie wedge to Markovian channels. *Rep. Math. Phys.*, 64(1-2):93–121, 2009. Cited p. 113.

[8] D. Elliott. *Bilinear Control Systems: Matrices in Action*. Springer, 2009. Cited pp. 103 and 106.

[9] J. Gauthier and G. Bornard. Contrôlabilité des systèmes bilinéaires. *SIAM J. Control and Optimization*, 20(3):377–384, 1982. Cited p. 110.

[10] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan. Completely positive dynamical semigroups of N-level systems. *J. Mathematical Phys.*, 17(5):821–825, 1976. Cited pp. 104 and 112.

[11] V. Jurdjevic. *Geometric Control Theory*. Cambridge University Press, 1997. Cited pp. 103, 104, 105, 106, and 108.

[12] V. Jurdjevic and I. Kupka. Control systems on semi-simple Lie groups and their homogeneous spaces. *Ann. Inst. Fourier, Grenoble*, 31(4):151–179, 1981. Cited pp. 103, 104, 107, 108, 109, and 110.

[13] V. Jurdjevic and H. J. Sussmann. Control systems on Lie groups. *J. Differential Equations*, 12:313–329, 1972. Cited p. 106.

[14] A. W. Knapp. *Lie Groups Beyond an Introduction*. Birkhäuser, second edition, 2002. Cited pp. 108, 109, and 111.

[15] I. Kurniawan. *Controllability aspects of the Lindblad-Kossakowski master equations, a Lie-theoretical approach*. PhD thesis, University of Würzburg, 2010. Cited pp. 110 and 113.

[16] I. Kurniawan, G. Dirr, and U. Helmke. Controllability aspects of quantum dynamics: A unified approach for closed and open systems. To appear in IEEE Transaction on Automatic Control, special issue on Quantum Control. Cited p. 113.

[17] G. Lindblad. On the generators of quantum dynamical semigroups. *Comm. Math. Phys.*, 48(2):119–130, 1976. Cited pp. 104 and 112.

[18] R. R. Mohler. *Bilinear control processes. With applications to engineering, ecology, and medicine*. Academic Press, 1973. Cited p. 103.

[19] F. Silva Leite. Pairs of generators for compact real forms of the classical Lie algebras. *Lin. Algebra and its Applications*, 121:123–133, 1989. Cited p. 110.

[20] F. Silva Leite and P. Crouch. Controllability on classical Lie groups. *Math. Control Signals Systems*, 1:31–42, 1988. Cited p. 110.

[21] E. Sontag. *Mathematical Control Theory*. Springer, 1990. Cited pp. 103 and 107.

[22] H. J. Sussmann. Some properties of vector field systems that are not altered by small perturbations. *J. Differential Equations*, 20:292–315, 1976. Cited p. 103.

Detection of motion direction of targets using a turtle retinal patch model

Mervyn P. B. Ekanayake
Department of Electrical and
Electronics Engineering
Univ. of Peradeniya
Sri Lanka

Bijoy K. Ghosh
Department of Mathematics
and Statistics
Texas Tech University
Lubbock, TX 79409, USA

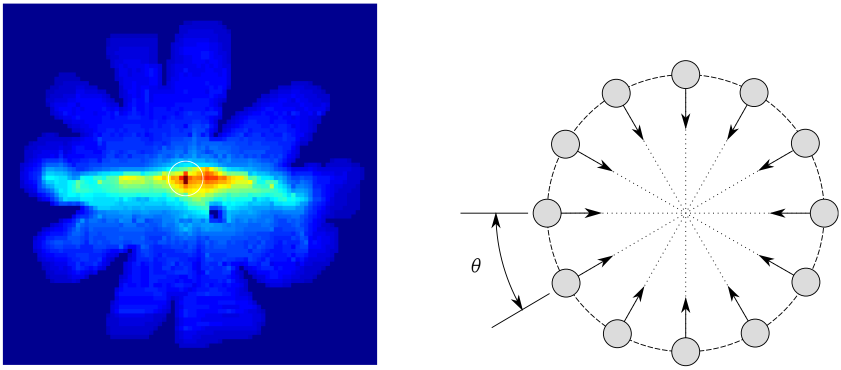
1 Introduction

Direction selectivity is an important feature of visual systems that has caught the attention of neuroscientists for over 100 years [8]. Directionally selective responses have been recorded by Hubel [10] in the primary visual cortex of an awake cat. Subsequently, Barlow and Levick [3] studied direction selectivity in the retinal ganglion cells of rabbits. Our interest in this paper is to study direction selectivity as a part of our ongoing study of modeling the visual pathway of freshwater turtles. In [16], we show that visual inputs produce waves that propagate across the visual cortex of freshwater turtles and visual information is encoded in the cortical waves. In all of our prior models the visual input was directly incident on the lateral geniculate, completely bypassing the retina. The purpose of this paper is to add a model of the retinal cell to the pathway and to study how retinal signals encode a moving point target incident on a small retinal patch. The target is moving with a constant, possibly unknown, velocity along directions that are spread across the entire 360°.

Turtle Retinal ganglion cells are either ON type, OFF type or ON-OFF type. The ON type cells have an excitatory center and inhibitory surrounding. The OFF type cells have an inhibitory center and excitatory surrounding. Finally the ON-OFF type cells have an excitatory center, an inhibitory annulus followed by an outermost excitatory surrounding (e.g. see [15]). Some of the turtle retinal cells are sensitive to the direction of the optical flow of an image sequence, (e.g. [1, 4]). These cells are called, *direction sensitive cells* or the B cells. The other cells, which are not sensitive to the direction of motion (but are sensitive to the intensity of the target), are called the A cells. The A cells can be ON or OFF type, whereas, the B cells are ON-OFF type. The A cells have a larger cell body (soma size) as well as a larger dendritic arborization. This results in a larger receptor field, compared to the B cells (see [15]). The A cells are smaller in number compared to the B cells on the turtle retina [18].

The turtle retina effectively encodes the motion parameters of moving targets in its visual space (see [19]). The retinal ganglion cells encode input signals using a sequence of spikes. We reproduce the spike generation process using a set of filters which model layers of rods and cones in the retina. The A cells have a similar block diagram except the directional filters are absent. The output of the filters are incident on a single compartment spike generating neuronal cell, with primarily sodium and potassium channels, modeled using Hodgkin-Huxley equations (see [9]). For a physiologically detailed model of a single cell that includes many additional channels (such as transient AHP channel, sustained calcium channel, calcium dependent potassium channel and transient calcium channel), see [7].

We consider a patch of the retina (see Fig. 1a) and circular targets that are moving with unknown constant speed and motion direction (see Fig. 1b). Our objective is described in the following two problems. In the first problem we consider targets that are moving along an unknown direction with fixed speed that are assumed to be known a priori. Our goal is to detect the motion direction of the unknown target. In the second problem we assume that both the direction and the speed of the targets are unknown. Our objective is to first estimate the speed and use this information to detect the motion direction. We remark that the second problem is biologically realistic but begin our analysis with the first problem because it sheds light on the detectability of B cells in comparison to the A cells. The first problem is also a prerequisite to solving the second.



(a) Cells on the turtle retina

(b) Paths of the incident light spot

Figure 1: (Left) Distribution of all cells on the retina showing the visual streak. The circle on the center of the streak indicates the location of the retinal patch. (Right) The input to the retina is a circular spot of light moving from one end of the patch to the other in a straight line.

Two methods to detect the unknown motion direction of the target are now described. In the first method, we use Principal Component Analysis (PCA) (see [16], [11]). The spike sequence generated by each cell in the model patch is low pass filtered using a second order linear filter. The filter output is a continuous signal that approximates the spike rate of the corresponding cell in the patch. The vector of spike rate functions over a suitable time window are projected as points on a cartesian coordinate system using PCA. We model these points as realizations of a Gaussian process, conditioned on the direction of motion, and detect the target motion direction using the well known Maximum Likelihood Estimation Method ([22]). In the second method we hypothesize that the pattern of spiking activity in a cell can be described by a class of point process, called Self Exciting Poisson Process (see [20]). We use the fact that a collection of such processes can be pooled together, and under an appropriate hypothesis (see [20]), can be modeled as an inhomogeneous Poisson Process. We pool together the spiking activities of a subpatch of cells in a patch and represent the pooled activities as an inhomogeneous vector Poisson Process. We estimate intensity functions for each component of this process, conditioned on the input direction of the target. The direction of motion of an unknown input can be detected using estimates of the intensity function vector.

2 Retinal cell modeling and construction of a retinal patch

Turtles, being vertebrates, have a multi layered retinal structure. From the point of view of visual signal processing, it has layers of photoreceptive cells consisting of cones and rods. These cells are synaptic to a layer of ganglion cells which give rise to the optic nerve fibers (see [19]).

We model the layer of photoreceptive cells as a cascade of filters which represent key functions, including the spatial and temporal variation of the receptor field (see [5]) and direction selectivity only for the B cells (see [4]). The ganglion cells are modeled as firing neurons using the Hodgkin-Huxley (HH) model calibrated with parameters from [13] and [14]. Noise is modeled as a zero mean Gaussian current input to the HH model. In the following subsections the function of the major components of the filter model are described. This model was originally reported by Baker [2] and we refer the reader to [7] for details.

The ganglion cells on a turtle retina are distributed in such a way that it is possible to observe a high density of cells along a line called the visual streak. The spatial distribution of turtle ganglion cells on the retina has been studied in [17]. It reports the cell density over a multitude of vertical and horizontal transects as to cover the entire retina. We interpolate the data provided using a two dimensional cubic spline to compute the cell density (both A and B types combined) over the whole retina.

In a subsequent paper [18], the distribution of the size of ganglion cells at some selected sites of the retina has been detailed. By inspecting this data, we can conclude that the histogram of the cell body size is bimodal. We fit this histogram data with sum of two Gaussian distributions. Additionally, we observe from [15], that the A cells have a large soma size compared to B cells. Therefore, we claim that in the bimodal distribution, the A cells are distributed with higher mean cell size and the B cells are distributed with lower mean cell size. The percentage of A cells calculated at each site are interpolated over the entire retina using a two dimensional cubic spline. Multiplying the percentage of A cells with the cell density data calculated as above, we obtain the distribution of A cells over the entire retina. This procedure is repeated for the B cells. Fig. 1a shows the distribution of the entire population of retinal ganglion cells.

Since the turtle retina has 350–390 thousand cells, we use about 1% of that for constructing large-scale models of the full retina. The majority of cells are B cells. The B cells are divided into three equal groups, corresponding to three distinct direction preferences. The A cells are divided in to two equal groups based on their receptor field structure, known as the A-ON and the A-OFF. The three groups of the B cells and the two groups of A cells are randomly sprinkled over the retina to match the computed density functions.

A circular retinal patch has been used to obtain results reported in this paper. The patch is taken to be the cells which are contained in a three millimeter circular disc centered at the location with maximum cell density on the visual streak. It has a total of 520 cells, of which 54 are A-ON cells and 55 are A-OFF cells. The B cell counts are 134, 136 and 141 for the three angle preferences of 180°, 40° and −75° respectively. These are the means of the groups of directional sensitive cells reported in [4].

3 Two simulations using the retinal patch

In this section we detail two different yet related simulations on the model retinal patch. In the first simulation we collect data to determine the unknown motion direction of a point target assuming that the speed of the target is known. In the second simulation we collect data to estimate the speed and use this information to detect motion direction of a point target assuming that both of these parameters are unknown. In both simulations the input is a spot of light on a dark background. The patch is circular of diameter three millimeters and the size of the spot is one tenth the size of the patch.

In Simulation I, we consider a circular retinal patch (shown in Fig. 1a) and assume that a point target moves with a constant velocity through the center of the patch. The target takes 0.8 seconds to cross the patch. The simulation pool consists of motion directions between 0° (i.e. the target moves from left to right) and 358° at steps of 2°. It follows that we have 180 different directions of motion. The objective of this simulation is to study how different cell types discriminate directions of motion. Simulation I is repeated twice, once under the assumption that the B cells have a perfect knowledge of θ. In the second instance, we assume that the B cells are able to observe θ up to a random variable θ^*. Each motion direction is simulated 60 times in the first instance. For the second instance, the directions are simulated 30 times.

In Simulation II, we use twelve different angles from 0° to 330° at steps of 30°. We have the target move along each direction at nine different speeds. As the speed varies, the target takes between 0.4 seconds to 2 seconds to diametrically cross the patch. In all, we have 108 different speed/angle combinations (i.e. 108 different velocities). Each combination is simulated 60 times. In addition to these evenly spaced simulation points, we also generate intermediate test points with five intermediate speeds and 60 intermediate angles. These intermediate points are each simulated 10 times.

4 Two main tools for analysis

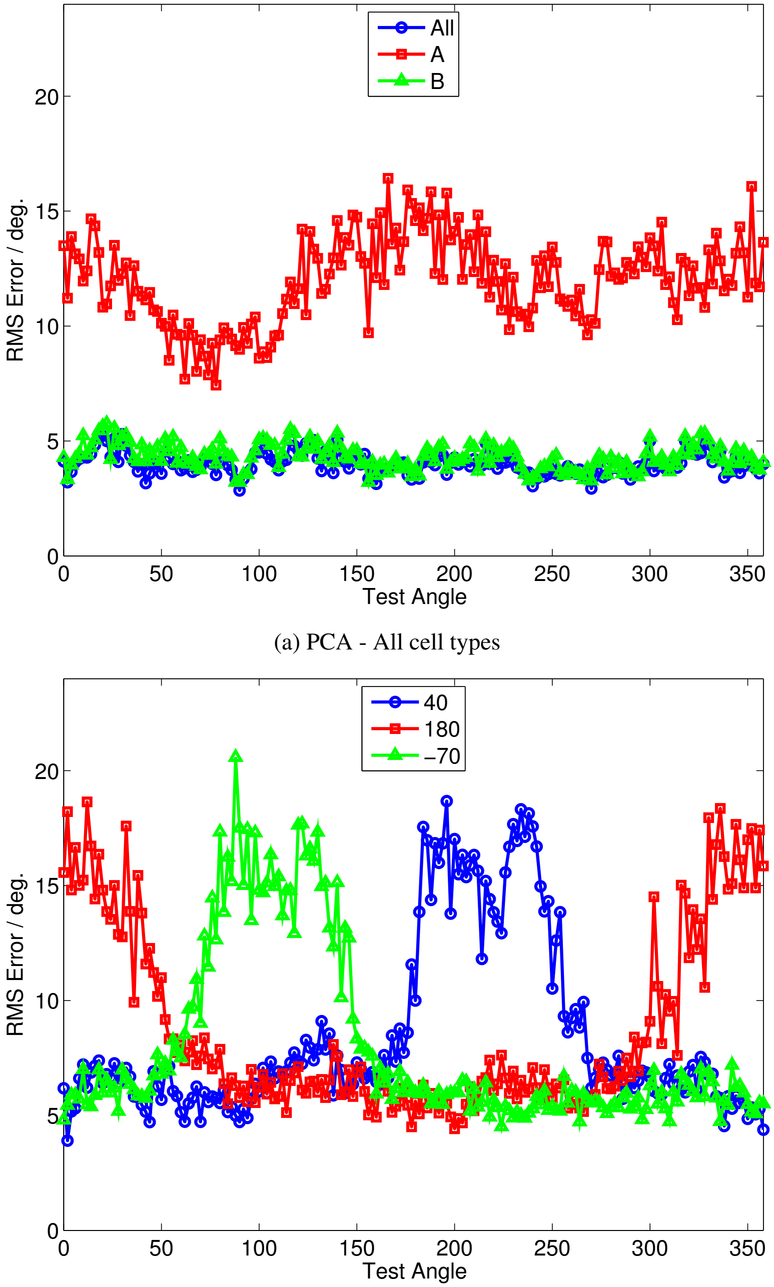
The two main tools for analysis we use are derived from the PCA [11] and the Models PPM arising from self exciting point processes [20]. In the Simulation I, the activities of the cells in the patch are low pass filtered individually. The smoothed activity functions are represented using PCA by considering the entire patch as the spatial window and over a suitable sliding time window. The spatiotemporal activity of the retinal patch is thus represented as a strand conditioned on the target direction. Maximum likelihood detection is performed (see Van Trees [22]) assuming that the strand is a Gaussian random process. The details are similar to what had been done by Du et. al. [6] on the turtle visual cortex. Alternatively, in the PPM approach, the spike activities of the cells are pooled over a subpatch and the intensity function of the pooled process, a Poisson Process, is computed. This step is repeated over a vector of subpatches. The obtained vector of intensity functions is now used to detect the target motion direction.

In the second simulation, the speed of the target is estimated from the intensity function of the pooled spike activities. The pooling process is similar to what was described for the first simulation. The speed estimation is carried out from the *half height pulse widths* of the associated intensity functions. Using the estimated speeds, the target directions are inferred as follows. The intensity functions are first computed over a vector of subpatches and are subsequently rescaled, using the estimated target speeds, to be distributed over an unit length in seconds. The target motion directions are detected from the rescaled intensity functions using PCA over each subpatch. The PCA is carried out over a moving time window and we assume that the principal component points form a Gaussian process. Over every subpatch, target detection is carried out by running a maximum likelihood detection algorithm for Gaussian processes (as performed for data in Simulation I). The final target direction is inferred using a majority vote over the subpatches (see [12, 21]).

5 Results

In addition to providing a model of the A and B cells, one of the main result of this paper is to illustrate the extent to which retinal cells are able to detect direction of target motion. The B cells out-perform the A cells in terms of their ability to discriminate motion direction, measured using Root Mean Square of the detection error. This fact is entirely obvious along the preferred direction of the directionally selective B cells. A priori, it is not clear why an ensemble of three directionally selective families of B cells would have a superior performance for targets moving along any direction. The superiority of the B cell performance over A cell, is particularly enhanced when the target speed is assumed to be unknown. In this case the speed is estimated from the retinal response data. Finally as a population, the B cells out-perform the A cells, even when B cells observe the target direction up to a large noise variance of (~ 30°).

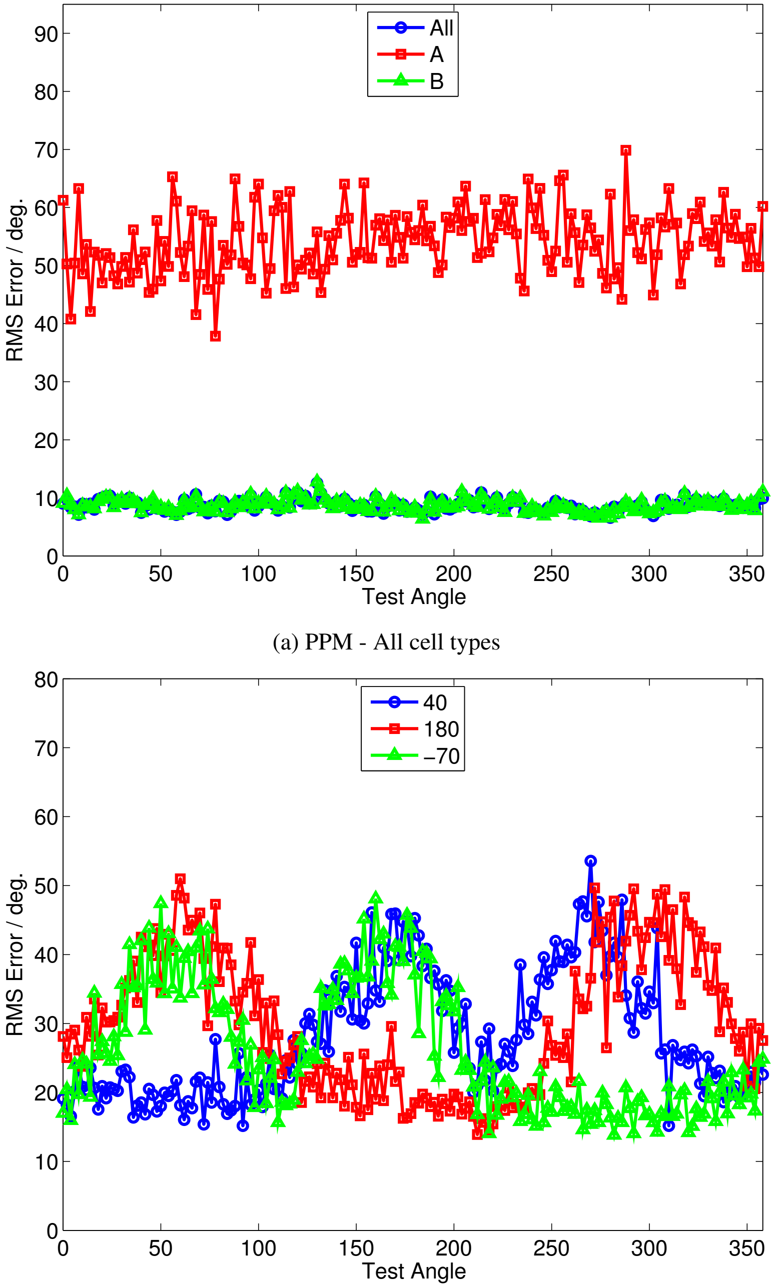
For Simulation I, it can be observed from Fig. 2b on the next page that the direction sensitive B cells detect motion directions with less error close to their preferred directions. If we consider the B cells together (shown in Fig. 2a on the next page), then the detection error is constant throughout all the motion angles. A cells, on the other hand, do show a higher level of detection error and some amount of variability with the motion direction. We suspect that this variation is purely due to the distribution of A cells in this specific patch under consideration. In Figs. 3a and 3b on page 121 we plot the Root Mean Square Error using PPM. The displayed results are obtained using a 20 *ms* window and the window starts at 400 *ms*, the mid point of the motion of the target in visual space. Fig. 3a clearly shows the effect of direction sensitive B cells. They out-perform the A cells in terms of having a lower Root Mean Square Error of detection. Also note from Figs. 2a and 3a that, when all three direction preferences of the B cells are taken together, the overall Root Mean Square Error is much lower than any single type and it remains constant over all motion directions.



(a) PCA - All cell types

(b) PCA - B cells

Figure 2: Root Mean Square Error of detection using Principal Component Analysis (PCA) for Simulation I.

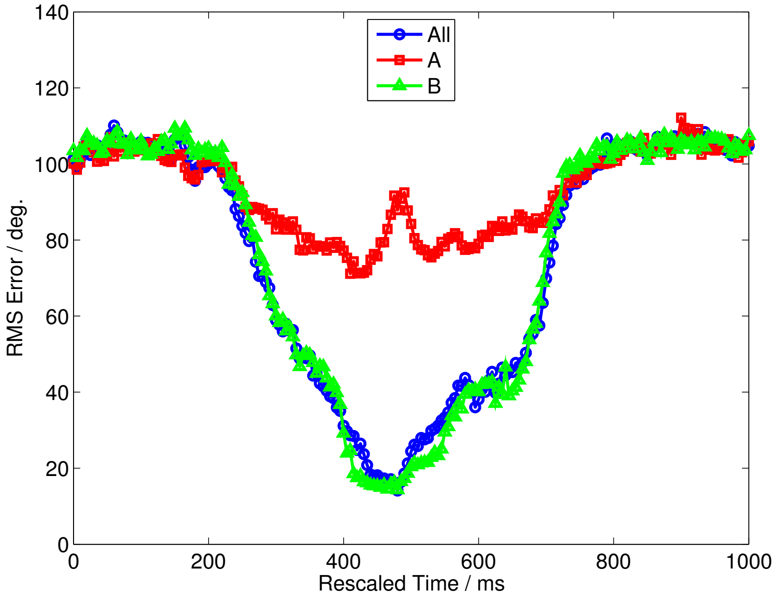


(a) PPM - All cell types

(b) PPM - B cells

Figure 3: Root Mean Square Error of detection using Poisson Process Model (PPM) for Simulation I.

We have omitted (see [7]) discussing the problem of estimating the speed and motion direction using Simulation II. When the speed is unknown and is estimated from the data, the root mean square error for motion direction is larger compared to what is observed in Simulation I. This fact has been illustrated in Figures 4 and 5 where the root mean square error has been plotted as a function of time as the target enters the patch at different speeds.

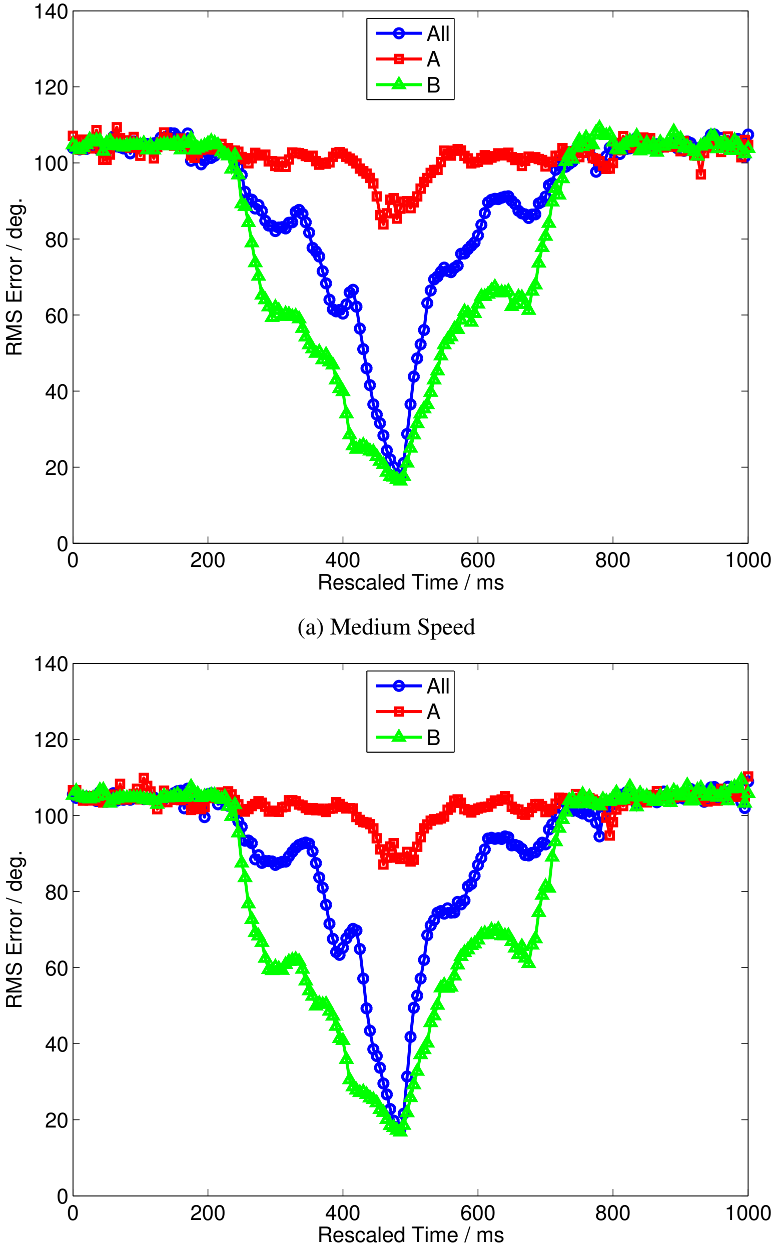


(a) High Speed

Figure 4: For Simulation II variation of the Root Mean Square Error of detection over rescaled time using all cells (blue), A cells (red) and B cells (green). Original *high speed* takes 400 *ms* to cross the patch, *medium speed* and *low speed* are shown in Figure 5.

6 Conclusion

Using Root Mean Square of the detection error as a criterion for measuring detectability, we show in this paper that – for the purpose of discriminating motion directions of targets, the direction-selective B cells are superior compared to the intensity sensitive A cells along their preferred direction, with no particular advantage along the null direction. Taken as a collection, B cells with three specific preferred directions (observed in the turtle retina) have a superior performance compared to the A cells for any target direction. The performance of a *B cell family* remains relatively unaltered under noisy conditions even when individual B cells observe target directions up to a zero mean Gaussian random variable with a large angular variance. All the above properties remain qualitatively intact when the speed of the target is uncertain,



(a) Medium Speed

(b) Low Speed

Figure 5: For Simulation II variation of the Root Mean Square Error of detection over rescaled time using all cells (blue), A cells (red) and B cells (green). Original *high speed* (Figure 4) takes 400 *ms*, *medium speed* (top) takes 1200 *ms*, *low speed* (bottom) takes 2000 *ms* to cross the patch.

although the actual values of the RMS errors rise. In this case we show that the RMS error can be decreased by using a voting algorithm that combines detection across multiple subpatches. All the claims made in this paper have been verified using two distinct decoding algorithms – the PCA and PPM.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under the grant number BIC 0523983 and ECCS 1029178. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Bibliography

[1] J. Ammermuller and H. Kolb. The organization of the turtle inner retina. I. On and off center pathways. *Journal of Comparative Neurology*, 358(1):1–34, 1995. Cited p. 115.

[2] T. I. Baker and P. S. Ulinski. Models of direction-selective and non-direction selective turtle retinal ganglion cells. *Society for Neuroscience Abstract*, 2001. Cited p. 117.

[3] H. B. Barlow and W. R. Levick. The mechanism of directionally selective units in rabbit's retina. *The Journal of Physiology*, 178:477–504, 1965. Cited p. 115.

[4] D. B. Bowling. Light responses of ganglion cells in the retina of the turtle. *The Journal of Physiology*, 299:173–196, 1980. Cited pp. 115, 117, and 118.

[5] J. R. Dearworth and A. M. Granda. Multiplied functions unify shapes of ganglion-cell receptive fields in retina of turtle. *Journal of Vision*, 2(3):204–217, 2002. Cited p. 117.

[6] X. Du, B. K. Ghosh, and P. S. Ulinski. Encoding and decoding target locations with waves in the turtle visual cortex. *IEEE Transactions in Biomedical Engineering*, 52(4):566–577, 2005. Cited p. 118.

[7] M. P. B. Ekanayake. *Decoding the Speed and Motion Direction of Moving Targets Using a Turtle Retinal Patch Model*. PhD thesis, Texas Tech University, 2011. Cited pp. 116, 117, and 122.

[8] S. Exner. *Entwurf zu einer physiologischen Erklärung der psychischen Erscheinungen, 1. Theil*. F. Deuticke, 1894. Cited p. 115.

[9] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117:500–544, 1952. Cited p. 115.

[10] D. H. Hubel. Single unit activity in striate cortex of unrestrained cats. *The Journal of Physiology*, 147:226–238, 1959. Cited p. 115.

[11] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002. Cited pp. 116 and 118.

[12] L. Lam and C. Y. Suen. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man and Cybernetics - Part A*, 27(5):553–568, 1997. Cited p. 119.

[13] E. M. Lasater and P. Witkovsky. Membrane currents of spiking cells isolated from turtle retina. *Journal of Comparative Physiology A*, 167(1):11–21, 1990. Cited p. 117.

[14] Y. Liu and E. M. Lasater. Calcium currents in turtle retinal ganglion cells. I. The properties of T- and L-type currents. *Journal of Neurophysiology*, 71(2):733–742, 1994. Cited p. 117.

[15] P. L. Marchiafava and R. Weiler. Intracellular analysis and structural correlates of the organization of inputs to ganglion cells in the retina of the turtle. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 208(1170):103–113, June 1980. Cited pp. 115 and 117.

[16] Z. Nenadic, B. K. Ghosh, and P. S. Ulinski. Modeling and estimation problems in the turtle visual cortex. *IEEE Transactions in Biomedical Engineering*, 49(8):753–762, 2002. Cited pp. 115 and 116.

[17] E. H. Peterson and P. S. Ulinski. Qunatitative studies of retinal ganglion cells in a turtle, pseudemys scripta elegans I: Number and distribution of ganglion cells. *Journal of Comparative Neurology*, 186:17–42, 1979. Cited p. 117.

[18] E. H. Peterson and P. S. Ulinski. Qunatitative studies of retinal ganglion cells in a turtle, pseudemys scripta elegans II: Size spectrum of ganglion cells and its regional variation. *Journal of Comparative Neurology*, 208:157–168, 1982. Cited pp. 115 and 117.

[19] R. D. Rodieck. *The First Steps in Seeing*. Sinauer, 1998. Cited pp. 115 and 117.

[20] D. L. Snyder. *Random Point Processes*. John Wiley & Sons, 1975. Cited pp. 116 and 118.

[21] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 4th edition, 2008. Cited p. 119.

[22] H. L. Van Trees. *Detection, Estimation and Modulation Theory, Part I*. John Wiley & Sons, 1968. Cited pp. 116 and 118.

On invariant subspaces and intertwining maps

Paul A. Fuhrmann
Department of Mathematics
Ben-Gurion University of the Negev
Beer Sheva, Israel
fuhrmannbgu@gmail.com

Abstract. The purpose of the present paper is the representation of invariant subspaces of a linear transformation as kernels and images of maps commuting with it. This extends a result of Halmos [15]. We focus also on the study of how the existence of complementary invariant subspaces is related to the invertibility of certain linear maps. This analysis connects to the concept of skew-primeness, introduced in Wolovich [20], as well as to a theorem of Roth [18].

1 Introduction

The present paper can be considered as a follow up to Fuhrmann and Helmke [14], filling in gaps left open in that paper. As in the above mentioned paper, the context in which we work is that of polynomial models, introduced in Fuhrmann [4]. There are several advantages in taking a polynomial approach. First and foremost, it is an efficient one and allows us to pass easily from the level of arithmetic of polynomial matrices to the geometric level of invariant subspaces. The polynomial model theory not only provides a characterization of the commutant of a given transformation as well as all maps intertwining two given ones, but at the same time characterizes, in terms of coprimeness of polynomial matrices, the invertibility properties of these maps. The main topic of the present paper is the representation of invariant subspaces of a linear transformation as kernels and images of maps commuting with it. This extends a result of Halmos [15]. See also Domanov [1], who presented a short proof based on elementary matrix calculations and a clever choice of coordinates, and the references therein. However, the present paper has a much broader scope, treating also the embeddability of quotient modules into the model, the relation between the invariant factors of a polynomial model and those of its submodules and quotient modules. We focus also on the study of how complementarity of invariant subspaces is related to the invertibility of linear maps. That such a connection exists is not surprising as both properties can be characterized in terms of coprimeness of polynomial matrices. This analysis connects to the concept of skew-primeness, introduced in Wolovich [20], as well as to a theorem of Roth [18]. For a geometric interpretation of skew-primeness, see Khargonekar, Georgiou and Özgüler [16]. Fuhrmann [8] contains an infinite dimensional generalization of skew-primeness. This opens up the possibility of establishing the analog of Halmos's theorem in the context of backward shift invariant subspaces. A different approach, based on dimension arguments, to Roth's theorem is given in Flanders and Wimmer [2].

The paper is structured as follows. In Section 2, we give a short, streamlined proof of the Halmos result. Section 3 is devoted to a brief description of the relevant results from the theory of polynomial models. Two related concepts of equivalence for polynomial matrices are introduced in Section 4. Section 5 is devoted to a block triangular representation based on a factorization of a nonsingular polynomial matrix. This result, the analog of representing a linear transformation, having an invariant subspace, in a block triangular form, has a simple proof yet is all important for everything that follows. In Section 6 we study the embeddability of a quotient module of a polynomial model in the polynomial model. Section 7 presents the polynomial model based analog of the Halmos result. Section 8 is devoted to the question, given a submodule of a polynomial model, to what extent is another submodule complementary to it. This leads to the study of skew-primness to which Section 9 is devoted. Finally, in Section 10, we present in terms of linear transformations some of the results obtained by polynomial methods.

Dedicated to my friend and colleague Uwe Helmke on the occasion of his 60th birthday.

2 A theorem of Halmos

Halmos [15] has shown that any invariant subspace $\mathcal{V}$ of an arbitrary complex $n \times n$ matrix A is the image of a complex matrix B, that commutes with A. Similarly, $\mathcal{V} = \operatorname{Ker} C$ for a matrix C commuting with A. Halmos uses the Hilbert space structure of $\mathbb{C}^n$, so his proof does not immediately extend to matrices over arbitrary fields. On the other hand, an essential part of his argument is based on the Frobenius theorem, stating that every square matrix is similar to its transpose $A^\top$. This result holds over any field. His presentation of the main proof idea is convoluted, due to an awkward notation and misleading comments. On the other hand, if one deletes all the unnecessary detours made by Halmos, i.e., using adjoints of complex matrices, allowing matrix multiplication on the right and not only on the left and avoiding basis descriptions, the proof condenses to an extremely short argument that is presented below. The proof holds for an arbitrary field $\mathbb{F}$ and is taken, verbatim, from Fuhrmann and Helmke [14].

Theorem 1. *Let A denote an arbitrary $n \times n$ matrix over a field $\mathbb{F}$ and $\mathcal{V}$ denote an invariant subspace of A. Then there exist matrices B, C, both commuting with A, such that* $\operatorname{Im} B = \mathcal{V}$ *and* $\operatorname{Ker} C = \mathcal{V}$.

Proof. Let $\mathcal{V}$ be a subspace invariant under A and let X be a basis matrix for $\mathcal{V}$. By invariance, there exists a matrix Λ for which

$$AX = X\Lambda, \tag{1}$$

and

$$\mathcal{V} = \operatorname{Im} X. \tag{2}$$

By a theorem of Frobenius [3], see also Fuhrmann [7], every matrix $A \in \mathbb{F}^{n \times n}$ is similar to its transpose $A^\top$. Let S be such a similarity matrix, i.e., we have $S^{-1}A^\top S = A$. Analogously, there exists a matrix $T \in \mathbb{F}^{p \times p}$ for which $T\Lambda^\top T^{-1} = \Lambda$. Substituting into (1), we get $S^{-1}A^\top SX = XT\Lambda^\top T^{-1}$, or $A^\top(SXT) = (SXT)\Lambda^\top$. Setting $Y = SXT$, we have

$$A^\top Y = Y\Lambda^\top. \tag{3}$$

We define now $B = XY^\top$ and compute

$$AB = AXY^\top = X\Lambda Y^\top \quad \text{and} \quad BA = XY^\top A = X\Lambda Y^\top,$$

i.e., we have $AB = BA$. Now we note that both X and Y have full column rank. In particular, $Y^\top$ is surjective which implies

$$\operatorname{Im} B = \operatorname{Im} X. \tag{4}$$

Similarly, there exists a full row rank matrix Z for which we have the kernel representation

$$\mathcal{V} = \operatorname{Ker} Z. \tag{5}$$

This shows the existence of a matrix L for which

$$ZA = LZ. \tag{6}$$

Applying Frobenius' theorem once again, there exists a nonsingular matrix U for which $L = U^{-1}L^\top U$. Substituting in (6), thus $ZS^{-1}A^\top S = U^{-1}L^\top UZ$, or $UZS^{-1}A^\top = L^\top UZS^{-1}$. Setting $W = UZS^{-1}$, we have

$$WA^\top = L^\top W. \tag{7}$$

Defining $C = W^\top Z$ and noting that $W^\top$ is injective. We conclude, that

$$\mathcal{V} = \operatorname{Ker} C. \tag{8}$$

To show that C commutes with A, we note that

$$AC = AW^\top Z = W^\top LZ \quad \text{and} \quad CA = W^\top ZA = W^\top LZ,$$

i.e., we have $AC = CA$. □

3 Preliminaries

We begin by giving a brief review of the basic results on polynomial and rational models that will be used in the sequel. We omit some of the proofs which can be found in various papers, e.g. Fuhrmann [4, 13].

3.1 Polynomial models

Polynomial models are defined as concrete representations of quotient modules of the form $\mathbb{F}[z]^m/\mathcal{M}$, where $\mathcal{M} \subset \mathbb{F}[z]^m$ is a full submodule, i.e., that $\mathbb{F}[z]^m/\mathcal{M}$ is required to be a torsion module. It can be shown that this is equivalent to a representation $\mathcal{M} = D(z)\mathbb{F}[z]^m$ with $D(z) \in \mathbb{F}[z]^{m\times m}$ nonsingular. Defining a projection map $\pi_D : \mathbb{F}[z]^m \longrightarrow \mathbb{F}[z]^m$ by

$$\pi_D f = D\pi_- D^{-1} f \qquad f \in F[z]^m, \tag{9}$$

we have the isomorphism

$$X_D = \operatorname{Im}\pi_D \simeq \mathbb{F}[z]^m / D(z)\mathbb{F}[z]^m, \tag{10}$$

which gives concrete, but non canonical, representations for the quotient module. We note that $f(z) \in X_D$ if and only if $D(z)^{-1}f$ is strictly proper. The **shift operator** $S_D : X_D \longrightarrow X_D$ is defined by

$$S_D f = \pi_D z f = z f - D(z)\xi_f, \qquad f \in X_D, \tag{11}$$

where $\xi_f = (D^{-1}f)_{-1}$.

It is known that $\lambda \in \mathbb{F}$ is an eigenvalue of S_D if and only if $\operatorname{Ker} D(\lambda) \neq 0$. In fact, we have

$$\operatorname{Ker}(\lambda I - S_D) = \{\frac{D(z)\xi}{z-\lambda} | \xi \in \operatorname{Ker} D(\lambda)\} \tag{12}$$

The polynomial model X_D becomes an $\mathbb{F}[z]$-module by using the S_D-induced module structure, i.e.,

$$p \cdot f = \pi_D(pf), \qquad p \in \mathbb{F}[z], f \in X_D. \tag{13}$$

The following proposition allows us to translate results obtained in the context of polynomial models to statements about matrices or linear transformations.

Proposition 2. *Let $A \in \mathbb{F}^{n\times n}$. Then we have the isomorphism*

$$S_{zI-A} \simeq A. \tag{14}$$

Proof. Clearly $f(z) = \sum_{i=0}^k f_i z^i \in X_{zI-A}$ if and only if $(zI-A)^{-1}f(z)$ is strictly proper. Using the identity $z^i I - A^i = (zI-A)\sum_{j=0}^{i-1} z^j A^{i-j-1}$, we write

$$f(z) = \sum_{i=0}^k A^i f_i + \sum_{i=0}^k (z^i I - A^i) f_i = \xi + (zI-A)g(z),$$

where $\xi = \sum_{i=0}^k A^i f_i$. This shows that $f(z) \in X_{zI-A}$ if and only if $f(z)$ is a constant polynomial., i.e., $X_{zI-A} = \mathbb{F}^n$. Next, we compute for $\xi \in \mathbb{F}^n$

$$S_{zI-A}\xi = \pi_{zI-A} z\xi = \pi_{zI-A}(zI-A+A)\xi = \pi_{zI-A}A\xi = A\xi,$$

which proves the isomorphism. □

3.2 Lattice of invariant subspaces

The next theorem explores the close relationship between factorizations of polynomial matrices and invariant subspaces, thereby providing a link between geometry and arithmetic. It is one of the principal results which makes the study of polynomial models so useful.

Theorem 3. *Let $D(z) \in \mathbb{F}[z]^{m\times m}$ be nonsingular. Then*

1. *A subset $\mathcal{V} \subset X_D$ is a submodule, or equivalently an S_D-invariant subspace, if and only if $\mathcal{V} = D_1 X_{D_2}$ for some factorization*
$$D(z) = D_1(z)D_2(z), \tag{15}$$
with $D_i(z) \in \mathbb{F}[z]^{m\times m}$ also nonsingular.

2. *We have*
$$S_D | D_1 X_{D_2} = D_1 S_{D_2} D_1^{-1}. \tag{16}$$

3. *We have the following isomorphism*
$$X_{D_1} \simeq X_{D_1 D_2} / D_1 X_{D_2}. \tag{17}$$

The factorization (15) can always be changed into

$$D(z) = (D_1(z)U(z)^{-1})(U(z)D_2(z)),$$

where $U(z)$ is unimodular. We use this freedom to insure that $D_2(z)^{-1}$ is proper. The simplest way to do this is to reduce $D_2(z)$ to row proper form. Throughout this paper, we will assume that.

Proposition 4. *Let $D(z) \in \mathbb{F}[z]^{m\times m}$ be nonsingular. Given the factorization (15) and under the assumption that $D_2(z)^{-1}$ is proper, we have the direct sum decomposition*

$$X_{D_1 D_2} = X_{D_1} \oplus D_1 X_{D_2}. \tag{18}$$

Proof. Clearly, $D_1 X_{D_2} \subset X_{D_1 D_2}$. For $f(z) \in X_{D_1}$, we compute

$$(D_1(z)D_2(z))^{-1} f(z) = D_2(z)^{-1}(D_1(z)^{-1} f(z)).$$

Since $D_1(z)^{-1} f(z)$ is strictly proper and $D_2(z)^{-1}$ is proper, it follows that the product $(D_1(z)D_2(z))^{-1} f(z)$ is strictly proper and hence we have the inclusions $X_{D_1} \subset X_{D_1 D_2}$ and $X_{D_1} + D_1 X_{D_2} \subset X_{D_1 D_2}$. Assume now $f(z) \in X_{D_1} \cap D_1 X_{D_2}$, then $D_1(z)^{-1} f(z)$ is both polynomial and strictly proper, so necessarily it is zero and we have $X_{D_1} \oplus D_1 X_{D_2} \subset X_{D_1 D_2}$. To prove the inverse inclusion, we assume $f(z) \in X_{D_1 D_2}$. Defining $f_1 = \pi_{D_1} f$, and, since $f(z) - f_1(z) \in \operatorname{Ker} \pi_{D_1} = D_1 \mathbb{F}[z]^m$, writing $f(z) = f_1(z) + D_1(z) f_2(z)$, necessarily $f_2(z) \in X_{D_2}$ and hence (18) follows. □

Theorem 5. *Let $\mathcal{V}_i$, $i = 1, \ldots, s$ be submodules of X_D, i.e. S_D-invariant subspaces, having the representations $\mathcal{V}_i = E_i X_{F_i}$, that correspond to the factorizations*

$$D(z) = E_i(z) F_i(z).$$

Then the following statements are true.

1. *$\mathcal{V}_1 \subset \mathcal{V}_2$ if and only if $E_1(z) = E_2(z)R(z)$, i.e., if and only if $E_2(z)$ is a left factor of $E_1(z)$.*

2. *$\cap_{i=1}^s \mathcal{V}_i$ has the representation $E_\nu X_{F_\nu}$ with $E_\nu(z)$ the l.c.r.m. of the $E_i(z)$ and $F_\nu(z)$ the g.c.r.d. of the $F_i(z)$.*

3. *$\mathcal{V}_1 + \cdots + \mathcal{V}_s$ has the representation $E_\mu X_{F_\mu}$ with $E_\mu(z)$ the g.c.l.d. of the $E_i(z)$ and $F_\mu(z)$ the l.c.l.m. of all the $F_i(z)$.*

Proof.

1. Assume $E_1(z) = E_2(z)R(z)$. Clearly $D(z) = E_1(z)F_1(z) = E_2(z)R(z)F_1(z) = E_2(z)F_2(z)$, so $R(z)F_1(z) = F_2(z)$. Hence $E_1X_{F_1} = E_2RX_{F_1} \subset E_2X_{RF_1} = E_2X_{F_2}$.

   Conversely, assume
   $$E_1X_{F_1} \subset E_2X_{F_2} \tag{19}$$
   Both $E_1X_{F_1} + D\mathbb{F}[z]^m$ and $E_2X_{F_2} + D\mathbb{F}[z]^m$ are submodules of $\mathbb{F}[z]^m$. We compute,
   $$E_iX_{F_i} + D\mathbb{F}[z]^m = E_iX_{F_i} + E_iF_i\mathbb{F}[z]^m = E_i[X_{F_i} + F_i\mathbb{F}[z]^m] = E_i\mathbb{F}[z]^m$$
   So (19) implies the inclusion
   $$E_1\mathbb{F}[z]^m \subset E_2\mathbb{F}[z]^m$$
   From this the factorization $E_1(z) = E_2(z)R(z)$ follows.

2. The intersection of submodules is a submodule. Hence, letting $\mathcal{V}_\nu = \cap_{i=1}^s \mathcal{V}_i$, we have $\mathcal{V}_\nu = E_\nu X_{F_\nu}$ for some factorization $D(z) = E_\nu(z)F_\nu(z)$. Since $\mathcal{V}_\nu \subset \mathcal{V}_i$, for $i = 1, \ldots, s$, we have $E_\nu X_{F_\nu} \subset E_iX_{F_i}$ and hence the factorizations
   $$E_\nu(z) = E_i(z)R_i(z). \tag{20}$$
   These imply
   $$F_i(z) = R_i(z)F_\nu(z). \tag{21}$$
   This shows that $E_\nu(z)$ is a common right multiple of the $E_i(z)$ and $F_\nu(z)$ a common right divisor of the $F_i(z)$. Clearly, $D(z)$ is a common left multiple of all the $E_i(z)$ and hence the least common right multiple of all the $E_i(z)$ must be a left factor of $D(z)$. So, let $E(z)$ be any common right multiple of the $E_i(z)$ which is also a left factor of $D(z)$. Thus $E(z) = E_i(z)Q_i(z)$ and $D(z) = E(z)F(z)$. Clearly
   $$EX_F = E_iQ_iX_F \subset E_iX_{Q_iF} = E_iX_{F_i}$$
   so
   $$EX_F \subset \cap_{i=1}^s E_iX_{F_i} = E_\nu X_{F_\nu}$$
   and this implies $E(z) = E_\nu(z)G(z)$. The last equality shows that $E_\nu(z)$ is the l.c.r.m. of the $E_i(z)$.

Similarly, let $F(z)$ be any other common right divisor of the $F_i(z)$. Thus, there exist factorizations $D(z) = E(z)F(z) = E_\nu(z)F_\nu(z)$ and clearly $E_\nu X_{F_\nu} \supset EX_F$. In particular, $F(z)$ is a right divisor of $F_\nu(z)$ which shows that $F_\nu(z)$ is the greatest common right divisor of the $F_i(z)$.

3. Let $\mathcal{V}_\mu = \mathcal{V}_1 + \cdots + \mathcal{V}_s = E_\mu X_{F_\mu}$. Since $\mathcal{V}_i \subset \mathcal{V}_\mu$ we have $E_i(z) = E_\mu(z)S_i(z)$ for all i. This means that $E_\mu(z)$ is a common left divisor of all $E_i(z)$. Let $E(z)$ be any other common left divisor of the $E_i(z)$. Then

$$E_i(z) = E(z)R_i(z) \tag{22}$$

and

$$E_i(z)F_i(z) = E(z)R_i(z)F_i(z) = E(z)F(z)$$

with

$$F(z) = R_i(z)F_i(z), \qquad 1 \le i \le s,$$

Now equalities (22) imply $E_iX_{F_i} \subset EX_F$ and hence

$$E_\mu X_{F_\mu} = \mathcal{V}_1 + \cdots + \mathcal{V}_s \subset EX_F$$

But this implies, by part (1), that $E_\mu(z) = E(z)G(z)$ and hence that $E_\mu(z)$ is a g.c.l.d. of the $E_i(z)$. Similarly, we can show that $F_\mu(z)$ the l.c.l.m. of all the $F_i(z)$. □

Corollary 6. *Given the factorizations $D(z) = E_i(z)F_i(z)$, for $i = 1,\ldots,s$, then*

1. *We have*

$$X_D = E_1X_{F_1} + \cdots + E_sX_{F_s}$$

if and only if the $E_i(z)$ are left coprime.

2. *We have $\cap_{i=1}^s E_iX_{F_i} = 0$ if and only if the $F_i(z)$ are right coprime.*

3. *Given the factorizations $D(z) = E_1(z)F_1(z) = E_2(z)F_2(z)$ of a nonsingular $D(z) \in F[z]^{m\times m}$, then we have the direct sum representation*

$$X_D = E_1X_{F_1} \oplus E_2X_{F_2} \tag{23}$$

if and only if $F_1(z), F_2(z)$ are right coprime and $E_1(z), E_2(z)$ are left coprime.

Proof.

1. Follows from the previous theorem.

2. Follows from the previous theorem.

3. The left coprimeness condition is equivalent to $X_D = E_1X_{F_1} + E_2X_{F_2}$, whereas the right coprimeness condition is equivalent to $E_1X_{F_1} \cap E_2X_{F_2} = \{0\}$. □

3.3 $\mathbb{F}[z]$-Homomorphisms

Polynomial models have two basic structures, that of an $\mathbb{F}$-vector space and that of an $\mathbb{F}[z]$-module. The $\mathbb{F}[z]$-homomorphisms are of particular importance in interpolation and the following theorem gives their characterization.

Theorem 7. *Let $D_1(z) \in \mathbb{F}[z]^{m\times m}$ and $D_2(z) \in \mathbb{F}[z]^{p\times p}$ be nonsingular. An $\mathbb{F}$–linear map $Z : X_{D_1} \longrightarrow X_{D_2}$ is an $\mathbb{F}[z]$-homomorphism, or a map intertwining S_{D_1} and S_{D_2}, i.e., it satisfies*

$$S_{D_2}Z = ZS_{D_1} \tag{24}$$

if and only if there exist $N_1(z), N_2(z) \in \mathbb{F}[z]^{p\times m}$ such that

$$N_2(z)D_1(z) = D_2(z)N_1(z) \tag{25}$$

and

$$Zf = \pi_{D_2}N_2 f. \tag{26}$$

Theorem 8. *Let $Z : X_{D_1} \longrightarrow X_{D_2}$ be the $\mathbb{F}[z]$-module homomorphism defined by*

$$Zf = \pi_{D_2}N_2 f. \tag{27}$$

with

$$N_2(z)D_1(z) = D_2(z)N_1(z) \tag{28}$$

holding. Then

1. $\operatorname{Ker} Z = E_1X_{F_1}$*, where $D_1(z) = E_1(z)F_1(z)$ and $F_1(z)$ is a g.c.r.d. of $D_1(z)$ and $N_1(z)$.*
2. $\operatorname{Im} Z = E_2X_{F_2}$*, where $D_2(z) = E_2(z)F_2(z)$ and $E_2(z)$ is a g.c.l.d. of $D_2(z)$ and $N_2(z)$.*
3. *Z is invertible if and only if $D_1(z)$ and $N_1(z)$ are right coprime and $D_2(z)$ and $N_2(z)$ are left coprime.*
4. *$D_1(z)$ and $N_1(z)$ are right coprime and $D_2(z)$ and $N_2(z)$ are left coprime if and only if there exist polynomial matrices $X_1(z), Y_1(z), X_2(z), Y_2(z)$ for which the following doubly coprime factorization holds*

$$\begin{gathered}\begin{bmatrix} Y_2(z) & -X_2(z) \\ -N_2(z) & D_2(z)\end{bmatrix}\begin{bmatrix} D_1(z) & X_1(z) \\ N_1(z) & Y_1(z)\end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I\end{bmatrix}, \\ \begin{bmatrix} D_1(z) & X_1(z) \\ N_1(z) & Y_1(z)\end{bmatrix}\begin{bmatrix} Y_2(z) & -X_2(z) \\ -N_2(z) & D_2(z)\end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I\end{bmatrix}.\end{gathered} \tag{29}$$

5. *In terms of the doubly coprime factorizations (29), $Z^{-1} : X_{D_2} \longrightarrow X_{D_1}$ is given by*

$$Z^{-1}g = -\pi_{D_1}X_1 g, \qquad g \in X_{D_2}. \tag{30}$$

4 On equivalence of polynomial matrices

Definition 9. Let $D(z)$ and $E(z)$ be polynomial matrices in $\mathbb{F}[z]^{m\times n}$. We say $D(z)$ and $E(z)$ are **unimodularly equivalent** if there exist unimodular polynomial matrices $U(z)\in\mathbb{F}[z]^{m\times m}$ and $V(z)\in\mathbb{F}[z]^{n\times n}$ such that $D(z)=U(z)E(z)V(z)$.

Clearly, unimodular equivalence is a bona fide equivalence relation, i.e., it is a reflexive symmetric and transitive relation. We proceed to show how to choose a canonical representative in each equivalence class. This is done via the invariant factor algorithm, which may be considered as a generalization of the Euclidean algorithm.

Applying the invariant factor algorithm, see Fuhrmann [9], every polynomial matrix is unimodularly equivalent to its **Smith canonical form**, i.e., to a diagonal polynomial matrix with the invariant factors $d_i(z)$ on the diagonal. We will always assume that the invariant factors are ordered so that $d_i(z)|d_{i-1}(z)$.

The characterization of module isomorphisms, given by Theorems 7 and 8, allows us to generalize the concept of unimodular equivalence to the case of nonsingular polynomial matrices of differing orders. Clearly, similarity of linear transformations is an equivalence relation, i.e., it is a reflexive, symmetric and transitive relation. The similarity relation for linear transformations induces an equivalence relation in the set of all nonsingular polynomial matrices. We formalize this by the following, based on Fuhrmann [5].

Definition 10. Let $D_1(z)$ and $D_2(z)$ be nonsingular polynomial matrices in $\mathbb{F}[z]^{m\times m}$ and $\mathbb{F}[z]^{p\times p}$ respectively. We say $D_1(z)$ and $D_2(z)$ are **coprime equivalent** if there exist polynomial matrices $N_2(z)$ and $N_1(z)$ such that

$$N_2(z)D_1(z)=D_2(z)N_1(z) \tag{31}$$

and

1. $N_2(z)$ and $D_2(z)$ are left coprime,
2. $D_1(z)$ and $N_1(z)$ are right coprime.

To justify the preceeding definition we need to establish the following.

Theorem 11. *Coprime equivalence is a bona fide equivalence relation, namely it is reflexive, symmetric and transitive.*

Proof. One can prove this result directly from the coprimeness assumptions and the use of Bezout equations. However, it follows also from the fact that similarity is an equivalence relation together with the characterization of the isomorphisms of polynomial models given in Theorems 7 and 8. □

Clearly, as a unimodular matrix is right and left coprime with any other matrix, unimodular equivalence implies coprime equivalence. Indeed, given $D(z)\in\mathbb{F}[z]^{m\times m}$, with invariant factors $d_i(z)$, $i=1,\dots,m$, we have the direct sum representation $X_D\simeq\oplus_{i=1}^m X_{d_i}$. We can use equivalence for a further reduction. Let

$$d_i(z)=\Pi_{j=1}^{n_i}p_{ij}(z)^{\nu_{ij}} \tag{32}$$

be the **primary decomposition** of the i-th invariant factor. The polynomials $p_{ij}(z)^{\nu_{ij}}$ are the **elementary divisors** of $D(z)$. The diagonal polynomial matrix having the elementary divisors on the diagonal will be called the **polynomial Jordan form**. We will use the same name even if $\Delta(z)$ has larger size and has extra units on the diagonal. In fact, defining $\pi_{ij}(z) = \Pi_{k\neq j} p_{ik}(z)^{\nu_{ik}}$, and noting that

$$d_i(z) = \pi_{ij}(z) p_{ij}(z)^{\nu_{ij}}, \tag{33}$$

we have $X_D = \oplus_{i,j} \pi_{ij} X_{p_{ij}^{\nu_{ij}}}$, and the isomorphism $X_D \simeq \oplus X_{p_{ij}^{\nu_{ij}}}$ follows. By a suitable choice of basis in the polynomial models $X_{p_{ij}^{\nu_{ij}}}$, we get the **Jordan canonical form**, see Fuhrmann [9] for the details.

Proposition 12. *Let $D(z) \in \mathbb{F}[z]^{p\times p}$. Then there exists a nonnegative integer m for which we have the unimodular equivalence*

$$\begin{bmatrix} D(z) & 0 \\ 0 & I_m \end{bmatrix} \simeq \Delta(z), \tag{34}$$

where $\Delta(z)$ is the polynomial Jordan form of $D(z)$.

Proof. First we note, using the factorization (33), that we have

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} d_i(z) = \begin{bmatrix} p_{i1} & & \\ & \ddots & \\ & & p_{in_i}(z) \end{bmatrix} \begin{bmatrix} \pi_{i1}(z)^{\nu_{i1}} \\ \vdots \\ \pi_{in_i}(z)^{\nu_{in_i}} \end{bmatrix}. \tag{35}$$

It is easily checked that the polynomial matrices

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} p_{i1} & & \\ & \ddots & \\ & & p_{in_i}(z) \end{bmatrix}$$

are left coprime and

$$d_i(z) \quad \text{and} \quad \begin{bmatrix} \pi_{i1}(z)^{\nu_{i1}} \\ \vdots \\ \pi_{in_i}(z)^{\nu_{in_i}} \end{bmatrix}$$

are right coprime. This implies that

$$d_i(z) \quad \text{and} \quad \begin{bmatrix} p_{i1} & & \\ & \ddots & \\ & & p_{in_i}(z) \end{bmatrix}$$

are equivalent. Obviously, they can be unimodularly equivalent only in the case $n_i = 1$. But unimodular equivalence can be achieved by replacing $d_i(z)$ by the $n_i \times n_i$

diagonal polynomial matrix

$$\begin{bmatrix} d_i(z) & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

The general case follows by first reducing $D(z)$, using unimodular polynomial matrices, to Smith form. □

We note that the method of enlarging a polynomial matrix by adding units on the diagonal has been efficiently employed in Rosenbrock [17].

We conclude this section with the following key result that connects equivalence and similarity and allows us to go freely from the level of polynomial matrices to that of linear transformations. This result goes back to the work of Weiersrass and Kronecker on pencils of matrices.

Theorem 13. *Let A_1 and A_2 be two linear transformations in the vector space $\mathcal{X}$. Then A_1 and A_2 are similar if and only if $zI - A_1$ and $zI - A_2$ are unimodularly equivalent.*

Proof. If A_1 and A_2 are similar there exists a nonsingular map R such that

$$RA_1 = A_2R. \tag{36}$$

This in turn implies

$$R(zI - A_1) = (zI - A_2)R, \tag{37}$$

and hence, by the invertibility of R, the equivalence of $zI - A_1$ and $zI - A_2$.

Conversely, assume $zI - A_1$ and $zI - A_2$ are unimodularly equivalent. Thus there exist unimodular polynomial matrices $U(z), V(z)$ for which $U(z)(zI - A_1) = (zI - A_2)V(z)$. This implies the similarity of S_{zI-A_1} and S_{zI-A_2}. Now, by Proposition 2, A_i is similar to S_{zI-A_i}, $i = 1,2$, so the similarity of A_1 and A_2 follows by transitivity. □

5 On block triangulization

The factorization $D(z) = D_1(z)D_2(z)$ is not convenient for the simultaneous reduction of $D_1(z), D_2(z)$ to the polynomial Jordan form. The following simple proposition allows us to bypass this difficulty.

Proposition 14. *Let $D_i(z) \in \mathbb{F}[z]^{m\times m}$ and let $D(z) = D_1(z)D_2(z)$. We assume without loss of generality that $D_2(z)^{-1}$ is proper. Then*

1. *$D_1(z)D_2(z)$ and $\begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix}$ are coprime equivalent.*

2. *Define a map* $Z : X_{D_1D_2} \longrightarrow X_{\left[\begin{smallmatrix} D_1(z) & 0 \\ -I & D_2(z)\end{smallmatrix}\right]}$ *by*

$$Zf = \pi_{\left[\begin{smallmatrix} D_1(z) & 0 \\ -I & D_2(z)\end{smallmatrix}\right]} \begin{bmatrix} I \\ 0 \end{bmatrix} f, \qquad f \in X_{D_1D_2}. \tag{38}$$

Then Z is an $\mathbb{F}[z]$*-isomorphism that preserves the following isomorphic direct sum decompositions*

$$X_{D_1D_2} = X_{D_1} \oplus D_1 X_{D_2}, \tag{39}$$

and

$$X_{\left[\begin{smallmatrix} D_1(z) & 0 \\ -I & D_2(z)\end{smallmatrix}\right]} = X_{\left[\begin{smallmatrix} D_1(z) & 0 \\ -I & I\end{smallmatrix}\right]} \oplus \begin{bmatrix} D_1(z) & 0 \\ -I & I \end{bmatrix} X_{\left[\begin{smallmatrix} I & 0 \\ 0 & D_2(z)\end{smallmatrix}\right]}. \tag{40}$$

3. *We have the following isomorphism*

$$S_{D_1D_2} \simeq \begin{bmatrix} S_{D_1} & 0 \\ \pi_{D_2}\Phi f & S_{D_2} \end{bmatrix}, \tag{41}$$

where $\Phi : X_{D_1} \longrightarrow \mathbb{F}^m$ *is defined by*

$$\Phi f = \xi_f = \pi_+ z D_1^{-1} f. \tag{42}$$

Proof.

1. Clearly, we have

$$\begin{bmatrix} I \\ 0 \end{bmatrix} D_1(z) D_2(z) = \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \begin{bmatrix} D_2(z) \\ I \end{bmatrix}. \tag{43}$$

It is easily verified that $\left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} D_1(z) & 0 \\ -I & D_2(z)\end{smallmatrix}\right]$ are left coprime and $D_1(z)D_2(z)$ and $\left[\begin{smallmatrix} D_2(z) \\ I \end{smallmatrix}\right]$ right coprime, which proves the claimed equivalence.

2. That Z, defined in (38), is an isomorphism follows from (43), the associated coprimeness conditions and Theorems 7 and 8.

We note that

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \in X_{\left[\begin{smallmatrix} D_1(z) & 0 \\ -I & I\end{smallmatrix}\right]} \quad \text{if and only if} \quad \begin{bmatrix} D_1(z) & 0 \\ -I & I \end{bmatrix}^{-1} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} D_1(z)^{-1} f_1 \\ D_1(z)^{-1} f_1 + f_2 \end{bmatrix}$$

is strictly proper. This is equivalent to $f_1 \in X_{D_1}$ and $f_2 = 0$. So, for $f \in X_{D_1}$, we compute

$$\begin{aligned} Zf &= \pi_{\left[\begin{smallmatrix} D_1(z) & 0 \\ -I & D_2(z)\end{smallmatrix}\right]} \begin{bmatrix} I \\ 0 \end{bmatrix} f \\ &= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \pi_- \begin{bmatrix} D_1(z)^{-1} & 0 \\ D_2(z)^{-1} D_1(z)^{-1} & D_2(z)^{-1} \end{bmatrix} \begin{bmatrix} f \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \begin{bmatrix} D_1(z)^{-1} f \\ D_2(z)^{-1} D_1(z)^{-1} f \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}. \end{aligned}$$

This implies

$$ZX_{D_1} = X_{\begin{bmatrix} D_1(z) & 0 \\ -I & I \end{bmatrix}}. \tag{44}$$

Similarly, let $f_2 \in X_{D_2}$. We compute

$$\begin{aligned}
ZD_1 f_2 &= \pi_{\begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix}} \begin{bmatrix} I \\ 0 \end{bmatrix} D_1(z) f_2 \\
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \pi_- \begin{bmatrix} D_1(z)^{-1} & 0 \\ D_2(z)^{-1} D_1(z)^{-1} & D_2(z)^{-1} \end{bmatrix} \begin{bmatrix} D_1(z) f_2 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \begin{bmatrix} f_2 \\ D_2(z)^{-1} f_2 \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \begin{bmatrix} 0 \\ D_2(z)^{-1} f_2 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ f_2 \end{bmatrix}.
\end{aligned}$$

This implies

$$Z(D_1 X_{D_2}) = \begin{bmatrix} D_1(z) & 0 \\ -I & I \end{bmatrix} X_{\begin{bmatrix} I & 0 \\ 0 & D_2 \end{bmatrix}}. \tag{45}$$

3. Recall that $S_{D_1} f_1 = z f_1 - D_1(z) \xi_{f_1}$, hence also

$$D_1(z)^{-1}(z f_1) = D_1(z)^{-1}(S_{D_1} f_1) + \xi_{f_1}.$$

For $f_1 \in X_{D_1}$, we compute

$$\begin{aligned}
S_{\begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix}} \begin{bmatrix} f_1 \\ 0 \end{bmatrix} &= \pi_{\begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix}} \begin{bmatrix} z f_1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \pi_- \begin{bmatrix} D_1(z)^{-1} & 0 \\ D_2(z)^{-1} D_1(z)^{-1} & D_2(z)^{-1} \end{bmatrix} \begin{bmatrix} z f_1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \pi_- \begin{bmatrix} D_1(z)^{-1} z f_1 \\ D_2(z)^{-1} D_1(z)^{-1} (z f_1) \end{bmatrix} \\
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \pi_- \begin{bmatrix} D_1(z)^{-1} z f_1 \\ D_2(z)^{-1} D_1(z)^{-1} (S_{D_1} f_1) + D_2(z)^{-1} \xi_{f_1} \end{bmatrix} \\
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \pi_- \begin{bmatrix} D_1(z)^{-1} (S_{D_1} f_1 + D_1 \xi_{f_1}) \\ D_2(z)^{-1} D_1(z)^{-1} (S_{D_1} f_1) + D_2(z)^{-1} \xi_{f_1} \end{bmatrix} \\
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \begin{bmatrix} D_1(z)^{-1} (S_{D_1} f_1) \\ D_2(z)^{-1} D_1(z)^{-1} (S_{D_1} f_1) + D_2(z)^{-1} \xi_{f_1} \end{bmatrix} \\
&= \begin{bmatrix} S_{D_1} f_1 \\ \pi_{D_2} \xi_{f_1} \end{bmatrix} = \begin{bmatrix} S_{D_1} f_1 \\ \pi_{D_2} \Phi f_1 \end{bmatrix}.
\end{aligned}$$

Finally, for $f_2 \in X_{D_2}$, we compute

$$
\begin{aligned}
S_{\begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix}} \begin{bmatrix} 0 \\ f_2 \end{bmatrix}
&= \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \pi_- \begin{bmatrix} D_1(z)^{-1} & 0 \\ D_2(z)^{-1} D_1(z)^{-1} & D_2(z)^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ z f_2 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ \pi_{D_2} z f_2 \end{bmatrix} = \begin{bmatrix} 0 \\ S_{D_2} f_2 \end{bmatrix}.
\end{aligned}
$$

Combining the two computations, (41) follows. □

Proposition 15. *Given nonsingular $D_1(z), D_2(z) \in \mathbb{F}[z]^{m \times m}$, let $Z : X_{D_1} \longrightarrow X_{D_1 D_2}$ be an injective $\mathbb{F}[z]$-homomorphism given by*

$$Zg = \pi_{D_1 D_2} X g, \qquad g(z) \in X_{D_1} \tag{46}$$

where $X(z), Y(z) \in \mathbb{F}[z]^{m \times m}$ satisfy

$$X(z) D_1(z) = (D_1(z) D_2(z)) Y(z). \tag{47}$$

Then

1. *$\overline{Z} : X_{D_1 D_2} \longrightarrow X_{D_1 D_2}$ given by*

$$\overline{Z} f = Z \pi_{D_1} f, \qquad f(z) \in X_{D_1 D_2} \tag{48}$$

is an $\mathbb{F}[z]$-homomorphism with

$$\mathrm{Ker}\, \overline{Z} = D_1 X_{D_2}. \tag{49}$$

2. *The projection $\pi_{D_1} : X_{D_1 D_2} \longrightarrow X_{D_1}$ is a surjective $\mathbb{F}[z]$-homomorphism.*

Proof.

1. From (47) we have

$$X(z)(D_1(z) D_2(z)) = (D_1(z) D_2(z))(Y(z) D_2(z)). \tag{50}$$

Since Z is assumed injective, it follows from Theorem 7 and (50) that the g.c.r.d. of $D_1(z), Y(z)$ is I. This implies that the g.c.r.d. of $(D_1(z) D_2(z))$ and $(Y(z) D_2(z))$ is $D_2(z)$. By Theorem 8, we obtain (49).

2. Follows by applying Theorem 8 and using the left coprimeness of $I, D_1(z)$ and the factorization $I(D_1(z) D_2(z)) = D_1(z) D_2(z)$. □

6 On embedding quotient modules

Given a submodule $\mathcal{N}$ of a module $\mathcal{M}$, it is trivially embedded in $\mathcal{M}$. However, if we consider the related question of embedding the quotient module $\mathcal{M}/\mathcal{N}$ in $\mathcal{M}$, this is not always possible. A simple example is that of the polynomial ring $\mathbb{F}[z]$ as a module over itself. A nontrivial submodule is of the form $\mathcal{N} = d\mathbb{F}[z]$, with $d(z) \in \mathbb{F}[z]$ of positive degree. Clearly, the quotient module $\mathbb{F}[z]/d\mathbb{F}[z]$ is a torsion module, hence cannot be embedded in $\mathbb{F}[z]$ which is torsion-free.

Proposition 15 immediately raises the question of whether there exists an injective $\mathbb{F}[z]$-homomorphism $Z : X_{D_1} \longrightarrow X_{D_1D_2}$. Since we have the isomorphism $X_{D_1} \simeq X_{D_1D_2}/D_1X_{D_2}$, this is equivalent to the embeddability of the quotient module $X_{D_1D_2}/D_1X_{D_2}$ into $X_{D_1D_2}$.

Theorem 16. *Given nonsingular $D_i(z) \in \mathbb{F}[z]^{m\times m}$, $i = 1,2$ and let $\Delta_i(z)$, $i = 1,2$, be the polynomial Jordan form of $D_i(z)$ Then*

1. *There exists a polynomial matrix $\Delta_3(z)$ for which we have the following equivalence:*
$$\begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \simeq \begin{bmatrix} \Delta_1(z) & 0 \\ \Delta_3(z) & \Delta_2(z) \end{bmatrix}.$$
2. *The elementary divisors of $D_i(z)$, $i = 1,2$, divide the corresponding elementary divisors of $D_1(z)D_2(z)$.*

Proof.

1. Appying Proposition 14, we have the following series of coprime equivalences.
$$\begin{aligned} D_1(z)D_2(z) &\simeq \begin{bmatrix} D_1(z) & 0 \\ -I & D_2(z) \end{bmatrix} \simeq \left[\begin{array}{cc|cc} D_1(z) & 0 & 0 & 0 \\ -I & D_2(z) & 0 & 0 \\ \hline 0 & 0 & I_1 & 0 \\ 0 & 0 & 0 & I_2 \end{array}\right] \\ &\simeq \left[\begin{array}{cc|cc} D_1(z) & 0 & 0 & 0 \\ 0 & I_1 & 0 & 0 \\ \hline -I & 0 & D_2(z) & 0 \\ 0 & 0 & 0 & I_2 \end{array}\right] \end{aligned} \tag{51}$$
Let $U_i(z), V_i(z)$ be unimodular polynomial matrices for which
$$U_i(z) \begin{bmatrix} D_i(z) & 0 \\ 0 & I_i \end{bmatrix} V_i(z) = \Delta_i(z), \qquad i = 1,2, \tag{52}$$
where $\Delta_i(z)$ is the polynomial Jordan form of $D_i(z)$. In turn, this implies
$$\begin{aligned} &\begin{bmatrix} U_1(z) & 0 \\ 0 & U_2(z) \end{bmatrix} \left[\begin{array}{cc|cc} D_1(z) & 0 & 0 & 0 \\ 0 & I_1 & 0 & 0 \\ \hline -I & 0 & D_2(z) & 0 \\ 0 & 0 & 0 & I_2 \end{array}\right] \begin{bmatrix} V_1(z) & 0 \\ 0 & V_2(z) \end{bmatrix} \\ &\qquad\qquad \simeq \begin{bmatrix} \Delta_1(z) & 0 \\ \Delta_3(z) & \Delta_2(z) \end{bmatrix}. \end{aligned} \tag{53}$$

Here $\Delta_3(z) = U_2(z)\begin{bmatrix} -I & 0 \\ 0 & 0 \end{bmatrix} V_1(z)$.

2. Assume now that $\Delta_1(z) = \operatorname{diag} \pi_i(z)^{\nu_i}$ and $\Delta_2(z) = \operatorname{diag} \rho_j(z)^{\mu_j}$. Consider the elementary block $\begin{bmatrix} \pi_i^{\nu_i}(z) & 0 \\ f_{ij}(z) & \rho_j^{\mu_j}(z) \end{bmatrix}$, where $\Delta_3(z) = \begin{bmatrix} f_{11}(z) & f_{12}(z) \\ f_{21}(z) & f_{22}(z) \end{bmatrix}$. If $\pi_i(z), \rho_j(z)$ are coprime, so are $\pi_i(z)^{\nu_i}$ and $\rho_j(z)^{\mu_j}$. In this case the Bezout equation $a(z)\pi_i(z)^{\nu_i} + b(z)\rho_j(z)^{\mu_j} = f_{ij}$ is polynomially solvable and, by applying appropriate elementary row and column operations, we have the equivalence

$$\begin{bmatrix} \pi_i^{\nu_i}(z) & 0 \\ f_{ij}(z) & \rho_j^{\mu_j}(z) \end{bmatrix} \simeq \begin{bmatrix} \pi_i^{\nu_i}(z) & 0 \\ 0 & \rho_j^{\mu_j}(z) \end{bmatrix}.$$

If, on the other hand, $\pi_i(z), \rho_j(z)$ are not coprime, we must have $\pi_i(z) = \rho_j(z) = \pi(z)$ and we factor $f_{ij}(z) = \pi(z)^{\alpha} g_{ij}(z)$ with $g_{ij}(z), \pi(z)$ coprime. The elementary divisors of $\begin{bmatrix} \pi^{\nu_i}(z) & 0 \\ \pi(z)^{\alpha} g_{ij}(z) & \pi^{\mu_j}(z) \end{bmatrix}$ are

$$\begin{cases} \pi(z)^{\nu_i}, \pi(z)^{\mu_j} & \text{if} \quad \alpha \geq \min(\nu_i, \mu_j) \\ \pi(z)^{\nu_i + \mu_j - \alpha}, \pi(z)^{\alpha} & \text{if} \quad \alpha < \min(\nu_i, \mu_j). \end{cases}$$

Clearly, we have the inequalities $\nu_i + \mu_j - \alpha \geq \nu_i$ as well as $\nu_i + \mu_j - \alpha \geq \mu_j$. Repeating the process for all elements of $\Delta_3(z)$, the result follows. □

We note that the method of proof that uses elementary divisors and the polynomial Sylvester equation follows that used in Roth [18].

Corollary 17. *Given a nonsingular $D(z) \in \mathbb{F}[z]^{m \times m}$. Then*

1. *The nontrivial elementary divisors of a submodule of X_D divide the corresponding elementary divisors of $D(z)$.*

2. *The nontrivial elementary divisors of a quotient module of X_D divide the corresponding elementary divisors of $D(z)$.*

Proof.

1. By Theorem 3, a submodule $\mathcal{V} \subset X_D$ is of the form $\mathcal{V} = D_1 X_{D_2}$ corresponding to the factorization (15). The claimed result follows by applying Theorem 16.

2. Follows from the isomorphism $X_{D_1 D_2}/D_1 X_{D_2} \simeq X_{D_1}$ and Theorem 16. □

Example 18. Assume $\alpha, \beta, \gamma \in \mathbb{F}$ are distinct. Let $D_1(z) = (z-\alpha)(z-\beta)$, $D_2(z) = (z-\alpha)(z-\gamma)$, and their respective polynomial Jordan forms $\Delta_1(z) = \begin{bmatrix} z-\alpha & 0 \\ 0 & z-\beta \end{bmatrix}$, $\Delta_2(z) = \begin{bmatrix} z-\alpha & 0 \\ 0 & z-\gamma \end{bmatrix}$. Clearly, since $D_1(z)D_2(z) = (z-\alpha)^2(z-\beta)(z-\gamma)$, the elementary divisors of $D_1(z)D_2(z)$ are $(z-\alpha)^2, (z-\beta), (z-\gamma)$.

Next, we show the unimodular equivalence

$$\begin{bmatrix} (z-\alpha)(z-\beta) & 0 \\ 0 & 1 \end{bmatrix} \simeq \begin{bmatrix} z-\alpha & 0 \\ 0 & z-\beta \end{bmatrix}$$

which follows from

$$\begin{bmatrix} z-\beta & z-\alpha \\ \frac{1}{\alpha-\beta} & \frac{1}{\alpha-\beta} \end{bmatrix} \begin{bmatrix} z-\alpha & 0 \\ 0 & z-\beta \end{bmatrix} \begin{bmatrix} \frac{z-\beta}{\alpha-\beta} & -1 \\ -\frac{z-\alpha}{\alpha-\beta} & 1 \end{bmatrix} = \begin{bmatrix} (z-\alpha)(z-\beta) & 0 \\ 0 & 1 \end{bmatrix}.$$

From this we obtain

$$U_1(z) = \begin{bmatrix} \frac{1}{\alpha-\beta} & -(z-\alpha) \\ -\frac{1}{\alpha-\beta} & (z-\beta) \end{bmatrix}, \quad V_1(z) = \begin{bmatrix} 1 & 1 \\ \frac{z-\alpha}{\alpha-\beta} & \frac{z-\beta}{\alpha-\beta} \end{bmatrix}.$$

Similarly

$$U_2(z) = \begin{bmatrix} (z-\gamma) & -(z-\alpha) \\ -\frac{1}{\alpha-\gamma} & \frac{1}{\alpha-\gamma} \end{bmatrix}, \quad V_2(z) = \begin{bmatrix} \frac{z-\gamma}{\alpha-\gamma} & -1 \\ -\frac{z-\alpha}{\alpha-\gamma} & 1 \end{bmatrix}.$$

We compute

$$\begin{bmatrix} f_{11}(z) & f_{12}(z) \\ f_{21}(z) & f_{22}(z) \end{bmatrix} = U_2(z)V_1(z) = \begin{bmatrix} \frac{1}{\alpha-\gamma} - \frac{(z-\alpha)^2}{\alpha-\beta} & \frac{1}{\alpha-\gamma} - \frac{(z-\alpha)(z-\beta)}{\alpha-\beta} \\ -\frac{1}{\alpha-\gamma} + \frac{(z-\alpha)(z-\gamma)}{\alpha-\beta} & -\frac{1}{\alpha-\gamma} + \frac{(z-\beta)(z-\gamma)}{\alpha-\beta} \end{bmatrix}$$

It follows that

$$\begin{bmatrix} \Delta_1(z) & 0 \\ \Delta_3(z) & \Delta_2(z) \end{bmatrix} = \left[\begin{array}{cc|cc} (z-\alpha) & 0 & 0 & 0 \\ 0 & (z-\beta) & 0 & 0 \\ \hline \frac{1}{\alpha-\gamma} - \frac{(z-\alpha)^2}{\alpha-\beta} & \frac{1}{\alpha-\gamma} - \frac{(z-\alpha)(z-\beta)}{\alpha-\beta} & (z-\alpha) & 0 \\ -\frac{1}{\alpha-\gamma} + \frac{(z-\alpha)(z-\gamma)}{\alpha-\beta} & -\frac{1}{\alpha-\gamma} + \frac{(z-\beta)(z-\gamma)}{\alpha-\beta} & 0 & (z-\gamma) \end{array}\right]$$

After further reductions by elementary operations, we have

$$\begin{bmatrix} \Delta_1(z) & 0 \\ \Delta_3(z) & \Delta_2(z) \end{bmatrix} = \left[\begin{array}{cc|cc} (z-\alpha) & 0 & 0 & 0 \\ 0 & (z-\beta) & 0 & 0 \\ \hline 1 & 0 & (z-\alpha) & 0 \\ 0 & 0 & 0 & (z-\gamma) \end{array}\right].$$

From this we can read off the elementary divisors of $D_1(z)D_2(z)$ which, of course, are as before $(z-\alpha)^2, (z-\beta), (z-\gamma)$.

We conclude this section by answering the question raised at its beginning.

Theorem 19. *Let $D_1(z), D_2(z) \in \mathbb{F}[z]^{m\times m}$ be nonsingular. Then*

1. *There exists an injective $\mathbb{F}[z]$-homomorphism $Z : X_{D_1} \longrightarrow X_D$ if and only if the invariant factors of $D_1(z)$ divide those of $D(z)$.*

2. *There exists an $\mathbb{F}[z]$-isomorphism between X_E and a quotient module of X_D if and only if the invariant factors of $E(z)$ divide those of $D(z)$.*

3. *There exists an surjective $\mathbb{F}[z]$-homomorphism $Z : X_D \longrightarrow X_E$ if and only if the invariant factors of $E(z)$ divide those of $D(z)$.*

Proof.

1. Assume there exists an injective $\mathbb{F}[z]$-homomorphism $Z : X_{D_1} \longrightarrow X_D$. Clearly $\operatorname{Im} Z$ is a submodule of X_D, thus it has a representation $\operatorname{Im} Z = \overline{D}_2 X_{\overline{D}_1}$ that corresponds to a factorization $D(z) = \overline{D}_2(z)\overline{D}_1(z)$. Since $S_{\overline{D}_1} \simeq S_{\overline{D}_2\overline{D}_1} | \overline{D}_2 X_{\overline{D}_1}$, the isomorphism $S_{D_1} \simeq S_{\overline{D}_1}$ follows. This implies the coprime equivalence of the polynomial matrices $D_1(z)$ and $\overline{D}_1(z)$. In particular, it follows that they have the same nontrivial elementary divisors. By Theorem 16, the elementary divisors of $\overline{D}_1(z)$ divide those of $D(z)$, hence also those of $D_1(z)$. This is, of course, also equivalent to the division relation between the respective invariant factors.

   To prove the converse, we can assume without loss of generality that both polynomial matrices are in Smith form. Let $d_i(z)$ and $e_i(z)$ be the nontrivial invariant factors of $D_2(z)$ and $D(z)$ respectively. We assume $d_i(z)|d_{i-1}(z)$ and $e_i(z)|e_{i-1}(z)$ By our assumption, there exist factorizations $e_i(z) = c_i(z)d_i(z)$. We use now the isomorphisms $X_{D_2} \simeq X_{d_1} \oplus \cdots \oplus X_{d_s}$ and $X_D \simeq X_{e_1} \oplus \cdots \oplus X_{e_s}$. Clearly, we must have $s \leq q$. Defining $Z : X_{d_1} \oplus \cdots \oplus X_{d_s} \longrightarrow X_{e_1} \oplus \cdots \oplus X_{e_s}$ by

$$Z \begin{bmatrix} f_1 \\ \vdots \\ f_s \end{bmatrix} = \begin{bmatrix} c_1 f_1 \\ \vdots \\ c_s f_s \\ 0 \end{bmatrix}, \qquad f_i(z) \in X_{d_i},$$

   we have the required homomorphism.

2. Assume that such an isomorphism exists. This implies that there exists a submodule of X_D, necessarily of the form $\mathcal{V} = D_1 X_{D_2}$ for some factorization $D(z) = D_1(z)D_2(z)$, for which $X_E \simeq X_{D_1D_2}/D_1X_{D_2}$. However, having the isomorphism $X_{D_1} \simeq X_{D_1D_2}/D_1X_{D_2}$, the isomorphism $X_E \simeq X_{D_1}$ follows and hence the invariant factors of $E(z)$ and $D_1(z)$ are equal. Applying Theorem 16, the invariant factors of $E(z)$ divide those of $D(z)$.

   In order to prove the converse, we can assume, as before, that without loss of generality both polynomial matrices are in Smith form. Let $d_i(z)$ and $e_i(z)$ be the nontrivial invariant factors of $E(z)$ and $D(z)$ respectively. By our assumption, there exist factorizations $e_i(z) = d_i(z)c_i(z)$. Since we have the isomorphism $X_D/D_1X_{D_2} \simeq X_{D_1} = \oplus_{i=1}^{s} X_{e_i}/d_i X_{c_i}$, with $D_1(z) = \operatorname{diag}(d_1, \ldots, d_s)$ and $D_2(z) = \operatorname{diag}(c_1, \ldots, c_s)$. Clearly the map $Z : X_E \longrightarrow X_D/D_1X_{D_2}$ given by

$$Z \begin{bmatrix} f_1 \\ \vdots \\ f_s \\ 0 \end{bmatrix} = \begin{bmatrix} [f_1]_{d_1 X_{c_1}} \\ \vdots \\ [f_s]_{d_s X_{c_s}} \\ 0 \end{bmatrix}, \qquad f_i(z) \in X_{d_i},$$

   provides the required homomorphism.

3. The proof follows along similar lines or can be obtained from the Part 1 by duality considerations. □

7 Kernel and image representations

We have now at hand the necessary machinery to prove the analog of Halmos' result in the context of polynomial models.

Theorem 20. *Let $\mathbb{F}$ be a field, $D(z) \in \mathbb{F}[z]^{m\times m}$ be nonsingular.*

1. *A subspace $\mathcal{V} \subset X_D$ is an S_D-invariant subspace if and only if it is the kernel of a map T that commutes with S_D.*
2. *A subspace $\mathcal{V} \subset X_D$ is an S_D-invariant subspace if and only if it is the image of a map T that commutes with S_D.*

Proof.

1. The "if" part is trivial.

   To prove the "only if" part, we assume that $\mathcal{V} \subset X_D$ is an S_D-invariant subspace. By Theorem 3, we have the representation $\mathcal{V} = D_1 X_{D_2}$ for a factorization

   $$D(z) = D_1(z)D_2(z). \tag{54}$$

   By Theorem 16, the elementary divisors of $D_1(z)$ divide the corresponding elementary divisors of $D_1(z)D_2(z)$. By Theorem 19, there exists an injective $\mathbb{F}[z]$-homomorphism $X : X_{D_1} \longrightarrow X_{D_1D_2}$. Hence there exist polynomial matrices $X_1(z), Y_1(z)$ for which $X_1(z)D_1(z) = (D_1(z)D_2(z))Y_1(z)$, $D_1(z), Y_1(z)$ are right coprime and, for $f(z) \in X_{D_1}$, we have $Xf = \pi_{D_1D_2}X_1 f$. We note that by applying Theorems 7 and 8 and using the identity

   $$I \cdot (D_1(z)D_2(z)) = D_1(z) \cdot D_2(z), \tag{55}$$

   the map $\pi_{D_1} : X_{D_1D_2} \longrightarrow X_{D_1}$ is a surjective $\mathbb{F}[z]$-homomorphism with $\operatorname{Ker}\pi_{D_1}|X_{D_1D_2} = D_1X_{D_2}$. We define $T : X_{D_1D_2} \longrightarrow X_{D_1D_2}$ by

   $$Tf = X\pi_{D_1}f = \pi_{D_1D_2}X_1 f, \qquad f \in X_{D_1D_2}. \tag{56}$$

   Clearly, T, as a product of $\mathbb{F}[z]$-homomorphisms, is also one and, as Ξ is injective, it follows that $\operatorname{Ker} T = D_1X_{D_2}$.

2. As before, the "if" part is trivial.

   To prove the "only if" part, we assume that $\mathcal{V} \subset X_D$ is an S_D-invariant subspace, hence, by Theorem 3, there exists a factorization

   $$D(z) = D_1(z)D_2(z), \tag{57}$$

   for which $\mathcal{V} = D_1X_{D_2}$. By Theorem 16, the elementary divisors of $D_2(z)$ divide the corresponding elementary divisors of $D_1(z)D_2(z)$. By Theorem 19, there exists a surjective $\mathbb{F}[z]$-homomorphism $Y : X_{D_1D_2} \longrightarrow X_{D_2}$. Applying Theorems 7 and 8 and using the identity

   $$D_1(z) \cdot D_2(z) = (D_1(z)D_2(z)) \cdot I, \tag{58}$$

   the embedding map $i_{D_2} : X_{D_2} \longrightarrow X_{D_1D_2}$ defined, for $f(z) \in X_{D_2}$, by $i_{D_2}f = \pi_{D_1D_2}D_1 f = D_1 f$ is an injective $\mathbb{F}[z]$-homomorphism with $\operatorname{Im} i_{D_2} = D_1X_{D_2}$. Next, we define $T : X_{D_1D_2} \longrightarrow X_{D_1D_2}$ by $T = i_{D_2}Y$. Clearly, T is an $\mathbb{F}[z]$-homomorphism with $\operatorname{Im} T = D_1X_{D_2}$. □

Noting that we have the isomorphism $A \simeq S_{zI-A}$, it follows that Theorem 1 is a consequence of Theorem 20. We also point out that the two statements in Throrem 20 are related by duality. We refrain from elaborating on this point in order to keep the scope of the paper within reasonable bounds. However, for those interested, the relevant duality theory can be found in Fuhrmann [6, 12].

Finally, we wish to point out that an infinite dimensional variant of Halmos' theorem is central to the development of behavioral system theory, namely the kernel representation of behaviors. Since the setting there is infinite dimensional, topology needs to be taken into account. We recall that σ is the backward shift operator in $z^{-1}\mathbb{F}[[z^{-1}]]^m$, a behavior is a linear, backward shift invariant subspace and closed subspace of $z^{-1}\mathbb{F}[[z^{-1}]]^m$. Here the topology is the w^* topology. Also, we note that a continuous map intertwining the shifts is necessarily of the form $P(\sigma)$. This leads to the following.

Theorem 21 (Willems)**.** *A subset $\mathcal{B} \subset z^{-1}\mathbb{F}[[z^{-1}]]^m$ is a behavior if and only if it admits a* **kernel representation***, i.e., there exists a $p \times m$ polynomial matrix $P(z)$ for which*

$$\mathcal{B} = \mathrm{Ker}\, P(\sigma) = \{h \in z^{-1}\mathbb{F}[[z^{-1}]]^m | \pi_- Ph = P(\sigma)h = 0\}. \tag{59}$$

For a proof, see Willems [19] and Fuhrmann [10].

8 Complementarity

We extend now the results of Section 7 to the case that the basic polynomial model relates to the polynomial matrix $D(z) = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$. Since, given a unimodular polynomial matrix $U(z)$, the polynomial models X_D and X_{DU} are isomorphic, and as

$$\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} \begin{bmatrix} I & 0 \\ -K(z) & I \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) - D_2(z)K(z) & D_2(z) \end{bmatrix},$$

we will always assume, without loss of generality, that $D_2(z)^{-1}D_3(z)$ is strictly proper. This allows us to do a finer analysis of the kernel and image representations involved. We proceed to study $\mathbb{F}[z]$-homomorphisms of $X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}$. Using the identity

$$\begin{bmatrix} 0 \\ I \end{bmatrix} D_2(z) = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} \begin{bmatrix} 0 \\ I \end{bmatrix}, \tag{60}$$

we define $J_2 : X_{D_2} \longrightarrow X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}$ by

$$J_2 f = \pi_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}} \begin{bmatrix} 0 \\ I \end{bmatrix} f, \qquad f(z) \in X_{D_2}. \tag{61}$$

We note that, in view of Theorems 7 and 8 and the right coprimeness of $D_2(z), \begin{bmatrix} 0 \\ I \end{bmatrix}$, J_2 is an injective $\mathbb{F}[z]$-homomorphism. We define

$$\mathcal{V} = \mathrm{Im}\, J_2. \tag{62}$$

Proposition 22. *Let $D_i(z) \in \mathbb{F}[z]^{m\times m}$ be nonsingular.*

1. (a) *A map $T : X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}} \longrightarrow X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}$ is an $\mathbb{F}[z]$-homomorphism, i.e., it satisfies*

$$TS_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}} = S_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}} T$$

if and only if there exist polynomial matrices $X_{ij}(z), Y_{ij}(z)$ for which

$$\begin{aligned}\begin{bmatrix} X_{11}(z) & X_{12}(z) \\ X_{21}(z) & X_{22}(z) \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}& \\ = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}&\begin{bmatrix} Y_{11}(z) & Y_{12}(z) \\ Y_{21}(z) & Y_{22}(z) \end{bmatrix}.\end{aligned} \tag{63}$$

Without loss of generality, we assume that

$$\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}^{-1}\begin{bmatrix} X_{11}(z) & X_{12}(z) \\ X_{21}(z) & X_{22}(z) \end{bmatrix}$$

is strictly proper. In these terms, T is given, for $\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \in X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}$, by

$$T\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \pi_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix}. \tag{64}$$

(b) *There exists polynomial matrices $X_{11}, X_{21}, Y_{11}, Y_{21}$ such that*

$$\begin{bmatrix} X_{11}(z) & 0 \\ X_{21}(z) & 0 \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}\begin{bmatrix} Y_{11}(z) & 0 \\ Y_{21}(z) & 0 \end{bmatrix}, \tag{65}$$

with $D_1(z), \begin{bmatrix} Y_{11}(z) \\ Y_{21}(z) \end{bmatrix}$ right coprime, such that the map

$$Y : X_{\begin{bmatrix} D_1 & 0 \\ D_3 & I \end{bmatrix}} \longrightarrow X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}},$$

defined by

$$Y\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \pi_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}\begin{bmatrix} X_{11} & 0 \\ X_{21} & 0 \end{bmatrix}\begin{bmatrix} f_1 \\ f_2 \end{bmatrix}. \tag{66}$$

is an injective $\mathbb{F}[z]$-homomorphism.

(c) *The map T, defined by* (64) *satisfies*

$$\mathrm{Ker}\, T = \mathcal{V} = \mathrm{Im}\, J_2, \tag{67}$$

if and only if

$$\begin{bmatrix} X_{11}(z) \\ X_{21}(z) \end{bmatrix} D_1(z) = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}\begin{bmatrix} Y_{11}(z) \\ Y_{21}(z) \end{bmatrix}. \tag{68}$$

and $D_1(z), \begin{bmatrix} Y_{11}(z) \\ Y_{21}(z) \end{bmatrix}$ are right coprime.

(d) *The map T, defined by (64) satisfies*

$$\operatorname{Im} T = \mathcal{V} = \operatorname{Im} J_2, \tag{69}$$

if and only if

$$\begin{aligned}&\begin{bmatrix} 0 & 0 \\ X_{21}(z) & X_{22}(z) \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}\\ &\qquad = \begin{bmatrix} I & 0 \\ 0 & D_2(z) \end{bmatrix}\begin{bmatrix} 0 & 0 \\ Y_{21}(z) & Y_{11}(z) \end{bmatrix},\end{aligned} \tag{70}$$

and $D_2(z), \begin{bmatrix} X_{21}(z) & X_{22}(z) \end{bmatrix}$ are left coprime.

2. *The following intertwining relations*

$$\begin{bmatrix} X_{11}(z) & 0 \\ X_{21}(z) & 0 \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}\begin{bmatrix} Y_{11}(z) & 0 \\ Y_{21}(z) & 0 \end{bmatrix}, \tag{71}$$

$$\begin{bmatrix} X_{11}(z) & 0 \\ X_{21}(z) & 0 \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}\begin{bmatrix} Y_{11}(z) & 0 \\ Y_{21}(z) & 0 \end{bmatrix}, \tag{72}$$

$$\begin{bmatrix} X_{11}(z) \\ X_{21}(z) \end{bmatrix} D_1(z) = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}\begin{bmatrix} Y_{11}(z) \\ Y_{21}(z) \end{bmatrix} \tag{73}$$

and

$$\begin{bmatrix} X_{11}(z) & 0 \\ X_{21}(z) & I \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}\begin{bmatrix} Y_{11}(z) & 0 \\ Y_{21}(z) & I \end{bmatrix} \tag{74}$$

are all equivalent to the following pair of equations.

$$\begin{aligned} X_{11}(z)D_1(z) &= D_1(z)Y_{11}(z) \\ X_{21}(z)D_1(z) &= D_3(z)Y_{11}(z) + D_2(z)Y_{21}(z). \end{aligned} \tag{75}$$

3. *Define a map $J_1 : X_{D_1} \longrightarrow X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}$ by*

$$J_1 f = \pi_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}} \begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} f, \qquad f(z) \in X_{D_1}. \tag{76}$$

We define

$$\mathcal{W} = \operatorname{Im} J_1. \tag{77}$$

J_1 is an $\mathbb{F}[z]$-homomorphism and it is injective if and only if $\begin{bmatrix} Y_{11}(z) \\ Y_{21}(z) \end{bmatrix}, D_1(z)$ are right coprime.

4. *Define a map $P_1 : X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}} \longrightarrow X_{D_1}$ by*

$$P_1 \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \pi_{D_1} \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = f_1. \tag{78}$$

Then

(a) *P_1 is a surjective $\mathbb{F}[z]$-homomorphism, with*

$$\operatorname{Ker} P_1 = \operatorname{Im} J_2 = \mathcal{V}. \tag{79}$$

(b) *For T defined by (64), we have*

$$T = J_1 P_1. \tag{80}$$

5. *Using (74), we define a map $J : X_{\begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}} \longrightarrow X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}$ by*

$$J \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} J_1 & J_2 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \pi_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}} \begin{bmatrix} X_{11}(z) & 0 \\ X_{21}(z) & I \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}. \tag{81}$$

Then

(a) *J is an $\mathbb{F}[z]$-homomorphism.*

(b) *We have $g(z) \in \operatorname{Im} J_1 \cap \operatorname{Im} J_2$ if and only if there exists $\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \in \operatorname{Ker} J$ for which $g(z) = J_1 f_1 = -J_2 f_2$.*

(c) *$\operatorname{Ker} J = \{0\}$ if and only if $D_1(z), Y_{11}(z)$ are right coprime.*

(d) *$\operatorname{Im} J = X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}$ if and only if $D_1(z), X_{11}(z)$ are left coprime.*

(e) *J is invertible is equivalent to the right coprimeness of $D_1(z), Y_{11}(z)$ which, in turn, is equivalent to the left coprimeness of $D_1(z), X_{11}(z)$.*

6. *Using (73), we define a map $Q_1 : X_{D_1} \longrightarrow X_{D_1}$ by*

$$Q_1 f_1 = \pi_{D_1} X_{11} f_1, \qquad f_1 \in X_{D_1}. \tag{82}$$

Then

(a) *Q_1 is a $\mathbb{F}[z]$-homomorphism.*

(b) *We have*

$$Q_1 = P_1 J_1. \tag{83}$$

(c) *Q_1 is invertible if and only if $X_{11}(z), D_1(z)$ are left coprime.*

(d) *Defining*

$$\mathcal{Q} = \operatorname{Im} Q_1, \tag{84}$$

we have the isomorphisms

$$\mathcal{Q} \simeq (\mathcal{W} + \mathcal{V})/\mathcal{V}, \tag{85}$$

and

$$\mathcal{Q} \simeq \mathcal{W}/(\mathcal{W} \cap \mathcal{V}). \tag{86}$$

7. *(a) J, defined in (*81*), is invertible if and only if Q_1, defined in (*82*) is invertible.*

   *(b) Assuming that J, defined in (*81*), is invertible, then without loss of generality we can take $X_{11} = Y_{11} = I$.*

Proof.

1. (a) Follows from Theorem 7.

   (b) By Proposition 15, there exists an injective $\mathbb{F}[z]$-homomorphism

$$Y : X_{\begin{bmatrix} D_1 & 0 \\ D_3 & I \end{bmatrix}} \longrightarrow X_{\begin{bmatrix} D_1 & 0 \\ D_3 & D_2 \end{bmatrix}}.$$

   By Theorem 7, there exist polynomial matrices satisfying

$$\begin{aligned} &\begin{bmatrix} X_{11}(z) & X_{12}(z) \\ X_{21}(z) & X_{22}(z) \end{bmatrix} \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix} \\ &\qquad = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} \begin{bmatrix} Y_{11}(z) & Y_{12}(z) \\ Y_{21}(z) & Y_{22}(z) \end{bmatrix}. \end{aligned} \tag{87}$$

   Wiithout loss of generality, we assume

$$\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}^{-1} \begin{bmatrix} X_{11}(z) & X_{12}(z) \\ X_{21}(z) & X_{22}(z) \end{bmatrix}$$

   is strictly proper. In particular, $D_1^{-1}X_{11}$ and $D_2^{-1}X_{22}$ are strictly proper. From (87) we obtain the following equalities.

$$\begin{aligned} X_{11}D_1 &= D_1Y_{11} \\ X_{12} &= D_1Y_{12} \\ X_{22} &= D_3Y_{12} + D_2Y_{22} \end{aligned} \tag{88}$$

   From the second equation we obtain $D_1^{-1}X_{12} = Y_{12}$. Since the left term is strictly proper and the right term polynomial, both must be zero. Similarly, we conclude that $X_{22}(z)$ and $Y_{22}(z)$ are both zero. Using these equations, we conclude that (65) holds. Since the map Y is injective, $D_1(z), \begin{bmatrix} Y_{11}(z) \\ Y_{21}(z) \end{bmatrix}$ are necessarily right coprime.

   (c) From (63) we obtain the following equalities.

$$\begin{aligned} X_{11}D_1 + X_{12}D_3 &= D_1Y_{11} \\ X_{12}D_2 &= D_1Y_{12} \\ X_{22}D_2 &= D_3Y_{12} + D_2Y_{22} \end{aligned} \tag{89}$$

Assumption (67) implies that for every $f(z) \in X_{D_2}$ we have

$$\begin{aligned}\begin{bmatrix}0\\0\end{bmatrix} &= \pi_{\begin{bmatrix}D_1(z) & 0\\ D_3(z) & D_2(z)\end{bmatrix}} \begin{bmatrix}X_{11} & X_{12}\\ X_{21} & X_{22}\end{bmatrix}\begin{bmatrix}0\\ f\end{bmatrix}\\ &= \begin{bmatrix}D_1(z) & 0\\ D_3(z) & D_2(z)\end{bmatrix}\pi_- \begin{bmatrix}D_1(z)^{-1} & 0\\ -D_2^{-1}D_3(z)D_1^{-1} & D_2(z)^{-1}\end{bmatrix}\begin{bmatrix}X_{12}f\\ X_{22}f\end{bmatrix}\\ &= \begin{bmatrix}D_1(z) & 0\\ D_3(z) & D_2(z)\end{bmatrix}\pi_- \begin{bmatrix}D_1(z)^{-1}X_{12}f\\ -D_2^{-1}D_3(z)D_1^{-1}X_{12}f + D_2(z)^{-1}X_{22}f\end{bmatrix}.\end{aligned}$$

Since from (89) we have $D_1(z)^{-1}X_{12}(z) = Y_{12}(z)D_1(z)^{-1}$, this implies that for $f(z) \in X_{D_2}$, we have

$$\pi_- D_1(z)^{-1}X_{12}f = \pi_- Y_{12}D_1(z)^{-1}f = 0.$$

The same identity holds for every $f(z) \in D_2\mathbb{F}[z]^m$, hence for all $f(z) \in \mathbb{F}[z]^m$. From this we conclude that $Y_{12}(z)D_1(z)^{-1}$ is a polynomial matrix. As it is also strictly proper, it follows that $Y_{12}(z) = 0$ and $X_{12}(z) = 0$ as well. Starting now from $\pi_- D_2^{-1}X_{22}f = \pi_- Y_{22}D_2^{-1}f = 0$, and using the same argument as before, we conclude that $Y_{22}(z) = 0$ and $X_{22}(z) = 0$. Using these identities, equation (63) can be rewritten as

$$\begin{bmatrix}X_{11}(z) & 0\\ X_{21}(z) & 0\end{bmatrix}\begin{bmatrix}D_1(z) & 0\\ D_3(z) & D_2(z)\end{bmatrix} = \begin{bmatrix}D_1(z) & 0\\ D_3(z) & D_2(z)\end{bmatrix}\begin{bmatrix}Y_{11}(z) & 0\\ Y_{21}(z) & 0\end{bmatrix},$$

which is equivalent to (68). In turn, this can be rewritten as the pair of equations (75).

(d) Assumption (69) implies that for every $f_1(z) \in X_{D_1}, f_2(z) \in X_{D_2}$, we have

$$\pi_{\begin{bmatrix}D_1 & 0\\ D_3 & D_2\end{bmatrix}}\begin{bmatrix}X_{11} & X_{12}\\ X_{21} & X_{22}\end{bmatrix}\begin{bmatrix}f_1\\ f_2\end{bmatrix} = \begin{bmatrix}0\\ g\end{bmatrix} \in \mathcal{V}.$$

Choosing $f_1(z) = 0$, we get

$$\begin{aligned}\begin{bmatrix}0\\ g\end{bmatrix} &= \pi_{\begin{bmatrix}D_1 & 0\\ D_3 & D_2\end{bmatrix}}\begin{bmatrix}X_{11} & X_{12}\\ X_{21} & X_{22}\end{bmatrix}\begin{bmatrix}0\\ f_2\end{bmatrix}\\ &= \pi_{\begin{bmatrix}D_1 & 0\\ D_3 & D_2\end{bmatrix}}\begin{bmatrix}X_{12}\\ X_{22}\end{bmatrix}\begin{bmatrix}f_1\\ f_2\end{bmatrix}.\end{aligned}$$

This implies $\pi_{D_1}X_{12}f_2 = 0$ for all $f_2 \in X_{D_2}$ and hence, as in the previous part, we conclude that $X_{12}(z) = 0$ and hence also $Y_{12}(z) = 0$. Redoing the argument with $f_2(z) = 0$, we obtain $\pi_{D_1}X_{11}f_1 = 0$ for all $f_1(z) \in X_{D_1}$. Since this holds also for all $f(z) \in D_1\mathbb{F}[z]^m$, we can conclude that $X_{11}(z) = Y_{11}(z) = 0$. Thus (63) reduces to (70). By Theorem 8, (69) holds if and only if $D_2(z), \begin{bmatrix}X_{21}(z) & X_{22}(z)\end{bmatrix}$ are left coprime.

2. Follows by a simple computation.

3. Follows from the intertwining relation (68) and Theorems 7 and 8.

4. (a) Using Theorems 7 and 8, this follows from the intertwining relation

$$\begin{bmatrix} I & 0 \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} = D_1(z)\begin{bmatrix} I & 0 \end{bmatrix} \tag{90}$$

Applying Theorem 8, (78) follows from the factorization

$$\begin{bmatrix} I & 0 \\ D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} = \begin{bmatrix} I & 0 \\ D_1(z) & 0 \\ D_3(z) & I \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & D_2(z) \end{bmatrix} \tag{91}$$

and the fact that the polynomial matrix $\left[\begin{smallmatrix} I & 0 \\ D_1(z) & 0 \\ D_3(z) & I \end{smallmatrix}\right]$ is right prime.

(b) This is immediate.

5. (a) Follows from (74).

(b) We have $g(z) \in \operatorname{Im} J_1 \cap \operatorname{Im} J_2$, if and only if there exist $f_i(z) \in X_{D_i}$ for which $g(z) = J_1 f_1 = J_2 f_2$. However, this is equivalent to $\left[\begin{smallmatrix} f_1 \\ -f_2 \end{smallmatrix}\right] \in \operatorname{Ker} J$.

(c) By Theorem 8, J is injective if and only if

$$\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}, \begin{bmatrix} Y_{11}(z) & 0 \\ Y_{21}(z) & I \end{bmatrix}$$

are right coprime. Clearly, this is the case if and only if $D_1(z), Y_{11}(z)$ are right coprime.

(d) Again, applying Theorem 8, J is surjective if and only if the polynomial matrices

$$\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}, \begin{bmatrix} X_{11}(z) & 0 \\ X_{21}(z) & I \end{bmatrix}$$

are left coprime. This is equivalent to the left coprimeness of $D_1(z)$ and $X_{11}(z)$.

(e) Since $\dim X_{\left[\begin{smallmatrix} D_1 & 0 \\ 0 & D_2 \end{smallmatrix}\right]} = \dim X_{\left[\begin{smallmatrix} D_1 & 0 \\ D_3 & D_2 \end{smallmatrix}\right]}$, J is invertible if and only if it is injective. The same holds for surjectivity.

6. (a) Follows from the first, intertwining, relation in (75) and Theorem 7.

(b) We compute, for $f(z) \in X_{D_1}$,

$$\begin{aligned} P_1 J_1 f &= \pi_{D_1}\begin{bmatrix} I & 0 \end{bmatrix}\pi_{\left[\begin{smallmatrix} D_1 & 0 \\ D_3 & D_2 \end{smallmatrix}\right]}\begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} f = \pi_{D_1}\begin{bmatrix} I & 0 \end{bmatrix}\begin{bmatrix} X_{11} \\ X_{21} \end{bmatrix} f \\ &= \pi_{D_1} X_{11} f = Q_1 f. \end{aligned}$$

(c) Follows from the intertwining relation $X_{11}(z)D_1(z) = D_1(z)Y_{11}(z)$ and Theorem 8, noting that in this case surjectivity is equivalent to invertibility.

(d) Restricting P_1 to $\mathcal{W}+\mathcal{V}=\operatorname{Im}J$, its image is $\mathcal{Q}$ and its kernel is $\mathcal{V}$, which proves the isomorphism (85). The isomorphism (86) follows from (85) by the standard module isomorphism $(\mathcal{W}+\mathcal{V})/\mathcal{V}\simeq\mathcal{W}/(\mathcal{W}\cap\mathcal{V})$. However, we give also a direct proof. We use (81), i.e., $Q_1=P_1J_1$. Let $\begin{bmatrix} f_1\\ f_2\end{bmatrix}\in\operatorname{Im}J_1=\mathcal{W}$. Then $\begin{bmatrix} f_1\\ f_2\end{bmatrix}\in\operatorname{Ker}P_1|_{\mathcal{W}}$ if and only if $f_1=0$, that is $\begin{bmatrix} f_1\\ f_2\end{bmatrix}\in\operatorname{Im}J_2$. Conversely, if $\begin{bmatrix} f_1\\ f_2\end{bmatrix}\in\mathcal{W}\cap\mathcal{V}$, then necessarily $f_1=0$ which shows $\operatorname{Ker}P_1|_{\mathcal{W}}=\operatorname{Im}J_1\cap\operatorname{Im}J_2$, and hence (86) follows.

7. (a) In both cases, the invertibility is equivalent to the left coprimeness of $D_1(z),X_{11}(z)$.

(b) Invertibility of J is equivalent to the right coprimeness of $D_1(z),Y_{11}(z)$ and, alternatively, to the left coprimeness of $D_1(z),X_{11}(z)$. Thus the intertwining relation $X_{11}D_1=D_1Y_{11}$ can be embedded in the doubly coprime factorization

$$\begin{aligned}\begin{bmatrix} K_{11}(z) & K_{12}(z)\\ -X_{11}(z) & D_1(z)\end{bmatrix}\begin{bmatrix} D_1(z) & L_{12}(z)\\ Y_{11}(z) & L_{22}(z)\end{bmatrix}&=\begin{bmatrix} I & 0\\ 0 & I\end{bmatrix}\\ \begin{bmatrix} D_1(z) & L_{12}(z)\\ Y_{11}(z) & L_{22}(z)\end{bmatrix}\begin{bmatrix} K_{11}(z) & K_{12}(z)\\ -X_{11}(z) & D_1(z)\end{bmatrix}&=\begin{bmatrix} I & 0\\ 0 & I\end{bmatrix}.\end{aligned}\tag{92}$$

In particular, we have the following identities

$$\begin{aligned} D_1(z)K_{12}(z)&=-L_{12}(z)D_1(z)\\ D_1(z)K_{11}(z)-L_{12}X_{11}(z)&=I\end{aligned}\tag{93}$$

From equation (73) and using the identities in (93), we have

$$X_{21}(z)D_1(z)-D_1(z)Y_{21}(z)=D_3(z)Y_{11}(z),$$

which implies

$$\begin{aligned} X_{21}(z)D_1(z)K_{12}(z)&=D_1(z)Y_{21}(z)K_{12}(z)+D_3(z)Y_{11}(z)K_{12}(z)\\ &=D_1(z)Y_{21}(z)K_{12}(z)+D_3(z)(I-L_{22}(z)D_1(z)).\end{aligned}$$

From this we see that the Sylvester equation

$$X(z)D_1(z)-D_1(z)Y(z)=I\tag{94}$$

is polynomially solvable, taking $X(z)=D_3(z)L_{22}(z)-X_{21}(z)L_{12}(z)$ and $Y(z)=Y_{21}(z)K_{12}(z)$. □

The space $\mathcal{Q}\subset X_{D_1}$ measures to what extent $\mathcal{W}$ is complementary to the subspace $\mathcal{V}$ of $X_{\begin{bmatrix} D_1 & 0\\ D_3 & D_2\end{bmatrix}}$. Clearly, we have $\dim(\mathcal{W}+\mathcal{V})=\dim\mathcal{W}+\dim\mathcal{V}-\dim(\mathcal{W}\cap\mathcal{V})$. Using (85) and (86), it follows that

$$\dim(\mathcal{W}+\mathcal{V})-\dim\mathcal{V}=\dim\mathcal{W}-\dim(\mathcal{W}\cap\mathcal{V})=\dim\mathcal{Q},$$

or, equivalently,

$$\dim(\mathcal{W}+\mathcal{V})/\mathcal{V}=\dim\mathcal{W}/(\mathcal{W}\cap\mathcal{V})=\dim\mathcal{Q}.$$

Thus $\mathcal{W}$ is complementary to $\mathcal{V}$ if and only if $\dim(\mathcal{W}+\mathcal{V})=\dim\mathcal{W}+\dim\mathcal{V}$. This in turn is equivalent to $\dim\mathcal{W}=\dim\mathcal{Q}$. Thus we have the following.

Corollary 23. *With the notation of Proposition 22, the following equivalent statements hold.*

1. *The invariant subspace $\mathcal{W}$ is complementary to the invariant subspace $\mathcal{V}$.*

2. *The polynomial matrices $X_{11}(z)$ and $D_1(z)$ are left coprime.*

3. *The polynomial matrices $Y_{11}(z)$ and $D_1(z)$ are right coprime.*

Proof. The subspace $\mathcal{W}$ is complementary to $\mathcal{V}$ if and only if the map J, defined in (81), is invertible and this is equivalent to it being either injective or surjective. These properties are characterized by the right coprimeness of $Y_{11}(z)$ and $D_1(z)$ and the left coprimeness of $X_{11}(z)$ and $D_1(z)$ respectively. □

The previous corollary leads directly to the study of skew primeness.

9 Skew-primeness

Given a linear transformation A acting in the space $\mathcal{X}$, not every A-invariant subspace has a complementary A-invariant subspace. We proceed with the characterization of those invariant subspaces for which an invariant complement exists. Our starting point is the study of this problem for the case of polynomial models and the shift operator in them. Since, by Theorem 3, invariant subspaces are characterized in terms of the factorization of a nonsingular polynomial matrix, it is only to be expected that the characterization we are after is going to relate to factorization theory. We have seen, in Subsection 3.3, how the geometry of submodules of a polynomial model can be characterised in terms of the arithmetic of polynomial matrices and in particular of coprimeness properties. The same turns out to be true in the case of the characterization of the existence of a complementary invariant subspace. The relevant condition is skew-primeness to be introduced shortly. Moreover, just as left or right coprimeness can be expressed in terms of the solvability of appropriate Bezout equations, skew-primeness will turn out to be equivalent to a Sylvester type equation. We recall the concept of skew-primeness of polynomial matrices and the principal result.

Definition 24. Let $D_1(z), D_2(z)\in\mathbb{F}[z]^{p\times p}$ be nonsingular polynomial matrices. The ordered pair $(D_1(z), D_2(z))$ is called **skew-prime** if there exist polynomial matrices $\overline{D}_1(z)$ and $\overline{D}_2(z)$ such that

1. $D_1(z)D_2(z)=\overline{D}_2(z)\overline{D}_1(z)$

2. $D_1(z)$ and $\overline{D}_2(z)$ are left coprime

3. $D_2(z)$ and $\overline{D}_1(z)$ are right coprime.

In this case we will say that the pair $(\overline{D}_2(z), \overline{D}_1(z))$ is a **skew-complement** of $(D_1(z), D_2(z))$. Note that a sufficient, but not necessary, condition for a pair $(D_1(z), D_2(z))$ to be skew-prime is that $\det D_1(z), \det D_2(z)$ are coprime.

For the following result, which we state without proof, see Fuhrmann [11]. The geometric interpretation of skew-primeness is due to Khargonekar, Georgiou and Özgüler [16].

Theorem 25. *Let $D_1(z)$, $D_2(z) \in \mathbb{F}[z]^{m\times m}$ be nonsingular polynomial matrices. Then the following statements are equivalent.*

1. *$D_1(z)$ and $D_2(z)$ are skew-prime.*
2. *The submodule $D_1 X_{D_2} \subset X_{D_1 D_2}$ is an $\mathbb{F}[z]$–direct summand, i.e., it has a complementary submodule.*
3. *The equation*

$$X(z)D_1(z) + D_2(z)Y(z) = I \tag{95}$$

has a polynomial solution.

Proof. The equivalence of 1. and 2. was proved in Corollary 6. So it suffices to prove the equivalence of 1. and 3.

Assume the pair $(D_1(z), D_2(z))$ is skew-prime. Hence, by Definition 24, there exist polynomial matrices $\overline{D}_2(z)$ and $\overline{D}_1(z)$ satisfying

$$D_1(z)D_2(z) = \overline{D}_2(z)\overline{D}_1(z), \tag{96}$$

with $D_1(z), \overline{D}_2(z)$ left coprime and $D_2(z), \overline{D}_1(z)$ right coprime. We apply now Theorems 7 and 8 to conclude the existence of an invertible map $Z : X_{\overline{D}_1} \longrightarrow X_{D_1}$ intertwining $S_{\overline{D}_1}$ and S_{D_1}, i.e., satisfying $ZS_{\overline{D}_1} = S_{D_1}Z$. The map Z is given, for $f \in X_{\overline{D}_1}$, by

$$Zf = \pi_{D_1}\overline{D}_2 f. \tag{97}$$

Clearly, $Z^{-1} : X_{D_1} \longrightarrow X_{\overline{D}_1}$ is an invertible map satisfying

$$Z^{-1}S_{D_1} = S_{\overline{D}_1}Z^{-1}.$$

Since $D_1(z), \overline{D}_2(z)$ are left coprime, there exist polynomial matrices $\overline{Y}(z), X(z)$ for which

$$\overline{D}_2(z)\overline{Y}(z) + D_1(z)X(z) = I. \tag{98}$$

Similarly, by the right coprimeness of $D_2(z), \overline{D}_1(z)$, there exist polynomial matrices $Y(z), \overline{X}(z)$ for which

$$\overline{X}(z)\overline{D}_1(z) + Y(z)D_2(z) = I. \tag{99}$$

Putting equations (96)-(99) into matrix form, we get

$$\begin{bmatrix} \overline{D}_2(z) & D_1(z) \\ \overline{X}(z) & -Y(z) \end{bmatrix} \begin{bmatrix} \overline{Y}(z) & \overline{D}_1(z) \\ X(z) & -D_2(z) \end{bmatrix} = \begin{bmatrix} I & 0 \\ K(z) & I \end{bmatrix}.$$

Multiplying on the left by $\left[\begin{smallmatrix} I & 0 \\ -K(z) & I \end{smallmatrix}\right]$ and appropriately redefining $\overline{X}(z)$ and $Y(z)$, we obtain the doubly coprime factorization

$$\begin{aligned} \begin{bmatrix} \overline{D}_2(z) & D_1(z) \\ \overline{X}(z) & -Y(z) \end{bmatrix} \begin{bmatrix} \overline{Y}(z) & \overline{D}_1(z) \\ X(z) & -D_2(z) \end{bmatrix} &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}, \\ \begin{bmatrix} \overline{Y}(z) & \overline{D}_1(z) \\ X(z) & -D_2(z) \end{bmatrix} \begin{bmatrix} \overline{D}_2(z) & D_1(z) \\ \overline{X}(z) & -Y(z) \end{bmatrix} &= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \end{aligned} \tag{100}$$

In particular, this implies the equality

$$\overline{Y}(z)\overline{D}_2(z) + \overline{D}_1(z)\overline{X}(z) = I. \tag{101}$$

From (99) we get $D_2(z)\overline{Y}(z) + D_1(z)X(z) = I$. We multiply this equality by $D_1(z)^{-1}$ on the left and by $D_1(z)$ on the right to obtain

$$D_1(z)^{-1}\overline{D}_2(z)\overline{Y}(z)D_1(z) + X(z)D_1(z) = I.$$

We use now the equalities $D_1(z)D_2(z) = \overline{D}_2(z)\overline{D}_1(z)$ and $\overline{Y}(z)D_1(z) = \overline{D}_1(z)Y(z)$ to obtain

$$X(z)D_1(z) + D_2(z)Y(z) = I.$$

Conversely, assume the existence of polynomial matrices $X(z)$ and $Y(z)$ that solve the polynomial Sylvester equation (95). Clearly $D_1(z)$ and $Y(z)$ are right coprime. The rational right coprime matrix fraction $Y(z)D_1(z)^{-1}$ has also a left coprime matrix fraction representation $\overline{D}_1(z)^{-1}\overline{Y}(z)$. This implies the equality

$$\overline{Y}(z)D_1(z) = \overline{D}_1(z)Y(z) \tag{102}$$

and, using Theorems 7 and 8, we conclude that the map $T : X_{D_1} \longrightarrow X_{\overline{D}_1}$, defined, for $f \in X_{D_1}$, by $Tf = \pi_{\overline{D}_1}\overline{Y}f$ is invertible. In particular, this implies

$$\dim X_{D_1} = \deg\det D_1 = \deg\det \overline{D}_1 = \dim X_{\overline{D}_1}. \tag{103}$$

Multiplying (95) on the left by $D_1(z)$ and on the right by $D_1(z)^{-1}$, we have

$$I = D_1(z)X(z) + D_1(z)D_2(z)Y(z)D_1(z)^{-1} = D_1(z)X(z) + D_1(z)D_2(z)\overline{D}_1(z)^{-1}\overline{Y}(z).$$

Since $D_1(z)D_2(z)\overline{D}_1(z)^{-1}\overline{Y}(z) = I - D_1(z)X(z)$ is a polynomial matrix and $\overline{D}_1(z)$, $\overline{Y}(z)$ are left coprime, we conclude that there exists a, necessarily nonsingular, polynomial matrix $\overline{D}_2(z)$ for which

$$\overline{D}_2(z)\overline{D}_1(z) = D_1(z)D_2(z). \tag{104}$$

This implies the equality $D_1(z)X(z)+\overline{D}_2(z)\overline{Y}(z)=I$, which shows that $D_1(z),\overline{D}_2(z)$ are left coprime. Defining the map $Z:X_{\overline{D}_1}\longrightarrow X_{D_1}$ by (97) it follows from Theorem 8 that Z is surjective. Since by (103) $\dim X_{\overline{D}_1}=\dim X_{D_1}$, it is injective as well and hence, again by Theorem 8, $D_2(z),\overline{D}_1(z)$ are right coprime, which proves the skew-primeness of $(D_1(z),D_2(z))$. This completes the proof of the theorem. We note however that there exist polynomial matrices $\overline{X}(z),Y(z)$ for which $\overline{X}(z)\overline{D}_1(z)+Y(z)D_2(z)=I$. Modifying the definition of $\overline{X}(z),Y(z)$ we obtain the doubly coprime factorization (100). □

We recall that in Proposition 14 we proved the equivalence of the polynomial matrices $D_1(z)D_2(z)$ and $\begin{bmatrix} D_1(z) & 0\\ -I & D_2(z)\end{bmatrix}$. Thus, the following result can be expected.

Proposition 26. *Let $D_i(z)\in\mathbb{F}[z]^{m\times m}$ be nonsingular. Then $D(z)=D_1(z)D_2(z)$ is a skew-prime factorization if and only if*

$$\begin{bmatrix} D_1(z) & 0\\ -I & D_2(z)\end{bmatrix}=\begin{bmatrix} D_1(z) & 0\\ -I & I\end{bmatrix}\begin{bmatrix} I & 0\\ 0 & D_2(z)\end{bmatrix}\tag{105}$$

is a skew-prime factorization.

Proof. Actually, the statement follows from Proposition 14 and (45) in particular. However, we include also a direct proof.

Assume $D(z)=D_1(z)D_2(z)$ is a skew-prime factorization. By Theorem 25, there exist polynomial matrices $X(z),Y(z)$ for which

$$X(z)D_1(z)+D_2(z)Y(z)=I.\tag{106}$$

This implies

$$\begin{bmatrix} 0 & 0\\ X(z) & I\end{bmatrix}\begin{bmatrix} D_1(z) & 0\\ -I & I\end{bmatrix}+\begin{bmatrix} I & 0\\ 0 & D_2(z)\end{bmatrix}\begin{bmatrix} I & 0\\ Y(z) & 0\end{bmatrix}=\begin{bmatrix} I & 0\\ 0 & I\end{bmatrix},$$

which, by Theorem 25, shows that (105) is a skew prime factorization. In fact, the factorization

$$\begin{bmatrix} D_1(z) & 0\\ -I & D_2(z)\end{bmatrix}=\begin{bmatrix} I & 0\\ -X(z) & D_2(z)\end{bmatrix}\begin{bmatrix} D_1(z) & 0\\ -Y(z) & I\end{bmatrix}\tag{107}$$

is complementary to the factorization (105) as the coprimeness conditions are easily checked.

Conversely, assume (105) is a left skew-prime factorization. Thus there exist polynomial matrices X_{ij},Y_{ij} such that

$$\begin{bmatrix} X_{11}(z) & X_{12}(z)\\ X_{21}(z) & X_{22}(z)\end{bmatrix}\begin{bmatrix} D_1(z) & 0\\ -I & I\end{bmatrix}+\begin{bmatrix} I & 0\\ 0 & D_2(z)\end{bmatrix}\begin{bmatrix} Y_{11}(z) & Y_{12}(z)\\ Y_{21}(z) & Y_{22}(z)\end{bmatrix}=\begin{bmatrix} I & 0\\ 0 & I\end{bmatrix},$$

This leads to the system of equations

$$\begin{aligned} X_{21}(z)D_1(z) + D_2(z)Y_{21}(z) &= X_{22}(z) \\ X_{11}(z)D_1(z) - X_{12}(z) + Y_{11}(z) &= I \\ X_{12}(z) + Y_{12}(z) &= 0 \\ X_{22}(z) + D_2(z)Y_{22}(z) &= I. \end{aligned}$$

From the first and last equations we get $X_{21}(z)D_1(z) + D_2(z)Y_{21}(z) = I - D_2(z)Y_{22}(z)$ or $X_{21}(z)D_1(z) + D_2(z)(Y_{21}(z) + Y_{22}(z)) = I$, i.e., (106) is polynomially solvable with $X(z) = X_{21}(z)$ and $Y(z) = Y_{21}(z) + Y_{22}(z)$. □

The following result is a slight generalization of Theorem 25.

Theorem 27. *Let $D_1(z) \in \mathbb{F}[z]^{m\times m}$, $D_2(z) \in \mathbb{F}[z]^{p\times p}$ be nonsingular and $D_3(z) \in \mathbb{F}[z]^{p\times m}$. Then the following statements are equivalent.*

1. *The polynomial matrices $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$ and $\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}$ are unimodularly equivalent.*
2. *The polynomial Sylvester equation*
$$X(z)D_1(z) + D_2(z)Y(z) = D_3(z) \tag{108}$$
has a polynomial solution.
3. *The factorization*
$$\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & D_2(z) \end{bmatrix} \tag{109}$$
is skew-prime.
4. *The invariant subspace*
$$\mathcal{V} = \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix} X_{\left[\begin{smallmatrix} I & 0 \\ 0 & D_2(z) \end{smallmatrix}\right]} = \left\{ \begin{bmatrix} 0 \\ f(z) \end{bmatrix} | f(z) \in X_{D_2} \right\}$$
is a direct summand of $X_{\left[\begin{smallmatrix} D_1 & 0 \\ D_3 & D_2 \end{smallmatrix}\right]}$.
5. *The elementary divisors of $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$ are those of $D_1(z)$ together with those of $D_2(z)$.*

Proof. 1. ⇔ 2.
Assume $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$ and $\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}$ are equivalent. We will show that the equivalence can be given by unimodular matrices in the form

$$\begin{bmatrix} I & 0 \\ -X(z) & I \end{bmatrix} \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} \begin{bmatrix} I & 0 \\ -Y(z) & I \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix} \tag{110}$$

Conversely, assume $X(z), Y(z)$ is a polynomial solution of equation (108). We compute

$$\begin{bmatrix} I & 0 \\ X(z) & I \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}\begin{bmatrix} I & 0 \\ Y(z) & I \end{bmatrix} = \begin{bmatrix} D_1(z) & 0 \\ X(z)D_1(z)+D_2(z)Y(z) & D_2(z) \end{bmatrix}$$
$$= \begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$$

As $\begin{bmatrix} I & 0 \\ X(z) & I \end{bmatrix}, \begin{bmatrix} I & 0 \\ Y(z) & I \end{bmatrix}$ are both unimodular, the equivalence of

$$\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}$$

follows.

2. ⇔ 3.

Next, consider the factorization (109). Assume equation (108) has a solution. Then

$$\begin{bmatrix} 0 & 0 \\ X(z) & I \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & D_2(z) \end{bmatrix}\begin{bmatrix} I & 0 \\ Y(z) & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

i.e., $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix}$ and $\begin{bmatrix} I & 0 \\ 0 & D_2(z) \end{bmatrix}$ are skew-prime.

Conversely, assume 3. Thus there exist polynomial matrices $X_{ij}(z)$ and $Y_{ij}(z)$ such that

$$\begin{bmatrix} X_{11}(z) & X_{12}(z) \\ X_{21}(z) & X_{22}(z) \end{bmatrix}\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & I \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & D_2(z) \end{bmatrix}\begin{bmatrix} Y_{11}(z) & Y_{12}(z) \\ Y_{21}(z) & Y_{22}(z) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

This implies the following equations

$$\begin{aligned} X_{21}(z)D_1(z)+X_{22}(z)D_3(z)+D_2(z)Y_{21}(z) &= 0 \\ X_{11}(z)D_1(z)+X_{12}(z)D_3(z)+Y_{11}(z) &= I \\ X_{12}(z)+Y_{12}(z) &= 0 \\ X_{22}(z)+D_2(z)Y_{22}(z) &= I. \end{aligned}$$

From these equations we obtain by substitution

$$X_{21}(z)D_1(z)+D_2(z)(Y_{21}(z)-Y_{22}(z)D_3(z))+D_3(z) = 0,$$

i.e., the Sylvester equation (108) is polynomially solvable.

3. ⇔ 4.

Follows from the factorization (109) and Theorem 25.

1. ⇔ 5.

The polynomial matrices $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$ and $\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}$ are equivalent if and only if they have the same elementary divisors. However, the elementary divisors of $\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}$ are those of $D_1(z)$ together with those of $D_2(z)$, hence the same is true for $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$.

The converse follows by the same argument used in the proof of Theorem 16. □

We point out that factorizations complementary to (109) are given by

$$\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix} = \begin{bmatrix} I & 0 \\ X(z) & D_2(z) \end{bmatrix} \begin{bmatrix} D_1(z) & 0 \\ Y(z) & I \end{bmatrix}, \tag{111}$$

where $X(z), Y(z)$ solve the polynomial Sylvester equation (108).

The previous theorem was stated in terms of polynomial matrices. It has an interpretation in terms of linear transformations. Before stating it, we prove a technical lemma.

Lemma 28. *Let $D_1(z)$ and $D_2(z)$ be nonsingular polynomial matrices such that $(D_1(z)D_2(z))^{-1}$ is strictly proper. Then if*

$$X(z)D_1(z) + D_2(z)Y(z) = I \tag{112}$$

has a polynomial matrix solution then it has one with $D_2(z)^{-1}X(z)$ and $YD_1(z)^{-1}$ strictly proper.

Proof. Let $X(z) = X_1(z) + D_2(z)X_2(z)$ and $Y(z) = Y_1(z) + Y_2(z)D_1(z)$ with

$$D_2(z)^{-1}X_1(z) \quad \text{and} \quad Y_1(z)D_1(z)^{-1}$$

strictly proper. Then (112) implies

$$X_1(z)D_1(z) + D_2(z)Y_1(z) + D_2(z)(X_2(z) + Y_2(z))D_1(z) = I$$

or

$$\begin{aligned} D_2(z)^{-1}X_1(z) + Y_1(z)D_1(z)^{-1} + (X_2(z) + Y_2(z)) &= D_2(z)^{-1}D_1(z)^{-1} \\ &= (D_1(z)D_2(z))^{-1}. \end{aligned}$$

This implies $X_2(z) + Y_2(z) = 0$ and $X_1(z)D_1(z) + D_2(z)Y_1(z) = I$. □

Theorem 27 has the following simple consequence.

Theorem 29. *Let $D_1(z) \in \mathbb{F}[z]^{p\times p}, D_2(z) \in \mathbb{F}[z]^{m\times m}$ be nonsingular polynomial matrices. Then $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$ and $\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}$ are equivalent for any polynomial matrix $D_3(z) \in \mathbb{F}[z]^{m\times p}$ if and only if $\det D_1(z)$ and $\det D_2(z)$ are coprime.*

Proof. Assume $\det D_1(z)$ and $\det D_2(z)$ are coprime. Then there exist polynomials $a_1(z)$ and $a_2(z)$ such that $a_1(z)\det D_1(z)I + a_2(z)\det D_2(z)I = I$. Since $\det D(z) = D(z)\operatorname{adj} D(z)$ we have

$$D_1(z)(a_1(z)\operatorname{adj} D_1(z)) + D_2(z)(a_2(z)\operatorname{adj} D_2(z)) = I$$

and hence, by Theorem 27, equivalence follows.

We prove the converse by contradiction. Assume $\begin{bmatrix} D_1(z) & 0 \\ D_3(z) & D_2(z) \end{bmatrix}$ and $\begin{bmatrix} D_1(z) & 0 \\ 0 & D_2(z) \end{bmatrix}$ are equivalent for all $D_3(z)$, which implies that the corresponding sets of elementary divisors are equal. Without loss of generality, we may assume that $D_1(z), D_2(z)$ are diagonal with their elementary divisors on the respective diagonals. If the coprimeness assumption is not satisfied, then there exists a pair of elementary divisors $e_i(z)$ of $D_1(z)$ and $f_j(z)$ of $D_2(z)$ which are powers of the same prime polynomial, say $e_i(z) = \pi(z)^{\mu_i}$ and $f_j(z) = \pi(z)^{\nu_j}$. Choosing $D_3(z)$ to be zero but for 1 as the ij element, and noting that the elementary divisors of $\begin{bmatrix} \pi(z)^{\mu_i} & 0 \\ 1 & \pi(z)^{\nu_j} \end{bmatrix}$ are $\pi(z)^{\mu_i+\nu_j}, 1$, we get a contradiction. □

10 Matrix representations

In this section we translate some of the results obtained in the context of polynomial models to the language of matrices with entries in the underlying field $\mathbb{F}$.

Proposition 30. *Let $\mathcal{A} : \mathbb{F}^n \longrightarrow \mathbb{F}^n$ be linear and let $\mathcal{V} \subset \mathbb{F}^n$ be a k-dimensional $\mathcal{A}$–invariant subspace. Choose a basis so that $\mathcal{A} = \begin{bmatrix} A & 0 \\ C & B \end{bmatrix}$, with $A \in \mathbb{F}^{(n-k)\times(n-k)}$, $B \in \mathbb{F}^{k\times k}$, $C \in \mathbb{F}^{k\times(n-k)}$, $X \in \mathbb{F}^{(n-k)\times(n-k)}$ and $Z \in \mathbb{F}^{k\times(n-k)}$, i.e.,*

$$\mathcal{V} = \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix} = \left\{ \begin{bmatrix} 0 \\ \xi \end{bmatrix} | \xi \in \mathbb{F}^k \right\}.$$

Then

1. *(a) There exists a matrix $\begin{bmatrix} X & U \\ Z & Y \end{bmatrix}$ that commutes with $\begin{bmatrix} A & 0 \\ C & B \end{bmatrix}$ and that fulfils (b)* Ker $\begin{bmatrix} X & U \\ Z & Y \end{bmatrix} = \mathcal{V}$, *if and only if $U = 0$, $Y = 0$, $\begin{bmatrix} X \\ Z \end{bmatrix}$ is left invertible and the following equations are satisfied*

$$\begin{cases} XA = AX \\ ZA = BZ + CX. \end{cases} \tag{113}$$

2. *Equation (*113*) can be rewritten as one of the following matrix equations.*

$$\begin{bmatrix} X \\ Z \end{bmatrix} A = \begin{bmatrix} A & 0 \\ C & B \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix}, \tag{114}$$

or

$$\begin{bmatrix} A & 0 \\ C & B \end{bmatrix} \begin{bmatrix} X & 0 \\ Z & I \end{bmatrix} = \begin{bmatrix} X & 0 \\ Z & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}. \tag{115}$$

3. *The subspace $\mathcal{W} = \operatorname{Im} \begin{bmatrix} X \\ Z \end{bmatrix}$ is an $\begin{bmatrix} A & 0 \\ C & B \end{bmatrix}$-invariant subspace. $\mathcal{W}$ is complementary to $\mathcal{V}$ if and only if X is invertible.*

4. *We have the isomorphism*

$$A \simeq \begin{bmatrix} A & 0 \\ C & B \end{bmatrix} |_{\mathbb{F}^n/\mathcal{V}}, \tag{116}$$

i.e., A is isomorphic to the map induced by $\mathcal{A}$ in the quotient space $\mathbb{F}^n/\mathcal{V}$.

5. *If* (115) *holds with X nonsingular, then we can assume, without loss of generality, that for some Z we have*

$$\begin{bmatrix} A & 0 \\ C & B \end{bmatrix}\begin{bmatrix} I & 0 \\ Z & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ Z & I \end{bmatrix}\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}. \tag{117}$$

Proof.

1. The existence follows from Theorem 20. Condition (b) implies that $U = 0$, $Y = 0$ and $\left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right]$ is left invertible. The commutativity condition (a) translates into

$$\begin{bmatrix} X & 0 \\ Z & 0 \end{bmatrix}\begin{bmatrix} A & 0 \\ C & B \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & B \end{bmatrix}\begin{bmatrix} X & 0 \\ Z & 0 \end{bmatrix}, \tag{118}$$

   which is equivalent to the pair of equations (113).

2. This is immediate.

3. Follows from (114).

4. Follows from the block-triangular representation of $\mathcal{A}$.

5. Multiplying (115) on the right by $\left[\begin{smallmatrix} X^{-1} & 0 \\ 0 & I \end{smallmatrix}\right]$ and redefining Z. □

Clearly, the first statement of Proposition 30 yields Halmos' theorem. We consider now some special cases:

1. The characteristic polynomials of A and B are coprime.

   Under this assumption, the Sylvester equation $ZA - BZ = C$ has a unique solution for every C. Choose $X = I$.

2. The matrices A, B are similar.

   In this case, there exists a nonsingular R for which $RA = BR$. Choose $X = 0$, $Z = R$ and we are done.

3. A special case of the previous item is $C = 0$.

   Choose $Z = 0$ and $X = I$.

In all these cases, the constructed matrix $\left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right]$ is left invertible.

Probably, the most interesting special consequence of the previous result is Roth's Theorem [18]. This is the case when (115) holds with X invertible. In that case the matrices $\left[\begin{smallmatrix} A & 0 \\ C & B \end{smallmatrix}\right], \left[\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}\right]$ are similar. Our intention is to clarify this connection. Roth did not consider the geometric aspects of his result nor did he consider the concept of skew-primeness, which was introduced a quarter century later in Wolovich [20]. The geometric interpretation of skew-primeness was given in Khargonekar, Georgiou and Özgüler [16]. Fuhrmann [8] contains an infinite dimensional generalization of skew-primeness. This opens up the possibility of establishing the analog of Halmos's theorem in the context of backward shift invariant subspaces.

Theorem 31 (Roth). *Given matrices $A \in \mathbb{F}^{(n-k)\times(n-k)}$, $B \in \mathbb{F}^{k\times k}$ and $C \in \mathbb{F}^{k\times(n-k)}$. Let $\mathcal{A} = \left[\begin{smallmatrix} A & 0 \\ C & B \end{smallmatrix}\right]$. Then the following statements are equivalent:*

1. *We have the following similarity*
$$\begin{bmatrix} A & 0 \\ C & B \end{bmatrix} \simeq \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}. \tag{119}$$

2. *The subspace $\mathcal{V} = \operatorname{Im} \left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$ has a complementary $\left[\begin{smallmatrix} A & 0 \\ C & B \end{smallmatrix}\right]$-invariant subspace.*

3. *There exists a solution of the following Sylvester equation*
$$ZA - BZ = C. \tag{120}$$

4. *There exists a matrix commuting with $\mathcal{A}$ whose kernel is $\mathcal{V}$ and whose image is complementary to $\mathcal{V}$.*

5. *The elementary divisors of $\mathcal{A}$ are those of A together with those of B.*

6. *The following*
$$\begin{bmatrix} zI - A & 0 \\ -C & zI - B \end{bmatrix} = \begin{bmatrix} zI - A & 0 \\ -C & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & zI - B \end{bmatrix} \tag{121}$$
is a skew-prime factorization. In that case, a complementary factorization is given by
$$\begin{bmatrix} zI - A & 0 \\ -C & zI - B \end{bmatrix} = \begin{bmatrix} I & 0 \\ Z & zI - B \end{bmatrix} \begin{bmatrix} zI - A & 0 \\ -Z & I \end{bmatrix}, \tag{122}$$
*where Z is a solution of the Sylvester equation (*120*).*

Proof. 4. ⇒ 2.

By Proposition 30, there exist matrices X, Z such that $\left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right]$ is left invertible and the commutativity relation (118) holds. Equation (118) is equivalent to equation (114), hence $\mathcal{W} = \operatorname{Im} \left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right]$ is an $\left[\begin{smallmatrix} A & 0 \\ C & B \end{smallmatrix}\right]$-invariant subspace. The complementarity assumption implies that X is nonsingular.

3. ⇒ 1.

Assume Z solves the Sylvester equation (120). This implies the identity (117) and hence the similarity (119).

2. ⇒ 1.

Let $\operatorname{Im} \left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right]$ be an $\left[\begin{smallmatrix} A & 0 \\ C & B \end{smallmatrix}\right]$-invariant subspace which is complementary to $\operatorname{Im} \left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$. Without loss of generality, we can assume that $\left[\begin{smallmatrix} X \\ Z \end{smallmatrix}\right]$ is left invertible. This implies the existence of a matrix K for which
$$\begin{bmatrix} A & 0 \\ C & B \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix} = \begin{bmatrix} X \\ Z \end{bmatrix} K. \tag{123}$$

The complementarity assumption implies that X is nonsingular. From equation (123) we obtain

$$\begin{bmatrix} A & 0 \\ C & B \end{bmatrix} \begin{bmatrix} I \\ ZX^{-1} \end{bmatrix} = \begin{bmatrix} I \\ ZX^{-1} \end{bmatrix} (XKX^{-1}). \tag{124}$$

This implies $XKX^{-1} = A$. Redefining Z, we have

$$\begin{bmatrix} A & 0 \\ C & B \end{bmatrix} \begin{bmatrix} I \\ Z \end{bmatrix} = \begin{bmatrix} I \\ Z \end{bmatrix} A. \tag{125}$$

Applying Proposition 30.2, we get (117), which proves (119).

2. ⇒ 3.

The Sylvester equation (120) follows from (125).

3. ⇒ 4.

Let Z be a solution of the Sylvester equation (120). This implies (125). In turn, we have

$$\begin{bmatrix} A & 0 \\ C & B \end{bmatrix} \begin{bmatrix} I & 0 \\ Z & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ Z & 0 \end{bmatrix} \begin{bmatrix} A & 0 \\ C & B \end{bmatrix}. \tag{126}$$

This shows that $\begin{bmatrix} I & 0 \\ Z & 0 \end{bmatrix}$ commutes with $\begin{bmatrix} A & 0 \\ C & B \end{bmatrix}$. Clearly, $\operatorname{Ker} \begin{bmatrix} I & 0 \\ Z & 0 \end{bmatrix} = \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$.

5. ⇔ 1.

Follows from Theorem 27.

1. ⇒ 3.

The similarity (119) implies the equivalence of the polynomial matrices $\begin{bmatrix} zI-A & 0 \\ -C & zI-B \end{bmatrix}$ and $\begin{bmatrix} zI-A & 0 \\ 0 & zI-B \end{bmatrix}$. By Theorem 27, the polynomial Sylvester equation $Z(z)(zI-A) + (zI-B)Y(z) = -C$ is solvable. Applying Lemma 28, we can assume that Z and Y are constant matrices. This implies $Y = -Z$ and the solvability of the Sylvester equation (120).

2. ⇒ 6.

This follows from a direct computation.

6. ⇒ 1.

By Theorem 27, this implies the equivalence of the polynomial matrices $\begin{bmatrix} zI-A & 0 \\ -C & zI-B \end{bmatrix}$ and $\begin{bmatrix} zI-A & 0 \\ 0 & zI-B \end{bmatrix}$ and hence, by Theorem 13, the similarity (119). □

Theorem 32. *Let $A \in \mathbb{F}^{(n-k)\times(n-k)}$, $B \in \mathbb{F}^{k\times k}$. Then $\begin{bmatrix} A & 0 \\ C & B \end{bmatrix}$ and $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ are similar for all $C \in \mathbb{F}^{k\times(n-k)}$ if and only if the characteristic polynomials of A and B are coprime.*

Proof. Suppose $\begin{bmatrix} A & 0 \\ C & B \end{bmatrix}$ and $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ are similar for all C. This implies that $\begin{bmatrix} zI-A & 0 \\ -C & zI-B \end{bmatrix}$ and $\begin{bmatrix} zI-A & 0 \\ 0 & zI-B \end{bmatrix}$ are equivalent for all matrices C. Since any polynomial matrix $D(z)$ can be written, uniquely, as $D(z) = C + (zI-B)E(z)$, we can assume, without loss of generality, that $\begin{bmatrix} zI-A & 0 \\ D(z) & zI-B \end{bmatrix}$ and $\begin{bmatrix} zI-A & 0 \\ 0 & zI-B \end{bmatrix}$ are equivalent for all polynomial matrices $D(z)$. By Theorem 29, the characteristic polynomials of A and B are coprime.

Conversely, if the characteristic polynomials of A and B are coprime, then the polynomial matrices $\begin{bmatrix} zI-A & 0 \\ -C & zI-B \end{bmatrix}$ and $\begin{bmatrix} zI-A & 0 \\ 0 & zI-B \end{bmatrix}$ are equivalent for all C and hence the similarity part follows. □

Bibliography

[1] I. Domanov. On invariant subspaces of matrices: A new proof of a theorem of Halmos. *Linear Algebra and its Applications*, 433:2255–2256, 2010. Cited p. 127.

[2] H. Flanders and H. K. Wimmer. On the matrix equations AX-XB = C and AX-YB = C. *SIAM Journal on Applied Mathematics*, 32:707–710, 1977. Cited p. 127.

[3] F. G. Frobenius. Über die mit einer Matrix vertauschbaren Matrizen. In J.-P. Serre, editor, *Ferdinand Georg Frobenius gesammelte Abhandlungen, Band III*, pages 415–427. Springer, 1968. Cited p. 128.

[4] P. A. Fuhrmann. Algebraic system theory: An analyst's point of view. *Journal of the Franklin Institute*, 301:521–540, 1976. Cited pp. 127 and 129.

[5] P. A. Fuhrmann. On strict system equivalence and similarity. *International Journal of Control*, 25:5–10, 1977. Cited p. 135.

[6] P. A. Fuhrmann. Duality in polynomial models with some applications to geometric control theory. *IEEE Transactions on Automatic Control*, 26:284–295, 1981. Cited p. 146.

[7] P. A. Fuhrmann. On symmetric rational matrix functions. *Linear Algebra and its Applications*, 50:167–250, 1983. Cited p. 128.

[8] P. A. Fuhrmann. On skew primeness of inner functions. *Linear Algebra and its Applications*, 208–209:539–551, 1994. Cited pp. 127 and 162.

[9] P. A. Fuhrmann. *A Polynomial Approach to Linear Algebra*. Springer, 1996. Cited pp. 135 and 136.

[10] P. A. Fuhrmann. A study of behaviors. *Linear Algebra and its Applications*, 351–352:303–380, 2002. Cited p. 146.

[11] P. A. Fuhrmann. Autonomous subbehaviors and output nulling subspaces. *International Journal of control*, 78:1378–1411, 2005. Cited p. 155.

[12] P. A. Fuhrmann. On duality in some problems of geometric control. *Acta Applicandae Mathematicae*, 91:207–251, 2006. Cited p. 146.

[13] P. A. Fuhrmann. On tangential matrix interpolation. *Linear Algebra and its Applications*, 433:2018–2059, 2010. Cited p. 129.

[14] P. A. Fuhrmann and U. Helmke. On theorems of halmos and roth. In H. Dym, M. C. de Oliveira, and M. Putinar, editors, *Mathematical Methods in Systems, Optimization, and Control. Festschrift in Honor of J. William Helton*, pages 173–187. Springer, 2012. Cited pp. 127 and 128.

[15] P. R. Halmos. Eigenvectors and adjoints. *Linear Algebra and its Applications*, 4:11–15, 1971. Cited pp. 127 and 128.

[16] P. P. Khargonekar, T. T. Georgiou, and A. B. Özgüler. Skew-prime polynomial matrices: The polynomial model approach. *Linear Algebra and its Applications*, 50:403–435, 1983. Cited pp. 127, 155, and 162.

[17] H. H. Rosenbrock. *State-space and Multivariable Theory*. John Wiley, 1970. Cited p. 137.

[18] W. E. Roth. The equations AX - YB = C and AX - XB = C in matrices. *Proceedings of the American Mathematical Society*, 3:392–396. Cited pp. 127, 142, and 162.

[19] J. C. Willems. From time series to linear systems. Part I: Finite-dimensional linear time invariant systems. *Automatica*, 22:561–580, 1986. Cited p. 146.

[20] W. A. Wolovich. Skew prime polynomial matrices. *IEEE Transactions on Automatic Control*, 23:880–887, 1978. Cited pp. 127 and 162.

On MacWilliams identities for codes over rings

Heide Gluesing-Luerssen
University of Kentucky
Lexington, KY
heide.gl@uky.edu

Abstract. This note provides a unified approach to MacWilliams identities for various weight enumerators of linear block codes over Frobenius rings. Such enumerators count the number of codewords having a pre-specified property. MacWilliams identities yield a transformation between such an enumerator and the corresponding enumerator of the dual code. All identities are derived from a MacWilliams identity for the full weight enumerator using the concept of an F-partition, as introduced by Zinoviev and Ericson (1996). With this approach, all well-known identities can easily be recovered.

1 Introduction

Two of the most famous results in block coding theory are the MacWilliams Identity Theorem and the MacWilliams Equivalence Theorem [17]. Both of them deal with linear block codes over finite fields. The MacWilliams Identity relates the Hamming weight enumerator of a code to that of its dual, whereas the Equivalence Theorem states that two codes are isometric with respect to the Hamming weight if and only if they are monomially equivalent (that is, they differ only by a permutation and rescaling of the codeword coordinates). The theoretical as well as practical impact of these results is well known: for instance for high dimensional codes, MDS codes, the entire theory of self-dual codes [18, Chs. 11.3, 6.5, and 19.2], [19], or the classification of constant weight codes in [11, Thm. 7.9.5].

The central role of the Hamming weight makes an understanding of weight enumerators and isometries a must for the analysis of any class of block codes. After the discovery of the importance of linear block codes over $\mathbb{Z}_4$ for nonlinear binary codes, the entirely new area of codes over rings began to develop, and both the Identity Theorem and the Equivalence Theorem have enjoyed various generalizations to other weight functions and many classes of rings; see, for instance, [6, 25, 26] and the references therein for the Equivalence Theorem and [1, 7, 9, 12, 18, 26, 28] for the Identity Theorem; more literature will be mentioned later on.

This note deals with MacWilliams identities. From a general viewpoint such identities tell us that, and how, a particular type of information about a code fully determines the same type of information for the dual code. In the classical MacWilliams identity, this type of information is the Hamming weight enumerator which encodes, for any possible Hamming weight, the number of codewords attaining this weight. The MacWilliams transform allows us to compute the Hamming weight enumerator of the dual code from the enumerator of the primal code without further knowledge of the actual codewords.

It is natural to ask whether other weight functions have the same duality property. In this note we give an overview of the various MacWilliams identities that can be found in the literature. We show how they can be derived using a uniform approach based on F-partitions introduced by Zinoviev and Ericson in [28]. We restrict ourselves to codes over finite commutative Frobenius rings. This includes all finite fields, all integer residue rings, and all commutative finite chain rings. The restriction to Frobenius rings is a consequence of Wood's result [27, Cor. 12.4.2] stating that there cannot be a MacWilliams identity for the Hamming weight enumerator of codes over non-Frobenius rings. By restricting ourselves to commutative rings, we deliberately do not present the results in most generality, but pay attention only to the most important and best known cases. This also allows us to identify the character-theoretic dual of a code with the usual dual with respect to the dot product, and consequently, we are in the most familiar situation where codes and their duals are submodules of the same ambient space.

The central tool of our approach is the notion of an F-partition on R^n, where R is a Frobenius ring. F-partitions are based on the Fourier transform and have been introduced in [28]. In that paper Zinoviev and Ericson showed already how F-partitions can be utilized to derive MacWilliams identities. However, they presented the identities in terms of linear maps between the weight distributions of the code and its dual, and not as a transform between weight enumerator polynomials. The latter is achieved by also computing the Krawtchouk coefficients as explicitly as possible. By doing so, we recover all the familiar MacWilliams identities as well as some lesser known identities, and we illustrate how further identities can be derived.

Except for some basic results on character theory, this note is self-contained. In the next section we define commutative finite Frobenius rings in a way that is most suitable for our purposes, namely based on the existence of a generating character. We then go on and define and discuss the Fourier transforms, the Poisson summation formula, and F-partitions. We then give a MacWilliams identity for the full weight enumerator of a code. This "enumerator" is a copy of the code within a suitable polynomial ring, and thus contains all information about the code. Its MacWilliams identity is the blueprint for all other identities. They are derived, in Section 3, by specializing the full weight enumerator to the desired enumerator. For F-partitions the particular specialization does indeed lead to a well-defined MacWilliams transform for the desired enumerators. In Section 4 we present the familiar identities as special cases of our results as well as some new identities. In Section 5 we generalize the results to the case of different weight functions on various blocks of the codewords.

Before going on with ring and coding theory, let me add some personal words. A chapter on codes over rings is probably not the first thing one expects in a Festschrift in honor of Uwe Helmke. May it exemplify the long time that has passed since the day I first met Uwe. When, decades ago, I became the newest addition to the mathematical systems theory family raised by Didi Hinrichsen, my academic sibling Uwe had already left the nest and came only sporadically to Bremen. His own long and successful professional career was just taking off. Even though I worked on subjects close to Uwe's many interests, I never managed, unfortunately, to collaborate with him in this area. But I am a proud co-author of Uwe's on a paper in coding theory!

How does one transition from systems theory to coding theory? One answer is “convolutional codes”, thus linear discrete-time systems over finite fields. Joachim Rosenthal introduced me (and many others) to this subject many years ago, and, coincidentally, we wrote a joint contribution about convolutional codes and systems theory for the Festschrift on the occasion of Didi’s 60th birthday! Now it’s Uwe’s turn. It is a great pleasure and honor to contribute to the Festschrift for his 60th birthday. Here’s to Uwe!

2 Frobenius rings and the Fourier transform

In this section we collect some material on character theory, the Fourier transform, and F-partitions. We present the MacWilliams identity for the full weight enumerator of a given code. The latter is simply a copy of the code inside a polynomial ring. The identity will be our blueprint for all other MacWilliams identities later on.

Throughout this note, let R be a finite commutative ring with identity. Moreover, let $\hat{R} := \mathrm{Hom}(R, \mathbb{C}^*)$ be the character module of R, that is, $\hat{R}$ consists of all group homomorphisms from $(R, +)$ to $(\mathbb{C}^*, \cdot)$. The R-module structure of $\hat{R}$ is given by the addition $(\chi_1 + \chi_2)(a) := \chi_1(a)\chi_2(a)$ and the scalar multiplication $r\chi(a) := \chi(ra)$. The trivial map $\chi \equiv 1$ is called the *principal character* of R.

The additive groups of R and $\hat{R}$ are isomorphic [24]. For our purposes, however, we will need that even the R-modules R and $\hat{R}$ are isomorphic. This is guaranteed for the class of Frobenius rings as defined next. These rings are usually defined in a different way, namely via their socle, see [15]. For finite rings, however, this is equivalent to our definition below, see [8]. Since this property is exactly what we need in this note, we simply use this as our definition.

Definition 1. The finite commutative ring R is called *Frobenius* if there exists a character $\chi \in \hat{R}$ such that $\alpha : R \longrightarrow \hat{R},\ r \longmapsto r\chi$ is an R-isomorphism. Any character χ with this property is called a *generating character* of R.

The terminology *generating character* has been cast by Klemm [13]. Claasen and Goldbach [2] called such characters *admissible* and Frobenius rings are called *admissible rings*. Since in the literature of codes over rings, the nomenclature Frobenius ring and generating character became prevalent, we will continue this practice.

Example 2. The integer residue rings $\mathbb{Z}_m$, where $m \in \mathbb{N}$, are Frobenius (a generating character is given by $\chi(a) := \zeta^a$, where $\zeta \in \mathbb{C}$ is an m-th primitive root of unity). Every finite field is Frobenius (every non-principal character is a generating character). Further examples of Frobenius rings are finite chain rings, finite group rings over Frobenius rings, direct products of Frobenius rings, and Galois rings.

From now on let R be a finite, commutative Frobenius ring and let χ be a generating character of R. We will identify R and $\hat{R}$ via $r \mapsto r\chi$. This isomorphism extends to an R-isomorphism between R^n and its character module $\widehat{R^n} \cong \hat{R}^n$ given by $\alpha : R^n \to \widehat{R^n}$, $x \mapsto \chi(\langle x, \cdot \rangle)$, where $\langle x, y \rangle := \sum_{i=1}^n x_i y_i$ denotes the dot product on R^n.

The following properties have been derived by Claasen and Goldbach [2, Cor. 3.6], or are standard results that can easily be derived or be found in, e.g., [26, App. A].

Proposition 3. *The only ideal contained in* $\ker\chi := \{r \in R \mid \chi(r) = 1\}$ *is the zero ideal. Furthermore,* $\sum_{y\in R^n} \chi(\langle x,y\rangle) = 0$ *if* $x \neq 0$ *and* $\sum_{y\in R^n} \chi(\langle x,y\rangle) = |R^n|$ *if* $x = 0$.

Let us now turn to codes and their duals. Throughout, a code over R of length n will be a submodule of R^n. For a code $\mathcal{C} \subseteq R^n$ we define the *dual* as

$$\mathcal{C}^\perp = \{w \in R^n \mid \langle w, v\rangle = 0 \text{ for all } v \in \mathcal{C}\}.$$

It is an easy consequence of Proposition 3 that the dual code can also be described as $\mathcal{C}^\perp = \{w \in R^n \mid \chi(\langle w, v\rangle) = 1 \text{ for all } v \in \mathcal{C}\}$. In other words, the dual $\mathcal{C}^\perp$ coincides with the character-theoretic dual $\{\phi \in \widehat{R^n} \mid \phi(v) = 1 \text{ for all } v \in \mathcal{C}\}$.

The main tool for proving MacWilliams identities is the Poisson summation formula for maps on R^n and their Fourier transforms. Let V be any complex vector space and $f : R^n \to V$ be any map. Recall the generating character χ on R. In our setting, the *Fourier transform* of f simply becomes

$$f^+ : R^n \longrightarrow V, \quad v \longmapsto \sum_{w\in R^n} \chi(\langle v,w\rangle) f(w), \tag{1}$$

and the *Poisson summation formula* [24, p. 199] reads as

$$\sum_{w\in\mathcal{C}^\perp} f(w) = \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}} f^+(v) \tag{2}$$

for any code $\mathcal{C} \subseteq R^n$.

Our presentation will be based on the notion of an F-partition (Fourier-invariant partition), which has been introduced by Zinoviev and Ericson [28]. They define F-partitions to be those partitions for which the linear space generated by the indicator functions of the partition sets is invariant under the Fourier-transform. In [28, Lem. 1] it is shown that this equivalent to the following invariance, which we will use as our definition.

Definition 4. Let $\mathcal{Q} = (Q_m)_{m\in M}$ be a partition of R^n. Then $\mathcal{Q}$ is called an *F-partition* if for all $l, m \in M$ and all $x \in Q_m$ the sum $\sum_{y\in Q_l} \chi(\langle x,y\rangle)$ depends only on the indices l and m and not on the specific choice of $x \in Q_m$. For an F-partition $\mathcal{Q}$ we define the *generalized Krawtchouk coefficients* $k_{m,l}$ as

$$k_{m,l} = \sum_{y\in Q_l} \chi(\langle x,y\rangle), \text{ where } x \text{ is any element in } Q_m. \tag{3}$$

One should note the relation to the classical Krawtchouk polynomials $K_l^{(n,q)}(x)$ which satisfy $\sum_{w\in\mathbb{Z}_q^n, \mathrm{wt}(w)=l} \chi(\langle v,w\rangle) = K_l^{(n,q)}(m)$ for each vector $v \in \mathbb{Z}_q^n$ of Hamming weight $\mathrm{wt}(v) = m$; see [11, Lem. 2.6.2].

It is worth noting that the property of being an F-partition depends of the choice of the generating character.

The following result appeared first for the rings $\mathbb{Z}_N$ in [5]. It can easily be verified.

Proposition 5. *Let* $U \subseteq R^*$ *be a subgroup of the group of units of* R*. Then the partition* $\mathcal{P}$ *given by the orbits of the group action of* U *on* R *is an* F*-partition on* R*.*

We close this section with a first instance of a MacWilliams identity. It will be the basis for deriving all other identities. The full weight enumerator, defined in the following theorem, is simply a copy of the code within the polynomial ring $\mathbb{C}[X_v \mid v \in R^n]$. Evidently, it carries all information about the code and therefore fully determines the dual code, hence the dual full weight enumerator. The MacWilliams identity in (4) simply tells us the precise transformation between the two.

Theorem 6. *Let $\mathcal{C} \subseteq R^n$ be a code. Consider the "weight function"*

$$f : R^n \longrightarrow \mathbb{C}[X_v \mid v \in R^n],\ v \longmapsto X_v.$$

The polynomial $\mathrm{fwe}_{\mathcal{C}} := \sum_{v\in\mathcal{C}} X_v$ *is called the full weight enumerator of $\mathcal{C}$. It satisfies the MacWilliams Identity*

$$\mathrm{fwe}_{\mathcal{C}^\perp} = \frac{1}{|\mathcal{C}|}\mathcal{M}(\mathrm{fwe}_{\mathcal{C}}), \tag{4}$$

where the MacWilliams transform $\mathcal{M} : \mathbb{C}[X_v \mid v \in R^n] \longrightarrow \mathbb{C}[X_v \mid v \in R^n]$ is defined as the algebra homomorphism satisfying $\mathcal{M}(X_v) = \sum_{w\in R^n} \chi(\langle v, w\rangle)X_w$ for all $v \in R^n$.

Proof. The Poisson summation formula applied to the map f yields

$$\begin{aligned}\mathrm{fwe}_{\mathcal{C}^\perp} &= \sum_{w\in\mathcal{C}^\perp} f(w) = \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}} f^+(v) = \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}}\sum_{w\in R^n}\chi(\langle v,w\rangle)X_w \\ &= \frac{1}{|\mathcal{C}|}\sum_{v\in\mathcal{C}}\mathcal{M}(X_v) = \frac{1}{|\mathcal{C}|}\mathcal{M}(\mathrm{fwe}_{\mathcal{C}}).\end{aligned}$$ □

3 MacWilliams identities for composition enumerators

A partition on R induces two specific partitions on R^n: the product partition and the symmetrized partition. Both partitions give naturally rise to enumerators. We show that if the partition on R is an F-partition, then both these enumerators satisfy a MacWilliams identity. Examples will be presented in the next section.

Let $\mathcal{P} = \{P_1, \ldots, P_L\}$ be an F-partition on R. For $\alpha \in R$ denote by $[\alpha] := [\alpha]_{\mathcal{P}}$ the index of the partition set containing $\alpha \in R$.

Definition 7. (a) The *induced product partition* of R^n is defined as

$$\mathcal{P}^n := (P_{l_1} \times \ldots \times P_{l_n})_{(l_1,\ldots,l_n)\in\{1,\ldots,L\}^n}.$$

(b) The *composition vector* of $v = (v_1, \ldots, v_n) \in R^n$ is defined as

$$\mathrm{comp}_{\mathcal{P}}(v) = (s_1, \ldots, s_L), \text{ where } s_l = |\{t \mid v_t \in P_l\}|.$$

It is contained in the set $\mathcal{S} := \{(s_1, \ldots, s_L) \in \mathbb{N}_0^L \mid \sum_{l=1}^L s_l = n\}$. The *induced symmetrized partition* on R^n is defined as

$$\mathcal{P}^n_{\mathrm{sym}} = (Q_s)_{s\in\mathcal{S}}, \text{ where } Q_s = \{v \in R^n \mid \mathrm{comp}_{\mathcal{P}}(v) = s\}.$$

Note that the partition sets in the product partition $\mathcal{P}^n$ collect all vectors for which each entry is contained in a prescribed partition set, whereas the sets in the symmetrized partition contain all vectors that have the same number of entries (disregarding position) in a given partition set.

For a given code $\mathcal{C} \subseteq R^n$ we may now define two types of partition enumerators. The following two results show that both of them satisfy a MacWilliams identity. We start with the product partition. Recall the notation $[\alpha]$ for $\alpha \in R$.

Theorem 8. *Let $\mathcal{C} \subseteq R^n$ be a code. The polynomial* $\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}} := \sum_{v\in\mathcal{C}} \prod_{i=1}^n Y_{t,[v_t]}$, *contained in the polynomial ring*

$$\widetilde{V} := \mathbb{C}[Y_{t,j} \mid t = 1,\dots,n,\ j = 1,\dots,L],$$

is called the product partition enumerator *of $\mathcal{C}$ with respect to $\mathcal{P}$. The coefficient of $\prod_{t=1}^n Y_{t,l_t}$ equals the cardinality of $\mathcal{C} \cap (P_{l_1} \times \ldots \times P_{l_n})$. The product partition enumerator satisfies the MacWilliams Identity*

$$\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}^\perp} = \frac{1}{|\mathcal{C}|}\widetilde{\mathcal{M}}(\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}}), \tag{5}$$

where the MacWilliams transform $\widetilde{\mathcal{M}} : \widetilde{V} \longrightarrow \widetilde{V}$ is defined as the algebra homomorphism satisfying $\widetilde{\mathcal{M}}(Y_{t,[\alpha]}) = \sum_{\beta\in R}\chi(\alpha\beta)Y_{t,[\beta]} = \sum_{l=1}^L \sum_{\beta\in P_l}\chi(\alpha\beta)Y_{t,l}$ for all $t = 1,\dots,n$ and $\alpha \in R$. In particular, $\widetilde{\mathcal{M}}$ is well-defined.

Proof. First of all, notice that $\sum_{\beta\in R}\chi(\alpha\beta)Y_{t,[\beta]} = \sum_{l=1}^L \sum_{\beta\in P_l}\chi(\alpha\beta)Y_{t,l}$. By Definition 4, the coefficient $\sum_{\beta\in P_l}\chi(\alpha\beta)$ does not depend on the choice of α in its partition set $P_{[\alpha]}$, and this establishes the well-definedness of $\widetilde{\mathcal{M}}$.

Consider now the situation of Theorem 6, and let $\phi : \mathbb{C}[X_v \mid v \in R^n] \longrightarrow \widetilde{V}$ be the substitution homomorphism defined via $\phi(X_v) = \prod_{t=1}^n Y_{t,[v_t]}$. Using the group homomorphism property of the character χ one computes

$$\phi \circ \mathcal{M}(X_v) = \sum_{w\in R^n} \chi(\langle v,w\rangle) \prod_{t=1}^n Y_{t,[w_t]} = \prod_{t=1}^n \sum_{\beta\in R} \chi(v_t\beta)Y_{t,[\beta]} = \widetilde{\mathcal{M}} \circ \phi(X_v).$$

Now Theorem 6 implies $\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}^\perp} = \phi(\mathrm{fwe}_{\mathcal{C}^\perp}) = \frac{1}{|\mathcal{C}|}\phi \circ \mathcal{M}(\mathrm{fwe}_{\mathcal{C}}) = \frac{1}{|\mathcal{C}|}\widetilde{\mathcal{M}}(\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}})$, as desired. □

Notice that we may write the identity (5) in the form

$$\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}^\perp}(Y_{t,l} \mid t = 1,\dots,n, l = 1,\dots,L) = \frac{1}{|\mathcal{C}|}\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}}(K\mathbf{Y}_t \mid t = 1,\dots,n),$$

where $\mathbf{Y}_t = (Y_{t,1},\dots,Y_{t,L})^\mathsf{T}$ and $K = (k_{m,l}) \in \mathbb{C}^{L\times L}$ is the Krawtchouk matrix of the partition $\mathcal{P}$ with entries defined in (3).

Next we present the MacWilliams identity for induced symmetrized partitions.

Theorem 9. *For a code $\mathcal{C}\subseteq R^n$ we define the* symmetrized partition enumerator *with respect to $\mathcal{P}$ as $\mathrm{PE}_{\mathcal{P}^n_{sym},\mathcal{C}} := \sum_{v\in\mathcal{C}}\prod_{t=1}^n Z_{[v_t]}$. It is a homogeneous polynomial of degree n in the polynomial ring $\widehat{V} := \mathbb{C}[Z_j \mid j=1,\ldots,L]$. The coefficient of the monomial $\prod_{j=1}^L Z_j^{s_j}$ in $\mathrm{PE}_{\mathcal{P}^n_{sym},\mathcal{C}}$ equals the cardinality $|\{v\in\mathcal{C}\mid comp_{\mathcal{P}}(v)=(s_1,\ldots,s_L)\}|$. The symmetrized partition enumerator satisfies the MacWilliams Identity*

$$\mathrm{PE}_{\mathcal{P}^n_{sym},\mathcal{C}^\perp} = \frac{1}{|\mathcal{C}|}\widehat{\mathcal{M}}(\mathrm{PE}_{\mathcal{P}^n_{sym},\mathcal{C}}), \tag{6}$$

where the MacWilliams transform $\widehat{\mathcal{M}}:\widehat{V}\longrightarrow\widehat{V}$ is the (well-defined) algebra homomorphism given by $\widehat{\mathcal{M}}(Z_{[\alpha]}) = \sum_{\beta\in R}\chi(\alpha\beta)Z_{[\beta]} = \sum_{l=1}^L\sum_{\beta\in P_l}\chi(\alpha\beta)Z_l$ for all $\alpha\in R$.

Proof. Again, the well-definedness of $\widehat{\mathcal{M}}$ follows from the fact that $\mathcal{P}$ is an F-partition on R. Let $\widetilde{V}$ and $\widetilde{\mathcal{M}}$ be as in Theorem 8 and consider the substitution homomorphism $\psi:\widetilde{V}\longrightarrow\widehat{V}$ given by $\psi(Y_{t,j})=Z_j$. Then

$$\psi\circ\widetilde{\mathcal{M}}(Y_{t,[\alpha]}) = \psi\Big(\sum_{\beta\in R}\chi(\alpha\beta)Y_{t,[\beta]}\Big) = \sum_{\beta\in R}\chi(\alpha\beta)Z_{[\beta]} = \widehat{\mathcal{M}}\circ\psi(Y_{t,[\alpha]})$$

for all $Y_{t,[\alpha]}$. Now Theorem 8 yields $\mathrm{PE}_{\mathcal{P}^n_{\mathrm{sym}},\mathcal{C}^\perp} = \psi(\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}^\perp}) = \frac{1}{|\mathcal{C}|}\psi\circ\widetilde{\mathcal{M}}(\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}})$ $= \frac{1}{|\mathcal{C}|}\widehat{\mathcal{M}}\circ\psi(\mathrm{PE}_{\mathcal{P}^n,\mathcal{C}}) = \frac{1}{|\mathcal{C}|}\widehat{\mathcal{M}}(\mathrm{PE}_{\mathcal{P}^n_{\mathrm{sym}},\mathcal{C}})$. □

Just as for the product partition we may write the identity (6) in the form

$$\mathrm{WE}_{\mathcal{P}^n_{\mathrm{sym}},\mathcal{C}^\perp}(Z_1,\ldots,Z_L) = \frac{1}{|\mathcal{C}|}\mathrm{WE}_{\mathcal{P}^n_{\mathrm{sym}},\mathcal{C}}(K(Z_1,\ldots,Z_L)^{\mathsf{T}}),$$

where again $K=(k_{m,l})\in\mathbb{C}^{L\times L}$ is the Krawtchouk matrix of the partition $\mathcal{P}$. For codes over fields this identity appeared already in MacWilliams and Sloane [18, Ch. 5, Thm. 10].

4 Examples

The MacWilliams identities for symmetrized partitions in Theorem 9 lead to the best known examples. Therefore, we cover these first and start with the most famous one.

Example 10. Consider the partition $\{0\}\cup(R\backslash\{0\})$ on R. This is indeed an F-partition as follows immediately from Proposition 3. For the resulting symmetrized partition on R^n the composition vector as defined in Definition 7(b) is $\mathrm{comp}(v)=(n-\mathrm{wt}(v),\mathrm{wt}(v))$, where $\mathrm{wt}(v)$ denotes the *Hamming weight* of $v\in R^n$. Hence the induced symmetrized partition on R^n is simply $(Q_l)_{l=0,\ldots,n}$, where $Q_l=\{v\in R^n\mid \mathrm{wt}(v)=l\}$, and the symmetrized partition enumerator of a code $\mathcal{C}$ in R^n is the classical *Hamming weight enumerator* $\mathrm{hwe}_{\mathcal{C}} = \sum_{v\in\mathcal{C}} Z_0^{n-\mathrm{wt}(v)}Z_1^{\mathrm{wt}(v)}$. Proposition 3 along with Theorem 9 show that the MacWilliams transform amounts to $Z_0\mapsto Z_0+(|R|-1)Z_1$, $Z_1\mapsto Z_0-Z_1$, and thus we have the familiar identity

$$\mathrm{hwe}_{\mathcal{C}^\perp}(Z_0,Z_1) = \frac{1}{|\mathcal{C}|}\mathrm{hwe}_{\mathcal{C}}(Z_0+(|R|-1)Z_1, Z_0-Z_1).$$

For fields, this identity is the classical result of MacWilliams [17]. For codes over the ring $\mathbb{Z}_4$ it has been derived by Hammons et al. [7, p. 303] (see also the references therein), for arbitrary residue rings $\mathbb{Z}_m$ by Klemm [14], for arbitrary finite Frobenius rings by Nechaev and Kuzmin [20], and finally for non-commutative finite Frobenius rings by Wood [26, Thm. 8.3].

For the following examples, recall from Proposition 5 that each subgroup $U \subseteq R^*$ gives rise to an F-partition.

Example 11. Let $U = \{1\}$ be the trivial group. Then U induces the partition $\mathcal{P}$ consisting of the singletons $\{a\}, a \in R$. The resulting symmetrized partition enumerator is the *complete weight enumerator* $\text{cwe}_{\mathcal{C}} := \sum_{v \in \mathcal{C}} \prod_{t=1}^{n} Z_{v_t} \in \mathbb{C}[Z_\alpha \mid \alpha \in R]$. The coefficient of $\prod_{\alpha \in R} Z_\alpha^{s_\alpha}$ is the number of codewords having exactly s_α entries equal to α. The MacWilliams identity is given by $\text{cwe}_{\mathcal{C}^\perp}(\mathbf{Z}) = \frac{1}{|\mathcal{C}|}\text{cwe}_{\mathcal{C}}(K_c\mathbf{Z})$, where $\mathbf{Z} = (z_\alpha \mid \alpha \in R)^{\mathsf{T}}$ and $K_c = \big(\chi(\alpha\beta)\big)_{\alpha,\beta \in R}$. For codes over fields, the identity appears already in the textbook [18] by MacWilliams and Sloane.
Consider the following two special cases.
1) Let $R = \mathbb{F}_4 = \{0, 1, a, a+1\}$, where $a^2 = a+1$. A generating character of $\mathbb{F}_4$ is given by $\chi(0) = \chi(1) = 1$ and $\chi(a) = \chi(a^2) = -1$. Thus,

$$K_c = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix},$$

and the MacWilliams identity for the complete weight enumerator reads as

$$\begin{aligned} \text{cwe}_{\mathcal{C}^\perp}(Z_0, Z_1, Z_a, Z_{a^2}) = \frac{1}{|\mathcal{C}|}\text{cwe}_{\mathcal{C}}(&Z_0 + Z_1 + Z_a + Z_{a^2}, Z_0 + Z_1 - Z_a - Z_{a^2}, \\ &Z_0 - Z_1 - Z_a + Z_{a^2}, Z_0 - Z_1 + Z_a - Z_{a^2}). \end{aligned} \tag{7}$$

2) For $R = \mathbb{Z}_4$ a generating character is given by $\chi(a) = i^a$ for $a = 0, \ldots, 3$, and the MacWilliams identity is

$$\begin{aligned} \text{cwe}_{\mathcal{C}^\perp}(Z_0, Z_1, Z_2, Z_3) = \frac{1}{|\mathcal{C}|}\text{cwe}_{\mathcal{C}}(&Z_0 + Z_1 + Z_2 + Z_3, Z_0 + iZ_1 - Z_2 - iZ_3, \\ &Z_0 - Z_1 + Z_2 - Z_3, Z_0 - iZ_1 - Z_2 + iZ_3). \end{aligned} \tag{8}$$

It appeared in [7, p. 303] as well as a special case of [14, Satz 1.2].

Example 12. Let $R = \mathbb{Z}_m$ for some $m \in \mathbb{N}$. Put $U = \{1, -1\}$. Then the orbits of the action of U on R are given by $P_0 = \{0\}$ and $P_a = \{a, -a\}$ (which may be a singleton). By Proposition 5 the orbits form an F-partition $\mathcal{P}$. It consists of $L := \lfloor m/2 \rfloor$ nonzero sets. The induced symmetrized partition enumerator on R^n is called the *symmetrized Lee weight enumerator*; thus $\text{slwe}_{\mathcal{C}} := \text{WE}_{\mathcal{P}^n_{\text{sym}}, \mathcal{C}} = \sum_{v \in \mathcal{C}} \prod_{t=1}^{n} Z_{[v_t]}$, where, as usual, $[v_t]$ is the index of the partition set $\{v_t, -v_t\}$. It enumerates the codewords having the same coordinates up to sign and ordering. Theorem 9 provides the according

MacWilliams identity.
Consider for example, the ring $R = \mathbb{Z}_4$. We may choose again $\chi(a) = i^a$, $a = 0,\dots,3$. In this case we have the partition sets $P_0 = \{0\}, P_1 = \{1,3\}, P_2 = \{2\}$ and obtain the transform $\widetilde{\mathcal{M}}(Z_l) = Z_0 + (\chi(l) + \chi(-l))Z_1 + \chi(2l)Z_2$ for $l = 0,1,2$. This results in the identity

$$\text{slwe}_{\mathcal{C}^\perp}(Z_0,Z_1,Z_2) = \frac{1}{|\mathcal{C}|}\text{slwe}_{\mathcal{C}}(Z_0+2Z_1+Z_2, Z_0-Z_2, Z_0-2Z_1+Z_2), \tag{9}$$

as it has been presented already in [7, p. 303] as well as in [12, Satz 1.2] as a special case.

Example 13. (1) This example has been studied by Klemm [12]. It generalizes (9) in a particular way. Consider $R = \mathbb{Z}_m$ and let $U = \mathbb{Z}_m^*$. Then the orbits of U in R are $P_d := \{a \in \mathbb{Z}_m \mid \gcd(a,m) = d\}$ for all divisors d of m. Note that $P_1 = U$ and $P_m = \{0\}$. Hence the coefficient of $\prod_{d|m} Z_d^{s_d}$ in the symmetrized partition enumerator equals the number of codewords having exactly s_d entries with additive order md^{-1}. The MacWilliams identity in Theorem 9 tells us that this information about the code fully determines the same information of the dual code.
(2) This example appeared in [10] by Huber. Let $R = \mathbb{F}_q = \mathbb{F}_{p^m}$ be a field of odd characteristic p and such that m is even if $p \equiv 3 \mod 4$. Then $(q-1)/4 \in \mathbb{Z}$ and thus there exists an element i in $\mathbb{F}_q$ such that $i^2 = -1$. Define $U := \langle i \rangle = \{1,-1,i,-i\}$. Its orbits in $\mathbb{F}_q$ form an F-partition $\mathcal{P}$ consisting of $L := (q-1)/4+1$ sets. For a vector $v \in \mathbb{F}_q^n$, the composition vector $\text{comp}_{\mathcal{P}}(v) \in \mathbb{N}_0^L$ counts the number of entries of v in each orbit, see Definition 7(b). It is called the *Gaussian weight* in [10]. Theorem 9 provides a MacWilliams identity for the resulting partition enumerator, which has already been presented in [10, Thm. 2].

Let us now turn to examples for the MacWilliams identity for product partition enumerators as derived in Theorem 8. We obtain the well-known identity for the exact weight enumerator as well as some other, lesser known, identities.

Example 14. This is the de-symmetrized version of the complete weight enumerator discussed in Example 11. Let U be the trivial group $\{1\}$, which induces the partition $\mathcal{P}$ consisting of the singletons $\{a\}$. The resulting product partition enumerator $\text{PE}_{\mathcal{P}^n,\mathcal{C}}$ is the *exact weight enumerator* $\text{ewe}_{\mathcal{C}} := \sum_{v\in\mathcal{C}} \prod_{t=1}^n Y_{t,v_t} \in \mathbb{C}[Y_{t,\alpha} \mid t = 1,\dots,n,\ \alpha \in R]$. The monomials of this polynomial are in bijection to the codewords. Just like the full weight enumerator in Theorem 6, the exact weight enumerator carries all information about the code (this time, the information is encoded in a polynomial in $n|R|$ indeterminates, whereas the full weight enumerator is a polynomial in $|R|^n$ indeterminates). It is thus clear that the exact weight enumerators of the code and its dual must determine each other, and the MacWilliams identity from Theorem 8 simply makes this explicit. The associated MacWilliams transform is given by $\widetilde{\mathcal{M}}(Y_{t,\alpha}) = \sum_{\beta\in R}\chi(\alpha\beta)Y_{t,\beta}$ for all t. For codes over fields this identity appears already in [18, Ch. 5, Thm. 14] by MacWilliams and Sloane.

The next examples cover in particular the Lee weight.

Example 15. 1) This is the de-symmetrized version of Example 12. Let R, U and the partition $\mathcal{P}$ be as in Example 12. The product partition enumerator is based on the weight function given by $v \mapsto \prod_{t=1}^{n} Z_{t,[v_t]} \in \mathbb{C}[Z_{t,l} \mid t = 1,\ldots,n, l = 0,\ldots,L]$, where L is the number of nonzero U-orbits. The enumerator keeps track of the entries of v up to sign, but including their position t.
As a special case, let $R = \mathbb{Z}_m$ and $\chi : \mathbb{Z}_m \to \mathbb{C}, a \mapsto \zeta^a$, where $\zeta \in \mathbb{C}$ is a primitive m-th root of unity. We call the resulting product partition enumerator $\text{plwe}_{\mathcal{C}}$ the *Product Lee Weight Enumerator*. The coefficient of a monomial $\prod_{t=1}^{n} Z_{t,l_t}$ equals the number of all codewords for which the t-th entry is $\pm l_t$. According to Theorem 8, $\text{plwe}_{\mathcal{C}}$ and $\text{plwe}_{\mathcal{C}^\perp}$ satisfy a MacWilliams identity with MacWilliams transform given by

$$\widetilde{M}(Z_{t,l}) = \begin{cases} Z_{t,0} + \sum_{s=1}^{L}\left(\zeta^{l\cdot s} + \zeta^{-l\cdot s}\right)Z_{t,s} & \text{if } m \text{ is odd} \\ Z_{t,0} + \sum_{s=1}^{L-1}\left(\zeta^{l\cdot s} + \zeta^{-l\cdot s}\right)Z_{t,s} + \zeta^{l\cdot L}Z_{t,L} & \text{if } m \text{ is even} \end{cases}$$

2) In the same way there exists a de-symmetrized version of Example 13(1). The analogous product partition enumerator keeps track of the additive order of each individual codeword coordinate. Again the MacWilliams identity tells us that this information fully determines the same type of information for the dual code.

The following is the de-symmetrized version of the Hamming weight.

Example 16. Consider again the F-partition $\{0\} \cup (R\backslash\{0\})$ on R as in Example 10. The induced product partition on R^n leads to the weight function given by $v \mapsto \prod_{t=1}^{n} S_{t,[v_t]}$, where $[0] = 0$ and $[a] = 1$ for $a \neq 0$. As a consequence, the product partition enumerator is the *Support Tracker* $\text{supp-we}_{\mathcal{C}} := \sum_{v\in\mathcal{C}} \prod_{t=1}^{n} S_{t,[v_t]}$. The coefficient of the monomial $\prod_{t=1}^{n} S_{t,l_t}$ enumerates the codewords having support equal to the set $\{t \mid l_t = 1\} \subseteq \{1,\ldots,n\}$. The MacWilliams identity obtained from Theorem 8 has been discussed already by Zinoviev and Ericson [28, Ex. 3] and Honold and Landjev [9, Ex. 23].

Example 17. This example does not immediately fit into our setting, but can be dealt with via a simple adjustment. Let $\mathbb{F}$ be a finite field. For a nonzero vector $v = (v_1,\ldots,v_n) \in \mathbb{F}^n$ define $\rho(v) := \max\{i \mid v_i \neq 0\}$ and put $\rho(0) = 0$. Then ρ induces a metric on $\mathbb{F}^n$, the *Rosenbloom-Tsfasman weight*, which has been introduced by Rosenbloom and Tsfasman in [21]. This metric plays a specific role for matrices, to which it generalizes straightforwardly by taking the sum of $\rho(x)$ over all rows x of the matrix (see also [3]). For the relevance of the Rosenbloom-Tsfasman weight for detecting matrix codes with large Hamming distance see [23].

We will now derive a MacWilliams identity for the Rosenbloom-Tsfasman enumerator of codes in $\mathbb{F}^n$ and their reversed dual. For $i = 0,\ldots,n$ let $P_i = \{v \in \mathbb{F}^n \mid \rho(v) = i\}$, and let $\mathcal{P}$ be the partition $(P_i)_{i=0}^{n}$ of $\mathbb{F}^n$. Note that $\mathcal{P}$ is not induced by a partition on $\mathbb{F}$ as in Definition 7. For a code $\mathcal{C} \subseteq \mathbb{F}^n$ define the corresponding Rosenbloom-Tsfasman enumerator $\text{RT}_{\mathcal{C}} = \sum_{i=0}^{n} |\mathcal{C} \cap P_i| Z_i$. Then $\text{RT}_{\mathcal{C}} \in \mathbb{C}[Z_0,\ldots,Z_n]_{\text{hom},1}$, the space of all homogeneous polynomials of degree 1 in $n+1$ variables.

Examples show immediately that this enumerator does not satisfy a MacWilliams identity for a code $\mathcal{C}$ and its dual $\mathcal{C}^\perp$. However, if one redefines the dual by applying a coordinate reversal, then a MacWilliams identity can be derived. Define

$$J := \begin{bmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{bmatrix} \in \mathbb{F}^{n\times n}. \tag{10}$$

Then $J = J^{-1} = J^{\mathsf{T}}$. The partition $\mathcal{P}$ has the following invariance property. First of all, it is not hard to see that the sets P_i form the orbits of the group action of the invertible lower triangular matrices on $\mathbb{F}^n$. Next, let $v, v' \in P_i$ for some i, thus $v' = vA$ for some invertible lower triangular matrix A. Notice that $JA^{\mathsf{T}}J$ is again lower triangular. Hence, $P_j(JA^{\mathsf{T}}J) = P_j$ for each partition set P_j, and one easily derives

$$\sum_{w\in P_j} \chi(\langle w, v'J\rangle) = \sum_{w\in P_j} \chi(\langle w, vJ\rangle).$$

As a consequence, the (generalized) Krawtchouk coefficients $k_{ij} = \sum_{w\in P_j} \chi(\langle w, vJ\rangle)$, where $v \in P_i$, depend only on i, j, and not on the specific choice of $v \in P_i$.

For a code $\mathcal{C} \subseteq \mathbb{F}^n$ define the reversed dual code $\mathcal{C}^{\underline{\perp}} := \mathcal{C}^{\perp}J := \{wJ \mid w \in \mathcal{C}^{\perp}\}$. Now it is easy to derive a MacWilliams identity between $\mathrm{RT}_{\mathcal{C}}$ and $\mathrm{RT}_{\mathcal{C}^{\underline{\perp}}}$. Consider again the full weight enumerator from Theorem 6, and for $v \in \mathbb{F}^n$ let $\tau(X_v) = Z_{\rho(v)}$. As usual, we extend τ to an algebra homomorphism on $\mathbb{C}[X_v \mid v \in \mathbb{F}^n]$. Using the MacWilliams transform $\mathcal{M}$ from Theorem 6 one obtains

$$\tau \circ \mathcal{M}(X_{vJ}) = \sum_{w\in\mathbb{F}^n} \chi(\langle w, vJ\rangle) Z_{\rho(w)} = \sum_{j=0}^{n} \sum_{w\in P_j} \chi(\langle w, vJ\rangle) Z_j = \mathcal{M}' \circ \tau(X_v),$$

where the transform $\mathcal{M}'$ on $\mathbb{C}[Z_0, \ldots, Z_n]_{\mathrm{hom},1}$ is defined as the vector space isomorphism given by $\mathcal{M}'(Z_i) = \sum_{j=0}^{n} k_{ij} Z_j$. Note that this is a linear map. Along with Theorem 6 all of this now leads to the MacWilliams Identity

$$\begin{aligned} \mathrm{RT}_{\mathcal{C}^{\underline{\perp}}} = \tau(\mathrm{fwe}_{\mathcal{C}^{\perp}J}) &= \frac{1}{|\mathcal{C}|} \tau \circ \mathcal{M}(\mathrm{fwe}_{(\mathcal{C}^{\perp}J)^{\perp}}) = \frac{1}{|\mathcal{C}|} \tau \circ \mathcal{M}(\mathrm{fwe}_{\mathcal{C}J}) \\ &= \frac{1}{|\mathcal{C}|} \mathcal{M}' \circ \tau(\mathrm{fwe}_{\mathcal{C}}) = \frac{1}{|\mathcal{C}|} \mathcal{M}'(\mathrm{RT}_{\mathcal{C}}). \end{aligned}$$

The Krawtchouk coefficients for the transform can be computed explicitly. Considering $\langle w, vJ\rangle$ and making use of the properties in Proposition 3, one derives

$$k_{i,j} = \begin{cases} 1 & \text{if } j = 0 \\ |P_j| = q^{j-1}(q-1) & \text{if } 1 \le j \le n-i \\ -q^{n-i} & \text{if } j = n-i+1 \\ 0 & \text{if } j > n-i+1 \end{cases}$$

Thus the MacWilliams transform reads as

$$\mathcal{M}'(Z_i) = Z_0 + \sum_{j=1}^{n-i} q^{j-1}(q-1) Z_j - q^{n-i} Z_{n-i+1}. \tag{11}$$

This has also been derived in [3, Thm. 3.1]. We come back to this in Example 21, when it will be extended to the space of matrices $\mathbb{F}^{s\times n}$.

5 Split partition enumerators

In this section we describe how the previous results generalize to MacWilliams identities for codes in $R^{n_1}\times R^{n_2}$, where one considers different induced partitions on R^{n_1} and R^{n_2}. The ideas generalize straightforwardly to any finite number of factors. Several of such cases for split enumerators have been investigated in the literature. They are discussed below. The main idea of this section has been used already in [18, Ch. 5, § 6] for the split Hamming weight enumerator.

Suppose $\mathcal{P}$ and $\mathcal{Q}$ are partitions of R. Moreover, let $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{Q}}$ be induced partitions on R^{n_1} and R^{n_2} in the sense of Definition 7, respectively. Thus, either one may be the induced product or symmetrized partition. Write $\tilde{\mathcal{P}}=(P_l)_{l=1}^{L}$ and $\tilde{\mathcal{Q}}=(Q_m)_{m=1}^{M}$, and for $v\in R^{n_1}$ (resp. $v\in R^{n_2}$) denote by $[v]$ the index of the partition set in $\tilde{\mathcal{P}}$ (resp. $\tilde{\mathcal{Q}}$) containing v. Let the resulting enumerating functions for $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{Q}}$ take values in $\mathbb{C}[\mathbf{Z}]$ and $\mathbb{C}[\mathbf{T}]$, respectively, where $\mathbf{Z}$ and $\mathbf{T}$ are appropriate lists of indeterminates; see Theorems 8 and 9. For any $v\in R^{n_1}$ let $\mathbf{Z}_{[v]}$ denote the monomial associated with the partition set $P_{[v]}$ and similarly for the vectors in R^{n_2}. Then the partition enumerators are given by $\sum_{v\in\mathcal{C}}\mathbf{Z}_{[v]}$ for codes $\mathcal{C}$ in R^{n_1} and $\sum_{v\in\mathcal{C}}\mathbf{T}_{[v]}$ for codes in R^{n_2}.

Definition 18. The *split partition enumerator* of a code $\mathcal{C}\subseteq R^{n_1}\times R^{n_2}$ with respect to the partition $\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}}$ is defined as

$$\mathrm{SPE}_{\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}},\mathcal{C}}=\sum_{(v,w)\in\mathcal{C}}\mathbf{Z}_{[v]}\mathbf{T}_{[w]}\in\mathbb{C}[\mathbf{Z},\mathbf{T}].$$

The coefficient of a monomial $\mathbf{Z}_l\mathbf{T}_m$ equals the cardinality $|\mathcal{C}\cap P_l\times Q_m|$.

It is not hard to see [28, Thm. 3] that if both $\mathcal{P}$ and $\mathcal{Q}$ are F-partitions on R, then $\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}}$ is an F-partition on $R^{n_1}\times R^{n_2}$. As a consequence, there is a MacWilliams identity between the split partition enumerators of a code and its dual. We can make this identity precise by combining the previous results.

From Theorems 8 and 9 we know that there exist transforms $\mathcal{M}_1:\mathbb{C}[\mathbf{Z}]\to\mathbb{C}[\mathbf{Z}]$ and $\mathcal{M}_2:\mathbb{C}[\mathbf{T}]\to\mathbb{C}[\mathbf{T}]$ such that

$$\mathrm{PE}_{\tilde{\mathcal{P}},\mathcal{C}_1^\perp}=\frac{1}{|\mathcal{C}_1|}\mathcal{M}_1\big(\mathrm{PE}_{\tilde{\mathcal{P}},\mathcal{C}_1}\big)\ \text{ and }\ \mathrm{PE}_{\tilde{\mathcal{Q}},\mathcal{C}_2^\perp}=\frac{1}{|\mathcal{C}_2|}\mathcal{M}_2\big(\mathrm{PE}_{\tilde{\mathcal{Q}},\mathcal{C}_2}\big)$$

for all codes $\mathcal{C}_i\subseteq R^{n_i}$, $i=1,2$. Now we can formulate

Theorem 19. *Define the transform $\bar{\mathcal{M}}:\mathbb{C}[\mathbf{Z},\mathbf{T}]\to\mathbb{C}[\mathbf{Z},\mathbf{T}]$ as the $\mathbb{C}$-algebra homomorphism given by $\mathbf{Z}_l\mathbf{T}_m\mapsto\mathcal{M}_1(\mathbf{Z}_l)\mathcal{M}_2(\mathbf{T}_m)$. Then we have the MacWilliams identity*

$$\mathrm{SPE}_{\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}},\mathcal{C}^\perp}=\frac{1}{|\mathcal{C}|}\bar{\mathcal{M}}\big(\mathrm{SPE}_{\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}},\mathcal{C}}\big)$$

for each code $\mathcal{C}\subseteq R^{n_1}\times R^{n_2}$.

Sketch of Proof: We start again with the MacWilliams identity for the full weight enumerator given in Theorem 6. From the proofs of Theorems 8 and 9 we know that $\phi_i\circ\mathcal{M}=\mathcal{M}_i\circ\phi_i$ for $i=1,2$, where $\mathcal{M}$ is as in Theorem 6 and where

$$\phi_1:\mathbb{C}[X_v\mid v\in R^{n_1}]\longrightarrow\mathbb{C}[\mathbf{Z}]\text{ and }\phi_2:\mathbb{C}[X_w\mid w\in R^{n_2}]\longrightarrow\mathbb{C}[\mathbf{T}]$$

are given by $\phi_1(X_v) = \mathbf{Z}_{[v]}$ and $\phi_2(X_w) = \mathbf{T}_{[w]}$. Furthermore, define

$$\phi : \mathbb{C}[X_{(v,w)} \mid (v,w) \in R^{n_1} \times R^{n_2}] \longrightarrow \mathbb{C}[\mathbf{Z}, \mathbf{T}], \quad X_{(v,w)} \longmapsto \mathbf{Z}_{[v]}\mathbf{T}_{[w]}.$$

Using the group homomorphism property of the character χ it is not hard to see that

$$\phi\big(\mathcal{M}(X_{(v,w)})\big) = \mathcal{M}_1(\phi_1(X_v))\mathcal{M}_2(\phi_2(X_w)) = \bar{\mathcal{M}} \circ \phi(X_{(v,w)}).$$

Now the identity in (4) along with $\mathrm{SPE}_{\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}},\mathcal{C}} = \phi(\mathrm{fwe}_{\mathcal{C}})$ leads to the desired identity. □

This result covers several results known from the literature.

Example 20. (1) Consider the Hamming weight on both R^{n_1} and R^{n_2}. In this case we obtain the simple the identity

$$\begin{aligned}&\mathrm{SPE}_{\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}},\mathcal{C}^\perp}(W_0,W_1,Z_0,Z_1)\\ &\quad= \frac{1}{|\mathcal{C}|}\mathrm{SPE}_{\tilde{\mathcal{P}}\times\tilde{\mathcal{Q}},\mathcal{C}}(W_0+(|R|-1)W_1, W_0-W_1, Z_0+(|R|-1)Z_1, Z_0-Z_1), \quad (12)\end{aligned}$$

where the coefficient of $W_1^iW_0^{n_1-i}Z_1^jZ_0^{n_2-j}$ equals the number of codewords having Hamming weight i on the first n_1 coordinates and Hamming weight j on the last n_2 coordinates. This identity has been derived already by MacWilliams and Sloane for codes over the binary field in [18, Ch. 5, Eq. (52)] and by Simonis [22, Eq. (3')], where the codewords are divided into t blocks of coordinates. A similar identity can be found in [4] by El-Khamy and McEliece. The latter authors also observe that if $\mathcal{C}$ is a systematic $[n,k]$ code, then the split weight enumerator for the Hamming weight on both parts is the input-redundancy weight enumerator which keeps track of the input weights in combination with the corresponding redundancy weight. This allows them to apply their identity to MDS codes in order to derive further results on the bit error probability for systematic RS codes. In [16] this weight enumerator has been used to derive a MacWilliams identity for the input-output weight enumerators of direct-product single-partity-check codes.
(2) One should note that the support-tracker discussed in Example 16 as well as the Product Lee Weight Enumerator in Example 15 are special cases of the split weight enumerator, where we partition the codewords into n blocks of length 1.

We close this note with the Rosenbloom-Tsfasman weight for matrix codes. The following result can also be found in [3] by Dougherty and Skriganov.

Example 21. Recall the Rosenbloom-Tsfasman metric ρ from Example 17. Consider the vector space $\mathbb{F}^{s\times n}$ of all $s\times n$-matrices over the field $\mathbb{F} = \mathbb{F}_q$. Denote the rows of a matrix $M \in \mathbb{F}^{s\times n}$ by $M_1,\ldots,M_s \in \mathbb{F}^n$. For $r := (r_1,\ldots,r_s) \in \mathcal{N} := \{0,\ldots,n\}^s$ define the set $P_r := \{M \in \mathbb{F}^{s\times n} \mid \rho(M_i) = r_i, i = 1,\ldots,s\}$. Then $(P_r)_{r\in\mathcal{N}}$ forms a partition on $\mathbb{F}^{s\times n}$. It is the direct product of the Rosenbloom-Tsfasman partition on $\mathbb{F}^n$ extended to $\mathbb{F}^n \times\ldots\times \mathbb{F}^n$. Define the split weight enumerator of a code $\mathcal{C} \subseteq \mathbb{F}^{s\times n}$ with respect to the RT-metric on each factor $\mathbb{F}^n$ as

$$\mathrm{SPE}_{\mathcal{C}} := \sum_{M\in\mathcal{C}}\prod_{i=1}^{s} Z_{i,\rho(M_i)} \in \mathbb{C}[Z_{1,0},\ldots,Z_{1,n},\ldots,Z_{s,0},\ldots,Z_{s,n}].$$

Then $\mathrm{SPE}_{\mathcal{C}}$ is a homogeneous polynomial of degree s, and for $r := (r_1,\ldots,r_s) \in \mathcal{N}$ the coefficient of $\prod_{i=1}^{s} Z_{i,r_i}$ equals the number of codewords in $\mathcal{C} \cap P_r$. In [3, Sec. 3] this enumerator has been coined the T-enumerator.

For a code $\mathcal{C} \subseteq \mathbb{F}^{s\times n}$ we define the reversed dual as

$$\mathcal{C}^{\perp} := \{B \in \mathbb{F}^{s\times n} \mid \textstyle\sum_{i=1}^{s} \langle B_i, A_i J\rangle = 0 \text{ for all } A \in \mathcal{C}\},$$

where the matrix J is as in (10); see also [3, p. 83]. The same line of reasoning as in the proof of Theorem 19 along with the MacWilliams identity derived in Example 17 leads to the identity

$$\mathrm{SPE}_{\mathcal{C}^{\perp}} = \frac{1}{|\mathcal{C}|}\mathcal{M}'(\mathrm{SPE}_{\mathcal{C}}),$$

where $\mathcal{M}'(\prod_{i=1}^{s} Z_{i,r_i}) = \prod_{i=1}^{s}(Z_{i,0} + \sum_{j=1}^{n-r_i} q^{j-1}\big(q-1)Z_{i,j} - q^{n-r_i}Z_{i,n-r_i+1}\big)$, see (11). This reproduces Theorem 3.1 in [3].

It is worth mentioning that the cumulative Rosenbloom-Tsfasman weight $\rho(M) := \sum_{i=1}^{s} \rho(M_i)$ does not satisfy a MacWilliams identity. In [3] a pair of codes with the same cumulative Rosenbloom-Tsfasman weight enumerator are given, and where the reversed dual codes have different enumerators.

Acknowledgments

The author was partially supported by National Science Foundation grant #DMS-0908379.

Bibliography

[1] E. Byrne, M. Greferath, and M. E. O'Sullivan. The linear programming bound for codes over finite Frobenius rings. *Des. Codes Cryptography*, 42:289–301, 2007. Cited p. 167.

[2] H. L. Claasen and R. W. Goldbach. A field-like property of finite rings. *Indag. Math.*, 3:11–26, 1992. Cited p. 169.

[3] S. Dougherty and M. Skriganov. MacWilliams duality and the Rosenbloom-Tsfasman metric. *Moscow Mathematical Journal*, 2(1):81–97, 2002. Cited pp. 176, 177, 179, and 180.

[4] M. El-Khamy and R. J. McEliece. The partition weight enumerator of MDS codes and its applications. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 926–930, 2005. Cited p. 179.

[5] T. Ericson, J. Simonis, H. Tarnanen, and V. Zinoviev. F-partitions of cyclic groups. *Appl. Algebra Engrg. Comm. and Comput.*, 8:387–393, 1997. Cited p. 170.

[6] M. Greferath and S. E. Schmidt. Finite ring combinatorics and MacWilliams' Equivalence Theorem. *J. Combin. Theory Ser. A*, 92:17–28, 2000. Cited p. 167.

[7] A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Solé. The $\mathbb{Z}_4$-linearity of Kerdock, Preparata, Goethals, and related codes. *IEEE Trans. Inform. Theory*, IT-40:301–319, 1994. Cited pp. 167, 174, and 175.

[8] Y. Hirano. On admissible rings. *Indag. Math.*, 8:55–59, 1997. Cited p. 169.

[9] T. Honold and I. Landjev. MacWilliams identities for linear codes over finite Frobenius rings. In *Proceedings of The Fifth International Conference on Finite Fields and Applications*, pages 276–292. Springer, 2001. Cited pp. 167 and 176.

[10] K. Huber. The MacWilliams theorem for two-dimensional modulo metrics. *Appl. Algebra Engrg. Comm. Comput.*, 8:41–48, 1997. Cited p. 175.

[11] W. C. Huffman and V. Pless. *Fundamentals of Error-Correcting Codes*. Cambridge University Press, 2003. Cited pp. 167 and 170.

[12] M. Klemm. Über die Identität von MacWilliams für die Gewichtsfunktion von Codes. *Arch. Math (Basel)*, 49:400–406, 1987. Cited pp. 167 and 175.

[13] M. Klemm. Eine Invarianzgruppe für die vollständige Gewichtsfunktion selbstdualer Codes. *Arch. Math (Basel)*, 53:332–336, 1989. Cited p. 169.

[14] M. Klemm. Selbstduale Codes über dem Ring der ganzen Zahlen modulo 4. *Arch. Math (Basel)*, 53:201–207, 1989. Cited p. 174.

[15] T. Y. Lam. *Lectures on Modules and Rings*. Springer, 1999. Cited p. 169.

[16] H. Lu, P. V. Kumar, and E. Yang. On the input-output weight enumerators of product accumulate codes. *IEEE Comm. Letters*, 8:520–522, 2004. Cited p. 179.

[17] F. J. MacWilliams. *Combinatorial problems of elementary abelian groups*. PhD thesis, Harvard University, 1962. Cited pp. 167 and 174.

[18] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 1977. Cited pp. 167, 173, 174, 175, 178, and 179.

[19] G. Nebe, E. M. Rains, and N. J. A. Sloane. *Self-dual codes and invariant theory*. Springer, 2006. Cited p. 167.

[20] A. A. Nechaev and A. S. Kuzmin. Formal duality of linearly presentable codes over a Galois field. In T. Mora and H. Mattson, editors, *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, volume 1255 of *Lecture Notes in Computer Science*, pages 263–276. Springer, 1997. Cited p. 174.

[21] M. Rosenbloom and M. Tsfasman. Codes for the m-metric. *Problemy Peredachi Informatsii*, 33(1):55–63, 1997. Cited p. 176.

[22] J. Simonis. MacWilliams identities and coordinate positions. *Lin. Algebra Appl.*, 216:81–91, 1995. Cited p. 179.

[23] M. Skriganov. On linear codes with large weights simultaneously for the Rosenbloom-Tsfasman and Hamming metrics. *J. Complexity*, 23:926–936, 2007. Cited p. 176.

[24] A. Terras. *Fourier Analysis on finite groups and applications*. Cambridge University Press, 1999. Cited pp. 169 and 170.

[25] J. A. Wood. Extension theorems for linear codes over finite rings. In T. Mora and H. Mattson, editors, *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, volume 1255 of *Lecture Notes in Computer Science*, pages 329–340. Springer, 1997. Cited p. 167.

[26] J. A. Wood. Duality for modules over finite rings and applications to coding theory. *Americ. J. of Math.*, 121:555–575, 1999. Cited pp. 167, 169, and 174.

[27] J. A. Wood. Foundations of linear codes defined over finite modules: The extension theorem and the MacWilliams identities. In P. Solè, editor, *Codes over Rings, Proceedings of the CIMPA Summer School, Ankara (2008)*, volume 6 of *Series on Coding Theory and Cryptology*, pages 124–190. Singapore: World Scientific, 2009. Cited p. 168.

[28] V. A. Zinoviev and T. Ericson. On Fourier invariant partitions of finite abelian groups and the MacWilliams identity for group codes. *Problems Inform. Transmission*, 32:117–122, 1996. Cited pp. 167, 168, 170, 176, and 178.

Lyapunov function based step size control for numerical ODE solvers with application to optimization algorithms

Lars Grüne
University of Bayreuth
Bayreuth, Germany
lars.gruene@uni-bayreuth.de

Iasson Karafyllis
Technical University of Crete
Chania, Greece
ikarafyl@enveng.tuc.gr

Abstract. We present and analyze an abstract step size selection algorithm which ensures asymptotic stability of numerical approximations to asymptotically stable ODEs. A particular implementation of this algorithm is proposed and tested with two numerical examples. The application to ODEs solving nonlinear optimization problems on manifolds is explained and illustrated by means of the Rayleigh quotient flow for computing eigenvalues of symmetric matrices.

1 Introduction

Step size control algorithms are nowadays standard in numerical methods for solving ordinary differential equations (ODEs). Due to the fact that the characteristics of the vector field depend on the state (and possibly also on time), adaptive step sizes should be used as the solution evolves. Using efficient implementations, the additional computational effort for the online computation of suitable step sizes is typically negligible compared to the gain in computational efficiency. Usually, the adaptive step sizes are selected on the basis of local error estimates and the corresponding algorithms are classical and can be found in any text book on numerical methods for differential equations, as, e.g., in [4, 14, 15, 25].

While error estimation based step size selection schemes achieve very good results in ensuring accurate approximations on finite time intervals, they do not necessarily guarantee that the asymptotic behavior of the numerical solution equals that of the exact solution. In this paper, we will investigate this case for ODEs exhibiting an asymptotically stable equilibrium. It is well known that any consistent and stable numerical scheme for ODEs inherits the asymptotic stability of the original equation in a practical sense, even for more general attractors than equilibria, see for instance [11, 12, 20] and [25, Chapter 7] for fixed step size and [7, 21] for schemes with variable step size. However, in general the numerical approximation need not be asymptotically stable in the usual sense. Instead, it may happen that the numerical solution does not converge to the equilibrium but only to a small neighborhood thereof and this can happen not only for fixed step sizes but also when error based step size control techniques are used, as [18, Example 2.11] shows.

A popular approach to ensure "true" asymptotic stability of the numerical scheme is the use of specialized numerical schemes, like (typically implicit) schemes having the A-stability or B-stability property which guarantee asymptotic stability for certain classes of ODEs, cf. e.g., [4, 15, 25], or geometric integration methods which preserve structural properties of the ODE also on infinite integration intervals, cf. e.g., [13] or [10]. Here, we build upon a different approach which was recently proposed in [18]. In this reference, general consistent Runge-Kutta schemes (explicit or implicit) were represented as hybrid control systems such that the step size selection problem could be reformulated as a nonlinear feedback stabilization problem. Consequently, nonlinear feedback design techniques like the small gain methodology or Lyapunov function based design, e.g., via backstepping, could then be applied to solve the problem under suitable assumptions on the system and, in case of Lyapunov function based design, on the corresponding Lyapunov function. Although the methods proposed in [18] may not necessarily outperform specialized tailored methods for particular problem classes, they form a systematic and versatile approach which can be applied to many different problems. While the majority of the results in [18] were focused on existence issues or explicit state dependent step size formulas, it was also observed that if a Lyapunov function for the ODE is known, then an online step size control algorithm similar to classical error estimation based step size control schemes can be designed.

In this paper, we will further investigate and refine this concept. Specifically, we will present an abstract Lyapunov function based step size control algorithm and prove asymptotic stability of the generated numerical approximations under general assumptions on the functions adjusting the step sizes. We then propose an implementation of this abstract algorithm in which the adjustment of the step sizes is obtained using ideas from consistency order based step size control. In this context it is important to note that the discretization error introduced by the Runge-Kutta scheme may not necessarily destroy asymptotic stability. On the contrary, it may well happen that the numerical approximation converges to the equilibrium at a faster rate than the exact solution and this effect may be even stronger if large time steps are used. A particular feature of our algorithm — which will also be visible in our numerical examples — is that it is able to detect this situation and then allows for large step sizes. Of course, proceeding this way, we can no longer guarantee that the numerical solution faithfully reproduces the exact solution. However, the algorithm still ensures convergence to the correct equilibrium and may thus be able to reach a small neighborhood of this equilibrium with considerably less steps than an approximation which aims at a correct reproduction of the exact solution during the transient phase.

The algorithm is thus particularly suited for applications in which the numerical computation of an asymptotically stable equilbrium — but not necessarily the path along which this equilibrium is approached — is of interest. A typical class of problems in which this is the case are ODEs which are desiged for solving nonlinear optimization problems. Since such ODEs, moreover, often come with a canonical Lyapunov function (in the simplest case the function to be optimized, itself), our algorithm is readily applicable. For optimization problems appearing in mathematical systems theory, the monograph of Helmke and Moore [16] presents a variety of ODEs for optimization and we will illustrate the use of our algorithm in this area by applying it to the Rayleigh quotient flow for computing the minimal eigenvalue of a symmetric matrix which is a particular example from [16].

The remainder of the paper is organized as follows. After introducing the necessary notation at the end of this introduction, Section 2 defines the systems under consideration as well as Runge-Kutta approximations and their representation via hybrid systems and introduces the precise problem formulation. Moreover, preliminary Lyapunov function results from [18] are recalled for convenience of the reader. In Section 3 we first present and analyze our abstract step size control algorithm and then discuss a particular implementation of this algorithm and illustrate its performance by means of two numerical examples. Section 4 then discusses the application nonlinear optimization, gives a brief survey of approaches from the literature to which our algorithm applies and finally illustrates the performance of our algorithm for the Rayleigh quotient flow.

1.1 Notation

By $C^0(A\,;\,\Omega)$, we denote the class of continuous functions on $A \subseteq \mathbb{R}^n$, which take values in $\Omega \subseteq \mathbb{R}^m$. By $C^k(A\,;\,\Omega)$, where $k \geq 1$ is an integer, we denote the class of differentiable functions on A with continuous derivatives up to order k, which take values in Ω.

For a vector $x \in \mathbb{R}^n$ we denote by $\|x\|$ the Euclidean norm and by $x^\top$ its transpose. By $B_\varepsilon(x)$, where $\varepsilon > 0$ and $x \in \mathbb{R}^n$, we denote the ball of radius $\varepsilon > 0$ centered at $x \in \mathbb{R}^n$, i.e., $B_\varepsilon(x) := \{y \in \mathbb{R}^n : |y - x| < \varepsilon\}$.

$\mathbb{R}^+$ denotes the set of non-negative real numbers and $\mathbb{Z}^+$ the set of non-negative integer numbers. By $\mathcal{K}_\infty$ we denote the set of all strictly increasing and continuous functions $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ with $\rho(0) = 0$ and $\lim_{s\to+\infty} \rho(s) = +\infty$.

For every scalar continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}$, $\nabla V(x)$ denotes the gradient of V at $x \in \mathbb{R}^n$, i.e., $\nabla V(x) = (\frac{\partial V}{\partial x_1}(x), \ldots, \frac{\partial V}{\partial x_n}(x))$. We say that a function $V : \mathbb{R}^n \to \mathbb{R}^+$ is positive definite if $V(x) > 0$ for all $x \neq 0$ and $V(0) = 0$. We say that a continuous function $V : \mathbb{R}^n \to \mathbb{R}^+$ is radially unbounded if for every $M > 0$ the set $\{x \in \mathbb{R}^n : V(x) \leq M\}$ is compact.

2 Setting and problem formulation

We consider autonomous differential equations of the type

$$\dot{z}(t) = f(z(t))\,,\; z(t) \in \mathbb{R}^n \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitz vector field for which there exists $x^* \in \mathbb{R}^n$ with $f(x^*) = 0$. Without loss of generality we may assume $x^* = 0$. For every $z_0 \in \mathbb{R}^n$ and $t \geq 0$, the solution of (1) with initial condition $z(0) = z_0$ will be denoted by $z(t)$ or by $z(t, z_0)$ if we want to emphasize the dependence on the initial value z_0.

2.1 Runge-Kutta schemes

A standard way of obtaining a numerical approximation of this solution is via a Runge-Kutta scheme. Here we summarize the facts for these schemes we need in this paper. Proofs and more details can be found, e.g., in the monographs [4], [14, 15] or [25].

Given an approximation $\tilde{z} \approx z(t)$ for some $t \geq 0$ and a time step $h > 0$, an approximation $\Phi(\tilde{z},h) \approx z(t+h)$ via an s-stage Runge-Kutta method is given by

$$k_j \quad = \quad f\left(\tilde{z} + h\sum_{l=1}^{s} a_{jl}k_l\right), \quad j = 1,\dots,s \tag{2}$$

$$\Phi(\tilde{z},h) \quad := \quad \tilde{z} + h\sum_{j=1}^{s} b_j k_j \tag{3}$$

Here a_{jl}, b_j, $j,l = 1,\dots,s$, are the *coefficients* of the scheme and $k_1,\dots,k_s$ are called the *stages* of the scheme. Some popular examples for Runge-Kutta schemes can be found in Section 3.3, below. If $a_{jl} = 0$ for all $l \geq j$ and all $j = 1,\dots,s$, then the scheme is called *explicit* and the equations (2) can be evaluated recursively. Otherwise, the scheme is called *implicit* and the equations (2) form a system of (in general nonlinear) equations. Under the Lipschitz assumption on f one can show using Banach's fixed point theorem that there exists a continuous function $h_{max} : \mathbb{R}^n \to \mathbb{R}^+$ such that a unique solution of (2) exists for each $\tilde{z} \in \mathbb{R}^n$ and each $h \in (0, h_{max}(\tilde{z})]$, see, e.g. [18]. In a practical implementation, (2) needs to be solved by some numerical method, e.g., a fixed-point iteration or Newton's method. Even though this introduces additional numerical effort in computing $\Phi(\tilde{z},h)$, this effort may pay off when solving stiff equations.

Given the initial time $\tau_0 = 0$, an initial value $z_0 \in \mathbb{R}^n$ and a sequence of time steps $h_i > 0$ we recursively define the times $\tau_{i+1} := \tau_i + h_i$. Then, one can generate approximations $\tilde{z}_i \approx z(\tau_i, z_0)$ at the times τ_i via the iteration

$$\tilde{z}_0 := z_0, \quad \tilde{z}_{i+1} := \Phi(\tilde{z}_i, h_i).$$

In order to analyze the convergence of a Runge-Kutta scheme, one looks at the approximation error $e_i := \|\tilde{z}_i - z(\tau_i, z_0)\|$. For estimating this error, the concept of *consistency* is used.

Definition 1. A Runge-Kutta scheme is called *consistent* with *order* $p \geq 1$, if for each compact set $K \subset \mathbb{R}^n$ there exists $\bar{h} > 0$ and a constant $C > 0$ such that the inequality

$$\|\Phi(z_0,h) - z(h,z_0)\| \leq Ch^{p+1} \tag{4}$$

holds for all $z_0 \in K$ and all $h \in (0,\bar{h}]$, where $z(h,z_0)$ denotes the solution of (1) and $\bar{h} > 0$ is chosen such that this solution exists for all $z_0 \in K$ and all $h \in (0,\bar{h}]$.

The consistency and the order of consistency depends on the coefficients of the scheme. For instance, the condition $\sum_{j=1}^{s} b_j = 1$ ensures consistency with order $p = 1$ for continuously differentiable vector fields f. Additional conditions on the coefficients a_{jl} and b_j ensure higher order consistency, i.e., (4) with $p \geq 2$, provided the vector field f is sufficiently smooth. Consistency together with a Lipschitz-type stability condition (which holds for any Runge-Kutta scheme provided f in (1) is Lipschitz) then implies convergence of the scheme. More precisely, if the scheme is consistent and if the solution $z(t,z_0)$ exists for $t \in [0,T]$, then there exists a constant

$C_T > 0$, such that for any selection of time steps $h_0, \ldots, h_{N-1} > 0$ satisfying $\tau_i \in [0,T]$ for $i = 0, \ldots, N$ the inequality

$$\max_{i=0,\ldots,N} e_i \le C_T h^p \tag{5}$$

holds for all sufficiently small $h > 0$, where $h := \max_{i=0,\ldots,N} h_i$ and p is the order of consistency of the scheme.

2.2 Runge-Kutta schemes as hybrid systems

Our goal in this paper is to analyze the dynamical behavior of the numerical approximation, more precisely its asymptotic stability at the origin. To this end, we need to interpret the values $\tilde{z}_i$ as states of a dynamical system. This is a relatively easy task if the time steps h_i are constant, i.e., $h_i \equiv h$ for all $i = 0, \ldots, N$, since in this case $\tilde{z}_{i+1} = \Phi(\tilde{z}_i, h)$ defines a standard discrete time dynamical system. However, if the h_i are time varying — which is the case we consider in this paper — the situation becomes more complicated. Varying time steps can, for instance, be handled as part of an extended state space, cf. [22], or by defining the discrete time system on the nonuniform time grid $\{\tau_0, \tau_1, \tau_2, \ldots\}$ induced by the time steps, cf. [21] or [7]. Here, we choose another option, namely to represent the numerical approximation by a hybrid dynamical system of the form

$$\begin{aligned} &\dot{x}(t) = F(h_i, x(\tau_i)) \, , \, t \in [\tau_i, \tau_{i+1}) \\ &\tau_0 = 0 \, , \, \tau_{i+1} = \tau_i + h_i \\ &h_i = \varphi(x(\tau_i)) \exp(-u(\tau_i)) \\ &x(t) \in \mathbb{R}^n \, , u(t) \in [0, +\infty) \end{aligned} \tag{6}$$

where $\varphi \in C^0(\mathbb{R}^n; (0,r])$, $r > 0$ is a constant, $F : \bigcup_{x \in \mathbb{R}^n} ([0, \varphi(x)] \times \{x\}) \to \mathbb{R}^n$ is a (not necessarily continuous) vector field with $F(h,0) = 0$ for all $h \in [0, \varphi(0)]$, $\lim_{h \to 0^+} F(h,z) = f(z)$, for all $z \in \mathbb{R}^n$. The function $u : \mathbb{R}^+ \to \mathbb{R}^+$ is a locally bounded input to the system whose meaning will be described below.

The solution $x(t)$ of the hybrid system (6) is obtained for every such u by setting $\tau_0 = 0$, $x(0) := x_0$ and then proceeding iteratively for $i = 0, 1, \ldots$ as follows (cf. [17]):

1. Given τ_i and $x(\tau_i)$, calculate τ_{i+1} according to $\tau_{i+1} = \tau_i + \varphi(x(\tau_i)) \exp(-u(\tau_i))$
2. Compute the state trajectory $x(t)$, $t \in (\tau_i, \tau_{i+1}]$ as the solution of the differential equation $\dot{x}(t) = F(h_i, x(\tau_i))$, i.e., $x(t) = x(\tau_i) + (t - \tau_i) F(h_i, x(\tau_i))$ for $t \in (\tau_i, \tau_{i+1}]$.

We denote the resulting trajectory by $x(t, x_0, u)$ or briefly $x(t)$ when x_0 snd u are clear from the context.

Any Runge-Kutta scheme can be represented by a hybrid system (6) by defining

$$F(h,x) := h^{-1}(\Phi(x,h) - x) = \sum_{j=1}^{s} b_j k_j \tag{7}$$

Indeed, from the explicit solution formula in Step 2, above, it immediately follows that the solutions of the hybrid system using this F and $x_0 = z_0$ satisfy

$$x(\tau_i, x_0, u) = \tilde{z}_i.$$

The corresponding time steps $h_i = \varphi(x(\tau_i))\exp(-u(\tau_i))$ are determined via the state dependent function $\varphi(x(\tau_i))$ and the time dependent factor $\exp(-u(\tau_i)) \in (0,1]$. Hence, $\varphi(x(\tau_i))$ can be interpreted as the maximal allowable step size for the state $x(\tau_i)$ (with global upper bound $r > 0$) and $u(\tau_i)$ can be used to arbitrarily reduce this value. Note that for implicit Runge-Kutta schemes typically an upper bound on the step size is needed in order to ensure solvability of the system of equations (2) and the function φ can be used for this purpose, for details we refer to [18].

We will further assume that there exists a continuous, non-decreasing function $M : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\|F(h,x)\| \leq \|x\| M(\|x\|) \text{ for all } x \in \mathbb{R}^n \text{ and } h \in [0, \varphi(x)] \tag{8}$$

This condition implies that (6) has the “Boundedness-Implies-Continuation” property and thus for each locally bounded input $u : \mathbb{R}^+ \to \mathbb{R}^+$ and $x_0 \in \mathbb{R}^n$ there exists a unique absolutely continuous solution function $[0, +\infty) \ni t \to x(t) \in \mathbb{R}^n$ with $x(0) = x_0$, see [17]. Appropriate step size restriction can always guarantee that (8) holds for F from (7), cf. [18].

Modeling numerical schemes (and particularly Runge-Kutta schemes) as hybrid systems is nonstandard but has certain advantages compared to the alternative discrete time formulations approaches from [7, 21, 22]. For instance, here we aim at stability statements for all step size sequences $(h_i)_{i \in \mathbb{N}_0}$ with $h_i > 0$ and $h_i \leq \varphi(x(\tau_i))$, cf. the discussion after Definition 3, below. Once φ is fixed, for the hybrid system (6) this is equivalent to ensuring the desired stability property for all locally bounded functions $u : \mathbb{R}^+ \to \mathbb{R}^+$. Hence, our hybrid approach leads to an explicit condition (“for all u”) while the discrete time approach leads to a more technical implicit condition (“for all h_i satisfying $h_i \leq \varphi(x(\tau_i))$”). Moreover, the formulation via hybrid models enables us to use readily available stability results from the hybrid control systems literature, while for other formulations we would have to rely on ad hoc arguments.

2.3 Problem formulation

Our general aim is to ensure asymptotic stability of the origin for (6), (7) for suitable choices of φ and all locally bounded inputs u, provided the origin is asymptotically stable for (1). To this end, we first precisely define these stability properties.

For the differential equation (1) we use the following condition, cf. [23] (see also [17, 19]).

Definition 2. We say that the origin $0 \in \mathbb{R}^n$ is *uniformly globally asymptotically stable* (UGAS) for (1) if it is

(i) *Lyapunov stable*, i.e., for each $\varepsilon > 0$ there exists $\delta > 0$ such that $\|z(t, z_0)\| \leq \varepsilon$ for all $t \geq 0$ and all $z_0 \in \mathbb{R}^n$ with $\|z_0\| \leq \delta$ and

(ii) *uniformly attractive*, i.e., for each $R > 0$ and $\varepsilon > 0$ there exists $T > 0$ such that $\|z(t, z_0)\| \leq \varepsilon$ for all $t \geq T$ and all $z_0 \in \mathbb{R}^n$ with $\|z_0\| \leq R$.

The next definition formalizes the condition that (6) is asymptotically stable for all locally bounded inputs u, cf. [17].

Definition 3. We say that the origin $0 \in \mathbb{R}^n$ is *uniformly robustly globally asymptotically stable* (URGAS) for (6) if it is

(i) *robustly Lagrange stable*, i.e., for each $R > 0$ it holds that $\sup\{\|x(t,x_0,u)\| \,|\, t \geq 0, \|x_0\| \leq R, u : \mathbb{R}^+ \to \mathbb{R}^+$ locally bounded$\} < \infty$,

(ii) *robustly Lyapunov stable*, i.e., for each $\varepsilon > 0$ there exists $\delta > 0$ such that $\|x(t,x_0,u)\| \leq \varepsilon$ for all $t \geq 0$, all $x_0 \in \mathbb{R}^n$ with $\|x_0\| \leq \delta$ and all locally bounded $u : \mathbb{R}^+ \to \mathbb{R}^+$, and

(iii) *robustly uniformly attractive*, i.e., for each $R > 0$ and $\varepsilon > 0$ there exists $T > 0$ such that $\|x(t,x_0,u)\| \leq \varepsilon$ for all $t \geq T$, all $x_0 \in \mathbb{R}^n$ with $\|x_0\| \leq R$ and all locally bounded $u : \mathbb{R}^+ \to \mathbb{R}^+$.

Contrary to the ordinary differential equation (1), for the hybrid system (6) Lyapunov stability and attraction do not necessarily imply Lagrange stability. This is why — in contrast to Definition 2 — we explicitly included this property in Definition 3.

Note that our choice $\varphi \in C^0(\mathbb{R}^n;(0,r])$ implies $\inf_{x\in\mathcal{N}} \varphi(x) > 0$ for any bounded neighborhood $\mathcal{N}$ of the origin. This implies that the asymptotic stability property can be achieved for a sequence of step sizes h_i which is bounded from below by a positive value. This avoids the undesirable property that the discretization step sizes tend to 0 as $i \to +\infty$. However, as we will see, it will also be possible to make rigorous statements in situations where such a φ does not exist, cf. Theorem 7 and the discussion after its proof.

The stability property in Definition 3 is called *robust* because it requires the respective stability properties uniformly for all locally bounded inputs u and thus for all (positive) step sizes $h_i \leq \varphi(x(\tau_i))$. This is an important feature because it allows us to couple our method with other step size selection schemes. For instance, we could use the step size $\min\{\varphi(x(\tau_i)),\tilde{h}_i\}$ where $\tilde{h}_i$ is chosen such that a local error bound is guaranteed. Such methods are classical, cf. [14] or any other textbook on numerical methods for ODEs. Proceeding this way results in a numerical solution which is asymptotically stable and at the same time maintains a pre-defined accuracy. Note that our approach will not incorporate error bounds, hence the approximation may deviate from the true solution, at least in the transient phase, i.e., away from 0. On the other hand, as [18, Example 2.1] shows, local error based step size control does in general not guarantee asymptotic stability of the numerical approximation. Thus, a coupling of both approaches may be needed in order to ensure both accuracy and asymptotic stability.

Assuming that Definition 2 is satisfied, Definition 3 now gives rise to several problems. The first of these is the following existence problem.

(P1) Existence *Is there a continuous function* $\varphi : \mathbb{R}^n \to (0,r]$*, such that* $0 \in \mathbb{R}^n$ *is URGAS for system* (6), (7)*?*

Provided the answer to **(P1)** is positive, one may then look at the following design problems.

(P2) Global Design *Construct a continuous function* $\varphi : \mathbb{R}^n \to (0,r]$*, such that* $0 \in \mathbb{R}^n$ *is URGAS for system* (6), (7).

(P3) Trajectory based Design *For a given initial value x_0, construct a sequence of step sizes $h_i > 0$ satisfying $h_i \leq \varphi(x(\tau_i))$ for all $i \in \mathbb{N}$ and the function φ from* **(P1)**.

A variety of results related to Problems **(P1)** and **(P2)** can be found in [18]. In this context we note that any consistent and stable numerical scheme for ODEs inherits the asymptotic stability of the original equation in a practical sense, even for more general attractors than equilibria, see for instance [11, 12] or [25, Chapter 7]. Practical asymptotic stability means that the system exhibits an asymptotically stable set close to the original attractor, i.e., in our case a small neighborhood around the equilibrium point, which shrinks down to the attractor as the time step h tends to 0. In contrast to this, the property defined in Definition 3 is "true" asymptotic stability, a stronger property which cannot in general be deduced from practical stability. In [25, Chapter 5], several results for our problem for specific classes of ODEs are derived using classical numerical stability concepts like A-stability, B-stability and the like. In [18], it was observed that Problems **(P1)** and **(P2)** can be interpreted as feedback stabilization problems for the system (6), (7) in which φ plays the role of the stabilizing feedback law. Consequently, various methods from nonlinear control theory, like small-gain and Lyapunov function techniques, have been applied to these problems in [18] generalizing the results from [25, Chapter 5] to more general classes of systems and to systems with different structural properties. In contrast to [18], in this paper our focus lies on Problem **(P3)** and applications thereof.

2.4 Lyapunov functions

Lyapunov functions are the main technical tool we are going to use in this paper. In this section we collect and extend some results from [18] which form the basis for our algorithm and analysis. The first lemma gives a sufficient Lyapunov condition for the URGAS property for hybrid systems of the form (6). For its proof we refer to [18, Lemma 4.1].

Lemma 4. *Consider system* (6) *and suppose that there exist a continuous, positive definite and radially unbounded function $V : \mathbb{R}^n \to \mathbb{R}^+$ and a continuous, positive definite function $W : \mathbb{R}^n \to \mathbb{R}^+$ such that for every $x \in \mathbb{R}^n$ the following inequality holds for all $h \in [0, \varphi(x)]$.*

$$V(x + hF(h,x)) \leq V(x) - hW(x) \tag{9}$$

Then the origin $0 \in \mathbb{R}^n$ is URGAS for system (6).

In the following section, we will use inequality (9) in order to construct adaptive step sizes h_i online while computing the numerical solution. To this end, we need to know a Lyapunov function V. Similar to [18], we will use a Lyapunov function for the continuous-time system (1) for this purpose. Such a Lyapunov function is defined as follows.

Definition 5. A positive definite, radially unbounded function $V \in C^1(\mathbb{R}^n;\mathbb{R}^+)$ is called a *Lyapunov function* for system (1) if the inequality

$$\nabla V(x) f(x) < 0 \tag{10}$$

holds for all $x \in \mathbb{R}^n \setminus \{0\}$.

As we will see in the proof of Theorem 7, below, such a Lyapunov function for (1) can indeed be used in order to establish (9) for F from (7). As a prerequisite for this proof, in the remainder of this section we establish bounds on the decay of V along the solutions of (1). To this end, observe that the equation

$$\lim_{\substack{h\to 0\\ h>0}} \frac{V(z(h,x))-V(x)}{h} = \nabla V(x)f(x)$$

(which follows by the chain rule) together with $\nabla V(x)f(x)<0$ for $x\neq 0$ implies that for each $\lambda\in(0,1)$, each $x\in\mathbb{R}^n$ and each sufficiently small $h>0$ the inequality

$$V(z(h,x))-V(x)\le h\lambda\nabla V(x)f(x) \tag{11}$$

holds. The following lemma makes the statement "sufficiently small $h>0$" in this observation more precise.

Lemma 6. *Let $V\in C^1(\mathbb{R}^n;\mathbb{R}^+)$ be a Lyapunov function for system* (1) *and $\lambda\in(0,1)$. Then the following statements hold.*

(i) For each two constants $R>\varepsilon>0$ there exists $h_{\varepsilon,R}>0$ such that (11) *holds for all $x\in\mathbb{R}^n$ with $R\ge\|x\|\ge\varepsilon$ and all $h\in(0,h_{\varepsilon,R}]$.*

(ii) Assume that $W(x):=-\nabla V(x)f(x)$ is locally Lipschitz and that there exist constants $b>1$, $\varepsilon,c>0$ and a continuous positive definite function $\ell:\mathbb{R}^n\to\mathbb{R}^+$ satisfying

$$\ell(x)\ge\sup\left\{\frac{|W(y)-W(z)|}{\|y-z\|} : y,z\in\mathbb{R}^n,\, y\neq z,\, \max\{\,\|y\|\,,\,\|z\|\,\}\le b\,\|x\|\right\}$$

for all $x\in\mathbb{R}^n\setminus\{0\}$ and

$$\|x\|\,\ell(x)\le cW(x) \tag{12}$$

for all $x\in B_\varepsilon(0)$. Then there exists a continuous function $\varphi\in C^0(\mathbb{R}^n;(0,r])$ for some $r>0$ such that (11) *holds for all $x\in\mathbb{R}^n$ and all $h\in(0,\varphi(x))$.*

Proof. (i) Fix an arbitrary $\eta\in(0,\varepsilon)$ and consider the compact sets $K:=\{x\in\mathbb{R}^n\,|\,R\ge\|x\|\ge\varepsilon\}$ and $K_\eta:=\{x\in\mathbb{R}^n\,|\,R+\eta\ge\|x\|\ge\varepsilon-\eta\}$ and the map $x\mapsto\nabla V(x)f(x)$. This map is continuous and attains negative values on K, hence the value $\alpha:=\max_{x\in K}\nabla V(x)f(x)$ exists and satisfies $\alpha<0$. Moreover, since any continuous map is uniformly continuous on each compact set, the map is uniformly continuous on K_η, i.e., for given $\varepsilon'>0$ there exists $\delta>0$ such that

$$|\nabla V(x)f(x)-\nabla V(y)f(y)|\le\varepsilon'$$

holds for all $x,y\in K_\eta$ with $\|x-y\|\le\delta$.

Since the vector field f is also continuous, its norm $\|f(x)\|$ is bounded on K_η by some $M>0$ and thus for all $t\in[0,\eta/M]$ we obtain that $z(t,x)\in K_\eta$ and $\|z(t,x)-x\|\le tM$ for all $x\in K$.

Now we set $\varepsilon' = (\lambda - 1)\alpha$, pick $\delta > 0$ from the uniform continuity property (without loss of generality we may choose $\delta \le \eta$) and set $h_{\varepsilon,R} := \delta/M$. Then for all $x \in K$ and all $t \in (0, h_{\varepsilon,R}]$ we obtain $\|z(t,x) - x\| \le \delta$ and thus for all $h \in (0, h_{\varepsilon,R}]$ we can estimate

$$\begin{aligned} V(z(h,x)) - V(x) &= \int_0^h \nabla(z(t,x)) f(z(t,x)) dt \\ &\le \int_0^h \nabla(x) f(x) + \varepsilon' dt \;= h\,\nabla(x) f(x) + h(\lambda - 1)\alpha \\ &\le h\nabla(x) f(x) + h(\lambda - 1)\nabla(x) f(x) \;=\; h(1-\lambda)\nabla(x) f(x) \end{aligned}$$

which shows the claim.

(ii) Follows from [18, Lemma 4.3]. □

3 Lyapunov function based step size control

In this section we present our Lyapunov function based step size control algorithm. We start by stating and analyzing an “abstract” version of this algorithm and then describe the details of our implementation and illustrate it by means of two numerical examples.

3.1 An abstract step size control algorithm

The following algorithm provides an abstract step size selection method based on a Runge-Kutta scheme for (1), expressed via (7) as a hybrid system (6), a Lyapunov function V for (1) according to Definition 5 and its derivative ∇V. Moreover, we assume that we have defined two functions

$$h_{reduce} : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^+ \quad \text{and} \quad h_{new} : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^+,$$

which for a given $h > 0$ and $x \in \mathbb{R}^n$ produce a reduced step size $h_{reduce}(h,x) < h$ and a new step size $h_{new}(h,x) > 0$. In order to simplify the presentation, the algorithm uses a maximal step size $h_{max} > 0$ which does not depend on the state x, but the inclusion of a state dependent upper bound is straightforward.

Algorithm 3: Lyapunov function based step size control algorithm

```
1: inputs
      Initial value $x_0 \in \mathbb{R}^n$, initial step size $h_0 > 0$, maximal step size $h_{max} > 0$,
      parameter $\lambda \in (0,1)$
2: set $x(0) := x_0$, $\tau_0 := 0$, $i := 0$
3: set $h_i := \min\{h_i, h_{max}\}$ and compute $x(\tau_i + h_i) := \Phi(x(\tau_i), h_i)$
4: while $V(x(\tau_i + h_i)) - V(x(\tau_i)) > \lambda h \nabla V(x(\tau_i)) f(x(\tau_i))$ do
5:    set $h_i := h_{reduce}(h_i, x(\tau_i))$ and recompute $x(\tau_i + h_i) := \Phi(x(\tau_i), h_i)$
6: end while
7: set $h_{i+1} := h_{new}(h_i, x(\tau_i))$, $\tau_{i+1} := \tau_i + h_i$, $i := i + 1$ and goto step 2
```

Note that this algorithm does not have a termination criterion and hence — in principle — produces infinitely many values $x(\tau_i)$. Of course, if only a solution on some interval $[0, T]$ is desired, the algorithm could be modified accordingly.

Since the above algorithm only produces one single sequence of step sizes h_i for each initial value x_0, it does not make sense to talk about robust stability concepts anymore. Moreover, since the Lyapunov function condition in step (2) is only ensured at the discrete times τ_i, we cannot in general expect stability properties for all $t \geq 0$ (although under suitable conditions they could be recovered by continuity arguments, see, e.g., [24]). However, under appropriate assumptions on h_{reduce} and h_{new} we can still recover the required properties from Definition 3 at the discrete time instants τ_i, as the following theorem states.

Theorem 7. *Consider Algorithm* 3 *in which* Φ *is a consistent Runge-Kutta scheme with order* $p \geq 1$. *Let* $V \in C^1(\mathbb{R}^n;\mathbb{R}^+)$ *be a Lyapunov function for system* (1) *and* $\lambda \in (0,1)$. *Assume that* $h_{\max} > 0$ *is chosen such that* $\Phi(x,h)$ *is defined for all* $x \in \mathbb{R}^n$ *and* $h \in (0,h_{max}]$ *and that the functions* h_{reduce} *and* h_{new} *satisfy*

- *there exist real values* $0 < \rho_1 \leq \rho_2 < 1$ *such that* $h_{reduce}(x,h) \in [\rho_1 h, \rho_2 h]$ *holds for all* $x \in \mathbb{R}^n$ *and all* $h > 0$ *for which the condition in Step (2) of Algorithm* 3 *is satisfied*
- $h_{new}(x,h) \geq h$ *holds for all* $x \in \mathbb{R}^n$ *and all* $h > 0$ *for which the condition in Step (2) of Algorithm* 3 *is not satisfied.*

Then, for each initial value Algorithm 3 *generates an infinite sequence of time steps* h_i *(i.e., the while loop in steps (2)–(4) always terminates) and at the times* τ_i *the resulting trajectories are*

(i) Lagrange stable*, i.e., for each* $R > 0$ *it holds that* $\sup\{\|x(\tau_i,x_0)\| \,|\, i \in \mathbb{N}, \|x_0\| \leq R\} < \infty$.

(ii) Lyapunov stable*, i.e., for each* $\varepsilon > 0$ *there exists* $\delta > 0$ *such that* $\|x(\tau_i,x_0)\| \leq \varepsilon$ *for all* $i \in \mathbb{N}$ *and all* $x_0 \in \mathbb{R}^n$ *with* $\|x_0\| \leq \delta$

(iii) uniformly attractive*, i.e., for each* $R > 0$ *and* $\varepsilon > 0$ *there exists* $T > 0$ *such that* $\|x(\tau_i,x_0)\| \leq \varepsilon$ *for all* $\tau_i \geq T$ *and all* $x_0 \in \mathbb{R}^n$ *with* $\|x_0\| \leq R$.

If, in addition, there exists a continuous function $\varphi \in C^0(\mathbb{R}^n;(0,r])$ *for some* $r > 0$ *such that*

$$V(\Phi(x,h)) - V(x) \leq h\lambda \nabla V(x) f(x) \tag{13}$$

holds for all $x \in \mathbb{R}^n$ *and all* $h \in (0,\varphi(x))$*, then for each initial value* $x_0 \in \mathbb{R}^n$ *there exists* $h_{min} > 0$ *such that* $h_i \geq h_{min}$ *holds for all* $i \in \mathbb{N}$. *In particular, in this case the sequence of times* τ_i *generated by Algorithm* 3 *tends to* ∞ *as* $i \to \infty$.

Proof. Let $(a_k)_{k\in\mathbb{N}}$ be an arbitrary sequence with $0 < a_{k+1} < a_k$, $\lim_{k\to\infty} a_k \to 0$ and pick $\tilde{\lambda} \in (\lambda,1)$ arbitrarily. Define the sets $M := \{x \in \mathbb{R}^n | V(x) \leq a_1\}$, $M_k := \{x \in \mathbb{R}^n | V(x) \in [a_{k+1}, a_k]\}$ and let $k \in \mathbb{N}$ be arbitrary. Since V is continuous, positive definite and radially unbounded, it follows from Lemma 3.5 in [19] that there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ with

$$\alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|) \quad \text{for all } x \in \mathbb{R}^n. \tag{14}$$

This implies that the sets M and M_k are compact and there exists $R_k > \varepsilon_k > 0$ such that $R_k \geq \|x\| \geq \varepsilon_k$ holds for all $x \in M_k$. Thus we can apply Lemma 6(i) which implies that there exists $h_{\varepsilon_k,R_k} \in (0,h_{max}]$ such that the condition (11) holds for $\tilde{\lambda}$ in place of λ for all $x \in M_k$ and all $h \in (0,h_{\varepsilon_k,R_k}]$.

Since the scheme is consistent and V is Lipschitz, this implies

$$V(\Phi(h,x)) - V(x) \le V(z(h,x)) - V(x) + LCh^{p+1} \le h\tilde{\lambda}\nabla V(x)f(x) + LCh^{p+1}$$

for all $x \in M_k$ and all $h \in (0,\min\{\bar{h},h_{\varepsilon_k,R_k}\}]$, where $\bar{h}$ is the step size from the consistency condition for the compact set M and L is the Lipschitz constant of V on the set $K = \{\Phi(x,h)\,|\,x \in M,\, h \in [0,\bar{h}]\} \cup \{z(h,x)\,|\,x \in M,\, h \in [0,\bar{h}]\}$, which is compact since $M \times [0,h]$ is compact and both $z(\cdot,\cdot)$ and $\Phi(\cdot,\cdot)$ are continuous. Setting

$$\gamma_k := \max_{x\in M_k} \nabla V(x)f(x) < 0 \quad \text{and} \quad h_k' := \left(\frac{(\lambda - \tilde{\lambda})\gamma}{LC}\right)^{\frac{1}{p}} > 0,$$

for $\bar{h}_k := \min\{\bar{h}, h_{\varepsilon,R}, h'\}$ we thus obtain

$$V(\Phi(x,h)) - V(x) \le h\lambda\nabla V(x)f(x) \le h\lambda\gamma_k \tag{15}$$

for all $x \in M_k$ and all $h \in (0,\bar{h}_k]$.

Now consider the while loop in the steps (2)–(4) of Algorithm 3 for some $x(\tau_i) \in M_k$ and denote by $h_{i,old}$ the time step for which step (1) is executed before entering this loop. Inequality (15) and the inequalities for h_{reduce} then implies that for any $x(\tau_i) \in M_k$ the loop terminates with a time step $h_i \ge \min\{h_{i,old}, \rho_1\bar{h}_k\}$ for which $V(\Phi(x(\tau_i),h_i)) \le V(x(\tau_i)) \le h_i\lambda\gamma_k$ holds. Moreover, since $h_{new}(h_i,x(\tau_i)) \ge h_i$, the subsequent time steps h_j, $j \ge i+1$, will satisfy the same lower bound as h_i as long as $x_j \in M_k$ holds. Hence, as long as $x_j \in M_k$ holds, $V(x(\tau_i))$ decreases monotonically with a fixed amount of decay and a uniform positive lower bound on the time step. Thus, by definition of the sets M_k, for each $k \in \mathbb{N}$ there exists a time $t_k > 0$ such that whenever $x(\tau_i) \in M_k$ there exists a $\tau_j \le \tau_i + t_k$ such that either $x(\tau_j) = 0$ (and thus $x(\tau_m) = 0$ for all $\tau_m \ge \tau_j$) or $x(\tau_j) \in M_l$ holds for some $l \ge k+1$. By induction, for each $k \in \mathbb{N}$ one thus obtains a time $T_k > 0$ such that for each $x_0 \in M$ there exists some $i_k \in \mathbb{N}$ such that $\tau_{i_k} \le T_k$ and either $x(\tau_i) \in M_k$ or $x(\tau_i) = 0$ holds for all $i \ge i_k$.

The three stability properties (i)–(iii) are now readily concluded from this property:

(i) Given $R > 0$ we choose $a_1 = \alpha_2(R)$ which implies that each $x_0 \in \mathbb{R}^n$ with $\|x_0\| \le R$ lies in M. Then the whole solution $x(\tau_i)$ lies in M which implies $\|x(\tau_i)\| \le \alpha_1^{-1}(\alpha_2(R))$ which implies Lagrange stability.

(ii) Given $\varepsilon > 0$ we choose $\delta = \alpha_2^{-1}(\alpha_1(\varepsilon))$. Then the inequality needed for Lyapunov stability follows as in (i) with δ in place of R.

(iii) Given R and ε, choose a_1 as in (i) and k so large that $a_k \le \alpha_1(\varepsilon)$ holds. Then, for $T = T_k$ and all $\tau_i \ge T$, the solution $x(\tau_i)$ is either equal to 0 or lies in M_l for some $l \ge k$. This implies $\|x(\tau_i)\| \le \alpha_1^{-1}(V(x(\tau_i))) \le \alpha_1^{-1}(a_k) = \varepsilon$ for all $\tau_i \ge T$.

We finish the proof by proving the additional property of the h_i in case that $\varphi \in C^0(\mathbb{R}^n;(0,r])$ exists such that (13) holds. To this end, pick an arbitrary $x_0 \in \mathbb{R}^n$ and choose a_1 so large that $x_0 \in M$ holds. Then, (13) implies that the values $\bar{h}_k$ defined in the proof, above, can be bounded from below by $h^* := \min_{x\in M}\varphi(x) > 0$. By induction, the inequality $h_i \ge \min\{h_{i,old}, \rho_1\bar{h}_k\}$ then implies $h_i \ge \min\{h_0, \rho_1 h^*\} > 0$ for all $i \in \mathbb{N}$ which yields the desired positive lower bound on the step sizes h_i. □

We note that various sufficient conditions ensuring the existence of $\varphi \in C^0(\mathbb{R}^n;(0,r])$ with (13) can be found in [18, Section 4]. We emphasize, however, that even without this condition the stability properties defined in Theorem 7 and in particular the convergence $x(\tau_i) \to 0$ is ensured. In particular, the numerical solution will converge to the origin even if Problem **(P1)** does not have a solution.

3.2 Implementation of the algorithm

There are various ways of defining h_{reduce} and h_{new} such that the conditions of Theorem 7 are satisfied. A simple way, proposed in [18] is to define

$$h_{reduce}(h,x) := h/2 \quad \text{and} \quad h_{new}(h,x) := 2h.$$

This choice produces reasonable results (cf. [18]) but due to the "try and error" nature of the resulting algorithm it has the disadvantage that typically the while loop is executed at least once for each i, implying that Φ is usually evaluated at least twice for each i.

A more efficient way of defining h_{reduce} and h_{new} is obtained by using ideas from the classical error estimation based step size control, cf. e.g. [14, Section II.4]. To this end, define the Lyapunov differences

$$\Delta V(x,h) := V(z(h,x)) - V(x) \quad \text{and} \quad \widetilde{\Delta V}(x,h) := V(\Phi(x,h)) - V(x)$$

for the exact solution and the numerical approximation. If V and f are sufficiently smooth, then for a p-th order scheme there exists a constant $c \in \mathbb{R}$ such that the approximate equality

$$\widetilde{\Delta V}(x,h) \approx \Delta V(x,h) + ch^{p+1}$$

holds for all sufficiently small $h > 0$. Hence, we can approximately compute c as

$$c \approx \frac{\widetilde{\Delta V}(x,h) - \Delta V(x,h)}{h^{p+1}}.$$

We now intend to find a time step $\tilde{h} > 0$ such that

$$\widetilde{\Delta V}(x,\tilde{h}) \le \lambda \tilde{h} \nabla V(x) f(x)$$

holds, which will be approximately satisfied if the inequality

$$\Delta V(x,\tilde{h}) + c\tilde{h}^{p+1} \le \lambda \tilde{h} \nabla V(x) f(x)$$

holds, i.e, if

$$\tilde{h} \le \left(\frac{\lambda \nabla V(x) f(x) - \Delta V(x,\tilde{h})/\tilde{h}}{c} \right)^{\frac{1}{p}}$$

holds. Inserting the approximate value for c, we obtain the condition

$$\tilde{h} \le h \left(\frac{\lambda \nabla V(x) f(x) - \Delta V(x,\tilde{h})/\tilde{h}}{\widetilde{\Delta V}(x,h)/h - \Delta V(x,h)/h} \right)^{\frac{1}{p}}.$$

This suggests to use the expression on the right hand side of this inequality as a candidate for a new step size in our algorithm for $h = h_i$. However, this expression is implicit (as it contains the unknown $\tilde{h}$) and contains the values $\Delta V(x,h)$ which are not available in practice as they depend on the exact solution.

Both problems vanish if we replace the term $\Delta V(x,h)$ by its first order approximation $h\nabla V(x)f(x)$ (both for h and $\tilde{h}$) which leads to the expression

$$h\left(\frac{(\lambda-1)\nabla V(x)f(x)}{\widetilde{\Delta V}(x,h)/h-\nabla V(x)f(x)}\right)^{\frac{1}{p}}.$$

Although the first order approximation $\Delta V(x,h) \approx h\nabla V(x)f(x)$ introduces an error of higher order than the error of the scheme, the resulting step size control mechanism shows very good results (cf. the discussion at the end of Example 8), probably due to the fact that the choice of $\lambda < 1$ introduces some tolerance against additional approximation errors.

For the practical implementation, we moreover need to take into account that the denominator $\widetilde{\Delta V}(x,h)/h-\nabla V(x)f(x)$ may become negative or 0 — this happens if the discretization error speeds up the convergence to the origin instead of slowing down the convergence. To this end, we replace the denominator by $\max\{\widetilde{\Delta V}(x,h)/h-\nabla V(x)f(x), \varepsilon(\lambda-1)\nabla V(x)f(x)\}$, where $\varepsilon > 0$ is a small positive constant. Finally, in order to compensate for the various approximations during the derivation, we multiply our formula with a security factor $\rho \in (0,1)$. All in all, we end up with

$$h_{reduce}(h,x) := \rho h\left(\frac{(\lambda-1)\nabla V(x)f(x)}{\max\{\widetilde{\Delta V}(x,h)/h-\nabla V(x)f(x), \varepsilon(\lambda-1)\nabla V(x)f(x)\}}\right)^{\frac{1}{p}}. \quad (16)$$

For h_{new} we may use the same formula, i.e.,

$$h_{new}(h,x) := h_{reduce}(h,x), \quad (17)$$

although this formula does not rigorously ensure the condition $h_{new}(x,h) \geq h$ imposed in Theorem 7 (it would satisfy this condition if all approximate equations in the derivation were exact). As an alternative, one might use the definition $h_{new}(h,x) := \max\{h, h_{reduce}(h,x)\}$, however, the difference between these two choices turned out to be marginal in our numerical simulations and since (17) yields a slightly lower number of evaluations of Φ we have used this variant in our simulations.

3.3 Examples

In our simulations we run Algorithm 3 with (16) for the Euler, the Heun and the classical Runge-Kutta scheme. All these schemes are explicit and thus satisfy $a_{jl} = 0$ for all $j,l = 1,\ldots,s$ with $l \geq j$. The Euler scheme is a scheme of order $p = 1$ with $s = 1$ stages and coefficient $b_1 = 1$, the Heun scheme is an $s = 2$ stage scheme of order $p = 2$ with

$$s = 2, a_{21} = 1, b_1 = b_2 = 1/2$$

and the classical Runge-Kutta scheme (henceforth abbreviated as RK4) is of order $p = 4$ with $s = 4$ stages and

$$a_{21} = a_{32} = 1/2, a_{43} = 1, a_{41} = a_{42} = a_{31} = 0, b_1 = b_4 = 1/6, b_2 = b_3 = 1/3.$$

The standard parameters in all examples were $h_0 = 0.1$, $h_{max} = 1$, and $\rho = 0.9$ and $\varepsilon = 0.01$ in Formula 16.

Example 8. The first example we investigate is the 2d differential equation

$$\dot{z}_1 = -z_1 + z_2^2, \quad \dot{z}_2 = -z_2 - z_1 z_2$$

with Lyapunov function $V(x) = \|x\|^2$, cf. [18, Example 4.15]. The Figures 1–3 show simulation results (phase portrait, Lyapunov function $V(x(\tau_i))$ over time and time steps) on the time interval $[0,20]$ with $\lambda = 0.5$ and $x_0 = (5,5)^\top$.

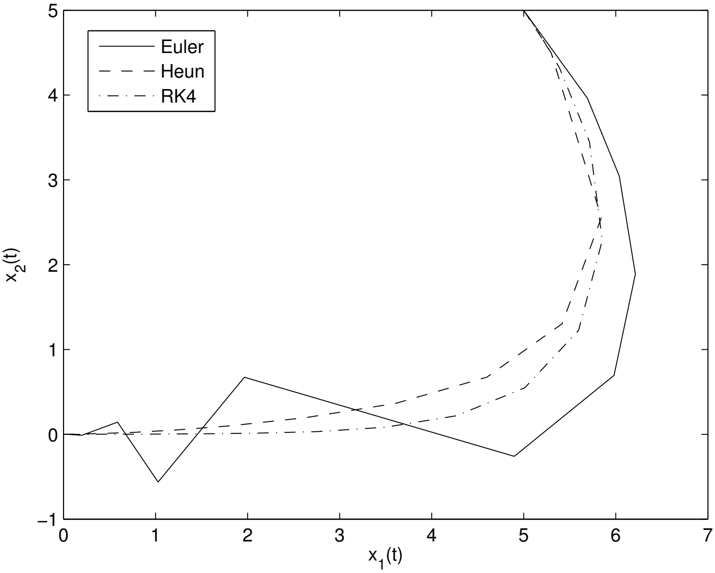


Figure 1: Phase portrait for Example 8 with $\lambda = 0.5$ and $x_0 = (5,5)^\top$

All solutions approach the equilibrium $x = 0$ very quickly. The total number of steps for the three schemes on the time interval $[0,20]$ were 28 for the Euler scheme, 42 for the Heun scheme and 52 for the RK4 scheme. Here, two facts are worth noting.

First, although the Euler scheme is the scheme with the lowest order, it allows for the largest steps and needs the fewest total number of steps. This is due to the fact that for asymptotic stability not only the size of the error matters but also the direction in which the error distorts the solution. Here, the error in the Euler scheme speeds up the convergence towards 0 and hence there is no reason for the scheme to reduce the time step. In contrast to this, while the errors for the Heun and the RK4 scheme are smaller, they have a tendency to slow down the convergence to 0 and hence the time steps have to be chosen more cautiously.

Second, it is clearly visible that our step size control Algorithm 3 does by no means ensure that the numerical solution is close to the exact solution during the transient phase, i.e., until a small neighborhood of the equilibrium is reached. In fact, the

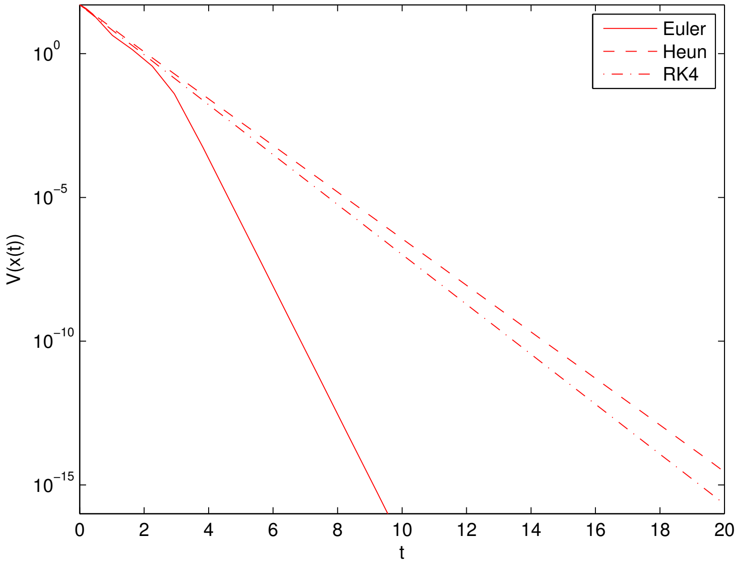


Figure 2: Lyapunov function (logarithmic scale) for Ex. 8, $\lambda = 0.5$, $x_0 = (5,5)^\top$

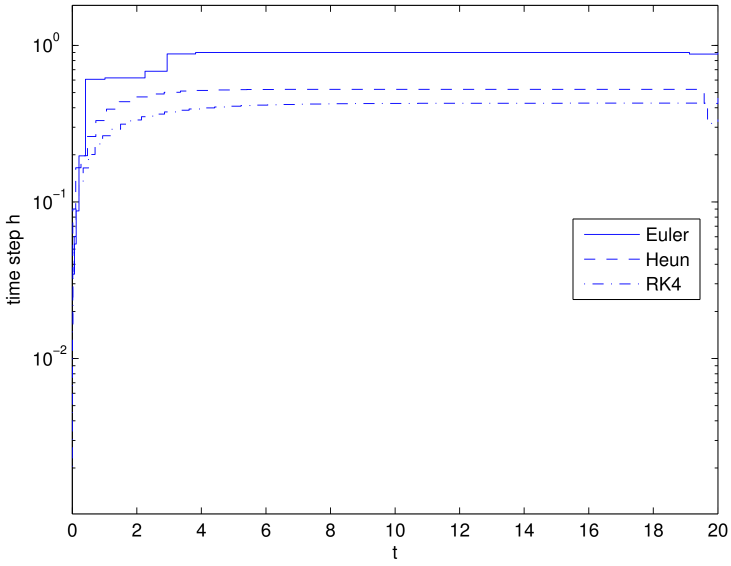


Figure 3: Time steps (logarithmic scale) for Example 8 with $\lambda = 0.5$ and $x_0 = (5,5)^\top$

three numerical solutions differ considerably and the Euler solution actually looks quite irregular. This is not a drawback of our algorithm but actually intended, since all the algorithm cares about is the convergence to $x = 0$ which is perfectly achieved for all schemes. If, in addition, a faithful reproduction of the exact solution during the transient phase is desired, our step size control algorithm could be coupled with traditional error estimation based techniques.

In order to illustrate the effects of different choices of $\lambda \in (0,1)$, Figure 4 shows the time steps for the RK4 scheme for $\lambda = 0.1, 0.5, 0.9$.

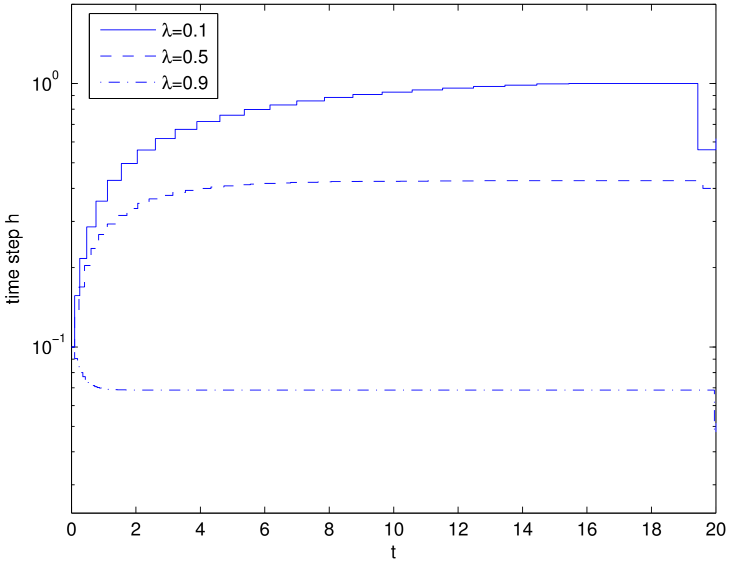


Figure 4: Time steps from Algorithm 3 (logarithmic scale) using the RK4 scheme applied to Example 8 with $\lambda = 0.1, 0.5, 0.9$ and $x_0 = (5,5)^\top$

As expected, the time steps become the smaller the closer the value λ is to 1, i.e., the more decay of the Lyapunov function is supposed to be guaranteed. The total number of steps for the simulations on the time interval $[0,20]$ was 28 for $\lambda = 0.1$, 52 for $\lambda = 0.5$ and 290 for $\lambda = 0.9$.

In order to investigate the efficiency of Formula (16), we have changed the definition of h_{new} in (17) by using Formula (16) with $\rho = 1.1$ instead of 0.9 (the ρ in the formula for h_{reduce} remains unchanged). With this enlarged ρ, it turns out that the condition in step (2) of Algorithm 3 is violated in more than 90% of the iterations (similar results have been obtained for Example 9, below). In contrast to that, with the choice of $\rho = 0.9$ this typically happens in less than 5% of the iterations, which shows that Formula (16) with $\rho = 0.9$ very precisely predicts a good (i.e., small enough but not too small) time step h_{i+1} for the next time step.

Example 9. Our second example is [18, Example 4.11(ii)], given by the 2d differential equation

$$\dot{z}_1 = -\|z\|^2 z_1 + z_2, \quad \dot{z}_2 = -z_1 - \|z\|^2 z_2$$

with Lyapunov function $V(z) = \|z\|^2$. Since the vector field for this example satisfies $\|f(z)\| = O(\|z\|)$ in a neighborhood of $z = 0$, one can show that in a neighborhood of $z = 0$ the consistency error of a p-th order Runge-Kutta scheme satisfies

$$\|\Phi(z_0,h) - z(h,z_0)\| = O(h^{p+1}\|z_0\|)$$

which, since V is quadratic, implies

$$\begin{aligned} \left|\widetilde{\Delta V}(z_0,h) - \Delta V(z_0,h)\right| &= \left|\Big(V(\Phi(z_0,h)) - V(z_0)\Big) - \Big(V(z(h,z_0)) - V(z_0)\Big)\right| \\ &= O(h^{p+1}\|z_0\|^2). \end{aligned}$$

On the other hand, writing the system in polar coordinates one verifies that

$$\Delta V(z_0,h) = O(h\|z_0\|^4),$$

again in a neighborhood of 0. Hence, for each fixed $h > 0$ and all z_0 sufficiently close to 0 the inequality

$$\widetilde{\Delta V}(z_0,h) < 0 \tag{18}$$

can *not* be guaranteed from these estimates, since the Lyapunov difference consistency error $|\widetilde{\Delta V}(z_0,h) - \Delta V(z_0,h)|$ is not guaranteed to be smaller than the exact decay $\Delta V(z_0,h)$. Since the analysis in [18] uses similar estimates, this explains why none of the sufficient conditions in [18] guaranteeing asymptotic stability for $h \not\to 0$ is satisfied for this example.

However, the fact that (18) is not guaranteed by this analysis does, of course, not imply that this inequality does not hold. Indeed, the fact that the difference $\|\Phi(z_0,h) - z(h,z_0)\|$ is large does not necessarily imply that the difference $|\widetilde{\Delta V}(z_0,h) - \Delta V(z_0,h)|$ is large: it may well happen that the error included in $\Phi(z_0,h)$ is large compared to $\Delta V(z_0,h)$ but nevertheless does not act destabilizing, because it changes the exact solution $z(h,z_0)$ into a direction in which V does not grow — or does even further decrease. In fact, we already observed this behavior in Example 8 for the Euler scheme and will also observe it for this example, but now for the RK4 scheme.

The Figures 5–7 show the simulation results (phase portrait, Lyapunov function $V(x(\tau_i))$ over time and the time steps) on the time interval $[0,200]$ with $\lambda = 0.5$ and $x_0 = (5,5)^\top$.

The total number of steps is 24925 for the Euler scheme, 621 for the Heun scheme and 240 for the RK2 scheme. Hence, in this example the benefit of using higher order schemes is clearly visible.

However, the advantage of the RK4 is not only due to the higher order. Looking at the step sizes one sees that for the Euler and the Heun scheme the step size is strictly decreasing after the first couple of time steps. Longer simulations indicate that the sequences indeed converge to 0 which is in accordance with the discussion above, i.e., that decay of the Lyapunov function can only be guaranteed for vanishing step

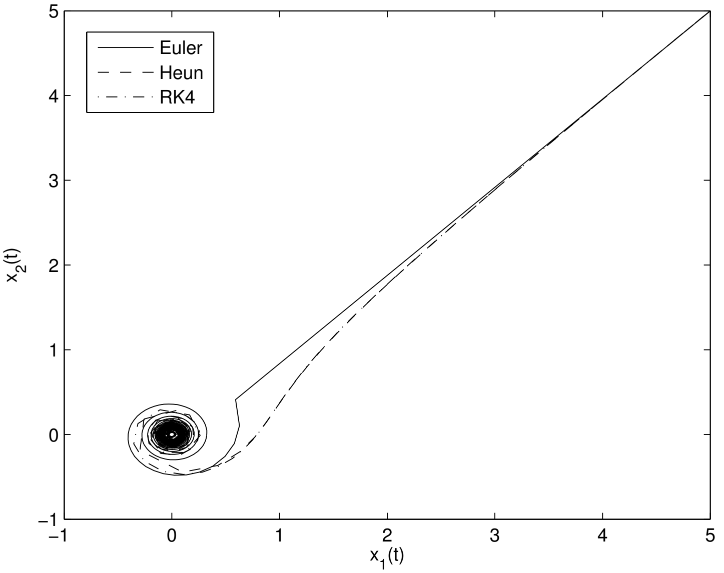


Figure 5: Phase portrait for Example 9 with $\lambda = 0.5$ and $x_0 = (5,5)^\top$

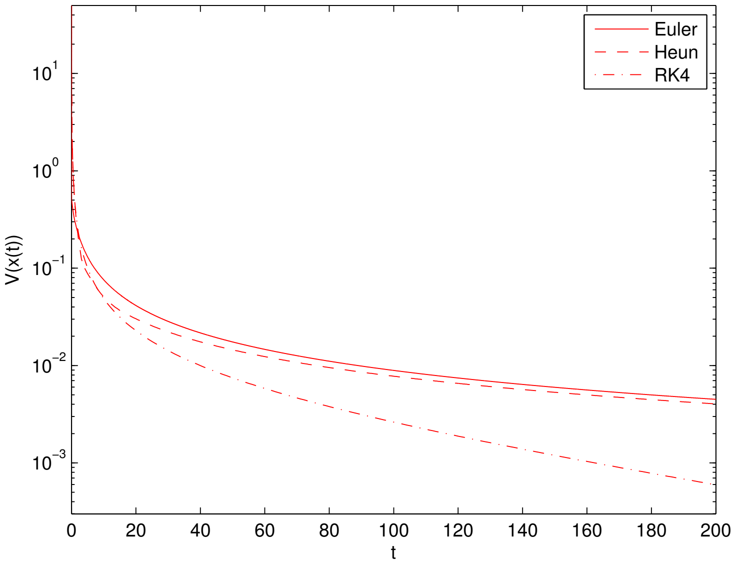


Figure 6: Lyapunov function (logarithmic scale) for Ex. 9, $\lambda = 0.5$, $x_0 = (5,5)^\top$

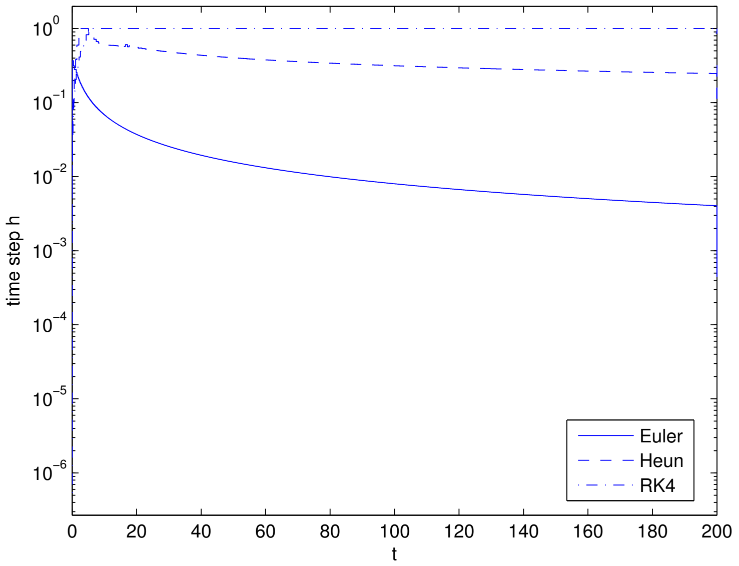


Figure 7: Time steps (logarithmic scale) for Example 9, $\lambda = 0.5$, $x_0 = (5,5)^\top$

size h if the discretization error acts destabilizing, which appears to be the case for these two schemes. In contrast to this, the error in the RK4 scheme has a stabilizing effect, because we observe a much faster decay of the Lyapunov function V than in the other examples (even faster than for the exact solutions), while the step sizes are constantly equal to the maximal allowed step size $h_{max} = 1$.

Summarizing, our examples show that the step size control scheme manages to obtain asymptotically stable solutions for different numerical schemes. A particular feature of the scheme is that step sizes are only reduced if a large error has destabilizing effect, while the scheme allows for large step sizes (and errors) as long as they do not affect the decay of the Lyapunov function.

4 Application to optimization

An obvious limitation of Algorithm 3 is that a Lyapunov function for (1) needs to be known. There are, however, several settings in which a Lyapunov function is known and yet finding a solution of (1) which converges to an equilibrium x^* of f (which in this section is typically $\neq 0$) is a meaningful problem. Examples can be found in [18, Section 6]. Here we focus on the solution of a nonlinear optimization problem (also called a nonlinear program), which are defined as follows.

$$\begin{aligned}
&\min_{x\in\mathbb{R}^m} F(x)\\
&\text{subject to}\\
&h_i(x) = 0, \quad i = 1,\dots,m\\
&g_j(x) \le 0, \quad j = 1,\dots,k.
\end{aligned} \tag{19}$$

Here $F:\mathbb{R}^m \to \mathbb{R}$, $h_i, g_j:\mathbb{R}^m \to \mathbb{R}$ for $i=1,\dots,p$ and $j=1,\dots,q$ are twice continuously differentiable functions.

The Problem (19) is well posed, e.g., if its feasible set

$$\Omega := \{x \in \mathbb{R}^m \,|\, h_i(x) = 0, i = 1,\dots,p, g_j(x) \le 0, j = 1,\dots,q\}$$

is nonempty and F is radially unbounded, or if Ω is nonempty and compact.

4.1 Differential equation approach to nonlinear optimization

The idea to solve (19) which fits our setting is now as follows: Design a differential equation

$$\dot{z} = f(z) \tag{20}$$

with state $z = (x,\bar{x}) \in \mathbb{R}^{m+k}$, which exhibits an asymptotically stable equilibrium $z^* = (x^*,\bar{x}^*)$ such that x^* is a minimizer of (19).

In order to explain how this can be accomplished, let us first look at an unconstrained problem, i.e., a problem (19) with $p = q = 0$. Then, a candidate for (20) (with $z = x \in \mathbb{R}^m$) is the (negative) gradient flow

$$f(x) := -\nabla F(x).$$

Using $V(x) := f(x) - f(x^*)$, we obtain

$$\nabla V(x) f(x) = -(\nabla F(x))^2$$

and if f is radially unbounded and $\nabla F(x) \neq 0$ for all $x \neq x^*$ (which means that x^* is the only critical point of F), then V is a Lyapunov function for f and global asymptotic stability of x^* follows from standard Lyapunov function arguments. Moreover, even though x^* and $f(x^*)$ are unknown, the Lyapunov function $V(x) := f(x) - f(x^*)$ can be used in Algorithm 3 since the term $f(x^*)$ vanishes in the derivative $\nabla V(x)$ and cancels out in the Lyapunov difference $V(\Phi(h,x)) - V(x)$. For further information on this approach for unconstrained problems we refer to [9].

For constrained problems, there are different ways to incorporate h_i and g_j into the definition of f in (20). Typically, these approaches include the constraints via suitable penalization terms in (20). In order to illustrate this concept in a simple setting, let us consider the case where no inequality constraints are present (i.e., $q = 0$) and the equality constraints are linear, i.e., of the form $Ax = b$ for a matrix A and a vector b of suitable dimension. For this setting, it was shown in [18, Section 6.1] that — under appropriate conditions on F and A — the system

$$\dot{z} = \begin{bmatrix} -(\nabla^2 F(x)(\nabla F(x) + \bar{x}^\top A)^\top + A^\top(Ax - b)) \\ -A(\nabla F(x) + \bar{x}^\top A)^\top \end{bmatrix} =: f(z)$$

has a unique asymptotically stable equilibrium $z^* = (x^*,\bar{x}^*)$ where x^* minimizes F. The corresponding Lyapunov function $V(z) = \|\nabla F(x) - \bar{x}^\top A\|^2 + \|Ax - b\|^2$ does not require the knowledge of x^* and is thus implementable in Algorithm 3. Similar constructions can be made for more general constraints, see, e.g., [1–3, 26, 27], however, not all of these approaches provide a Lyapunov function implementable in Algorithm 3 and sometimes the dynamics are only locally defined. Of course, suitable conditions on the data of (19) are needed in order to ensure that the (extended) minimizer is indeed an asymptotically stable equilibrium of (20). For this purpose, linear independence conditions on the derivatives of the constraint functions and sufficient conditions like KKT or Fritz John conditions can be employed, which we will not discuss in detail here (the interested reader is referred, e.g., to [5, 6, 8]). However, the interplay between these conditions for the approaches just cited and Algorithm 3 is still largely unexplored and will be addressed in future research.

Finally, we remark that — unless certain convexity assumptions are satisfied — is in general overly optimistic to assume that the global optimum x^* is the only equilibrium of (20). However, as our example in the next section shows, one may still be able to solve (20) using Algorithm 3 if the initial value lies in the domain of attraction. Again, the precise behavior of Algorithm 3 in this situation is subject to future research.

4.2 Optimization on manifolds: the example of the Rayleigh quotient flow

Optimization problems of the type (19) can be designed in order to solve various problems in systems theory. A comprehensive account of such techniques can be found in Helmke and Moore [16]. For many of the problems discussed in this monograph gradient flows are presented and analyzed. Typically, the optimization problems presented in [16] are posed on suitable manifolds $M \subset \mathbb{R}^n$ and the constraints in (19) then represent the condition $x \in M$.

As an example, let us look at the problem of computing the smallest eigenvalue λ_{min} of a symmetric real matrix $A \in \mathbb{R}^{n\times n}$. The minimal eigenvalue can then be computed as the minimum of the Rayleigh quotient

$$r_A(x) = \frac{x^\top Ax}{\|x\|^2}$$

over all $x \in M = \mathbb{S}^{n-1} := \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ and the minimizer $x^* \in \mathbb{S}^{n-1}$ is an associated eigenvector. Hence, λ_{min} and an associated eigenvector can be computed by solving (19) with $F(x) = x^\top Ax$ and $h_1(x) = \|x\|^2 - 1$.

The gradient flow associated to this minimization problem is the Rayleigh quotient flow

$$\dot{x} = -(A - r_A(x)I_n)x, \tag{21}$$

where I_n is the $n \times n$ identity matrix and the derivative of r_A at a point $x \in \mathbb{S}^{n-1}$ applied to ξ from the tangent space $T_x\mathbb{S}^{n-1}$ is given by

$$\nabla r_A(x)\xi = 2x^\top A\xi$$

(details on the derivation of these formulas can be found in [16, Section 1.3]).

Similar gradient flows are provided and analyzed in [16] for various other problems on manifolds M. All these flows have in common that the solution of the gradient flow stays in M, i.e., that the vector field in (20) satisfies $f(x) \in T_xM$ for all $x \in M$. Hence, for the exact solution to (20) the constraints are automatically satisfied.

However, when applying a Runge-Kutta scheme to (20), due to the discretization error $x \in M$ does in general not imply $\Phi(h,x) \in M$. One remedy for this problem is to incorporate the constraints which keep the system on M into the definition of f in (20) and to consider Φ as an "exterior" approximation to the gradient flow on M in the ambient $\mathbb{R}^n$. However, our attempt to do this for the Rayleigh quotient flow so far did not yield satisfactory results, since the solution deteriorates due to additional equilibria appearing outside $M = \mathbb{S}^{n-1}$.

Hence, as an alternative approach we use an "interior" approximation, in which we modify Φ in order to keep the numerical solution on M (for more information on this approach see [13, Chapter IV] and for its relation to differential algebraic equations see [15, Chapter VII]). This approach is possible if we can define (and implement) a projection operator $\mathbb{P}$ which maps each point in a neighborhood $\mathcal{N}$ of M to M. For $M = \mathbb{S}^{n-1}$ such a projection is simply given by $\mathbb{P}x = x/\|x\|$ for all $x \in \mathcal{N} = \mathbb{R}^n \setminus \{0\}$. Then, we may replace $\Phi(h,x)$ by $\mathbb{P}\Phi(h,x)$ and if $\mathbb{P}$ satisfies the inequality $\|\mathbb{P}x - x\| \le C\|y - x\|$ for all $x \in \mathcal{N}$, all $y \in M$ and some constant $C > 0$, then one easily verifies that for sufficiently small $h > 0$ the projected approximation $\mathbb{P}\Phi(x,h)$ is well defined and consistent with the same order of consistency as $\Phi(x,h)$. Proceding this way leads to very good results for the Rayleigh quotient flow, as the following example shows.

Example 10. We applied Algorithm 3 with $\mathbb{P}\Phi = \Phi/\|\Phi\|$ and Φ obtained by applying the Euler, Heun and RK4 scheme introduced on Section 3.3 to the Rayleigh quotient flow (21). As Lyapunov function we use $V(x) = r_A(x) - \lambda_{min}$, which, as explained in Section 4.1, can be implemented in Algorithm 3 without the knowledge of λ_{min}. Here, we use the (randomly chosen) symmetric 3×3 matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 4 \\ 3 & 4 & 11 \end{bmatrix}$$

with $\lambda_{min} \approx -0.046732641945883$ and associated eigenvector

$$x^* \approx \begin{bmatrix} 0.954876958271786 \\ -0.242466419355919 \\ -0.171522680851079 \end{bmatrix} \in \mathbb{S}^2.$$

Since the Rayleigh quotient flow has several equilibria (in fact, each eigenvalue of A is a critical value of r_A), the system is not UGAS. However, it is still UGAS on each compact subset of the domain of attraction of either x^* or $-x^*$ and if we start in such a set then the guaranteed decay of V ensures that we stay in this set. Moreover, since the set of exceptional points (i.e., the set of initial values for which the solution does not converge to either x^* or $-x^*$) is a set of lower dimension, picking a "random" initial value (in our simulation $x_0 = (1,0,0)^\top$) the probability of starting in a compact subset of the domain of attraction is very high. Due to this fact, we did not observe any problems in our numerical simulations (which showed comparable results for several other matrices we have tested).

The Figures 8–10 show the phase portrait (projected into the (x_1,x_2)-plane), the values $r_A(x(t)) - \lambda_{min}$ and the corresponding time steps. For each scheme, the simulation was stopped if the condition $|r_A(x(\tau_{i+1})) - r_A(x(\tau_i))| < 10^{-10}$ was satisfied.

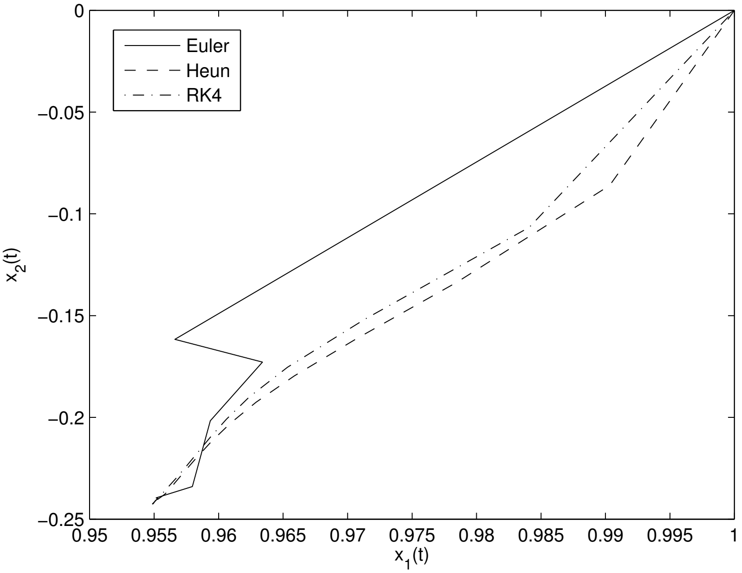


Figure 8: Phase portrait (in (x_1,x_2)-plane) for Ex. 10 with $\lambda = 0.4$ and $x_0 = (1,0,0)^\top$

The total number of steps in the computation was 13 for the Euler scheme, 32 for the Heun scheme and 26 for RK4. The value $\lambda = 0.4$ in the simulations was chosen because it turned out to yield termination in a smaller number of steps than larger or smaller choices of λ.

It is interesting to note that — as in Example 8 — the Euler scheme turns out to yield the best results since it delivers the approximation of the minimal eigenvalue λ_{min} up to the desired accuracy of 10^{-10} in the smallest number of steps, even though the solution (as clearly visible in Figure 8) is obviously not a good approximation during the transient phase. Hence, the example shows that if the emphasis lies on a numerically cheap computation of the minimum, i.e., the equilibrium, then high order schemes may not necessarily be advantageous.

For the Euler scheme (and to a lesser extent also for the Heun scheme), Figure 10 shows that the step size constantly changes between larger and smaller values. This behavior is typical for the application of Lyapunov based step size control with explicit schemes to stiff equations (cf. e.g., [18, Example 6.1]) which may lead to the conjecture that the Rayleigh quotient flow for the particular matrix A we have chosen is a moderately stiff system (even though we did not check this rigorously).

Since the Rayleigh quotient flow is a well studied systems and there are many

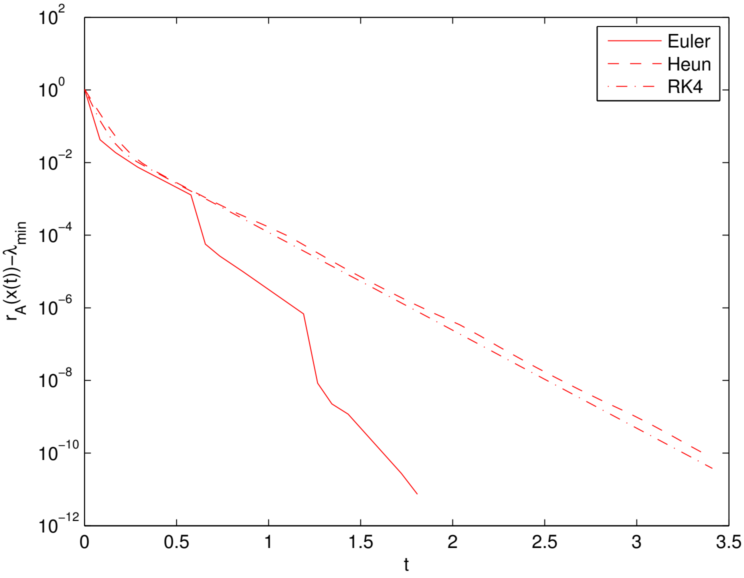


Figure 9: $r_A(x(t)) - \lambda_{min}$ (logarithmic scale) for Ex. 10, $\lambda = 0.4$, $x_0 = (1,0,0)^\top$

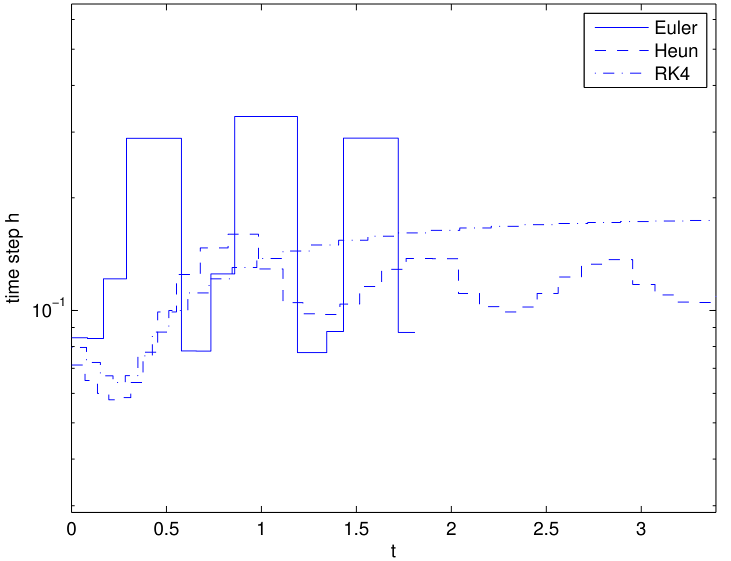


Figure 10: Time steps (logarithmic scale) for Example 10, $\lambda = 0.4$, $x_0 = (1,0,0)^\top$

known techniques for its discretization (again, we refer to [16]), we do not expect Algorithm 3 to outperform more sophisticated methods particularly tailored for the Rayleigh quotient flow. Still, our method produces very reasonable results and moreover provides valuable insights into the performance of different discretization methods.

Bibliography

[1] A. S. Antipin. Minimization of convex functions on convex sets by means of differential equations. *Differ. Equ.*, 30:1365–1375, 1995. Cited p. 204.

[2] A. A. Brown and M. C. Bartholomew-Biggs. ODE versus SQP methods for constrained optimization. *J. Optim. Theory Appl.*, 62(3):371–386, 1989. Cited p. 204.

[3] A. Cabot. The steepest descent dynamical system with control. Applications to constrained minimization. *ESAIM Control Optim. Calc. Var.*, 10(2):243–258 (electronic), 2004. Cited p. 204.

[4] P. Deuflhard and F. Bornemann. *Scientific computing with ordinary differential equations*. Springer, 2002. Cited pp. 183, 184, and 185.

[5] A. V. Fiacco and G. P. McCormick. *Nonlinear programming*. SIAM, second edition, 1990. Cited p. 204.

[6] R. Fletcher. *Practical methods of optimization*. Wiley-Interscience, second edition, 2001. Cited p. 204.

[7] B. M. Garay and K. Lee. Attractors under discretization with variable stepsize. *Discrete Contin. Dyn. Syst.*, 13(3):827–841, 2005. Cited pp. 183, 187, and 188.

[8] P. E. Gill, W. Murray, and M. H. Wright. *Practical optimization*. Academic Press, 1981. Cited p. 204.

[9] B. S. Goh. Algorithms for unconstrained optimization problems via control theory. *J. Optim. Theory Appl.*, 92(3):581–604, 1997. Cited p. 203.

[10] V. Grimm and G. R. W. Quispel. Geometric integration methods that preserve Lyapunov functions. *BIT*, 45(4):709–723, 2005. Cited p. 184.

[11] L. Grüne. *Asymptotic behavior of dynamical and control systems under perturbation and discretization*, volume 1783 of *Lecture Notes in Mathematics*. Springer, 2002. Cited pp. 183 and 190.

[12] L. Grüne. Attraction rates, robustness, and discretization of attractors. *SIAM J. Numer. Anal.*, 41(6):2096–2113, 2003. Cited pp. 183 and 190.

[13] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*. Springer, second edition, 2006. Cited pp. 184 and 205.

[14] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I Nonstiff Problems*. Springer, second edition, 1993. Cited pp. 183, 185, 189, and 195.

[15] E. Hairer and G. Wanner. *Solving ordinary differential equations. II Stiff and differential-algebraic problems*. Springer, second edition, 1996. Cited pp. 183, 184, 185, and 205.

[16] U. Helmke and J. B. Moore. *Optimization and dynamical systems*. Springer, 1994. Cited pp. 184, 204, and 208.

[17] I. Karafyllis. A system-theoretic framework for a wide class of systems. I. Applications to numerical analysis. *J. Math. Anal. Appl.*, 328(2):876–899, 2007. Cited pp. 187, 188, and 189.

[18] I. Karafyllis and L. Grüne. Feedback stabilization methods for the numerical solution of ordinary differential equations. *Discrete Contin. Dyn. Syst. Ser. B*, 16(1):283–317, 2011. Cited pp. 183, 184, 185, 186, 188, 189, 190, 192, 195, 197, 200, 202, 203, and 206.

[19] H. K. Khalil. *Nonlinear systems*. Prentice Hall, third edition, 2002. Cited pp. 188 and 193.

[20] P. E. Kloeden and J. Lorenz. Stable attracting sets in dynamical systems and in their one-step discretizations. *SIAM J. Numer. Anal.*, 23(5):986–995, 1986. Cited p. 183.

[21] P. E. Kloeden and B. Schmalfuss. Lyapunov functions and attractors under variable time-step discretization. *Discrete Contin. Dynam. Systems*, 2(2):163–172, 1996. Cited pp. 183, 187, and 188.

[22] H. Lamba. Dynamical systems and adaptive timestepping in ODE solvers. *BIT*, 40(2):314–335, 2000. Cited pp. 187 and 188.

[23] Y. Lin, E. D. Sontag, and Y. Wang. A smooth converse Lyapunov theorem for robust stability. *SIAM J. Control Optim.*, 34(1):124–160, 1996. Cited p. 188.

[24] D. Nešić, A. R. Teel, and E. D. Sontag. Formulas relating $\mathcal{KL}$ stability estimates of discrete-time and sampled-data nonlinear systems. *Syst. Control Lett.*, 38(1):49–60, 1999. Cited p. 193.

[25] A. M. Stuart and A. R. Humphries. *Dynamical systems and numerical analysis*. Cambridge University Press, 1996. Cited pp. 183, 184, 185, and 190.

[26] H. Yamashita. A differential equation approach to nonlinear programming. *Math. Programming*, 18(2):155–168, 1980. Cited p. 204.

[27] L. Zhou, Y. Wu, L. Zhang, and G. Zhang. Convergence analysis of a differential equation approach for solving nonlinear programming problems. *Appl. Math. Comput.*, 184(2):789–797, 2007. Cited p. 204.

Algebraic criteria for circuit realisations

Timothy H. Hughes
Department of Engineeering
University of Cambridge
Cambridge CB2 1PZ, UK
thh22@cam.ac.uk

Malcolm C. Smith
Department of Engineeering
University of Cambridge
Cambridge CB2 1PZ, UK
mcs@eng.cam.ac.uk

Abstract. This paper provides algebraic criteria for the number of inductors and capacitors which must be present in a realisation of a given electrical impedance function. The criteria are expressed in terms of the rank and signature of the associated Hankel, or Sylvester, or Bezoutian matrix, or equivalently in terms of an extended Cauchy index.

1 Introduction

The purpose of this paper is to provide algebraic criteria for the number of reactive elements that are needed in the realisation of a given impedance function in electrical circuits. The basis for these results is the paper of Youla and Tissi [20] which introduced the method of reactance extraction in network synthesis. There it was shown that the number of capacitors and inductors needed to realise a given driving-point behaviour is the same for *any* minimally reactive reciprocal realisation and is related to the number of positive and negative entries in a certain “reactance signature matrix” associated with the scattering matrix. In this paper we rework this result starting with the more familiar impedance function. We first relate the number of capacitors and inductors to the number of positive and negative eigenvalues of the Hankel matrix. In turn this is related to conditions on the Sylvester and Bezoutian matrices. The criteria for the latter matrices, and also in terms of an extended Cauchy index, are shown to be valid for non-proper impedances. The case of non-minimally reactive networks is also considered and the generalisation to multi-ports is discussed.

We are grateful for the opportunity provided by this Festschrift volume to acknowledge the contributions of Uwe Helmke to the field of Dynamical Systems and Control Theory in his many elegant results and papers. It is also an opportunity to thank him for his initiative in organising the workshop on “Mathematical Aspects of Network Synthesis” at the Institut für Mathematik, Universität Würzburg, 27-28 September 2010, which brought together researchers with common interests in this field, and which led to a second workshop being held on the theme in Cambridge the following year. Happy Birthday Uwe!

Mit herzlichen Glückwünschen an Professor Uwe Helmke
anlässlich seines 60. Geburtstags.

2 Notation

We denote the *rank* of a matrix by $r(\cdot)$ and the *determinant* of a square matrix by $|\cdot|$. For a real symmetric matrix we denote the number of strictly positive and strictly

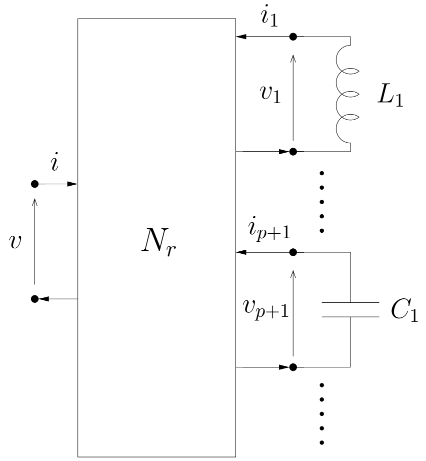


Figure 1: One-port network N with reactive elements extracted.

negative eigenvalues by $\pi(\cdot)$ and $\nu(\cdot)$ respectively. The *signature* $\sigma(\cdot)$ of a real symmetric matrix is defined by $\sigma(\cdot) = \pi(\cdot) - \nu(\cdot)$. Let $x_1,\ldots,x_r$ be a sequence of non-zero real numbers. We define $\mathbf{P}(x_1,\ldots,x_r)$ to be the number of permanences of sign and $\mathbf{V}(x_1,\ldots,x_r)$ to be the number of variations of sign in the sequence $x_1,\ldots,x_r$. We denote the set of real-rational functions in the variable s by $\mathbb{R}(s)$. The subset of *proper* rational functions, denoted by $\mathbb{R}_p(s)$, are those which are bounded at $s = \infty$. We similarly denote the set of real-rational matrix functions with r rows and c columns by $\mathbb{R}^{r\times c}(s)$ and the corresponding subset of proper real-rational matrix functions by $\mathbb{R}_p^{r\times c}(s)$. We denote the *McMillan degree* [3, Section 3.6] of a function $F(s) \in \mathbb{R}^{r\times c}(s)$ by $\delta(F(s))$. If $F(s) = a(s)/b(s) \in \mathbb{R}(s)$ with $a(s)$ and $b(s)$ coprime then $\delta(F(s)) = \max\{\deg(a(s)), \deg(b(s))\}$. The *extended Cauchy index* of a rational function or a symmetric rational matrix function (see Definitions 5 and 12) is denoted by $\gamma(F(s))$. We call a factorisation of a function $F(s) \in \mathbb{R}^{r\times c}(s)$ into the form $F(s) = B^{-1}(s)A(s)$ for $A(s)$, $B(s)$ real polynomial matrices in s a *left matrix factorisation*. For a symmetric matrix $F(s) \in \mathbb{R}^{m\times m}(s)$ with left matrix factorisation $F(s) = B^{-1}(s)A(s)$ we denote the Bezoutian by $\mathcal{B}(B,A)$ (see Sections 6 and 9). We denote by $X\dot{+}Y$ the block diagonal matrix with diagonal blocks X and Y.

3 Reactance extraction and the Hankel matrix

We begin with a function $Z(s) \in \mathbb{R}_p(s)$ with $\delta(Z(s)) = n$. Suppose $Z(s)$ is the impedance of a one-port network N containing only transformers, resistors and reactive elements (inductors and capacitors) with positive values, hereafter referred to as a *reciprocal network*. Then N contains no fewer than n reactive elements [3, Theorem 4.4.3], and is called *minimally reactive* if it contains exactly this many.

Suppose that N contains exactly p inductors and q capacitors and is minimally reactive, so $p+q = n$. Using the procedure of reactance extraction [20] N takes the form of Figure 1 where the network N_r possesses a hybrid matrix M such that

$$\begin{bmatrix} v \\ \mathbf{v_a} \\ \mathbf{i_b} \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix} \begin{bmatrix} i \\ \mathbf{i_a} \\ \mathbf{v_b} \end{bmatrix}, \tag{1}$$

where $\mathbf{i_a} = \begin{bmatrix} i_1, \ldots, i_p \end{bmatrix}^\top$ is the vector of (Laplace-transformed) currents through the inductors in N with $\mathbf{v_a}$ the corresponding voltages, $\mathbf{v_b} = \begin{bmatrix} v_{p+1}, \ldots, v_{p+q} \end{bmatrix}^\top$ is the vector of (Laplace-transformed) voltages across the capacitors in N with $\mathbf{i_b}$ the corresponding currents, and the matrix M is partitioned compatibly with the pertinent vectors. The existence of a hybrid matrix in the form (1) follows from [3, Section 4.4] and is discussed in greater detail in Section 8 of this paper. Since N_r is a reciprocal network then, by [3, Theorem 2.8.1],

$$(1 \dotplus \Sigma) M = M^\top (1 \dotplus \Sigma), \tag{2}$$

where $\Sigma = \left(I_p \dotplus -I_q\right)$. When terminated on the reactive elements we have

$$\begin{bmatrix} \mathbf{v_a} \\ \mathbf{i_b} \end{bmatrix} = -s\Lambda \begin{bmatrix} \mathbf{i_a} \\ \mathbf{v_b} \end{bmatrix},$$

where $\Lambda = \operatorname{diag}\{L_1, \ldots, L_p, C_1, \ldots, C_q\}$. Then it can readily be seen that $Z(s) = J + H(sI - F)^{-1}G$ where

$$F = -\Lambda^{-1} \begin{bmatrix} M_{22} & M_{23} \\ M_{32} & M_{33} \end{bmatrix} \in \mathbb{R}^{n \times n}, \tag{3}$$

$$G = -\Lambda^{-1} \begin{bmatrix} M_{21} \\ M_{31} \end{bmatrix} \in \mathbb{R}^{n \times 1}, \tag{4}$$

$$H = \begin{bmatrix} M_{12} & M_{13} \end{bmatrix} \in \mathbb{R}^{1 \times n}, \tag{5}$$

$$J = M_{11} \in \mathbb{R}, \tag{6}$$

and, since $\Sigma^2 = I_n$, and Σ and Λ are both diagonal, from (2) we have

$$F = \Lambda^{-1} \Sigma F^\top \Sigma \Lambda, \tag{7}$$

$$G = -\Lambda^{-1} \Sigma H^\top. \tag{8}$$

Consider the controllability and observability matrices

$$V_c = \begin{bmatrix} G, FG, \ldots, F^{n-1}G \end{bmatrix}, \tag{9}$$

$$V_o = [H^\top, F^\top H^\top, \ldots, \left(F^\top\right)^{n-1} H^\top]^\top. \tag{10}$$

Since $\delta\left(Z(s)\right) = n$ the state-space realisation (3-6) must be controllable and observable and hence V_o and V_c both have rank n. Furthermore from (7,8) we have

$$V_c = -\Lambda^{-1} \Sigma V_o^\top. \tag{11}$$

We introduce the Hankel matrix

$$\mathcal{H}_n = V_o V_c = \begin{bmatrix} h_0 & h_1 & \ldots & h_{n-1} \\ h_1 & h_2 & \ldots & h_n \\ \vdots & \vdots & \ddots & \vdots \\ h_{n-1} & h_n & \ldots & h_{2n-2} \end{bmatrix}, \tag{12}$$

where $h_i = HF^iG$ for $i = 0, 1, 2, \ldots$ are the Markov parameters, which are also directly defined from the Laurent expansion

$$Z(s) = h_{-1} + \frac{h_0}{s} + \frac{h_1}{s^2} + \frac{h_2}{s^3} + \ldots. \tag{13}$$

It follows from (11) that

$$\mathcal{H}_n = V_o\left(-\Lambda^{-1}\Sigma\right)V_o^{\top}. \tag{14}$$

From (14) and Sylvester's law of inertia [15] we deduce the following.

Theorem 1. *Let $Z(s) \in \mathbb{R}_p(s)$ with $\delta(Z(s)) = n$ and let $\mathcal{H}_n$ be as in (12) for $Z(s)$ as in (13). If $Z(s)$ is the impedance of a reciprocal network containing exactly p inductors and q capacitors with $p + q = n$ then $\pi(\mathcal{H}_n) = q$ and $\nu(\mathcal{H}_n) = p$.*

Define the infinite Hankel matrix

$$\mathcal{H} = \begin{bmatrix} h_0 & h_1 & h_2 & \ldots \\ h_1 & h_2 & h_3 & \ldots \\ h_2 & h_3 & h_4 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{15}$$

and the corresponding finite Hankel matrices

$$\mathcal{H}_k = \begin{bmatrix} h_0 & h_1 & \ldots & h_{k-1} \\ h_1 & h_2 & \ldots & h_k \\ \vdots & \vdots & \ddots & \vdots \\ h_{k-1} & h_k & \ldots & h_{2k-2} \end{bmatrix}, \tag{16}$$

for $k = 1, 2, \ldots$. Then it is known that $\mathcal{H}$ has finite rank equal to n and $|\mathcal{H}_n| \neq 0$ [10, p. 206-7]. From (14) and [9, Theorem 24, p. 343] we have the following.

Theorem 2. *Let $Z(s) \in \mathbb{R}_p(s)$ with $\delta(Z(s)) = n$ and let $\mathcal{H}_k$ be as in (16) for $Z(s)$ as in (13). If $Z(s)$ is the impedance of a reciprocal network containing exactly p inductors and q capacitors with $p + q = n$ then $|\mathcal{H}_n| \neq 0$, $|\mathcal{H}_k| = 0$ for $k > n$, and*

$$q = \mathbf{P}(1, |\mathcal{H}_1|, \ldots, |\mathcal{H}_n|), \tag{17}$$

$$p = \mathbf{V}(1, |\mathcal{H}_1|, \ldots, |\mathcal{H}_n|). \tag{18}$$

In any subsequence of zero values, $|\mathcal{H}_k| \neq 0$, $|\mathcal{H}_{k+1}| = |\mathcal{H}_{k+2}| = \ldots = 0$, signs are assigned to the zero values as follows: $\operatorname{sign}\left(|\mathcal{H}_{k+j}|\right) = (-1)^{\frac{j(j-1)}{2}} \operatorname{sign}(|\mathcal{H}_k|)$.

4 The Cauchy index and the Sylvester matrix

The *Cauchy index* of a real-rational function $F(s)$ between limits $-\infty$ and $+\infty$, denoted $I_{-\infty}^{+\infty}F(s)$, is the difference between the number of jumps of $F(s)$ from $-\infty$ to $+\infty$ and the number of jumps from $+\infty$ to $-\infty$ as s is increased in $\mathbb{R}$ from $-\infty$ to $+\infty$. From [10, Theorem 9, p. 210], if $F(s) \in \mathbb{R}_p(s)$ then $I_{-\infty}^{+\infty}F(s)$ is equal to the signature of the corresponding Hankel matrix. From Theorem 1 we deduce the following.

Theorem 3. *Let $Z(s) \in \mathbb{R}_p(s)$ be the impedance of a reciprocal network containing exactly p inductors and q capacitors and with $p+q = \delta(Z(s))$. Then*

$$q - p = I_{-\infty}^{+\infty} Z(s).$$

We now write

$$Z(s) = \frac{a(s)}{b(s)} = \frac{a_n s^n + a_{n-1}s^{n-1} + \ldots + a_0}{b_n s^n + b_{n-1}s^{n-1} + \ldots + b_0}. \tag{19}$$

Multiplying by $b(s)$ in (13) and equating coefficients of s we obtain

$$\begin{aligned} h_{-1}b_n &= a_n, \\ h_{-1}b_{n-1} + h_0 b_n &= a_{n-1}, \\ &\vdots \\ h_{-1}b_0 + h_0 b_1 + \ldots + h_{n-2}b_{n-1} + h_{n-1}b_n &= a_0, \\ h_r b_0 + h_{r+1}b_1 + \ldots + h_{r+n-1}b_{n-1} + h_{r+n}b_n &= 0, \quad (r = 0, 1, \ldots). \end{aligned}$$

Define the matrices

$$\mathcal{S}_{2k} = \begin{bmatrix} b_n & b_{n-1} & \ldots & b_{n-k+1} & b_{n-k} & \ldots & b_{n-2k+1} \\ a_n & a_{n-1} & \ldots & a_{n-k+1} & a_{n-k} & \ldots & a_{n-2k+1} \\ 0 & b_n & \ldots & b_{n-k+2} & b_{n-k+1} & \ldots & b_{n-2k+2} \\ 0 & a_n & \ldots & a_{n-k+2} & a_{n-k+1} & \ldots & a_{n-2k+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & b_n & b_{n-1} & \ldots & b_{n-k} \\ 0 & 0 & \ldots & a_n & a_{n-1} & \ldots & a_{n-k} \end{bmatrix}, \tag{20}$$

for $k = 1, 2, \ldots$, in which we put $a_j = 0$, $b_j = 0$ for $j < 0$. Following [10, p. 214] we observe that $\mathcal{S}_{2k} = \Gamma_{2k} U_{2k}$ where

$$\Gamma_{2k} = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ h_{-1} & h_0 & \ldots & h_{k-2} & h_{k-1} & \ldots & h_{2k-2} \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & h_{-1} & \ldots & h_{k-3} & h_{k-2} & \ldots & h_{2k-3} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & h_{-1} & h_0 & \ldots & h_{k-1} \end{bmatrix},$$

$$U_{2k} = \begin{bmatrix} b_n & b_{n-1} & b_{n-2} & \ldots & b_{n-2k+1} \\ 0 & b_n & b_{n-1} & \ldots & b_{n-2k+2} \\ 0 & 0 & b_n & \ldots & b_{n-2k+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & b_n \end{bmatrix}.$$

Since a sequence of $k(k-1)$ pairwise row permutations carries Γ_{2k} into a block lower triangular matrix with diagonal blocks I_k and $\mathcal{H}_k$ then

$$|\mathcal{S}_{2k}| = b_n^{2k} |\mathcal{H}_k|. \tag{21}$$

It may be observed that $|\mathcal{S}_{2n}|$ is the Sylvester resultant of $a(s)$ and $b(s)$, which is well known to be non-zero when $a(s)$ and $b(s)$ are coprime. Accordingly we will refer to the matrices $\mathcal{S}_{2k}$ in (20) as *Sylvester matrices*. If $Z(s) \in \mathbb{R}_p(s)$ then $b_n \neq 0$ and from (21) and Theorem 2 we obtain the following.

Theorem 4. *Let $Z(s) \in \mathbb{R}_p(s)$ with $\delta(Z(s)) = n$ and let $|\mathcal{S}_{2k}|$ be as in* (20) *for $Z(s)$ as in (*19*). If $Z(s)$ is the impedance of a reciprocal network containing exactly p inductors and q capacitors with $p + q = n$ then $|\mathcal{S}_{2n}| \neq 0$, $|\mathcal{S}_{2k}| = 0$ for $k > n$, and*

$$q = \mathbf{P}(1, |\mathcal{S}_2|, |\mathcal{S}_4|, \ldots, |\mathcal{S}_{2n}|),$$
$$p = \mathbf{V}(1, |\mathcal{S}_2|, |\mathcal{S}_4|, \ldots, |\mathcal{S}_{2n}|).$$

In any subsequence of zero values, $|\mathcal{S}_{2k}| \neq 0$, $|\mathcal{S}_{2(k+1)}| = |\mathcal{S}_{2(k+2)}| = \ldots = 0$, signs are assigned to the zero values as follows: $\operatorname{sign}\left(|\mathcal{S}_{2(k+j)}|\right) = (-1)^{\frac{j(j-1)}{2}} \operatorname{sign}\left(|\mathcal{S}_{2k}|\right)$.

We remark that Theorem 4 still holds when the polynomials $a(s)$ and $b(s)$ in (19) are not coprime providing we replace n with $r = \delta(a(s)/b(s))$ in the above theorem statement. Indeed the conditions $|\mathcal{S}_{2r}| \neq 0$ and $|\mathcal{S}_{2k}| = 0$ for all $k > r$ hold if and only if the function $Z(s)$ in (19) has $\delta(Z(s)) = r$ or equivalently the polynomials $a(s)$ and $b(s)$ have exactly $n - r$ roots in common.

5 Non-proper impedances and the extended Cauchy index

We consider the extension of the previous results to general rational functions (without the assumption of properness). We first introduce the following.

Definition 5. For $F(s) \in \mathbb{R}(s)$ we define the *extended Cauchy index* $\gamma(F(s))$ to be the difference between the number of jumps of $F(s)$ from $-\infty$ to $+\infty$ and the number of jumps from $+\infty$ to $-\infty$ as s increases from a point a through $+\infty$ and then from $-\infty$ to a again, for any $a \in \mathbb{R}$ which is not a pole of $F(s)$.

If $F(s)$ is proper or has a pole of even multiplicity at $s = \infty$ then $\gamma(F(s)) = I_{-\infty}^{+\infty} F(s)$. If $F(s)$ is non-proper and has a pole of odd multiplicity at $s = \infty$ then $\gamma(F(s))$ differs from $I_{-\infty}^{+\infty} F(s)$ by ±1. Note that Definition 5 does not depend on the choice of a. It is straightforward to verify the following.

Lemma 6. *Let $F(s)$, $F_1(s)$, $F_2(s) \in \mathbb{R}(s)$. Then*

1. $\gamma(F(s)) = -\gamma(1/F(s))$.
2. *If $F(s) = F_1(s) + F_2(s)$ and $\delta(F(s)) = \delta(F_1(s)) + \delta(F_2(s))$ then $\gamma(F(s)) = \gamma(F_1(s)) + \gamma(F_2(s))$.*

Now suppose that a non-proper $Z(s)$ with $\delta(Z(s)) = n$ is the impedance of a minimally reactive reciprocal network containing p inductors and q capacitors. Then $1/Z(s)$ is (strictly) proper and is the admittance of the network. Again following [3, Section 4.4, Theorem 2.8.1], reactance extraction provides a hybrid matrix M satisfying (1) with v and i interchanged, and with (2) satisfied for $\Sigma = \left(-I_p \dotplus I_q\right)$. If we now form the Hankel matrix $\mathcal{H}_n^{\dagger}$ corresponding to $1/Z(s)$ we can deduce that

$$p - q = \sigma(\mathcal{H}_n^{\dagger}) = \gamma(1/Z(s)),$$

where we have used the same reasoning as for Theorem 1 (noting the change in sign due to the change in sign in Σ) and the fact that the extended Cauchy index for a proper rational function is equal to the signature of the corresponding Hankel matrix [10, p. 210]. Hence using Lemma 6(1.) and combining with Theorem 3 for the case that $Z(s)$ is proper we obtain the following result.

Theorem 7. *Let $Z(s) \in \mathbb{R}(s)$ be the impedance of a reciprocal network containing exactly p inductors and q capacitors and with $p+q=\delta(Z(s))$. Then*

$$q-p=\gamma(Z(s)).$$

We further consider a non-proper $Z(s)$. As in Section 3 we can form Hankel matrices $|\mathcal{H}_k^\dagger|$ corresponding to $1/Z(s)$. It can then be seen that Theorem 2 holds with $Z(s)$ replaced by $1/Z(s)$, the expressions for q and p in (17,18) interchanged, and $|\mathcal{H}_k|$ replaced everywhere by $|\mathcal{H}_k^\dagger|$. Now if $Z(s)$ is written in the form (19) then $a_n \neq 0$ and we can define Sylvester matrices $\mathcal{S}_{2k}^\dagger$ corresponding to $1/Z(s)$. As in Section 4 it follows that

$$|\mathcal{S}_{2k}^\dagger| = a_n^{2k}|\mathcal{H}_k^\dagger|. \tag{22}$$

We further note that $\mathcal{S}_{2k}^\dagger$ differs from $\mathcal{S}_{2k}$ by the interchange of row i with row $i+1$ for i odd. Therefore

$$|\mathcal{S}_{2k}^\dagger| = (-1)^k|\mathcal{S}_{2k}|. \tag{23}$$

Combining the modified form of Theorem 2 with (22) and (23) we obtain the following.

Theorem 8. *Theorem* 4 *(and its subsequent remark) holds for any* $Z(s) \in \mathbb{R}(s)$.

6 The Bezoutian matrix

Let $Z(s) \in \mathbb{R}(s)$ be written as in (19). The *Bezoutian* matrix is a symmetric matrix $\mathcal{B} = \mathcal{B}(b,a)$ whose elements $\mathcal{B}_{ij}$ satisfy

$$a(w)b(z)-b(w)a(z) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{B}_{ij}z^{i-1}(z-w)w^{j-1}. \tag{24}$$

If $Z(s) \in \mathbb{R}_p(s)$ then, for $\mathcal{H}_k$ as in (16) with $Z(s)$ written as in (13), the matrix $\mathcal{B}(b,a)$ is congruent to $\mathcal{H}_n$ [8, equation 8.58]. It follows that $\gamma(Z(s)) = \sigma(\mathcal{H}_n) = \sigma(\mathcal{B}(b,a))$ and $\delta(Z(s)) = r(\mathcal{H}_n) = r(\mathcal{B}(b,a))$, these relationships holding irrespective of whether $a(s)$ and $b(s)$ are coprime. If $Z(s)$ is not proper then, since $b(s)/a(s)$ is proper and $\mathcal{B}(b,a) = -\mathcal{B}(a,b)$, we have that $\gamma(Z(s)) = -\gamma(1/Z(s)) = -\sigma(\mathcal{B}(a,b)) = \sigma(\mathcal{B}(b,a))$ and $\delta(Z(s)) = r(\mathcal{B}(a,b)) = r(\mathcal{B}(b,a))$. There is also a close relationship between the Bezoutian matrix and the Sylvester matrix. Let $Z(s)$ be as in (19) and let $\mathcal{B}_k$ be the matrix formed from the final k rows and columns of $\mathcal{B}(b,a)$, i.e.

$$\mathcal{B}_k = (\mathcal{B}_{ij})_{i,j=n-k+1}^{n}, \tag{25}$$

for $k = 1, 2, \dots, n$. Define matrices $T, P_{11}, P_{12}, P_{21}, P_{22} \in \mathbb{R}^{k \times k}$ where

$$T = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

and

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} a_{n-k} & \cdots & a_{n-2k+1} & b_{n-k} & \cdots & b_{n-2k+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & \cdots & a_{n-k} & b_{n-1} & \cdots & b_{n-k} \\ a_n & \cdots & a_{n-k+1} & b_n & \cdots & b_{n-k+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & a_n & 0 & \cdots & b_n \end{bmatrix},$$

in which we put $a_j = 0$, $b_j = 0$ for $j < 0$. Then, following [8, Theorem 8.44], the matrices P_{21} and P_{22} commute and, using a Gohberg-Semencul formula [12, Theorem 5.1], we find

$$|P| = |P_{11}P_{22} - P_{12}P_{21}| = |\mathcal{B}_k||T|.$$

Since a sequence of $k(k-1)/2$ pairwise column permutations carries T into I_k, and a sequence of $k(3k-1)/2$ pairwise column permutations followed by $k(2k-1)$ pairwise row permutations carries P into $\mathcal{S}_{2k}^{\mathsf{T}}$, it follows that

$$|\mathcal{S}_{2k}| = |\mathcal{B}_k|,$$

for $k = 1, 2, \dots, n$. Theorems 7 and 8 then lead to the following result.

Theorem 9. *Let $Z(s) \in \mathbb{R}(s)$ be as in (19) with $\delta(Z(s)) = n$. Further let $\mathcal{B}_k$ be as in (25) for $\mathcal{B}_{ij}$, $\mathcal{B}(b,a)$ defined via (24). If $Z(s)$ is the impedance of a reciprocal network containing exactly p inductors and q capacitors with $p + q = n$ then*

$$q = \frac{1}{2}(\delta(Z(s)) + \gamma(Z(s))) = \pi(\mathcal{B}(b,a)) = \mathbf{P}(1, |\mathcal{B}_1|, \dots, |\mathcal{B}_n|),$$
$$p = \frac{1}{2}(\delta(Z(s)) - \gamma(Z(s))) = \nu(\mathcal{B}(b,a)) = \mathbf{V}(1, |\mathcal{B}_1|, \dots, |\mathcal{B}_n|).$$

In any subsequence of zero values, $|\mathcal{B}_k| \neq 0, |\mathcal{B}_{k+1}| = |\mathcal{B}_{k+2}| = \dots = 0$ signs are assigned to the zero values as follows: $\operatorname{sign}\left(|\mathcal{B}_{k+j}|\right) = (-1)^{\frac{j(j-1)}{2}} \operatorname{sign}(|\mathcal{B}_k|)$.

We remark that the above theorem still holds when the polynomials $a(s)$ and $b(s)$ are not coprime providing we replace n with $r = \delta(a(s)/b(s))$ in the theorem statement.

7 Biquadratic functions

Despite their apparent simplicity the realisation of biquadratic functions has been much studied by circuit theorists. Accordingly we write down explicitly the conditions obtained in this paper which apply to this class. Let

$$Z(s) = \frac{a_2 s^2 + a_1 s + a_0}{b_2 s^2 + b_1 s + b_0}. \tag{26}$$

The Sylvester matrix $\mathcal{S}_4$ takes the form

$$\mathcal{S}_4 = \begin{bmatrix} b_2 & b_1 & b_0 & 0 \\ a_2 & a_1 & a_0 & 0 \\ 0 & b_2 & b_1 & b_0 \\ 0 & a_2 & a_1 & a_0 \end{bmatrix},$$

and we have

$$|\mathcal{S}_2| = b_2a_1 - b_1a_2,$$
$$|\mathcal{S}_4| = (b_2a_1 - b_1a_2)(b_1a_0 - b_0a_1) - (b_2a_0 - b_0a_2)^2.$$

The realisability conditions implied by Theorem 8 are shown in Table 1. Note that $|\mathcal{S}_4| > 0$ implies $|\mathcal{S}_2| \neq 0$. The conditions take the identical form if Theorem 9 is used together with the Bezoutian

$$\mathcal{B}_2 = \begin{bmatrix} b_1a_0 - a_1b_0 & b_2a_0 - a_2b_0 \\ b_2a_0 - a_2b_0 & b_2a_1 - a_2b_1 \end{bmatrix}.$$

In Table 1 it may be observed that whether the reactive elements are of the same kind, or of different kind, is determined by the sign of the resultant $|\mathcal{S}_4|$. This fact is stated by Foster [7] but no proof is provided, as noted by Kalman [14]. Also, for the case that $|\mathcal{S}_4| > 0$, [7] differentiates the 2 cases in Table 1 according to $\operatorname{sign}(b_2a_0 - a_2b_0)$ rather than $\operatorname{sign}(|\mathcal{S}_2|)$, which is easily shown to be equivalent.

Table 1 does not contain any information about synthesis, namely whether a reciprocal realisation exists for a given impedance function $Z(s)$, only the properties that a minimally reactive reciprocal realisation must satisfy if it does exist. It is well known that a function is realisable by a passive network if and only if it is positive-real. For the biquadratic (26) this is equivalent to

$$b_1a_1 - \left(\sqrt{b_0a_2} - \sqrt{b_2a_0}\right)^2 \geq 0,$$

and all coefficients in (26) have the same sign. Under this condition it is known that minimally reactive reciprocal realisations always exist [20], [3] and that transformers are not needed if $|\mathcal{S}_4| > 0$ (see Section 10). On the other hand, transformers are needed for some functions if $|\mathcal{S}_4| < 0$ [16]. Results on the classification of transformerless, minimally reactive reciprocal realisations of the biquadratic can be found in [13].

	$\lvert\mathcal{S}_2\rvert > 0$	$\lvert\mathcal{S}_2\rvert < 0$	$\lvert\mathcal{S}_2\rvert = 0$
$\lvert\mathcal{S}_4\rvert > 0$	(0,2)	(2,0)	-
$\lvert\mathcal{S}_4\rvert < 0$	(1,1)	(1,1)	(1,1)
$\lvert\mathcal{S}_4\rvert = 0$	(0,1)	(1,0)	(0,0)

Table 1: The number of reactive elements (# inductors, # capacitors) in a minimally reactive reciprocal realisation of a biquadratic.

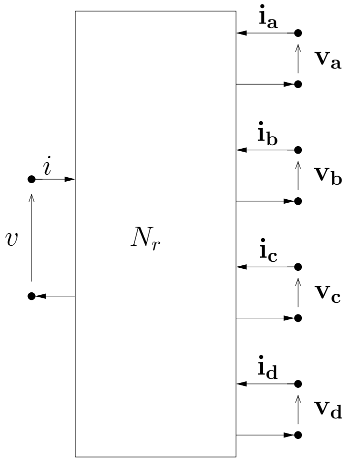


Figure 2: The network N_r obtained by removing all reactive elements from N.

8 Non-minimally reactive networks

Youla and Tissi use the scattering matrix formalism to establish lower bounds on the number of capacitors and inductors which are needed in reciprocal realisations (possibly non-minimally reactive) of a given scattering matrix [20, Theorem 2]. In this section we derive an equivalent result using the reactance extraction procedure as described in Anderson and Vongpanitlerd [3].

Let $Z(s) \in \mathbb{R}_p(s)$ be the impedance matrix of a one-port reciprocal network N containing exactly p inductors and q capacitors. Using the procedure in [3, Section 4.4], upon removal of the reactive elements in N we are left with the network N_r in Fig. 2 possessing a hybrid matrix M [3, equation 4.4.56] such that

$$\begin{bmatrix} v \\ \mathbf{v_a} \\ \mathbf{i_b} \\ \mathbf{i_c} \\ \mathbf{v_d} \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} & M_{15} \\ M_{21} & M_{22} & M_{23} & M_{24} & M_{25} \\ M_{31} & M_{32} & M_{33} & M_{34} & M_{35} \\ -M_{14}^{\mathrm{T}} & -M_{24}^{\mathrm{T}} & -M_{34}^{\mathrm{T}} & 0 & 0 \\ -M_{15}^{\mathrm{T}} & -M_{25}^{\mathrm{T}} & -M_{35}^{\mathrm{T}} & 0 & 0 \end{bmatrix} \begin{bmatrix} i \\ \mathbf{i_a} \\ \mathbf{v_b} \\ \mathbf{v_c} \\ \mathbf{i_d} \end{bmatrix},$$

where $(\mathbf{i_a}, \mathbf{v_a}), \ldots, (\mathbf{i_d}, \mathbf{v_d})$ are pairs of Laplace-transformed vectors of currents and voltages of dimensions p', q', $p-p'$, $q-q'$ respectively, and M is partitioned compatibly with the pertinent vectors. The network N is obtained upon terminating the ports corresponding to $(\mathbf{i_a}, \mathbf{v_a})$, $(\mathbf{i_c}, \mathbf{v_c})$ with inductors and the ports $(\mathbf{i_b}, \mathbf{v_b})$, $(\mathbf{i_d}, \mathbf{v_d})$ with capacitors. Then we have

$$\begin{bmatrix} \mathbf{v_a} \\ \mathbf{i_b} \end{bmatrix} = -s \begin{bmatrix} \mathcal{L}_2 & 0 \\ 0 & \mathcal{C}_3 \end{bmatrix} \begin{bmatrix} \mathbf{i_a} \\ \mathbf{v_b} \end{bmatrix},$$

$$\begin{bmatrix} \mathbf{v_c} \\ \mathbf{i_d} \end{bmatrix} = -s \begin{bmatrix} \mathcal{L}_4 & 0 \\ 0 & \mathcal{C}_5 \end{bmatrix} \begin{bmatrix} \mathbf{i_c} \\ \mathbf{v_d} \end{bmatrix},$$

where $\mathcal{L}_2 = \operatorname{diag}\left(L_1,\ldots,L_{p'}\right)$, $\mathcal{C}_3 = \operatorname{diag}\left(C_1,\ldots,C_{q'}\right)$, $\mathcal{L}_4 = \operatorname{diag}\left(L_{p'+1},\ldots,L_p\right)$ and $\mathcal{C}_5 = \operatorname{diag}\left(C_{q'+1},\ldots,C_q\right)$. It follows that equations (4.4.60) and (4.4.61) in [3, p. 195] must hold.

Since N_r is reciprocal then, by [3, Theorem 2.8.1],

$$\left(1\dotplus I_{p'}\dotplus -I_{q'}\dotplus -I_{p-p'}\dotplus I_{q-q'}\right)M = M^\top\left(1\dotplus I_{p'}\dotplus -I_{q'},\dotplus -I_{p-p'}\dotplus I_{q-q'}\right). \tag{27}$$

which implies that all entries in M_{15}, M_{25} and M_{34} are zero. Furthermore since $Z(s)$ is proper we require $M_{14} = 0$. It may then be verified that $Z(s)$ has a state-space realisation with state vector $\left[\mathbf{i_a}^\top, \mathbf{v_b}^\top\right]^\top$ with dimension $n = p' + q'$ and with $Z(s) = J + H\left(sI - F\right)^{-1}G$ where

$$F = -R\begin{bmatrix} M_{22} & M_{23} \\ M_{32} & M_{33} \end{bmatrix} \in \mathbb{R}^{n\times n}, \tag{28}$$

$$G = -R\begin{bmatrix} M_{21} \\ M_{31} \end{bmatrix} \in \mathbb{R}^{n\times 1}, \tag{29}$$

$$H = \begin{bmatrix} M_{12} & M_{13} \end{bmatrix} \in \mathbb{R}^{1\times n}, \tag{30}$$

$$J = M_{11} \in \mathbb{R}. \tag{31}$$

Here

$$R = \begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix},$$

with

$$R_{11} = \left(\mathcal{L}_2 + M_{24}\mathcal{L}_4M_{24}^\top\right)^{-1} \in \mathbb{R}^{p'\times p'},$$

$$R_{22} = \left(\mathcal{C}_3 + M_{35}\mathcal{C}_5M_{35}^\top\right)^{-1} \in \mathbb{R}^{q'\times q'},$$

where existence of $R_{11} > 0$ and $R_{22} > 0$ is guaranteed since both $\left(\mathcal{L}_2 + M_{24}\mathcal{L}_4M_{24}^\top\right)$ and $\left(\mathcal{C}_3 + M_{35}\mathcal{C}_5M_{35}^\top\right)$ are positive definite.

Let $\Sigma = \left(I_{p'}\dotplus -I_{q'}\right)$. It is straightforward to verify that $\Sigma^2 = I_n$, $\Sigma R = R\Sigma$, and both R and Σ are symmetric. Then from (27-31) we have $F = R\Sigma F^\top\Sigma R^{-1}$ and $G = -R\Sigma H^\top$. Let V_c and V_o be as in (9,10) with $\mathcal{H}_n$ as in (12). It is straightforward to show that $V_c = -R\Sigma V_o^\top$ and hence

$$\mathcal{H}_n = V_o\left(-R\Sigma\right)V_o^\top.$$

From [15, Theorem 2], the number of positive and negative eigenvalues of $\mathcal{H}_n$ cannot exceed the corresponding quantities for $-R\Sigma$. Since $-R\Sigma = \left(-R_{11}\dotplus R_{22}\right)$ with $-R_{11} < 0$ and $R_{22} > 0$, it follows that $-R\Sigma$ has exactly q' positive and p' negative eigenvalues. From the dimension of the state vector it follows that the McMillan degree of $Z(s)$ is no greater than $n = p' + q'$. Hence, for $\mathcal{H}_k$ as in (16), we have $\pi\left(\mathcal{H}_n\right) = \pi\left(\mathcal{H}_k\right)$ and $\nu\left(\mathcal{H}_n\right) = \nu\left(\mathcal{H}_k\right)$ for all $k \geq \delta\left(Z(s)\right)$, and so $\pi\left(\mathcal{H}_k\right) \leq q' \leq q$ and $\nu\left(\mathcal{H}_k\right) \leq p' \leq p$ for all $k \geq \delta\left(Z(s)\right)$.

Using the argument in Section 5 about the existence of either a proper impedance or a proper admittance we obtain the following theorem which holds irrespective of whether the network is minimally reactive or whether $a(s)$ and $b(s)$ are coprime.

Theorem 10. *Let $Z(s) \in \mathbb{R}(s)$ be as in (19). If $Z(s)$ is the impedance of a reciprocal network containing exactly p inductors and q capacitors then*

$$q \geq \frac{1}{2}\left(\delta\left(Z(s)\right)+\gamma(Z(s))\right)=\pi\left(\mathcal{B}\left(b,a\right)\right),$$
$$p \geq \frac{1}{2}\left(\delta\left(Z(s)\right)-\gamma(Z(s))\right)=\nu\left(\mathcal{B}\left(b,a\right)\right).$$

Here $\pi\left(\mathcal{B}(b,a)\right)$ and $\nu\left(\mathcal{B}(b,a)\right)$ can be calculated in accordance with Theorem 9 providing we replace n with $r=\delta\left(a(s)/b(s)\right)$.

9 Multi-port networks, generalised Bezoutians, and the extended matrix Cauchy index

The results in this paper generalise in a natural way to multi-port networks. In contrast to the one-port case there is no guarantee of existence of a proper impedance or a proper admittance function. However from [2] any reciprocal m-port network N possesses a scattering matrix description $S(s)$ where

$$\begin{bmatrix} v_1-i_1 \\ v_2-i_2 \\ \vdots \\ v_m-i_m \end{bmatrix} = S(s)\begin{bmatrix} v_1+i_1 \\ v_2+i_2 \\ \vdots \\ v_m+i_m \end{bmatrix}, \tag{32}$$

and $i_1, v_1, \ldots$ are the Laplace-transformed currents and voltages at the m ports. It is well known that $S(s) \in \mathbb{R}_p^{m\times m}(s)$ and is symmetric [20, Section 2].

Consider the transformation

$$\phi(s)=\frac{s+\alpha}{s-\alpha}, \qquad \alpha>0, \tag{33}$$

for which

$$\phi^{-1}(s)=\frac{\alpha\left(s+1\right)}{s-1},$$

which maps the left half of the s-plane onto the interior of the unit circle in the ϕ-plane. Let

$$\hat{S}(s)=S(\phi^{-1}(s)).$$

It follows from [20, Section 3] that $\hat{S}(s) \in \mathbb{R}_p^{m\times m}(s)$ is symmetric and has a realisation $\hat{S}(s)=J+H\left(sI-F\right)^{-1}G$ satisfying $J=J^{\top}$, $\Sigma F=F^{\top}\Sigma$, and $\Sigma G=H^{\top}$ where $\Sigma=\left(I_p \dotplus -I_q\right)$ with p (respectively q) the number of inductors (respectively capacitors) in N. It may then be shown that $V_c=\Sigma V_o^{\top}$ where V_c, V_o are as in (9,10) for $n=p+q \geq \delta\left(\hat{S}(s)\right)$.

Consider now the infinite Hankel matrix for $\hat{S}(s)$

$$\mathcal{H}=\begin{bmatrix} W_0 & W_1 & W_2 & \cdots \\ W_1 & W_2 & W_3 & \cdots \\ W_2 & W_3 & W_4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{34}$$

together with the finite Hankel matrices

$$\mathcal{H}_k = \begin{bmatrix} W_0 & W_1 & \cdots & W_{k-1} \\ W_1 & W_2 & \cdots & W_k \\ \vdots & \vdots & \ddots & \vdots \\ W_{k-1} & W_k & \cdots & W_{2k-2} \end{bmatrix},$$

for $k = 1,2,\ldots$ where $W_i = HF^iG$ for $i = 0,1,2\ldots$ which coincide with the matrices in the Laurent series expansion of $\hat{S}(s)$

$$\hat{S}(s) = W_{-1} + \frac{W_0}{s} + \frac{W_1}{s^2} + \frac{W_2}{s^3} + \ldots. \tag{35}$$

Then from [20, Appendix 1], $r(\mathcal{H}) = r(\mathcal{H}_k) = \delta\left(\hat{S}(s)\right)$ for all $k \geq \delta\left(\hat{S}(s)\right)$ (and indeed for all $k \geq r$ where $r \leq \delta\left(\hat{S}(s)\right)$ is the degree of the least common multiple of all denominators of $\hat{S}(s)$). Furthermore if $\hat{S}(s)$ is symmetric then so too is $\mathcal{H}$ and, as shown in [4], for $k \geq \delta\left(\hat{S}(s)\right)$ we also have $\sigma(\mathcal{H}) = \sigma(\mathcal{H}_k)$. Since $\mathcal{H}_n = V_oV_c = V_o\Sigma V_o^\top$ and $n \geq \delta\left(\hat{S}(s)\right)$ then from [15, Theorem 2] (upon a suitable bordering of the matrices $\mathcal{H}_n$ and V_o to make them square and compatible) we have the following.

Theorem 11. *Let $S(s)$ be the scattering matrix of a reciprocal m-port network containing exactly p inductors and q capacitors. Further let $\hat{S}(s) = S\left(\phi^{-1}(s)\right)$ for $\phi(s)$ as in (33). Then $\hat{S}(s) \in \mathbb{R}_p^{m\times m}(s)$ is symmetric and, with $\mathcal{H}$ as in (34) for $\hat{S}(s)$ written as in (35), we have $p \geq \pi(\mathcal{H})$ and $q \geq \nu(\mathcal{H})$.*

For $\mathcal{H}$ as in (34) with $\hat{S}(s) \in \mathbb{R}_p^{m\times m}(s)$ symmetric and written as in (35), $\sigma(\mathcal{H})$ is equal to the matrix Cauchy index of $\hat{S}(s)$ [4]. To extend these results to the case of non-proper rational matrix functions we introduce the following generalisation of the extended Cauchy index.

Definition 12. For a symmetric matrix $F(s) \in \mathbb{R}^{m\times m}(s)$ we define the extended matrix Cauchy index $\gamma(F(s))$ to be the difference between the number of jumps in the eigenvalues of $F(s)$ from $-\infty$ to $+\infty$ less the number of jumps in the eigenvalues of $F(s)$ from $+\infty$ to $-\infty$ as s increases from a point a through $+\infty$ and then from $-\infty$ to a again, for any $a \in \mathbb{R}$ which is not a pole of $F(s)$.

We remark that $\gamma(F(s))$ is well defined since the eigenvalues of $F(s)$ are defined by algebraic functions [5], and since $F(s)$ has real eigenvalues for any real s, the local power series defining them will not possess fractional powers, hence we can define an extended Cauchy index for each eigenvalue individually and then take the sum.

Definition 12 coincides with the extended Cauchy index of Definition 5 in the scalar case. Furthermore, if $F(s) \in \mathbb{R}_p^{m\times m}(s)$ then $\gamma(F(s))$ coincides with the matrix Cauchy index defined in [4]. Using results in [4] it is straightforward to show the following generalisation of Lemma 6.

Lemma 13. *Let $F(s), F_1(s), F_2(s) \in \mathbb{R}^{m\times m}(s)$ be symmetric. Then*

1. $\gamma(F(s)) = -\gamma\left(F^{-1}(s)\right)$ *when $F^{-1}(s)$ exists.*

2. *If* $F(s) = F_1(s) + F_2(s)$ *and* $\delta(F(s)) = \delta(F_1(s)) + \delta(F_2(s))$ *then* $\gamma(F(s)) = \gamma(F_1(s)) + \gamma(F_2(s))$.

Similar to the scalar case there is a correspondence between the matrix extended Cauchy index and a matrix Bezoutian. If $F(s)$ is a symmetric matrix with a left matrix factorisation $F(s) = B^{-1}(s)A(s)$ ($A(s)$ and $B(s)$ need not be left coprime) then, consistently with [4], we define the matrix Bezoutian $\mathcal{B}(B,A)$ as the symmetric matrix with block entries $\mathcal{B}_{ij}$ satisfying

$$B(z)A^{\top}(w) - A(z)B^{\top}(w) = \sum_{i=1}^{n}\sum_{i=1}^{n}\mathcal{B}_{ij}z^{i-1}(z-w)w^{j-1}.$$

This definition coincides with the definition in Section 6 in the scalar case. If $F(s) \in \mathbb{R}_p^{m\times m}(s)$ is symmetric and with left matrix factorisation $F(s) = B^{-1}(s)A(s)$ then, from [1] we have

$$\delta(F(s)) = r(\mathcal{B}(B,A)),$$

and from [4] we obtain

$$\gamma(F(s)) = \sigma(\mathcal{B}(B,A)).$$

We remark that these properties hold irrespective of whether $B(s)$ and $A(s)$ are left coprime. If $F(s)$ is not proper then consider the transformation $\phi(s)$ in (33) for any α which is not a pole of $F(s)$. Then the function $\hat{F}(s) = F\left(\phi^{-1}(s)\right) \in \mathbb{R}_p^{m\times m}(s)$ and we have $\delta\left(\hat{F}(s)\right) = \delta(F(s))$. Since $\phi(s)$ is a monotonically decreasing function of s except at $s = \alpha$, and $\phi(s)$ is rational and bounded at $s = \infty$, it follows that $\gamma\left(\hat{F}(s)\right) = -\gamma(F(s))$. Suppose in addition that $F(s)$ has a left matrix factorisation $F(s) = B^{-1}(s)A(s)$ and let n be the maximum of the degrees of the entries in the matrices $A(s)$ and $B(s)$. It follows that $\hat{F}(s)$ has a left matrix factorisation $\hat{F}(s) = \hat{B}^{-1}(s)\hat{A}(s)$ where

$$\begin{aligned}&\hat{B}(z)\hat{A}^{\top}(w) - \hat{A}(z)\hat{B}^{\top}(w)\\ &\qquad = (z-1)^n\left(B\left(\phi^{-1}(z)\right)A^{\top}\left(\phi^{-1}(w)\right) - A\left(\phi^{-1}(z)\right)B^{\top}\left(\phi^{-1}(w)\right)\right)(w-1)^n.\end{aligned}$$

Then it is straightforward to verify that

$$\mathbf{z}^{\top}\mathcal{B}\left(\hat{B},\hat{A}\right)\mathbf{w} = -2\alpha\hat{\mathbf{z}}^{\top}\mathcal{B}(B,A)\,\hat{\mathbf{w}},$$

for all z, w, where

$$\begin{aligned}\mathbf{z}^{\top} &= \begin{bmatrix}1, & z, & \cdots, & z^{n-1}\end{bmatrix},\\ \mathbf{w}^{\top} &= \begin{bmatrix}1, & w, & \cdots, & w^{n-1}\end{bmatrix},\\ \hat{\mathbf{z}}^{\top} &= (z-1)^{n-1}\begin{bmatrix}1, & \alpha\frac{z+1}{z-1}, & \cdots, & \left(\alpha\frac{z+1}{z-1}\right)^{n-1}\end{bmatrix},\\ \hat{\mathbf{w}}^{\top} &= (w-1)^{n-1}\begin{bmatrix}1, & \alpha\frac{w+1}{w-1}, & \cdots, & \left(\alpha\frac{w+1}{w-1}\right)^{n-1}\end{bmatrix}.\end{aligned}$$

It may be verified that $\hat{\mathbf{w}} = T_1 T_2 \mathbf{w}$ and $\hat{\mathbf{z}} = T_1 T_2 \mathbf{z}$ for

$$T_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ \alpha & 2^1\alpha & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha^{n-2} & \binom{n-2}{1}2^1\alpha^{n-2} & \cdots & 2^{n-2}\alpha^{n-2} & 0 \\ \alpha^{n-1} & \binom{n-1}{1}2^1\alpha^{n-1} & \cdots & \binom{n-1}{n-2}2^{n-2}\alpha^{n-1} & 2^{n-1}\alpha^{n-1} \end{bmatrix},$$

$$T_2 = \begin{bmatrix} (-1)^{n-1} & \binom{n-1}{1}(-1)^{n-2} & \cdots & \binom{n-1}{n-2}(-1) & 1 \\ (-1)^{n-2} & \binom{n-2}{1}(-1)^{n-3} & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ -1 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where $\binom{k}{r} = k!/(r!(k-r)!)$. It follows that

$$\mathcal{B}\left(\hat{B},\hat{A}\right) = (T_1T_2)^{\mathsf{T}}\left(-2\alpha\mathcal{B}(B,A)\right)(T_1T_2),$$

and hence $\gamma(F(s)) = -\gamma\left(\hat{F}(s)\right) = -\sigma\left(\mathcal{B}\left(\hat{B},\hat{A}\right)\right) = \sigma(\mathcal{B}(B,A))$ and $\delta(F(s)) = \delta\left(\hat{F}(s)\right) = r\left(\mathcal{B}\left(\hat{B},\hat{A}\right)\right) = r(\mathcal{B}(B,A))$. We have shown the following.

Lemma 14. *Let $F(s) \in \mathbb{R}^{m\times m}(s)$ be symmetric with left matrix factorisation $F(s) = B^{-1}(s)A(s)$. Then*

$$\begin{aligned} \delta(F(s)) &= r(\mathcal{B}(B,A)), \\ \gamma(F(s)) &= \sigma(\mathcal{B}(B,A)). \end{aligned}$$

We conclude by considering the case when a hybrid matrix description of the behaviour of N is available. By rearranging equation (32) we find

$$(I - \Sigma_e S(s))\begin{bmatrix}\mathbf{v}_\alpha \\ \mathbf{i}_\beta\end{bmatrix} = (I + \Sigma_e S(s))\begin{bmatrix}\mathbf{i}_\alpha \\ \mathbf{v}_\beta\end{bmatrix},$$

where $\mathbf{i}_\alpha$, $\mathbf{v}_\alpha$ are the Laplace-transformed vectors of current and voltage across the first m_1 ports, $\mathbf{i}_\beta$, $\mathbf{v}_\beta$ are the Laplace-transformed vectors of current and voltage across the remaining m_2 ports, and $\Sigma_e = \left(I_{m_1} \dotplus -I_{m_2}\right)$. Hence providing the pertinent inverse exists we have

$$\begin{bmatrix}\mathbf{v}_\alpha \\ \mathbf{i}_\beta\end{bmatrix} = M(s)\begin{bmatrix}\mathbf{i}_\alpha \\ \mathbf{v}_\beta\end{bmatrix}, \tag{36}$$

where

$$M(s)\Sigma_e = -\Sigma_e + 2\left(\Sigma_e - S(s)\right)^{-1},$$

which is symmetric. Such a Σ_e is commonly referred to as an *external signature matrix*, e.g. [19]. From the properties of the McMillan degree [3, Section 3.6] we have

$$\delta(M(s)\Sigma_e) = \delta(S(s)) = \delta\left(\hat{S}(s)\right),$$

and from Lemma 13 and the previous discussion it is straightforward to verify that

$$\gamma(M(s)\Sigma_e) = \gamma(S(s)) = -\gamma\big(\hat{S}(s)\big).$$

Combining this with Lemma 14 and Theorem 11 we obtain the following theorem which holds irrespective of whether the network is minimally reactive or whether $A(s)$ and $B(s)$ are left coprime.

Theorem 15. *Let $M(s)$ be the hybrid matrix of an m-port reciprocal network containing exactly p inductors and q capacitors, with current excitation at the first m_1 ports and voltage excitation at the remaining m_2 ports as in (36), and let $\Sigma_e = \big(I_{m_1} \dotplus -I_{m_2}\big)$. Then $M(s)\Sigma_e \in \mathbb{R}^{m\times m}(s)$ is symmetric and, with $M(s)\Sigma_e$ written as a left matrix factorisation $M(s)\Sigma_e = B^{-1}(s)A(s)$, we have*

$$q \geq \frac{1}{2}\left(\delta\left(M(s)\Sigma_e\right) + \gamma(M(s)\Sigma_e)\right) = \pi\left(\mathcal{B}(B,A)\right),$$
$$p \geq \frac{1}{2}\left(\delta\left(M(s)\Sigma_e\right) - \gamma(M(s)\Sigma_e)\right) = \nu\left(\mathcal{B}(B,A)\right).$$

10 Notes

1. (Networks with only one kind of reactive element). It follows from Theorem 7 that any minimally reactive reciprocal one-port network which contains only one kind of reactive element has an impedance function $Z(s) \in \mathbb{R}(s)$ which satisfies $\gamma(Z(s)) = \pm\delta(Z(s))$. This implies that the poles and zeros of $Z(s)$ are real and interlace each other. This is a well-known property of networks with only one kind of reactive element [18]. It is also well-known that any such impedance function can be realised without the aid of transformers in the Cauer and Foster canonical forms.

2. (Poles and zeros of impedance functions). More generally than in 1. Theorem 7 allows connections to be drawn between pole and zero locations of an impedance function $Z(s)$ and the number of inductors and capacitors in any minimally reactive reciprocal realisation of $Z(s)$. In particular, knowledge of all real axis poles and zeros and their multiplicities (including those at infinity) is sufficient to compute the extended Cauchy Index of a positive-real function.

3. (Mechanical networks). The results in this paper apply equally to mechanical networks comprising springs, dampers, inerters and levers with a direct correspondence being provided by the force-current analogy [17].

4. (Identification). The role of the Cauchy index of a proper rational function, equivalently the signature of the corresponding Hankel matrix, is well known in the subject of identification. In [11] it is shown that the $2n$-dimensional parameter space of a strictly proper rational function is divided into $n+1$ connected regions in which there are no pole-zero cancellations, with each such region being characterised by the Cauchy index, and the disconnected regions being separated by rational functions of lower McMillan degree. The original observation is credited to R.W. Brockett [11].

5. (Balanced model order reduction). The Cauchy index of a proper rational function $F(s) = d + c\left(sI - A\right)^{-1} b$ is also equal to the signature of the cross-gramian matrices $W_{co}(T) = \int_0^T e^{At}bce^{At}dt$ for $T \geq 0$, and provides insight into the effects of balanced model order reduction on the structural properties of the function [6].

Acknowledgments

This work was supported in part by the Engineering and Physical Sciences Research Council grant number EP/G066477/1.

Bibliography

[1] B. D. O. Anderson and E. I. Jury. Generalized Bezoutian and Sylvester matrices in multivariable linear control. *IEEE Transactions on Automatic Control*, 21(4):551–556, 1976. Cited p. 224.

[2] B. D. O. Anderson and R. W. Newcomb. Cascade connection for time-invariant n-port networks. *Proceedings of the IEE*, 113(6):970–974, 1966. Cited p. 222.

[3] B. D. O. Anderson and S. Vongpanitlerd. *Network Analysis and Synthesis*. Prentice-Hall, 1973. Cited pp. 212, 213, 216, 219, 220, 221, and 225.

[4] R. R. Bitmead and B. D. O. Anderson. The matrix Cauchy index: Properties and applications. *SIAM Journal on Applied Mathematics*, 33(4):655–672, 1977. Cited pp. 223 and 224.

[5] G. A. Bliss. *Algebraic functions*. AMS, 1933. Cited p. 223.

[6] K. V. Fernando and H. Nicholson. On the Cauchy index of linear systems. *IEEE Transactions on Automatic Control*, 28(2):222–224, 1983. Cited p. 226.

[7] R. M. Foster. Academic and theoretical aspects of circuit theory. *Proceeding of the IRE*, 50:866–871, 1962. Cited p. 219.

[8] P. A. Fuhrmann. *A Polynomial Approach to Linear Algebra*. Springer, second edition, 2012. Cited pp. 217 and 218.

[9] F. R. Gantmacher. *The Theory of Matrices*, volume I. Chelsea, 1980. Cited p. 214.

[10] F. R. Gantmacher. *The Theory of Matrices*, volume II. Chelsea, 1980. Cited pp. 214, 215, and 217.

[11] K. Glover. Some geometrical properties of linear systems with implications in identification. *Proceedings of the 6th IFAC World Congress, Boston*, August 24–30 1975. Cited p. 226.

[12] U. Helmke and P. A. Fuhrmann. Bezoutians. *Linear Algebra Appl.*, 122-124:1039–1097, 1989. Cited p. 218.

[13] J. Z. Jiang and M. C. Smith. Regular positive-real functions and five-element network synthesis for electrical and mechanical networks. *IEEE Transactions on Automatic Control*, 56(6):1275–1290, 2011. Cited p. 219.

[14] R. Kalman. Old and new directions of research in system theory. In J. C. Willems, S. Hara, Y. Ohta, and H. Fujioka, editors, *Perspectives in Mathematical System Theory, Control, and Signal Processing*, volume 398 of *LNCIS*, pages 3–13. Springer, 2010. Cited p. 219.

[15] A. M. Ostrowski. A quantitative formulation of Sylvester's law of inertia. *Proceedings of the National Academy of Sciences of the United States of America*, 45(5):740–744, 1959. Cited pp. 214, 221, and 223.

[16] M. Reichert. Die kanonisch und übertragerfrei realisierbaren Zweipolfunktionen zweiten Grades (Transformerless and canonic realisation of biquadratic immittance functions). *Arch. Elek. Übertragung*, 23:201–208, 1969. Cited p. 219.

[17] M. C. Smith. Synthesis of mechanical networks: The inerter. *IEEE Transactions on Automatic Control*, 47(10):1648–1662, 2002. Cited p. 226.

[18] M. E. V. Valkenburg. *Introduction to Modern Network Synthesis*. John Wiley & Sons, 1960. Cited p. 226.

[19] J. C. Willems. Realization of systems with internal passivity and symmetry constraints. *Journal of the Franklin Institute*, 301(6):605–621, 1976. Cited p. 225.

[20] D. C. Youla and P. Tissi. N-port synthesis via reactance extraction, Part I. *IEEE International Convention Record*, 14(7):183–205, 1966. Cited pp. 211, 212, 219, 220, 222, and 223.

Decentralized tracking of interconnected systems

Achim Ilchmann
Ilmenau Technical University
Ilmenau, Germany
achim.ilchmann@tu-ilmenau.de

Abstract. Decentralized funnel controllers are applied to finitely many interacting single-input single-output, minimum phase, relative degree one systems in order to track reference signals of each system within a prespecified performance funnel. The reference signals as well as the systems belong to a fairly large class. The result is a generalization of the work by [2].

1 Introduction

We generalize the early work by Helmke, Prätzel-Wolters, and Schmidt [2] who exploited the standard high-gain adaptive controller $u(t = -k(t)y(t)$, $\dot{k}(t) = y(t)^2$ (for linear minimum phase systems with relative degree one and positive high-frequency gain) to track reference signals of N systems which are interconnected. This approach, including the class of systems, the class of reference signals and internal models, the control objective, and the control strategy, is briefly summarized in Section 2.

In the present note we generalize Helmke's approach by the high-gain "funnel controller" as follows: We consider the **class of systems** described by $i = 1, \ldots, N$ interconnected single-input single-output controlled functional differential systems of the form

$$\boxed{\dot{y}_i(t) \; = \; T_i\big(y_1(\cdot), \ldots, y_N(\cdot)\big)(t) + \gamma_i v_i(t), \qquad y_i|_{[-h,0]} = y_i^0 \in C^\infty([-h,0],\mathbb{R})} \tag{1}$$

where, loosely speaking, $h \geq 0$ quantifies the "memory" of the system, $\gamma_i > 0$, and the nonlinear causal operators T_i belong to the operator class $\mathcal{T}_h^{N,1}$; see Definition 2. Note that interconnections without any structure are incorporated since every T_i depends on all $y_1(\cdot), \ldots, y_N(\cdot)$.

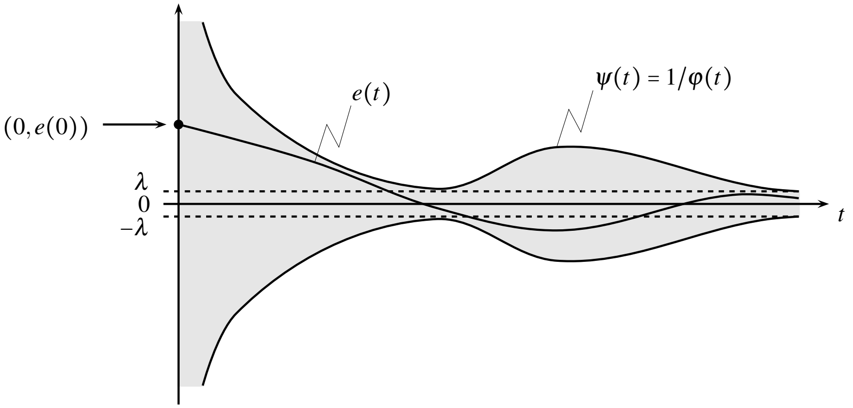
The **class of reference signals** $\mathcal{Y}_{\mathrm{ref}}$, we allow for, are all absolutely continuous functions which are bounded with essentially bounded derivative

$$\mathcal{Y}_{\mathrm{ref}} := W^{1,\infty}(\mathbb{R}_{\geq 0},\mathbb{R}) := \{y_{\mathrm{ref}}\colon \mathbb{R}_{\geq 0} \to \mathbb{R} \text{ is abs. cont.} | y_{\mathrm{ref}}, \dot{y}_{\mathrm{ref}} \in L^\infty(\mathbb{R}_{\geq 0},\mathbb{R})\} \tag{2}$$

where $L^\infty_{\mathrm{loc}}(I,\mathbb{R})$ (resp. $L^1_{\mathrm{loc}}(I,\mathbb{R})$) denote the space of measurable, locally essentially bounded (resp. locally integrable) functions $I \to \mathbb{R}$.

For the concept of "funnel control", we prespecify admissible functions φ belonging to

$$\Phi := \left\{ \varphi \in W^{1,\infty}(\mathbb{R}_{\geq 0},\mathbb{R}_{\geq 0}) \;\middle|\; \begin{array}{l} \forall\, t > 0 \;:\; \varphi(t) > 0,\ \liminf_{t\to\infty} \varphi(t) > 0, \\ \forall\, \delta > 0 \;:\; \varphi|_{[\delta,\infty)}(\cdot)^{-1} \text{ is globally Lipschitz} \end{array} \right\} \tag{3}$$



"Infinite" funnel, that is the funnel defined on $(0,\infty)$ with pole at $t = 0$.

Figure 1: Error evolution in a funnel $\mathcal{F}_\varphi$ with boundary $\psi(t) = 1/\varphi(t)$ for $t > 0$.

so that φ describes the reciprocal of the funnel boundary of the funnel

$$\mathcal{F}_\varphi := \{(t,e) \in \mathbb{R}_{\geq 0} \times \mathbb{R} \mid \varphi(t)|e| < 1\} . \tag{4}$$

See Figure 1, and Section 3.2 for a variety of funnels.

We will show that the simple **funnel controllers**

$$\boxed{v_i(t) = \frac{-\varphi_i(t)}{1-\varphi_i(t)|e_i(t)|}\, e_i(t), \qquad e_i(\cdot) = y_i(\cdot) - y_{\mathrm{ref},i}(\cdot), \qquad i = 1,\dots,N,} \tag{5}$$

achieve the **control objective**: for N prespecified performance funnels $\mathcal{F}_{\varphi_i}$, the N proportional output error feedback laws (5) applied to (1) yield a closed-loop system which has only bounded trajectories and, most importantly, each error $e_i(\cdot)$ evolves within the performance funnel $\mathcal{F}_{\varphi_i}$, for $i = 1,\dots,N$; see Figure 1 and Figure 2.

Funnel control seems advantageous when compared to high-gain adaptive control: the gain is no longer monotone but increases if necessary to exploit the high-gain property of the system and decreases if a high gain is not necessary. Most importantly, prespecified transient behaviour of the output error is addressed. Although asymptotic tracking of the reference signals is not guaranteed, the error is forced into an arbitrarily small strip; therefore, from a practical point of view this difference is negligible since the width of the funnel (see (23)) may be chosen arbitrarily small. Moreover, funnel control allows for much more general system classes and reference classes than in [2] and the interconnection between the subsystems is not limited as in [2]. If an identical reference trajectory is chosen for every subsystem, our control strategy could be called synchronization of interconnected systems. Decentralized funnel control for interconnected systems is the main contribution of the present note and it is treated in Section 3. We finalize the paper with some illustrative simulation in Section 4.

2 The approach by Uwe Helmke and coworkers

In the present section, the approach by Helmke, Prätzel-Wolters, and Schmidt [2] is summarized; the generalized approach will then be related to the latter in Section 3.

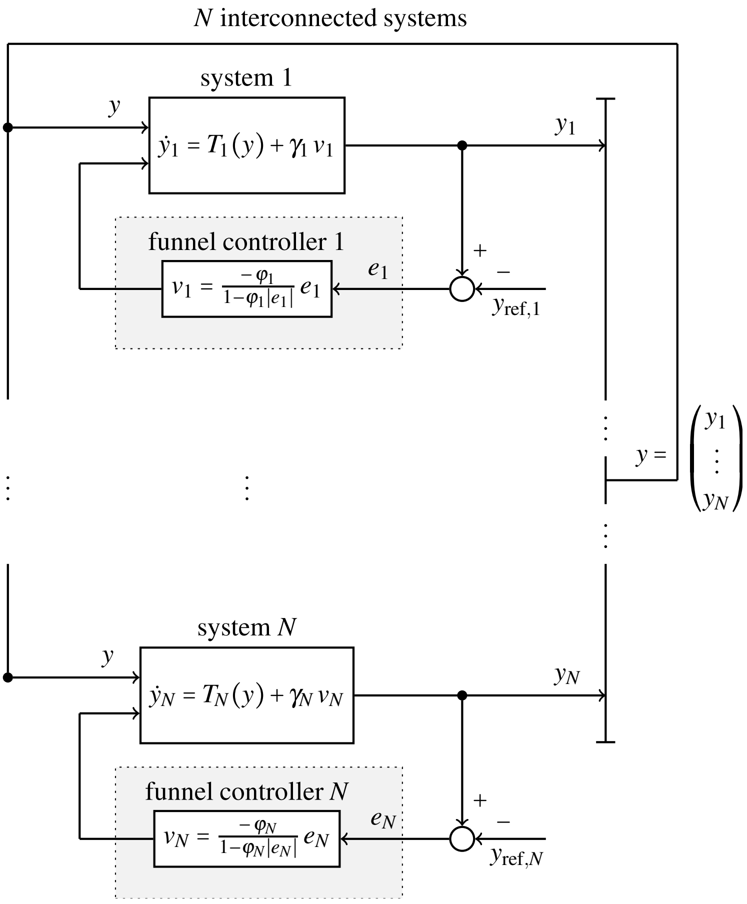

N interconnected systems
system 1
y
$\dot y_1 = T_1(y) + \gamma_1 v_1$
y_1
funnel controller 1
$v_1 = \frac{-\varphi_1}{1-\varphi_1|e_1|} e_1$
e_1
+
−
$y_{\text{ref},1}$
$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$
system N
y
$\dot y_N = T_N(y) + \gamma_N v_N$
y_N
funnel controller N
$v_N = \frac{-\varphi_N}{1-\varphi_N|e_N|} e_N$
e_N
+
−
$y_{\text{ref},N}$

Figure 2: Decentralized funnel control of N interconnected systems

Roughly speaking, the underlying idea is to combine adaptive high-gain controllers and internal models (generating the signals to be tracked) to interconnected high-gain stabilizable, relative degree one systems; then tracking of reference signals of each subsystem is achieved if the interconnection has a certain structure which preserves for the interconnected system the minimum phase property inherited from the subsystems.

2.1 Class of linear systems

Consider $i = 1, \ldots, N$ interconnected single-input single-output systems of the form

$$\boxed{\begin{aligned} \dot x_i(t) &= A_i x_i(t) + b_i u_i(t) \\ y_i(t) &= c_i x_i(t) \end{aligned}} \tag{6}$$

which all satisfy, for (unknown) $A_i \in \mathbb{R}^{n_i \times n_i}$, $b_i, c_i^\top \in \mathbb{R}^{n_i}$, the structural properties

$$\textit{positive high-frequency gain} \text{ and } \textit{relative degree one}, \text{ i.e. } c_i b_i > 0 \tag{7}$$

$$\textit{minimum phase}, \text{ i.e. } \quad \det \begin{bmatrix} sI_n - A_i & b_i \\ c_i & 0 \end{bmatrix} \neq 0 \ \forall s \in \overline{\mathbb{C}}_+ \tag{8}$$

$$u(t) = Fy(t) + v(t), \quad \text{for some } F \in \mathbb{R}^{N \times N}$$
$$\text{with } \textit{interconnection structure } f_{ij} \ker c_j \subset \operatorname{im} b_i \text{ for } i \neq j, \tag{9}$$

where $u(t) = (u_1(t), \dots, u_N(t))^\top$, $y(t) = (y_1(t), \dots, y_N(t))^\top$, $v(t) = (v_1(t), \dots, v_N(t))^\top$, and v denotes the N-dimensional input of the interconnected system.

It was well-known in the high-gain adaptive control community that the structural properties (7) and (8) allow for a simple adaptive high-gain controller

$$u_i(t) = -k_i(t) y_i(t), \qquad \dot{k}_i(t) = y_i(t)^2, \tag{10}$$

which, if applied to (6) for arbitrary initial data $x_i(0) = x_i^0 \in \mathbb{R}^{n_i}$, $k_i(0) = k_i^0 \in \mathbb{R}$, yields in a closed-loop system (6), (10), and this system has a unique global solution and satisfies

$$\lim_{t \to \infty} y_i(t) = 0, \quad \lim_{t \to \infty} k_i(t) = k_i^\infty \in \mathbb{R}, \quad x_i(\cdot) \in L^\infty(\mathbb{R}_{\geq 0}, \mathbb{R}^{n_i});$$

see, for example, [7] or [12]. One important issue of this approach is that no information on the system entries of (6) are incorporated in the feedback controller. However, one drawback is monotonically increasing gain functions $t \mapsto k_i(t)$ which may have a large limit and so possible noise in the output measurement is amplified.

2.2 Control objective

Let $y_{\text{ref},i} : \mathbb{R}_{\geq 0} \to \mathbb{R}$ denote N reference signals which are periodic and satisfy a linear differential equation

$$y_{\text{ref},i}(\cdot) \in \ker P_i(\tfrac{\mathrm{d}}{\mathrm{d}t}) := \left\{ \zeta(\cdot) \in C^\infty(\mathbb{R}_{\geq 0}, \mathbb{R}) \mid P_i(\tfrac{\mathrm{d}}{\mathrm{d}t})\zeta = 0 \right\}$$

for given $P_i(s) \in \mathbb{R}[s]$, $i = 1, \dots, N$. The control objective is to find N decentralized adaptive controllers depending on the tracking error

$$e_i(\cdot) := y_i(\cdot) - y_{\text{ref},i}(\cdot) \mapsto v_i(\cdot)$$

in combination with an internal model (depending on $P_1(s), \dots, P_N(s)$) so that the closed-loop system has only bounded trajectories and the tracking errors satisfy, for any initial conditions,

$$\lim_{t \to \infty} e_i(t) = 0 \qquad \forall \, i = 1, \dots, N.$$

2.3 Adaptive high-gain controller

Before we state the main result of [2] which is the following Theorem 1, we stress the underlying ideas of this result:

- The N systems in (6) may be written as one system with N inputs u and N outputs y, and the latter has strict relative degree one with high-frequency gain matrix $diag\{c_1b_1,\dots,c_nb_n\}$ and it inherits the minimum phase property.
- The polynomials $P_i(s)$ allow to design an internal model so that the reference signals are, for suitable initial values, the output of the internal model.
- The special interconnection structure by F in (9) preserves the strict relative degree one and minimum phase property of the multi-input multi-output system $v \mapsto y$.
- The adaptive high-gain controllers in (10) are applicable.

Theorem 1. *[2, Th. 2.4]*
Consider N interconnected systems as in (6)-(9). *Let* $P_i(s) \in \mathbb{R}[s]$ *such that* $\ker P_i(\frac{\mathrm{d}}{\mathrm{d}t})$ *contains periodic solutions only;* $i = 1,\dots,N$. *Choose a Hurwitz polynomial* $Q(s) \in \mathbb{R}[s]$ *such that*

$$\ell := \deg Q(s) = \deg P(s) \qquad \textit{where } P(s) = \operatorname{lcm}\{P_1(s)\dots,P_N(s)\}$$

and a minimal realization $(A_r,b_r,c_r) \in \mathbb{R}^{\ell\times\ell}\times\mathbb{R}^{\ell}\times\mathbb{R}^{1\times\ell}$ *such that*

$$c_r(sI_\ell - A_r)^{-1}b_r + 1 \;=\; \frac{Q(s)}{P(s)}\,. \tag{11}$$

Then for any reference signals $y_{\mathrm{ref},i}(\cdot) \in \ker P_i(\frac{\mathrm{d}}{\mathrm{d}t})$ *and any initial conditions* $x_i^0 \in \mathbb{R}^{n_i}$, $z_i^0 \in \mathbb{R}^{\ell}$, $k_i^0 \in \mathbb{R}$, *the N decentralized high-gain controllers* $e_i := y_i - y_{\mathrm{ref},i} \mapsto v_i$ *given by*

$$\boxed{\begin{array}{ll} \dot z_i(t) = A_r z_i(t) - b_r k_i(t)\,e_i(t), & z_i(0) = z_i^0 \\ \dot k_i(t) = e_i(t)^2, & k_i(0) = k_i^0 \\ y_i(t) = c_i x_i(t) & \\ v_i(t) = c_r z_i(t) - k_i(t)\,e_i(t)\,, & e_i(\cdot) = y_i(\cdot) - y_{\mathrm{ref},i}(\cdot) \end{array}} \tag{12}$$

applied to (6), (9) *yield a closed-loop system* (6), (9), (12) *which has solution, this solution is global and unique and satisfies, for* $i = 1,\dots,N$,

$$\lim_{t\to\infty} e_i(t) = 0, \quad \lim_{t\to\infty} k_i(t) \in \mathbb{R}, \quad x_i(\cdot) \in L^\infty(\mathbb{R}_{\ge 0},\mathbb{R}^{n_i}), \quad z_i(\cdot) \in L^\infty(\mathbb{R}_{\ge 0},\mathbb{R}^{\ell})\,.$$

3 Main result

In this section we show how to generalize Theorem 1 in the following sense: The restriction of the interconnection (9) between the systems is superfluous. We allow for systems described by functional differential equations encompassing nonlinear systems, infinite dimensional systems, systems with hysteresis such as relay or backlash. The class of reference signals are arbitrary signals which are bounded and have essentially bounded derivative; an internal model as in (11) is not needed. Furthermore, the control strategy does not involve a monotonically increasing gain $k_i(\cdot)$ as in (10) but a gain which is large if "necessary" and decreases thereafter. The control strategy will obey prespecified transient behaviour.

3.1 Class of systems

We consider $i = 1, \dots, N$ interconnected single-input single-output systems described by controlled functional differential equations of the form (1) where, loosely speaking, $h \geq 0$ quantifies the "memory" of the system, $\gamma_i > 0$, and the nonlinear causal operators T_i belong to the following operator class $\mathcal{T}_h^{N,q}$. Note that interconnections without any structure are incorporated since every T_i depends on all $y_1(\cdot), \dots, y_N(\cdot)$.

Definition 2 (Operator class $\mathcal{T}_h^{N,q}$)**.** [3]
Let $h \geq 0$, $N, q \in \mathbb{N}$. An operator T is said to be of class $\mathcal{T}_h^{N,q}$ if, and only if, the following hold:

(i) $T : C([-h,\infty),\mathbb{R}^N) \to L^\infty_{\mathrm{loc}}(\mathbb{R}_{\geq 0},\mathbb{R}^q)$ is a causal operator.

(ii) $\forall\, t \geq 0\ \forall\, w \in C([-h,t],\mathbb{R}^N)\ \exists\, \tau > t,\ \exists\, \delta, \Delta > 0\ \forall\, y, z \in C(w; h, t, \tau, \delta, N)\quad :$

$$\operatorname*{ess-sup}_{s\in[t,\tau]} \|(Ty)(s) - (Tz)(s)\| \ \leq\ \Delta \cdot \max_{s\in[t,\tau]} \|y(s) - z(s)\|\,,$$

where $C(w; h, t, \tau, \delta, N)$ denotes the space of all continuous extensions z of $w \in C([-h,t],\mathbb{R}^N)$ to the interval $[-h,\tau]$ with the property that $\|z(s) - w(t)\| \leq \delta$.

(iii) $\forall\, \delta > 0\ \exists\, \Delta > 0\ \forall\, y \in C([-h,\infty),\mathbb{R}^N)$ with $\sup_{s\in[-h,\infty)} \|y(s)\| \leq \delta \quad :$

$$\|(Ty)(t)\| \leq \Delta \ \text{ for almost all } t \geq 0\,.$$

The crucial property is Property (iii): a bounded-input, bounded-output assumption on the operator T. Property (ii) is a technical assumption of local Lipschitz type which is used in establishing well-posedness of the closed-loop system. To interpret this assumption correctly, we need to give meaning to Ty for a function $y \in C(I,\mathbb{R}^N)$ on a bounded interval I of the form $[-h,\rho)$ or $[-h,\rho]$, where $0 < \rho < \infty$. This we do by showing that T "localizes" to an operator $\tilde{T} : C(I,\mathbb{R}^N) \to L^\infty_{\mathrm{loc}}(J,\mathbb{R}^N)$, where $J := I \setminus [-h,0)$. Let $y \in C(I)$. For each $\sigma \in J$, define $y_\sigma \in C([-h,\infty),\mathbb{R}^N)$ by

$$y_\sigma(t) := \begin{cases} y(t), & t \in [-h,\sigma], \\ y(\sigma), & t > \sigma\,. \end{cases}$$

By causality, we may define $\tilde{T}y \in L^\infty_{\mathrm{loc}}(J,\mathbb{R}^N)$ by the property $\tilde{T}y|_{[0,\sigma]} = Ty_\sigma|_{[0,\sigma]}$ for all $\sigma \in J$. Henceforth, we will not distinguish notationally an operator T and its "localization" $\tilde{T}$: the correct interpretation being clear from context.

In the following we will show the wide range of system classes which can be written in the form (1) with an operator $T_i\big(y_1(\cdot), \dots, y_N(\cdot)\big)$ belonging to the class $\mathcal{T}_h^{N,q}$.

3.1.1 Linear systems

We first study the linear prototype of systems of the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + bu(t), \quad x(0) = x^0 \\ y(t) &= cx(t) \end{aligned} \tag{13}$$

with $A \in \mathbb{R}^{n\times n}$, $b, c^\top \in \mathbb{R}^n$, $x^0 \in \mathbb{R}^n$ and relative degree one, i.e. $cb \neq 0$. We show that the interconnected systems (1) is a generalization of the interconnected system (6), (9). In our setup, with a slightly different control objective (see Section 2.2) than 1, the special assumption on F in (9) is superfluous.

Clearly, (13) has relative degree one if, and only if, $\mathbb{R}^n = \operatorname{im} b \oplus \ker c$. If this is the case, then there exists $V \in \mathbb{R}^{n\times(n-1)}$ with $\operatorname{im} V = \ker C$ such that the coordinate transformation

$$x \mapsto \begin{bmatrix} y \\ z \end{bmatrix} := S^{-1}x \quad \text{where } S := \left[b(cb)^{-1}, V\right]$$

takes (13) into the equivalent form

$$\begin{aligned} \dot{y}(t) &= A_1 y(t) + A_2 z(t) + cbu(t), \quad & y(0) &= y^0 \\ \dot{z}(t) &= A_3 y(t) + A_4 z(t), & z(0) &= z^0, \end{aligned} \tag{14}$$

with $z(t) \in \mathbb{R}^{n-1}$ and real matrices A_1, A_2, A_3, A_4 of conforming formats. This allows to rewrite (13) in terms (14) and the linear and causal operator,

$$\begin{aligned} T^{z^0} : C(\mathbb{R}_{\geq 0}, \mathbb{R}) &\to C(\mathbb{R}_{\geq 0}, \mathbb{R}) \\ y(\cdot) &\mapsto \left(t \mapsto A_1 y(t) + A_2 \left[e^{A_4 t} z^0 + \int_0^t e^{A_4(t-\tau)} A_3 y(\tau) d\tau\right]\right), \end{aligned} \tag{15}$$

parametrized by $z^0 \in \mathbb{R}^{n-1}$, as a functional differential equation in $y(\cdot)$ only:

$$\dot{y}(t) \ = T^{z^0} y(\cdot)(t) + cbu(t), \qquad y(0) = y^0. \tag{16}$$

If (13) is minimum phase (see (8)), then equivalently $\sigma(A_4) \subset \mathbb{C}_-$; and hence the operator T^{z^0} has the crucial property

$$\forall\, \delta > 0\ \exists \Delta > 0\ \forall\, y(\cdot) \in L^\infty(\mathbb{R}_{\geq 0}, \mathbb{R}) \text{ with } \|y\|_\infty < \delta\ :\ \|T^{z^0} y\|_\infty < \Delta, \tag{17}$$

and it is readily checked that T^{z^0} belongs to the class $\mathcal{T}_0^{1,1}$. Therefore, each minimum phase system (6) with positive high-frequency gain $c_i b_i > 0$ can be equivalently written in the form (1).

Next we consider the class of systems (6) which satisfy the structural properties (7) and (8), write them in the form

$$\dot{y}_i(t) \ = \ T_i^{z_i^0}\big(y_i(\cdot)\big)(t) + c_i b_i u_i(t), \qquad y_i = cx_i^0 \tag{18}$$

and interconnect them with the feedback (9). Writing $F = \begin{bmatrix} f^1 \\ \cdots \\ f^N \end{bmatrix}$, this results in

$$\dot y_i(t) \; = \; \underbrace{T_i^{z_i^0}\big(y_i(\cdot)\big)(t) + c_i b_i \left[f^i y(t)\right]}_{=:T_i^{z^0}(y(\cdot))(t)} + c_i b_i v_i(t)\,, \qquad y_i = c x_i^0$$

and the so defined operator $y(\cdot) \mapsto T_i^{z^0}(y)(\cdot)$ also belongs to class $\mathcal{T}_0^{1,1}$ and we arrive at the structure of (1).

3.1.2 Infinite dimensional linear systems

The finite-dimensional class of systems of the form (13) can be extended to infinite dimensions by reinterpreting the operators A_j in (14) as the generating operators of a regular linear system (regular in the sense of [11]). In the infinite-dimensional setting, A_4 is assumed to be the generator of a strongly continuous semigroup $\mathbf{S} = (\mathbf{S}_t)_{t \ge 0}$ of bounded linear operators and a Hilbert space X with norm $\|\cdot\|_X$. Let X_1 denote the space $\operatorname{dom}(A_4)$ endowed with the graph norm and let X_{-1} denote the completion of X with respect to the norm $\|z\|_{-1} = \|(s_0 I - A_4)^{-1} z\|_X$, where s_0 is any fixed element of the resolvent set of A_4. Then A_3 is assumed to be a bounded linear operator from $\mathbb{R}$ to X_{-1} and A_2 is assumed to be a bounded linear operator from X_1 to $\mathbb{R}$. Assuming that the semigroup $\mathbf{A}_4$ is exponentially stable and that $\mathbf{A}_4$ extends to a bounded linear operator (again denoted by $\mathbf{A}_4$) from X to $\mathbb{R}$, then the operator T given by

$$(Ty)(t) \; := \; A_1(t) y(t) + A_2 \left[\mathbf{S}_t z^0 + \int_0^t \mathbf{S}_{t-\tau} A_3 y(\tau)\, \mathrm{d}\tau\right]$$

is of class $\mathcal{T}_0^{1,1}$ and we arrive at the structure of (1). For more details see [8], and for a similar but more general approach see [3, Appendix A.2].

3.1.3 Nonlinear systems

Consider the following nonlinear generalization of (14):

$$\begin{aligned} \dot y(t) &= f(p(t), y(t), z(t)) + g(y(t), z(t), u(t)), \qquad & y(0) &= y^0 \in \mathbb{R} \\ \dot z(t) &= h(t, y(t), z(t)), & z(0) &= z^0 \in \mathbb{R}^{n-1} \end{aligned} \tag{19}$$

with continuous

$$f : \mathbb{R}^P \times \mathbb{R} \times \mathbb{R}^{n-1} \to \mathbb{R}, \quad g : \mathbb{R} \times \mathbb{R}^{n-1} \times \mathbb{R} \to \mathbb{R}, \quad h : \mathbb{R}_{\ge 0} \times \mathbb{R} \times \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$$

having the properties: $h(\cdot, y, z)$ measurable for all $(y, z) \in \mathbb{R} \times \mathbb{R}^{n-1}$ and

$$\begin{aligned} &\forall \text{ compact } \mathcal{C} \subset \mathbb{R} \times \mathbb{R}^{n-1} \; \exists \; \kappa \in L^1_{\mathrm{loc}}(\mathbb{R}_{\ge 0}, \mathbb{R}) \text{ for a.a. } t \ge 0 \; \forall \; (y,z), (\bar y, \bar z) \in \mathcal{C} \\ &\qquad : \; \|h(t, y, z) - h(t, \bar y, \bar z)\| \le \kappa(t) \|(y, z) - (\bar y, \bar z)\|\,. \end{aligned}$$

Then, viewing the second of the differential equations in (19) in isolation (with input y), it follows that, for each $(z^0, y) \in \mathbb{R}^{n-1} \times L^\infty_{\mathrm{loc}}(\mathbb{R}_{\ge 0}, \mathbb{R})$, the initial-value

problem $\dot{z}(t) = h(t,y(t),z(t))$, $z(0) = z^0 \in \mathbb{R}^{n-1}$, has unique maximal solution, which we denote by $[0,\omega) \to \mathbb{R}^{n-1}$, $t \mapsto z(t;z^0,y)$.
In addition, we assume

$$\exists c_0 > 0 \; \exists q > 1 \; \forall (u,y,z) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n-1} \; : \; u \cdot g(y,z,u) \geq c_0 \, |u|^q \tag{20}$$

and

$$\begin{aligned} &\exists \theta \in C(\mathbb{R}_{\geq 0}, \mathbb{R}_{\geq 0}) \; \exists c > 0 \; \forall y \in L^\infty_{\text{loc}}(\mathbb{R}_{\geq 0}, \mathbb{R}) \; \forall t \in [0,\omega) \\ &\qquad : \; \|z(t,z^0,y)\| \leq c \, [1 + \operatorname*{ess-sup}_{s \in [0,t]} \theta(|y(s)|)] \end{aligned} \tag{21}$$

which, in turn, implies that $\omega = \infty$. Note that this is akin to, but weaker than, Sontag's [9] concept of input-to-state stability. Now fix $z^0 \in \mathbb{R}^{n-1}$ arbitrarily, and define the operator

$$T : C(\mathbb{R}_{\geq 0}, \mathbb{R}) \to L^\infty_{\text{loc}}(\mathbb{R}_{\geq 0}, \mathbb{R} \times \mathbb{R}^{n-1}), \quad y \mapsto Ty = (y(\cdot), z(\cdot, z^0, y)) \, .$$

In view of (21), Property (ii) of Definition 2 holds; setting $h = 0$, we see that Property (iii) of Definition 2 also holds. Arguing as in [8, Sect. 3.2.3], via an application of Gronwall's Lemma, it can be shown that Property (iii)(b) holds. Therefore, this construction yields a family (parameterized by the initial data z^0) of operators T of class $\mathcal{T}_0^{1,n}$. Therefore, (19) is equivalent to

$$\dot{y}(t) \; = \; f(p(t), (Ty)(t)) + g((Ty)(t), u(t)) \, . \tag{22}$$

Clearly, (22) is not of the form (1). However, the nonlinear function $g((Ty)(t), u(t))$ compared to $\gamma u(t)$ allows for high-gain stabilization since assumption (20) yields, for any compact set $\mathcal{C} \subset \mathbb{R}^P \times \mathbb{R}^{M \times L}$,

$$\forall u \in \mathbb{R} \; : \; \min_{(v,w) \in \mathcal{C}} \frac{u \, [f(v,w) + g(w,u)]}{|u|} \; \geq \; - \max_{(v,w) \in \mathcal{C}} |f(v,w)| + c_0 \, |u|^{q-1} \, ,$$

and further, this gives the following condition (akin to radial unboundedness or weak coercivity)

$$\forall (u_n) \in (\mathbb{R}^*)^{\mathbb{N}} \text{ with } \lim_{n \to \infty} |u_n| = \infty \; : \; \lim_{n \to \infty} \min_{(v,w) \in \mathcal{C}} \frac{u_n \, [f(v,w) + g(w,u_n)]}{|u_n|} = \infty \, .$$

Now our general result Theorem 3 can be shown if condition (20) holds, but we omit this to keep the presentation simple; for details see [5, Remark 4(iv)]. The other reason why (22) is not of the form (1) is the first summand in (22). Again, for technical reasons we omit to show how to incorporate this more general form but refer to Step 1 of the proof of Theorem 3: the arguments used their indicate how the right hand side of (1) could be generalized. Under the assumption that N systems of the form (19) can be written in a feasible form, we may interconnect them via

$$u(t) \; = \; F(y(t)) + v(t) \, , \quad \text{for some continuous } F : \mathbb{R}^N \to \mathbb{R}^N$$

and we arrive at the structure of (1).

3.1.4 Nonlinear delay systems

Let functions

$$\mathcal{G}_i : \mathbb{R} \times \mathbb{R}^{\ell} \to \mathbb{R}^q : (t, \zeta) \mapsto \mathcal{G}_i(t, \zeta), \qquad i = 0, \dots, n,$$

be measurable in t and locally Lipschitz in ζ uniformly with respect to t: precisely,

(i) $\forall \zeta \in \mathbb{R}^{\ell} \; : \; \mathcal{G}_i(\cdot, \zeta)$ is measurable;

(ii) $\forall$ compact $\mathcal{K} \subset \mathbb{R}^l \; \exists c > 0$ for a.a. $t \geq 0 \; \forall \zeta, \psi \in \mathcal{K}$
$: \; \|\mathcal{G}_i(t, \zeta) - \mathcal{G}_i(t, \psi)\| \leq c \, \|\zeta - \psi\|$.

For $i = 0, \dots, n$, let $h_i \geq 0$ and define $h := \max_i h_i$. The operator T, defined for $\zeta \in C([-h, \infty), \mathbb{R}^l)$ by

$$(T\zeta)(t) := \int_{-h_0}^{0} \mathcal{G}_0(s, \zeta(t+s)) \, \mathrm{d}s + \sum_{i=1}^{n} \mathcal{G}_i(t, \zeta(t - h_i)) \quad \forall t \geq 0.$$

is of class $\mathcal{T}_h^{\ell,q}$; for details see [8].

3.1.5 Systems with hysteresis

A general class of hysteresis operators, which includes many physically motivated hysteretic effects, is discussed in [6]. Examples of such operators include backlash hysteresis, elastic-plastic hysteresis, and Preisach operators. In [4], it is pointed out that these operators are of class $\mathcal{T}_0^{1,1}$. For illustration, we describe two particular examples of a hysteresis operators.

Relay hysteresis. Let $a_1 < a_2$ and let $\rho_1 : [a_1, \infty) \to \mathbb{R}$, $\rho_2 : (-\infty, a_2] \to \mathbb{R}$ be continuous, globally Lipschitz and satisfy $\rho_1(a_1) = \rho_2(a_1)$ and $\rho_1(a_2) = \rho_2(a_2)$. For a given input $y \in C(\mathbb{R}_{\geq 0}, \mathbb{R})$ to the hysteresis element, the output w is such that $(y(t), w(t)) \in \mathrm{graph}(\rho_1) \cup \mathrm{graph}(\rho_2)$ for all $t \geq 0$: the value $w(t)$ of the output at $t \geq 0$ is either $\rho_1(y(t))$ or $\rho_2(y(t))$, depending on which of the threshold values a_2 or a_1 was “last" attained by the input y. When suitably initialized, such a hysteresis element has the property that, to each input $y \in C(\mathbb{R}_{\geq 0}, \mathbb{R})$, there corresponds a unique output $w = Ty \in C(\mathbb{R}_{\geq 0}, \mathbb{R})$: the operator T, so defined, is of class $\mathcal{T}_0^{1,1}$.

Backlash hysteresis with a *backlash* or *play* operator of class $\mathcal{T}_0^{1,1}$ is also feasible: see [5, Sect. 4.5.2].

3.2 Control objective: funnel control

The *class of reference signals* $\mathcal{Y}_{\mathrm{ref}}$ is all absolutely continuous functions which are bounded, see (2). Obviously, the class $\mathcal{Y}_{\mathrm{ref}}$ is considerably larger than the class of periodic functions solving a time-invariant linear differential equation as in Section 2.2.

The *control objective* is met by decentralized funnel control (see Figure 2) as follows: The N decentralized proportional output error feedback **funnel controllers** (5) applied to (1) yield, for N prespecified performance funnels $\mathcal{F}_{\varphi_i}$ determined by $\varphi_i \in \Phi$

(see (3)) and arbitrary N reference signals $y_{\mathrm{ref},i}(\cdot) \in \mathcal{Y}_{\mathrm{ref}}$ (see (2)), a closed-loop system which has only bounded trajectories and, most importantly, each error $e_i(\cdot)$ evolves within the performance funnel $\mathcal{F}_{\varphi_i}$, for $i = 1,\ldots,N$; see Figure 1.

Note that, by assumption,

$$\lambda_\varphi := \inf_{t>0} \varphi(t)^{-1} = \frac{1}{\|\varphi\|_\infty} > 0, \qquad \forall\, \varphi \in \Phi; \tag{23}$$

and λ_φ describes the minimal width of the funnel bounded away from zero. If $\varphi(0) = 0$, then the width of the funnel is infinity at $t = 0$; see Figure 1. In the following we only treat "infinite" funnels for technical reasons; if the funnel is finite, i.e. $\varphi(0) > 0$, then we certainly need to assume that the initial error is within the funnel at $t = 0$, i.e. $\varphi(0)|Cx^0 - y_{\mathrm{ref}}(0)| < 1$, and this assumption suffices.

As indicated in Figure 1, we do not assume that the funnel boundary decreases monotonically; whilst in most situation the control designer will choose a monotone funnel, there are situations where widening the funnel at some later time might be beneficial: e.g., when it is known that the reference signal changes strongly or the system is perturbed by some calibration so that a large error would enforce a large control action.

A variety of funnels are possible; we describe some of them here.

1) For $a \in (0,1)$ and $b > 0$, the function

$$t \mapsto \varphi(t) = \begin{cases} \frac{1}{1-at} & , t \in \left[0, \frac{1-b}{a}\right] \\ \frac{1}{b} & , t \geq \frac{1-b}{a} \end{cases} \tag{24}$$

determines the funnel boundary $t \mapsto \varphi(t)^{-1} := \max\{1 - at, b\}$, which is defined on the whole of $\mathbb{R}_{\geq 0}$; hence $\mathcal{F}_\varphi$ is a "finite" funnel.

2) For $a > 0$ and $b \in (0,1)$, the function $t \mapsto \varphi(t) := \min\{at,\ b^{-1}\}$ determines the "infinite" funnel $\mathcal{F}_\varphi$ and the funnel boundary $t \mapsto \varphi(t)^{-1} = \max\left\{\frac{1}{at}, b\right\}$ is defined for all $t > 0$. The funnel boundary decays strictly monotonically in the transient phase on the interval $[0,(ab)^{-1}]$ and is equal to the constant value $b^{-1} > 0$ thereafter.

3) Let $M, \mu, \lambda > 0$ with $M > \lambda$. Then the function $t \mapsto \varphi(t)^{-1} := \max\{M e^{-\mu t}, \lambda\}$ determines a "finite" funnel and ensures error evolution with prescribed exponential decay in the transient phase $[0,T]$, $T = \ln(M/\lambda)/\mu$, and tracking accuracy $\lambda > 0$ thereafter. Note that with this choice we may capture the control objective of "practical (M,μ)-stability".

4) The choice $t \mapsto \varphi(t) = \min\{t/\tau, 1\}/\lambda$ with $\tau, \lambda > 0$, ensures that the modulus of the error decays at rate $\tau\lambda/t$ in the "initial (transient) phase" $(0,\tau]$, and, is bounded by λ in the "terminal phase" $[\tau, \infty)$.

The above examples are only given to illustrate the shape of the funnel boundary in the initial phase; it need not be constant or monotone in the terminal phase.

3.3 Funnel control

We are now in a position to state the main result; see Figure 2 for illustration. Note that in comparison to Theorem 1, we address prespecified transient behaviour, the gain is no longer monotone, and the class of reference signals as well as the class of systems is much larger. However, funnel control does not guarantee that the output errors $e_i(t)$ tend to zero asymptotically as t tends to infinity; but from a practical point of view this difference is negligible since the width of the funnel (see (23)) may be chosen arbitrarily small.

Theorem 3. *Consider N interconnected systems* (1) *for $T_i \in \mathcal{T}_h^{N,1}$ and $\gamma_i > 0$ and let, for $\varphi_i \in \Phi$, associated performance funnels $\mathcal{F}_{\varphi_i}$ be given, where $i = 1\ldots,N$. Then for any reference signals and initial data*

$$y_{\mathrm{ref},i}(\cdot) \in \mathcal{Y}_{\mathrm{ref}}, \qquad y_i|_{[-h,0]} = y_i^0 \in C^\infty([-h,0],\mathbb{R}), \qquad i = 1\ldots,N,$$

the N decentralized funnel controllers (5) *applied to* (1) *yield, for $i = 1\ldots,N$, a closed-loop initial value problem which has a solution, every solution can be maximally extended, and every maximal solution $y\colon[-h,\omega) \to \mathbb{R}^N$ has the following properties:*

(i) *$\omega = \infty$, i.e. no finite escape time;*

(ii) *The gains $\frac{\varphi_i(\cdot)}{1-\varphi_i(\cdot)\,|e_i(\cdot)|}$, the outputs $y_i(\cdot)$, and the inputs $v_i(\cdot)$ are all bounded on $\mathbb{R}_{\geq 0}$ for all $i = 1,\ldots,N$;*

(iii) *every tracking error $e_i(\cdot)$ evolves within the funnel $\mathcal{F}_{\varphi_i}$ and is uniformly bounded away from the funnel boundary in the sense:*

$$\forall\, i = 1,\ldots,N\ \exists\, \varepsilon_i > 0\ \forall\, t > 0\ :\ |e_i(t)| \leq \varphi_i(t)^{-1} - \varepsilon_i .$$

Proof. Step 1: We use the notation

$$y = (y_1,\ldots,y_N)^\top, \quad y_{\mathrm{ref}} = (y_{\mathrm{ref},1},\ldots,y_{\mathrm{ref,N}})^\top, \quad Ty := \big(T_1 y,\ldots,T_N y\big)^\top .$$

In view of the potential singularity in the feedback (5), some care is required in formulation of the closed-loop initial-value problem (1), (5). We therefore define

$$\mathcal{D} := \big\{(t,\zeta) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^N \big|\ \forall\, i = 1,\ldots,N\ :\ (t,\zeta_i - y_{\mathrm{ref},i}(t)) \in \mathcal{F}_{\varphi_i}\big\}$$

and

$$F\colon \mathcal{D} \times \mathbb{R}^N \to \mathbb{R}^N, \quad ((t,\zeta),w) \mapsto F\big((t,\zeta),w\big) = \big(F_1((t,\zeta),w),\ldots,F_N((t,\zeta),w)\big)^\top$$

where

$$F_i\big((t,\zeta),w\big) := w_i - \gamma_i \frac{\varphi_i(t)\,[\zeta_i - y_{\mathrm{ref},i}(t)]}{1-\varphi_i(t)\,|\zeta_i - y_{\mathrm{ref},i}(t)|}, \quad i = 1,\ldots,N.$$

In this case, the closed-loop, initial-value problem (1), (5) is formulated as

$$\dot y(t) = F\big((t,y(t)),(\mathbf{T}y)(t)\big), \qquad y|_{[-h,0]} = y^0. \tag{25}$$

Since $T_i \in \mathcal{T}_h^{N,1}$, it follows immediately from the definition of T that $T \in \mathcal{T}_h^{N,N}$; and since the function F is a Carathéodory function[1], we may apply [3, Theorem B.1][2] to conclude that the closed-loop initial-value problem (25) has a *solution* (a function $y \in C([-h,\omega),\mathbb{R}^N)$ where $\omega \in (0,\infty]$ such that $y|_{[-h,0]} = y^0$, $y|_{[0,\omega)}$ is locally absolutely continuous, with $(t,y(t)) \in \mathcal{D}$ for all $t \in [0,\omega)$ and (25) holds for almost all $t \in [0,\omega)$) and every solution can be extended to a maximal solution (that means it has no proper right extension that is also a solution); moreover, noting that F is locally essentially bounded, if $y : [-h,\omega) \to \mathbb{R}$ is a maximal solution, then the closure of $\operatorname{graph}\big(y|_{[0,\omega)}\big)$ is not a compact subset of $\mathcal{D}$.

Step 2: In the following let $y : [-h,\omega) \to \mathbb{R}^N$ for $\omega \in (0,\infty]$ be a maximal solution of the closed-loop, initial-value problem (25).

Then $e := y - y_{\text{ref}}$ evolves on $[0,\omega)$ within the funnel and is therefore bounded. Also, by definition of $\mathcal{D}$,

$$\forall\, i = 1,\ldots,N \;\forall\, t \in [0,\omega) \;:\; \varphi_i(t)|e_i(t)| < 1\,.$$

The initial-value problem (25) is equivalent to the system of $i = 1,\ldots,N$ functional initial-value problems

$$\dot e_1(t) = T_i\big(e(\cdot) - y_{\text{ref},i}(\cdot)\big)(t) - \dot y_{\text{ref},i}(t) - \gamma_i \frac{\varphi_i(t)\; e_i(t)}{1-\varphi_i(t)\,|e_i(t)|}, \quad e|_{[-h,0]} = y^0 - y_{\text{ref}}(0)\,. \tag{26}$$

Now define, for arbitrary but fixed $\delta \in (0,\omega)$ and $i = 1,\ldots,N$,

$$\begin{aligned}
\hat f_i &:= \sup_{t\in[0,\omega)} \big|(T_i y)(t) - \dot y_{\text{ref},i}(t)\big| \\
\lambda_i &:= \inf_{t\in(0,\omega)} \varphi_i(t)^{-1} \\
L_i &> 0 \quad \text{Lipschitz bound of } \varphi_i|_{[\delta,\infty)}(\cdot)^{-1} \\
k_i(t) &:= \frac{\varphi_i(t)}{1-\varphi_i(t)\,|e_i(t)|} \qquad \forall\, t \in [0,\omega) \\
\varepsilon_i &:= \min\left\{ \frac{\lambda_i}{2}, \frac{\gamma_i\lambda_i}{2[L_i + \hat f_i]}, \min_{t\in[0,\delta]} \left\{\varphi_i(t)^{-1} - |e_i(t)|\right\} \right\}.
\end{aligned} \tag{27}$$

We show that

$$\forall\, i = 1,\ldots,N \quad \forall\, t \in (0,\omega) \quad : \quad \varphi_i(t)^{-1} - |e_i(t)| \geq \varepsilon_i\,. \tag{28}$$

[1] Let $\mathcal{D}$ be a *domain* in $\mathbb{R}_+ \times \mathbb{R}$ (that is, a non-empty, connected, relatively open subset of $\mathbb{R}_+ \times \mathbb{R}$). A function $F : \mathcal{D} \times \mathbb{R}^q \to \mathbb{R}$, is deemed to be a *Carathéodory function* if, for every “rectangle" $[a,b] \times [c,d] \subset \mathcal{D}$ and every compact set $K \subset \mathbb{R}^q$, the following hold: (i) $F(t,\cdot,\cdot) : [c,d] \times K \to \mathbb{R}$ is continuous for all $t \in [a,b]$; (ii) $F(\cdot,x,w) : [a,b] \to \mathbb{R}$ is measurable for each fixed $(x,w) \in [c,d] \times K$; (iii) there exists an integrable function $\gamma : [a,b] \to \mathbb{R}_+$ such that $|F(t,x,w)| \leq \gamma(t)$ for almost all $t \in \mathbb{R}_+$ and all $(x,w) \in [c,d] \times K$.

[2] In [3, Theorem B.1] only the class $\mathcal{T}_h^{1,q}$ is considered. However, it is only a technicality to show the same result for the class $\mathcal{T}_h^{N,q}$.

The inequalities in (28) hold on $(0,\delta]$ by definition of ε_i. Seeking a contradiction, suppose that

$$\exists\, i \in \{1,\ldots,N\}\ \exists\, t_1 \in [\delta,\omega)\ :\ \varphi_i(t_1)^{-1} - |e_i(t_1)| < \varepsilon_i\,.$$

Then there exists

$$t_0 \ :=\ \max\left\{ t \in [\delta, t_1) \mid \varphi_i(t)^{-1} - |e_i(t)| = \varepsilon_i \right\}$$

and we readily conclude that, for all $t \in [t_0,t_1]$,

$$\varphi_i(t)^{-1} - |e_i(t)| \le \varepsilon_i \qquad \text{and} \qquad |e_i(t)| \ge \varphi_i(t)^{-1} - \varepsilon_i \ge \lambda_i - \varepsilon_i \overset{(27)}{\ge} \lambda_i/2$$

and

$$k(t)|e_i(t)| = \frac{|e_i(t)|}{\varphi_i(t)^{-1} - |e_i(t)|} \ge \frac{\lambda_i}{2\varepsilon_i}$$

so that

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}\tfrac{1}{2}\, e_i(t)^2 &= e_i(t)\left[(T_i y)(t) - \dot y_{\mathrm{ref},i}(t) - \gamma_i k(t) e_i(t) \right] \\ &\le -\gamma_i k(t) e_i(t)^2 + \hat f_i |e_i(t)| \le \left[-\gamma_i \frac{\lambda_i}{2\varepsilon_i} + \hat f_i \right] |e_i(t)| \le -L_i |e_i(t)| \end{aligned} \tag{29}$$

and therefore

$$\begin{aligned} |e_i(t_1)| - |e_i(t_0)| &= \int_{t_0}^{t_1} \frac{e_i(\tau)\dot e_i(\tau)}{|e_i(\tau)|}\,\mathrm{d}\tau \\ &\le -L_i(t_1 - t_0) \le -\left| \varphi_i(t_1)^{-1} - \varphi_i(t_0)^{-1} \right| \le \varphi_i(t_1)^{-1} - \varphi_i(t_0)^{-1} \end{aligned}$$

and we arrive at the contradiction

$$\varepsilon_i = \varphi_i(t_0)^{-1} - |e_i(t_0)| \le \varphi_i(t_1)^{-1} - |e_i(t_1)| < \varepsilon_i\,.$$

This proves (28).

Step 3: (28) is equivalent to $k(\cdot) \in L^\infty([0,\omega),\mathbb{R})$. Since the errors $e_i(\cdot)$ evolve within the funnels, they are bounded on $[0,\omega)$ and also the input functions satisfy $v_i(\cdot) \in L^\infty([0,\omega),\mathbb{R})$; since the reference signals $y_{\mathrm{ref},i}(\cdot)$ are bounded, it follows that $y_i(\cdot) \in L^\infty([0,\omega),\mathbb{R})$. Finally, boundedness of all functions and maximality of $[0,\omega)$ yields that $\omega = \infty$, whence Assertion (i) and Assertion (iii); and Assertion (ii) is a consequence of (28). This completes the proof of the theorem. □

Step 2 of the proof of Theorem 3 is "compact"; a more intuitive, but slightly more technical, alternative would go as follows:
Suppose, after the definition of t_0, that $e_i(t_0) > 0$. Then $e_i(t) > 0$ for all $t \in [t_0,t_1]$ and (29) may be replaced by

$$\tfrac{\mathrm{d}}{\mathrm{d}t} e_i(t) \ \le\ -L_i \ \le\ \tfrac{\mathrm{d}}{\mathrm{d}t}\varphi_i(t)^{-1} \qquad \forall\, t \in [t_0,t_1]\,.$$

This shows that the increase of $e_i(t)$ is smaller than the increase of the funnel boundary $\varphi_i(t)^{-1}$ at each $t \in [t_0,t_1]$; hence the error evolution cannot hit the funnel boundary on $[t_0,t_1]$; this violates the definition of t_1. The case $e_i(t_0) < 0$ is then treated analogously.

4 Illustrative simulation

We consider the same set of $N = 4$ single-input, single-output minimum phase systems with high-frequency gain 1 as in [1, Sect. 4] given by transfer functions $u_i \mapsto y_i$:

$$\begin{aligned} g_1(s) &= \frac{s+1}{s^2-2s+1}, & g_2(s) &= \frac{s^3+4s^2+5s+2}{s^4-5s^3+3s^2+4s-1}, \\ g_3(s) &= \frac{1}{s-1}, & g_4(s) &= \frac{s^2+2+1}{s^3+2s^2+3s-2} \end{aligned} \tag{30}$$

and interconnection matrix

$$F = \begin{bmatrix} 0 & 2 & 1 & 1/2 \\ 1 & 0 & 1/3 & 1/4 \\ 1/2 & 1 & 0 & 1 \\ 1/4 & 3/4 & 3/2 & 0 \end{bmatrix} \tag{31}$$

for (9). In [1, Sect. 4], the reference signals $y_{\mathrm{ref},i}(t) = \sin(t + (i-1)\pi/4)$ and the internal model $\frac{P(s)}{Q(s)} = \frac{s^2+16}{(s+\pi)^2}$ is chosen according to (11) for $i = 1,2,3,4$, resp. We have confirmed, for applying the high-gain controllers (12) to (30), the same simulation results, but not depicted here. Instead, for purposes of illustration we have chosen a randomly generated matrix

$$F \approx \begin{pmatrix} 8.15 & 6.32 & 9.58 & 9.57 \\ 9.06 & 0.98 & 9.65 & 4.85 \\ 1.27 & 2.78 & 1.58 & 8 \\ 9.13 & 5.47 & 9.71 & 1.42 \end{pmatrix} \tag{32}$$

with no special structure as in (9), no internal model (11), and (chaotic) reference signals

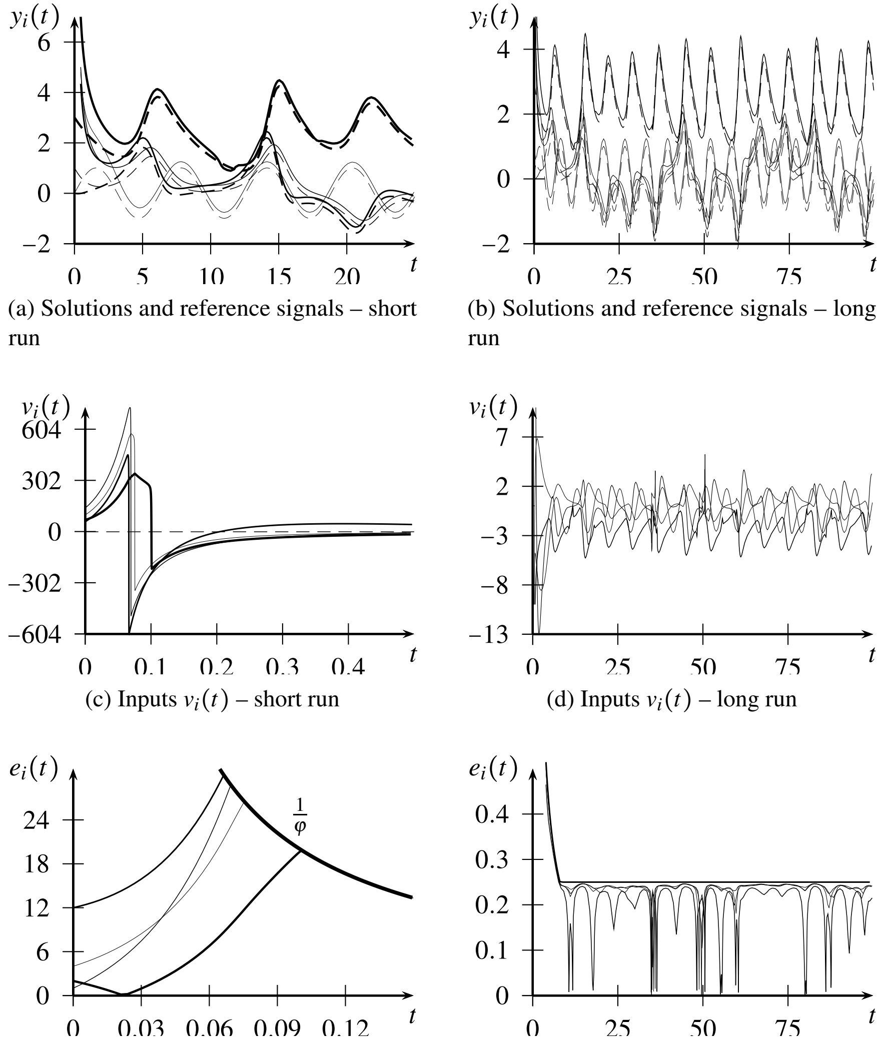
$$y_{\mathrm{ref},1} = \xi_1, \quad y_{\mathrm{ref},2} = \xi_2, \quad y_{\mathrm{ref},3} = \xi_3, \quad y_{\mathrm{ref},4}(t) = \sin(t\pi/4), \tag{33}$$

where (ξ_1, ξ_2, ξ_3) is the solution of initial-value problem for the following Lorenz system:

$$\begin{aligned} \dot{\xi}_1 &= \xi_2 - \xi_1, & \xi_1(0) &= 1 \\ \dot{\xi}_2 &= (28\xi_1/10) - (\xi_2/10) - \xi_1\xi_3, & \xi_2(0) &= 0 \\ \dot{\xi}_3 &= \xi_1\xi_2 - (8\xi_3/30), & \xi_3(0) &= 3\,. \end{aligned} \tag{34}$$

It is well known that the unique global solution of (34) is bounded with bounded derivative; see, for example, [10]. The function φ as in (24) with parameters $a = 0.5$ and $b = 0.25$ has been chosen to specify the performance $\mathcal{F}_\varphi$.

The results of Theorem 3 have been confirmed by the simulations depicted in Figure 3 on the next page. Due to the rapidly decreasing funnel in the transient phase $[0, 0.1]$, all errors tends to the funnel boundary, and hence the gain increases to preclude boundary contact; this makes the gain very large and yields $|u_i(t)| \approx 600$. After that the $u_i(t)$ take moderate values in $[-5, 5]$ and the errors stay away from the funnel boundary.



(a) Solutions and reference signals – short run

(b) Solutions and reference signals – long run

(c) Inputs $v_i(t)$ – short run

(d) Inputs $v_i(t)$ – long run

(e) Errors $|e_i(t)|$ and performance funnel $\mathcal{F}_\varphi$ – short run

(f) Errors $|e_i(t)|$ and performance funnel $\mathcal{F}_\varphi$ – long run

Figure 3: Simulation of solutions $y_i(t)$, reference signals $y_{\text{ref,i}}(t)$, and errors $e_i(t)$ (from thickest to thinnest) with respect to $i = 4, 1, 2, 3$ and performance funnel $\mathcal{F}_\varphi$

Acknowledgments

I am indebted to Stephan Trenn (TU Kaiserslautern) for helping to improve the note, and to Stefan Brechtken (TU Ilmenau) for carefully working out the simulations and the diagrams.

Bibliography

[1] U. Helmke, D. Prätzel-Wolters, and S. Schmidt. Adaptive synchronization of interconnected linear systems. Technical Report 37, University of Kaiserslautern, Department of Mathematics, 1990. Cited p. 243.

[2] U. Helmke, D. Prätzel-Wolters, and S. Schmidt. Adaptive synchronization of interconnected linear systems. *IMA Journal of Mathematical Control and Information*, 8:397–408, 1991. Cited pp. 229, 230, and 233.

[3] A. Ilchmann and E. P. Ryan. Performance funnels and tracking control. *International Journal of Control*, 82(10):1828–1840, 2009. Cited pp. 234, 236, and 241.

[4] A. Ilchmann, E. P. Ryan, and C. J. Sangwin. Systems of controlled functional differential equations and adaptive tracking. *SIAM Journal on Control and Optimization*, 40(6):1746–1764, 2002. Cited p. 238.

[5] A. Ilchmann, E. P. Ryan, and C. J. Sangwin. Tracking with prescribed transient behaviour. *ESAIM: Control, Optimisation and Calculus of Variations*, 7:471–493, 2002. Cited pp. 237 and 238.

[6] H. Logemann and A. D. Mawby. Low-gain integral control of infinite dimensional regular linear systems subject to input hysteresis. In F. Colonius, U. Helmke, D. Prätzel-Wolters, and F. Wirth, editors, *Advances in Mathematical Systems Theory*, pages 255–293. Birkhäuser, 2000. Cited p. 238.

[7] A. S. Morse. Recent problems in parameter adaptive control. In I. D. Landau, editor, *Outils et Modèles Mathématiques pour l'Automatique, l'Analyse de Systèmes et le Traitment du Signal*, pages 733–740. CNRS, 1983. Cited p. 232.

[8] E. P. Ryan and C. J. Sangwin. Controlled functional differential equations and adaptive stabilization. *International Journal of Control*, 74(1):77–90, 2001. Cited pp. 236, 237, and 238.

[9] E. D. Sontag. Smooth stabilization implies coprime factorization, ISS to iISS. *IEEE Transactions on Automatic Control*, 34(4):435–443, 1989. Cited p. 237.

[10] C. Sparrow. *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*. Springer, 1982. Cited p. 243.

[11] G. Weiss. Transfer functions of regular linear systems, Part I: Characterizations of Regularity. *Transactions of the American Mathematical Society*, 342(2):827–854, 1994. Cited p. 236.

[12] J. C. Willems and C. I. Byrnes. Global adaptive stabilization in the absence of information on the sign of the high frequency gain. In A. Bensoussan and J. L. Lions, editors, *Analysis and Optimization of Systems, Proceedings of the 6th INRIA Conference, Nice, France*, pages 49–57. Springer, 1984. Cited p. 232.

Series solutions of HJB equations

Arthur J. Krener
Department of Applied
Mathematics
Naval Postgraduate School
Monterey, CA, USA

Cesar O. Aguilar
Department of Applied
Mathematics
Naval Postgraduate School
Monterey, CA, USA

Thomas W. Hunt
Department of Applied
Mathematics
Naval Postgraduate School
Monterey, CA, USA

Abstract. We examine three methods for solving the Hamilton Jacobi Bellman PDE that arises in infinite horizon optimal control problems.

1 Introduction

The Hamilton Jacobi Bellman Partial Differential Equation (HJB PDE) characterizes the solution of an optimal control problem. Consider the problem of finding a control trajectory $u(t)$, $0 \leq t \leq \infty$ that minimizes the integral of a Lagrangian

$$\int_0^\infty l(x,u)\, dt$$

subject to the dynamic constraints

$$\dot{x} = f(x,u), \qquad x(0) = x^0$$

where $x \in \mathbb{R}^{n\times 1}$, $u \in \mathbb{R}^{m\times 1}$.

If f, l are smooth and the optimal cost is a smooth function $\pi(x^0)$ of the initial condition then the optimal control is given by an optimal feedback $u(t) = \kappa(x(t))$ and the HJB PDE is satisfied,

$$0 = \min_u \left\{ \frac{\partial \pi}{\partial x}(x) f(x,u) + l(x,u) \right\},$$

$$\kappa(x) \in \operatorname*{argmin}_u \left\{ \frac{\partial \pi}{\partial x}(x) f(x,u) + l(x,u) \right\}.$$

If we further assume that the control Hamiltonian

$$H(\lambda,x,u) = \lambda f(x,u) + l(x,u)$$

is strictly convex in u for all $\lambda \in \mathbb{R}^{1\times n}$ and $x \in \mathbb{R}^{n\times 1}$ then the HJB PDE can be rewritten as

$$0 = \frac{\partial \pi}{\partial x}(x) f(x,\kappa(x)) + l(x,\kappa(u)), \tag{1}$$

$$0 = \frac{\partial \pi}{\partial x}(x) \frac{\partial f}{\partial u}(x,\kappa(x)) + \frac{\partial l}{\partial u}(x,\kappa(u)). \tag{2}$$

The simplest example of this is the so called Linear Quadratic Regulator (LQR) where the dynamics is linear and the Lagrangian is quadratic

$$f(x,u) = Fx + Gu, \qquad l(x,u) = \frac{1}{2}\left(x^\top Qx + 2x^\top Su + u^\top Ru\right)$$

where

$$\begin{bmatrix} Q & S \\ S^\top & R \end{bmatrix}$$

is nonnegative definite and R is positive definite. If $F,\ G$ is stabilizable and $Q^{\frac{1}{2}},\ F$ is detectable then the HJB PDE has a unique solution

$$\pi(x) = \frac{1}{2}x^\top Px, \qquad \kappa(x) = Kx$$

where P is the unique nonnegative definite solution of the algebraic Riccati equation

$$F^\top P + PF + Q - (PG+S)R^{-1}(PG+S)^\top = 0 \tag{3}$$

and

$$K = -R^{-1}(PG+S)^\top. \tag{4}$$

Moreover all the eigenvalues of $F + GK$ are in the open left half plane so the closed loop dynamics

$$\dot{x} = (F + GK)x \tag{5}$$

is exponentially stable.

We return to the nonlinear problem. Perhaps the principle reason for trying to solve an optimal control problem is to find a feedback $u = \kappa(x)$ that makes the closed loop system

$$\dot{x} = f(x, \kappa(x)) \tag{6}$$

asymptotically stable. If the HJB PDE can be solved for $\pi(x),\ \kappa(x)$ then the closed loop system can be shown to be asymptotically stable in some region around the origin by a Lyapunov argument,

$$\frac{d}{dt}\pi(x(t)) = \frac{\partial \pi}{\partial x}(x(t))f(x(t),\kappa(x(t))) = -l(x(t),\kappa(u(t))) \le 0.$$

Given an approximate solution $\pi(x)$ to the HJB PDE we seek the largest punctured sublevel set of $\pi(x)$ where $\pi(x) > 0$ and $\frac{\partial \pi}{\partial x}(x)f(x,\kappa(x)) < 0$. Then we know that this punctured sublevel set is in the basin of attraction of the origin for the closed loop dynamics (6).

Suppose the Lagrangian and the dynamics have Taylor series expansions

$$f(x,u) = Fx + Gu + f^{[2]}(x,u) + f^{[3]}(x,u) + \ldots, \tag{7}$$

$$l(x,u) = \frac{1}{2}\left(x^\top Qx + u^\top Ru\right) + l^{[3]}(x,u) + l^{[4]}(x,u) + \ldots, \tag{8}$$

where ${}^{[d]}$ denotes polynomial vector fields homogeneous of degree d in $x,\ u$.

Various methods have been proposed in the literature ([2] and references) to find similar series expansions of the optimal cost and/or the stabilizing optimal feedback,

$$\pi(x) = \frac{1}{2}x^{\mathsf{T}}Px + \pi^{[3]}(x) + \pi^{[4]}(x) + \ldots, \tag{9}$$

$$\kappa(x) = Kx + \kappa^{[2]}(x) + \kappa^{[3]}(x) + \ldots. \tag{10}$$

We shall examine three of them, Al'brecht's method [1], the state dependent Riccati equation method [3, 4] and Garrard's method [5–7]. We shall describe these methods and see how well they do on a simple example.

This paper is dedicated to our esteemed colleague and good friend Uwe Helmke on the occasion of his sixtieth birthday.

2 Al'brecht's Method

Al'brecht's method has been discussed and used in [9, 11, 13] and many other papers. Al'brecht plugged the series expansions (7–10) into the HJB equations (1, 2) and collected terms degree by degree. The lowest terms of the first HJB equation (1) are of degree 2 and the lowest terms of the second HJB equation (2) are of degree 1. They reduce to the algebraic Riccati equation (3) and the formula for the linear gain (4). Therefore Al'brecht assumed that F, G, Q, R, S satisfied the assumptions of the Linear Quadratic Regulator discussed above so that these equations have a unique solution.

Having found P, K we turn to the degree 3 terms of (1) and the degree 2 terms of (2),

$$0 = \frac{\partial \pi^{[3]}}{\partial x}(x)(F+GK)x + x^{\mathsf{T}}Pf^{[2]}(x,Kx) + l^{[3]}(x,Kx), \tag{11}$$

$$0 = \frac{\partial \pi^{[3]}}{\partial x}(x)G + x^{\mathsf{T}}P\frac{\partial f^{[2]}}{\partial u}(x,Kx) + \frac{\partial l^{[3]}}{\partial u}(x,Kx) + \left(\kappa^{[2]}(x)\right)^{\mathsf{T}}R. \tag{12}$$

The unknowns in these equations are $\pi^{[3]}(x)$ and $\kappa^{[2]}(x)$ and the equations are triangular, the second unknown does not appear in the first equation. To decide the solvability of the first, we study the linear operator

$$\pi^{[3]}(x) \mapsto \frac{\partial \pi^{[3]}}{\partial x}(x)(F+GK)x$$

from cubic polynomials to cubic polynomials. Its eigenvalues are of the form $\lambda_i + \lambda_j + \lambda_k$ where λ_i, λ_j, λ_k are eigenvalues of $F+GK$. A cubic resonance occurs when such a sum equals zero. But all the eigenvalues of $F+GK$ are in the open left half plane so there are no cubic resonances.

Hence there is a unique solution to the first equation for $\pi^{[3]}(x)$ and then the second equation yields

$$\kappa^{[2]}(x) = -R^{-1}\left(\frac{\partial \pi^{[3]}}{\partial x}(x)G + x^{\mathsf{T}}P\frac{\partial f^{[2]}}{\partial u}(x,Kx) + \frac{\partial l^{[3]}}{\partial u}(x,Kx)\right)^{\mathsf{T}}$$

Then we find $\pi^{[4]}(x)$ from the degree 4 terms in (1)

$$\begin{aligned} 0 = {} & \frac{\partial \pi^{[4]}}{\partial x}(x)(F+GK)x + \frac{\partial \pi^{[3]}}{\partial x}(x)\left(f(x,Kx+\kappa^{[2]}(x))\right)^{[2]} \\ & + x^{\mathrm{T}}P\left(f(x,Kx+\kappa^{[2]}(x))\right)^{[3]} + l^{[4]}(x,Kx) + \left(l^{[3]}(x,Kx+\kappa^{[2]}(x))\right)^{[4]} \end{aligned}$$

where $(\cdot)^{[d]}$ denotes the degree d part of the expression in the parenthesis. This equation is always solvable because the map

$$\pi^{[4]}(x) \mapsto \frac{\partial \pi^{[4]}}{\partial x}(x)(F+GK)x$$

from quartic polynomials to quartic polynomials has eigenvalues of the form $\lambda_i + \lambda_j + \lambda_k + \lambda_l$ where the λ's are eigenvalues of $F+GK$. Then the degree 3 part of (2) yields

$$\begin{aligned} \kappa^{[3]}(x) = -R^{-1} \Bigg(& \frac{\partial \pi^{[4]}}{\partial x}(x)G + \frac{\partial \pi^{[3]}}{\partial x}(x)\frac{\partial f^{[2]}}{\partial u}(x,Kx) \\ & + x^{\mathrm{T}}P\left(\frac{\partial f^{[3]}}{\partial u}(x,Kx+\kappa^{[2]}(x))\right)^{[2]} + \frac{\partial l^{[3]}}{\partial u}(x,Kx)\Bigg)^{\mathrm{T}}. \end{aligned}$$

The higher degree terms are found in a similar fashion. The MATLAB based Nonlinear Systems Toolbox [8] that was written by one of authors contains a routine “hjb.m" that implements Al'brecht method. It runs very fast when n, m, d are small to medium. For example when $n=6$, $m=3$, $d=3$ the routine takes 0.076734 seconds on a MacBook Pro with an 2.66 GHz Intel Core Duo processor. When d is increased to 5 it takes 3.422941 seconds.

Al'brecht's method generates a candidate Lyapunov function $\pi(x)$ for closed loop dynamics

$$\dot{x} = f(x, \kappa(x))$$

because

$$\begin{aligned} \frac{d}{dt}\pi(x(t)) &= \frac{\partial \pi}{\partial x}(x(t))f(x(t), \kappa(x(t))) \\ &= -l(x(t), \kappa(x(t))) + O(x(t))^{d+2} \end{aligned}$$

One seeks the largest sub level set $\{x : \pi(x) \le c\}$ where for $x \neq 0$

$$\begin{aligned} \pi(x) &> 0, \\ \frac{\partial \pi}{\partial x}(x)f(x, \kappa(x)) &< 0. \end{aligned} \tag{13}$$

In the second inequality the true $f(x,u)$ should be used, not its Taylor expansion. The second inequality can be relaxed using the LaSalle invariance principle.

A modification of Al'brecht's method can also be used to generate a candidate Lyapunov function for an uncontrolled dynamics

$$\dot{x} = f(x) = Fx + f^{[2]}(x) + f^{[3]}(x) + \dots.$$

The first step is to solve for P a linear Lyapunov equation of the form

$$FP + PF + Q = 0,$$

where Q is chosen to be positive definite. The candidate Lyapunov function is

$$\pi(x) = \frac{1}{2}x^\top Px + \pi^{[3]}(x) + \pi^{[4]}(x) + \dots,$$

where π is the solution of the nonlinear Lyapunov equation

$$0 = \frac{\partial \pi}{\partial x}(x)f(x) + \frac{1}{2}x^\top Qx.$$

This equation is a degenerate HJB equation with no control and so it can also be solved term by term. The method is due to Zubov [14] and is implemented by "zbv.m" in the Nonlinear Systems Toolbox.

3 State Dependent Riccati Equation Method

The state dependent Riccati equation (SDRE) method can be used on problems of the form

$$\begin{aligned} f(x,u) &= F(x)x + G(x)u, \\ l(x,u) &= \frac{1}{2}\left(x^\top Q(x)x + 2x^\top S(x)u + u^\top R(x)u\right). \end{aligned}$$

Many nonlinear optimal control problems can be written in this form. To do so the only additional restrictions on (7, 8) are that the dynamics $f(x,u)$ be linear in u and the Lagrangian be quadratic in u. Usually there are many different ways to choose $F(x)$, $G(x)$, $Q(x)$, $R(x)$, $S(x)$ and little seems to be known about which choices are better than others.

One assumes that the optimal cost and optimal feedback have similar nonunique representations

$$\pi(x) = \frac{1}{2}x^\top P(x)x, \qquad \kappa(x) = K(x)x.$$

Then the HJB equations become

$$\begin{aligned} 0 = x^\top \big(&F^\top(x)P(x) + P(x)F(x) + Q(x) \\ &-(P(x)G(x) + S(x))R^{-1}(x)(P(x)G(x) + S(x))^\top\big)x \\ &+ \sum_{ij} \frac{\partial P_{ij}}{\partial x}(x)\left(F(x)x + G(x)K(x)x\right)x_i x_j, \\ 0 = x^\top &(P(x)G(x) + S(x)) + x^\top K^\top(x)R(x) \\ &+ \sum_{ij} \frac{\partial P_{ij}}{\partial x}(x)G(x)x_i x_j. \end{aligned}$$

In the SDRE method one ignores the last sum in each of these equations to obtain

$$\begin{aligned}0 &= x^\top \big(F^\top(x)P(x)+P(x)F(x)+Q(x)\\&\qquad -(P(x)G(x)+S(x))R^{-1}(x)(P(x)G(x)+S(x))^\top\big)x,\\0 &= x^\top(P(x)G(x)+S(x))+x^\top K^\top(x)R(x),\end{aligned}$$

which reduce to the state dependent Riccati equation and a formula for the state dependent gain

$$\begin{aligned}0 &= F^\top(x)P(x)+P(x)F(x)+Q(x)\\&\qquad -(P(x)G(x)+S(x))R^{-1}(x)(P(x)G(x)+S(x))^\top, &(14)\\K(x) &= -R^{-1}(x)(P(x)G(x)+S(x))^\top. &(15)\end{aligned}$$

To our knowledge the mathematical justification for omitting the last sums has never been clearly explained. But the result is to replace a nonlinear partial differential equation (HJB) with a nonlinear functional equation (SDRE). Whether this is a true simplification is questionable. There have been several recommendations about how to solve SDRE [3]. A symbolic software package such as Maple or Mathematica may be able to solve simple systems with special structure. Another possibility is to solve it online at a relatively high bit rate. Or perhaps it can be solved offline at a large number of states and then gain scheduling is used in between. In [12] an equation similar to the SDRE is solved by series expansion in a small parameter.

We shall show that it can also be solved by series expansion in the state vector. Assume there are the following series expansions.

$$\begin{aligned}F(x) &= F^{[0]}+F^{[1]}(x)+F^{[2]}(x)+\dots,\\G(x) &= G^{[0]}+G^{[1]}(x)+G^{[2]}(x)+\dots,\\Q(x) &= Q^{[0]}+Q^{[1]}(x)+Q^{[2]}(x)+\dots,\\R(x) &= R^{[0]}+R^{[1]}(x)+R^{[2]}(x)+\dots, \qquad (16)\\S(x) &= S^{[0]}+S^{[1]}(x)+S^{[2]}(x)+\dots,\\P(x) &= P^{[0]}+P^{[1]}(x)+P^{[2]}(x)+\dots,\\K(x) &= K^{[0]}+K^{[1]}(x)+K^{[2]}(x)+\dots,\end{aligned}$$

where the superscript $^{[d]}$ denotes a matrix valued polynomial that is homogeneous of degree d in x.

The first step is to expand $R^{-1}(x)$. It is not hard to verify that

$$R^{-1}(x) = T^{[0]}+T^{[1]}(x)+T^{[2]}(x)+\dots,$$

where

$$\begin{aligned}T^{[0]} &= (R^{[0]})^{-1},\\T^{[1]}(x) &= -(R^{[0]})^{-1}R^{[1]}(x)(R^{[0]})^{-1},\\T^{[2]}(x) &= -(R^{[0]})^{-1}R^{[2]}(x)(R^{[0]})^{-1}+(R^{[0]})^{-1}R^{[1]}(x)(R^{[0]})^{-1}R^{[1]}(x)(R^{[0]})^{-1}.\end{aligned}$$

If we plug these expansions into (14, 15) and collect the degree 0 terms we get the familiar algebraic Riccati and gain equations for the linear quadratic part of the problem

$$0 = (F^{[0]})^\top P^{[0]} + P^{[0]} F^{[0]} + Q^{[0]} - (P^{[0]} G^{[0]} + S^{[0]}) T^{[0]} (P^{[0]} G^{[0]} + S^{[0]})^\top_; \tag{17}$$

$$K^{[0]} = -T^{[0]} (P^{[0]} G^{[0]} + S^{[0]})^\top. \tag{18}$$

Having solved these equations for $P^{[0]}$, $K^{[0]}$ we collect the terms of degree 1 in (14, 15),

$$\begin{aligned} 0 = {} & (F^{[0]} + G^{[0]} K^{[0]})^\top P^{[1]}(x) + P^{[1]}(x)(F^{[0]} + G^{[0]} K^{[0]}) \\ & + (F^{[1]}(x))^\top P^{[0]} + P^{[0]} F^{[1]}(x) + Q^{[1]}(x) \\ & - (P^{[0]} G^{[1]}(x) + S^{[1]}) T^{[0]} (P^{[0]} G^{[0]}(x) + S^{[0]})^\top \cdot \\ & \qquad \cdot (P^{[0]} G^{[0]}(x) + S^{[0]}) T^{[0]} (P^{[0]} G^{[1]}(x) + S^{[1]})^\top \\ & - (P^{[0]} G^{[0]}(x) + S^{[0]}) T^{[1]}(x) (P^{[0]} G^{[0]}(x) + S^{[0]})^\top, \end{aligned} \tag{19}$$

$$K^{[1]}(x) = -T^{[0]} (P^{[1]} G^{[0]}(x) P^{[0]} G^{[1]} + S^{[10]})^\top - T^{[1]}(x) (P^{[0]} G^{[0]}(x) + S^{[0]})^\top_. \tag{20}$$

Notice that (19) is a linear Lyapunov equation in the unknown $P^{[1]}(x)$. If all the eigenvalues of $F^{[0]} + G^{[0]} K^{[0]}$ are in the open left half plane then this equation is always solvable because the eigenvalues of

$$P^{[1]}(x) \mapsto (F^{[0]} + G^{[0]} K^{[0]})^\top P^{[1]}(x) + P^{[1]}(x)(F^{[0]} + G^{[0]} K^{[0]})$$

are sums of pairs of eigenvalues of $F^{[0]} + G^{[0]} K^{[0]}$ and so none of them can be zero if the linear quadratic part of the problem satisfies the standard LQR assumptions. The higher degree terms are found in a similar fashion.

The SDRE method also yields a candidate Lyapunov function $x^\top P(x) x$ for the closed loop dynamics

$$\dot{x} = (F(x) + G(x) K(x)) x.$$

The Lyapunov derivative is

$$\begin{aligned} \frac{d}{dt} x^\top(t) P(x(t)) x(t) = {} & x^\top(t) \left(F(x(t)) + G(x(t)) K(x(t))\right)^\top P(x(t)) \\ & + P(x(t)) \left(F(x(t)) + G(x(t)) K(x(t))\right) x(t) \\ & + \sum_{ij} \frac{\partial P_{ij}}{\partial x}(x(t)) \left(F(x(t)) + G(x(t)) K(x(t))\right) x(t) x_i(t) x_j(t), \end{aligned}$$

which reduces to

$$\begin{aligned} \frac{d}{dt} x^\top(t) P(x(t)) x(t) = {} & -x^\top(t) \big(Q(x(t)) + (P(x(t)) G(x(t)) + S(x(t))) \cdot \\ & \qquad \cdot R^{-1}(x(t)) (P(x(t)) G(x(t)) + S(x(t)))^\top \big) x(t) \\ & + \sum_{ij} \frac{\partial P_{ij}}{\partial x}(x(t)) \left(F(x(t)) + G(x(t)) K(x(t))\right) x(t) x_i(t) x_j(t). \end{aligned}$$

So the quadratic part of the Lyapunov derivative is nonpositive but the higher terms may be positive.

4 Garrard's Method

Garrard's method is a simplification of Al'brecht's method that was developed when computing resources were more limited. Garrard considered a reduced set of problems where

$$l(x,u) = \frac{1}{2}\left(x^\top Qx + u^\top Ru\right), \tag{21}$$

$$f(x,u) = Fx + Gu + f^{[d]}(x), \tag{22}$$

where d is either 2 or 3. His method does not yield an approximation to the optimal cost but it does yield an approximation to the optimal feedback,

$$\kappa(x) = Kx + \kappa^{[d]}(x).$$

As with all the series methods that we consider, Garrard assumed that $F,\ G,\ Q,\ R$ satisfied the assumptions of the Linear Quadratic Regulator discussed above. The first step of the method is to find $P,\ K$ as before.

Suppose $d = 2$, the next step is to solve (11). He rewrote this equation assuming (21, 22) as

$$0 = \left(\frac{\partial \pi^{[3]}}{\partial x}(x)(F+GK) + (f^{[2]}(x))^\top P\right)x \tag{23}$$

and ignored the fact that $\frac{\partial \pi^{[3]}}{\partial x}(x)$ is the gradient of a function. He treated it as an arbitrary row vector valued polynomial homogeneous of degree 2. Then (23) has multiple solutions. Since $F + GK$ is invertible one simple solution of (23) is

$$\frac{\partial \pi^{[3]}}{\partial x}(x) = -(f^{[2]}(x))^\top P(F+GK)^{-1}, \tag{24}$$

but this is usually not the gradient of a function because its mixed partials do not commute

$$\frac{\partial^2 \pi^{[3]}}{\partial x_i \partial x_j}(x) \neq \frac{\partial^2 \pi^{[3]}}{\partial x_j \partial x_i}(x).$$

Garrard set

$$\kappa^{[2]}(x) = -R^{-1}\left(\frac{\partial \pi^{[3]}}{\partial x}(x)G\right)^\top. \tag{25}$$

When $d = 3$ then $\pi^{[3]}(x) = 0,\ \kappa^{[2]}(x) = 0$ and the relevant equation is

$$0 = \left(\frac{\partial \pi^{[4]}}{\partial x}(x)(F+GK) + (f^{[3]}(x))^\top P\right)x.$$

Again if we ignore the fact that $\frac{\partial\pi^{[4]}}{\partial x}(x)$ is a gradient then one solution of (24) and (25) is

$$\frac{\partial\pi^{[4]}}{\partial x}(x) = -(f^{[3]}(x))^{\top}P(F+GK)^{-1},$$
$$\kappa^{[3]}(x) = -R^{-1}\left(\frac{\partial\pi^{[4]}}{\partial x}(x)G\right)^{\top}.$$

As we mentioned above Garrard only used his method to solve problems with one degree of nonlinearity in the dynamics but the method can be easily generalized to problems with multiple degrees of nonlinearity provided they are in the SDRE form (16).

We can solve for $\frac{\partial\pi^{[3]}}{\partial x}(x)$ by ignoring the fact that it is a gradient, cf. (24), and then use it to define $\kappa^{[2]}(x)$, cf. (25). We put $\kappa^{[2]}(x)$ in the form

$$\kappa^{[2]}(x) = K^{[1]}(x)x,$$

where $K^{[1]}(x)$ is an $m\times n$ matrix valued polynomial homogeneous of degree 1. Again this can always be done, usually in many ways.

At the next level the relevant equation is

$$0 = \left(\frac{\partial\pi^{[4]}}{\partial x}(x)(F^{[0]}+G^{[0]}K^{[0]}) + \frac{\partial\pi^{[3]}}{\partial x}(x)\left(F^{[1]}(x)+G^{[0]}K^{[1]}(x)\right) + (F^{[2]}(x)x)^{\top}P^{[0]}\right)x.$$

If we ignore the fact that $\frac{\partial\pi^{[4]}}{\partial x}(x)$ is a gradient this has a solution

$$\frac{\partial\pi^{[4]}}{\partial x}(x) = -\left(\frac{\partial\pi^{[3]}}{\partial x}(x)\left(F^{[1]}(x)+G^{[0]}K^{[1]}(x)+G^{[1]}(x)K^{[0]}\right) + x^{\top}(F^{[2]}(x))^{\top}P^{[0]}\right)(F+GK)^{-1}.$$

This can be continued to higher degrees but there is one significant disadvantage of this method. The function $\pi(x)$ is never computed so we don't have a potential Lyapunov function to check the basin of attraction of the closed loop system. One way around this is given the closed loop dynamics, use Zubov's method to compute a candidate Lyapunov function to determine the basin of attraction. But Zubov's method is a simplification of Al'brecht's method so why not just use Al'brecht's method?

5 Example

We apply the three methods described above to a simple problem where we know the exact solution. Consider the LQR problem of minimizing

$$\frac{1}{2}\int_0^{\infty} |z|^2 + u^2\,dt$$

subject to

$$\begin{aligned}\dot z_1 &= z_2,\\ \dot z_2 &= u.\end{aligned}$$

The optimal cost and optimal feedback are

$$\begin{aligned}\pi(z) &= \frac{1}{2} z^{\mathsf{T}} \begin{bmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{bmatrix} z,\\ u &= -\begin{bmatrix} 1 & \sqrt{3} \end{bmatrix} z.\end{aligned}$$

If we make the nonlinear change of coordinates

$$\begin{aligned}z_1 &= \sin x_1,\\ z_2 &= x_2 - \frac{x_1^3}{3},\end{aligned}$$

then the problem becomes nonlinear, minimize

$$\frac{1}{2}\int_0^\infty \sin^2 x_1 + \left(x_2 - \frac{x_1^3}{3}\right)^2 + u^2 \, \mathrm{dt}$$

subject to

$$\begin{aligned}\dot x_1 &= \left(x_2 - \frac{x_1^3}{3}\right)\sec x_1,\\ \dot x_2 &= \left(x_1^2 x_2 - \frac{x_1^5}{3}\right)\sec x_1 + u.\end{aligned} \tag{26}$$

But we know the true solution,

$$\begin{aligned}\pi(x) &= \frac{\sqrt{3}}{2}\sin^2 x_1 + \left(x_2 - \frac{x_1^3}{3}\right)\sin x_1 + \frac{\sqrt{3}}{2}\left(x_2 - \frac{x_1^3}{3}\right)^2,\\ \kappa(x) &= -\sin x_1 - \sqrt{3}\left(x_2 - \frac{x_1^3}{3}\right).\end{aligned} \tag{27}$$

Notice that the change of coordinates is a nonsingular mapping from $-\frac{\pi}{2} < x_1 < \frac{\pi}{2}$, $-\infty < x_2 < \infty$ to $-1 < z_1 < 1$, $-\infty < z_2 < \infty$. The nonlinear system (26) is only defined on the strip $-\frac{\pi}{2} < x_1 < \frac{\pi}{2}$, $-\infty < x_2 < \infty$ even though (27) defines $\pi(x)$ and $\kappa(x)$ on $-\infty < x_1 < \infty$, $-\infty < x_2 < \infty$.

We applied the power series methods described above to this nonlinear problem. For Al'brecht's method and the SDRE method we computed $\pi(x)$ to degree 4 and $\kappa(x)$ to degree 3. For the SDRE method

$$\pi(x) = x^{\mathsf{T}} P(x) x.$$

For Garrard's method we first computed $\kappa(x)$ to degree 3 and then found $\pi(x)$ of degree four by Zubov's method. Here are the results.

Method	Time (sec)	Norm π Error	Norm κ Error
Al'brecht	0.0090	$1.1771e-15$	$1.3476e-15$
SDRE	0.0136	0.4707	0.8951
Garrard	0.0154	$7.4470e-16$	1.8735

The times are essentially the same for the three methods. The π errors are the l_2 norms of the differences between the vectors of coefficients of the computed π's and the Taylor polynomial of degree 4 of the true π. The κ errors are the l_2 norms of the differences between the vectors of coefficients of the computed κ's and the Taylor polynomial of degree 3 of the true κ. The Al'brecht method computes the polynomials π and κ essentially to machine precision. The SDRE method makes substantial errors in both. Garrard's method also makes a substantial error in the computation of κ but π, computed by Zubov's method, corrects this error to machine precision. It is an open question whether this always happens.

Perhaps more important are the sizes of the basin of attraction of the closed loop dynamics of the three methods. So we computed these basins as follows. We plugged each third degree polynomial $\kappa(x)$ into the nonlinear dynamics (26) and computed the largest sublevel set of the corresponding fourth degree polynomial $\pi(x)$ where $\pi(x) \geq 0$ and the Lyapunov derivative of $\pi(x)$ is nonpositive. The results are shown in the figures on the following pages. The Al'brecht and Garrard basins of attraction appear identical perhaps because the corresponding π's are nearly equal while the SDRE basin of attraction is considerably smaller. It is perhaps a surprise that all of these basins are relatively small. After all, the LQR feedback globally stabilizes the linear system. So we computed the basin of attraction for Al'brecht's method where the optimal cost is computed to degree 6 and the optimal feedback is computed to degree 5. The computation took 0.210 seconds and the basin of attraction is shown in Figure 4. Notice the different scale from the other figures.

6 Conclusion

We have discussed three power series methods for approximately solving the HJB equation that arises in the infinite horizon optimal control problem. The computational burdens are roughly equivalent but only Al'becht's method can be mathematically justified. Therefore we recommend it.

We have seen in an example even when the Taylor polynomials of the optimal cost and optimal feedback are computed to machine precision, the closed loop dynamics may fail to have a large basin of attraction. This is because of the truncation of the higher order terms. One can increase the degree of the Taylor approximations but this does not always lead to a larger basin of attraction. Therefore we are developing patchy methods to remedy this [10].

Acknowledgments

The first author's research was supported in part by AFOSR and NSF.

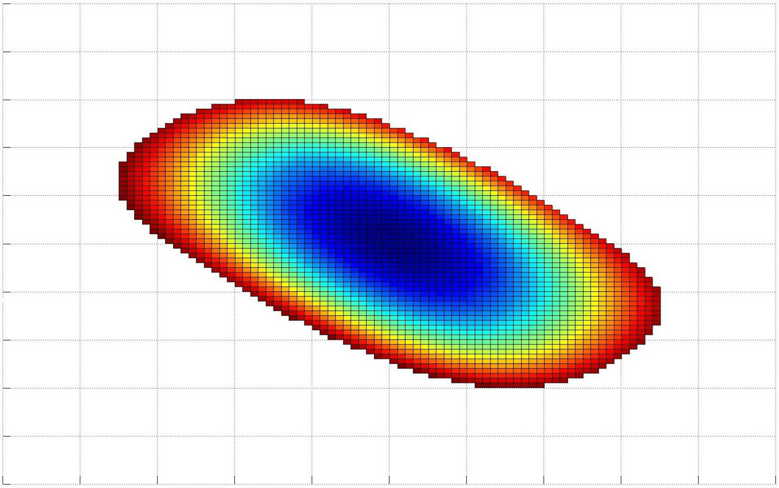


Figure 1: Al'brecht Basin of Attraction with $d = 3$, the region shown is $-1 \leq x_2 \leq 1$ on the vertical axis and $-1 \leq x_1 \leq 1$ on the horizontal axis.

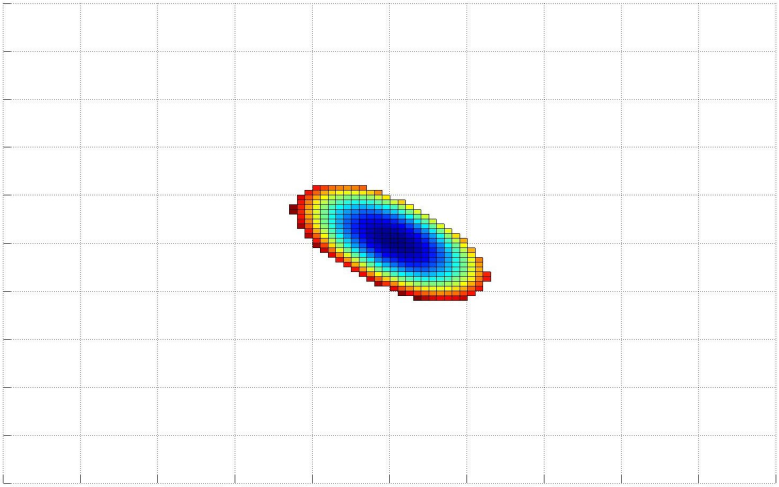


Figure 2: SDRE Basin of Attraction with $d = 3$, the region shown is $-1 \leq x_2 \leq 1$ on the vertical axis and $-1 \leq x_1 \leq 1$ on the horizontal axis.

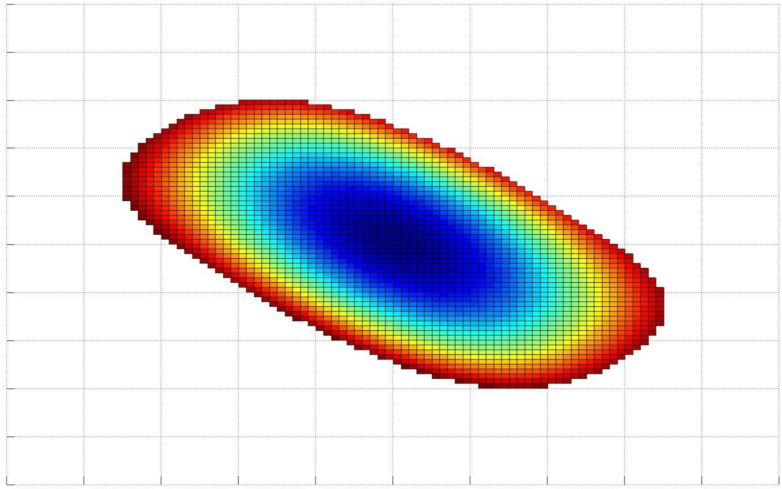


Figure 3: Garrard Basin of Attraction with $d = 3$, the region shown is $-1 \leq x_2 \leq 1$ on the vertical axis and $-1 \leq x_1 \leq 1$ on the horizontal axis.

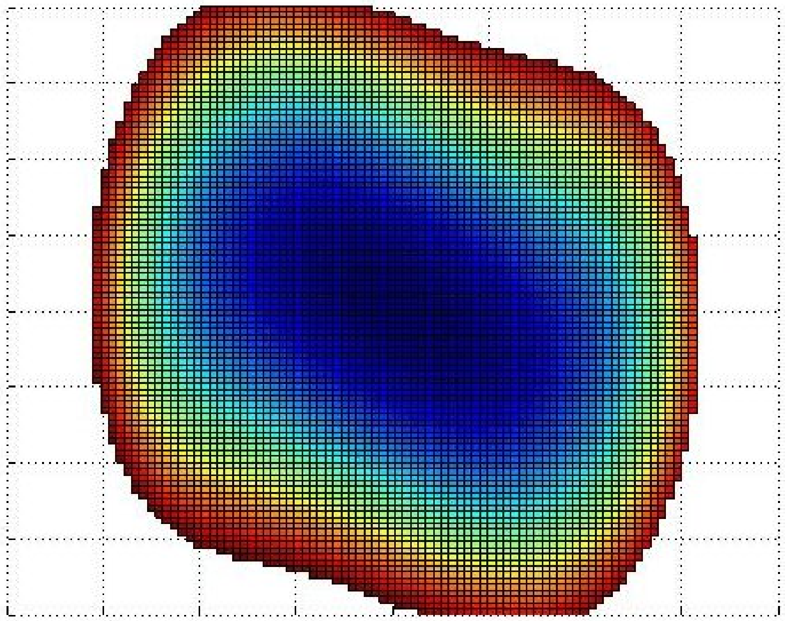


Figure 4: Al'brecht Basin of Attraction with $d = 5$, the region shown is $-2 \leq x_2 \leq 2$ on the vertical axis and $-2 \leq x_1 \leq 2$ on the horizontal axis.

Bibliography

[1] E. G. Al'brecht. On the optimal stabilization of nonlinear systems. *J. Appl. Math. Mech.*, 25:1254–1266, 1961. Cited p. 249.

[2] S. C. Beeler, H. T. Tran, and H. T. Banks. Feedback control methodologies for nonlinear systems. *Journal of Optimization Theory and Applications*, 107:1–33, 2000. Cited p. 249.

[3] J. R. Cloutier. State-dependent Riccati equation techniques: An overview. In *Proceedings of the American Control Conference*, pages 932–936, 1997. Cited pp. 249 and 252.

[4] J. R. Cloutier, C. N. D'Souza, and C. P. Mracek. Nonlinear regulation and nonlinear H-infinity control via the state-dependent Riccati equation technique; part 1, theory; part 2, examples. In *Proceedings of the International Conference on Nonlinear Problems in Aviation and Aerospace*, pages 117–141, 1996. Cited p. 249.

[5] W. L. Garrard. Suboptimal feedback control for nonlinear systems. *Automatica*, 8:219–221, 1972. Cited p. 249.

[6] W. L. Garrard, D. F. Enns, and A. Snell. Nonlinear feedback control of highly manoeuvrable aircraft. *International Journal of Control*, 56:799–812, 1992. Cited p. 249.

[7] W. L. Garrard and J. M. Jordan. Design of nonlinear automatic flight control systems. *Automatica*, 13:497–505, 1977. Cited p. 249.

[8] A. J. Krener. Nonlinear Systems Toolbox V. 1.0. MATLAB based toolbox available by request from `ajkrener@ucdavis.edu`, 1997. Cited p. 250.

[9] D. L. Lukes. Optimal regulation of nonlinear dynamical systems. *SIAM J. Contr.*, 7:75–100, 1969. Cited p. 249.

[10] C. Navasca and A. J. Krener. Patchy solutions of Hamilton Jacobi Bellman partial differential equations. In A. Chiuso, A. Ferrante, and S. Pinzoni, editors, *Modeling, Estimation and Control*, volume 364 of *Lecture Notes in Control and Information Sciences*, pages 251–270. Cited p. 257.

[11] B. F. Spencer Jr., T. L. Timlin, M. K. Sain, and S. J. Dyke. Series solution of a class of nonlinear regulators. *Journal of Optimization Theory and Applications*, 91:321–345, 1996. Cited p. 249.

[12] A. Wernli and G. Cook. Suboptimal control for the nonlinear quadratic regulator problem. *Automatica*, 11:75–84, 1975. Cited p. 252.

[13] T. Yoshida and K. A. Loparo. Quadratic regulator theory for analytic non-linear systems with additive controls. *Automatica*, 25:531–544, 1989. Cited p. 249.

[14] V. I. Zubov. *Methods of A. M. Lyapunov and their application.* P. Noordhoff, Groningen, 1964. Cited p. 251.

Optimisation geometry

Jonathan H. Manton
The University of Melbourne
Victoria, 3010, Australia
`j.manton@ieee.org`

Abstract. This article demonstrates how an understanding of the geometry of a family of cost functions can be used to develop efficient numerical algorithms for real-time optimisation. Crucially, it is not the geometry of the individual functions which is studied, but the geometry of the family as a whole. In some respects, this challenges the conventional divide between convex and non-convex optimisation problems because none of the cost functions in a family need be convex in order for efficient numerical algorithms to exist for optimising in real-time any function belonging to the family. The title "Optimisation Geometry" comes by analogy from the study of the geometry of a family of probability distributions being called information geometry.

1 Introduction and motivation

Classical optimisation theory is concerned with developing algorithms that scale well with increasing problem size and is therefore well-suited to "one-time" optimisation tasks such as encountered in the planning and design phases of an engineering endeavour. Techniques from classical optimisation theory are often applied to "real-time" optimisation tasks in signal processing applications, yet real-time optimisation problems have their own exploitable characteristics.

The often overlooked perspective this article brings to real-time optimisation problems is that the family of cost functions should be studied as a whole. This leads to a nascent theory of real-time optimisation that explores the theoretical and practical consequences of understanding the topology and geometry of how a collection of cost functions fit together.

For the purposes of this article, real-time optimisation is the challenge of developing a numerical algorithm taking a parameter value $\theta \in \Theta$ as input, and returning relatively quickly a suitable approximation to an element of

$$\left\{ x_* \in X \mid f(x_*) = \min_x f(x;\theta) \right\} \tag{1}$$

where the parametrised cost functions $f(\cdot;\theta)$ are known in advance. Since combinatorial and other non-smooth optimisation problems are less amenable to the methods introduced in this article, for the moment it may be assumed that X and Θ are differentiable manifolds and $f : X \times \Theta \to \mathbb{R}$ is a smooth function. (An important generalisation involving smooth fibre bundles will be introduced in Section 2.)

An example of real-time optimisation in signal processing is maximum-likelihood estimation, where x is the parameter to be estimated from the observation θ and $f(x;\theta)$

is the negative logarithm of the statistical likelihood function. In a communications system, if the transmitted message is x and the received packet is θ then each time a new packet is received the optimisation problem (1) must be solved to recover x from θ.

The distinguishing features setting apart real-time optimisation from classical optimisation are: the class of cost functions $f(\cdot;\theta)$ is known in advance; the class is relatively small (meaning Θ is finite-dimensional); an autonomous algorithm is required that quickly and efficiently optimises $f(\cdot;\theta)$ for (almost) any value of θ.

Real-time optimisation problems also differ from adaptive problems in that global robustness is important. Real-time algorithms must be capable of handling in turn any sequence of values for the parameter θ, whereas adaptive algorithms can assume successive values of θ will be close to each other, thereby simplifying the problem to that of tracking perturbations. Nevertheless, there are similarities because it is proposed here, in essence, to solve real-time optimisation problems by reducing them to tracking problems. Geometry facilitates this reduction.

The recent popularity of convex optimisation methods in signal processing exemplifies the earlier remark that classical optimisation theory is often applied to real-time optimisation problems. While great benefit has come from the realisation that classes of signal processing problems can be converted into convex optimisation problems such as Second-Order Cone Programming problems, this approach does not exploit the relationships between the different cost functions in the same family.

Although convexity currently determines the dichotomy of optimisation — convex problems are “easy” and non-convex problems are “hard” [12] — this is irrelevant for real-time optimisation because the complexity of real-time algorithms can be reduced by using the results of offline computations made during the design phase. An extreme example is when all the cost functions $f(\cdot;\theta)$ are just translated versions of a cost function $h(\cdot)$, such as $f(x;\theta) = h(x-\theta)$. The cost function h might be difficult to optimise, but once its minimum x_* has been found, the real-time optimisation algorithm itself is trivial: given θ, the minimum of $f(x;\theta) = h(x-\theta)$ is immediately computed to be $x_* + \theta$.

This line of reasoning extends to more general situations. For concreteness, take the parameter space Θ to be the circle S^1 (or, in fact, any compact manifold). As before, each individual cost function $f(\cdot;\theta)$ might be difficult to optimise, but provided the location of the minimum varies smoothly for almost every value of θ, the following (simplified) algorithm presents itself. Choose a finite number of parameter values $\theta_1,\cdots,\theta_n \in \Theta$. Using whatever means possible, compute beforehand the minima $x_1,\cdots,x_n \in X$ of the cost functions $f(x;\theta_i)$, that is, $f(x_i) = \min_x f(x;\theta_i)$. The minimum of $f(\cdot;\theta)$ generally can be found quickly and reliably by determining the θ_i closest to θ, and starting with the pair (x_i,θ_i), applying a homotopy method [1] to find the minimum of successive cost functions $f(\cdot;\theta_i + k\varepsilon)$ for $k = 1,\cdots,K$, where $\varepsilon = (\theta - \theta_i)/K$; see Section 5 for details. Thus, the overall complexity of real-time optimisation is determined by how the cost functions $f(\cdot;\theta)$ change as θ is varied, and not by any classical measure of the difficulty of optimising a particular cost function in the family $\{f(\cdot;\theta) \mid \theta \in \Theta\}$.

Another reason for believing in advance that the geometry of the family of cost funtions as a whole will help determine the computational complexity of real-time optimisation is that work on topological complexity and real complexity theory has already demonstrated that the geometry of a problem provides vital clues for its numerical solution [4, 13, 14]. (Another example of the efficacy of using geometry to develop numerical solutions is [3].)

Shifting from a Euclidean-based perspective of optimisation to a manifold-based perspective is expected to facilitate the development of a complexity theory for real-time optimisation. Moving to a differential geometric setting accentuates the geometric aspects while attenuating artifacts introduced by specific choices of coordinate systems used to describe an optimisation problem [7, 8, 10]. Furthermore, a wealth of problems occur naturally on manifolds [5, 8, 9], and coaxing them into a Euclidean framework is artificial and not necessarily beneficial.

The flat, unbounded geometry of Euclidean space places no topological restrictions on cost functions $f : \mathbb{R}^n \to \mathbb{R}$. Focusing on compact manifolds creates a richer structure for algorithms to exploit while maintaining practical relevance: compact Lie groups, and Grassmann and Stiefel manifolds occur in a range of signal processing applications. To the extent that no algorithm can search an unbounded region in finite time, the restriction to compact manifolds is not necessarily that restrictive. As a first step then, this article focuses on optimisation problems on compact manifolds.

One way to visualise how the cost functions in a family fit together is to imagine the mapping $\theta \mapsto f(\cdot;\theta)$ carving out a subset of the space of all (smooth) functions. This is essentially the approach taken in information geometry [2], where $f(\cdot;\theta)$ is a probability density function rather than a cost function. It seems appropriate then to use Optimisation Geometry as the title of this article.

Tangentially, it is remarked that even for one-time optimisation problems, it is not clear to the author that convexity is the fundamental divide separating easy from hard problems. Convexity might be an artifact of focusing on optimisation problems on $\mathbb{R}^n$ rather than on compact manifolds. There do not exist any nontrivial convex functions $f : M \to \mathbb{R}$ on a compact connected manifold M — if f is convex [15] then it is necessarily a constant — yet if M were a circle or a sphere, presumably there are numerous classes of cost functions that can be "easily" optimised.

Not only has Uwe brought happiness into my personal life with his genuine friendship and good humour, Uwe has played a pivotal role in my academic life. It is with all the more pleasure and sincerity then that I dedicate this article to Uwe Helmke on the occasion of his 60th birthday.

2 A fibre bundle formulation of optimisation

A real-time optimisation algorithm computes a possibly discontinuous mapping g from Θ to X. Given an input $\theta \in \Theta$, the algorithm returns $g(\theta) \in X$ where g satisfies

$$f(g(\theta);\theta) = \min_x f(x;\theta) \tag{2}$$

for all, or almost all, $\theta \in \Theta$. (Randomised algorithms are not considered here.) In a certain sense then, the additional information contained in the cost functions $f(\cdot;\theta)$

is irrelevant; if a closed-form expression for g can be determined then the original functions f can be discarded.

However, often in practice it is too hard (or not worth the effort) to find g explicitly. Optimisation algorithms therefore typically make use of the cost function, finding the minimum by moving downhill, for example. With the caveat that there is no need to remain with the original cost functions $f(\cdot;\theta)$ — they can be replaced by any other family provided there is no consequential change to the "optimising function" g — a first attempt at studying the complexity of real-time optimisation problems can be made by endeavouring to link the geometry of f with the computational complexity of evaluating the optimising function g.

Define M to be the product manifold $M = X \times \Theta$, and let $\pi : M \to \Theta$ denote the projection $(x,\theta) \mapsto \theta$. The family of cost functions $f(\cdot;\theta)$ can be thought of as a single function $f : M \to \mathbb{R}$, that is, as a scalar field on M. Provided $f : M \to \mathbb{R}$ is smooth, the manifold M relates to how the cost functions fit together.

If f is not smooth, a reparametrisation of the family of cost functions could be sought to make it smooth; in essence, a parametrisation $\theta \mapsto f(\cdot;\theta)$ is required for which smooth perturbations of θ result in smooth perturbations of the corresponding cost functions. To increase the chance of this being possible, an obvious and notationally convenient generalisation of the real-time optimisation problem is introduced.

Definition 1 (Fibre bundle optimisation problem)**.** Let M be a smooth fibre bundle over the base space Θ with typical fibre X and canonical projection $\pi : M \to \Theta$. Let $f : M \to \mathbb{R}$ be a smooth function. The fibre bundle optimisation problem is to devise an algorithm computing an *optimising function* $g : \Theta \to M$ that satisfies $(\pi \circ g)(\theta) = \theta$ and $(f \circ g)(\theta) = \min_{p \in \pi^{-1}(\theta)} f(p)$ for all $\theta \in \Theta$.

Standing Assumptions: For mathematical simplicity, it is assumed throughout that M, Θ and X in Definition 1 are compact. Smoothness means C^∞-smoothness.

If $M = X \times \Theta$ then the only difference from before is that the output of the algorithm is now a tuple $(x_*, \theta) \in M$ rather than merely $x_* \in X$. Allowing M to be a non-trivial bundle is useful in practice, as now demonstrated.

Example 2. Let M and Θ be compact connected manifolds. If $\pi : M \to \Theta$ is a submersion then it is necessarily surjective and makes M a fibre bundle. Given a smooth $f : M \to \mathbb{R}$, the fibre bundle optimisation problem is equivalent to the constrained optimisation problem of minimising $f(p)$ subject to $\pi(p) = \theta$.

Example 3. Let $\mathrm{St}(k,n) = \{X \in \mathbb{R}^{n\times k} \mid X^T X = I\}$ denote a Stiefel manifold and $\mathrm{O}(k) = \{X \in \mathbb{R}^{k\times k} \mid X^T X = I\}$ an orthogonal group. The Grassmann manifold $\mathrm{Gr}(k,n)$ is a quotient space of $\mathrm{St}(k,n)$, and in particular, $\mathrm{St}(k,n)$ decomposes as a bundle $\pi : \mathrm{St}(k,n) \to \mathrm{Gr}(k,n)$ with typical fibre $\mathrm{O}(k)$. Given a smooth function $f : \mathrm{St}(k,n) \to \mathbb{R}$, the corresponding fibre bundle optimisation problem is to minimise $f(X)$ subject to the range-space of X being fixed (that is, that $\pi(X)$ is known). A related optimisation problem (involving a constraint on the kernel rather than the range-space of X) occurs naturally in low-rank approximation problems [11].

The optimisation problem in Example 3 can be written in parametrised form by changing f to $\tilde{f} : \mathrm{Gr}(k,n) \times O(k) \to \mathbb{R}$, but if f is smooth then $\tilde{f}$ need not be continuous. Fibre bundles allow for twists in the global geometry.

Example 4. Another decomposition of $\mathrm{St}(k,n)$ is $\pi : \mathrm{St}(k,n) \to S^{n-1}$ where $\pi(X)$ is the first column of X. This corresponds to interpreting an element $X \in \mathrm{St}(k,n)$ as a point in the $(k-1)$-dimensional orthogonal frame bundle of the $(n-1)$-dimensional sphere. More generally, fibre bundle optimisation problems arise whenever a smooth function f is defined on a tangent bundle, sphere bundle, (orthogonal) frame bundle or normal bundle of a manifold M, and it is required to optimise $f(p)$ subject to p being constrained to lie above a specified point on M.

Remark 5. Fibre bundle optimisation problems (Definition 1) decompose into lower-dimensional fibre bundle optimisation problems. If $\tilde{\Theta}$ is a submanifold of Θ then the restriction of π to $\pi^{-1}(\tilde{\Theta})$ makes $M \cap \pi^{-1}(\tilde{\Theta})$ into a fibre bundle over $\tilde{\Theta}$. Conversely, a fibre bundle optimisation problem can be embedded in a higher-dimensional fibre bundle optimisation problem.

The optimising function g in Definition 1 would be a section if it were smooth, but in general g need not be everywhere continuous much less smooth. This is handled by imposing a niceness constraint on the optimisation problem.

Definition 6 (Niceness)**.** The fibre bundle optimisation problem in Definition 1 is deemed to be nice if there exist a finite number of connected open sets $\Theta_i \subset \Theta$ whose union is dense in Θ, and there exist smooth local sections $g_i : \Theta_i \to M$ such that $(f \circ g_i)(\theta) = \min_{p \in \pi^{-1}(\theta)} f(p)$ whenever $\theta \in \Theta_i$.

The requirement that the g_i are sections means $\pi(g_i(\theta)) = \theta$ for every $\theta \in \Theta_i$. The smallest number of connected open sets required in Definition 6 can be considered to be the topological complexity of the optimisation problem by analogy with the definition of topological complexity in [14]; note though that the g_i are required to be smooth in Definition 6 whereas Smale required only continuity.

Remark 7. A more practical definition of niceness might require the Θ_i in Definition 6 to be semialgebraic sets, perhaps even with a limit placed on the number of function evaluations required to test if θ is in Θ_i. This is not seen as a major issue though because it is always possible to evaluate more than one of the g_i at θ and choose the one which gives the lowest value of $f(g_i(\theta))$; the algorithm for computing g_i can return whatever it likes if $\theta \notin \Theta_i$. See also Section 5.

Whereas Section 1 only required a real-time optimisation algorithm to compute the correct answer for *almost* all values of θ, the standing assumption of compactness together with restricting attention to nice problems means the algorithm can be required to work for all θ; see Remarks 8 and 14.

Remark 8. The compactness of M means that if $\theta_n \in \Theta_i$, $\theta_n \to \theta$ then $\{g_i(\theta_n)\}_{n=1}^{\infty}$ has at least one limit point, call it q. Then $\pi(q) = \theta$ and $f(q) = \min_{p \in \pi^{-1}(\theta)} f(p)$. Therefore, if a fibre bundle optimisation problem is nice (Definition 6) then an optimising function exists on the whole of Θ (Definition 1).

Remark 9. In Definition 1, the geometry of the optimisation problem is encoded jointly by M and f. It is straightforward to reduce f to a canonical form by replacing M with the graph $\Gamma = \{(p, f(p)) \in M \times \mathbb{R} \mid p \in M\}$. Then f becomes the height function $(x,y) \mapsto y$ and the geometry of the optimisation problem is encoded in how Γ sits inside $M \times \mathbb{R}$.

As a visual aid, it can be assumed, from Remark 9 and the Whitney embedding theorem, that M is embedded in Euclidean space and the level sets $f^{-1}(c)$ are horizontal slices of M.

3 The torus

To motivate subsequent developments, this section primarily considers fibre bundle optimisation problems on the product bundle $M = S^1 \times S^1$. The function $f : S^1 \times S^1 \to \mathbb{R}$ can be thought of as defining the temperature at each point of a torus. Definitions and results will be stated in generality though, for arbitrary M.

3.1 Fibre-wise Morse functions

Minimising $f : S^1 \times S^1 \to \mathbb{R}$ restricted to a fibre is simply the problem of minimising a real-valued function on a circle. The smoothness of f and the compactness of S^1 ensure the existence of at least one global minimum per fibre.

To give more structure to the set of critical points, it is common to restrict attention either to real-analytic functions or Morse functions. Optimisation of real-analytic functions will not be considered here, but may well prove profitable for the study of gradient-like algorithms for fibre bundle optimisation problems.

If $h : S^1 \to \mathbb{R}$ is Morse, meaning all its critical points are non-degenerate, then its critical points are isolated and hence finite in number. Furthermore, the Newton method for optimisation converges locally quadratically to non-degenerate critical points. These are desirable properties that will facilitate the development of optimisation algorithms in Sections 4 and 5.

Definition 10 (Fibre-wise Morse function)**.** A *fibre-wise critical point* p of the function f in Definition 1 is a critical point of $f|_{\pi^{-1}(\pi(p))}$, the restriction of f to the fibre $\pi^{-1}(\pi(p))$ containing p. It is non-degenerate if the Hessian of $f|_{\pi^{-1}(\pi(p))}$ at p is non-singular. If all fibre-wise critical points of f are non-degenerate then f is a fibre-wise Morse function.

Remark 11. Note that f being fibre-wise Morse differs from f being Morse; a non-degenerate fibre-wise critical point need not be a critical point of f, and even if it were, it need not be non-degenerate as a critical point of f.

Lemma 12. *Let $f : M \to \mathbb{R}$ be a fibre-wise Morse function (Definition* 10*) on the bundle $\pi : M \to \Theta$ (Definition* 1*). The set N of fibre-wise critical points is a submanifold of M with the same dimension as Θ. It intersects each fibre $\pi^{-1}(\theta)$ transversally.*

Proof. It suffices to work locally; let $U \subset \Theta$ be open. Denote by VM the vertical bundle of M; it is a subbundle of the tangent bundle TM. Let $s_1, \cdots, s_k : \pi^{-1}(U) \to VM$ be a local basis, where $k = \dim M - \dim \Theta$. (The s_i are local smooth sections of VM such that $\{s_1(p), \cdots, s_k(p)\}$ is a basis for V_pM for every $p \in \pi^{-1}(U)$.) Define $e : \pi^{-1}(U) \to \mathbb{R}^k$ by $e(p) = (df(s_1(p)), \cdots, df(s_k(p)))$. Then the set of fibre-wise critical points is given locally by $N \cap \pi^{-1}(U) = e^{-1}(0)$. Fix $p \in N$. Since f is fibre-wise Morse, de_p restricted to V_pM is non-singular. Therefore de_p is surjective and $\ker de_p + V_pM = T_pM$. Thus, $e^{-1}(0)$ is an embedded submanifold of M, it has dimension $\dim M - k = \dim \Theta$, and it intersects each fibre transversally. □

The situation is especially nice on the torus: Lemma 12 implies that the set N of fibre-wise critical points of a fibre-wise Morse function is a disjoint union of a finite number of circles, with each circle winding its way around the torus the same number of times. Precisely, there is an integer b such that for any θ, each connected component of N intersects the fibre $\pi^{-1}(\theta) = S^1 \times \{\theta\}$ precisely b times.

As soon as the fibre-wise critical points of f are known at a single fibre $\pi^{-1}(\theta)$, the fibre-wise critical points of f at another fibre $\pi^{-1}(\theta')$ can be determined by tracking each of the points in $N \cap \pi^{-1}(\theta)$ as θ moves along a continuous path to θ'. This is referred to as following the circles in N from one fibre to another.

Investing more effort beforehand can obviate the need to follow more than one circle. A lookup table can record the circle in N on which the minimum lies based on which region contains θ. Proposition 13 formalises this. (In practice, there may be reasons for deciding to track more than one circle; see Remark 7.)

Proposition 13. *If f in Definition* 1 *is fibre-wise Morse (Definition* 10*) then the fibre bundle optimisation problem is nice (Definition* 6*).*

Proof. Let N be the set of fibre-wise critical points of f. For $\theta \in \Theta$, $N \cap \pi^{-1}(\theta)$ is a finite set of points because $\pi^{-1}(\theta)$ is compact and $N \pitchfork \pi^{-1}(\theta)$ with $\dim N + \dim \pi^{-1}(\theta) = \dim M$; see Lemma 12. Therefore there exist an open neighbourhood $U_\theta \subset \Theta$ of θ and local smooth sections $s_1^{(\theta)}, \cdots, s_{k_\theta}^{(\theta)} : U_\theta \to M$ such that $N \cap \pi^{-1}(U_\theta) = \cup_{i=1}^{k_\theta} s_i^{(\theta)}(U_\theta)$; pictorially, each section traces out a distinct component of $N \cap \pi^{-1}(U_\theta)$. Let $V_\theta \subset \Theta$ be an open neighbourhood of θ whose closure $\overline{V_\theta}$ is contained in U_θ. By compactness there exist a finite number of the V_θ which cover Θ; denote these sets by V_{θ_i}. Let $J_{ij} = \{\theta \in \overline{V_{\theta_i}} \mid f(s_j^{(\theta_i)}(\theta)) = h(\theta)\}$ where $h(\theta) = \min_{p \in \pi^{-1}(\theta)} f(p)$. Each J_{ij} is a closed subset of $\overline{V_{\theta_i}}$ because h is continuous. Furthermore, $\cup_j J_{ij} = \overline{V_{\theta_i}}$ and hence $\cup_{ij} J_{ij} = \Theta$. Let Θ_{ij} denote the interior of J_{ij}. Since $J_{ij} \setminus \Theta_{ij}$ is nowhere dense, $\cup_{ij} \Theta_{ij}$ is dense in Θ. The requirements of Definition 6 are met with $g_{ij}(\theta) = s_j^{(\theta_i)}(\theta)$. □

Remark 14. A stronger definition of niceness could have been adopted: each g_i in Definition 6 could have been required to be a smooth optimising function on $\overline{\Theta_i}$, the closure of Θ_i. Also, because there are only a finite number of sets involved, $\cup_i \Theta_i$ is dense in Θ if and only if $\cup_i \overline{\Theta_i} = \Theta$.

3.2 Connection with Morse theory

It is natural to ask what role Morse theory plays in real-time optimisation. After all, Morse theory contributes to one-time optimisation problems by providing information about the number, type and to some extent the location of critical points.

The short answer is the connection between Morse theory and real-time optimisation is more subtle than for one-time optimisation. The fibre bundle formulation of real-time optimisation highlights that real-time optimisation is concerned with *constrained* optimisation. It is not the level sets $\{p \in M \mid f(p) = c\}$ that are important for real-time optimisation but how they intersect the fibres $\pi^{-1}(\theta)$. From an algorithmic perspective, whereas one-time optimisation algorithms are required to find (isolated) critical points, real-time optimisation algorithms (at least from the viewpoint of this article) are required to track the critical points from fibre to fibre.

Nevertheless, for completeness, this section recalls what classical Morse theory says about the torus. Let $f : M \to \mathbb{R}$ be a smooth Morse function on $M = S^1 \times S^1$ with distinct critical points having distinct values. This is a mild assumption in practice because an arbitrarily small perturbation of f can always be found to enforce this.

Morse theory explains how the level sets $f^{-1}(c)$ fit together to form M. The fibre bundle optimisation problem is to find the smallest c for which $f^{-1}(c)$ intersects the submanifold $\pi^{-1}(\theta)$ for a given θ.

If $p \in \pi^{-1}(\theta)$ is a local minimum of f then it is also a local minimum of $f|_{\pi^{-1}(\theta)}$, and similarly for a local maximum. In both cases, p is an isolated critical point of $f|_{\pi^{-1}(\theta)}$. This need not be true though if p is a saddle point of f.

Let $p_0, \cdots, p_{n-1}$ denote the critical points of f ordered so the values $c_i = f(p_i)$ ascend. The genus of the torus dictates that the number of saddle points equals the total number of local minima and maxima, therefore $n \geq 4$.

For $c \in [c_0, c_{n-1}]$ a regular value of f, $f^{-1}(c)$ is a compact one-dimensional manifold and hence diffeomorphic to a finite number of circles. The number of circles changes by one as c passes through a critical value. In particular, $f^{-1}(c_0)$ is a single point, $f^{-1}(c)$ for $c \in (c_0, c_1)$ is diffeomorphic to S^1, and $f^{-1}(c_1)$ is either diffeomorphic to a circle plus a distinct point, or it is diffeomorphic to two circles joined at a single point. In general, $f^{-1}(c_i)$ is either diffeomorphic to zero or more copies of a circle plus a distinct point, or it is diffeomorphic to zero or more copies of a circle plus two circles joined at a single point. The former occurs when p_i is a local extremum and the latter occurs when p_i is a saddle point.

Not only is $f^{-1}(c)$ diffeomorphic to a finite number of circles for c a regular value, but $\pi^{-1}(\theta)$ is also diffeomorphic to a circle. Visually then, increasing c corresponds to sliding one or more rubber bands along the surface of the torus, and of interest is when one of these rubber bands first hits the circle $\pi^{-1}(\theta)$. The point of first contact is either a critical point of f or a non-transversal intersection point of $f^{-1}(c) \cap \pi^{-1}(\theta)$. Indeed, if $p \in f^{-1}(c) \cap \pi^{-1}(\theta)$ is not a critical point of f then p is a critical point of $f|_{\pi^{-1}(\theta)}$ if and only if p is a non-transversal intersection point of $f^{-1}(c)$ with $\pi^{-1}(\theta)$. This connects with Definition 10.

4 Newton's method and approximate critical points

The Newton method is the archtypal iterative algorithm for function minimisation. Whereas its global convergence properties are intricate — domains of attraction can be fractal — the local convergence properties of the Newton method are well understood. The advantage of real-time optimisation over one-time optimisation is it suffices to study local convergence properties of iterative algorithms because suitable initial conditions can be calculated offline.

The concept of an approximate zero was introduced in [4]. An equivalent concept will be used here, however subsequent developments differ. In [4], attention was restricted to analytic functions and global constants were sought for use in one-time algorithms (for solving polynomial equations), as opposed to the focus here on real-time optimisation algorithms.

The Newton iteration for finding a critical point of $h: \mathbb{R}^n \to \mathbb{R}$ is $x_{k+1} = x_k - [h''(x_k)]^{-1} h'(x_k)$. Its invariance to affine changes of coordinates means it suffices to assume in this section that the critical point of interest is located at the origin. The Euclidean norm and Euclidean inner product are used throughout for $\mathbb{R}^n$.

Definition 15 (Approximate critical point)**.** Let $h : \mathbb{R}^n \to \mathbb{R}$ be a smooth function with a non-degenerate critical point at the origin: $Dh(0) = 0$ and $D^2 h(0)$ is non-singular. A point x is an *approximate critical point* if, when started at $x_0 = x$, the Newton iterates x_k at least double in accuracy per iteration: $\|x_{k+1}\| \leq \frac{1}{2}\|x_k\|$.

Provided the critical point is non-degenerate, the set of approximate critical points contains a neighbourhood of the critical point. For the development of homotopy-based algorithms in Section 5, it is desirable to have techniques for finding a $\rho > 0$ such that all points within ρ of the critical point are approximate critical points. Two techniques will be explored, starting with the one-dimensional case for simplicity.

Example 16. Let $h(x) = x^2 + x^3$. Then $h'(x) = 2x + 3x^2$ and $h''(x) = 2 + 6x$. The Newton iterate is $x \mapsto x - \frac{2x+3x^2}{2+6x} = \frac{3x^2}{2+6x}$. Graphing this function shows that the largest interval $[-\rho, \rho]$ containing only approximate critical points is constrained by the equation $\frac{3x^2}{2+6x} = -\frac{x}{2}$ for $x < 0$. In particular, $\rho = \frac{1}{6} \approx 0.17$ is the best possible.

Explicit calculation as in Example 16 is generally not practical. It will be assumed that on an interval I containing the origin the first few derivatives of h are bounded. Since $h'(0) = 0$, a basic approximation for $h'(x)$ on I is $h'(x) = xh''(\bar{x})$ for some $\bar{x} \in I$. It follows that if $h''(\bar{x})/h''(x)$ is bounded between $\frac{1}{2}$ and $\frac{3}{2}$ for $x, \bar{x} \in I$ then all points in I are approximate critical points. Moreover, $h'''(x)$ can be used to bound the change in $h''(x)$. This makes plausible the following lemma.

Lemma 17. *Let $h : \mathbb{R} \to \mathbb{R}$ be a smooth function with $h'(0) = 0$ and $h''(0) \neq 0$. Let I be an interval containing the origin and $\alpha = \sup_{x \in I} |h'''(x)|$. Let $\rho = \frac{|h''(0)|}{2\alpha}$. Then every point in the interval $[-\rho, \rho] \cap I$ is an approximate critical point of h.*

Proof. Follows from Proposition 22 upon noting that $h''(x) - h''(y) = h'''(\bar{x})(x - y)$ for some $\bar{x}$ lying between x and y. □

Example 18. In Example 16, $h'''(x) = 6$. Applying Lemma 17 gives $\rho = \frac{1}{6} \approx 0.17$, coincidentally agreeing with the best possible bound.

The second technique is to look at the derivative of the Newton map $x \mapsto x - \frac{h'(x)}{h''(x)}$, which is $\frac{h'(x)h'''(x)}{[h''(x)]^2}$. Provided the magnitude of this derivative does not exceed $\frac{1}{2}$ then x is an approximate critical point.

Example 19. In Example 16, $\frac{h'(x)h'''(x)}{[h''(x)]^2} = \frac{3x(2+3x)}{(1+3x)^2}$. Its magnitude does not exceed $\frac{1}{2}$ provided $|x| \leq \frac{3-\sqrt{6}}{9} \approx 0.06$.

The need for evaluating $\frac{h'(x)h'''(x)}{[h''(x)]^2}$ can be avoided by using bounds on derivatives; an upper bound on $|h'''(x)|$ gives a lower bound, linear in x, on $h''(x)$, and an upper bound, quadratic in x, on $|h'(x)|$. Nevertheless, the first technique appears to be preferable, and will be the one considered further.

Lemma 20. *Let $h : \mathbb{R}^n \to \mathbb{R}$ have a non-degenerate critical point at the origin. Let $H_x \in \mathbb{R}^{n\times n}$, a symmetric matrix, denote its Hessian at x, that is, $D^2h(x)\cdot(\xi,\xi) = \langle H_x\xi,\xi\rangle$. Let $\bar{H}_x$ denote the averaged Hessian $\bar{H}_x = \int_0^1 H_{tx}\,dt$. Then x is an approximate critical point if $\|H_x^{-1}\bar{H}_x - I\| \le \frac{1}{2}$, where the norm is the operator norm.*

Proof. The gradient of h at x is $\int_0^1 H_{tx}x\,dt = \bar{H}_x x$. Therefore the Newton map is $x \mapsto x - H_x^{-1}\bar{H}_x x$. If $\|H_x^{-1}\bar{H}_x - I\| \le \frac{1}{2}$ then $\|x - H_x^{-1}\bar{H}_x x\| \le \frac{1}{2}\|x\|$, as claimed. □

Lemma 21. *With notation as in Lemma 20, if $\|H_x - H_0\| < \|H_0^{-1}\|^{-1}$ then*

$$\|H_x^{-1}\bar{H}_x - I\| \le \frac{\|\bar{H}_x - H_x\|}{\|H_0^{-1}\|^{-1} - \|H_x - H_0\|}. \tag{3}$$

Proof. Let $A = -(H_x - H_0)H_0^{-1}$. Then $\|A\| \le \|H_x - H_0\|\|H_0^{-1}\| < 1$. Therefore $\|(I - A)^{-1}\| = \|I + A + A^2 + \cdots\| \le 1 + \|A\| + \|A\|^2 + \cdots = (1 - \|A\|)^{-1}$. Moreover, $\|H_x^{-1}\bar{H}_x - I\| = \|H_0^{-1}(I-A)^{-1}(\bar{H}_x - H_x)\| \le \|H_0^{-1}\|(1 - \|A\|)^{-1}\|\bar{H}_x - H_x\|$. Finally, note $(1 - \|A\|)^{-1} \le (1 - \|H_x - H_0\|\|H_0^{-1}\|)^{-1}$. □

A bound on the third-order derivative yields a Lipschitz constant for the Hessian.

Proposition 22. *Define h and H_x as in Lemma 20. Let I be a star-shaped region about the origin. Let $\alpha \in \mathbb{R}$ be such that $\|H_x - H_y\| \le \alpha\|x - y\|$ for $x,y \in I$. Let $\rho = (2\alpha\|H_0^{-1}\|)^{-1}$. If $x \in I$ and $\|x\| \le \rho$ then x is an approximate critical point.*

Proof. First, $\|\bar{H}_x - H_x\| \le \int_0^1 \|H_{tx} - H_x\|\,dt \le \alpha\|x\|\int_0^1 1 - t\,dt = \frac{\alpha}{2}\|x\|$. Also, $\|H_x - H_0\| \le \alpha\|x\| \le \frac{1}{2}\|H_0^{-1}\|^{-1}$. Lemma 21 implies $\|H_x^{-1}\bar{H}_x - I\| \le \frac{(4\|H_0^{-1}\|)^{-1}}{\|H_0^{-1}\|^{-1} - (2\|H_0^{-1}\|)^{-1}}$. The result now follows from Lemma 20. □

5 A homotopy-based algorithm for optimisation

This section outlines how a homotopy-based algorithm can solve fibre bundle optimisation problems efficiently.

Homotopy-based algorithms have a long history [1]. Attention has mainly focused on one-time problems where little use can be made of results such as Proposition 22 requiring the prior calculation of various bounds on derivatives and locations of critical points. Time spent on prior calculations is better spent on solving the one-time problem directly. The reverse is true for real-time algorithms. The more calculations performed offline, the more efficient the real-time algorithm can be made, up until when onboard memory becomes a limiting factor.

Definition 6 may make it appear that nice optimisation problems are not necessarily that nice if the sets Θ_i are complicated. However, it is always straightforward to find fibre-wise critical points by path following. The worst that can happen if the Θ_i are complicated is that the algorithm may need to follow more than one path because it cannot be sure which path contains the sought after global minimum.

Proposition 23. *With notation as in Lemma* 12*, let* $\gamma\colon[0,1]\to\Theta$ *be a smooth path. Let* $p\in N\cap\pi^{-1}(\gamma(0))$*. Then* γ *lifts to a unique smooth path* $\tilde{\gamma}\colon[0,1]\to N$ *such that* $\tilde{\gamma}(0)=p$ *and* $\pi(\tilde{\gamma}(t))=\gamma(t)$ *for* $t\in[0,1]$.

Proof. Follows from Lemma 12 in a similar way Proposition 13 did. □

Corollary 24. *With notation as in Lemma* 12*, the number of points in the set* $N\cap\pi^{-1}(\theta)$ *is constant for all* $\theta\in\Theta$.

Different paths with the same end points can have different lifts. Nevertheless, as the number of fibre-wise critical points is constant per fibre, as soon as the fibre-wise critical points on one fibre are known, the fibre-wise critical points on any other fibre can be found by following any path from one fibre to another. Furthermore, only paths containing local minima need be followed to find a global minimum.

Proposition 25. *With notation as in Lemma* 12*, let* p *and* q *lie on a connected component of* N*. Then* p *is a fibre-wise local minimum if and only if* q *is a fibre-wise local minimum.*

Proof. Fibre-wise, each critical point is assumed non-degenerate. Therefore, along a continuous path, the eigenvalues of the Hessian cannot change sign and the index is preserved. □

Referring to Proposition 25, define $\tilde{N}\subset N$ to be the connected components of N corresponding to fibre-wise local minima.

An outline of a homotopy-based algorithm for fibre bundle optimisation problems can now be sketched. It will be refined presently. It relies on several lookup tables, the first of which has entries $(\theta,\pi^{-1}(\theta)\cap\tilde{N})$ for $\theta\in\{\theta_1,\cdots,\theta_n\}\subset\Theta$. That is to say, the set of all local minima of f restricted to the fibres over $\theta_1,\cdots,\theta_n$, have been determined in advance.

1. Given θ as input, determine an appropriate starting point θ_i from the finite set $\{\theta_1,\cdots,\theta_n\}$.

2. Determine an appropriate path γ from θ_i to θ.

3. Track each fibre-wise critical point $p\in\pi^{-1}(\theta_i)\cap\tilde{N}$ along the path γ (i.e., numerically compute the lift $\tilde{\gamma}$ defined in Proposition 23).

4. Evaluate the cost function f at the fibre-wise local minima on the fibre $\pi^{-1}(\theta)$ to determine which are global minima. Return one or all of the global minima.

Step 3 can be accomplished with a standard path-following scheme [1]. A refinement is to utilise Proposition 22, as now explained. Using a suitably chosen local coordinate chart, the cost function f restricted to a sufficiently small segment of the path γ can be represented locally by a function $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$. Here, h should be thought of as a parametrised cost function, with $h(\cdot;0)$ the starting function having a non-degenerate critical point at the origin, and the objective being to track that critical point all the way to the cost function $h(\cdot;1)$. An *a priori* bound on the location of the critical point of $h(\cdot;t)$ is readily available; see for example [6, Chapter 16]. Similarly, Proposition 22 gives a bound on how far away from the critical point the initial point can be whilst ensuring the Newton method converges rapidly. Therefore, these two bounds enable the determination of the largest value of $t \in [0,1]$ such that, starting at the origin, the Newton method is guaranteed to converge rapidly to the critical point of $h(\cdot;t)$. Once that critical point has been found, a new local chart can be chosen and the process repeated.

These same bounds, which are pre-computed and stored in lookup tables, permit the determination of the number of Newton steps required to get sufficiently close to the critical point. For intermediate points along the path, it is not necessary for the critical points to be found accurately. Provided the algorithm stays within the bound determined by Proposition 22, the correct path is guaranteed of being followed.

The fact that M may be a manifold presents no conceptual difficulty. As in [8], it suffices to work in local coordinates, and change charts as necessary, as already mentioned earlier.

Steps 1 and 2 of the algorithm pose three questions. How should the set $\{\theta_1, \cdots, \theta_n\}$ be chosen, how should a particular θ_i be selected based on θ, and what path should be chosen for moving from θ_i to θ? Importantly, the algorithm will work regardless of what choices are made. Nevertheless, expedient choices can significantly enhance the efficiency of the algorithm.

Another refinement is to limit in Step 3 the number of paths that are followed. Proposition 13 ensures that it is theoretically possible to determine beforehand which path the global minimum will lie on. Therefore, with the use of another lookup table, the number of paths the algorithm must track can be reduced; see Remark 7.

6 Conclusion

A nascent theory of optimisation geometry was propounded for studying real-time optimisation problems. It was demonstrated that irrespective of how difficult an individual cost function might be to optimise offline, a simple and reliable homotopy-based algorithm can be used for the real-time implementation.

Real-time optimisation problems were reformulated as fibre bundle optimisation problems (Definition 1). The geometry inherent in this fibre bundle formulation provides information about the problem's intrinsic computational complexity. An advantage of studying the geometry is it prevents any particular choice of coordinates from dominating, so there is a possibility of seeing through obfuscations caused by the chosen formulation of the problem.

That geometry helps reveal the true complexity of an optimisation problem can be demonstrated by referring back to the discussion of the fibre bundle optimisation problem on the torus in Section 3. Irrespective of how complicated the individual cost functions are (but with the proviso that they be fibre-wise Morse), the fibre-wise critical points will lie on a finite number of circles that wind around the torus, and because these circles cannot cross each other, or become tangent to a fibre, they each wind around the torus the same number of times. Therefore, in terms of where the fibre-wise critical points lie, the intrinsic complexity is encoded by just two integers: the number of circles, and the number of times each circle intersects a fibre.

Although this article lacked the opportunity to explore this aspect, a crucial observation is even though it may appear that some problems are more complicated than others because the paths of fibre-wise critical points locally "fluctuate" more, a smooth transformation can be applied to iron out these fluctuations. Smooth transformations cannot change the *intrinsic complexity* whereas they can, by definition, eliminate *extrinsic complexity*.

The second determining aspect of complexity is the number of times the fibre-wise minimum jumps from one circle to another. This is precisely what is counted by the topological complexity, mentioned just after Definition 6.

For higher dimensional problems, attention can always be restricted to compact one-dimensional submanifolds of the parameter space Θ, in which case the situation is essentially the same as for the torus; see Remark 5. The only difference is the circles may become intertwined. The theory of links and braids may play a role in further investigations, for if two circles are linked then no smooth transformation can separate them.

Another potentially interesting direction for further work is to explore the possibility of replacing a family of cost functions with an equivalent family which is computationally simpler to work with but which gives the same answer.

There are myriad other opportunities for refinements and extensions. The theory presented in this article was the first that came to mind and may well be far from optimal.

Acknowledgments

This work was supported in part by the Australian Research Council.

Bibliography

[1] E. L. Allgower and K. Georg. *Introduction to Numerical Continuation Methods*. SIAM, 2003. Cited pp. 262, 270, and 272.

[2] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS, 2000. Cited p. 263.

[3] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus: From Hodge theory to numerical stability. *Bulletin of the American Mathematical Society*, 47(2):281–354, 2010. Cited p. 263.

[4] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer, 1997. Cited pp. 263 and 268.

[5] G. S. Chirikjian and A. B. Kyatkin. *Applications of Noncommutative Harmonic Analysis: With Emphasis on Rotation and Motion Groups*. CRC Press, 2000. Cited p. 263.

[6] M. W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, 1974. Cited p. 272.

[7] H. T. Jongen, P. Jonker, and F. Twilt. *Nonlinear Optimisation in Finite Dimensions: Morse Theory, Chebyshev Approximation, Transversality, Flows, Parametric Aspects*. Kluwer, 2000. Cited p. 263.

[8] J. H. Manton. Optimisation algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, 2002. Cited pp. 263 and 272.

[9] J. H. Manton. On the role of differential geometry in signal processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 1021–1024, 2005. Cited p. 263.

[10] J. H. Manton. A centroid (Karcher mean) approach to the joint approximate diagonalisation problem: The real symmetric case. *Digital Signal Processing*, 16:468–478, 2006. Cited p. 263.

[11] J. H. Manton, R. Mahony, and Y. Hua. The geometry of weighted low rank approximations. *IEEE Transactions on Signal Processing*, 51(2):500–514, 2003. Cited p. 264.

[12] J. L. Nazareth. *Differentiable Optimization and Equation Solving: A Treatise on Algorithmic Science and the Karmarkar Revolution*. Springer, 2003. Cited p. 262.

[13] S. Smale. The fundamental theorem of algebra and complexity theory. *Bulletin of the American Mathematical Society*, 4(1):1–36, 1981. Cited p. 263.

[14] S. Smale. On the topology of algorithms, I. *Journal of Complexity*, 3:81–89, 1987. Cited pp. 263 and 265.

[15] C. Udrişte. *Convex Functions and Optimization Methods on Riemannian Manifolds*. Kluwer, 1994. Cited p. 263.

Active noise control with sampled-data filtered-x adaptive algorithm

Masaaki Nagahara
Kyoto University, Graduate School
of Informatics, Kyoto, Japan
nagahara@ieee.org

Kenichi Hamaguchi
Kyoto University, Graduate School
of Informatics, Kyoto, Japan
hamaguchi@acs.i.kyoto-u.ac.jp

Yutaka Yamamoto
Kyoto University, Graduate School
of Informatics, Kyoto, Japan
yy@i.kyoto-u.ac.jp

Abstract. Analysis and design of filtered-*x* adaptive algorithms are conventionally done by assuming that the transfer function in the secondary path is a discrete-time system. However, in real systems such as active noise control, the secondary path is a continuous-time system. Therefore, such a system should be analysed and designed as a hybrid system including discrete- and continuous- time systems and AD/DA devices. In this article, we propose a hybrid design taking account of continuous-time behaviour of the secondary path via lifting (continuous-time polyphase decomposition) technique in sampled-data control theory.

1 Introduction

Recent development of digital technology enables us to make digial signal processing (DSP) systems much more robust, flexible, and cheaper than analog systems. Owing to the recent digital technology, advanced adaptive algorithms with fast DSP devices are used in *active noise control* (ANC) systems [2, 8]; air conditioning ducts [5], noise cancelling headphones [6], and automotive applications [12], to name a few.

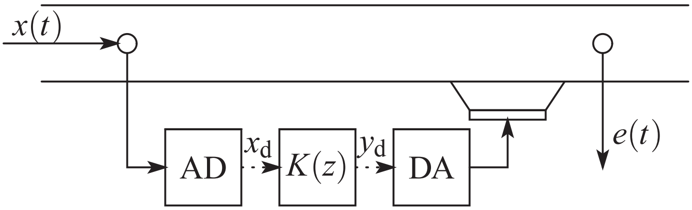


Figure 1: Active noise control system

Fig. 1 shows a standard active noise control system. In this system, $x(t)$ represents continuous-time noise which we want to eliminate during it goes through the duct. Precisely, we aim at diminishing the noise at the point C. For this purpose, we set a loudspeaker near the point C which emits antiphase sound signals to cancel the noise. Since the noise is unknown in many cases, it is almost impossible to determine antiphase signals *a priori*. Hence, we set a microphone at the point A to measure the continuous-time noise, and adopt a digital filter $K(z)$ with AD (analog-to-digital) and DA (digital-to-analog) devices. Namely, the continuous-time signal $x(t)$ is discretized to produce a discrete-time signal x_{d}, which is processed by the digital filter $K(z)$ to produce another discrete-time signal y_{d}. Then a DA converter and a loudspeaker at the point B are used to emit antiphase signals to cancel the noise in the duct.

In active noise control, it is important to compensate the distortion by the transfer characteristic of the secondary path (from B to C). To compensate this, a standard adaptive algorithm uses a filtered signal of the noise x, and is called *filtered-x algorithm* [9]. This filter is usually chosen by a discrete-time model of the secondary path [2, 9]. Consequently, the adaptive filter $K(z)$ optimizes the norm (or the variance in the stochastic setup) of the discretized signal $e(nh)$, $n = 0,1,2,\ldots$ where h is the sampling period of AD and DA device. This is proper if the secondary path is also a discrete-time system. However, in reality, the path is a *continuous-time* system, and hence the optimization should be executed taking account of the behavior of the continuous-time error signal $e(t)$. Such an optimization may seem to be difficult because the system is a *hybrid* system containing both continuous- and discrete-time signals.

Recently, several articles have been devoted to the design considering a continuous-time behavior. In [13], a hybrid controller containing an analog filter and a digital adaptive filter has been proposed. Owing to the analog filter, a robust performance is attained against the variance of the secondary path. However, an analog filter is often unwelcome because of its poor reliability or maintenance cost. Another approach has been proposed in [8]. In this paper, they assume that the noise $x(t)$ is a linear combination of a finite number of sinusoidal waves. Then the adaptive algorithm is executed in the frequency domain based on the frequency response of the continuous-time secondary path. This method is very effective if we *a priori* know the frequencies of the noise. However, unknown signal with other frequencies cannot be eliminated. If we prepare adaptive filters considering many frequencies to avoid such a situation, the complexity of the controller will be very high.

The same situation has been considered in control systems theory. The modern *sampled-data control theory* [1] has been developed in 90's [15], which gives an exact design/analysis method for hybrid systems containing continuous-time plants and discrete-time controllers. The key idea is *lifting*. Lifting is a transformation of continuous-time signals to an infinite-dimensional (i.e., function-valued) discrete-time signals. The operation can be interpreted as a *continuous-time polyphase decomposition*. In multirate signal processing, the (discrete-time) polyphase decomposition enables the designer to perform all computations at the lowest rate [14]. In the same way, by lifting, continuous-time signals or systems can be represented in the discrete-time domain with no errors.

The lifting approach is recently applied to digital signal processing [4, 10, 16, 17], and proved to provide an effective method for digital filter design. Motivated these works, this article focuses on a new scheme of filtered-x adaptive algorithm which takes account of the continuous-time behavior. More precisely, we define the problem of active noise control as design of the digital filter which minimizes a continuous-time cost function. By using the lifting technique, we derive the Wiener solution for this problem, and a steepest descent algorithm based on the Wiener solution. Then we propose an LMS (least mean square) type algorithm to obtain a causal system. The LMS algorithm involves an integral computation on a finite interval, and we adopt an approximation based on lifting representation. The approximated algorithm can be easily executed by a (linear, time-invariant, and finite dimensional) digital filter.

The paper is organized as follows: Section 2 formulates the problem of active noise control. Section 3 gives the Wiener solution, the steepest descent algorithm, and the LMS-type algorithm with convergence theorems. Section 4 proposes an approximation method for computing an integral of signals for the LMS-type algorithm. Section 5 shows simulation results to illustrate the effectiveness of the proposed method. Section 6 concludes the paper.

This paper is dedicated to Uwe Helmke on the occasion of his 60th birthday.

Notation

$\mathbb{R}$, $\mathbb{R}_+$	the sets of real numbers and non-negative real numbers, resp.
$\mathbb{Z}$, $\mathbb{Z}_+$	the sets of integers and non-negative integers, resp.
$\mathbb{R}^n$, $\mathbb{R}^{n\times m}$	the sets of n-dimensional vectors and $n \times m$ matrices over $\mathbb{R}$, resp.
L^2, $L^2[0,h)$	the sets of all square integrable functions on $\mathbb{R}_+$ and $[0,h)$, resp.
$M^\top$	transpose of a matrix M
$\overline{a}$	the complex conjugate of a complex number a
s	the symbol for Laplace transform
z	the symbol for Z transform

2 Problem formulation

In this section, we formulate the design problem of active noise control. Let us consider the block diagram shown in Fig. 2 which is a model of the active noise control system shown in Fig. 1. In this diagram, $P(s)$ is the transfer function of the primary path from A to C in Fig. 1. The transfer function of the secondary path from B to C is represented by $F(s)$. Note that $P(s)$ and $F(s)$ are continuous-time systems. We model the AD device by the ideal sampler $\mathcal{S}_h$ with a sampling period h defined by

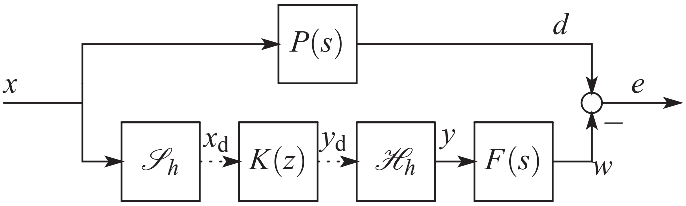


Figure 2: Block diagram of active noise control system

$$(\mathcal{S}_h x)[n] := x(nh), \quad n \in \mathbb{Z}_+.$$

That is, the ideal sampler $\mathcal{S}_h$ converts continuous-time signals to discrete-time signals. Then, the DA device is modeled by the zero-order hold $\mathcal{H}_h$ with the same period h defined by

$$(\mathcal{H}_h y)(t) := \sum_{n=0}^{\infty} \phi_0(t-nh)y[n], \quad t \in [0,\infty),$$

where $\phi_0(t)$ is the zero-order hold function or the box function defined by

$$\phi_0(t) := \begin{cases} 1, & t \in [0,h), \\ 0, & \text{otherwise.} \end{cases}$$

That is, the zero-order hold $\mathcal{H}_h$ converts discrete-time signals to continuous-time signals.

With the setup, we forumulate the design problem as follows:

Problem 1. Find the optimal FIR (finite impulse response) filter

$$K(z) = \sum_{k=0}^{N-1} \alpha_k z^{-k}$$

which minimizes the continuous-time cost function

$$J = \int_0^{\infty} e(t)^2 \mathrm{d}t. \tag{1}$$

Instead of the conventional adaptive filter design [3], this problem deals with the continuous-time behavior of the error signal $e(t)$. To solve such a hybrid problem (i.e., a problem for a mixed continuous- and discrete-time system), we introduce the lifting approach based on the sampled-data control theory [1].

In what follows, we assume the following:

Assumption 2. The following properties hold:

1. The noise x is unknown but causal, that is, $x(t) = 0$ if $t < 0$, and belongs to L^2.
2. The primary path $P(s)$ is unknown, but proper and stable.
3. The secondary path $F(s)$ is known, proper and stable.

3 Sampled-data filtered-x algorithm

In this section, we discretize the continuous-time cost function (1) without any approximation, and derive optimal filters. We also give convergence theorems for the proposed adaptive filters. The key idea to derive the results in this section is the *lifting* technique [1, 15].

3.1 Wiener solution

In this subsection, we derive the optimal filter coefficients $\alpha_0, \alpha_1, \ldots, \alpha_{N-1}$ which minimize the cost function J in (1).

First, we split the time domain $[0,\infty)$ into the union of sampling intervals $[nh,(n+1)h)$, $n \in \mathbb{Z}_+$, as

$$[0,\infty) = [0,h) \cup [h,2h) \cup [2h,3h) \cup \cdots.$$

By this, the cost function (1) is transformed into the sum of the $L^2[0,h)$-norm of $e(t)$ on the intervals:

$$J = \int_0^\infty e(t)^2 \mathrm{d}t = \sum_{n=0}^{\infty} \int_0^h e(nh+\theta)^2 \mathrm{d}\theta = \sum_{n=0}^{\infty} \int_0^h e_n(\theta)^2 \mathrm{d}\theta, \tag{2}$$

where $e_n(\theta) = e(nh+\theta)$, $\theta \in [0,h)$, $n \in \mathbb{Z}_+$. The sequence $\{e_n\}$ of functions $e_1, e_2, \ldots$ on $[0,h)$ is called the *lifted signal* [1, 15] of the continuous-time signal $e \in L^2$, and we denote the *lifting operator* by $\mathcal{L}$, that is, $\{e_n\} = \mathcal{L}e$. In what follows, we use the notion of lifting to derive the optimal coefficients.

Next, we assume that a state space realization is given for $F(s)$ as

$$F : \begin{cases} \dot{\zeta}(t) = A\zeta(t) + By(t), \\ w(t) = C\zeta(t), \quad t \in \mathbb{R}_+ \end{cases}$$

where $\zeta(0) = 0$, $A \in \mathbb{R}^{\nu\times\nu}$, $B \in \mathbb{R}^{\nu\times 1}$, and $C \in \mathbb{R}^{1\times\nu}$. By Fig. 2, the continuous-time signal w is given by

$$w = Fy = F\mathcal{H}_h y_\mathrm{d}$$

where y_d is a discrete-time signal $y_\mathrm{d} = \{y_\mathrm{d}[n]\}$ which is produced by the filter $K(z)$. Let $w_n(\theta) := w(nh+\theta)$, $\theta \in [0,h)$, $n \in \mathbb{Z}_+$ (i.e., $\{w_n\} := \mathcal{L}w$). Then, the sequence of functions $\{w_n\}$ is obtained as

$$\{w_n\} = \mathcal{L}F\mathcal{H}_h y_\mathrm{d}.$$

Let $\mathcal{F}_h := \mathcal{L}F\mathcal{H}_h$. Then the system $\mathcal{F}_h$ is a discrete-time system as shown in the following lemma [1, Sec. 10.2]:

Lemma 3. *$\mathcal{F}_h$ is a linear time-invariant discrete-time (infinite-dimensional) system with the following state-space representation:*

$$\mathcal{F}_h : \begin{cases} \xi[n+1] = A_h\xi[n] + B_h y_\mathrm{d}[n], \\ \quad w_n = \mathcal{C}_h\xi[n] + \mathcal{D}_h y_\mathrm{d}[n], \quad n \in \mathbb{Z}_+, \end{cases} \tag{3}$$

where

$$\begin{aligned} &A_h := \mathrm{e}^{Ah} \in \mathbb{R}^{\nu\times\nu}, \quad B_h := \int_0^h \mathrm{e}^{A\theta} B \mathrm{d}\theta \in \mathbb{R}^{\nu\times 1}, \\ &\mathcal{C}_h : \mathbb{R}^\nu \ni \xi \mapsto C\mathrm{e}^{A\bullet}\xi \in L^2[0,h), \quad \mathcal{D}_h : \mathbb{R} \ni y_\mathrm{d} \mapsto \int_0^\bullet C\mathrm{e}^{A\tau} B \mathrm{d}\tau \cdot y_\mathrm{d} \in L^2[0,h) \end{aligned} \tag{4}$$

The LTI property of $\mathcal{F}_h$ in Lemma 3 gives

$$\begin{aligned}\{w_n\} = \mathcal{F}_h\{y_\mathrm{d}[n]\} = \mathcal{F}_h\left(\left\{\sum_{k=0}^{N-1}\alpha_k z^{-k} x_\mathrm{d}[n]\right\}\right) = \sum_{k=0}^{N-1}\alpha_k\mathcal{F}_h\left(\{z^{-k}x_\mathrm{d}[n]\}\right)\\ = \left\{\sum_{k=0}^{N-1}\alpha_k u_{n-k}\right\},\end{aligned} \tag{5}$$

where $\{u_n\} := \mathcal{F}_h\{x_\mathrm{d}[n]\}$. Note that $\{u_n\}$ is the lifted signal of the continuous-time signal $u = F\mathcal{H}_h x_\mathrm{d}$, that is,

$$\{u_n\} = \mathcal{L}(F\mathcal{H}_h x_\mathrm{d}) = \mathcal{L}u.$$

The relation (5) gives the continuous-time relation as

$$w(t) = \sum_{k=0}^{N-1}\alpha_k u(t-kh), \quad t \in \mathbb{R}_+.$$

By using this relation, we obtain the following theorem for the optimal filter.

Theorem 4 (Wiener solution)**.** *Let $u := (F\mathcal{H}_h)x_\mathrm{d}$. Define a matrix Φ and a vector β as*

$$\Phi := [\Phi_{kl}]_{k,l=0,1,\dots,N-1} \in \mathbb{R}^{N\times N}, \quad \beta := [\beta_k]_{k=0,1,\dots,N-1} \in \mathbb{R}^N,$$

where for $k,l = 0,1,\dots,N-1$,

$$\Phi_{kl} := \int_0^\infty u(t-kh)u(t-lh)\mathrm{d}t, \quad \beta_k := \int_0^\infty d(t)u(t-kh)\mathrm{d}t.$$

Assume the matrix Φ is nonsingular. Then the gradient of J defined in (1) *is given by*

$$\nabla_\alpha J = 2\left(\Phi\alpha - \beta\right), \quad \alpha := [\alpha_0, \alpha_1, \dots, \alpha_{N-1}]^\top, \tag{6}$$

and the optimal FIR parameter $\alpha^\mathrm{opt} = [\alpha_0^\mathrm{opt}, \alpha_1^\mathrm{opt}, \dots, \alpha_{N-1}^\mathrm{opt}]^\top$ which minimizes J is given by

$$\alpha^\mathrm{opt} = \Phi^{-1}\beta. \tag{7}$$

Proof. Let $\{d_n\} := \mathcal{L}d$. By the equations (2), (5), and $e_n = d_n - w_n$, we have

$$\begin{aligned}J = \sum_{n=0}^\infty\int_0^h d_n(\theta)^2\mathrm{d}\theta - 2\sum_{k=0}^{N-1}\alpha_k\sum_{n=0}^\infty\int_0^h d_n(\theta)u_{n-k}(\theta)\mathrm{d}\theta\\ + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\alpha_k\alpha_l\sum_{n=0}^\infty\int_0^h u_{n-k}(\theta)u_{n-l}(\theta)\mathrm{d}\theta.\end{aligned} \tag{8}$$

Computing the gradient $\nabla_\alpha J$ and applying the inverse lifting, we obtain (6). Then, if the matrix Φ is nonsingular, the optimal parameter (7) is given by solving the Wiener-Hopf equation $\Phi\alpha - \beta = 0$. □

We call the optimal parameter α^opt the *Wiener solution*.

3.2 Steepest descent algorithm

In this subsection, we derive the *steepest descent algorithm* (SD algorithm) [3] for the Wiener solution obtained in Theorem 4. This algorithm is a base for adaptation of the ANC system discussed in the next subsection.

According to the identity (6) in Theorem 4 for the gradient of J, the steepest descent algorithm is described by

$$\begin{aligned} \alpha[n+1] &= \alpha[n] - \frac{\mu}{2}\nabla_{\alpha[n]}J \\ &= \alpha[n] + \mu\left(\beta - \Phi\alpha[n]\right), \quad n \in \mathbb{Z}_+, \end{aligned} \tag{9}$$

where $\mu > 0$ is the *step-size parameter*.

We then analyse the stability of the above recursive algorithm. Before deriving the stability condition, we give an upper bound of the eigenvalues of the matrix Φ.

Lemma 5. *Let $\lambda_1,\dots,\lambda_N$ be the eigenvalues of the matrix Φ. Let $\hat{u}$ denote the Fourier transform of $u = F\mathcal{H}_h x_{\rm d}$, and define*

$$S(\mathrm{j}\omega) := \frac{1}{h}\sum_{n=-\infty}^{\infty}\left|\hat{u}\left(\mathrm{j}\omega + \frac{2n\pi\mathrm{j}}{h}\right)\right|^2.$$

Then we have

$$0 \le \lambda_i \le \|S\|_\infty = \sup\left\{S(\mathrm{j}\omega) \mid \omega \in \left(-\tfrac{\pi}{h}, \tfrac{\pi}{h}\right)\right\}, \tag{10}$$

for $i = 1,2,\dots,N$.

Proof. First, we prove $\lambda_i \ge 0$ for $i = 1,2,\dots,N$. Let

$$U(t) = \left[u(t), u(t-h), \dots, u(t-Nh+h)\right]^\top.$$

Then, for non-zero vector $v \in \mathbb{R}^N$, we have

$$v^\top\Phi v = v^\top\left(\int_0^\infty U(t)U(t)^\top \mathrm{d}t\right)v = \int_0^\infty \left|v^\top U(t)\right|^2 \mathrm{d}t \ge 0.$$

Thus $\Phi \ge 0$ and hence $\lambda_i \ge 0$ for $i = 1,2,\dots,N$. Next, since $u(t) = 0$ for $t < 0$, we have

$$\Phi_{kl} = \int_0^\infty u(t-kh)u(t-lh)\mathrm{d}t = \int_0^\infty u\big(t-(k-l)h\big)u(t)\mathrm{d}t.$$

By Parseval's identity,

$$\begin{aligned} \Phi_{kl} &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{\hat{u}(\mathrm{j}\omega)}\hat{u}(\mathrm{j}\omega)\mathrm{e}^{\mathrm{j}\omega(k-l)h}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\sum_{n=-\infty}^{\infty}\int_{-h/\pi}^{h/\pi}\left|\hat{u}\left(\mathrm{j}\omega + \frac{2n\pi\mathrm{j}}{h}\right)\right|^2 \mathrm{e}^{\mathrm{j}\omega(k-l)h}\mathrm{d}\omega \\ &= \frac{h}{2\pi}\int_{-h/\pi}^{h/\pi} S(\mathrm{j}\omega)\mathrm{e}^{\mathrm{j}\omega(k-l)h}\mathrm{d}\omega. \end{aligned}$$

Then, let $v = [v_0, v_1, \ldots, v_{N-1}]^\top$ be a nonzero vector in $\mathbb{R}^N$. Let $\hat{v}$ denote the discrete Fourier transform of v, that is,

$$\hat{v}(\mathrm{j}\omega) := \sum_{k=0}^{N-1} v_k \mathrm{e}^{-\mathrm{j}\omega kh}, \quad \omega \in (-\pi/h, \pi/h).$$

Perseval's identity again gives

$$v^\top v = \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} \overline{\hat{v}(\mathrm{j}\omega)} \hat{v}(\mathrm{j}\omega) \mathrm{d}\omega.$$

Then we have

$$\begin{aligned} v^\top \Phi v &= \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} v_k v_l \Phi_{kl} = \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} v_k v_l \cdot \frac{h}{2\pi} \int_{-h/\pi}^{h/\pi} S(\mathrm{j}\omega) \mathrm{e}^{\mathrm{j}\omega(k-l)h} \mathrm{d}\omega \\ &= \frac{h}{2\pi} \int_{-\pi/h}^{\pi/h} S(\mathrm{j}\omega) \overline{\hat{v}(\mathrm{j}\omega)} \hat{v}(\mathrm{j}\omega) \mathrm{d}\omega \leq \|S\|_\infty \cdot v^\top v. \end{aligned}$$

It follows that

$$\max_{1 \leq i \leq N} \lambda_i = \max\{v^\top \Phi v \mid v \in \mathbb{R}^N, \quad v^\top v = 1\} \leq \|S\|_\infty.$$

□

By this lemma, we derive a sufficient condition on the step size μ for convergence.

Theorem 6 (Stability of SD algorithm). *Suppose that* $\Phi > 0$ *and the step size* μ *satisfies*

$$0 < \mu < 2\|S\|_\infty^{-1}. \tag{11}$$

Then the sequence $\{\alpha[n]\}$ *produced by the iteration* (9) *converges to the Wiener solution* α^{opt} *for any initial vector* $\alpha[0] \in \mathbb{R}^N$.

Proof. The iteration (9) is rewritten as

$$\alpha[n+1] = (I - \mu\Phi)\alpha[n] + \mu\beta.$$

Suppose $\Phi > 0$. Let $\lambda_{\max}$ denote the maximum eigenvalue of Φ. Then $\lambda_{\max} > 0$ since $\Phi > 0$. The condition (11) and the inequality (10) in Lemma 5 give $0 < \mu < 2\lambda_{\max}^{-1}$, which is equivalent to $|1 - \mu\lambda_i| < 1$, $i = 1, 2, \ldots, N$. It follows that the eigenvalues of the matrix $I - \mu\Phi$ lie in the open unit disk in the complex plane, and hence the iteration (9) is asymptotically stable. The final value

$$\alpha_\infty := \lim_{n \to \infty} \alpha[n]$$

of the iteration is clearly given by the solution of the equation $\Phi\alpha_\infty = \beta$. Thus, since $\Phi > 0$, we have $\alpha_\infty = \Phi^{-1}\beta = \alpha^{\mathrm{opt}}$. □

3.3 LMS-type algorithm

The steepest decent algorithm assumes that the matrix Φ and the vector β are known *a priori*. That is, the noise $\{x(t)\}_{t \in \mathbb{R}_+}$ and the primary path $P(s)$ are assumed to be known. However, in practice, the noise $\{x(t)\}_{t \in \mathbb{R}_+}$ cannot be fixed before we run the

ANC system. In other words, the ANC system should be *noncausal* for running the steepest descent algorithm. Moreover, we cannot produce arbitrarily noise $\{x(t)\}_{t\in\mathbb{R}_+}$ (this is why x is *noise*), we cannot identify the primary path $P(s)$. Hence, the assumption is difficult to be satisfied.

In the sequel, we can only use data up to the present time for causality and we cannot use the model of $P(s)$. Under this limitation, we propose to use an LMS-type adaptive algorithm using the filtered noise $u = F\mathcal{H}_h x_{\mathrm{d}}$ and the error e up to the present time. First, by the equation (5) and the relation $e = d - w$, we have

$$\frac{\partial J}{\partial \alpha_k} = -2\left(\beta_k - \sum_{l=0}^{N-1} \Phi_{kl}\alpha_l\right) = -2\int_0^\infty e(t)u(t-kh)\mathrm{d}t, \quad k = 0,1,\ldots,N-1.$$

Based on this, we propose the following adaptive algorithm:

$$\alpha[n+1] = \alpha[n] + \mu\delta[n], \quad n \in \mathbb{Z}_+, \tag{12}$$

where $\delta[n] = \left[\delta_0[n], \delta_1[n], \ldots, \delta_{N-1}[n]\right]^\top$ with

$$\delta_k[n] := \int_0^{nh} e(t)u(t-kh)\mathrm{d}t, \quad k = 0,1,\ldots,N-1.$$

The update direction vector $\delta[n]$ can be recursively computed by

$$\delta[n+1] = \delta[n] + \int_{nh}^{(n+1)h} e(t)u(t)\mathrm{d}t, \quad n \in \mathbb{Z}_+, \tag{13}$$

where

$$u(t) := \left[u(t), u(t-h), \ldots, u\big(t-(N-1)h\big)\right]^\top.$$

This means that to obtain the vector $\delta[n]$ one needs to measure the error e and the signal $u = F\mathcal{H}_h x_{\mathrm{d}}$ on the interval $[(n-1)h, nh)$ and compute the integral in (13). We call this scheme the *sampled-data filtered-x adaptive algorithm*. The term "sampled-data" comes from the use of sampled-data x_{d} of the continuous-time signal x. The sampled-data filtered-x adaptive algorithm is illustrated in Fig. 3. As shown in this figure, in order to run the adaptive algorithm, we should use the signal u which is "filtered" x_{d} by $F\mathcal{H}_h$, and also use the error signal e.

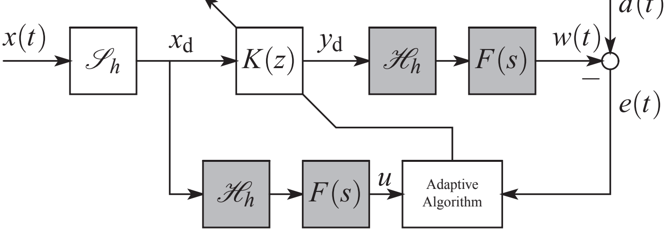


Figure 3: Sampled-data filtered-x adaptive algorithm

To analyse the convergence of the iteration, we consider the following autonomous system:

$$\alpha[n+1] = \big(I - \mu\Phi[n]\big)\alpha[n], \quad n \in \mathbb{Z}_+, \tag{14}$$

where $\Phi[n] = \big[\Phi_{kl}[n]\big]_{k,l=0,1,\dots,N-1}$ with

$$\Phi_{kl}[n] := \int_0^{nh} u(t-kh)u(t-lh)\mathrm{d}t.$$

Then we have the following lemma:

Lemma 7. *Suppose the following conditions:*

1. *The sequence $\{\Phi[n]\}$ is uniformly bounded, that is, there exists $\gamma > 0$ such that*
$$\|\Phi[n]\| \le \gamma, \quad \forall n \in \mathbb{Z}_+.$$
2. *The step-size parameter μ satisfies*
$$0 < \mu < 2\left(\max_{n\in\mathbb{Z}_+} \lambda_{\max}\big(\Phi[n]\big)\right)^{-1},$$
where $\lambda_{\max}\big(\Phi[n]\big)$ is the maximum eigenvalue of $\Phi[n]$.
3. *The sequence $\{\mu\Phi[n]\}$ is slowly-varying, that is, there exists a sufficiently small $\varepsilon > 0$ such that*
$$\big\|\mu\big(\Phi[n] - \Phi[n-1]\big)\big\| \le \varepsilon, \quad \forall n \in \mathbb{Z}_+.$$

Then the autonomous system (14) *is uniformly exponentially stable*[1].

Proof. Let $\Psi[n] := I - \mu\Phi[n]$, $n \in \mathbb{Z}_+$. By the assumption 1, we have

$$\big\|\Psi[n]\big\| = \big\|I - \mu\Phi[n]\big\| \le N + \mu\big\|\Phi[n]\big\| \le N + \mu\gamma.$$

Thus, the sequence $\{\Psi[n]\}$ is uniformly bounded. By the assumption 2, we have

$$\big|\lambda_{\max}\big(\Psi[n]\big)\big| < 1, \quad \forall n \in \mathbb{Z}_+.$$

Also, by the assumption 3, we have

$$\big\|\Psi[n] - \Psi[n-1]\big\| \le \varepsilon,$$

that is, the sequence $\{\Psi[n]\}$ is slowly varying. With these inequalities, the uniform exponential stability of the system (14) follows from Theorem 24.8 in [11]. □

[1] The system (14) is said to be *uniformly exponentially stable* [11] if there exist a finite positive constant c and a constant $0 \le r < 1$ such that for any n_0 and $\alpha_0 = \alpha[0] \in \mathbb{R}^N$, the corresponding solution satisfies $\|\alpha[n]\| \le cr^{n-n_0}\|\alpha_0\|$ for all $n \ge n_0$.

By Lemma 7, we have the following theorem:

Theorem 8 (Stability of LMS algorithm)**.** *Suppose the conditions 1–3 in Lemma* 7*. Then the sequence* $\{\alpha[n]\}$ *converges to the Wiener solution* α^{opt}.

Proof. Let $\beta[n] := \big[\beta_k[n]\big]_{k=0,1,\dots,N-1} \in \mathbb{R}^N$ with

$$\beta_k[n] := \int_0^{nh} d(t)u(t-kh)\mathrm{d}t.$$

Put $c[n] := \alpha[n] - \alpha^{\mathrm{opt}}$ and $q[n] := \beta[n] - \Phi[n]\alpha^{\mathrm{opt}}$. Then, $\Phi[n] \to \Phi$ and $\beta[n] \to \beta$ as $n \to \infty$, and hence

$$q[n] \to \infty \ \text{ as } \ n \to \infty. \tag{15}$$

By Lemma 7, the autonomous system (14) is uniformly exponentially stable and from (15) it follows that $c[n] \to 0$ as $n \to \infty$. Thus, we have $\alpha[n] \to \alpha^{\mathrm{opt}}$ as $n \to \infty$. □

4 Approximation method

To run the algorithm (12) with (13), we have to calculate the integral in (13). It is usual that the error signal *e* is given as sampled data, and hence the exact value of this integral is difficult to obtain in practice. Therefore, we introduce an approximation method for this computation.

First, we split the interval $[0,h)$ into L short intervals as

$$[0,h) = [0,h/L) \cup [h/L, 2h/L) \cup \cdots \cup [h-h/L, h).$$

Assume that the error *e* is constant on each short interval. Then we have,

$$\int_{nh}^{(n+1)h} e(t)u(t-kh)\mathrm{d}t = \sum_{l=0}^{L-1} \int_{lh/L+nh}^{(l+1)h/L+nh} e(t)u(t-kh)\mathrm{d}t = e[n]^{\mathrm{T}} U[n-k],$$

where

$$e[n] := \begin{bmatrix} e(nh) \\ e(h/L+nh) \\ \vdots \\ e(h-h/L+nh) \end{bmatrix}, \quad U[n] := \begin{bmatrix} \int_0^{h/L} u(\theta+nh)\mathrm{d}\theta \\ \int_{h/L}^{2h/L} u(\theta+nh)\mathrm{d}\theta \\ \vdots \\ \int_{(L-1)h/L}^{h} u(\theta+nh)\mathrm{d}\theta \end{bmatrix}.$$

Then the integral in $U[n]$ can be computed via the state-space representation of $\mathcal{F}_h$ given in (3). In fact, $U[n]$ can be computed by the following digital filter:

$$F_h \begin{cases} \eta[n+1] = A_h\eta[n] + B_h x_{\mathrm{d}}[n], \\ \quad U[n] = C_h\eta[n] + D_h x_{\mathrm{d}}[n], \quad n \in \mathbb{Z}_+ \end{cases}$$

where A_h and B_h are given in (4), C_h and D_h are matrices defined by

$$C_h := \begin{bmatrix} \int_0^{h/L} Ce^{A\theta} \mathrm{d}\theta \\ \int_{h/L}^{2h/L} Ce^{A\theta} \mathrm{d}\theta \\ \vdots \\ \int_{(L-1)h/L}^{h} Ce^{A\theta} \mathrm{d}\theta \end{bmatrix}, \quad D_h := \begin{bmatrix} \int_0^{h/L} \int_0^{\theta} Ce^{A\tau} \mathrm{d}\tau \mathrm{d}\theta \\ \int_{h/L}^{2h/L} \int_0^{\theta} Ce^{A\tau} \mathrm{d}\tau \mathrm{d}\theta \\ \vdots \\ \int_{(L-1)h/L}^{h} \int_0^{\theta} Ce^{A\tau} \mathrm{d}\tau \mathrm{d}\theta \end{bmatrix}.$$

Note that the integrals in B_h, C_h, and D_h can be effectively computed by using matrix exponentials [1, 7].

Let us summarise the proposed adaptive algorithm. The continuous-time error $e(t)$ is sampled with the fast sampling period h/L and blocked to become the discrete-time signal $e[n]$, and the signal $x(t)$ is sampled with the sampling period h to become $x_\mathrm{d}[n]$. Then the sampled signal x_d is filtered by $F_h(z)$ and the signal $U[n]$ is obtained. By using $e[n]$ and $\{U[n], U[n-1], \ldots, U[n-N+1]\}$, we update the filter coefficient $\alpha[n]$ by (12) and (13) with

$$\int_{nh}^{(n+1)h} e(t)u(t)\mathrm{d}t \approx \begin{bmatrix} e[n]^\top U[n] \\ e[n]^\top U[n-1] \\ \vdots \\ e[n]^\top U[n-N+1] \end{bmatrix}.$$

We show the proposed adaptive scheme in Fig. 4.

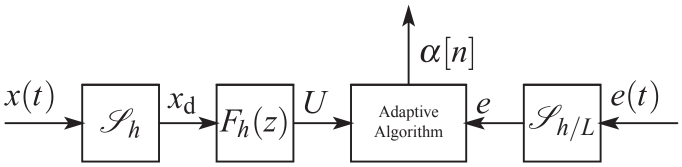


Figure 4: filtered-x adaptive scheme

5 Simulation

In this section, we show simulation results of active noise control. The analog systems $F(s)$ and $P(s)$ are given by

$$F(s) = \frac{1}{s+1.1} \cdot \frac{1}{20} \sum_{k=1}^{4} \frac{k^2}{s^2 + 2\zeta ks + k^2},$$

$$P(s) = \frac{1.2 \times 1.3}{(s+1.2)(s+1.3)} \cdot \frac{1}{20} \sum_{k=1}^{4} \frac{(1.2k)^2}{s^2 + 2\zeta(1.2k)s + (1.2k)^2}.$$

The Bode gain plots of these systems are shown in Fig. 5. The gain $|F(\mathrm{j}\omega)|$ has peaks at $\omega = 1,2,3,4$ (rad/sec) and $|P(\mathrm{j}\omega)|$ has peaks at $\omega = 1.2, 2.4, 3.6, 4.8$ (rad/sec). We set the sampling period $h = 1$ (sec) and the fast-sampling ratio $L = 8$. Note that the systems $F(s)$ and $P(s)$ are stable and have peaks beyond the Nyquist frequency $\omega = \pi$ (rad/sec).

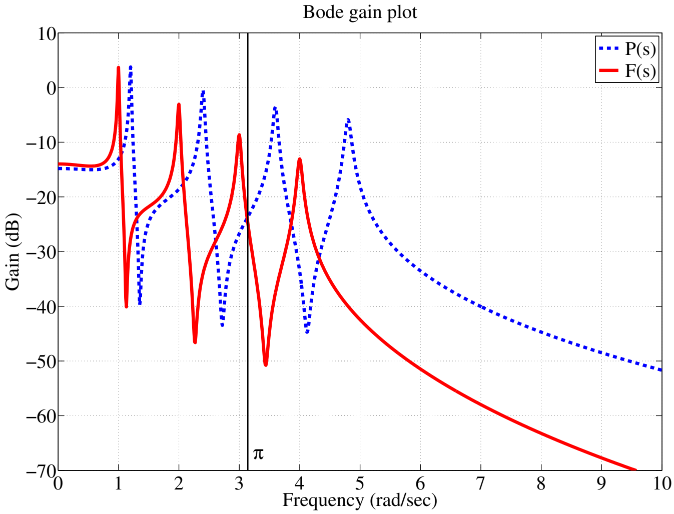


Figure 5: Frequency response of $F(s)$ (dash) and $P(s)$ (solid). The vertical line indicates the Nyquist frequency π (rad/sec).

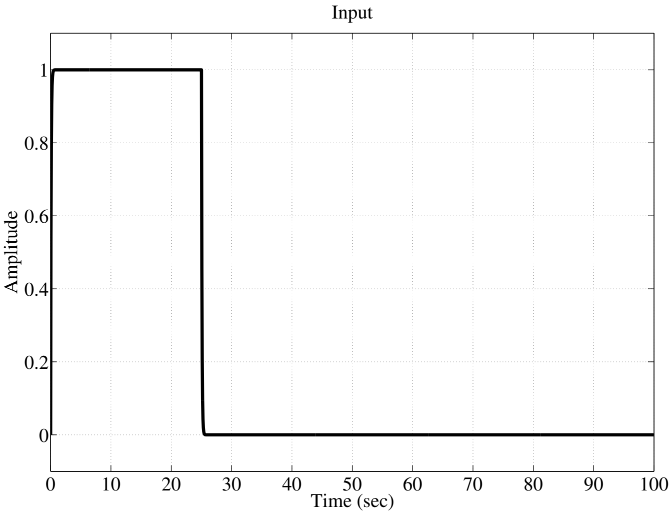


Figure 6: Input signal $x(t)$ with $0 \leq t \leq 100$ (sec).

Then we run a simulation of active noise control by the proposed method with the input signal $x(t)$ shown in Fig. 6. Note that the input $x(t)$ belongs to L^2 and satisfies our assumption. To compare with the proposed method, we also run a simulation by a standard discrete-time LMS algorithm [2], which is obtained by setting the fast-sampling parameter L to be 1. The step-size parameter μ in the coefficient update in (12) is set to be 0.1.

Fig. 7 shows the absolute values of error signal $e(t)$ (see Fig. 1 or Fig. 2). The errors by the conventional design is much larger than that by the proposed method. In fact, the L^2 norm of the error signal $e(t)$, $0 \le t \le 100$ (sec) is 2.805 for the conventional method and 1.392 for the proposed one, which is improved by about 49.6%. The result shows the effectiveness of our method.

Fig. 8 shows the L^2 norm of the error $e(t)$, $0 \le t \le 100$ (sec) with some values of the step-size parameter μ. Fig. 8 shows that the error by the proposed method is equal to or smaller than that by the conventional method for almost all values of μ. Moreover, the error by the proposed method can be small for much wider interval than that by the conventional method. In fact, the L^2 norm of the error $\|e\|_2 < 10$ if $\mu \in (0, 0.7257)$ by the proposed method, while $\|e\|_2 < 10$ if $\mu \in (0, 0.4051)$ by the conventional method. That is, the interval by the proposed method is about 1.8 times wider than that by the conventional method.

In summary, the simulation results show that the proposed method gives better performance for wider interval of the step-size parameter μ on which the adaptive system is stable than the conventional method.

6 Conclusion

In this article, we have proposed a hybrid design of filtered-x adaptive algorithm via lifting method in sampled-data control theory. The proposed algorithm can take account of the continuous-time behavior of the error signal. We have also proposed an approximation of the algorithm, which can be easily implemented in DSP. Simulation results have shown the effectiveness of the proposed method.

Acknowledgments

This research is supported in part by the JSPS Grant-in-Aid for Scientific Research (B) No. 2136020318360203, Grant-in-Aid for Exploratory Research No. 22656095, and the MEXT Grant-in-Aid for Young Scientists (B) No. 22760317.

Bibliography

[1] T. Chen and B. Francis. *Optimal Sampled-Data Control Systems.* Springer, 1995. Cited pp. 276, 278, 279, and 286.

[2] S. J. Elliott and P. A. Nelson. Active noise control. *IEEE Signal Processing Mag.*, 10-4:12–35, 1993. Cited pp. 275, 276, and 287.

[3] S. Haykin. *Adaptive Filter Theory.* Prentice Hall, 1996. Cited pp. 278 and 281.

[4] K. Kashima, Y. Yamamoto, and M. Nagahara. Optimal wavelet expansion via sampled-data control theory. *IEEE Signal Processing Lett.*, 11-2:79–82, 2004. Cited p. 276.

[5] Y. Kobayashi and H. Fujioka. Active noise cancellation for ventilation ducts using a pair of loudspeakers by sampled-data H_∞ optimization. *Advances in Acoustics and Vibration*, Article ID 253948, 2008. Cited p. 275.

[6] S. Kuo, S. Mitra, and W.-S. Gan. Active noise control system for headphone applications. *IEEE Trans. Contr. Syst. Technol.*, 14(2):331 –335, 2006. Cited p. 275.

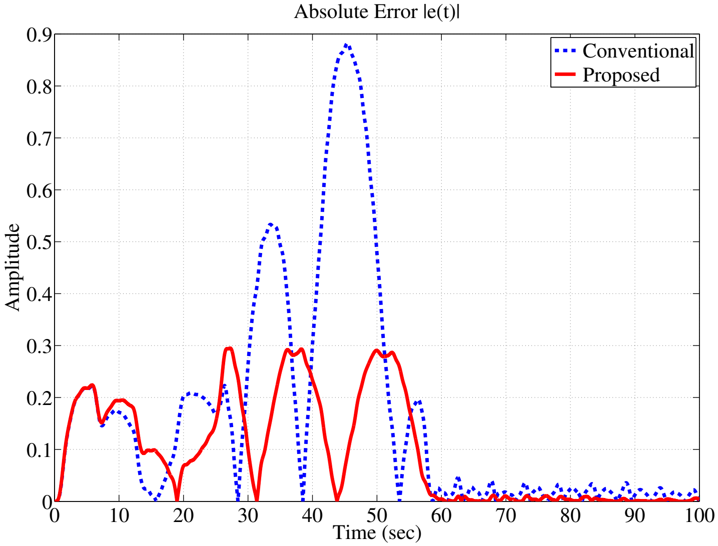


Figure 7: Absolute values of error signal $e(t)$: conventional (dash) proposed (solid).

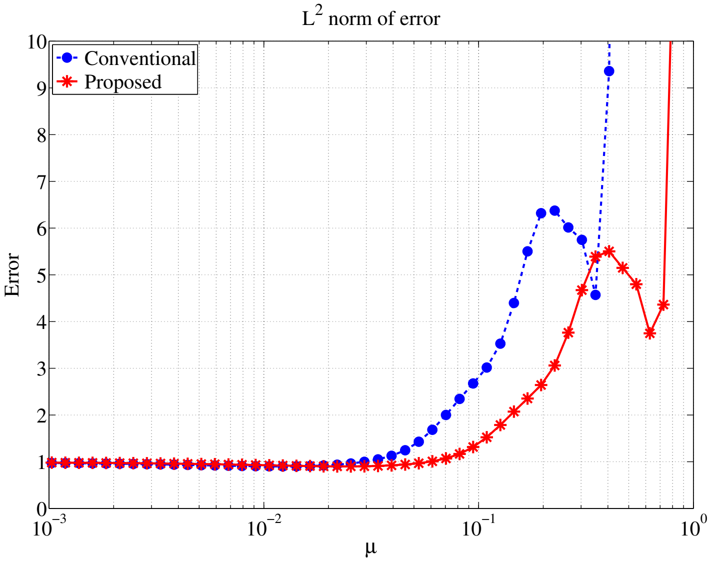


Figure 8: L^2 norm of the error $e(t)$: conventional (dash) and proposed (solid).

[7] C. F. V. Loan. Computing integrals involving the matrix exponential. *IEEE Trans. Automat. Contr.*, 23:395–404, 1994. Cited p. 286.

[8] T. Meurers, S. M. Veres, and S. J. Elliott. Frequency selective feedback for active noise control. *IEEE Signal Processing Mag.*, 22-4:32–41, 2002. Cited pp. 275 and 276.

[9] D. R. Morgan. An analysis of multiple correlation cancellation loops with a filter in the auxiliary path. *IEEE Trans. Signal Processing*, ASSP-28:454–467, 1980. Cited p. 276.

[10] M. Nagahara and Y. Yamamoto. Optimal design of fractional delay FIR filters without band-limiting assumption. *Proc. of IEEE ICASSP*, 4:221–224, 2005. Cited p. 276.

[11] W. J. Rugh. *Linear System Theory*. Prentice-Hall, 2nd edition, 1996. Cited p. 284.

[12] R. Shoureshi and T. Knurek. Automotive applications of a hybrid active noise and vibration control. *IEEE Control Syst. Mag.*, 16(6):72 –78, 1996. Cited p. 275.

[13] Y. Song, Y. Gong, and S. M. Kuo. A robust hybrid feedback active noise cancellation headset. *IEEE Trans. Signal Processing*, 13:607–617, 2005. Cited p. 276.

[14] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993. Cited p. 276.

[15] Y. Yamamoto. A function space approach to sampled-data control systems and tracking problems. *IEEE Trans. Automat. Contr.*, 39:703–712, 1994. Cited pp. 276, 278, and 279.

[16] Y. Yamamoto, M. Nagahara, and P. P. Khargonekar. Signal reconstruction via H^∞ sampled-data control theory — Beyond the shannon paradigm. *IEEE Trans. Signal Processing*, 60(2):613–625, 2012. Cited p. 276.

[17] D. Yasufuku, Y. Wakasa, and Y. Yamamoto. Adaptive digital filtering based on a continuous-time performance index. *SICE Transactions*, 39-6, 2003. Cited p. 276.

Linear switching systems and random products of matrices

Masaki Ogura
Department of Mathematics and
Statistics, Texas Tech University,
Broadway and Boston
Lubbock, TX 79409-1042, USA
masaki.ogura@ttu.edu

Clyde F. Martin
Department of Mathematics and
Statistics, Texas Tech University,
Broadway and Boston
Lubbock, TX 79409-1042, USA
clyde.f.martin@ttu.edu

Abstract. Dayawansa and Martin proved that the switching system $\dot{x} = (\delta(t)A + (1 - \delta(t)B)x(t)$, $\delta(t) \in \{0,1\}$ is stable for all switching sequences if and only if the two systems $\dot{x} = Ax$ and $\dot{x} = Bx$ have a common Lyapunov function. This theorem can be interpreted in the following manner. Let $G_1 = \{e^{A\tau} : \tau \in \mathbb{R}^+\}$ and $G_2 = \{e^{B\tau} : \tau \in \mathbb{R}^+\}$ where $\mathbb{R}^+ = \{r \in \mathbb{R} : r \geq 0\}$. Let $X_n = D_n \cdots D_1$ where $D_{2i} \in G_2$ and $D_{2i+1} \in G_1$. This is constructed by random sampling from $G_1 \cup G_2$. The theorem of Dayawansa and Martin can then be interpreted as $\|X_n\| \to 0$ if and only if $\dot{x} = Ax$ and $\dot{x} = Bx$ have a common Lyapunov function.

1 Introduction

The theory of switching systems in control theory [11] is a well-developed field of research as is the theory of random products of matrices [3] in probability theory. However there has been very little overlap between the two branches of mathematics. In this paper we explore some of the connections between the two fields.

Consider first the system

$$\dot{x} = (\delta(t)A + (1 - \delta(t))B)x$$

where A and B are real $d \times d$ matrices and $\delta(t) \in \{0,1\}$. This system has been studied quite extensively, see for example [6] or [8]. Now consider the semigroups

$$G_1 = \{e^{A\tau} : \tau \in \mathbb{R}^+\}$$

and

$$G_2 = \{e^{B\tau} : \tau \in \mathbb{R}^+\}.$$

We construct a product of the form

$$P_n = e^{A\tau_n} e^{B\tau_{n-1}} \cdots e^{A\tau_1}$$

by sampling randomly from $G_1 \cup G_2$. Let $X_n = D_n \cdots D_1$ where $D_{2i} \in G_2$ and $D_{2i+1} \in G_1$. This is constructed by random sampling from $G_1 \cup G_2$. The theorem of Dayawansa and Martin can then be interpreted as $\|X_n\| \to 0$ if and only if $\dot{x} = Ax$ and $\dot{x} = Bx$ have a common Lyapunov function. Discrete time systems can be interpreted in a similar manner using the corresponding semigroups.

The control theory literature has been primarily concerned with two problems. First and foremost is the problem of stability and this problem was addressed in [6] and in many other papers. The second problem is the problem of controllability. This problem is addressed for continuous-time switching systems in [1] and in other papers. The controllability problem for discrete time systems seems to be considerably harder.

This paper is organized as follows. Section 2 studies the stability of discrete-time switching systems. The equivalence between the two stability notions considered in [9] and [14] is proved. In Section 3 an estimate of the growth rate of switching systems, not necessarily stable, is shown. Section 4 studies the controllability of switching systems.

1.1 Notations and conventions

For a real number x let $\log^+(x) := \max(\log x, 0)$. For a matrix $M \in \mathbb{R}^{d\times d}$, its maximal singular value is denoted by $\|M\|$. Let $\mathcal{M}$ be a subset of $\mathbb{R}^{d\times d}$. Define $\|\mathcal{M}\|$ by

$$\|\mathcal{M}\| := \sup_{B\in\mathcal{M}} \|B\|.$$

For a positive integer k let $\mathcal{M}^k$ be the set of k-products of the matrices in $\mathcal{M}$; i.e.,

$$\mathcal{M}^k := \{B_1 \cdots B_k : B_1, \dots, B_k \in \mathcal{M}\}.$$

If all the matrices in $\mathcal{M}$ are invertible then we write

$$\mathcal{M}^{-1} := \{B^{-1} : B \in \mathcal{M}\}.$$

Finally let $\mathrm{Gl}(d,\mathbb{R})$ denote the multiplicative group of invertible $d \times d$ real matrices.

2 Stability

Consider the discrete-time switching system

$$x(k+1) = A_{\sigma_k} x(k), \quad A_{\sigma_k} \in \mathcal{M} \tag{1}$$

with $\mathcal{M}$ consisting of finite number of matrices:

$$\mathcal{M} = \{A_1, \dots, A_m\} \subset \mathbb{R}^{d\times d}.$$

The *joint spectral radius* [12] of $\mathcal{M}$ is defined by

$$\rho(\mathcal{M}) := \lim_{k\to\infty} \max_{B\in\mathcal{M}^k} \|B\|^{1/k}.$$

It is well known that this quantity characterizes the stability of the switching system [2, 13]:

Theorem 1. *The switching system* (1) *converges to the origin for any initial point* x_0 *if and only if*

$$\rho(\mathcal{M}) < 1.$$

Jungers et. al. [9] extended the notion of joint spectral radius and then introduced a novel notion of stability in the following way. We here quote some definitions from [9]. For a parameter $p \in [1, \infty]$ the *L^p-norm joint spectral radius* (*p-radius*, in short) is defined by

$$\rho_p := \lim_{k\to\infty} \left[m^{-k} \sum_{B \in \mathcal{M}^k} \|B\|^p \right]^{1/(pk)},$$
$$\rho_\infty := \rho(\mathcal{M}).$$

It was observed [9] that ρ_p is a nondecreasing function of p so that

$$\rho_1 \leq \rho_p \leq \rho(\mathcal{M})$$

for every p. The switching system is said to be *p-weakly stable* if

$$\rho_p < 1 < \rho(\mathcal{M}).$$

On the other hand, Wang et. al. [14] studied the stability of mean and variance of the solution of the stochastic switching system

$$X(k+1) = A_{\sigma_k} X(k),\ X(0) \in \mathbb{R}^{d\times d},\ A_{\sigma_k} \in \mathcal{M}$$

where the constant probability $1/m$ is assigned for all the matrices $A_1, \ldots, A_m$. The mean E_k and the variance V_k at time k are given by

$$E_k = m^{-k} \sum_{B \in \mathcal{M}^k} B,$$
$$V_k = m^{-k} \sum_{B \in \mathcal{M}^k} (B - E_k)^\top (B - E_k).$$

It was shown [14] that the stability of E_k and V_k are characterized by the spectral radiuses of certain matrices, which are easily computable.

We here remark that a similar result was obtained in [4], which studied switching systems as Markov jump linear systems. Also [5] characterized the so called mean square stability of the switching system (1) with respect to a special transition probability using the norm of a finite product of matrices $A_1, \ldots, A_m$.

The aim of this section is to prove the next theorem that shows the equivalence between the 2-weak stability and the stability of mean and variance.

Theorem 2. *E_k and V_k converges to 0 exponentially as $k \to \infty$ if and only if $\rho_2 < 1$.*

This theorem gives another interpretation of the stability of the system studied in [14].

Example 3. In [14] the authors studied the switching system (1) given by the matrices

$$A_1 = \begin{bmatrix} 0.9739 & 0.0098 \\ -0.9772 & 0.9739 \end{bmatrix},\ A_2 = \begin{bmatrix} 0.9719 & 0.0975 \\ -0.0975 & 0.9719 \end{bmatrix}.$$

The JSR (Joint Spectral Radius) Toolbox of MATLAB gives

$$\rho_\infty \in [1.0395, 1.0856],$$

where the method by Blondel et. al. [2] is used. On the other hand, using Proposition 2.3 of [10] we can see that

$$\rho_2 = 0.9876 < 1.$$

Therefore, by Theorem 2, the mean E_k and variance V_k converges to 0 exponentially fast, which was originally proved in [14] by showing that the spectral radiuses of certain matrices are less than 1.

Remark 4. One may be tempted to conjecture that E_k exponentially converges to 0 if and only if $\rho_1 < 1$. This is not true. Consider the scalar switching system (1) with $\mathcal{M} = \{1, -1\}$. Then the mean E_k is always equal to 0 but we have $\rho_1 = 1$.

From Theorem 2 and this remark one may conjecture the following:

Conjecture 5. Let m be a positive integer. All the moments with order less than or equal to $2m$ exponentially converge to 0 if and only if $\rho_{2m} < 1$.

Now let us prove Theorem 2. We need the following easy lemma.

Lemma 6. *If B is a $d \times d$ matrix then*

$$\|B\|^2 \le 2\sum_{n=1}^{d} \|Be_n\|^2, \tag{2}$$

where $e_1, \dots, e_d$ are the standard basis of $\mathbb{R}^d$.

Also we will use the equality

$$V_k = m^{-k} \sum_{B \in \mathcal{M}^k} (B^\top B) - E_k^\top E_k \tag{3}$$

that was shown in [14].

Proof of Theorem 2. Suppose that E_k and V_k exponentially converge to 0 as $k \to \infty$. Then (3) yields that $m^{-k}\sum_{B\in\mathcal{M}^k}(B^\top B)$ exponentially converges to 0 as $k \to \infty$. Therefore there exists $C > 0$, $\lambda < 1$, and K such that, for every $n = 1, 2, \dots, d$, if $k > K$ then

$$m^{-k} \sum_{B\in\mathcal{M}^k} \|Be_n\|^2 < C\lambda^k.$$

Then, if $k > K$,

$$\begin{aligned} m^{-k} \sum_{B\in\mathcal{M}^k} \|B\|^2 &\le m^{-k} \sum_{B\in\mathcal{M}^k} 2\sum_{n=1}^{d} \|Be_n\|^2 \quad \text{by (2)} \\ &= 2\sum_{n=1}^{d} m^{-k} \sum_{B\in\mathcal{M}^k} \|Be_n\|^2 \\ &< 2dC\lambda^k. \end{aligned}$$

Therefore

$$\left[m^{-k}\sum_{B\in\mathcal{M}^k}\|B\|^2\right]^{1/2k} < (2dC)^{1/2k}\sqrt{\lambda}$$

and hence

$$\rho_2 \le \sqrt{\lambda} < 1.$$

On the other hand assume $\rho_2 < 1$. By the monotonicity of the p-radius we have $\rho_1 < 1$. Therefore there exists K such that if $k > K$ then

$$\left[m^{-k}\sum_{B\in\mathcal{M}^k}\|B\|\right]^{1/k} < c < 1.$$

Therefore

$$\|E_k\| \le m^{-k}\sum_{B\in\mathcal{M}^k}\|B\| < c^k.$$

Moreover, because $\rho_2 < 1$, there exists K' such that if $k > K'$ then

$$\left[m^{-k}\sum_{B\in\mathcal{M}^k}\|B\|^2\right]^{1/(2k)} < c' < 1.$$

Hence if $k > \max(K, K')$ then

$$\|V_k\| \le m^{-k}\sum_{B\in\mathcal{M}^k}(\|B\|^2) + \|E_k\|^2 \le c'^{2k} + c^{2k}.$$

This completes the proof. □

3 Growth rate

Let $X_1, X_2, \cdots$ be a stationary stochastic process with values in the space of $d\times d$ real matrices. Furstenberg et. al. [7] proved a fundamental result about the growth rate of the products of random matrices.

Theorem 7 ([7, Theorem 3.9]). *Let μ be a probability measure on* $\mathrm{Gl}(d,\mathbb{R})$ *satisfying*

$$\int [\log^+\|g\| + \log^+\|g^{-1}\|]\,d\mu(g) < \infty.$$

Then there is a sequence of subspaces

$$0 \subset L_r \subset L_{r-1} \subset \cdots \subset L_2 \subset L_1 \subset L_0 = \mathbb{R}^d$$

and a sequence of values

$$\beta^0(\mu) > \beta'(\mu) > \beta''(\mu) > \cdots > \beta^{(r)}(\mu)$$

such that if $x_0 \in L_i \backslash L_{i+1}$ then with probability one

$$\lim_{N\to\infty}\frac{1}{N}\log\|X_N X_{N-1}\cdots X_1 x_0\| = \beta^{(i)}(\mu).$$

We can give an estimate for the numbers $\beta^{(n)}$:

Theorem 8. *We have*

$$\beta^{0}(\mu) \le \log\|\mathcal{M}\|. \tag{4}$$

Moreover if all the matrices in $\mathcal{M}$ are invertible then

$$-\log\|\mathcal{M}^{-1}\| \le \beta^{(r)}(\mu). \tag{5}$$

Proof. Because $\|x(k+1)\| \le \|A_{\sigma_k}\|\|x_k\| \le \|\mathcal{M}\|\|x_k\|$ for every k, we have $\|x_k\| \le \|\mathcal{M}\|^k$ $\|x_0\|$ and hence

$$\frac{1}{k}\log\|x_k\| \le \log\|\mathcal{M}\| + \frac{\|x_0\|}{k}$$

so that

$$\limsup_{k\to\infty}\frac{1}{k}\log\|x_k\| \le \log\|\mathcal{M}\|.$$

This shows (4)

Then assume that all the matrices in $\mathcal{M}$ are invertible. Then $x(k) = A_{\sigma_k}^{-1}x(k+1)$ so that $\|x(k)\| \le \|A_{\sigma_k}^{-1}\|\|x(k+1)\|$ and hence

$$\begin{aligned}\|x(k+1)\| &\ge \|A_{\sigma}^{-1}\|^{-1}\|x(k)\| \\ &\ge \inf_{\sigma}\|A_{\sigma}^{-1}\|^{-1}\|x(k)\| \\ &= (\sup\|A_{\sigma}^{-1}\|)^{-1}\|x(k)\| \\ &= \|\mathcal{M}^{-1}\|^{-1}\|x(k)\|.\end{aligned}$$

The same argument above gives us

$$\liminf_{k\to\infty}\frac{1}{k}\log\|x(k)\| \ge -\log\|\mathcal{M}^{-1}\|.$$

This proves (5). □

4 Controllability

4.1 Continuous-time case

The switching system

$$\dot{x}(t) = A_{\sigma_t}x(t),\ A_{\sigma_t} \in \mathcal{M} \subset \mathbb{R}^{d\times d} \tag{6}$$

is said to be *globally controllable* [1] if for every $x_0, x_f \in \mathbb{R}^d\setminus\{0\}$ there exists a switching signal σ such that the solution of (6) satisfies $x(0) = x_0$ and $x(T) = x_f$ for some $T > 0$.

Let us see some examples of globally controllable switching systems.

Example 9. Let $\mathcal{M} = \{A_1, A_2\}$ where

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ A_2 = \begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix}.$$

The associated switching system Σ_2 can be checked to be globally controllable.

Example 10. Let us generalize the above example. Consider the switching system Σ_d ($d \geq 2$) associated with d matrices $\{B_1, B_2, \dots, B_d\} \subset \mathbb{R}^{d \times d}$ defined by

$$B_1 = \begin{bmatrix} A_1 & \\ & O_{d-2} \end{bmatrix},$$

$$B_2 = \begin{bmatrix} A_2 & \\ & O_{d-2} \end{bmatrix},$$

$$B_k(i,j) = \begin{cases} -1, & (i,j) = (k-1,k) \\ 1, & (i,j) = (k,k-1) \\ 0, & \text{otherwise} \end{cases} \qquad (k \geq 3)$$

where O_{d-2} is the zero matrix of size $(d-2) \times (d-2)$. We can prove the global controllability of Σ_d inductively. First we know that Σ_2 is globally controllable by Example 9. Then assume that Σ_d is globally controllable and consider Σ_{d+1}. Let $\tilde{\Sigma}_d$ be the switching system defined by

$$\tilde{\Sigma}_d := \begin{bmatrix} \Sigma_d & \\ & 0 \end{bmatrix}$$

with the state space $\mathbb{R}^{d+1}$. We can observe that an initial state $x \neq 0$ can be steered to arbitrary $y \neq 0$ as

$$\begin{bmatrix} x_1 \\ \vdots \\ x_{d-1} \\ x_d \\ x_{d+1} \end{bmatrix} \xrightarrow{\dot{x} = B_{d+1}x} \begin{bmatrix} x_1 \\ \vdots \\ x_{d-1} \\ \sqrt{x_d^2 + x_{d+1}^2} \\ 0 \end{bmatrix} \xrightarrow{\tilde{\Sigma}_d} \begin{bmatrix} y_1 \\ \vdots \\ y_{d-1} \\ \sqrt{y_d^2 + y_{d+1}^2} \\ 0 \end{bmatrix} \xrightarrow{\dot{x} = B_{d+1}x} \begin{bmatrix} y_1 \\ \vdots \\ y_{d-1} \\ y_d \\ y_{d+1} \end{bmatrix}.$$

This example shows that we can construct a globally controllable switching system using at most d subsystems. A conjecture is that:

Conjecture 11. If the switching system (6) is globally controllable then $\mathcal{M}$ has at least d matrices.

To prove this conjecture we might be able to use a result from [1]. Altafini [1] rewrote the switching system (6) as

$$\dot{x} = \mathcal{F}(x,u) = \sum_{i=1}^{m} u_i A_i x$$

where $u_i \in \{0,1\}$ and $\sum_{i=1}^{m} u_i = 1$ and studied this equation as a bilinear system, giving the next result.

Theorem 12 ([1]). *The switching system* (6) *is globally controllable if and only if*

$$\operatorname{rank}(\mathrm{Lie}(\mathcal{F})) = d.$$

4.2 Discrete-time case

This subsection studies the controllability of discrete-time switching system (1), in particular the scalar switching system

$$x(k+1) = (\delta(k)\lambda + (1-\delta(k))\mu)x(k) \tag{7}$$

where $\lambda, \mu \in \mathbb{R}$ and δ is a $\{0,1\}$-valued switching signal.
This system cannot be globally controllable because the reachable set

$$\{\lambda^p \mu^q x_0\}_{p\geq 0,\ q\geq 0}$$

is countable. This observation leads us to the following general definition.

Definition 13. The switching system (1) is said to be *globally approximately controllable* if for every $x_0, x_f \in \mathbb{R}^d \backslash \{0\}$ and $\varepsilon > 0$ there exists a switching signal σ such that the solution of (1) satisfies $x(0) = x_0$ and $\|x(T) - x_f\| < \varepsilon$ for some $T > 0$.

The main result of this subsection is the following characterization of the global approximate controllability.

Theorem 14. *The scalar switching system* (7) *is globally approximately controllable if and only if the following two conditions hold:*

- *Either one of λ and μ is negative;*
- *The number $\frac{\log|\lambda|}{\log|\mu|}$ is negative and irrational.*

Proof of the necessity of Theorem 14. Assume that the scalar switching system (7) is globally approximately controllable. If both λ and μ are positive then we can see that the sign of $x(k)$ is the same as that of x_0 so that the trajectory of the system is contained in one of the half lines $(-\infty, 0]$ and $[0, \infty)$, which contradicts to the controllability. Therefore either λ or μ is negative.
Secondly if $\frac{\log|\lambda|}{\log|\mu|}$ is positive then either $|\lambda|, |\mu| > 1$ or $|\lambda|, |\mu| < 1$ holds and hence we have $|x(k)| \geq |x_0|$ or $|x(k)| \leq |x_0|$, respectively, which again contradict to the controllability.
Finally assume that $\frac{\log|\lambda|}{\log|\mu|} = q/p$ for some integers p, q. Then $|\lambda| = |\mu|^{q/p}$ so that the set of possible values of $|x(k)|$ is included in $\{|\mu|^{n/p}|x_0|\}_{n\in\mathbb{Z}}$, which is not dense in $\mathbb{R}$. This completes the proof. □

For the proof of sufficiency we need the next lemma:

Lemma 15. *If α is a negative irrational number then the set $\{m + n\alpha\}_{m,n\geq 0}$ is dense in $\mathbb{R}$.*

Proof. This lemma is a simple application of the equidistribution theorem: If α is an irrational number then the sequence

$$\alpha,\ 2\alpha,\ 3\alpha, \ldots \bmod 1$$

is uniformly distributed on the unit interval.

Let $z \in \mathbb{R}$ and $\varepsilon > 0$ be arbitrary. Let $\{z\}$ be the fractional part of z. Because $\{n(-\alpha)\}_{n\geq 0}$ is dense in $[0,1]$ modulo 1, there exists $n \geq 0$ such that

$$-\varepsilon < n(-\alpha) + \{z\} < \varepsilon \quad \bmod 1. \tag{8}$$

This n can be taken sufficiently large so without loss of generality we assume that

$$n\alpha - z < -\varepsilon. \tag{9}$$

By (8) there exists an integer m such that

$$-\varepsilon < (n(-\alpha) + \{z\}) - m < \varepsilon.$$

and hence

$$-\varepsilon < ((m + \lfloor z \rfloor) + n\alpha) - z < \varepsilon.$$

Because $(m + \lfloor z \rfloor) \geq 0$ by (9) and $n \geq 0$ this completes the proof. □

Now we are at the position of proving the sufficiency of Theorem 14.

Proof of the sufficiency of Theorem 14. Suppose that the two conditions in Theorem 14 hold. It is sufficient to show that the set

$$R := \{\lambda^m \mu^n x_0\}_{m,n\geq 0}$$

is dense in $\mathbb{R}$. Without loss of generality assume that $x_0 > 0$ and $\lambda < 0$.

Since $\alpha = \frac{\log \mu^2}{\log \lambda^2} = \frac{\log|\mu|}{\log|\lambda|}$ is a negative irrational number, by the above lemma the set

$$\left\{(\log \lambda^2)(m + n\alpha)\right\}_{m,n\geq 0}$$

is dense in $\mathbb{R}$ and hence its image by the exponential mapping

$$\{\lambda^{2m}\mu^{2n}\}_{m,n\geq 0}$$

is dense in $[0,\infty)$. Therefore, because $\lambda < 0$, the set

$$\{\lambda^{2m+1}\mu^{2n}\}_{m,n\geq 0}$$

is dense in $(-\infty, 0]$. Hence the set R is dense in $\mathbb{R}$. This completes the proof. □

Define the distance between the switching systems

$$\Sigma : x(k+1) = (\delta(k)\lambda + (1-\delta(k))\mu)x(k),$$
$$\Sigma' : x(k+1) = (\delta(k)\lambda' + (1-\delta(k))\mu')x(k)$$

by

$$d(\Sigma, \Sigma') := |\lambda - \lambda'| + |\mu - \mu'|.$$

Corollary 16. *Let Σ be a globally approximately controllable switching system. For every $\varepsilon > 0$ there exists a not globally approximately controllable switching system Σ' such that $d(\Sigma, \Sigma') < \varepsilon$.*

Controllability of higher dimensional switching systems seems to be complicated.

Bibliography

[1] C. Altafini. The reachable set of a linear endogenous switching system. *Systems and Control Letters*, 47:343–353, 2002. Cited pp. 292, 296, 297, and 298.

[2] V. D. Blondel and Y. Nesterov. Computationally efficient approximations of the joint spectral radius. *SIAM Journal on Matrix Analysis and Applications*, 27(1):256–272, 2005. Cited pp. 292 and 294.

[3] P. Bougerol and J. Lacroix. *Products of Random Matrices with Applications to Schrödinger Operators*. Progress in Probability and Statistics. Birkhäuser, 1985. Cited p. 291.

[4] O. L. V. Costa, M. D. Fragoso, and R. P. Marques. *Discrete-Time Markov Jump Linear Systems*. Springer, 2005. Cited p. 293.

[5] X. Dai, Y. Huang, and M. Xiao. Almost sure stability of discrete-time switched linear systems: A topological point of view. *SIAM Journal on Control and Optimization*, 47(4):2137–2156, 2008. Cited p. 293.

[6] W. P. Dayawansa and C. F. Martin. A converse Lyapunov theorem for a class of dynamical systems which undergo switching. *IEEE Transactions on Automatic Control*, 44(4):751–760, 1999. Cited pp. 291 and 292.

[7] H. Furstenberg and Y. Kifer. Random matrix products and measures on projective spaces. *Israel Journal of Mathematics*, 46(1):12–32, 1983. Cited p. 295.

[8] B. Hanlon, N. Wang, M. Egerstedt, and C. Martin. Switched linear systems: Stability and the convergence of random products. Under review. Cited p. 291.

[9] R. M. Jungers and V. Y. Protasov. Weak stability of switching dynamical systems and fast computation of the p-radius of matrices. In *Proceedings of the 49th IEEE Conference on Decision and Control*, pages 7328–7333, 2010. Cited pp. 292 and 293.

[10] R. M. Jungers and V. Y. Protasov. Fast methods for computing the p-radius of matrices. *SIAM Journal on Scientific Computing*, 33(3):1246–1266, 2011. Cited p. 294.

[11] D. Liberzon. *Switching in Systems and Control*. Birkhäuser, 2003. Cited p. 291.

[12] G.-C. Rota and W. G. Strang. A note on the joint spectral radius. *Indagationes Mathematicae*, 22(4):379–381, 1960. Cited p. 292.

[13] J. Theys. *Joint Spectral Radius: Theory and approximations*. PhD thesis, Université catholique de Louvain, 2005. Cited p. 292.

[14] N. Wang, M. Egerstedt, and C. Martin. Stability of switched linear systems and the convergence of random products. In *Proceedings of the 48h IEEE Conference on Decision and Control*, pages 3721–3726, 2009. Cited pp. 292, 293, and 294.

Some remarks on discrete-time unstable spectral factorization

Giorgio Picci
University of Padova
Italy
picci@dei.unipd.it

Abstract. This paper is a simple variation on the theme of rational spectral factorization allowing for unstable spectral factors. As an application we discuss the structure of all-pass rational functions.

1 Introduction

The following is a classical problem in System Theory. Given a discrete-time purely non deterministic (p.n.d.) m-dimensional stationary process $\mathbf{y}$ with a rational spectral density matrix $\Phi(z)$, one wants to describe and classify the state space realizations of $\mathbf{y}$; i.e. the representations of the type

$$\begin{cases} \mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{w}(t), & \\ \quad \mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{w}(t), & t \in \mathbb{Z} \end{cases} \tag{1}$$

where $\mathbf{w}$ is a normalized white noise process and (A, B, C, D) are constant matrices. When $\dim \mathbf{x}(t)$ is the smallest possible the representation is called a *minimal realization* of $\mathbf{y}$. The solution involves computing the rational spectral factors of $\Phi(z)$ parametrized in the form $W(z) = C(zI - A)^{-1}B + D$. By definition *minimal spectral factors* are spectral factors of minimal McMillan degree. Here we will not worry about constructing the noise process $\mathbf{w}$. Traditionally this problem is solved by looking for a parametrization of all minimal *analytic (or causal)* spectral factors in terms of solutions of a certain Linear Matrix Inequality. When A is asymptotically stable, which we shall write $|\lambda(A)| < 1$, the state space model (1) is a *causal* or *forward* representation since the equations can be solved to yield

$$\mathbf{x}(t) = \sum_{s=-\infty}^{t-1} A^{t-1-s} B\mathbf{w}(s), \qquad \mathbf{y}(t) = \sum_{s=-\infty}^{t-1} CA^{t-1-s} B\mathbf{w}(s) + D\mathbf{w}(t)$$

which represent $\mathbf{x}(t)$ and $\mathbf{y}(t)$ as causal functions of the past $\{\mathbf{w}(s)\,;\, s \leq t\}$. However this is just one particular representation out of many possible others. A random process is just a *flow*; i.e. has no intrinsic causality. State space representations could be anti-causal or even of "mixed" causal-anticausal type. The only condition which is really needed is that the entries of the impulse response of the system (1) should be in $\ell^2(\mathbb{Z})$, compare e.g. [9]. Equivalently the spectrum of $W(z)$ should satisfy $\sigma(W(z)) \cap \{|z| = 1\} = \emptyset$ but it could otherwise be quite arbitrary. In this paper we want to study the realization problem without causality constraints. This has some interesting applications as we shall see. Acausal realizations in continuous-time were studied in [7], however so far, the discrete-time problem seems to have remained unsolved.

This paper is dedicated to Uwe Helmke a long-time friend and colleague.

2 A quick introduction to rational spectral factorization

A rational $m \times m$ matrix function $\Phi(z)$ is a *spectral density* iff it satisfies the following conditions

1. the parahermitian symmetry $\Phi(z) = \Phi(z^{-1})^\top$,
2. $\Phi(e^{j\theta})$ is integrable on the unit circle; hence has no poles on the unit circle,
3. $\Phi(z)$ is positive semidefinite on the unit circle $\Phi(e^{j\theta}) \geq 0$.

When $\Phi(z)$ satisfies the first two conditions, but it is not necessarily positive on the unit circle, we shall simply say that it is a *parahermitian function.* Every rational parahermitian function has a Laurent expansion in a neighborhood of the unit circle,

$$\Phi(z) = \sum_{\tau=-\infty}^{+\infty} \Lambda(\tau) z^{-\tau}$$

where Λ is an $m \times m$ summable matrix function satisfying the symmetry relation $\Lambda(-\tau) = \Lambda(\tau)^\top$. For a causal realization ($|\lambda(A)| < 1$) one obtains the well-known formulas

$$\Lambda(\tau) = \begin{cases} CA^{\tau-1}\,\bar{C}^\top & \tau > 0 \\ C\Sigma C^\top + DD^\top & \tau = 0 \end{cases}$$

where Σ is the state covariance matrix satisfying the *Discrete-time Lyapunov Equation*

$$\Sigma = A\Sigma A^\top + BB^\top \tag{DLE}$$

and $\bar{C}^\top := A\Sigma C^\top + BD^\top$. Denoting for typographical reasons, $\Lambda(0)$ as Λ_0, we obtain the decomposition

$$\begin{aligned} \Phi(z) &= \left[C(zI - A)^{-1}\bar{C}^\top + \Lambda_0/2\right] + \left[\Lambda_0/2 + \bar{C}(z^{-1}I - A^\top)^{-1}C^\top\right] \\ &=: \qquad\qquad \Phi_+(z) \qquad\qquad + \qquad\qquad \Phi_+(z^{-1})^\top. \end{aligned} \tag{2}$$

which is an analytic (on $\{|z| \geq 1\}$) + co-analytic; i.e. causal + anticausal decomposition. When in addition, $\Phi(e^{j\theta}) \geq 0$ the analytic component $\Phi_+(z)$, is called the *Positive Real part* of $\Phi(z)$, see e.g. [1] for the definition and implications of this property.

For general A the formulas above do not hold. In particular the state covariance does not satisfy the Lyapunov equation (DLE). One can however have many other additive decompositions of a parahermitian function $\Phi(z)$. Note in fact that the poles of $\Phi(z)$ have reciprocal symmetry; i.e. if $\Phi(z)$ has a pole in $z = p_k$ then $1/p_k$ must also be a pole (be it finite or not) of the same multiplicity. Hence the set of poles, $\sigma(\Phi)$, of a $\Phi(z)$ of degree $2n$ can be split in two reciprocal subsets σ_1 and σ_2 each containing n complex numbers (repeated according to multiplicity), such that $\sigma_2 = 1/\sigma_1$. In general, this decomposition of the spectrum yields, by partial fraction expansion, a rational additive decomposition of $\Phi(z)$ of the type

$$\Phi(z) = Z(z) + Z(z^{-1})^\top, \tag{3}$$

where $Z(z)$ is a rational function which we shall still write as $C(zI-A)^{-1}\bar{C}^{\top}+\Lambda_0/2$, with poles in σ_1 while those of $Z(z^{-1})^{\top}$ are necessarily in $\sigma_2 = 1/\sigma_1$. Naturally, here A need not be asymptotically stable.

Recall that an $n\times n$ matrix A has *unmixed spectrum* if $\sigma(A)$, does not contain reciprocal pairs counting multiplicity. Assuming minimality of $(C,A,\bar{C}^{\top})$, then A has unmixed spectrum if and only if the selected pole set $\sigma_1 \equiv \sigma(A)$ has no self-reciprocal elements. It is obvious that this happens if and only if $\sigma_1 \cap \sigma_2 = \varnothing$.

Example 1. The parahermitian function

$$\Phi(z) = \frac{K^2}{(z-\frac{1}{2})^2(z^{-1}-\frac{1}{2})^2}$$

has $\sigma(\Phi) = \{\frac{1}{2},\frac{1}{2},2,2\}$ so the only non intersecting subsets of two elements are $\{\frac{1}{2},\frac{1}{2}\}$ and $\{2,2\}$. Hence in this case either $Z(z)$ is stable and coincides with $\Phi(z)_+$ or is totally antistable with poles in $\{2,2\}$.

Note that the unmixed spectrum condition is exactly the condition insuring that the Lyapunov equation $P-APA^{\top}=Q$ has a (unique) solution for arbitrary Q [10].

The rational spectral factorization problem is, given a parahermitian rational matrix $\Phi(z)$, find conditions for existence of rational matrix functions $W(z)$ such that $\Phi(z) = W(z)W(z^{-1})^{\top}$. In the classical setting one starts from the additive decomposition $\Phi(z) = \Phi_+(z)+\Phi_+(z^{-1})^{\top}$ with $\Phi_+(z)$ positive real and looks for *analytic* spectral factors. We want to generalize this problem. Start from a parahermitian matrix given by a general additive decomposition (3), where the function $Z(z) := C(zI-A)^{-1}\bar{C}^{\top} + \frac{1}{2}\Lambda_0$ is not necessarily positive real and look for rational spectral factors *with the same poles of* $Z(z)$. Note that the existence of spectral factors is not automatically guaranteed and is in fact *equivalent to the positivity of* $\Phi(e^{j\theta})$ since, irrespective of analiticity, if a spectral factor W exists, then $\Phi(e^{j\theta}) = W(e^{j\theta})W(e^{-j\theta})^{\top} \geq 0$.

Now it is obvious that the additive decomposition (3) can formally be rewritten as

$$\Phi(z) = \begin{bmatrix} C(zI-A)^{-1} & I \end{bmatrix} \begin{bmatrix} 0 & \bar{C}^{\top} \\ \bar{C} & \Lambda_0 \end{bmatrix} \begin{bmatrix} (z^{-1}I-A^{\top})^{-1}C^{\top} \\ I \end{bmatrix} \tag{4}$$

but, by a well-known identity, described in the following lemma, the constant matrix in the middle can be modified.

Lemma 2. *Pick any $n\times n$ symmetric matrix P and form the array*

$$N(P) = \begin{bmatrix} P-APA^{\top} & -APC^{\top} \\ -CPA^{\top} & -CPC^{\top} \end{bmatrix},$$

then we have

$$\begin{bmatrix} C(zI-A)^{-1} & I \end{bmatrix} N(P) \begin{bmatrix} (z^{-1}I-A^{\top})^{-1}C^{\top} \\ I \end{bmatrix} \equiv 0$$

identically.

Hence we can rewrite (4) as

$$\Phi(z) = \begin{bmatrix} C(zI-A)^{-1} & I \end{bmatrix} \begin{bmatrix} P-APA^{\top} & \bar{C}^{\top}-APC^{\top} \\ \bar{C}-CPA^{\top} & \Lambda_0-CPC^{\top} \end{bmatrix} \begin{bmatrix} (z^{-1}I-A^{\top})^{-1}C^{\top} \\ I \end{bmatrix} \tag{5}$$

from which the following sufficient condition for spectral factorization of $\Phi(z)$ follows.

Lemma 3. *If there exists $P = P^{\top}$ solving the following* Linear Matrix Inequality

$$M(P) := \begin{bmatrix} P-APA^{\top} & \bar{C}^{\top}-APC^{\top} \\ \bar{C}-CPA^{\top} & \Lambda_0-CPC^{\top} \end{bmatrix} \geq 0 \tag{LMI}$$

then the parahermitian matrix $\Phi(z)$ admits spectral factors. In fact, let B,D be defined by the factorization

$$M(P) = \begin{bmatrix} B \\ D \end{bmatrix} \begin{bmatrix} B^{\top} & D^{\top} \end{bmatrix}, \tag{6}$$

then $W(z) = C(zI-A)^{-1}B + D$ is a spectral factor.

Hence if the inequality (LMI) has a symmetric solution, $\Phi(z)$ is actually a spectral density. Note that no stability of A nor minimality are required.

Proposition 4. *If A is unmixing and $C(zI-A)^{-1}\bar{C}^{\top}$ is a minimal realization then $C(zI-A)^{-1}B$ is also a minimal realization.*

Proof. (sketch) The proof can be based on the fact that $\sigma_1 \cap \sigma_2 = \varnothing$ implies that the McMillan degree $\delta(Z) = \frac{1}{2}\delta(\Phi) = n$ so that the dimension of A is $n \times n$ which implies that (A,B) must be reachable otherwise the dimension of a minimal realization of W would be smaller than n and hence $\Phi(z) = W(z)W(z^{-1})^{\top}$ would have McMillan degree smaller than $2n$ which is in contrast with the minimality of the realization of $Z(z)$. □

Note that if $\sigma_1 \equiv \sigma(A)$ has self-reciprocal elements there may be ambiguities in forming the decomposition $\Phi(z) = Z(z) + Z(z^{-1})^{\top}$. Actually in some case the decomposition may not even exist.

3 The linear matrix inequality: Necessity

Conversely, we want to show that for any spectral factor there is a $P = P^{\top}$ satisfying the LMI constructed with the parameters $(A, C, \bar{C}, \Lambda_0)$ of some $Z(z)$.

Theorem 5. *Let $W(z) = C(zI-A)^{-1}B + D$ be a rational spectral factor of $\Phi(z)$ with an unmixing A matrix. Then there is a corresponding additive decomposition $\Phi(z) = Z(z) + Z(z^{-1})^{\top}$ with $= C(zI-A)^{-1}\tilde{C}^{\top} + \frac{1}{2}\Lambda_0$ and a unique $P = P^{\top}$ satisfying the linear matrix inequality*

$$\begin{bmatrix} P-APA^{\top} & \tilde{C}^{\top}-APC^{\top} \\ \tilde{C}-CPA^{\top} & \Lambda_0-CPC^{\top} \end{bmatrix} \geq 0.$$

If $|\lambda(A)| < 1$ and (A,B) is reachable, then $P > 0$.

Proof. For analytic spectral factors ($|\lambda(A)| < 1$) this result is well-known and in fact nearly obvious. A short probabilistic proof goes as follows. Take a stationary realization (1) with transfer function $W(z)$, define $\mathbf{z}(t) := \begin{bmatrix}\mathbf{x}(t+1)^\top & \mathbf{y}(t)^\top\end{bmatrix}^\top$ and compute the variance of

$$\begin{bmatrix} B \\ D \end{bmatrix} \mathbf{w}(t) = \begin{bmatrix} \mathbf{x}(t+1) \\ \mathbf{y}(t) \end{bmatrix} - \begin{bmatrix} A \\ C \end{bmatrix} \mathbf{x}(t) = \mathbf{z}(t) - \hat{\mathbb{E}}\left[\mathbf{z}(t) \mid \mathbf{x}(t)\right]$$

and notice that the variance of the quantity on the left must obviously be positive semidefinite. By Pithagora's theorem the variance of the second member is

$$\begin{bmatrix} P & \bar{C}^\top \\ \bar{C} & \Lambda_0 \end{bmatrix} - \begin{bmatrix} APA^\top & APC^\top \\ CPA^\top & CPC^\top \end{bmatrix}$$

since

$$\mathbb{E}\,\mathbf{y}(t)\mathbf{x}(t+1)^\top = \mathbb{E}\,(C\mathbf{x}(t) + D\mathbf{w}(t))(\mathbf{x}(t)^\top A^\top + \mathbf{w}(t)^\top B^\top) = CPA^\top + DB^\top = \bar{C}$$

(this follows since the model is causal). In this case the solution P of the LMI is the variance of $\mathbf{x}(t)$.

For arbitrary A we can give an algebraic proof as follows. Note first that, whenever $W(z) = C(zI - A)^{-1}B + D$ is a spectral factor then we can write

$$\Phi(z) = \begin{bmatrix} C(zI-A)^{-1} & I \end{bmatrix} \begin{bmatrix} BB^\top & BD^\top \\ DB^\top & DD^\top \end{bmatrix} \begin{bmatrix} (z^{-1}I - A^\top)^{-1}C^\top \\ I \end{bmatrix} \tag{7}$$

The LMI is defined once we show that this $\Phi(z)$ has a parahermitian additive decomposition like (3), or equivalently (4). Now in force of Lemma 2, this amounts to showing that there are matrices $P, \tilde{C}$ and R such that

$$\begin{bmatrix} BB^\top & BD^\top \\ DB^\top & DD^\top \end{bmatrix} = \begin{bmatrix} 0 & \tilde{C}^\top \\ \tilde{C} & R \end{bmatrix} + \begin{bmatrix} P - APA^\top & -APC^\top \\ -CPA^\top & -CPC^\top \end{bmatrix}, \tag{8}$$

since by multiplying this equation on the left by $[C(zI-A)^{-1} \;\; I]$ and on the right by $[C(Iz^{-1} - A)^{-1} \;\; I]^\top$ and by using Lemma 2 we would conclude that $\Phi(z)$ has a parahermitian additive decomposition of the same form as (4) with $Z(z) = C(zI - A)^{-1}\tilde{C} + \frac{1}{2}R$. By the unmixing assumption the Lyapunov equation $P - APA^\top = BB^\top$ has a unique symmetric solution P. Therefore solving equation (8) with this fixed P yields

$$\begin{aligned} \tilde{C} &= CPA^\top + DB^\top \\ R &= CPC^\top + DD^\top \end{aligned}$$

whereby $\Phi(z) = C(zI-A)^{-1}\tilde{C}^\top + R + \tilde{C}(z^{-1}I - A^\top)^{-1}C^\top := \tilde{Z}(z) + R + \tilde{Z}(z^{-1})^\top$. Now by the unmixing assumption $\Phi(z)$ is analytic in some neighborhood of the unit circle and there admits a unique Laurent expansion. In a suitably small neighborhood of the unit circle $\tilde{Z}(z)$ has an expansion in powers of z^{-1} without constant term while $\tilde{Z}(z^{-1})^\top$ has an expansion in positive powers of z also without constant term. It follows that $R = \Lambda_0$ and there is a unique $P = P^\top$ satisfying the LMI corresponding to $Z(z) = C(zI-A)^{-1}\tilde{C} + \frac{1}{2}\Lambda_0$. □

When A is assumed to be asymptotically stable one gets the celebrated *Positive Real Lemma* attributed to Kalman [5], Yakubovich [11] and Popov [8].

Theorem 5 and Lemma 3 together provide a necessary and sufficient condition for spectral factorization and provide, at least in principle a way to compute spectral factors. Note that the matrix $\tilde{C} = CPA^\top + DB^\top$ must be the same for all minimal spectral factors $W(z) = C(zI - A)^{-1}B + D$ irrespective of which P is selected, since it is the "B" parameter of a minimal realization of $Z(z)$; hence it cannot depend on which spectral factor is chosen to form $\Phi(z)$. In other words, $\tilde{C}$ is an *invariant over the family of all minimal spectral factors expressed with a fixed* (C, A) *pair.* Recall that for a stable A we have $\tilde{C} = \bar{C} = \mathbb{E}\mathbf{y}(t)\mathbf{x}(t+1)^\top$ which is clearly an invariant quantity.

4 The algebraic Riccati equation

It is not difficult to see that the number of columns of the spectral factor $W(z)$ varies with $P \in \mathcal{P}$. In fact, assuming that $\begin{bmatrix} B^\top & D^\top \end{bmatrix}^\top$ in the factorization (6) is always taken of full column rank p, the corresponding $W(z)$ is of dimension $m \times p$, where $p := \operatorname{rank} M(P)$. The *rank minimizing* solutions form a subset $\mathcal{P}_0 \subset \mathcal{P}$ which is characterized as follows.

Proposition 6 (Theorem 4.1 in [3]). *The $m \times m$ matrix $\Delta(P) := \Lambda_0 - CPC^\top$ is non singular for all $P \in \mathcal{P}$ if and only if*

- *the normal rank of $\Phi(z)$ is full (equal to m),*
- $\Phi(z)$ *does not have zeros at $z = 0$ or at $z = \infty$.*

If and only if these conditions hold the minimum rank of $M(P)$; $P \in \mathcal{P}$ is equal to m.

The first condition is simply that the underlying process should be of full rank. The second condition means that both limits

$$\lim_{z \to 0} \Phi(z)^{-1}, \qquad \text{and} \qquad \lim_{z \to \infty} \Phi(z)^{-1}$$

exist and are non-singular (in fact if one limit exists so does the other). If this is the case the problem is called *regular.*

We shall henceforth assume regularity. Then, letting $T := -(\bar{C}^\top - APC^\top)\Delta(P)^{-1}$, one has a block-diagonalization of $M(P)$

$$\begin{bmatrix} I & T \\ 0 & I \end{bmatrix} M(P) \begin{bmatrix} I & 0 \\ T^\top & I \end{bmatrix} = \begin{bmatrix} R(P) & 0 \\ 0 & \Delta(P) \end{bmatrix},$$

where

$$R(P) = P - APA^\top - (\bar{C}^\top - APC^\top)\Delta(P)^{-1}(\bar{C}^\top - APC^\top)^\top.$$

Hence, $P \in \mathcal{P}$ if and only if it satisfies the *Algebraic Riccati Inequality*

$$R(P) \geq 0. \qquad \text{(ARI)}$$

Moreover, $p = \operatorname{rank} M(P) = m + \operatorname{rank} R(P) \geq m$. If P satisfies the *Algebraic Riccati Equation* $R(P) = 0$, i.e.

$$P = APA^\top + (\bar{C}^\top - APC^\top)\Delta(P)^{-1}(\bar{C}^\top - APC^\top)^\top, \tag{ARE}$$

then $\operatorname{rank} M(P) = m$ and the corresponding spectral factor $W(z)$ is square $m \times m$. The family of P's solving the ARE; i.e. corresponding to square spectral factors, form the subfamily $\mathcal{P}_0$ of $\mathcal{P}$. If $P \notin \mathcal{P}_0$, $W(z)$ is rectangular.

5 The structure of rational all-pass functions

A cute application of Theorem 5 and Lemma 3 is to rational spectral factors of the spectral density $\Phi(z) \equiv I$; i.e. to square rational matrix functions $Q(z) = C(zI - A)^{-1}B + D$ such that $Q(z)Q(z^{-1})^\top = I$. These are called (rational) *all-pass* functions. In this way we shall slightly generalize representation results in the literature, say [6] and in particular of Fuhrmann and Hoffmann [4], which were obtained for analytic all-pass; i.e. inner, functions. Fix a realization of $Z(z)$ in an additive decomposition of $\Phi(z)$ with (A, C) an observable pair so that the pole structure of $Q(z)$ is fixed. Then, since the McMillan degree of $Z(z)$ is zero, we must have $\bar{C} = 0$. Hence we shall have to look for the rank minimizing solutions of the spectral factorization LMI

$$\begin{bmatrix} P - APA^\top & -APC^\top \\ -CPA^\top & I - CPC^\top \end{bmatrix} \geq 0$$

where, in virtue of Proposition 6, $I - CPC^\top := DD^\top$ is non singular. Taking, without loss of generaity, full-rank solutions of the factorization equation (6), this permits to solve for B to get $B = -APC^\top D^{-\top}$ and leads to the *Homogeneous Algebraic Riccati Equation*

$$P - A\left[P + PC^\top(I - CPC^\top)^{-1}CP\right]A^\top = 0, \tag{HARE}$$

whose solutions parametrize in 1:1 way the square all-pass functions with the given denominator. Note that to the trivial solution $P = 0$, corresponds $Q(z) = D$, a constant $m \times m$ unitary matrix. The other solutions of the HARE parametrize the non trivial all pass functions with the given denominator.

We shall provisionally assume that A is invertible. Consider the zero-dynamics matrix

$$\Gamma := A - BD^{-1}C = A + APC^\top(I - CPC^\top)^{-1}C$$

using the Riccati equation one derives the invariance relation $\Gamma P = PA^{-\top}$ so that, if P is an invertible solution

$$P^{-1}\Gamma P = A^{-\top}. \tag{9}$$

So far we don't know if there are any invertible solutions of the homogeneous Riccati equation (HARE). By the matrix inversion lemma, they must satisfy

$$P - A\left[P^{-1} - C^\top C\right]^{-1}A^\top = 0$$

which, since A is invertible, turns into $P^{-1} = A^{-\top}P^{-1}A^{-1} - A^{-\top}C^{\top}CA^{-1}$ that is into the Lyapunov equation

$$P^{-1} = A^{\top}P^{-1}A + C^{\top}C, \tag{10}$$

which by observability and unmixing has a unique nonsingular solution [10]. In case of A asymptotically stable, P would be positive definite.

All other solutions of the homogeneous Riccati equation must be singular.

Theorem 7. *Let A be unmixing and nonsingular and (C,A) be observable. There is a 1:1 correspondence between square all pass rational matrix functions of the form $Q(z) = C(zI - A)^{-1}B + D$, defined modulo multiplication from the right by an arbitrary constant unitary matrix, and solutions $P = P^{\top}$ of the homogeneous Riccati equation* (HARE). *Consider the orthogonal direct sum decomposition*

$$\mathbb{R}^n = \operatorname{Im}P \oplus \operatorname{Ker}P, \tag{11}$$

then $\operatorname{Im}P$ *is an invariant subspace for* Γ *and* $\operatorname{Ker}P$ *is a left-invariant subspace for A which is orthogonal to the reachable subspace of (A,B). The McMillan degree of $Q(z)$ is then equal to* $\dim\{\operatorname{Im}P\}$. *In a basis adapted to the direct sum decomposition* (11), *$P = \operatorname{diag}\{\hat{P}_1,\, 0\}$ and the restrictions $\hat{A}_{11}^{-\top}$, $\hat{\Gamma}_{11}$ of $A^{-\top}$ and of Γ to* $\operatorname{Im}P$ *are similar; i.e.*

$$\hat{P}_1^{-1}\hat{\Gamma}_{11}\hat{P}_1 = \hat{A}_{11}^{-\top}.$$

Proof. The first statement is just a particularization of Theorem 5. The orthogonal direct sum decomposition (11) holds since P is symmetric. From the invariance relation $\Gamma P = PA^{-\top}$, it follows that for any $v \in \mathbb{R}^n$, $\Gamma Pv \in \operatorname{Im}P$ and hence $\operatorname{Im}P$ is invariant for Γ. Next, for any $x \in \operatorname{Ker}P$ we have $PA^{-\top}x = \Gamma Px = 0$ and hence $\operatorname{Ker}P$ is an invariant subspace for $A^{-\top}$. In fact $\operatorname{Ker}P$ is orthogonal to the reachable subspace for (A,B) as $x^{\top}B = -x^{\top}APC^{\top}D^{-\top} = 0$ since $\operatorname{Ker}P$ is also an invariant subspace for $A^{\top}$ and hence $x^{\top}A$ belongs to the left nullspace of P. Since P is symmetric there is an orthogonal basis of eigenvectors in which $P = \operatorname{diag}\{\hat{P}_1,\, 0\}$ with $\hat{P}_1$ non singular and the invariance relation can be written

$$\begin{bmatrix} \hat{P}_1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \hat{A}_{11}^{-\top} & \hat{A}_{12}^{-\top} \\ \hat{A}_{21}^{-\top} & \hat{A}_{22}^{-\top} \end{bmatrix} = \begin{bmatrix} \hat{\Gamma}_{11} & \hat{\Gamma}_{12} \\ \hat{\Gamma}_{21} & \hat{\Gamma}_{22} \end{bmatrix}\begin{bmatrix} \hat{P}_1 & 0 \\ 0 & 0 \end{bmatrix}$$

from which the similarity of $\hat{A}_{11}^{-\top}$ to $\hat{\Gamma}_{11}$ follows. In this basis $Q(z)$ has a realization $(\hat{C}_1, \hat{A}_{11}, \hat{B}_1, D)$ of dimension equal to $\dim\{\operatorname{Im}P\}$. Since $\hat{P}_1$ is non singular and satisfies the Lyapunov equation $\hat{P}_1 = \hat{A}_{11}\hat{P}_1\hat{A}_{11}^{\top} + \hat{B}_1\hat{B}_1^{\top}$ this realization must be reachable. □

We now analyze the case when A is singular[1] . Let $A = \begin{bmatrix} N & 0 \\ 0 & A_0 \end{bmatrix}$ where N is nilpotent and A_0 invertible. Since the HARE can be written as $P = \Gamma(P)PA^{\top}$, iterating we get,

[1] The following argument is taken from unpublished joint work with A. Ferrante [2].

$P = \Gamma(P)^k P(A^\top)^k$ for any integer $k \geq 0$, so that any solution of the HARE must have the form

$$P = \begin{bmatrix} 0 & 0 \\ 0 & P_0 \end{bmatrix}$$

where P_0 satisfies the reduced order HARE

$$P_0 - A_0 \left[P_0 + P_0 C_0^\top (I - C_0 P_0 C_0^\top)^{-1} C_0 P_0 \right] A_0^\top = 0, \tag{12}$$

with an obvious definition of C_0. In this way we have reduced the problem to one with a nonsingular A in a smaller dimensional space of dimension say n_0. In particular, (11) now becomes $\mathbb{R}^{n_0} = \operatorname{Im} P_0 \oplus \operatorname{Ker} P_0$ where $\operatorname{Im} P_0$ is an invariant subspace for Γ_0 and $\operatorname{Ker} P_0$ is a left-invariant subspace for A_0. All statements of Theorem 7 remain true in this reduced dimensional context.

Remark 8. The structure of all pass functions of full McMillan degree n can be obtained by a simple similarity argument based on the fact that $Q(z)^{-1} = Q(z^{-1})^\top$.

Bibliography

[1] B. D. O. Anderson and S. Vongpanitlerd. *Network Analysis and Synthesis*. Prentice-Hall, 1973. Cited p. 302.

[2] A. Ferrante and G. Picci. Unstable spectral factorization and related questions. Technical report, Department of Information Engineering, University of Padova, 2012. Cited p. 308.

[3] A. Ferrante, G. Picci, and S. Pinzoni. Silverman algorithm and the structure of discrete time stochastic systems. *Linear Algebra and its Applications*, 351–352:219–242, 2002. Cited p. 306.

[4] P. A. Fuhrmann and J. Hoffmann. Factorization theory for stable, discrete-time inner functions. *Journal of Mathematical Systems, Estimation, and Control*, 7(4):383–400, 1997. Cited p. 307.

[5] R. E. Kalman. Lyapunov functions for the problem of Luré in automatic control. *Proc. Nat. Acad. Sci. U.S.A.*, 49:201–205, 1963. Cited p. 306.

[6] G. Michaletzky. A note on the factorization of discrete-time inner functions. *Journal of Mathematical Systems, Estimation, and Control*, 8(4):479–482, 1998. Cited p. 307.

[7] G. Picci and S. Pinzoni. Acausal models and balanced realizations of stationary processes. *Linear Algebra and its Applications*, 205–206:957–1003, 1994. Cited p. 301.

[8] V. M. Popov. Hyperstability and optimality of automatic systems with several control functions. *Rev. Roumaine Sci. Tech. Sér. Électrotech. Énergét.*, 9:629–690, 1964. Cited p. 306.

[9] Y. Rozanov. *Stationary Random Processes*. Holden Days, 1963. Cited p. 301.

[10] H. K. Wimmer and A. D. Ziebur. Remarks on inertia theorems for matrices. *Checoslovak Mathematical Journal*, 25:556–561, 1975. Cited pp. 303 and 308.

[11] V. A. Yakubovich. The solution of some matrix inequalities encountered in automatic control theory. *Dokl. Akad. Nauk SSSR*, 143:1304–1307, 1962. Cited p. 306.

On the algebraic classification of bimodal piecewise linear systems

Xavier Puerta
Universitat Politecnica de Catalunya
Barcelona, Spain
francisco.javier.puerta@upc.edu

Abstract. Given a bimodal piecewise linear system, we study the equivalence relation given by simultaneous similarity and feedback of the linear systems of the plant. We obtain a reduced form of its equations and, in some particular cases, a complete system of invariants.

1 Introduction

A *bimodal piecewise linear system* is a control system defined by two linear systems, each operating in one of two regions of $\mathbb{R}^n$ that are separated by a hyperplane. Some canonical forms of this kind of systems with respect to changes of variables are obtained, for example, in [2] and [3], in the case where the two linear systems coincide on the hyperplane of separation. Here, as a continuation of [4], we consider also state feedbacks in the corresponding equivalence relation. As an application, following [1], we obtain a characterization of controllability in terms of the invariants of the system.

From the mathematical point of view, this contribution consists of the study of the orbits of a space of matrices with respect to a particular Lie group acting on it. I would like to recognize here how much I have learned on this subject by working together with Uwe Helmke. And I express my best wishes to him for his 60th birthday!

2 Algebraic objects associated to bimodal piecewise linear systems

A bimodal piecewise linear system is defined by equations of the form

$$\begin{aligned} \dot{x}(t) &= A_1x(t) + B_1u(t) \quad \text{if } c^\top x(t) \le 0, \\ \dot{x}(t) &= A_2x(t) + B_2u(t) \quad \text{if } c^\top x(t) > 0, \end{aligned} \tag{1}$$

where A_i and B_i, $i = 1,2$, are $n \times n$ and $n \times m$, $m \le n$, real matrices, respectively. We can assume $B_1 = B_2 = B$ by admitting impulsive inputs. In fact, we can extend the state vector with the input vector $u(t)$ and add the equation $\dot{u}(t) = v(t)$, where $v(t)$ is the new input vector. There is no loss of generality assuming $\operatorname{rank} B = m$.

On the other hand, some restrictions must be imposed on the matrices A_1 and A_2 in order that the above equations define a dynamical system. In fact, if $c^\top x(t) = 0$, then $c^\top A_1 x(t)$ and $c^\top A_2 x(t)$ must have the same sign. That is to say, $(c^\top A_1 x)(c^\top A_2 x) \ge 0$ if $x \in \operatorname{Ker} c^\top$, or equivalently, the quadratic form defined by the matrix

$$Q := A_1^\top cc^\top A_2 + A_2^\top cc^\top A_1$$

is non-negative definite on $\operatorname{Ker} c^\top$.

We remark that $\operatorname{rank} Q \leq 2$ and that the cone defined by $x^{\mathsf{T}} Q x = 0$ is the union of $\operatorname{Ker} c^{\mathsf{T}} A_1$ and $\operatorname{Ker} c^{\mathsf{T}} A_2$. Therefore, the above condition implies that $\operatorname{Ker} Q \subset \operatorname{Ker} c^{\mathsf{T}}$. Let $\mathcal{V} := \operatorname{Ker} Q = \operatorname{Ker} c^{\mathsf{T}} A_1 \cap \operatorname{Ker} c^{\mathsf{T}} A_2$ and $\mathcal{U} := \operatorname{Ker} c^{\mathsf{T}}$. If we discard bad conditioned systems in a neighborhood of $\mathcal{U}$, we have that $A_1|_{\mathcal{V}} = A_2|_{\mathcal{V}}$.

The system (1) is determined by the quadruple of matrices (A_1, A_2, B, c). Let $\sum$ be the set of this kind of quadruples, subjected to the above restrictions. We introduce in $\sum$ the following equivalence relation

$$(A_1, A_2, B, c) \sim (A_1', A_2', B', c')$$

if

$$c' = cS \quad \text{and} \quad \begin{bmatrix} A_i' & B' \end{bmatrix} = S^{-1} \begin{bmatrix} A_i & B \end{bmatrix} \begin{bmatrix} S & 0 \\ R & T \end{bmatrix} \tag{2}$$

for $i = 1, 2$, where $S \in \mathrm{Gl}(n)$, $T \in \mathrm{Gl}(m)$ and R a $m \times n$ matrix. This equivalence relation corresponds to simultaneous feedback equivalence of the two linear systems appearing in (1).

First of all, we remark that (2) implies that, with the above notations, $\mathcal{U}' = S(\mathcal{U})$ and $\mathcal{V}' = S(\mathcal{V})$. Let $\mathcal{V}$ be a fixed subspace of $\mathbb{R}^n$ of codimension $\nu \leq 2$. We consider the subgroup of the feedback group acting in $\sum$ as shown in (2) formed by the triples (S, R, T) with $S(\mathcal{V}) = \mathcal{V}$.

Our goal is to apply the Kronecker theory of pencils to the above equivalence relation. In order to do this, we associate to the quadruple (A_1, A_2, B, c) the following pair of linear maps

$$f, \sigma : \mathcal{V} \times \mathbb{R}^m \to \mathbb{R}^n = \mathcal{V} \times \mathbb{R}^\nu$$

defined by $\sigma(x, y) = x$ and $f(x, y) = A_i x + By$, $i = 1$ or 2.

3 Reduced form

If we have two equivalent quadruples $(A_1, A_2, B, c) \sim (A_1', A_2', B', c')$, it is clear that the corresponding pairs (σ, f), (σ, f') according to the above definition, are equivalent in the Kronecker sense. That is to say, there exist automorphisms ψ and φ of $\mathcal{V} \times \mathbb{R}^m$ and $\mathbb{R}^n$, respectively such that $\varphi\sigma = \sigma\psi$ and $\varphi f' = f\psi$. In fact, we can define $\varphi(x) := Sx$ and $\psi(x, y) = (Sx, Rx + Ty)$. Notice that $\varphi\sigma = \sigma\psi$ implies that $\psi(\mathcal{V}) = \mathcal{V}$. The theory of Kronecker gives a canonical form for the matrix representation of σ and f. More precisely, there exist bases of $\mathcal{V} \times \mathbb{R}^m$ and $\mathbb{R}^n$ (we call them *Kronecker bases*) in such a way that the matrix representations of σ and f with respect to these bases are

$$\begin{bmatrix} I_{n-\nu} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} F & H \\ G & E \end{bmatrix},$$

respectively, where $F = \operatorname{diag}\{F_J, N_o, N_c, N_b\}$, $E = \operatorname{diag}\{0, E_b\}$,

$$G = \begin{bmatrix} 0 & G_o & 0 & 0 \\ 0 & 0 & 0 & G_b \end{bmatrix} \in \mathbb{R}^{(n-\nu)\times\nu}, \quad H = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ H_c & 0 \\ 0 & H_b \end{bmatrix} \in \mathbb{R}^{\nu\times m},$$

and F_J is a $h \times h$ Jordan matrix, that is to say, $F_J = \operatorname{diag}(N_{h_1} - \lambda_1 I, \ldots, N_{h_s} - \lambda_s I)$, $h = \sum_{i=1}^{s} h_i$. We order the Jordan blocks in such a way that if $i < j$ and $\lambda_i = \lambda_j$ then $h_i \geq h_j$.

For the rest of the blocks, $N_b = \operatorname{diag}\{N_{l_1},\ldots,N_{l_t}\}$, $H_b = [\operatorname{diag}\{H_{l_1},\ldots,H_{l_t}\},0]$, $G_b = [\operatorname{diag}\{G_{l_1},\ldots,G_{l_t}\}^{\top},0]^{\top}$, $E_b = \operatorname{diag}\{0,I_{l_0}\} \in \mathbb{R}^{l\times l}$, $l = \sum_{i=0}^{t} l_i$, $N_c = \operatorname{diag}\{N_{k_1},\ldots,N_{k_r}\}$, $H_c = \operatorname{diag}\{H_{k_1},\ldots,H_{k_r}\}$, $k = \sum_{i=1}^{r} k_i$, $N_o = \operatorname{diag}\{N_{d_1},\ldots,N_{d_s}\}$, $G_o = \operatorname{diag}\{G_{d_1},\ldots,G_{d_s}\}$, $d = \sum_{i=1}^{s} d_i$, and

$$N_i = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{i\times i}, \quad H_i = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \in \mathbb{R}^{i\times 1}$$

and

$$G_i = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{1\times i}.$$

We call the invariants of the above form, the Kronecker invariants of the quadruple. In [5], for example, one can find a constructive method for obtaining the Kronecker bases as well as the Kronecker invariants. In this reference it is shown, moreover, that the Kronecker bases depend continuously on the pair (σ, f) if the Kronecker invariants are constant. The following theorem is an application of the above results.

Theorem 1. *Let $(A_1(u), A_2(u), B(u), c(u))$ be a family of quadruples defined on an open contractible set $\mathcal{U} \subset \mathbb{R}^{\alpha}$ with the same Kronecker invariants (in particular, this includes a constant system). Then, there exist continuous (resp. smooth) matrix families $S(u)$, $T(u)$ and $R(u)$ defined in $\mathcal{U}$, transforming the quadruples $(A_1(u), A_2(u), B(u), c(u))$, as in* (2) *giving*

$$\begin{aligned} c &\mapsto [0,\cdots,0,c_1(u),c_2(u)], \\ A_i &\mapsto \begin{bmatrix} F & X_i(u) \\ G & Y_i(u) \end{bmatrix}, i = 1,2, \\ B &\mapsto \begin{bmatrix} H \\ E \end{bmatrix}, \end{aligned} \tag{3}$$

where X_i, Y_i and c_i are continuous functions defined on $\mathcal{U}$. We call the transformed quadruple a reduced form.

Remark 2. We can replace the matrix F in (2) by $\operatorname{diag}\{F_J, N_o + \lambda I, N_c + \mu I, N_b + \gamma I\}$, with I the identity matrix of appropriate size, for any triple of scalars $\lambda, \mu, \gamma \in \mathbb{C}$. This follows taking into account that we can change the eigenvalues of N_c, N_o and N_b by a feedback and an output injection which render the block decomposition of F invariant, and from the way how the Kronecker bases are obtained. In the following, let us denote $X + \lambda I$ by X_{λ}, so the above matrix becomes $\operatorname{diag}\{F_J, N_{o\lambda}, N_{c\mu}, N_{b\gamma}\}$.

However, there is still room for making additional reductions on the matrices X_i and Y_i by means of the action of the feedback group. Let

$$X_i = \begin{bmatrix} X_{i11} & X_{i12} \\ X_{i21} & X_{i22} \\ X_{i31} & X_{i32} \\ X_{i41} & X_{i42} \end{bmatrix}, \quad Y_i = \begin{bmatrix} Y_{i11} & Y_{i12} \\ Y_{i21} & Y_{i22} \end{bmatrix}, \quad i = 1,2,$$

according to the Kronecker block partition of F. We prove the following theorem.

Theorem 3. *Any quadruple* $(A_1,A_2,B,c) \in \sum$ *is (generically) equivalent to the reduced form of Theorem* 1 *with*

$$X_{111} = 0,\ X_{131} = 0,\ X_{142} = 0 \text{ and } Y_{122} = 0.$$

Sketch of the proof. We begin with a simultaneous similarity transformation of A_i with a matrix of the form

$$S = \left[\begin{array}{cccc|cc} I_h & Q_1 & 0 & 0 & T_1 & 0 \\ 0 & I_d & 0 & 0 & 0 & 0 \\ 0 & Q_2 & I_k & 0 & T_2 & 0 \\ 0 & 0 & 0 & I_l & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & I_d & 0 \\ 0 & 0 & 0 & 0 & 0 & I_l \end{array}\right]$$

and choose λ and μ (cf. Remark 2) such that the Sylvester equations

$$F_J[Q_1 T_1] - [Q_1 T_1]\left[\begin{array}{cc} N_{o\lambda} & X_{121} \\ G_o & Y_{111} \end{array}\right] = [0 - X_{111}] \quad \text{and}$$

$$N_{c\mu}[Q_2 T_2] - [Q_2 T_2]\left[\begin{array}{cc} N_{o\lambda} & X_{121} \\ G_o & Y_{111} \end{array}\right] = [0 - X_{131}]$$

both have solutions. Then, one can check that if we take Q_1, Q_2, T_1 and T_3 as the (unique) solutions of the above equations, $[S^{-1}A_iS, S^{-1}B] = [A_i', B]$ have a Kronecker form with $X_{111}' = 0$, $X_{131}' = 0$. Notice that $S(\mathcal{V}) = \mathcal{V}$.

Next, we consider the following sub-matrices of $[A_i', B]$:

$$A_{ib} = \left[\begin{array}{cc} N_b & X_{i42} \\ G_b & Y_{i22} \end{array}\right], \quad B_b = \left[\begin{array}{c} H_b \\ E_b \end{array}\right], \quad i = 1,2.$$

The rest of the zero blocks of the reduced form of the theorem are obtained thanks to the following lemma. □

Lemma 4. *Given the quadruple* (A_1,A_2,B,c), *there exists an invertible matrix* T *with* $\mathrm{Ker}[0\ I_{d+l}] = \mathrm{Ker}[0\ I_{d+l}]T$ *and a feedback matrix* F, *such that* $T^{-1}B_b = B_bU$ *for a certain invertible matrix* U *and*

$$T^{-1}A_{ib}T + B_bF = \left[\begin{array}{cc} N_{b,\gamma} & X_{i42}' \\ G_b & Y_{i22}' \end{array}\right] \tag{4}$$

with $X_{142}' = 0$ *and* $Y_{122}' = \gamma I_{d+l}$.

Proof. We illustrate the proof with an example and leave the details for the general case to the reader. Let

$$[A_{1b} \| B_{1b}] = \left[\begin{array}{ccc|cc|cc||cc} 0 & 0 & 0 & 0 & 0 & a_1 & b_1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & a_2 & b_2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & a_3 & b_3 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & a_4 & b_4 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & a_5 & b_5 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & a_6 & b_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & a_7 & b_7 & 0 & 0 \end{array}\right]$$

and let

$$T=\left[\begin{array}{ccc|cc|cc} 1 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ 0 & 1 & s_1 & 0 & s_3 & s_2 & s_4 \\ 0 & 0 & 1 & 0 & 0 & s_1 & s_3 \\ \hline 0 & 0 & t_1 & 1 & t_2 & t_3 & t_4 \\ 0 & 0 & 0 & 0 & 1 & t_1 & t_2 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

It is clear that $\operatorname{Ker}[0\ I_{d+l}] = \operatorname{Ker}[0\ I_{d+l}]T$ and there exists U invertible such that $T^{-1}B_{1b} = B_{1b}U$. On the other hand, one can check that the equation

$$A_{1b}T = T\left[\begin{array}{ccc|cc|cc} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}\right] + B_{1b}\left[\begin{array}{ccc|cc|cc} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 \\ v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 \end{array}\right]$$

has the (unique!) solution $s_1 = -a_6$, $s_2 = -a_3$, $s_3 = -b_6$, $s_5 = -a_2$, $s_6 = -b_2$, $t_1 = -a_7$, $t_2 = -b_7$, $t_3 = -a_5$, $t_4 = -b_5$, $u_1 = -s_1$, $u_2 = -s_2$, $u_3 = -s_5$, $u_4 = -s_4$, $u_5 = -s_6$, $u_6 = a_1$, $u_7 = b_1$, $v_1 = 0$, $v_2 = -t_1$, $v_3 = -t_3$, $v_4 = -t_2$, $v_5 = -t_4$, $v_6 = a_4$, $v_7 = b_4$.

Generalizing the above equations for matrices of the form (3) (with an arbitrary number of nilpotent blocks of arbitrary sizes), one proves the lemma. □

4 Particular cases

Taking into account that $\nu = 1$ or 2, there are few possibilities for the reduced form of the last section. However, the obtention of a complete set of invariants is still a *wild* problem. We make the assumption that the transfer function of the system defined by the triples $(A_i, B, c^\top)$ is nontrivial. In [4], the case $\nu = 1$ is considered and, in the cases $m = 1$ and $m = 0$, we find a complete set of invariants. We recall that for $\nu = 1$ and $m = 1$, we can take $c = [0,\dots,0,1]$ and

$$[A_i, b] = \left[\begin{array}{cc|c} F_J & M_i & 0 \\ 0 & L_i & p \end{array}\right], \quad i = 1,2 \tag{5}$$

with

$$L_i = \begin{bmatrix} 0 & 0 & \dots & 0 & \sigma_{i0} \\ 1 & 0 & \dots & 0 & \sigma_{i1} \\ 0 & 1 & \dots & 0 & \sigma_{i2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & \sigma_{il} \end{bmatrix} \quad \text{and} \quad p = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and the M_i are $h \times (l+1)$ matrices with all the columns null except for the last one.

Let $\gamma = (\gamma_1,\dots,\gamma_s)$ be an s-tuple of integers, $0 \le \gamma_i \le h_i$. For $i = 1,\dots,s$, let d_i^γ be the γ_ith column vector of the canonical basis of $\mathbb{R}^{h_i}$ if $\gamma_i \ne 0$, or the null vector for $\gamma_i = 0$. We define

$$d(\gamma) = \begin{bmatrix} d_1^\gamma \\ \vdots \\ d_s^\gamma \end{bmatrix}.$$

We say that γ (and $d(\gamma)$) is *nice* (with respect to F_J) if for $i < j$, $\lambda_i = \lambda_j$ ($h_i \ge h_j$) and $\gamma_i = 1$, we have $\gamma_j = 0$.

Theorem 5. *Given a pair $[A_i,b]$ of the form* (5)*, there exists a nice s-tuple γ and a vector $g = (g_1,\dots,g_s)^\top$ with $g_i \in \mathbb{R}^{h_i}$ a vector whose entries are 0 except, possibly, for $h_i - \gamma_i + 1$ consecutive ones, $i = 1,\dots,s$, such that $[A_i,b]$ can be reduced in such a way that, with the above notations, $L_1 = N_l$, $M_1 = [0,\dots,0,g]$ and $M_2 = [0,\dots,0,g+d(\gamma)]$. Moreover, if the λ_i are distinct then, the above forms are canonical, that is to say, l, the s-tuples γ, g, the scalars $(\sigma_{2,0},\dots,\sigma_{2,l})$ and the Jordan invariants of F_J are a complete set of invariants of the pair.*

Since the controllability of the system does not depend on changes of variables and state feedbacks, one can characterize controllability in terms of the previous invariants. More precisely, one has the following characterization of controllability, obtained by translating the conditions of [1] in terms of the above invariants.

Theorem 6. *The system defined by the quadruple (A_1,A_2,b,c) is controllable if and only if its Kronecker form is of the type* (5)*, and its canonical form is such that $(F_J,[d(\gamma),g])$ is controllable and, if F_J has a real eigenvalue, the corresponding eigenvector, v_J satisfies $v_J^\top d(\gamma)(g+d(\gamma))^\top v > 0$.*

Likewise, for $\nu = 2$ and $m = 1$, one can take $c = [0,\dots,0,c_1,1]$ and M_i and L_i as in (5) with the last column a two column block vector.

Example 7. In $\mathbb{R}^3$, the possibilities for the pairs $[A_1,b],[A_2,b]$ are ($\star$ denotes a free parameter)

$$\left[\begin{array}{c|cc||c} 0 & \star & 0 & 1 \\ \hline 0 & \star & \star & 0 \\ 1 & \star & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc||c} 0 & \star & \star & 1 \\ \hline 0 & \star & \star & 0 \\ 1 & \star & \star & 0 \end{array}\right],$$

$$\left[\begin{array}{c|cc||c} 0 & 0 & \star & 1 \\ \hline 0 & \star & \star & 0 \\ 0 & \star & \star & 0 \end{array}\right], \left[\begin{array}{c|cc||c} 0 & \star & \star & 1 \\ \hline 0 & \star & \star & 0 \\ 0 & \star & \star & 0 \end{array}\right],$$

$$\left[\begin{array}{c|cc||c} \lambda & 0 & \star & 0 \\ \hline 0 & \star & \star & 0 \\ 0 & \star & 0 & 1 \end{array}\right], \left[\begin{array}{c|cc||c} \lambda & \star & \star & 0 \\ \hline 0 & \star & \star & 0 \\ 0 & \star & \star & 1 \end{array}\right],$$

$$\left[\begin{array}{c|cc||c} 0 & \star & \star & 0 \\ \hline 1 & \star & \star & 0 \\ 0 & \star & 0 & 1 \end{array}\right], \left[\begin{array}{c|cc||c} 0 & \star & \star & 0 \\ \hline 1 & \star & \star & 0 \\ 0 & \star & \star & 1 \end{array}\right].$$

Acknowledgments

This work was partially supported by DGICYT MTM2007-67812-C02-02.

Bibliography

[1] M. K. Camlibel, W. P. M. K. Heemels, and J. M. Schumacher. Algebraic necessary and sufficient conditions for controllability of conewise linear systems. *IEEE Transactions on Automatic Control*, 53(3):762–774, 2008. Cited pp. 311 and 316.

[2] V. Carmona, E. Freire, E. Ponce, and F. Torres. On simplifying and classifying piecewise-linear systems. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(5):609–620, 2002. Cited p. 311.

[3] M. di Bernardo, U. Montanero, and S. Santini. Canonical forms of generic piecewise linear continous systems. *IEEE Transactions on Automatic Control*, 56(8):1911–1915, 2011. Cited p. 311.

[4] X. Puerta. Feedback reduced and canonical forms for piecewise linear systems. In *Proceedings of 11th IEEE Workshop on Variable Structure Systems*, pages 256–259, 2010. Cited pp. 311 and 315.

[5] X. Puerta, F. Puerta, and J. Ferrer. Global reduction to the Kronecker canonical form for a C^r-family of time invariant linear systems. *Linear Algebra and its Applications*, 346(1):27–45, 2002. Cited p. 313.

On the state equivalence to classes of extended nonholonomic normal forms

Sandra Ricardo
University of Trás-os-Montes e Alto
Douro, Vila Real,
and
Instituto de Sistemas e Robótica
Coimbra, Portugal
`sricardo@utad.pt`

Abstract. This paper addresses the geometric characterisation, under state equivalence, of three classes of nonlinear systems obtained by extension of the Brockett nonholonomic integrator system. Necessary and sufficient conditions are given describing the local state equivalence of a general control-affine system with two controls to a system of those classes.

1 Introduction

In the last decades, nonholonomic systems has been the focus of an intensive and fruitful research activity, motivated both by the wide range of applications of this systems in real life and the mathematical richness intrinsically encoded in this category of nonlinear systems. The research activity was firstly devoted to first-order nonholonomic systems, and then, increasingly, also devoted to second-order nonholonomic systems.

First-order nonholonomic systems are systems subject to first-order nonholonomic constraints, that is, constraints on the generalized coordinates and velocities that are not integrable, i.e., constraints of the form $\Phi(q,\dot{q}) = 0$, which can not be written as the time derivative of some function of the generalized coordinates q. A wheeled mobile robot of unicycle type is an emblematic example of a system with first-order nonholonomic constraints [14].

Second-order nonholonomic systems are systems subject to second-order nonholonomic constraints, that is, constraints on the generalized coordinates, velocities and accelerations of the form $\Phi(q,\dot{q},\ddot{q}) = 0$ which are not integrable, i.e., can not be written as the time derivative of some function of the generalized coordinates q and velocities $\dot{q}$. Examples of second-order nonholonomic constraints are found in underactuated robot manipulators, for instance the Acrobot or the Pendubot (see [6, 22]), the aircraft PVTOL (see [11, 15, 21]), etc. The description of several examples of nonholonomic control systems can be found for instance in [7], and the references therein (see also [8]).

Many underactuated systems (i.e., control systems with fewer controls inputs (actuators) than the number of configuration variables) exhibit nonholonomic constraints which can be of first-order or of second-order. In [9], Brockett introduced the so-called Brockett nonholonomic integrator system (also called Heisenberg system) which has occupied a distinguishable place in the class of underactuated first-order nonholonomic systems. It has been pointed out as a prototypical system, located at the intersection of many different areas: control theory, sub-Riemannian geometry, the Hamilton-Jacobi theory, etc [7, 8, 23]. In [16, 17], the authors consider and studied classes of nonlinear systems with a drift term, obtained by cascading a drift-free nonholonomic system with a set of integrators. Relevant examples are systems obtained by cascading the Brockett nonholonomic integrator system, usually referred to as the extended nonholonomic normal forms. In the present paper we are interested in the geometric characterisation of those extended nonholonomic forms under state transformations, that is, we address the problem of finding necessary and sufficient conditions that a general nonlinear system must satisfy in order to be transformable, by a diffeomorphism of the state space, into those normal forms. Although our interest in this paper focuses on the classes of nonlinear systems obtained by expanding the Brockett nonholonomic integrator, the study of a certain class of second-order nonholonomic systems, that includes as particular cases the second-order chained forms, is closely related with our main purpose, and thus we shall briefly review some work in that topic.

This paper is organized as follows. In Section 2 we go through a brief literature review concerning related work. In particular, we present the nonholonomic normal forms we are interested in. Section 3 gives the main contributions of this paper, namely Theorem 3 and Theorem 5. The section starts with some preliminary notions and some related work is also discussed. The proofs of our main results are given in Section 4 and the paper ends up with some conclusions.

It is a great pleasure to contribute with this paper to the Festschrift in Honor of Uwe Helmke on the occasion of his 60th birthday. My first contact with Uwe happened in 1999, when I attended my first workshop out of Portugal, the Workshop on Lie Theory and Applications, chaired by Uwe Helmke and Knut Hüper. Two years later, I had the opportunity of spending four months at the Institut für Mathematik, University of Wuerzburg, working under the supervision of Uwe and within an EU TMR programme. This was my first experience of interaction with a research group out of Portugal. During those four months I had the pleasure of working directly with Uwe and sharing many good social moments with him and other researchers at his Institut. The work we developed together during my stay led to my first two publications, a conference paper at MTNS'2002 and a paper in the Journal Communications in Information and Systems (CIS). My first talk in English happened at the MTNS'2002, also to present joint work with Uwe. My blossoming as a researcher in mathematics is thus linked with Uwe and his research group. Thanks Uwe!

2 Nonholonomic normal forms

2.1 The Brockett nonholonomic integrator system

In his pioneering paper [9], Brockett introduced the system

$$\begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= u_2, \\ \dot{x}_3 &= x_1 u_2 - x_2 u_1, \end{aligned} \tag{1}$$

where $x = (x_1,x_2,x_3)^T \in \mathbb{R}^3$ represents the state of the system and $u = (u_1,u_2)^T \in \mathbb{R}^2$ the control. This system, quoted in the literature as the *Brockett nonholonomic integrator system* or *Heisenberg system*, is an interesting and useful local canonical form for the class of nonlinear controllable systems

$$\dot{x} = g_1(x)u_1 + g_2(x)u_2, \quad x \in \mathbb{R}^3, \quad u = (u_1,u_2) \in \mathbb{R}^2. \tag{2}$$

It is proved in [9] that there exists a local feedback and coordinate transformation of the form

$$\tilde{x} = \Psi(x), \quad \tilde{u} = \Theta(x)u,$$

about a point x_0, with Ψ a diffeomorphism and $\Theta(x)$ an orthogonal matrix, transforming the system (2) into the form (1).

The importance of the Brockett nonholonomic integrator system has been noticed both in nonlinear control and nonholonomic mechanics. In particular, it has been pointed out as a benchmark example of a first-order nonholonomic underactuated system. It mimics the kinematic model of a wheeled mobile robot of the unicycle type and displays all the basic properties of first-order nonholonomic systems.

Under simple transformation of coordinates, system (1) is converted into one of the following chain forms[1]

$$\begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= u_2, \\ \dot{x}_3 &= x_2u_1, \end{aligned} \qquad \text{or} \qquad \begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= u_2, \\ \dot{x}_3 &= x_1u_2. \end{aligned} \tag{3}$$

In [18] (see also [19, 24]) conditions have been found in order to check if a given first-order nonholonomic system can be transformed, via feedback and coordinate transformations, into the system

$$\begin{aligned} \dot{x}_1 &= u_1, \\ \dot{x}_2 &= u_2, \\ \dot{x}_3 &= x_2u_1, \\ &\vdots \\ \dot{x}_n &= x_{n-1}u_1, \end{aligned}$$

where $x = (x_1,\dots,x_n)^T$ is the state of the system and u_1 and u_2 are inputs. Such system is referred to in the literature as the *first-order chained form* or *Goursat normal form*.

2.2 Extended nonholonomic normal forms

In [2] the authors observed that the Brockett nonholonomic integrator (1) fails to capture the case where both the kinematics and dynamics of a wheeled robot must be taken into account. To tackle this realistic case, present in many physical systems

[1] Under the transformation of coordinates given, respectively, by $\tilde{x}_3 = \frac{1}{2}(x_1x_2 - x_3)$ and $\tilde{x}_3 = \frac{1}{2}(x_3 + x_1x_2)$.

(where forces and torques are the actual inputs), the authors proposed to extend the Brockett nonholonomic integrator system. Actually, it is shown in that paper that the dynamic equations of motion of a mobile robot of the unicycle type can be transformed into the system

$$\begin{aligned} \ddot{x}_1 &= u_1, \\ \ddot{x}_2 &= u_2, \\ \dot{x}_3 &= x_1\dot{x}_2 - x_2\dot{x}_1, \end{aligned} \tag{$\mathcal{ENDI}$}$$

with $(x_1,x_2,x_3,\dot{x}_1,\dot{x}_2)^T \in \mathbb{R}^5$ the state vector and $u = (u_1,u_2)^T \in \mathbb{R}^2$ the control. This system, which can be viewed as an extension of the Brockett nonholonomic integrator (1), is quoted in the literature as the *Extended Nonholonomic Double Integrator*, and will be denoted through the rest of the present paper by ($\mathcal{ENDI}$).

System ($\mathcal{ENDI}$) is locally strongly accessible for any $x \in \mathbb{R}^5$, controllable and small time locally controllable (STLC) at any equilibrium $x_e \in \{x \in \mathbb{R}^5 : \dot{x}_1 = \dot{x}_2 = 0\}$ (see [3, 16, 30]). The works [1–3] deal with the stabilization problem for system ($\mathcal{ENDI}$).

Equivalently, system ($\mathcal{ENDI}$) can be rewritten as the first-order system

$$\begin{aligned} \dot{x}_1 &= y_1, & \dot{y}_1 &= u_1, \\ \dot{x}_2 &= y_2, & \dot{y}_2 &= u_2, \\ \dot{x}_3 &= x_1y_2 - x_2y_1, \end{aligned}$$

with state $(x_1,x_2,x_3,y_1,y_2)^T \in \mathbb{R}^5$. This system falls into the class of control-affine systems with drift

$$\Sigma: \quad \dot{z} = F(z) + G_1(z)u_1 + G_2(z)u_2, \quad z \in M,$$

where M is a smooth manifold and the drift vector field F and the input vector fields G_1 and G_2 are smooth on M. The presence of the drift complicates the analysis, but makes it more challenging.

Two other systems that naturally fall into the above class of control-affine systems with drift, and are as well extensions of the Brockett nonholonomic integrator (1), are the second-order nonholonomic systems considered in the works [16, 17]:

$$\begin{aligned} \ddot{x}_1 &= u_1, \\ \ddot{x}_2 &= u_2, \\ \ddot{x}_3 &= x_1u_2 - x_2u_1, \end{aligned} \tag{$\mathcal{ENMS}$}$$

and

$$\begin{aligned} \ddot{x}_1 &= u_1, \\ \ddot{x}_2 &= u_2, \\ \ddot{x}_3 &= \dot{x}_1u_2 - \dot{x}_2u_1, \end{aligned} \tag{$\mathcal{ENNM}$}$$

with $(x_1,x_2,x_3,\dot{x}_1,\dot{x}_2,\dot{x}_3)^T \in \mathbb{R}^6$ the state vector and $u = (u_1,u_2)^T \in \mathbb{R}^2$ the control. System ($\mathcal{ENMS}$) is usually referred to as the mechanical extension of the Brockett nonholonomic integrator. Through the rest of the paper we shall call it the *Extended Nonholonomic mechanical system* and denote it by $(\mathcal{ENMS})$. As far as we are aware, there is not a usual quotation for system $(\mathcal{ENNM})$, but we point out that it has a crucial difference with respect to the other two extended systems: it is not a mechanical system, since the input vector fields depend on velocities (for a definition of a mechanical control system see, for instance, [10, 27]). Through this paper we shall call it the *Extended Nonholonomic Not Mechanical system* and denote it by $(\mathcal{ENNM})$. In [17] optimal trajectories were obtained for systems $(\mathcal{ENMS})$ and $(\mathcal{ENNM})$. The work [16] discusses tracking and stabilization problems for the three extended nonholonomic forms $(\mathcal{ENDI})$, $(\mathcal{ENMS})$ and $(\mathcal{ENNM})$. Accessibility and controllability results were also discussed in that paper.

2.3 Second-order nonholonomic chained forms

Systems with second-order nonholonomic constraints always include the drift-term. As already observed, the presence of the drift complicates the analysis, but offers challenging problems. Many researchers have been working on the stabilization problem and on the tracking control problem for this class of systems. See, for instance, the works [4, 5, 12, 21, 31]. In these works, a key procedure is to convert the second-order nonholonomic system under consideration, via feedback and coordinate transformations, into a special normal system of the form

$$\begin{aligned}\ddot{x}_1 &= u_1,\\ \ddot{x}_2 &= u_2,\\ \ddot{x}_3 &= x_2 u_1,\end{aligned} \qquad (\mathcal{SCF}_1) \qquad \text{or} \qquad \begin{aligned}\ddot{x}_1 &= u_1,\\ \ddot{x}_2 &= u_2,\\ \ddot{x}_3 &= x_1 u_2,\end{aligned} \qquad (\mathcal{SCF}_2)$$

with $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ the configuration of the system and $u = (u_1, u_2)^T \in \mathbb{R}^2$ the control. The obtained systems, here labeled $(\mathcal{SCF}_1)$ and $(\mathcal{SCF}_2)$, are known in the literature as the *second-order chained forms*. These forms simplify considerably the dynamical equations of the system, so being much more suitable to deal with than the original dynamical equations.

We observe that the second-order chained forms can be seen as the analogues of the first-order chained forms in the 3-dimensional case, see (3), reflecting similarities between the first-order nonholonomic constraints, respectively, $\dot{x}_3 = x_2\dot{x}_1$ and $\dot{x}_3 = x_1\dot{x}_2$, and the second-order nonholonomic constraints, respectively, $\ddot{x}_3 = x_2\ddot{x}_1$ and $\ddot{x}_3 = x_1\ddot{x}_2$. We can also look at system $(\mathcal{ENMS})$ (see the precedent subsection) as the analogue of the Brockett nonholonomic integrator, reflecting similarities between the first-order nonholonomic constraint $\dot{x}_3 = x_1\dot{x}_2 - x_2\dot{x}_1$ and the second-order nonholonomic constraint $\ddot{x}_3 = x_1\ddot{x}_2 - x_2\ddot{x}_1$. However, it is important to notice that, contrary to what happens with the three first-order nonholonomic forms described in (1) and (3), the three second-order nonholonomic forms $(\mathcal{ENMS})$, $(\mathcal{SCF}_1)$ and $(\mathcal{SCF}_2)$ are not state equivalent, that is, we cannot pass from one system to another by coordinate transformations.

In [29], the authors considered second-order nonholonomic systems described by the following equations,

$$\begin{aligned}
\ddot{x}_1 &= u_1,\\
\ddot{x}_2 &= u_2,\\
\ddot{x}_3 &= w\left(\frac{s_1}{2}\,(\dot{x}_1)^2 + \frac{s_2}{2}\,(\dot{x}_2)^2 + \dot{x}_1\dot{x}_2\right) + \frac{1}{2}(x_1u_2 - x_2u_1), \quad s_1, s_2, w \in \mathbb{R},
\end{aligned} \tag{4}$$

with $x = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ the configuration of the system and $u = (u_1, u_2)^T \in \mathbb{R}^2$ the control. It was proved in that paper that system (4) is a canonical form for the class of second-order nonholonomic mechanical control systems described by the equations

$$\begin{aligned}
\ddot{x}_1 &= u_1,\\
\ddot{x}_2 &= u_2,\\
\ddot{x}_3 &= -\dot{x}^T\Gamma\dot{x} + (Kx)^T u,
\end{aligned} \tag{5}$$

where

$$\Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & 0\\ \Gamma_{12} & \Gamma_{22} & 0\\ 0 & 0 & 0 \end{pmatrix} = \Gamma^T, \quad K = \begin{pmatrix} k_{11} & k_{12} & 0\\ k_{21} & k_{22} & 0\\ 0 & 0 & 0 \end{pmatrix}, \quad x = \begin{pmatrix} x_1\\ x_2\\ x_3 \end{pmatrix}, \quad u = \begin{pmatrix} u_1\\ u_2\\ 0 \end{pmatrix},$$

and $\Gamma_{ij}, k_{ij} \in \mathbb{R}$. The constants s_1, s_2 and w, in system (4), were proved to be invariants of the system and are related with the constants Γ_{ij}, k_{ij} by the following equalities

$$w = -\frac{2\Gamma_{12} + k_{12} + k_{21}}{k_{21} - k_{12}}, \quad s_1 = \frac{2(\Gamma_{11} + k_{11})}{2\Gamma_{12} + k_{12} + k_{21}}, \quad s_2 = \frac{2(\Gamma_{22} + k_{22})}{2\Gamma_{12} + k_{12} + k_{21}}.$$

Systems ($\mathcal{SCF}_1$), ($\mathcal{SCF}_2$) and ($\mathcal{ENMS}$) correspond thus to the values of the invariants $s_1 = s_2 = 0$ in the three cases and the invariant w is, respectively, $1, -1$ and 0.

In [29] (see also [26, 28]), the authors consider the problem of finding necessary and sufficient conditions in order to check if a general control-affine system can be transformable, under coordinate transformations, into a system belonging to the class of second-order nonholonomic systems (4). Two alternative answers are given to this question, but excluding the degenerate case of system ($\mathcal{ENMS}$). We return to this issue in Section 3.2.

3 S-equivalence to the nonholonomic normal forms

3.1 Preliminaries

Let M and $\widetilde{M}$ be smooth manifolds and consider the control-affine systems

$$\Sigma: \quad \dot{z} = F(z) + \sum_{r=1}^{m} u_r G_r(z) \qquad \text{and} \qquad \widetilde{\Sigma}: \quad \dot{\tilde{z}} = \tilde{F}(\tilde{z}) + \sum_{r=1}^{m} u_r \tilde{G}_r(\tilde{z}),$$

with $z \in M$ and $\tilde{z} \in \widetilde{M}$. We say that Σ and $\widetilde{\Sigma}$ are state equivalent, shortly *S-equivalent*, if there exists a diffeomorphism $\Phi: M \to \widetilde{M}$ such that

$$\mathrm{D}\Phi(z) \cdot F(z) = \tilde{F}(\Phi(z)) \quad \text{and} \quad \mathrm{D}\Phi(z) \cdot G_r(z) = \tilde{G}_r(\Phi(z)), \quad 1 \le r \le m,$$

(where DΦ stands for the differential of Φ) which we will denote as

$$\Phi_* F = \tilde{F} \quad \text{and} \quad \Phi_* G_r = \tilde{G}_r, \quad 1 \le r \le m.$$

The systems Σ and $\widetilde{\Sigma}$ are called locally S-equivalent, at $z_0 \in M$ and $\tilde{z}_0 \in \widetilde{M}$, respectively, if there exist neighborhoods U_{z_0} of z_0 and $\tilde{U}_{\tilde{z}_0}$ of $\tilde{z}_0$, such that Σ restricted to U_{z_0} and $\widetilde{\Sigma}$ restricted to $\tilde{U}_{\tilde{z}_0}$ are S-equivalent [13, 20, 25].

In this paper, our purpose is to give a set of necessary and sufficient conditions to characterise the state equivalence of a control-affine system with two controls to the extended nonholonomic normal forms considered in the precedent section. Our conditions will be given in terms of Lie brackets of vector fields. We recall that, the *Lie bracket* of two smooth vector fields X and Y, defined on a smooth manifold M, is a new smooth vector field denoted by $[X,Y]$ and defined, in coordinates, as

$$[X,Y](z) = \mathrm{D}Y(z)X(z) - \mathrm{D}X(z)Y(z), \quad z \in M,$$

where D$Y(z)$ and D$X(z)$ denote the Jacobi matrix of Y and X in z-coordinates, respectively. We will use often the notation

$$\mathrm{ad}_X^0 Y = Y \quad \text{and} \quad \mathrm{ad}_X^{j+1} Y = [X, \mathrm{ad}_X^j Y], \quad j \ge 0.$$

Notice that transforming a vector field via a diffeomorphism is compatible with Lie bracket, that is, $\Phi_*([X,Y]) = [\Phi_* X, \Phi_* Y]$.

3.2 S-equivalence to the second-order chained forms

As observed in Section 2.3, systems ($\mathcal{ENMS}$), ($\mathcal{SCF}_1$) and ($\mathcal{SCF}_2$) belong to the class of second-order nonholonomic mechanical control systems described, in the configuration space $\mathbb{R}^3$, by equations (4). Equivalently, this class can be described by the first-order equations, in the state space $\mathbb{R}^6$:

$$\begin{aligned} \dot{x}_1 &= y_1, & \dot{y}_1 &= u_1, \\ \dot{x}_2 &= y_2, & \dot{y}_2 &= u_2, \\ \dot{x}_3 &= y_3, & \dot{y}_3 &= w\left(\frac{s_1}{2}\, y_1^2 + \frac{s_2}{2}\, y_2^2 + y_1 y_2\right) + \frac{1}{2}(x_1 u_2 - x_2 u_1), \end{aligned}$$

where $(x_1,x_2,x_3,y_1,y_2,y_3)^T \in \mathbb{R}^6$ and $s_1, s_2, w \in \mathbb{R}$. In [29] the following question is addressed: *"Which conditions must a control-affine system, with two controls,*

$$\Sigma: \quad \dot{z} = F(z) + G_1(z)u_1 + G_2(z)u_2, \quad z \in \mathbb{R}^6,$$

satisfy in order to be locally S-equivalent to a system of the class considered above?" The next result is one of the two alternative answers given in that paper to this question.

Theorem 1. *[29] The system Σ is locally S-equivalent, at $z_0 \in \mathbb{R}^6$, to the system*

$$\begin{aligned} \dot{x}_1 &= y_1, & \dot{y}_1 &= u_1, \\ \dot{x}_2 &= y_2, & \dot{y}_2 &= u_2, \\ \dot{x}_3 &= y_3, & \dot{y}_3 &= w\left(\frac{s_1}{2}\, y_1^2 + \frac{s_2}{2}\, y_2^2 + y_1 y_2\right) + \frac{1}{2}(x_1 u_2 - x_2 u_1), \end{aligned}$$

where $w \in \mathbb{R}\backslash\{0\}$ *and* $s_1, s_2 \in \mathbb{R}$, *at the origin of* $\mathbb{R}^6$, *if and only if the system* Σ *satisfies, in a neighborhood of* z_0, *the conditions:*

(C1) $G_1, G_2, \mathrm{ad}_F G_1, \mathrm{ad}_F G_2, [G_1, \mathrm{ad}_F G_2], [\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]$ *are linearly independent;*

(C2) $G_1, G_2, [G_1, \mathrm{ad}_F G_2], [\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]$ *commute;*

(C3) *(i)* $[G_i, \mathrm{ad}_F G_i] = s_i [G_1, \mathrm{ad}_F G_2], \ i = 1,2,$
(ii) $[F, [G_1, \mathrm{ad}_F G_2]] = w[\mathrm{ad}_F G_1, \mathrm{ad}_F G_2];$

(C4) *(i)* $[\mathrm{ad}_F G_i, \mathrm{ad}_F^2 G_j] = 0, \ i, j = 1,2,$
(ii) $F(z_0) = \mathrm{ad}_F^2 G_i(z_0) = 0, \ i = 1,2.$

Obviously, the geometric characterisation of the second-order chained forms $(\mathcal{SCF}_1)$ and $(\mathcal{SCF}_2)$ are obtained immediately from Theorem 1, for the values of the invariants $s_1 = s_2 = 0$ and, respectively, $w = 1$ and $w = -1$.

As already observed, the system $(\mathcal{ENMS})$ corresponds to invariants $w = s_1 = s_2 = 0$. This degenerated case is excluded from the considerations in the above theorem. In fact, this system is the only member of the class of systems considered in the statement of the above theorem that do not satisfy property (C1), which results directly from the fact that $w = 0$. Systems of that class satisfying property (C1) are said to satisfy the geodesic accessibility property [29] (for a definition of geodesically accessible mechanical control systems see [27]).

Remark 2. In [29] the authors also proposed an alternative version of the above theorem, involving another set of necessary and sufficient conditions. Such alternative set of conditions has the advantage that the role those conditions play is well understood in terms of mechanical properties, but, in counterpart, some of those conditions are, in general, not easy to check. Theorem 1 has the advantage of involving a set of necessary and sufficient conditions that are easier to check, and moreover, leads to a constructive proof, allowing to get the diffeomorphism performing the state equivalence.

3.3 S-equivalence to the extended nonholonomic normal forms

In this section we are interested in answering the question: When is a general control-affine system, with two controls,

$$\Sigma: \quad \dot{z} = F(z) + G_1(z)u_1 + G_2(z)u_2, \tag{6}$$

locally S-equivalent to a system of the form $(\mathcal{ENDI})$, $(\mathcal{ENMS})$ and $(\mathcal{ENNM})$?

As observed before, these three systems can be viewed as extensions of the Brockett nonholonomic integrator (1). The extended nonholonomic forms $(\mathcal{ENDI})$ and $(\mathcal{ENMS})$ are both mechanical control systems. System $(\mathcal{ENDI})$ is first-order nonholonomic, whereas system $(\mathcal{ENMS})$ is second-order nonholonomic. System $(\mathcal{ENNM})$ is second-order nonholonomic but it is not a mechanical system, since the input vector fields depend on velocities (for a definition of a mechanical control system see [10, 27]).

3.3.1 Converting systems to the form $(\mathcal{ENDI})$

As already observed, it is shown in [2] that the dynamic equations of motion of a mobile robot of the unicycle type can be transformed into the system $(\mathcal{ENDI})$.

We see next which conditions a two input control-affine system (6) must satisfy to be S-equivalent to the form $(\mathcal{ENDI})$.

Theorem 3. *The system Σ is locally S-equivalent, at $z_0 \in \mathbb{R}^5$, to the system*

$$\begin{aligned} \ddot{x}_1 &= u_1, \\ \ddot{x}_2 &= u_2, \\ \dot{x}_3 &= x_1\dot{x}_2 - x_2\dot{x}_1, \end{aligned} \tag{$\mathcal{ENDI}$}$$

at the origin of $\mathbb{R}^5$, if and only if the system Σ satisfies in a neighborhood of z_0, the conditions:

(C1) $G_1, G_2, \mathrm{ad}_F G_1, \mathrm{ad}_F G_2\,[\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]$ *are linearly independent;*

(C2) $G_1, G_2, [\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]$ *commute;*

(C3) *(i)* $[G_i, \mathrm{ad}_F G_j] = 0,\ i,j = 1,2,$

(ii) $[\mathrm{ad}_F G_i, \mathrm{ad}_F^2 G_j] = 0,\ i,j = 1,2,$

(C4) $F(z_0) = \mathrm{ad}_F^2 G_i(z_0) = 0,\ i = 1,2.$

We observe that conditions (C1) - (C4) are easily checkable, since they involve only the calculation of Lie brackets in terms of the vector fields F, G_1 and G_2. For a proof see Section 4.1.

3.3.2 What about the system $(\mathcal{ENMS})$?

In state-space form, system $(\mathcal{ENMS})$, reads as

$$\begin{aligned} \dot{x}_1 &= y_1, & \dot{y}_1 &= u_1, \\ \dot{x}_2 &= y_2, & \dot{y}_2 &= u_2, \\ \dot{x}_3 &= y_3, & \dot{y}_3 &= x_1u_2 - x_2u_1, \end{aligned}$$

with $(x,y) = (x_1,x_2,x_3,y_1,y_2,y_3)^T \in \mathbb{R}^6$. Denoting by G_1, G_2 the input vector fields and by F the drift, we have

$$G_1 = \frac{\partial}{\partial y_1} - x_2\frac{\partial}{\partial y_3}, \quad G_2 = \frac{\partial}{\partial y_2} + x_1\frac{\partial}{\partial y_3}, \quad \text{and} \quad F = y_i\frac{\partial}{\partial x_i}, \quad i = 1,2,3,$$

where a summation is understood over the index i. By Lie brackets computations we can see that the vector fields

$$G_1,\ G_2,\ \mathrm{ad}_F G_1,\ \mathrm{ad}_F G_2,\ \text{and}\ [\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]$$

span a 5-dimensional space, which is indeed the highest possible dimension we can obtain. It is then clear that this system is never controllable in $\mathbb{R}^6$. Since

$$\ddot{x}_3 = x_1\ddot{x}_2 - x_2\ddot{x}_1 = \frac{d}{dt}(x_1\dot{x}_2 - x_2\dot{x}_1),$$

we conclude that the second-order constraint $\ddot{x}_3 = x_1\ddot{x}_2 - x_2\ddot{x}_1$ is indeed integrable and reduces to

$$\dot{x}_3 = (x_1\dot{x}_2 - x_2\dot{x}_1) + k, \quad k \in \mathbb{R}.$$

System ($\mathcal{ENMS}$) is only controllable on the submanifold of $\mathbb{R}^6$ given by

$$\mathcal{S} = \{(x,y) \in \mathbb{R}^6 \mid \dot{x}_3 - x_1\dot{x}_2 + x_2\dot{x}_1 = k\},$$

where the system ($\mathcal{ENMS}$) reduces to

$$\begin{aligned} \dot{x}_1 &= y_1, & \dot{y}_1 &= u_1,\\ \dot{x}_2 &= y_2, & \dot{y}_2 &= u_2.\\ \dot{x}_3 &= x_1y_2 - x_2y_1 + k, \end{aligned}$$

At equilibrium points we have $k = 0$ and the system reduces to system ($\mathcal{ENDI}$).

Remark 4. We observe that out of equilibria we obtain several mechanical structures (due to the parameter $k \in \mathbb{R}$), which is in accordance with results obtained in [27, 29], namely the fact that the system ($\mathcal{ENMS}$) does not satisfy the geodesic accessibility property - a crucial property for guaranteeing the uniqueness of the mechanical structure.

3.3.3 Converting systems to the form ($\mathcal{ENNM}$)

We shall now consider the second-order nonholonomic system ($\mathcal{ENNM}$). This system can be re-written in state-space form as the first-order system,

$$\begin{aligned} \dot{x}_1 &= y_1, & \dot{y}_1 &= u_1,\\ \dot{x}_2 &= y_2, & \dot{y}_2 &= u_2,\\ \dot{x}_3 &= y_3, & \dot{y}_3 &= y_1u_2 - y_2u_1, \end{aligned}$$

with $(x,y) = (x_1,x_2,x_3,y_1,y_2,y_3)^T \in \mathbb{R}^6$. Necessary and sufficient conditions for the S-equivalence to this form are given in next result.

Theorem 5. *The system Σ is locally S-equivalent, at $z_0 \in \mathbb{R}^6$, to the system*

$$\begin{aligned} \dot{x}_1 &= y_1, & \dot{y}_1 &= u_1,\\ \dot{x}_2 &= y_2, & \dot{y}_2 &= u_2,\\ \dot{x}_3 &= y_3, & \dot{y}_3 &= y_1u_2 - y_2u_1, \end{aligned} \tag{$\mathcal{ENNM}$}$$

at the origin of $\mathbb{R}^6$, if and only if the system Σ satisfies, in a neighborhood of z_0, the conditions:

(C1) $G_1, G_2, [G_1,G_2], \mathrm{ad}_F G_1, \mathrm{ad}_F G_2, [F,[G_1,G_2]]$ *are linearly independent;*

(C2) $\mathrm{ad}_F G_1, \mathrm{ad}_F G_2, [F,[G_1,G_2]], [G_1,G_2]$ *commute;*

(C3) *(i)* $\mathrm{ad}_F^2 G_i = [G_i, \mathrm{ad}_F G_i] = [G_i,[G_1,G_2]] = 0,\ i = 1,2,$

(ii) $[G_1, \mathrm{ad}_F G_2] = -[G_2, \mathrm{ad}_F G_1];$

(C4) $F(z_0) = 0.$

Remark 6. We observe that system ($\mathcal{ENNM}$) is of a different nature from that of systems ($\mathcal{ENDI}$) and ($\mathcal{ENMS}$). In fact, there exist a crucial structural difference due to the fact that in system ($\mathcal{ENNM}$) the input vector fields depend on velocities, what do not happen with the other systems.

The proof of Theorem 5 is given in Section 4.2.

4 Proofs of main results

In this section we give proofs of Theorem 3 and Theorem 5. In both proofs we use the well-known result:

Theorem 7 (The Simultaneous Flow Box Theorem)**.** *Let M be an n-dimensional smooth manifold and $z \in M$. If $g_1, g_2, \ldots, g_k$, $k \le n$, are smooth vector fields on M satisfying the conditions:*

(i) The k vectors $g_1(z), g_2(z), \ldots, g_k(z)$ are linearly independent,

(ii) $[g_i, g_j] = 0$, *for all* $1 \le i, j \le k$,

then there exists a local coordinate system $(x_1, x_2, \ldots, x_k)$ on an open neighborhood U of z in which

$$g_1 = \frac{\partial}{\partial x_1}, \quad g_2 = \frac{\partial}{\partial x_2}, \quad \cdots \quad , \quad g_k = \frac{\partial}{\partial x_k}.$$

4.1 Proof of Theorem 3

Proof of Theorem 3. Necessity is obvious, by a direct calculation. We prove sufficiency by showing that there exists a sequence of diffeomorphisms that brings the system to the desired form. By Theorem 7, the conditions (C1) and (C2) allow to conclude the existence of a diffeomorphism $\phi : U_{z_0} \subset \mathbb{R}^5 \to \mathbb{R}^5$, with U_{z_0} an open neighborhood of z_0 and $\phi(z) = (x,y) = (x_1,x_2,x_3,y_1,y_2)$[2] such that $\phi(z_0) = 0$,

$$\tilde{G}_1 = \phi_*(G_1) = \frac{\partial}{\partial y_1}, \quad \tilde{G}_2 = \phi_*(G_2) = \frac{\partial}{\partial y_2},$$

and

$$\left[\mathrm{ad}_{\tilde{F}} \tilde{G}_1, \mathrm{ad}_{\tilde{F}} \tilde{G}_2\right] = \phi_*([\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]) = \frac{\partial}{\partial x_3}.$$

[2] There is some advantage for the sequel in considering $\phi(z) = (x,y) = (x_1,x_2,x_3,y_1,y_2)$, instead of $\phi(z) = (x_1,x_2,x_3,x_4,x_5)$, which at this point could appear more natural. Actually, it allows to distinguish the special role that the input vector fields play in the proof.

In this system of local coordinates (x,y), we denote the drift as

$$\tilde{F} = f_1 \frac{\partial}{\partial x_1} + f_2 \frac{\partial}{\partial x_2} + f_3 \frac{\partial}{\partial x_3} + f_4 \frac{\partial}{\partial y_1} + f_5 \frac{\partial}{\partial y_2},$$

where the components $f_r = f_r(x,y)$, $1 \le r \le 5$, are smooth functions in a neighborhood of $0 \in \mathbb{R}^5$. The condition (C3)(i) says $[G_i, \mathrm{ad}_F G_j] = 0$, $i,j = 1,2$. Therefore,

$$\frac{\partial^2 f_r}{\partial y_i \partial y_j} = 0, \quad i,j = 1,2,$$

from which we obtain

$$f_r = k_r^1\, y_1 + k_r^2\, y_2 + k_r^3, \quad 1 \le r \le 5, \tag{7}$$

with $k_r^l = k_r^l(x)$, $1 \le l \le 3$, smooth functions depending on variables $x = (x_1,x_2,x_3)$. The Jacobi identity and property (C3)(ii) allow us to conclude that $[F,[\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]] = 0$. Therefore,

$$[\tilde{F},[\mathrm{ad}_{\tilde{F}} \tilde{G}_1, \mathrm{ad}_{\tilde{F}} \tilde{G}_2]] = 0,$$

that is, the drift $\tilde{F}$ does not depend on coordinate x_3. The above considerations imply that the drift vector field takes the form

$$\tilde{F} = \left(k_j^1\, y_1 + k_j^2\, y_2 + k_j^3\right) \frac{\partial}{\partial x_j} + \left(k_{i+3}^1\, y_1 + k_{i+3}^2\, y_2 + k_{i+3}^3\right) \frac{\partial}{\partial y_i},\ j = 1,2,3,\ i = 1,2,$$

where a summation is understood over the indices j and i, and $k_r^l = k_r^l(x_1,x_2)$, $1 \le r \le 5$, $l = 1,2,3$, are smooth functions of x_1, x_2. For simplicity, we skip the "tilde" notation, and, in particular, denote the vector fields $\tilde{F}$ and $\tilde{G}_i$ simply by F and G_i, $i = 1,2$. We compute

$$-\mathrm{ad}_F G_1 = k_j^1 \frac{\partial}{\partial x_j} + k_{i+3}^1 \frac{\partial}{\partial y_i} \quad \text{and} \quad -\mathrm{ad}_F G_2 = k_j^2 \frac{\partial}{\partial x_j} + k_{i+3}^2 \frac{\partial}{\partial y_i}, \quad j = 1,2,3,\ i = 1,2.$$

Consider the distribution $\mathcal{D} := \mathrm{span}\{\frac{\partial}{\partial x_3}\}$. We observe that

$$[\mathrm{ad}_F G_i, \mathcal{D}] = 0 \quad \text{and} \quad [G_i, \mathcal{D}] = 0, \quad i = 1,2,$$

implying that the distribution $\mathcal{D}$ is invariant under the vector fields $\mathrm{ad}_F G_i$ and G_i, $i = 1,2$. Thus, the projection $\tau : \mathbb{R}^5 \to \mathbb{R}^4$, $\tau(x,y) = (x_1,x_2,y_1,y_2)$, is well defined. We consider the projections of the vector fields $-\mathrm{ad}_F G_i$ and G_i on $\mathbb{R}^4$, given by $\tau_*(-\mathrm{ad}_F G_i)$ and $\tau_* G_i$ (which are well defined by the above property of $\mathcal{D}$). We have

$$\tau_*(-\mathrm{ad}_F G_1) = k_i^1 \frac{\partial}{\partial x_i} + k_{i+3}^1 \frac{\partial}{\partial y_i}, \qquad \tau_* G_1 = \frac{\partial}{\partial y_1},$$
$$\tau_*(-\mathrm{ad}_F G_2) = k_i^2 \frac{\partial}{\partial x_i} + k_{i+3}^2 \frac{\partial}{\partial y_i}, \qquad \tau_* G_2 = \frac{\partial}{\partial y_2}.$$

From the equality $[\mathrm{ad}_F G_1, \mathrm{ad}_F G_2] = \frac{\partial}{\partial x_3}$, we obtain $[\tau_*(-\mathrm{ad}_F G_1), \tau_*(-\mathrm{ad}_F G_2)] = 0$. We conclude that $\tau_*(-\mathrm{ad}_F G_1), \tau_*(-\mathrm{ad}_F G_2), \tau_* G_1$ and $\tau_* G_2$ are commuting vector fields on $\mathbb{R}^4$. By (C1), these vector fields are also independent. Therefore, there exists on $\mathbb{R}^4$, a local diffeomorphism ψ such that $\psi(x_1,x_2,y_1,y_2) = (\bar{x}_1,\bar{x}_2,\bar{y}_1,\bar{y}_2)$, with $\psi(0) = 0$ and

$$\psi_*(\tau_*(-\mathrm{ad}_F G_i)) = \frac{\partial}{\partial \bar{x}_i} \quad \text{and} \quad \psi_*(\tau_* G_i) = \frac{\partial}{\partial \bar{y}_i}, \quad i = 1,2. \tag{8}$$

The diffeomorphism ψ satisfying the above equalities must be of the form

$$\begin{aligned} \bar{x}_1 &= \psi_1(x_1,x_2), & \bar{y}_1 &= y_1 + \psi_3(x_1,x_2), \\ \bar{x}_2 &= \psi_2(x_1,x_2), & \bar{y}_2 &= y_2 + \psi_4(x_1,x_2), \end{aligned}$$

with ψ_i, $i = 1,2,3,4$, smooth functions of x_1 and x_2. In particular, the first equality of (8) imply

$$\begin{aligned} \frac{\partial \psi_1}{\partial x_1}k_1^1 + \frac{\partial \psi_1}{\partial x_2}k_2^1 &= 1, & \frac{\partial \psi_1}{\partial x_1}k_1^2 + \frac{\partial \psi_1}{\partial x_2}k_2^2 &= 0, \\ \frac{\partial \psi_2}{\partial x_1}k_1^1 + \frac{\partial \psi_2}{\partial x_2}k_2^1 &= 0, & \frac{\partial \psi_2}{\partial x_1}k_1^2 + \frac{\partial \psi_2}{\partial x_2}k_2^2 &= 1, \\ \frac{\partial \psi_3}{\partial x_1}k_1^1 + \frac{\partial \psi_3}{\partial x_2}k_2^1 + k_4^1 &= 0, & \frac{\partial \psi_3}{\partial x_1}k_1^2 + \frac{\partial \psi_3}{\partial x_2}k_2^2 + k_4^2 &= 0, \\ \frac{\partial \psi_4}{\partial x_1}k_1^1 + \frac{\partial \psi_4}{\partial x_2}k_2^1 + k_5^1 &= 0, & \frac{\partial \psi_4}{\partial x_1}k_1^2 + \frac{\partial \psi_4}{\partial x_2}k_2^2 + k_5^2 &= 0. \end{aligned}$$

Consider on $\mathbb{R}^5$, the coordinates transformation Ψ defined by

$$\begin{aligned} \bar{x}_1 &= \psi_1(x_1,x_2), & \bar{y}_1 &= y_1 + \psi_3(x_1,x_2), \\ \bar{x}_2 &= \psi_2(x_1,x_2), & \bar{y}_2 &= y_2 + \psi_4(x_1,x_2). \\ \bar{x}_3 &= x_3, \end{aligned} \tag{9}$$

The drift F is transformed via Ψ into

$$\bar{F} = \Psi_* F = (\bar{y}_i + \bar{k}_i)\,\frac{\partial}{\partial \bar{x}_i} + \bar{k}_{i+3}\,\frac{\partial}{\partial \bar{y}_i} + (\bar{k}_3^1\,\bar{y}_1 + \bar{k}_3^2\,\bar{y}_2 + \bar{k}_3^3)\,\frac{\partial}{\partial \bar{x}_3}, \quad i = 1,2,$$

with a summation understood over the index i, and $\bar{k}_i, \bar{k}_{i+3}$ and $\bar{k}_3^l$, $l = 1,2,3$, new smooth functions depending on the variables $\bar{x}_1,\bar{x}_2$. Clearly, the transformation of coordinates (9) preserves the vector fields G_1, G_2 and $[\mathrm{ad}_F G_1, \mathrm{ad}_F G_2]$, i.e.,

$$\begin{aligned} &\bar{G}_1 = \Psi_* G_1 = \frac{\partial}{\partial \bar{y}_1}, \quad \bar{G}_2 = \Psi_* G_2 = \frac{\partial}{\partial \bar{y}_2}, \quad \text{and} \\ &\left[\mathrm{ad}_{\bar{F}} \bar{G}_1, \mathrm{ad}_{\bar{F}} \bar{G}_2\right] = \Psi_*[\mathrm{ad}_F G_1, \mathrm{ad}_F G_2] = \frac{\partial}{\partial \bar{x}_3}. \end{aligned} \tag{10}$$

For simplicity, we skip the "bar" notation. Doing so, the vector fields $\bar{F}, \bar{G}_1, \bar{G}_2$ will be denoted simply by F, G_1, G_2. In particular, the last equation of equality (10), which now takes the form $[\mathrm{ad}_F G_1, \mathrm{ad}_F G_2] = \frac{\partial}{\partial x_3}$, implies

$$\frac{\partial k_3^2}{\partial x_1} - \frac{\partial k_3^1}{\partial x_2} = 1. \tag{11}$$

Set

$$\tilde{x}_3 = x_3 + \varphi(x_1, x_2), \tag{12}$$

with φ a smooth function to be determined. Then

$$\begin{aligned}\dot{\tilde{x}}_3 &= k_3^1\, y_1 + k_3^2\, y_2 + k_3^3 + \frac{\partial\varphi}{\partial x_1}(y_1 + k_1) + \frac{\partial\varphi}{\partial x_2}(y_2 + k_2)\\ &= (k_3^1 + \frac{\partial\varphi}{\partial x_1})y_1 + (k_3^2 + \frac{\partial\varphi}{\partial x_2})y_2 + (k_3^3 + \frac{\partial\varphi}{\partial x_1}k_1 + \frac{\partial\varphi}{\partial x_2}k_2).\end{aligned}$$

We observe that the following system of PDEs

$$\begin{cases} \frac{\partial\varphi}{\partial x_1} + k_3^1 = -\frac{1}{2}\, x_2 \\ \frac{\partial\varphi}{\partial x_2} + k_3^2 = \frac{1}{2}\, x_1 \end{cases} \tag{13}$$

satisfies the integrability conditions, i.e., $\frac{\partial^2\varphi}{\partial x_2 \partial x_1} = \frac{\partial^2\varphi}{\partial x_1 \partial x_2}$, $i = 1,2$ (because of (11)) and therefore, possess solution $\varphi(x_1, x_2)$. We plug that solution into (12) and obtain

$$\dot{\tilde{x}}_3 = -\frac{1}{2}\, x_2 y_1 + \frac{1}{2}\, x_1 y_2 + \tilde{k}_3,$$

where $\tilde{k}_3 = k_3^3 + \frac{\partial\varphi}{\partial x_1}k_1 + \frac{\partial\varphi}{\partial x_2}k_2$ is a function depending on variables x_1 and x_2. Once again, in order to simplify, we skip the "tilde" notation and denote the variable $\tilde{x}_3$ by x_3 and the function $\tilde{k}_3$ by k_3. The drift takes the form

$$F = (y_i + k_i)\,\frac{\partial}{\partial x_i} + k_{i+3}\,\frac{\partial}{\partial y_i} + (-\frac{1}{2}x_2 y_1 + \frac{1}{2}x_1 y_2 + k_3)\,\frac{\partial}{\partial x_3} \tag{14}$$

with a sum understood over the index $i = 1,2$. We still have

$$G_1 = \frac{\partial}{\partial y_1}, \quad G_2 = \frac{\partial}{\partial y_2}, \quad \text{and} \quad [\mathrm{ad}_F G_1, \mathrm{ad}_F G_2] = \frac{\partial}{\partial x_3},$$

since these vector fields are preserved by the transformation of coordinates (12).

We shall see now that, under condition $(C4)$, the functions k_r, $1 \le r \le 5$, vanish. Indeed, from (14) we obtain

$$\mathrm{ad}_F G_1 = -\frac{\partial}{\partial x_1} + \frac{1}{2}x_2\frac{\partial}{\partial x_3}, \qquad \mathrm{ad}_F G_2 = -\frac{\partial}{\partial x_2} - \frac{1}{2}x_1\frac{\partial}{\partial x_3},$$

and

$$\begin{aligned}\mathrm{ad}_F^2 G_1 &= \frac{\partial k_j}{\partial x_1}\frac{\partial}{\partial x_j} + \left(y_2 + \frac{\partial k_3}{\partial x_1} + \frac{1}{2}k_2\right)\frac{\partial}{\partial x_3} + \frac{\partial k_{j+3}}{\partial x_1}\frac{\partial}{\partial y_j},\\ \mathrm{ad}_F^2 G_2 &= \frac{\partial k_j}{\partial x_2}\frac{\partial}{\partial x_j} + \left(-y_1 + \frac{\partial k_3}{\partial x_2} - \frac{1}{2}k_1\right)\frac{\partial}{\partial x_3} + \frac{\partial k_{j+3}}{\partial x_2}\frac{\partial}{\partial y_j}.\end{aligned}$$

Then

$$\begin{aligned}
[\mathrm{ad}_F G_1, \mathrm{ad}_F^2 G_1] &= -\frac{\partial^2 k_j}{\partial x_1^2}\frac{\partial}{\partial x_j} - \left(\frac{\partial^2 k_3}{\partial x_1^2} + \frac{\partial k_2}{\partial x_1}\right)\frac{\partial}{\partial x_3} - \frac{\partial^2 k_{j+3}}{\partial x_1^2}\frac{\partial}{\partial y_j},\\
[\mathrm{ad}_F G_2, \mathrm{ad}_F^2 G_2] &= -\frac{\partial^2 k_j}{\partial x_2^2}\frac{\partial}{\partial x_j} - \left(\frac{\partial^2 k_3}{\partial x_2^2} - \frac{\partial k_1}{\partial x_2}\right)\frac{\partial}{\partial x_3} - \frac{\partial^2 k_{j+3}}{\partial x_2^2}\frac{\partial}{\partial y_j},
\end{aligned}$$

and

$$[\mathrm{ad}_F G_1, \mathrm{ad}_F^2 G_2] = -\frac{\partial^2 k_j}{\partial x_1 \partial x_2}\frac{\partial}{\partial x_j} - \left(\frac{\partial^2 k_3}{\partial x_1 \partial x_2} - \frac{1}{2}\frac{\partial k_1}{\partial x_1} + \frac{1}{2}\frac{\partial k_2}{\partial x_2}\right)\frac{\partial}{\partial x_3} - \frac{\partial^2 k_{j+3}}{\partial x_1 \partial x_2}\frac{\partial}{\partial y_j}.$$

The condition $[\mathrm{ad}_F G_i, \mathrm{ad}_F^2 G_i] = 0$, $i = 1,2$, implies

$$\frac{\partial^2 k_i}{\partial x_1^2} = \frac{\partial^2 k_{i+3}}{\partial x_1^2} = 0, \quad \frac{\partial^2 k_i}{\partial x_2^2} = \frac{\partial^2 k_{i+3}}{\partial x_2^2} = 0, \quad i = 1,2, \quad \text{and} \tag{15}$$

$$\frac{\partial^2 k_3}{\partial x_1^2} = -\frac{\partial k_2}{\partial x_1}, \qquad \frac{\partial^2 k_3}{\partial x_2^2} = \frac{\partial k_1}{\partial x_2}. \tag{16}$$

From the condition $[\mathrm{ad}_F G_i, \mathrm{ad}_F^2 G_j] = 0$, $i,j = 1,2$, $i \neq j$, we get

$$\frac{\partial^2 k_i}{\partial x_1 \partial x_2} = \frac{\partial^2 k_{i+3}}{\partial x_1 \partial x_2} = 0, \quad i = 1,2, \quad \text{and} \tag{17}$$

$$\frac{\partial^2 k_3}{\partial x_1 \partial x_2} = \frac{1}{2}\frac{\partial k_1}{\partial x_1} - \frac{1}{2}\frac{\partial k_2}{\partial x_2}. \tag{18}$$

Conditions (15) and (17) give

$$k_r(x_1, x_2) = \alpha_r x_1 + \beta_r x_2 + \theta_r, \qquad \alpha_r, \beta_r, \theta_r \in \mathbb{R}, \quad r = 1,2,4,5. \tag{19}$$

We rewrite the drift vector field as

$$\begin{aligned}
F = (y_i + \alpha_i x_1 + \beta_i x_2 + \theta_i)\frac{\partial}{\partial x_i} + (-\frac{1}{2}x_2 y_1 + \frac{1}{2}x_1 y_2 + k_3)\frac{\partial}{\partial x_3}\\
+ (\alpha_{i+3} x_1 + \beta_{i+3} x_2 + \theta_{i+3})\frac{\partial}{\partial y_i}.
\end{aligned}$$

From condition (C4) and its invariance under coordinates transformations, we conclude that the transformed vector fields, still denoted by F, G_1 and G_2, satisfy

$$F(0) = 0 \quad \text{and} \quad \mathrm{ad}_F^2 G_1(0) = \mathrm{ad}_F^2 G_2(0) = 0,$$

from which we get, respectively, $\theta_r = 0$, and

$$\frac{\partial k_r}{\partial x_1}(0) = \frac{\partial k_r}{\partial x_2}(0) = 0, \quad r = 1,2,4,5. \tag{20}$$

Equalities (20) imply $\alpha_r = \beta_r = 0$, and we conclude that $k_r = 0,\ r = 1,2,4,5$. From (16) and (18) we get

$$\frac{\partial^2 k_3}{\partial x_j^2} = \frac{\partial^2 k_3}{\partial x_1 x_2} = 0,\ \ j = 1,2,$$

and thus

$$k_3 = v_3^1 x_1 + v_3^2 x_2 + v_3^3, \quad v_3^1, v_3^2, v_3^3 \in \mathbb{R}.$$

From $F(0) = 0$ we obtain $v_3^3 = 0$ and condition $\mathrm{ad}_F^2 G_j(0) = 0,\ j = 1,2$, yields

$$\frac{\partial k_3}{\partial x_1}(0) = \frac{\partial k_3}{\partial x_2}(0) = 0.$$

Thus $v_3^1 = v_3^2 = 0$, and we obtain $k_3 = 0$. The drift can now be written as

$$F = y_j \frac{\partial}{\partial x_j} + \frac{1}{2}(x_1 y_2 - x_2 y_1) \frac{\partial}{\partial x_3}.$$

Finally, the transformation of coordinates given by

$$\tilde{x}_3 = 2x_3,$$

transforms the system into the desired form. □

4.2 Proof of Theorem 5

Proof of Theorem 5. Necessity is obvious, by a direct calculation. We prove sufficiency by showing that there exists a sequence of diffeomorphisms that brings the system to the desired form. By convenience for the sequel, we shall consider the vector fields $-\mathrm{ad}_F G_1, -\mathrm{ad}_F G_2, -[F,[G_1,G_2]]$ instead of $\mathrm{ad}_F G_1, \mathrm{ad}_F G_2, [F,[G_1,G_2]]$. In view of conditions (C1) and (C2), we can conclude the existence of a diffeomorphism $\phi : U_{z_0} \subset \mathbb{R}^6 \to \mathbb{R}^6$, with U_{z_0} an open neighborhood of z_0 and $\phi(z) = (x,y) = (x_1,x_2,x_3,y_1,y_2,y_3)$, such that $\phi(z_0) = 0$ and

$$-\mathrm{ad}_{\tilde{F}} \tilde{G}_1 = \phi_*(-\mathrm{ad}_F G_1) = \frac{\partial}{\partial x_1}, \tag{21}$$

$$-\mathrm{ad}_{\tilde{F}} \tilde{G}_2 = \phi_*(-\mathrm{ad}_F G_2) = \frac{\partial}{\partial x_2}, \tag{22}$$

$$-\left[\tilde{F},[\tilde{G}_1,\tilde{G}_2]\right] = \phi_*(-[F,[G_1,G_2]]) = \frac{\partial}{\partial x_3}, \qquad \text{and} \tag{23}$$

$$\left[\tilde{G}_1,\tilde{G}_2\right] = \phi_*([G_1,G_2]) = \frac{\partial}{\partial y_3}. \tag{24}$$

In this system of local coordinates (x,y), we denote the drift as

$$\tilde{F} = f_s \frac{\partial}{\partial x_s} + f_{s+3} \frac{\partial}{\partial y_s}, \quad s = 1,2,3,$$

and the input vector fields as

$$\tilde{G}_1 = g_{s,1} \frac{\partial}{\partial x_s} + g_{s+3,1} \frac{\partial}{\partial y_s}, \quad \text{and} \quad \tilde{G}_2 = g_{s,2} \frac{\partial}{\partial x_s} + g_{s+3,2} \frac{\partial}{\partial y_s}, \quad s = 1,2,3,$$

where a summation is understood over the index s and the functions $f_s, f_{s+3}, g_{s,i}$ and $g_{s+3,i}$, $i = 1,2$ are smooth functions (of variables (x,y)) in a neighborhood of $0 \in \mathbb{R}^6$. For simplicity, we skip the "tilde" notation, and, in particular, denote the vector fields $\tilde{F}$ and $\tilde{G}_i$ simply by F and G_i, $i = 1,2$. By condition (C3)(i), we have, in particular, $\mathrm{ad}_F^2 G_1 = \mathrm{ad}_F^2 G_2 = 0$, which allows to conclude that the component functions f_s and f_{s+3} do not depend on variables x_1 and x_2. Moreover, we can also see that $\mathrm{ad}_F^2[G_1,G_2] = 0$. Indeed, by the Jacobi identity and condition (C3)(ii), we get

$$\begin{aligned} \mathrm{ad}_F[G_1,G_2] &= [F,[G_1,G_2]] = [\mathrm{ad}_F G_1, G_2] + [G_1, \mathrm{ad}_F G_2] \\ &= -[G_2, \mathrm{ad}_F G_1]] + [G_1, \mathrm{ad}_F G_2]] = 2[G_1, \mathrm{ad}_F G_2]], \end{aligned} \tag{25}$$

and

$$\begin{aligned} \mathrm{ad}_F^2[G_1,G_2] &= [F, \mathrm{ad}_F[G_1,G_2]] = 2[F,[G_1,\mathrm{ad}_F G_2]]] \\ &= 2\left([\mathrm{ad}_F G_1, \mathrm{ad}_F G_2] + \left[G_1, \mathrm{ad}_F^2 G_2\right]\right) = 0, \end{aligned}$$

where the last equality follows by conditions (C2) and (C3)(i). Therefore,

$$[F,[F,[G_1,G_2]]] = \left[F, \frac{\partial}{\partial x_3}\right] = 0,$$

that is, the component functions of F do not depend also on coordinate x_3. From equalities (23) and (24) we obtain

$$\frac{\partial}{\partial x_3} = [[G_1,G_2],F] = \left[\frac{\partial}{\partial y_3}, F\right],$$

implying that

$$\frac{\partial f_s}{\partial y_3}\frac{\partial}{\partial x_s} + \frac{\partial f_{s+3}}{\partial y_3}\frac{\partial}{\partial y_s} = \frac{\partial}{\partial x_3},$$

that is,

$$\frac{\partial f_l}{\partial y_3} = 0, \quad \text{for} \quad l = 1,2,4,5,6, \qquad \text{and} \qquad \frac{\partial f_3}{\partial y_3} = 1.$$

The drift can then be re-written as

$$F = f_r(y_1,y_2)\frac{\partial}{\partial x_r} + (y_3 + k_3(y_1,y_2))\frac{\partial}{\partial x_3} + f_{s+3}(y_1,y_2)\frac{\partial}{\partial y_s} \quad r = 1,2,\ s = 1,2,3,$$

with a summation understood over the indices r and s and the smooth functions $f_l(y_1,y_2)$, $l = 1,2,4,5,6$, and $k_3(y_1,y_2)$ depending only on variables y_1 and y_2. Also, from condition (C3)(i), we have $[G_i, \mathrm{ad}_F G_i] = 0$, $i = 1,2$, which implies that

$$[G_1, \frac{\partial}{\partial x_1}] = [G_2, \frac{\partial}{\partial x_2}] = 0. \tag{26}$$

Furthermore, we can show that also $[G_i,[F,[G_1,G_2]]] = 0$, $i = 1,2$. Indeed, the Jacobi identity together with condition (C3)(ii) give

$$\begin{aligned} [G_1,[F[G_1,G_2]]] &= [G_1,-2[G_2,\mathrm{ad}_F G_1]]] = -2[G_1,[G_2,\mathrm{ad}_F G_1]]] \\ &= -2([[G_1,G_2],\mathrm{ad}_F G_1] + [G_2,[G_1,\mathrm{ad}_F G_1]]]) = 0, \end{aligned}$$

$$\begin{aligned} [G_2,[F[G_1,G_2]]] &= [G_2,2[G_1,\mathrm{ad}_F G_2]]] = 2[G_2,[G_1,\mathrm{ad}_F G_2]]] \\ &= 2([[G_2,G_1],\mathrm{ad}_F G_2] + [G_1,[G_2,\mathrm{ad}_F G_2]]]) = 0, \end{aligned}$$

where the last equalities follow by conditions (C2) and (C3)(i). Therefore,

$$[G_i, \frac{\partial}{\partial x_3}] = 0, \quad i = 1,2. \tag{27}$$

Since $[G_i, [G_1, G_2]] = 0$, $i = 1,2$ (by conditions (C3)(i)), we also get

$$[G_i, \frac{\partial}{\partial y_3}] = 0, \quad i = 1,2. \tag{28}$$

Conditions (26), (27) and (28) allow to re-write the input vector fields as

$$\begin{aligned} G_1 &= g_{s,1}(x_2, y_1, y_2)\frac{\partial}{\partial x_s} + g_{s+3,1}(x_2, y_1, y_2)\frac{\partial}{\partial y_s}, \quad \text{and} \\ G_2 &= g_{s,2}(x_1, y_1, y_2)\frac{\partial}{\partial x_s} + g_{s+3,2}(x_1, y_1, y_2)\frac{\partial}{\partial y_s}, \end{aligned}$$

with, as usual, a summation understood over the index $s = 1,2,3$. From equality (25), we obtain

$$\frac{\partial}{\partial x_3} = \left[\frac{\partial}{\partial x_1}, G_2\right] - \left[\frac{\partial}{\partial x_2}, G_1\right]. \tag{29}$$

Condition (C3)(ii) says that $[G_1, \mathrm{ad}_F G_2] = -[G_2, \mathrm{ad}_F G_1]$, that is

$$\left[\frac{\partial}{\partial x_2}, G_1\right] = -\left[\frac{\partial}{\partial x_1}, G_2\right]. \tag{30}$$

Together, equalities (29) and (30) give

$$\frac{\partial}{\partial x_3} = 2\left[\frac{\partial}{\partial x_1}, G_2\right] = 2\frac{\partial g_{s,2}}{\partial x_1}\frac{\partial}{\partial x_s} + 2\frac{\partial g_{s+3,2}}{\partial x_1}\frac{\partial}{\partial y_s}, \tag{31}$$

and, consequently,

$$\frac{\partial g_{3,2}}{\partial x_1} = \frac{1}{2}, \quad \text{and} \quad \frac{\partial g_{l,2}}{\partial x_1} = 0, \quad l = 1,2,4,5,6,$$

that is, the functions $g_{l,2}$, for $l = 1,2,4,5,6$, depend only on variables y_1, y_2, and

$$g_{3,2} = \frac{1}{2}x_1 + \hat{g}_{3,2}(y_1, y_2),$$

with $\hat{g}_{3,2}$ a new smooth function of y_1, y_2. Now,

$$\left[\frac{\partial}{\partial x_2}, G_1\right] = \frac{\partial g_{s,1}}{\partial x_2}\frac{\partial}{\partial x_s} + \frac{\partial g_{s+3,1}}{\partial x_2}\frac{\partial}{\partial y_s},$$

and, by (30) and (31),

$$\left[\frac{\partial}{\partial x_2}, G_1\right] = -\left[\frac{\partial}{\partial x_1}, G_2\right] = -\frac{1}{2}\frac{\partial}{\partial x_3}.$$

It follows

$$\frac{\partial g_{s,1}}{\partial x_2}\frac{\partial}{\partial x_s}+\frac{\partial g_{s+3,1}}{\partial x_2}=-\frac{1}{2}\frac{\partial}{\partial x_3}.$$

Therefore

$$\frac{\partial g_{3,1}}{\partial x_2}=-\frac{1}{2}, \quad \text{and} \quad \frac{\partial g_{l,1}}{\partial x_2}=0, \quad l=1,2,4,5,6,$$

that is, the functions $g_{l,1}$, for $l=1,2,4,5,6$, depend only on variables y_1,y_2, and

$$g_{3,1}=-\frac{1}{2}x_2+\hat{g}_{3,1}(y_1,y_2),$$

with $\hat{g}_{3,1}$ a new smooth function of y_1,y_2. The above conclusions allow to re-write the input vector fields as

$$\begin{aligned} G_1 &= g_{r,1}(y_1,y_2)\frac{\partial}{\partial x_r}+\left(-\frac{1}{2}x_2+\hat{g}_{3,1}(y_1,y_2)\right)\frac{\partial}{\partial x_3}+g_{s+3,2}(y_1,y_2)\frac{\partial}{\partial y_s},\\ G_2 &= g_{r,2}(y_1,y_2)\frac{\partial}{\partial x_r}+\left(\frac{1}{2}x_1+\hat{g}_{3,2}(y_1,y_2)\right)\frac{\partial}{\partial x_3}+g_{s+3,2}(y_1,y_2)\frac{\partial}{\partial y_s}, \end{aligned}$$

with a summation understood over the indices $r=1,2$ and $s=1,2,3$.
Consider the distribution $\mathcal{D}:=\operatorname{span}\{\frac{\partial}{\partial x_3},\frac{\partial}{\partial y_3}\}$. We observe that

$$[\mathrm{ad}_F G_i,\mathcal{D}]=0 \quad \text{and} \quad [G_i,\mathcal{D}]=0, \quad i=1,2,$$

implying that the distribution $\mathcal{D}$ is invariant under the vector fields $\mathrm{ad}_F G_i$ and G_i, $i=1,2$. Thus, the projection $\tau:\mathbb{R}^6\to\mathbb{R}^4$, $\tau(x,y)=(x_1,x_2,y_1,y_2)$, is well defined. We consider the projections of the vector fields $-\mathrm{ad}_F G_i$ and G_i on $\mathbb{R}^4$, given by $\tau_*(-\mathrm{ad}_F G_i)$ and $\tau_* G_i$ (which are well defined by the above property of $\mathcal{D}$). We have

$$\begin{aligned} \tau_*(-\mathrm{ad}_F G_1) &= \frac{\partial}{\partial x_1}, \qquad \tau_* G_1 = g_{r,1}(x_2,y_1,y_2)\frac{\partial}{\partial x_r}+g_{r+3,1}(x_2,y_1,y_2)\frac{\partial}{\partial y_r},\\ \tau_*(-\mathrm{ad}_F G_2) &= \frac{\partial}{\partial x_2}, \qquad \tau_* G_2 = g_{r,2}(y_1,y_2)\frac{\partial}{\partial x_r}+g_{r+3,2}(y_1,y_2)\frac{\partial}{\partial y_r}, \end{aligned}$$

where, as usual, a sum is understood over the index $r=1,2$. The projected vector fields

$$\tau_*(-\mathrm{ad}_F G_1),\ \tau_*(-\mathrm{ad}_F G_2),\ \tau_* G_1,\ \tau_* G_2$$

are independent (by condition (C1)) and pairwise commuting[3]. Therefore, there exists on $\mathbb{R}^4$, a local diffeomorphism ψ such that $\psi(x_1,x_2,y_1,y_2)=(\bar{x}_1,\bar{x}_2,\bar{y}_1,\bar{y}_2)$, with $\psi(0)=0$ and

$$\psi_*(\tau_*(-\mathrm{ad}_F G_i))=\frac{\partial}{\partial \bar{x}_i} \quad \text{and} \quad \psi_*(\tau_* G_i)=\frac{\partial}{\partial \bar{y}_i}, \quad i=1,2. \tag{32}$$

[3] Recall the Lie bracket relations on $\mathbb{R}^6$:

$$[G_1,G_2]=\frac{\partial}{\partial y_3},\ [G_1,\mathrm{ad}_F G_1]=[G_2,\mathrm{ad}_F G_2]=[\mathrm{ad}_F G_1,\mathrm{ad}_F G_2]=0,\ [G_2,\mathrm{ad}_F G_1]=\frac{1}{2}\frac{\partial}{\partial x_3}.$$

The first equality of (32) implies that the diffeomorphism ψ must be of the form

$$\begin{aligned} \bar{x}_1 &= x_1 + \psi_1(y_1, y_2), & \bar{y}_1 &= \psi_3(y_1, y_2), \\ \bar{x}_2 &= x_2 + \psi_2(y_1, y_2), & \bar{y}_2 &= \psi_4(y_1, y_2), \end{aligned}$$

with ψ_i, $i = 1,2,3,4$, smooth functions of y_1 and y_2. The second equality of (32) implies

$$\begin{aligned} g_{1,1} + \frac{\partial \psi_1}{\partial y_1} g_{4,1} + \frac{\partial \psi_1}{\partial y_2} g_{5,1} &= 0, & g_{1,2} + \frac{\partial \psi_1}{\partial y_1} g_{4,2} + \frac{\partial \psi_1}{\partial y_2} g_{5,2} &= 0, \\ g_{2,1} + \frac{\partial \psi_2}{\partial y_1} g_{4,1} + \frac{\partial \psi_2}{\partial y_2} g_{5,1} &= 0, & g_{2,2} + \frac{\partial \psi_2}{\partial y_1} g_{4,2} + \frac{\partial \psi_2}{\partial y_2} g_{5,2} &= 0, \\ \frac{\partial \psi_3}{\partial y_1} g_{4,1} + \frac{\partial \psi_3}{\partial y_2} g_{5,1} &= 1, & \frac{\partial \psi_3}{\partial y_1} g_{4,2} + \frac{\partial \psi_3}{\partial y_2} g_{5,2} &= 0, \\ \frac{\partial \psi_4}{\partial y_1} g_{4,1} + \frac{\partial \psi_4}{\partial y_2} g_{5,1} &= 0, & \frac{\partial \psi_4}{\partial y_1} g_{4,2} + \frac{\partial \psi_4}{\partial y_2} g_{5,2} &= 1. \end{aligned} \tag{33}$$

Consider on $\mathbb{R}^6$, the coordinates transformation Ψ defined by

$$\begin{aligned} \bar{x}_1 &= x_1 + \psi_1(y_1, y_2), & \bar{y}_1 &= \psi_3(y_1, y_2), \\ \bar{x}_2 &= x_2 + \psi_2(y_1, y_2), & \bar{y}_2 &= \psi_4(y_1, y_2), \\ \bar{x}_3 &= x_3, & \bar{y}_3 &= y_3. \end{aligned} \tag{34}$$

The drift F is transformed via Ψ into

$$\bar{F} = \Psi_* F = \bar{f}_r \frac{\partial}{\partial \bar{x}_r} + (\bar{y}_3 + \bar{k}_3) \frac{\partial}{\partial \bar{x}_3} + (\bar{f}_{s+3}) \frac{\partial}{\partial \bar{y}_s}, \quad r = 1,2, \ s = 1,2,3,$$

with $\bar{f}_r$, $\bar{f}_{s+3}$ and $\bar{k}_3$ new smooth functions of variables $\bar{y}_1, \bar{y}_2$. In view of (33), the input vector fields become

$$\begin{aligned} \bar{G}_1 &= \Psi_* G_1 = \left(-\frac{1}{2}\bar{x}_2 + \bar{g}_{3,1}\right) \frac{\partial}{\partial \bar{x}_3} + \frac{\partial}{\partial \bar{y}_1} + \bar{g}_{6,1} \frac{\partial}{\partial \bar{y}_3}, \\ \bar{G}_2 &= \Psi_* G_2 = \left(\frac{1}{2}\bar{x}_1 + \bar{g}_{3,2}\right) \frac{\partial}{\partial \bar{x}_3} + \frac{\partial}{\partial \bar{y}_2} + \bar{g}_{6,2} \frac{\partial}{\partial \bar{y}_3}, \end{aligned}$$

with $\bar{g}_{3,i}$ and $\bar{g}_{6,i}$, $i = 1,2$ new smooth functions of variables $\bar{y}_1, \bar{y}_2$. Clearly, the vector fields $-\mathrm{ad}_{\bar{F}} \bar{G}_1$ and $-\mathrm{ad}_{\bar{F}} \bar{G}_2$ are preserved, that is,

$$-\mathrm{ad}_{\bar{F}} \bar{G}_1 = \Psi_*(-\mathrm{ad}_F G_1) = \frac{\partial}{\partial \bar{x}_1}, \tag{35}$$

$$-\mathrm{ad}_{\bar{F}} \bar{G}_2 = \Psi_*(-\mathrm{ad}_F G_2) = \frac{\partial}{\partial \bar{x}_2}. \tag{36}$$

For simplicity, we skip the “bar” notation. Doing so, the vector fields $\bar{F}, \bar{G}_1, \bar{G}_2$ will be denoted simply by F, G_1, G_2. In particular, the equality (35), now re-written as

$-\mathrm{ad}_F G_1 = [G_1, F] = \frac{\partial}{\partial x_1}$, implies

$$\frac{\partial f_2}{\partial y_1} = \frac{\partial f_4}{\partial y_1} = \frac{\partial f_5}{\partial y_1} = 0, \tag{37}$$

$$\frac{\partial f_1}{\partial y_1} = 1, \tag{38}$$

$$f_4 \frac{\partial g_{3,1}}{\partial y_1} + f_5 \frac{\partial g_{3,1}}{\partial y_2} = \frac{1}{2} f_2 + \frac{\partial k_3}{\partial y_1} + g_{6,1}, \tag{39}$$

$$f_4 \frac{\partial g_{6,1}}{\partial y_1} + f_5 \frac{\partial g_{6,1}}{\partial y_2} = \frac{\partial f_6}{\partial y_1}. \tag{40}$$

Similarly, the equality $-\mathrm{ad}_F G_2 = [G_2, F] = \frac{\partial}{\partial x_2}$, see (36), implies

$$\frac{\partial f_1}{\partial y_2} = \frac{\partial f_4}{\partial y_2} = \frac{\partial f_5}{\partial y_2} = 0, \tag{41}$$

$$\frac{\partial f_2}{\partial y_2} = 1, \tag{42}$$

$$f_4 \frac{\partial g_{3,2}}{\partial y_1} + f_5 \frac{\partial g_{3,2}}{\partial y_2} = -\frac{1}{2} f_1 + \frac{\partial k_3}{\partial y_2} + g_{6,2}, \tag{43}$$

$$f_4 \frac{\partial g_{6,2}}{\partial y_1} + f_5 \frac{\partial g_{6,2}}{\partial y_2} = \frac{\partial f_6}{\partial y_2}. \tag{44}$$

Clearly, equalities (38), (42), (37) and (41), imply

$$f_1 = y_1 + k_1, \quad f_2 = y_2 + k_2, \quad f_4 = k_4, \quad \text{and} \quad f_5 = k_5, \quad k_i \in \mathbb{R}, i = 1,2,4,5.$$

Condition (C4) implies that $F(0) = 0$, which in its turn implies $k_i = 0$, $i = 1,2,4,5$, and thus

$$f_1 = y_1, \quad f_2 = y_2, \quad f_4 = f_5 = 0.$$

Now, relations (39) and (43) become, respectively,

$$g_{6,1} = -(\frac{1}{2} y_2 + \frac{\partial k_3}{\partial y_1}) \quad \text{and} \quad g_{6,2} = \frac{1}{2} y_1 - \frac{\partial k_3}{\partial y_2},$$

whereas relations (40) and (44) lead to

$$\frac{\partial f_6}{\partial y_1} = 0 \quad \text{and} \quad \frac{\partial f_6}{\partial y_2} = 0.$$

These equalities, together with the condition $F(0) = 0$, allow to get $f_6 = 0$. We summarize the above conclusions, rewriting

$$\begin{aligned} F &= y_1 \frac{\partial}{\partial x_1} + y_2 \frac{\partial}{\partial x_2} + (y_3 + k_3(y_1, y_2)) \frac{\partial}{\partial x_3}, \\ G_1 &= \left(-\frac{1}{2} x_2 + g_{3,1}(y_1, y_2) \right) \frac{\partial}{\partial x_3} + \frac{\partial}{\partial y_1} - \left(\frac{1}{2} y_2 + \frac{\partial k_3}{\partial y_1}(y_1, y_2) \right) \frac{\partial}{\partial y_3}, \\ G_2 &= \left(\frac{1}{2} x_1 + g_{3,2}(y_1, y_2) \right) \frac{\partial}{\partial x_3} + \frac{\partial}{\partial y_2} + \left(\frac{1}{2} y_1 - \frac{\partial k_3}{\partial y_2}(y_1, y_2) \right) \frac{\partial}{\partial y_3}. \end{aligned}$$

Let us consider the transformation of coordinates defined by $\tilde{y}_3 = y_3 + k_3(y_1, y_2)$. We obtain

$$\dot{\tilde{y}}_3 = -\left(\frac{1}{2}y_2 + \frac{\partial k_3}{\partial y_1}\right)u_1 + \left(\frac{1}{2}y_1 - \frac{\partial k_3}{\partial y_2}\right)u_2 + \frac{\partial k_3}{\partial y_1}u_1 + \frac{\partial k_3}{\partial y_2}u_2 = -\frac{1}{2}y_2u_1 + \frac{1}{2}y_1u_2.$$

In particular, this transformation preserves the vector fields $\mathrm{ad}_F G_1$, $\mathrm{ad}_F G_2$, $[G_1, G_2]$ and $[F, [G_1, G_2]]$. The drift and input vector fields, still denoted by F, G_1 and G_2 are transformed into

$$\begin{aligned}
F &= y_1\frac{\partial}{\partial x_1} + y_2\frac{\partial}{\partial x_2} + \tilde{y}_3\frac{\partial}{\partial x_3}, \\
G_1 &= \left(-\frac{1}{2}x_2 + g_{3,1}(y_1, y_2)\right)\frac{\partial}{\partial x_3} + \frac{\partial}{\partial y_1} - \frac{1}{2}y_2\frac{\partial}{\partial \tilde{y}_3}, \\
G_2 &= \left(\frac{1}{2}x_1 + g_{3,2}(y_1, y_2)\right)\frac{\partial}{\partial x_3} + \frac{\partial}{\partial y_2} + \frac{1}{2}y_1\frac{\partial}{\partial \tilde{y}_3}.
\end{aligned}$$

We then obtain

$$\frac{\partial}{\partial \tilde{y}_3} = [G_1, G_2] = \left(\frac{\partial g_{3,2}}{\partial y_1} - \frac{\partial g_{3,1}}{\partial y_2}\right)\frac{\partial}{\partial x_3} + \frac{\partial}{\partial \tilde{y}_3},$$

implying

$$\frac{\partial g_{3,2}}{\partial y_1} = \frac{\partial g_{3,1}}{\partial y_2}. \tag{45}$$

Now consider the transformation of coordinates defined by $\tilde{x}_3 = x_3 + \varphi_1(y_1, y_2)$, with φ_1 a smooth function to be determined. It follows

$$\begin{aligned}
\dot{\tilde{x}}_3 &= \tilde{y}_3 + \left(-\frac{1}{2}x_2 + g_{3,1}\right)u_1 + \left(\frac{1}{2}x_1 + g_{3,2}\right)u_2 + \frac{\partial \varphi_1}{\partial y_1}u_1 + \frac{\partial \varphi_1}{\partial y_2}u_2 \\
&= \tilde{y}_3 - \frac{1}{2}x_2u_1 + \frac{1}{2}x_1u_2 + \left(g_{3,1} + \frac{\partial \varphi_1}{\partial y_1}\right)u_1 + \left(g_{3,2} + \frac{\partial \varphi_1}{\partial y_2}\right)u_2.
\end{aligned}$$

Condition (45) guarantees that the system of PDEs

$$\begin{cases} \frac{\partial \varphi_1}{\partial y_1} + g_{3,1} = 0 \\ \frac{\partial \varphi_1}{\partial y_2} + g_{3,2} = 0 \end{cases}$$

satisfies the integrability conditions, i.e., $\frac{\partial^2 \varphi_1}{\partial y_2 \partial y_1} = \frac{\partial^2 \varphi_1}{\partial y_1 \partial y_2}$. We thus conclude that it exists a solution $\varphi_1(x_1, x_2)$ such that

$$\dot{\tilde{x}}_3 = \tilde{y}_3 - \frac{1}{2}x_2u_1 + \frac{1}{2}x_1u_2.$$

Set $\hat{x}_3 = x_3 + \frac{1}{2}x_2y_1 - \frac{1}{2}x_1y_2$. We obtain

$$\dot{\hat{x}}_3 = \tilde{y}_3.$$

Finally, the transformation of coordinates given by

$$\bar{x}_3 = 2\hat{x}_3, \quad \bar{y}_3 = 2\tilde{y}_3$$

transforms the system into the desired form. □

5 Conclusions

In this paper, we consider nonholonomic normal systems with drift obtained by extending the Brockett nonholonomic integrator and geometrically characterise, under state equivalence, those systems. A set of necessary and sufficient conditions for the local state equivalence of a general control-affine system with two input to those systems are given. Such conditions are extremely simple to be checked since they only involve Lie brackets of the drift and input vector fields. Moreover, they lead to a constructive procedure allowing to get the diffeomorphism performing the state equivalence.

Bibliography

[1] A. P. Aguiar, J. P. Hespanha, and A. M. Pascoal. Switched seesaw control for the stabilization of underactuated vehicles. *Automatica*, 43(12):1997–2008, 2007. Cited p. 322.

[2] A. P. Aguiar and A. M. Pascoal. Stabilization of the extended nonholonomic double integrator via logic-based hybrid control. In *Proceedings of the 6th IFAC Symposium on Robot Control*, 2000. 6 pages (no pagination). Cited pp. 321, 322, and 327.

[3] A. P. Aguiar and A. M. Pascoal. Practical stabilization of the extended nonholonomic double integrator. In *Proceedings of the 10th Mediterranean Conference on Control and Automation*, 2002. Paper no. 424 (no pagination). Cited p. 322.

[4] N. P. I. Aneke, H. Nijmeijer, and A. G. de Jager. Trajectory tracking by cascaded backstepping control for a second-order nonholonomic mechanical system. In A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, editors, *Nonlinear Control in the Year 2000*, pages 35–49. Springer, 2000. Cited p. 323.

[5] N. P. I. Aneke, H. Nijmeijer, and A. G. de Jager. Tracking control of second-order chained form systems by cascaded backstepping. *International Journal of Robust and Nonlinear Control*, 13:95–115, 2003. Cited p. 323.

[6] H. Arai, K. Tanie, and N. Shiroma. Nonholonomic control of a three-DOF planar underactuated manipulator. *IEEE Transactions on Robotics and Automation*, 14:681–695, 1998. Cited p. 319.

[7] A. M. Bloch. *Nonholonomic Mechanics and Control*. Springer, 2003. Cited pp. 319 and 320.

[8] A. M. Bloch, J. E. Marsden, and D. V. Zenkov. Nonholonomic dynamics. *Notices of the AMS*, 52:324–333, 2005. Cited pp. 319 and 320.

[9] R. W. Brockett. Control theory and singular Riemannian geometry. In P. J. Hilton and G. S. Young, editors, *New Directions in Applied Mathematics*, pages 11–27. Springer, 1981. Cited pp. 319, 320, and 321.

[10] F. Bullo and A. D. Lewis. *Geometric Control of Mechanical Systems*. Springer, 2004. Cited pp. 323 and 326.

[11] J. Hauser, S. Sastry, and G. Meyer. Nonlinear control design for slightly non-minimum phase systems: Application to V/STOL aircraft. *Automatica*, 28(4):665–679, 1992. Cited p. 319.

[12] J.-I. Imura, K. Kobayashi, and T. Yoshikawa. Nonholonomic control of 3 link planar manipulator with a free joint. In *Proceedings of the 35th Conference on Decision and Control*, pages 1435–1436, 1996. Cited p. 323.

[13] A. Isidori. *Nonlinear Control Systems*. Springer, 3rd edition, 1995. Cited p. 325.

[14] I. Kolmanovsky and N. H. McClamroch. Developments in nonholonomic control problems. *IEEE Control Systems Magazine*, 15(6):20–36, 1995. Cited p. 319.

[15] P. Martin, S. Devasia, and B. Paden. A different look at output tracking: control of a VTOL aircraft. In *Proc. of the 33rd IEEE Conference on Decision and Control*, pages 2376–2381, 1994. Cited p. 319.

[16] K. Morgansen. Controllability and trajectory tracking for classes of cascade-form second-order nonholonomic systems. In *Proceedings of the 40th IEEE Conference on Decision and Control*, pages 3031–3036, 2001. Cited pp. 320, 322, and 323.

[17] K. Morgansen and R. W. Brockett. Nonholonomic control based on approximate inversion. In *Proceedings of the American Control Conference*, pages 3515–3519, 1999. Cited pp. 320, 322, and 323.

[18] R. M. Murray. Nilpotent bases for a class of nonintegrable distributions with applications to trajectory generation for nonholonomic systems. *Mathematics of Control, Signals, and Systems*, 7(1):58–75, 1994. Cited p. 321.

[19] R. M. Murray and S. S. Sastry. Nonholonomic motion planning: Steering using sinusoids. *IEEE Transactions on Automatic Control*, 38(5):700–716, 1993. Cited p. 321.

[20] H. Nijmeijer and A. J. van der Schaft. *Nonlinear Dynamical Control Systems*. Springer, 1990. Cited p. 325.

[21] R. Olfati-Saber. Global configuration stabilization for the VTOL aircraft with strong input coupling. *IEEE Transactions on Automatic Control*, 47(11):1949–1952, 2002. Cited pp. 319 and 323.

[22] G. Oriolo and Y. Nakamura. Control of mechanical systems with second-order nonholonomic constraints: Underactuated manipulators. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pages 2398–2403, 1991. Cited p. 319.

[23] C. Park, D. J. Scheeres, V. Guibout, and A. Bloch. Globally optimal feedback control law of the underactuated Heisenberg system by generating functions. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 2687–2692, 2006. Cited p. 320.

[24] W. Pasillas-Lépine and W. Respondek. On the geometry of Goursat structures. *ESAIM: Control, Optimisation and Calculus of Variations*, 6:119–181, 2001. Cited p. 321.

[25] W. Respondek. Introduction to geometric nonlinear control; linearization, observability and decoupling. In A. Agrachev, editor, *Mathematical Control Theory*, number 1 in ICTP Lecture Notes, pages 169–222. 2002. Cited p. 325.

[26] S. Ricardo and W. Respondek. A geometric characterisation of the second-order nonholonomic chained form. In *Proceedings of the 9th Portuguese Conference on Automatic Control*, 2010. Cited p. 324.

[27] S. Ricardo and W. Respondek. When is a control system mechanical? *Journal of Geometric Mechanics*, 2(3):265–302, 2010. Cited pp. 323, 326, and 328.

[28] S. Ricardo and W. Respondek. State equivalence to the second-order nonholonomic chained form. In *Special Issue on Geometric Control Theory, a tribute to Fátima Silva Leite on the occasion of her 60th anniversary*. Department of Mathematics, University of Coimbra, 2011. Cited p. 324.

[29] S. Ricardo and W. Respondek. Second-order nonholonomic mechanical control systems. preprint, 2012. Cited pp. 323, 324, 325, 326, and 328.

[30] H. J. Sussmann. Local controllability and motion planning for some classes of systems with drift. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pages 1110–1114, 1991. Cited p. 322.

[31] T. Yoshikawa, K. Kobayashi, and T. Watanabe. Design of a desirable trajectory and convergent control for 3-DOF manipulator with a nonholonomic constraint. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1805–1810, 2000. Cited p. 323.

Purification of low-dimensional quantum systems subject to Lindblad dissipation

Patrick Rooney
University of Michigan
Ann Arbor, Michigan, USA
dprooney@umich.edu

Anthony Bloch
University of Michigan
Ann Arbor, Michigan, USA
abloch@umich.edu

Abstract. We present a result on the purification of quantum Lindblad systems for two and three dimensions. In both cases, it is shown that a necessary condition for purifiability is that all Lindblad operators must share a common eigenvector. A further necessary condition for purifiability is that the subspace orthogonal to the common eigenvector must not be invariant under at least one of the Lindblad operators. In order to show this, we assume we can construct arbitrary Hamiltonian functions, and we project the Lindblad equation to a control equation over the interior of the space of unitary orbits.

1 Introduction

In recent decades, technological advances have allowed for greater precision in the manipulation of quantum systems, both in physics and chemistry. This has given rise to the application of mathematical control theory to quantum systems [11, 16, 17, 21, 25]. One important goal is the construction of quantum computers, which have the power to perform algorithms not accessible to conventional computers [9, 15]. A major experimental obstacle to any implementation of such a computer, however, is the decoherence of the system under influence of the environment, and so an important task is how to determine the controllability properties of a given system. While much progress has been made on the controllability of closed quantum systems [1, 7], work on open quantum systems has proven to be more challenging [2, 3, 8, 23].

The state of a closed quantum system is described by a norm-one vector in a complex Hilbert space that evolves according to the Schrödinger equation:

$$\frac{d}{dt}|\psi\rangle = -iH(t)|\psi(t)\rangle \tag{1}$$

where the Hamiltonian operator H must be Hermitian and we set $\hbar = 1$. An open quantum system, on the other hand, is described by a trace-one, positive-semidefinite operator ρ on the Hilbert space, called the density operator. A state $|\psi\rangle$ in the Hilbert space corresponds to the rank-one density operator $|\psi\rangle\langle\psi|$,[1] and is called a *pure* state. Other states can be formed from linear superpositions of pure states, and are called *mixed* states.

[1] Dirac's bra-ket notation consists of writing vectors as "kets" $|\psi\rangle$ and their linear duals as "bras" $\langle\psi|$. Inner and outer products are written $\langle\psi_1|\psi_2\rangle$ and $|\psi_1\rangle\langle\psi_2|$, respectively.

In the absence of interaction with the environment, ρ obeys the von Neumann equation, which is the extension of the Schrödinger equation:

$$\frac{d\rho}{dt} = [-iH(t), \rho]. \tag{2}$$

An important issue is that certain relevant quantities are invariant under the von Neumann equation. The density matrix $\rho(t)$ at any time t can be written $\rho(t) = U(t)\rho(0)U^{-1}(t)$, where $U(t)$ is unitary. Since matrices at different times are similar, the eigenvalues of ρ are constant. The *purity* of the system is defined to be $P = \sqrt{\mathrm{Tr}(\rho^2)}$, so that $P = 1$ for pure states, and $P < 1$ for all others. It is also invariant under the von Neumann equation [24], since it is the 2-norm of the vector of eigenvalues. This has implications for quantum control. Control variables typically appear only in the Hamiltonian (although there is research towards engineering dissipation super-operators as well [4, 5, 14]), and thus the control dynamics cannot directly alter the eigenvalues of ρ and cannot purify the state.

To model a system that interacts with the environment, the von Neumann equation must be modified. A density operator whose dynamics are both Markovian and time-invariant obeys the Lindblad equation [10, 13]:

$$\begin{aligned} \frac{d\rho}{dt} &= [-iH(t), \rho] + \mathcal{L}_D(\rho) \\ \mathcal{L}_D(\rho) &= \sum_{j=1}^{M} \left(L_j \rho L_j^\dagger - \frac{1}{2}\{L_j^\dagger L_j, \rho\} \right) \end{aligned} \tag{3}$$

where the Lindblad operators $\{L_j\}$can be taken to be traceless[2], and $\{\cdot,\cdot\}$ denotes the anti-commutator. The Lindbladian dynamics are responsible for the system moving between unitary orbits, and so the structure of the Lindblad operators will determine to what extent a system can be purified (the set of pure states, having $P = 1$, constitute one of the orbits).

If one makes the simplifying assumption that the control structure allows for the construction of any Hamiltonian, motion along the orbit can be made arbitrarily faster than the motion between orbits [12, 22]. More precisely, we assume that $H(t) = H_0 + \sum_{j=1}^{n^2-1} u_j(t) H_j$, where H_0 is the drift Hamiltonian, $u_j(t)$ are control functions from $[0, \infty)$ to $\mathbb{R}$ and $\{H_j\}$ are a basis of $\mathfrak{su}(n)$. The norm of Hamiltonian term is unbounded, whereas the norm of the Lindblad super-operator $\mathcal{L}_D(\rho)$ is bounded, so arbitrarily large controls $u_j(t)$ can steer states to others on the same orbit with arbitrary precision. With this in mind, we can treat the position along the orbit as a control. The natural question to ask is: where is the best place along a given orbit to increase purity?

We present a theorem that determines whether purification, starting from any density operator on a two or three dimensional Hilbert space, is possible under a certain set of Lindblad operators $\{L_j\}$. Purification here means that the purity $P(t) \to 1$ as $t \to \infty$. Achieving a purity of precisely $P = 1$ in finite time is not possible. In section two, we

[2] Adding a multiple of the identity to a Lindblad operator is equivalent to adding a term to the Hamiltonian[6].

will prove this theorem for two dimensions, which draws from our work in [19]. In section three, we prove the theorem for three dimensions.

Theorem 1. *A Lindblad system of the form* (3) *in two or three dimensions is purifiable only if:*

- *All Lindblad operators L_j share a common eigenvector $|\psi_c\rangle$.*
- *At least one of the Lindblad operators is not a multiple of a Hermitian operator.*

Dedication: This paper is submitted in honor of Uwe Helmke for his inspiring work and friendship over the years

2 Two-dimensional systems

The structure of orbits in two dimensions is straightforward. ρ has two eigenvalues $\lambda_1 \geq \lambda_2$. Since the set of spectra is in bijection with the set of orbits, and since the eigenvalues must add to 1, the set of orbits can be indexed by the variable $r = \lambda_1 - \lambda_2$, which takes values in the interval $[0,1]$. The orbit corresponding to $r = 0$ is a singleton (called the completely mixed state). At that point, $\rho = \frac{1}{2}I$, so:

$$\frac{d\rho}{dt} = \frac{1}{2}\sum_j [L_j, L_j^\dagger] =: \Omega_{CM}. \tag{4}$$

Ω_{CM} is Hermitian and traceless, so it has real eigenvalues $\pm\omega$. While $\frac{dr}{dt}$ in general does not exist at $r = 0$, one can find the one-sided derivative when $r(t) = 0$:

$$\lim_{\delta t \to 0+} \frac{r(t+\delta t) - r(t)}{\delta t} = \lim_{\delta t \to 0+} \frac{(2\omega\delta t + o(\delta t) - 0}{\delta t} = 2\omega. \tag{5}$$

All other orbits are homeomorphic to the sphere S^1, and $\frac{dr}{dt}$ will vary depending on the location along the sphere. If $|\psi_i\rangle$ is the eigenvector corresponding to eigenvalue λ_i, we can write $r = \langle\psi_1|\rho|\psi_1\rangle - \langle\psi_2|\rho|\psi_2\rangle$, so: We can write:

$$\begin{aligned}
\frac{dr}{dt} =& \langle\dot{\psi}_1|\rho|\psi_1\rangle - \langle\dot{\psi}_2|\rho|\psi_2\rangle + \langle\psi_1|\dot{\rho}|\psi_1\rangle - \langle\psi_2|\dot{\rho}|\psi_2\rangle \\
&+ \langle\psi_1|\rho|\dot{\psi}_1\rangle - \langle\psi_2|\rho|\dot{\psi}_2\rangle \\
=& \lambda_1(\langle\dot{\psi}_1|\psi_1\rangle + \langle\psi_1|\dot{\psi}_1\rangle) - \lambda_2(\langle\dot{\psi}_2|\psi_2\rangle + \langle\psi_2|\dot{\psi}_2\rangle) \\
&+ \langle\psi_1|\dot{\rho}|\psi_1\rangle - \langle\psi_2|\dot{\rho}|\psi_2\rangle \\
=& \langle\psi_1|\dot{\rho}|\psi_1\rangle - \langle\psi_2|\dot{\rho}|\psi_2\rangle \\
=& \langle\psi_1|\mathcal{L}_D(\rho)|\psi_1\rangle - \langle\psi_2|\mathcal{L}_D(\rho)|\psi_2\rangle
\end{aligned}$$

where in the second-to-last step, the normalization of the vectors makes the quantities in parentheses vanish. In the last step, we have used Equation (3) and the fact that $\langle\psi_i|[-iH,\rho]|\psi_i\rangle = 0$. Now we have

$$\begin{aligned}
\frac{dr}{dt} = \sum_j \Big(& \langle\psi_1|L_j\rho L_j^\dagger|\psi_1\rangle - \frac{1}{2}\langle\psi_1|L_j^\dagger L_j\rho|\psi_1\rangle - \frac{1}{2}\langle\psi_1|\rho L_j^\dagger L_j|\psi_2\rangle \\
& - \langle\psi_2|L_j\rho L_j^\dagger|\psi_2\rangle + \frac{1}{2}\langle\psi_2|L_j^\dagger L_j\rho|\psi_2\rangle + \frac{1}{2}\langle\psi_2|\rho L_j^\dagger L_j|\psi_2\rangle \Big)
\end{aligned}$$

If one writes $\rho = \lambda_1|\psi_1\rangle\langle\psi_1| + \lambda_2|\psi_2\rangle\langle\psi_2|$, then inserts the identity operator $I = |\psi_1\rangle\langle\psi_1| + |\psi_2\rangle\langle\psi_2|$ between L_j and $L_j^\dagger$, and abbreviates $w_{ij} := \sum_k |\langle\psi_i|L_k|\psi_j\rangle|^2$, one gets:

$$\frac{dr}{dt} = (w_{12} - w_{21}) - r(w_{12} + w_{21}). \tag{6}$$

To prove that the first condition in the theorem is necessary for purification, note that Equation (6) reduces to $\frac{dr}{dt} = -2w_{21} \leq 0$ when $r = 1$ (which is the set of pure orbits). It should be clear that for purification to take place, $\lim_{r(t)\to 1} \dot{r}(t) \geq 0$. It follows that purification requires that $w_{21} = 0$. Since w_{21} is the sum of terms that are individually non-negative, each must be zero. But each term is of the form $|\langle\psi_2|L_k|\psi_1\rangle|^2$, it follows that $|\psi_1\rangle$ is an eigenvector of L_k for each k. So if we are to purify the system, it must be to a pure state $|\psi_c\rangle\langle\psi_c|$ such that $|\psi_c\rangle$ is an eigenvector of all Lindblad operators.

The second condition requires one of the Lindblad operators to not be a multiple of a Hermitian operator. If that were true (in which case there would only be one linearly independent Lindblad operator), we would have $w_{12} = w_{21}$ for any choice of $|\psi_1\rangle$, $|\psi_2\rangle$, and therefore $\frac{dr}{dt} = -2rw_{21} \leq 0$. Without this condition, therefore, purity can never increase.

3 Three-dimensional systems

The set of orbits in three dimensions can be mapped to a 2-simplex as follows. Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of ρ. Define co-ordinates in $\mathbb{R}^2$ as follows:

$$\begin{aligned} x_1 &= \lambda_1 - \lambda_2 \\ x_2 &= \frac{1}{\sqrt{3}}(\lambda_1 + \lambda_2 - 2\lambda_3) = \frac{1 - 3\lambda_3}{\sqrt{3}}. \end{aligned} \tag{7}$$

The image of the map, which we shall call T, is shown in Fig. 1. The three vertices correspond to (1) the completely mixed state (where $\lambda_1 = \lambda_2 = \lambda_3$), (2) the pure states (where $\lambda_1 = 1$, $\lambda_2 = \lambda_3 = 0$) and (3) the states completely mixed in a two-dimensional subspace, but pure with respect to the third (that is, $\lambda_1 = \lambda_2$ and $\lambda_3 = 0$). The three edges correspond to eigenvalue crossings ($\lambda_1 = \lambda_2$ or $\lambda_2 = \lambda_3$) or to $\lambda_3 = 0$. Also shown in the diagram are lines across which λ_1 is constant. These are lines where $x_1 + \frac{1}{\sqrt{3}}x_2$ are constant.

Note that in the interior of T, and along the top edge (but not its endpoints), the eigenvalues are distinct. Here the orbits are homemomorphic to $U(3)/[S^1 \times S^1 \times S^1]$ (see [20] for a discussion of the geometry of orbits), which is six-dimensional. On the side edges, as well as the two top vertices, the orbits are homeomorphic to $U(3)/[S^1 \times (U(2)]$, which is four-dimensional. The third vertex, the completely mixed state is a singleton.

When the eigenvalues are distinct we can write down differential equations analogous to Equation (6). We can write:

$$\begin{aligned} \frac{dx_1}{dt} &= \langle\psi_1|\mathcal{L}_D(\rho)|\psi_1\rangle - \langle\psi_2|\mathcal{L}_D(\rho)|\psi_2\rangle \\ \frac{dx_2}{dt} &= -\sqrt{3}\langle\psi_3|\mathcal{L}_D(\rho)|\psi_3\rangle. \end{aligned} \tag{8}$$

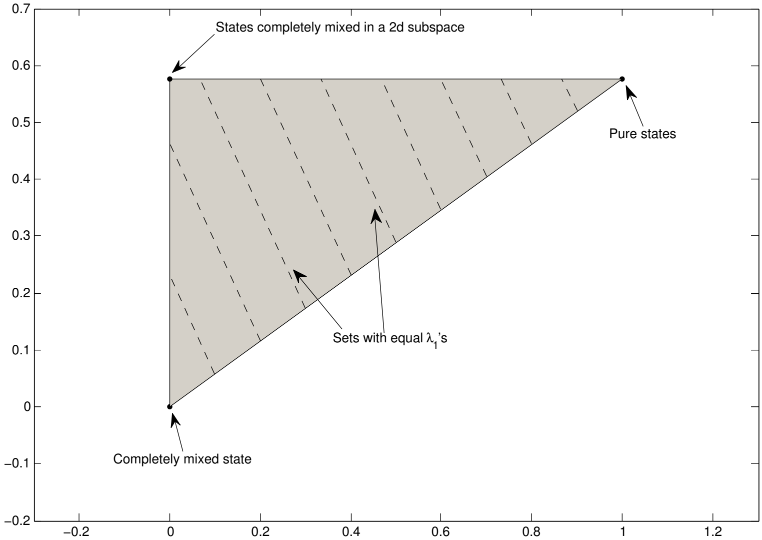


Figure 1: The set of orbits for three dimensions.

Upon substitution of $\mathcal{L}_D(\rho)$, as well as using

$$\rho = \frac{1}{3}I + \frac{x_1}{2}\left(|\psi_1\rangle\langle\psi_1| - |\psi_2\rangle\langle\psi_2|\right) + \frac{x_2}{2\sqrt{3}}\left(I - 3|\psi_3\rangle\langle\psi_3|\right), \tag{9}$$

and inserting the identity operator $I = \sum_k |\psi_k\rangle\langle\psi_k|$ between L_j and $L_j^\dagger$, we arrive at:

$$\begin{aligned}\frac{dx_1}{dt} &= \frac{1}{3}(2w_{12} - 2w_{21} + w_{13} - w_{23} + w_{32} - w_{31}) - \frac{x_1}{2}(2w_{12} + 2w_{21} + w_{32} + w_{31})\\ &\quad - \frac{x_2}{2\sqrt{3}}(2w_{21} - 2w_{12} + 2w_{13} - 2w_{23} + w_{31} - w_{32})\\ \frac{dx_2}{dt} &= \frac{1}{\sqrt{3}}(w_{13} + w_{23} - w_{32} - w_{31}) - \frac{\sqrt{3}x_1}{2}(w_{31} - w_{32})\\ &\quad - \frac{x_2}{2}(2w_{13} + 2w_{23} + w_{31} + w_{32}).\end{aligned} \tag{10}$$

At the orbit of pure states, where the co-ordinates are $(x_1, x_2) = (1, \frac{1}{\sqrt{3}})$, the velocities become:

$$\frac{dx_1}{dt} = -2w_{21} - w_{31}, \qquad \frac{dx_2}{dt} = -\sqrt{3}w_{31}. \tag{11}$$

In particular, $\frac{d\lambda_1}{dt} = \frac{dx_1}{dt} + \frac{1}{\sqrt{3}}\frac{dx_2}{dt} = -2(w_{21} + w_{31})$. Clearly, $\sup_{\lambda_1=1}\frac{d\lambda_1}{dt} = 0$ for purification, and so we must have $w_{21} = w_{31} = 0$. This can only be true if $|\psi_1\rangle$ is a common eigenvector $|\psi_c\rangle$ of all Lindblad operators, which satisfies the first condition.

To show the second condition, note, using Equation (10), that:

$$\begin{aligned}\frac{d\lambda_1}{dt} &= \frac{1}{2}\frac{dx_1}{dt} + \frac{1}{2\sqrt{3}}\frac{dx_2}{dt} \\ &= \frac{1}{3}\left(w_{12} - w21 + w_{13} - w_{31}\right) \\ &- \frac{x_1}{2}\left(w_{12} + w_{21} + 2w_{31}\right) - \frac{x_2}{2\sqrt{3}}\left(w_{21} - w_{12} + 2w_{13}\right).\end{aligned} \tag{12}$$

If the Lindblad operators are all multiples of Hermitian operators, we have $w_{12} = w_{21}$ and $w_{13} = w_{31}$. The above equation then reduces to

$$\frac{d\lambda_1}{dt} = -x_1\left(w_{21} + w_{31}\right) - \frac{x_2}{\sqrt{3}} w_{13}. \tag{13}$$

Regardless of what the eigenvectors of ρ, the above equation indicates purity cannot be increased. It follows that at least one of the Lindblad operators must be non-Hermitian.

4 Conclusion

We have provided a theorem describing necessary conditions for the purification of two and three-dimensional Lindblad systems. The Lindblad operators must all share a common eigenvector, which also serves as the destination pure state. Intuitively, Lindblad operators can be thought of as “jump" operators. Their eigenvectors are invariant under these jumps and thus do not decohere. The purification process then relies on finding a state that does not decohere under any of the Lindblad operators present.

A further necessary condition for purification is that one of the Lindblad operators must not be a multiple of a Hermitian operator. In other words, the eigenvectors of this operators must not be orthogonal. Moreover, without this condition, purity can never be increased regardless of what the initial density operator is.

We conjecture both of these conditions hold for arbitrary dimensions (including infinite dimensions), as the reasoning used is quite straightforward. Whether there are additional *sufficient* conditions for purification remains an ongoing research interest. More details on this work and the projection of Lindblad equation onto the space of orbits is in progress [18].

Acknowledgments

The authors would like to thank the National Science Foundation for support under grants DMS-090656 and DMS-1207683.

Bibliography

[1] F. Albertini and D. D’Alessandro. Notions of controllability for bilinear multilevel quantum systems. *IEEE Trans. Automatic Control*, 48(8):1399, 2003. Cited p. 345.

[2] C. Altafini. Controllability properties for finite dimensional quantum markovian master equations. *J. Math. Phys.*, 44(6):2357, 2003. Cited p. 345.

[3] C. Altafini. Coherent control of open quantum dynamical systems. *Phys. Rev. A*, 70(6):062321, 2004. Cited p. 345.

[4] D. Bacon et al. Universal simulation of Markovian quantum dynamics. *Phys. Rev. A*, 64:062302, 2001. Cited p. 346.

[5] J. Barreiro et al. An open-system quantum simulator with trapped ions. *Nature*, 470:486, 2011. Cited p. 346.

[6] H.-P. Breuer and F. Petruccione. *The Theory of Open Quantum Systems*. Oxford University Press, 2007. Cited p. 346.

[7] D. D'Alessandro. *Introduction to Quantum Control and Dynamics*. Chapman & Hall/CRC, 2008. Cited p. 345.

[8] G. Dirr, U. Helmke, I. Kurniawan, and T. Schulte-Herbrüggen. Lie-semigroup structures for reachability and control of open quantum systems: Kossakowski-Lindblad generators form Lie wedge to Markovian channels. *Rep. Math. Phys.*, 64:93, 2009. Cited p. 345.

[9] R. P. Feynman. Simulating physics with computers. *Int. J. Theo. Phys.*, 26(6):467, 1982. Cited p. 345.

[10] V. Gorini, A. Kossakowski, and E. Sudarshan. Completely positive dynamical semigroups of N-level systems. *J. Math. Phys.*, 17(5):821, 1976. Cited p. 346.

[11] G. M. Huang, T. J. Tarn, and J. W. Clark. On the controllability of quantum-mechanical systems. *J. Math. Phys.*, 24(11):2608, 1983. Cited p. 345.

[12] N. Khaneja, R. Brockett, and S. J. Glaser. Time optimal control in spin systems. *Phys. Rev. A*, 63:032308, 2001. Cited p. 346.

[13] G. Lindblad. On the generators of quantum dynamical semigroups. *Comm. Math. Phys.*, 48:119, 1976. Cited p. 346.

[14] S. Lloyd and L. Viola. Engineering quantum dynamics. *Phys. Rev. A*, 65:010101, 2001. Cited p. 346.

[15] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000. Cited p. 345.

[16] A. P. Peirce, M. A. Dahleh, and H. Rabitz. Optimal control of quantum-mechanical systems: Existence, numerical approximation, and applications. *Phys. Rev. A*, 37(12):030302, 1988. Cited p. 345.

[17] V. Ramakrishna, M. V. Salapaka, M. Dahleh, H. Rabitz, and A. Peirce. Controllability of molecular systems. *Phys. Rev. A*, 51(2):960, 1995. Cited p. 345.

[18] P. Rooney, A. Bloch, and C. Rangan. Projection of the controlled quantum Lindblad equation onto orbit space. In progress. Cited p. 350.

[19] P. Rooney, A. Bloch, and C. Rangan. Decoherence control and purification of two-dimensional quantum density matrices under Lindblad dissipation. *submitted to Phys. Rev. A*, 2012. arXiv:1201.0399v1 [quant-ph]. Cited p. 347.

[20] S. G. Schirmer, T. Zhang, and J. V. Leahy. Orbits of quantum states and geometry of bloch vectors for N-level systems. *J. Phys. A: Math. Gen.*, 37:1389, 2004. Cited p. 348.

[21] M. Shapiro and P. Brumer. Laser control of product quantum state populations in unimolecular reactions. *J. Phys. Chem.*, 84(7):4103, 1986. Cited p. 345.

[22] S. E. Sklarz, D. J. Tannor, and N. Khaneja. Optimal control of quantum dissipative dynamics: Analytic solution for cooling the three-level λ system. *Phys. Rev. A*, 69:053408, 2004. Cited p. 346.

[23] A. I. Solomon and S. G. Schirmer. Dissipative effects in multilevel systems. *J. Phys.: Conference Series*, 87:012015, 2007. Cited p. 345.

[24] D. J. Tannor and A. Bartana. On the interplay of control fields and spontaneous emission in laser cooling. *J. Phys. Chem. A*, 103:10359, 1999. Cited p. 346.

[25] D. J. Tannor and S. A. Rice. Control of selectivity of chemical reaction via control of wave packet evolution. *J. Chem. Phys.*, 83(10):5013, 1985. Cited p. 345.

Decoding of subspace codes, a problem of Schubert calculus over finite fields

Joachim Rosenthal
University of Zurich
Zurich, Switzerland
rosenthal@math.uzh.ch

Anna-Lena Trautmann
University of Zurich
Zurich, Switzerland
trautmann@math.uzh.ch

Abstract. Schubert calculus provides algebraic tools to solve enumerative problems. There have been several applied problems in systems theory, linear algebra and physics which were studied by means of Schubert calculus. The method is most powerful when the base field is algebraically closed. In this article we first review some of the successes Schubert calculus had in the past. Then we show how the problem of decoding of subspace codes used in random network coding can be formulated as a problem in Schubert calculus. Since for this application the base field has to be assumed to be a finite field new techniques will have to be developed in the future.

1 Introduction

Hermann Cäsar Hannibal Schubert (1848-1911) is considered the founder of enumerative geometry. He was a high school teacher in Hamburg, Germany. He studied questions of the type: Given four lines in projective three-space in general position, is there a line intersecting all given ones. This question can then be generalized to:

Problem 1. Given N k-dimensional subspaces $\mathcal{U}_i \subset \mathbb{C}^{k+m}$. Is there a subspace $\mathcal{V} \subset \mathbb{C}^{k+m}$ of complimentary dimension $m = \dim V$ such that

$$\mathcal{V} \bigcap \mathcal{U}_i \neq \{0\},\ i = 1,\dots,N. \tag{1}$$

Using a symbolic calculus he then came up with the following surprising result [21, 22]:

Theorem 2. *In case the subspaces $\mathcal{U}_i \subset \mathbb{C}^{k+m}, i = 1,\dots,N$ are in general position and in case $N = km$ there exist exactly*

$$d(k,m) = \frac{1!2!\cdots(k-1)!(km)!}{m!(m+1)!\cdots(m+k-1)!}. \tag{2}$$

different m dimensional subspaces $\mathcal{V} \subset \mathbb{C}^{k+m}$ which satisfy the intersection condition (1).

Note that two-dimensional subspaces in $\mathbb{C}^4$ describe lines in projective space $\mathbb{P}^3$ and Schubert hence claims in the case of four lines in three-space in general position that there are exactly $d(2,2) = 2$ lines intersecting all four given lines.

Schubert used in the derivation of Theorem 2 Poncelet's principle of preservation of numbers which was not considered a theorem of mathematics at the time. For this reason Schubert's results were not accepted by the mathematics community of the 19th century and Hilbert devoted the 15th of his famous 24 problems to the question if mathematicians can come up with rigorous techniques to prove or disprove the claims of Dr. Schubert. A rigorous verification of Theorem 2 was derived in the last century and we refer the interested reader to the survey article [14] by Kleiman, where the progress over time about Schubert calculus and the Hilbert problem 15 is described.

In the sequel we introduce the most important concepts from Schubert calculus.

Let $\mathbb{F}$ be an arbitrary field. Denote by $\mathrm{Grass}(k,n) = \mathrm{Grass}(k,\mathbb{F}^n)$ the Grassmann variety consisting of all k-dimensional subspaces of the vector space $\mathbb{F}^n$. $\mathrm{Grass}(k,n)$ can be embedded into projective space using the Plücker embedding:

$$\begin{aligned}\varphi : \mathrm{Grass}(k,\mathbb{F}^n) &\longrightarrow \mathbb{P}^{\binom{n}{k}-1}\\ \mathrm{span}(u_1,\ldots,u_k) &\longmapsto \mathbb{F}(u_1\wedge\ldots\wedge u_k).\end{aligned}$$

If one chooses a basis $\{e_1,\ldots,e_n\}$ of $\mathbb{F}^n$ and the corresponding canonical basis of $\Lambda^k\mathbb{F}^n$

$$\{e_{i_1}\wedge\ldots\wedge e_{i_k} \mid 1\le i_1<\ldots<i_k\le n\}$$

then one has an induced map of the coordinates. If U is a $k\times n$ matrix whose row space $\mathrm{rs}(U)$ describes the subspace $\mathcal{U} := \mathrm{span}(u_1,\ldots,u_k)$ and $U[i_1,\ldots,i_k]$ denotes the submatrix of U given by the columns $i_1,\ldots,i_k$, then one readily verifies that the Plücker embedding is given in terms of coordinates via:

$$\mathrm{rs}(U)\longmapsto[\det(U[1,...,k]):\det(U[1,...,k-1,k+1]):...:\det(U[n-k+1,...,n])].$$

The $k\times k$ minors $\det(U[i_1,\ldots,i_k])$ of the matrix U are called the *Plücker coordinates* of the subspace $\mathcal{U}$.

The image of this embedding describes indeed a variety and the defining equations are given by the so called "shuffle relations" (see e.g. [15, 19]). The shuffle relations are a set of quadratic equations in terms of the Plücker coordinates.

A flag $\mathcal{F}$ is a sequence of nested linear subspaces

$$\mathcal{F}:\ \{0\}\subset V_1\subset V_2\subset\ldots\subset V_n=\mathbb{F}^n$$

having the property that $\dim V_j = j$ for $j=1,\ldots,n$.

Denote by $\nu=(\nu_1,\ldots,\nu_k)$ an ordered index set satisfying

$$1\le\nu_1<\ldots<\nu_k\le n.$$

For every flag $\mathcal{F}$ one defines a Schubert variety

$$S(\nu;\mathcal{F}) := \{W\in\mathrm{Grass}(m,\mathbb{F}^n)\mid \dim(W\bigcap V_{\nu_i})\ge i \text{ for } i=1,\ldots,k\}. \tag{3}$$

The Schubert varieties are sub-varieties of the Grassmannian $\mathrm{Grass}(k,\mathbb{F}^n)$ and they contain a Zariski dense affine subset called Schubert cell and defined as:

$$C(\nu;\mathcal{F}) := \{W\in S(\nu;\mathcal{F})\mid \dim(W\bigcap V_{\nu_i-1}) = i-1;\text{ for } i=1,\ldots,k\}. \tag{4}$$

In terms of Plücker coordinates the defining equations of the Schubert variety $S(\nu;\mathcal{F})$ are given by the quadratic shuffle relations describing the Grassmann variety together with a set of linear equations (see [15]).

A fundamental question in Schubert calculus is now the following:

Problem 3. Given two Schubert varieties $S(\nu;\mathcal{F})$ and $S(\tilde{\nu};\tilde{\mathcal{F}})$. Describe as explicitly as possible the intersection variety

$$S(\nu;\mathcal{F})\cap S(\tilde{\nu};\tilde{\mathcal{F}}).$$

Schubert's Theorem 2 can actually also be formulated as an intersection problem of Schubert varieties. For this note that

$$\left\{\mathcal{V}\in \mathrm{Grass}(k,\mathbb{F}^{k+m}) \mid \mathcal{V}\bigcap\mathcal{U}_i \neq \{0\}\right\} \tag{5}$$

describes a Schubert variety with regard to some flag and the theorem then states that in the intersection of N Schubert varieties of above type one finds $d(k,m)$ m-dimensional subspaces as solutions in general.

In the case of an algebraically closed field one has rather precise information about this intersection variety. Topologically the intersection variety turns out to be a union of Schubert varieties of lower dimension and the multiplicities are governed by the Littlewood–Richardson rule [9]. When the field is not algebraically closed much less is known. There has been work done over the real numbers by Frank Sottile [25, 26]. Over general fields very little is known and we will show in this article that the decoding of subspace codes can be viewed as a Schubert calculus problem over some finite field. The following example illustrates the concepts.

Example 4. As a base field we take $\mathbb{F}=\mathbb{F}_2=\{0,1\}$ the binary field. Consider the Grassmannian $\mathrm{Grass}_2(2,\mathbb{F}^4)$ representing all lines in projective three-space $\mathbb{P}^3$. We would like to study Schubert's question in this situation: Given four lines in three-space, is there always a line intersecting all four given ones. Clearly there are many situations where the answer is affirmative, e.g. when the lines already intersect in some point. In general this is however not the case as we now demonstrate. Consider the following four lines in $\mathbb{P}^3$ represented as row spaces of the following four matrices:

$$\begin{bmatrix}1&0&0&0\\0&1&0&0\end{bmatrix},\begin{bmatrix}0&0&1&0\\0&0&0&1\end{bmatrix},\begin{bmatrix}1&1&0&1\\0&1&1&1\end{bmatrix},\begin{bmatrix}1&1&1&0\\1&0&1&1\end{bmatrix}.$$

We claim that there exists no line in projective three-space $\mathbb{P}^3$, i.e. no two-dimensional subspace in $\mathrm{Grass}_2(2,\mathbb{F}^4)$ intersecting all four given subspaces non-trivially. $\mathrm{Grass}_2(2,\mathbb{F}^4)$ is embedded in $\mathbb{P}^5$ via the Plücker embedding. Denote by

$$u_{i,j} := \det U[i,j], 1\le i<j\le 4$$

the Plücker coordinates of some subspace $\mathcal{U}\in\mathrm{Grass}_2(2,\mathbb{F}^4)$. The four lines impose the linear constraints:

$$\begin{aligned} u_{3,4} &= 0,\\ u_{1,2} &= 0,\\ u_{1,2}+u_{1,4}+u_{2,3}+u_{2,4}+u_{3,4} &= 0,\\ u_{1,2}+u_{1,3}+u_{1,4}+u_{2,3}+u_{3,4} &= 0. \end{aligned}$$

The points in $\mathbb{P}^5$ representing the image of $\mathrm{Grass}_2(2,\mathbb{F}^4)$ are described by one quadratic equation (shuffle relation):

$$u_{1,2}u_{3,4}+u_{1,3}u_{2,4}+u_{1,4}u_{2,3}=0.$$

Solving the 5 equations in the 6 unknowns results in one quadratic equation:

$$(u_{1,4})^2+u_{1,4}u_{2,3}+(u_{2,3})^2=0$$

which has no solutions over $\mathbb{F}_2$ in $\mathbb{P}^5$. Note that there are exactly $d(2,2)=2$ solutions over the algebraic closure as predicted by Schubert.

Readers who want to know more on the subject of Schubert calculus will find material in the survey article [15].

The paper is structured as follows: In Section 2 we present results which were derived by Schubert calculus. In Section 3 we introduce the main topic of this paper, namely subspace codes used in random network coding. In Section 4 we show that list decoding of random network codes is a problem of Schubert calculus over some finite field.

Many of the results we describe in this paper were derived by the first author in collaboration with Uwe Helmke. This collaboration was always very stimulating and the first author would like to thank Uwe Helmke for this continuing collaboration.

2 Results in systems theory and linear algebra derived by means of Schubert calculus

In the past Schubert calculus has been a very powerful tool for several problem areas in the applied sciences. In this section we review two such problem areas and we show to what extend Schubert calculus led to strong existence results and better understanding.

The pole placement problem

One of the most prominent problems in mathematical systems theory has been the pole placement problem. In the static situation the problem can be described as follows: Consider a discrete linear system

$$x(t+1)=Ax(t)+Bu(t),\ \ y(t)=Cx(t) \tag{6}$$

described by matrices A,B,C having size $n\times n$, $n\times m$ and $p\times n$ respectively. Consider a monic polynomial

$$\varphi(s):=s^n+a_{n-1}s^{n-1}+\cdots+a_1s+a_0\in\mathbb{F}[s]$$

of degree n having coefficients in the base field $\mathbb{F}$. In its simplest version the pole placement problem asks for the existence of a feedback law $u(t) = Ky(t)$ such that the resulting closed loop system

$$x(t+1) = (A+BKC)x(t) \tag{7}$$

has characteristic polynomial $\varphi(s)$.

At first glance this problem looks like a problem from matrix theory whose solution can be derived by means of linear algebra. Surprisingly, the problem is highly nonlinear and closely related to Schubert's Problem 1. This geometric connection was first realized in an interesting paper by Brockett and Byrnes [2] who showed that over the complex numbers arbitrary pole placement is generically possible as soon as $n \le mp$ and in case that the McMillan degree n is equal to mp then there are exactly $d(m,p)$ static feedback laws resulting in the closed loop characteristic polynomial $\varphi(s)$. The interested reader will find more details in a survey article by Byrnes [3].

The geometric insight one gained from the Grassmannian point of view was also helpful for deriving pole placement results over other base fields. Over the reals the most striking result was obtained by A. Wang in [31] where it was shown that arbitrary pole placement is possible with real compensators as soon as $n < mp$. Over a finite field some preliminary results were obtained by Gorla and the first author in [7].

U. Helmke in collaboration with X. Wang and the first author have been studying the pole placement problem in the situation when symmetries are involved [10]. This problem then leads to a Schubert type problem in the Lagrangian Grassmannian.

Sums of Hermitian matrices

Given Hermitian matrices $A_1,\dots,A_r \in \mathbb{C}^{n\times n}$ each with a fixed spectrum

$$\lambda_1(A_l) \ge \dots \ge \lambda_n(A_l), \quad l = 1,\dots,r \tag{8}$$

and arbitrary else. Is it possible to find then linear inequalities which describe the possible spectrum of the Hermitian matrix

$$A_{r+1} := A_1 + \cdots + A_r?$$

Questions of this type have a long history in operator theory and linear algebra. For example H. Weyl derived in 1912 the following famous inequality for any set of indices $1 \le i,j \le n$ with $1 \le i+j-1 \le n$:

$$\lambda_{i+j-1}(A_1 + A_2) \le \lambda_i(A_1) + \lambda_j(A_2). \tag{9}$$

In collaboration with U. Helmke the first author extended work by Johnson [13] and Thompson [27, 28] to derive a large set of eigenvalue inequalities. This was achieved through the use of Schubert calculus and we will say more in a moment. The obtained inequalities included in special cases not only the inequalities by H. Weyl but also the more extensive inequalities from Lidskii and Freede Thompson [28].

In order to make the connection to Schubert calculus we follow [9] and denote with $v_{1l},\ldots,v_{nl}$ the set of orthogonal eigenvectors of the Hermitian operator A_l, $l = 1,\ldots,r+1$.

Using these ordered set of eigenvectors one constructs for each Hermitian matrix A_l the flag:

$$\mathcal{F}_l: \quad \{0\} \subset V_{1l} \subset V_{2l} \subset \ldots \subset V_{nl} = \mathbb{C}^n \tag{10}$$

defined through the property:

$$V_{ml} := \operatorname{span}(v_{1l},\ldots,v_{ml}) \;\; m = 1,\ldots,n. \tag{11}$$

The connection to Schubert calculus is now established by the following result as it can be found in [9]. The theorem generalizes earlier results by Freede and Thompson [28].

Theorem 5. *Let $A_1,\ldots,A_r$ be complex Hermitian $n\times n$ matrices and denote with $\mathcal{F}_1,\ldots,\mathcal{F}_{r+1}$ the corresponding flags of eigenspaces defined by (11). Assume $A_{r+1} = A_1+\cdots+A_r$. and let $\underline{i}_l = (i_{1l},\ldots,i_{kl})$ be $r+1$ sequences of integers satisfying*

$$1 \le i_{1l} < \ldots < i_{kl} \le n, \;\; l = 1,\ldots,r+1. \tag{12}$$

Suppose the intersection of the $r+1$ Schubert subvarieties of $\operatorname{Grass}(k,\mathbb{C}^n)$ is non-empty, i.e.:

$$S(\underline{i}_1;\mathcal{F}_1)\bigcap\ldots\bigcap S(\underline{i}_{r+1};\mathcal{F}_{r+1}) \neq \varnothing. \tag{13}$$

Then the following matrix eigenvalue inequalities hold:

$$\sum_{j=1}^{k} \lambda_{n-i_{j,r+1}+1}(A_1+\cdots+A_r) \ge \sum_{l=1}^{r}\sum_{j=1}^{k} \lambda_{i_{jl}}(A_l) \tag{14}$$

$$\sum_{j=1}^{k} \lambda_{i_{j,r+1}}(A_1+\cdots+A_r) \le \sum_{l=1}^{r}\sum_{j=1}^{k} \lambda_{n-i_{jl}+1}(A_l). \tag{15}$$

In 1998 Klyachko could show that the inequalities coming from Schubert calculus as described in Theorem 5 are not only necessary but that they describe a Polytope of all possible inequalities. The interested reader will find Klyachko's result as well as much more in the survey article by Fulton [6].

A priori classical Schubert calculus provides very strong existence results. It is a different matter to derive effective numerical algorithms to compute the subspaces which satisfy the different Schubert conditions. For this reason Huber, Sottile and Sturmfels [12] developed effective numerical algorithms over the reals. As we will demonstrate in the next sections it would be very desirable to have effective numerical algorithms also in the case of Schubert type problems defined over some finite field.

3 Random network coding

In network coding one is looking at the transmission of information through a network with possibly several senders and several receivers. A lot of real-life applications of network coding can be found, e.g. data streaming over the Internet, where a source wants to send the same information to many receivers at the same time.

The network channel is represented by a directed graph with three different types of vertices, namely *sources*, i.e. vertices with no incoming edges, *sinks*, i.e. vertices with no outgoing edges, and *inner nodes*, i.e. vertices with incoming and outgoing edges. One assumes that at least one source and one sink exist. Under *linear* network coding the inner nodes are allowed to forward linear combinations of the incoming information vectors. The use of linear network coding possibly improves the transmission rate in comparison to just forwarding information at the inner nodes [1]. This can be illustrated in the example of the butterfly network: The source S wants to send the same information, a and b, to both receivers $R1$ and $R2$. Under forwarding every inner node forwards the incoming information and thus has to decide on either a or b (in this example on a) at the bottleneck vertex, marked above by x. Thus, $R1$ does not receive b. With linear network coding we allow the bottleneck vertex to send the sum of the two incoming informations, which allows both receivers to recover both a and b with a simple operation.

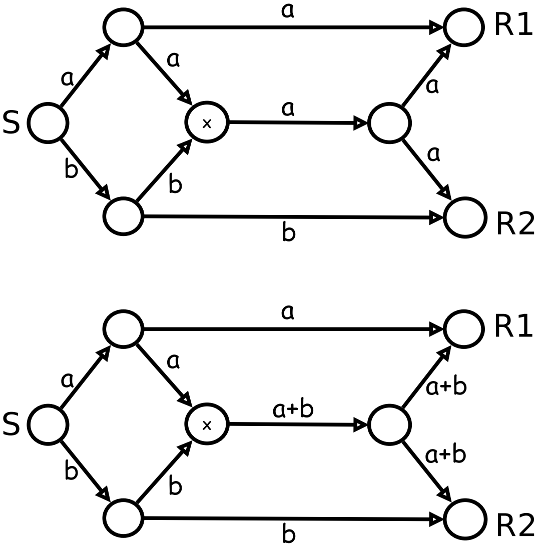


Figure 1: The butterfly network under the forwarding and the network coding model.

In this linear network coding setting, when the topology of the underlying network is unknown or time-varying, one speaks of *random* (linear) network coding. This setting was first studied in [11] and a mathematical model was introduced in [17], where the authors showed that it makes sense to use vector spaces instead of vectors over a finite field $\mathbb{F}_q$ as codewords. In this model the source injects a basis of the respective codeword into the network and the inner nodes forward a random linear combination of their incoming vectors. Therefore, each sink receives a linear combinations of the original vectors, which span the same vector space as the sent vectors, if no errors occurred during transmission.

In coding practice the base field is a finite field $\mathbb{F}_q$ having q elements, where q is a prime power. $\mathbb{F}_q^\times := \mathbb{F}_q \setminus \{0\}$ will denote the set of all invertible elements of $\mathbb{F}_q$. We will call the set of all subspaces of $\mathbb{F}_q^n$ the projective geometry of $\mathbb{F}_q^n$, denoted by $\mathcal{P}(q,n)$, and denote the Grassmannian $\mathrm{Grass}(k,\mathbb{F}_q^n)$ by $\mathrm{Grass}_q(k,n)$.

There are two types of errors that may occur during transmission, a decrease in dimension which is called an *erasure* and an increase in dimension, called an *insertion*. Assume $\mathcal{U} \in \mathcal{P}(q,n)$ was sent and erasures and insertions occurred during transmission, then the received word is of the type

$$\mathcal{R} = \bar{\mathcal{U}} \oplus \mathcal{E}$$

where $\bar{\mathcal{U}}$ is a subspace of $\mathcal{U}$ and $\mathcal{E} \in \mathcal{P}(q,n)$ is the error space. A random network coding channel in which both insertions and erasures can happen is called an *operator channel*.

In order to have a notion of decoding capability of some code a good metric is required on the set $\mathcal{P}(q,n)$: The *subspace distance* is a metric on $\mathcal{P}(q,n)$ given by

$$\begin{aligned} d_S(\mathcal{U},\mathcal{V}) &= \dim(\mathcal{U}+\mathcal{V}) - \dim(\mathcal{U}\cap\mathcal{V}) \\ &= \dim(\mathcal{U}) + \dim(\mathcal{V}) - 2\dim(\mathcal{U}\cap\mathcal{V}) \end{aligned}$$

for any $\mathcal{U},\mathcal{V} \in \mathcal{P}(q,n)$. Another metric on $\mathcal{P}(q,n)$ is the *injection distance*, defined as

$$d_I(\mathcal{U},\mathcal{V}) = \max\{\dim(\mathcal{U}),\dim(\mathcal{V})\} - \dim(\mathcal{U}\cap\mathcal{V}).$$

Note, that for $\mathcal{U},\mathcal{V} \in \mathrm{Grass}_q(k,n)$ it holds that $d_S(\mathcal{U},\mathcal{V}) = 2d_I(\mathcal{U},\mathcal{V})$. A *subspace code* $\mathcal{C}$ is simply a subset of $\mathcal{P}(q,n)$. If $\mathcal{C} \subseteq \mathrm{Grass}_q(k,n)$, we call it a *constant dimension code*. The minimum distance of a subspace code is defined in the usual way.

Different constructions of subspace codes have been studied, e.g. in [4, 5, 16–18, 20, 24, 30]. Some facts on isometry classes and automorphisms of these codes can be found in [29].

The set of all invertible $n\times n$-matrices with entries in $\mathbb{F}_q$, called the general linear group, is denoted by GL_n. Moreover, the set of all $k\times n$-matrices over $\mathbb{F}_q$ is denoted by $\mathbb{F}_q^{k\times n}$.

Let $U \in \mathbb{F}_q^{k\times n}$ be a matrix of rank k and

$$\mathcal{U} = \mathrm{rs}(U) := \text{row space}(U) \in \mathrm{Grass}_q(k,n).$$

One can notice that the row space is invariant under GL_k-multiplication from the left, i.e. for any $T \in GL_k$

$$\mathcal{U} = \mathrm{rs}(U) = \mathrm{rs}(TU).$$

Thus, there are several matrices that represent a given subspace. A unique representative of these matrices is the one in reduced row echelon form. Any $k \times n$-matrix can be transformed into reduced row echelon form by a $T \in GL_k$.

Given $U \in \mathbb{F}_q^{k\times n}$ of rank k, $\mathcal{U} \in \mathrm{Grass}_q(k,n)$ its row space and $A \in GL_n$, we define

$$\mathcal{U}A := \mathrm{rs}(UA).$$

Let $U, V \in \mathbb{F}_q^{k\times n}$ be matrices such that $\mathrm{rs}(U) = \mathrm{rs}(V)$. Then one readily verifies that $\mathrm{rs}(UA) = \mathrm{rs}(VA)$ for any $A \in GL_n$, hence the operation is well defined.

Decoding subspace codes

Given a subspace code $\mathcal{C} \subseteq \mathcal{P}(q,n)$ and a received codeword $\mathcal{R} \in \mathcal{P}(q,n)$, a *maximum likelihood decoder* decodes to a codeword $\mathcal{U} \in \mathcal{C}$ that maximizes the probability

$$P(\mathcal{R} \text{ received} \,|\, \mathcal{U} \text{ sent})$$

over all $\mathcal{U} \in \mathcal{C}$.

A *minimum distance decoder* chooses the closest codeword to the received word with respect to the subspace or injection distance. Let us assume that both the erasure and the insertion probability is less than some fixed ε. Then over an operator channel where the insertion probability is equal to the erasure probability, maximum likelihood decoding is equivalent to minimum distance decoding with respect to the subspace distance while in an adversarial model it is equivalent to minimum distance decoding with respect to the injection distance [23].

Assume the minimum (injection) distance of $\mathcal{C}$ is d, then if there exists $\mathcal{U} \in \mathcal{C}$ with $d_I(\mathcal{R},\mathcal{U}) \leq \frac{d-1}{2}$, then $\mathcal{U}$ is the unique closest codeword and the minimum distance decoder will always decode to $\mathcal{U}$.

Note, that a minimum subspace distance decoder is equivalent to a minimum injection distance decoder when $\mathcal{C}$ is a constant dimension code. Since we will investigate constant dimension codes in the remainder of this paper we will always use the injection distance. All results can then be carried over to the subspace distance.

A very important concept in coding theory is the problem of *list decoding* (see [8]). It is the goal of list decoding to come up with an algorithm which allows one to compute all code words which are within some distance of some received subspace.

For some $\mathcal{U} \in \mathcal{P}(q,n)$ we denote the ball of radius e with center $\mathcal{U}$ in $\mathcal{P}(q,n)$ by $B_e(\mathcal{U})$. If we want to describe the same ball inside $\mathrm{Grass}_q(k,n)$ we denote it by $B_e^k(\mathcal{U})$. Note that for a constant dimension code the ball $B_e^k(\mathcal{U})$ is nothing else than some Schubert variety of $\mathrm{Grass}_q(k,n)$.

Given a subspace code $\mathcal{C} \subseteq \mathcal{P}(q,n)$ and a received codeword $\mathcal{R} \in \mathcal{P}(q,n)$, a *list decoder with error bound e* outputs a list of codewords $\mathcal{U}_1,\dots,\mathcal{U}_m \in \mathcal{C}$ whose injection (resp. subspace) distance from $\mathcal{R}$ is at most e. In other words, the list is equal to the set

$$B_e(\mathcal{R}) \cap \mathcal{C}.$$

If $\mathcal{C}$ is a constant dimension code, then the output of the list decoder becomes $B_e^k(\mathcal{R}) \cap \mathcal{C}$.

4 List decoding in Plücker coordinates

As already mentioned before the balls of radius t (with respect to the injection distance) around some $\mathcal{U} \in \mathrm{Grass}_q(k,n)$ forms a Schubert variety over a finite field. In terms of Plücker coordinates it is possible to give explicit equations. For it we need the Bruhat order:

$$(i_1,\dots,i_k) \geq (j_1,\dots,j_k) \iff i_l \geq j_l \forall l \in \{1,\dots,k\}.$$

It is easy to compute the balls in the following special case.

Proposition 6. *Define* $\mathcal{U}_0 := \mathrm{rs}[\ I_{k\times k} \quad 0_{k\times n-k}\ ]$. *Then*

$$\begin{aligned} B_t^k(\mathcal{U}_0) = \{\mathcal{V} \in \mathrm{Grass}_q(k,n) \,|\varphi(\mathcal{V}) = [\mu_{1,\dots,k} : \dots : \mu_{n-k+1,\dots,n}], \\ \mu_{i_1,\dots,i_k} = 0\ \forall (i_1,\dots,i_k) \nleq (t+1,\dots,k,n-t+1,\dots,n)\} \end{aligned}$$

Proof. For $\mathcal{V}$ to be inside the ball it has to hold that

$$\begin{aligned} & d_I(\mathcal{U}_0,\mathcal{V}) \leq t \\ \iff & k - \dim(\mathcal{U}_0 \cap \mathcal{V}) \leq t \\ \iff & \dim(\mathcal{U}_0 \cap \mathcal{V}) \geq k-t \end{aligned}$$

i.e. $\mathcal{V}$ contains a $(k-t)$-dimensional subspace of $\mathcal{U}_0$. Therefore $\varphi(\mathcal{V})$ has to fulfill $\mu_{i_1,\dots,i_k} = 0$ if $(i_1,\dots,i_k) \nleq (t+1,\dots,k,n-t+1,\dots,n)$. □

With the knowledge of $B_t^k(\mathcal{U}_0)$ we can also express $B_t^k(\mathcal{U})$ for any $\mathcal{U} \in \mathrm{Grass}_q(k,n)$. For this note, that for any $\mathcal{U} \in \mathrm{Grass}_q(k,n)$ there exists an $A \in GL_n$ such that $\mathcal{U}_0 A = \mathcal{U}$. Moreover,

$$B_t^k(\mathcal{U}_0 A) = B_t^k(\mathcal{U}_0)A.$$

For simplifying the computations we define φ on GL_n, where we denote by $A_{i_1,\dots,i_k}$ the submatrix of A that consists of the rows $i_1,\dots,i_k$:

$$\begin{aligned} \varphi : GL_n &\longrightarrow GL_{\binom{n}{k}} \\ A &\longmapsto \begin{bmatrix} \det A_{1,\dots,k}[1,\dots,k] & \dots & \det A_{1,\dots,k}[n-k+1,\dots,n] \\ \vdots & & \vdots \\ \det A_{n-k+1,\dots,n}[1,\dots,k] & \dots & \det A_{n-k+1,\dots,n}[n-k+1,\dots,n] \end{bmatrix} \end{aligned}$$

Lemma 7. *Let* $\mathcal{U} \in \mathrm{Grass}_q(k,n)$ *and* $A \in GL_n$. *It holds that*

$$\varphi(\mathcal{U}A) = \varphi(\mathcal{U})\varphi(A).$$

Theorem 8. *Let* $\mathcal{U} = \mathcal{U}_0 A \in \mathrm{Grass}_q(k,n)$. *Then*

$$\begin{aligned} B_t^k(\mathcal{U}) &= B_t^k(\mathcal{U}_0 A) \\ &= \{\mathcal{V} \in \mathrm{Grass}_q(k,n) \,|\varphi(\mathcal{V})\varphi(A^{-1}) = [\mu_{1,\dots,k} : \dots : \mu_{n-k+1,\dots,n}], \\ &\qquad \mu_{i_1,\dots,i_k} = 0\ \forall (i_1,\dots,i_k) \nleq (t+1,\dots,k,n-t+1,\dots,n)\}. \end{aligned}$$

There are always several choices for $A \in GL_n$ such that $\mathcal{U}_0 A = \mathcal{U}$. Since $GL_{\binom{n}{k}}$ is very large we try to choose A as simple as possible. We will now explain one such construction:

1. The first k rows of A are equal to the matrix representation U of $\mathcal{U}$.
2. Find the pivot columns of U (assume that U is in reduced row echelon form).
3. Fill up the respective columns of A with zeros in the lower $n-k$ rows.
4. Fill up the remaining submatrix of size $n-k \times n-k$ with an identity matrix.

Then the inverse of A can be computed as follows:

1. Find a permutation $\sigma \in S_n$ that permutes the columns of A such that
$$\sigma(A) = \begin{bmatrix} I_k & U'' \\ 0 & I_{n-k} \end{bmatrix}.$$
2. Then the inverse of that matrix is
$$\sigma(A)^{-1} = \begin{bmatrix} I_k & -U'' \\ 0 & I_{n-k} \end{bmatrix}.$$
3. Apply σ on the rows of $\sigma(A)^{-1}$. The result is A^{-1}. One can easily see this if one represents σ by a matrix S. Then one gets $(SA)^{-1}S = A^{-1}S^{-1}S = A^{-1}$.

Example 9. In $\mathcal{G}_2(2,4)$ we want to find

$$B_1^2(\mathcal{U}) = \{\mathcal{V} \in \mathcal{G}_2(2,4) \mid \mathcal{V} \cap \mathcal{U} = 1\}$$

for

$$\mathcal{U} = \mathrm{rs}(U) = \mathrm{rs}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

We find the pivot columns $U[1,3]$ and build

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

Then we find the column permutation $\sigma = (23)$ such that

$$\sigma(A) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Now we can easily invert as described above and see that $\sigma(A)^{-1} = \sigma(A)$. We apply σ on the rows and get

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then

$$\varphi(A^{-1}) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

From Theorem 8 we know that

$$\begin{aligned} B_1^2(\mathcal{U}) &= \{\mathcal{V} \in \mathcal{G}_2(2,4) \mid \varphi(\mathcal{V})\varphi(A^{-1}) = [\mu_{1,2} : \dots : \mu_{3,4}],\ \mu_{i_1,i_2} = 0\ \forall (i_1,i_2) \nleq (2,4)\} \\ &= \{\mathcal{V} \in \mathcal{G}_2(2,4) \mid \varphi(\mathcal{V})\varphi(A^{-1}) = [\mu_{1,2} : \mu_{1,3} : \dots : \mu_{3,4}],\ \mu_{3,4} = 0\} \end{aligned}$$

Now let $\varphi(\mathcal{V}) = [v_{1,2} : v_{1,3} : v_{1,4} : v_{2,3} : v_{2,4} : v_{3,4}]$, then

$$\varphi(\mathcal{V})\varphi(A^{-1}) = [v_{1,3} : v_{1,2} : v_{1,3} + v_{1,4} : v_{2,3} : v_{3,4} : v_{2,3} + v_{2,4}]$$

and hence

$$\begin{aligned} B_1^2(\mathcal{U}) &= \{\mathcal{V} \in \mathcal{G}_2(2,4) \mid \varphi(\mathcal{V}) = [v_{1,2} : v_{1,3} : v_{1,4} : v_{2,3} : v_{2,4} : v_{3,4}],\ v_{2,3} + v_{2,4} = 0\} \\ &= \{\mathcal{V} \in \mathcal{G}_2(2,4) \mid \varphi(\mathcal{V}) = [v_{1,2} : v_{1,3} : v_{1,4} : v_{2,3} : v_{2,4} : v_{3,4}],\ v_{2,3} = v_{2,4}\}. \end{aligned}$$

Note, that we do not have to compute the whole matrix $\varphi(A^{-1})$ since in this case we only need the last column of it to find the equations that define $B_1^2(\mathcal{U})$.

5 Conclusion

The article explains the importance of Schubert calculus in various areas of systems theory and linear algebra. The strongest results in Schubert calculus require that the base field is algebraically closed. The problem of list decoding subspace codes is a problem of Schubert calculus where the underlying field is a finite field. It will be a topic of future research to come up with efficient algorithms to tackle this problem computationally.

Acknowledgments

Research partially supported by Swiss National Science Foundation Project no. 138080.

Bibliography

[1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung. Network information flow. *IEEE Transactions on Information Theory*, 46:1204–1216, 2000. Cited p. 359.

[2] R. W. Brockett and C. I. Byrnes. Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint. *IEEE Transanctions on Automatic Control*, 26:271–284, 1981. Cited p. 357.

[3] C. I. Byrnes. Pole assignment by output feedback. In H. Nijmeijer and J. M. Schumacher, editors, *Three Decades of Mathematical System Theory*, pages 31–78. Springer, 1989. Cited p. 357.

[4] T. Etzion and N. Silberstein. Error-correcting codes in projective spaces via rank-metric codes and Ferrers diagrams. *IEEE Transactions on Information Theory*, 55(7):2909–2919, 2009. Cited p. 360.

[5] T. Etzion and A. Vardy. Error-correcting codes in projective space. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 871–875, 2008. Cited p. 360.

[6] W. Fulton. Eigenvalues, invariant factors, highest weights, and Schubert calculus. *Bulletin of the AMS*, 37(3):209–249, 2000. Cited p. 358.

[7] E. Gorla and J. Rosenthal. Pole placement with fields of positive characteristic. In X. Hu, U. Jonsson, B. Wahlberg, and B. Ghosh, editors, *Three Decades of Progress in Control Sciences*, pages 215—231. Springer, 2010. Cited p. 357.

[8] V. Guruswami. *List Decoding of Error-Correcting Codes*, volume 3282 of *Lecture Notes in Computer Science*. Springer, 2004. Cited p. 361.

[9] U. Helmke and J. Rosenthal. Eigenvalue inequalities and Schubert calculus. *Mathematische Nachrichten*, 171:207–225, 1995. Cited pp. 355 and 358.

[10] U. Helmke, J. Rosenthal, and X. A. Wang. Output feedback pole assignment for transfer functions with symmetries. *SIAM Journal on Control and Optimization*, 45(5):1898–1914, 2006. Cited p. 357.

[11] T. Ho, R. Kötter, M. Medard, D. R. Karger, and M. Effros. The benefits of coding over routing in a randomized setting. *Proceedings of the IEEE International Symposium on Information Theory*, page 442, 2003. Cited p. 359.

[12] B. Huber, F. Sottile, and B. Sturmfels. Numerical Schubert calculus. *Journal of Symbolic Computation*, 26(6):767–788, 1998. Cited p. 358.

[13] S. Johnson. *The Schubert Calculus and Eigenvalue Inequalities for Sums of Hermitian Matrices*. PhD thesis, UC Santa Barbara, 1979. Cited p. 357.

[14] S. L. Kleiman. Problem 15: Rigorous foundations of Schubert's enumerative calculus. In *Proceedings of Symposia in Pure Mathematics*, volume 28, pages 445–482. AMS, 1976. Cited p. 354.

[15] S. L. Kleiman and D. Laksov. Schubert calculus. *American Mathematical Monthly*, 79:1061–1082, 1972. Cited pp. 354, 355, and 356.

[16] A. Kohnert and S. Kurz. Construction of large constant dimension codes with a prescribed minimum distance. In J. Calmet, W. Geiselmann, and J. Müller-Quade, editors, *MMICS*, volume 5393 of *Lecture Notes in Computer Science*, pages 31–42. Springer, 2008. Cited p. 360.

[17] R. Kötter and F. R. Kschischang. Coding for errors and erasures in random network coding. *IEEE Transactions on Information Theory*, 54(8):3579–3591, 2008. Cited pp. 359 and 360.

[18] F. Manganiello, E. Gorla, and J. Rosenthal. Spread codes and spread decoding in network coding. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 851–855, 2008. Cited p. 360.

[19] C. Procesi. A primer of invariant theory. Brandeis lecture notes, Brandeis University, 1982. Notes by G. Boffi. Cited p. 354.

[20] J. Rosenthal and A.-L. Trautmann. A complete characterization of irreducible cyclic orbit codes and their Plücker embedding. *Designs, Codes and Cryptography*. To appear, arXiv:1201.3825. Cited p. 360.

[21] H. Schubert. Anzahlbestimmung für lineare Räume beliebiger Dimension. *Acta Mathematica*, 8:97–118, 1886. Cited p. 353.

[22] H. Schubert. Beziehungen zwischen den linearen Räumen auferlegbaren charakteristischen Bedingungen. *Mathematische Annalen*, 38:598–602, 1891. Cited p. 353.

[23] D. Silva and F. R. Kschischang. On metrics for error correction in network coding. *IEEE Transactions on Information Theory*, 55(12):5479–5490, 2009. Cited p. 361.

[24] D. Silva, F. R. Kschischang, and R. Kötter. A rank-metric approach to error control in random network coding. *IEEE Transactions on Information Theory*, 54(9):3951–3967, 2008. Cited p. 360.

[25] F. Sottile. Enumerative geometry for the real Grassmannian of lines in projective space. *Duke Mathematical Journal*, 87(1):59–85, 1997. Cited p. 355.

[26] F. Sottile. Real Schubert calculus: Polynomial systems and a conjecture of Shapiro and Shapiro. *Experimental Mathematics*, 9(2):161–182, 2000. Cited p. 355.

[27] R. C. Thompson. The Schubert calculus and matrix spectral inequalities. *Linear Algebra and its Applications*, 117:176–179, 1989. Cited p. 357.

[28] R. C. Thompson and L. Freede. On the eigenvalues of a sum of Hermitian matrices. *Linear Algebra and its Applications*, 4:369–376, 1971. Cited pp. 357 and 358.

[29] A.-L. Trautmann. Isometry and automorphisms of constant dimension codes. arXiv:1205.5465, 2012. Cited p. 360.

[30] A.-L. Trautmann, F. Manganiello, M. Braun, and J. Rosenthal. Cyclic orbit codes. arXiv:1112.1238, 2011. Cited p. 360.

[31] X. Wang. Pole placement by static output feedback. *Journal of Mathematical Systems, Estimation, and Control*, 2(2):205–218, 1992. Cited p. 357.

Bilinear quantum control systems on Lie groups and Lie semigroups

Thomas Schulte-Herbrüggen
Technical University of Munich
Garching, Germany
tosh@ch.tum.de

Abstract. We set out to convey some of the Lie theoretical beauty of quantum control of bilinear systems as it has emerged within the last 15 years of contact, inspiration and exchange with the Helmke group. During this time, controllability criteria could be shifted from the well-known Lie-algebra rank condition to symmetry conditions in the branching diagrams for simple subalgebras of $\mathfrak{su}(N)$. Reachable sets of closed bilinear control systems were linked to the theory of C-numerical ranges. In coherently controlled open Markovian systems, the set of reachable directions (in physics known as Lindblad generators) form a Lie wedge generating a Lie semigroup (Markovian quantum map) that helps to approximate reachable sets in open systems. Once the reachable sets are known, gradient-flow algorithms have been devised to solve the abstract optimisation task on the reachable sets. They thus complement numerical algorithms that solve concrete optimal control problems on the manifold of admissible control amplitudes. The algorithmic tools have been presented in a unified programming framework.

How principles turn into practice has meanwhile emerged in a plethora of examples showing applications in solid-state devices, circuit-QED, ion traps, NV-centres in diamond, quantum dots, and in spin systems.

1 Introduction

This contribution is also meant as an invitation to the well-established community of mathematical systems theorists and *classical* control engineers to exchange with the vibrant developments in the field of *quantum* systems and control [19] in view of future technologies. These may be triggered by precise controls for, e.g., quantum simulation in order to improve the understanding of quantum phase transitions [49] between normal conducting and superconducting phases, or ferromagentic vs. antiferromagnetic phases to name just a few. Needless to say an operative thorough picture of these phenomena will booster advanced material design.

More precisely, an important issue in *quantum simulation* [1, 4, 18, 23, 29] is to manipulate all pertinent dynamical degrees of freedom of a system $\mathcal{A}$ of interest (which, however, all-too-often is experimentally not fully accessible) by a quantum system $\mathcal{B}$ that is in fact well controllable in practice and the dynamics of which are equivalent to those of $\mathcal{A}$. We will show how to characterise this situation algebraically in terms of quantum systems theory.

Besides the practical applications and implications, quantum systems should be of particular appeal to the (classical) control engineer, because nearly all sytems of interest boil down to the standard form of *bilinear control systems* [15, 22, 38, 59]

$$\dot{X}(t) = (A + \sum_j u_j B_j) X(t) \quad \text{with} \quad X_0 = X(0)\,. \tag{1}$$

Here one may take A, B as linear operators on the (finite-dimensional) Hilbert space of quantum states $|\psi(t)\rangle \in \mathcal{H}$. For n two-level spin-$\frac{1}{2}$ systems $\mathcal{H} = (\mathbb{C}^2)^{\otimes n}$. More precisely, A denotes the system or drift Hamiltonian iH_0, while the B_j are the control Hamiltonians iH_j governed by typically piece-wise constant control amplitudes $u_j \in \mathbb{R}$ (which need not be bounded). Thus Eqn. (1) captures all of the following important scenarios:

1. controlled Schrödinger equation
$$|\dot{\psi}(t)\rangle = -i(H_d + \sum_j u_j H_j)|\psi(t)\rangle \quad \text{with} \quad |\psi(0)\rangle = |\psi_0\rangle \tag{2}$$
2. quantum gate for closed system
$$\dot{U}(t) = -i(H_d + \sum_j u_j H_j) U(t) \quad \text{with} \quad U(0) = \mathbb{1} \tag{3}$$
3. quantum state in open quantum system
$$\dot{\rho}(t) = -(i\,\mathrm{ad}_{H_d} + i\sum_j u_j\,\mathrm{ad}_{H_j} + \Gamma_L)\,\mathrm{vec}(\rho(t)) \quad \text{with} \quad \rho(0) = \rho_0 \tag{4}$$
4. quantum map of open quantum system
$$\dot{F}(t) = -(i\,\mathrm{ad}_{H_d} + i\sum_j u_j\,\mathrm{ad}_{H_j} + \Gamma_L) F(t) \quad \text{with} \quad F(0) = \mathbb{1}\,, \tag{5}$$

where U denotes a unitary operator on $\mathcal{H}$ (e.g., used as quantum gate). F is the linear quantum map for open systems governed by the relaxation (super)operator Γ on $\mathcal{H} \otimes \mathcal{H}$ and ρ is the density operator (i.e. $\rho = \rho^\dagger \geq 0$ with $\mathrm{tr}\,\rho = 1$).

While the familiar *linear control systems* $\dot{x}(t) = Ax + Bu \quad \text{with} \quad x_0 = x(0)$ are fully controllable [32] if by $\mathrm{rank}[B, AB, A^2B, \ldots, A^{N-1}B] = N$ one has full rank, *bilinear systems* of Eqn. (1) are fully controllable over the compact connected Lie group $\mathbf{G}$ (generated by its Lie algebra $\mathfrak{g}$ via $\mathbf{G} = \langle \exp \mathfrak{g} \rangle$) whenever they satisfy the celebrated *Lie-algebra rank condition* [7, 8, 30, 31, 61]

$$\langle A, B_j \,|\, j = 1, 2, \ldots, m \rangle_{\mathrm{Lie}} = \mathfrak{g}\,. \tag{6}$$

Since in open systems (as in Eqns. (4) and (5)) $\mathfrak{g}$ is usually no longer compact, dissipative systems are obviously more subtle as will be seen in the concluding section.

For closed quantum systems of n spins-$\frac{1}{2}$, one has $\mathfrak{g} = \mathfrak{su}(N)$ with $N := 2^n$, which already shows that the state space and thereby the dynamic degrees of freedom in quantum systems scale *exponentially* in system size (as opposed to classical systems, where they scale linearly). Thus it is obvious that assessing controllability via an explicit Lie closure, though mathematically straight forward, becomes dramatically more tedious in quantum systems, and beyond seven qubits it is mostly prohibitive.

Overview

The contribution is organised as follows: In Sec. 2 we exemplify quantum systems theory of closed systems by shifting the paradigm of controllability from the celebrated Lie-algebra rank condition to symmetry conditions on the dynamic system algebra that are easier to assess in large quantum systems (2.1), while in (2.2) *quantum simulation* of fermionic, bosonic and spin systems is addressed by characterising their system algebras. Taking the system algebra as generator of the dynamic group, (2.3.) then establishes the *reachable sets* as subgroup orbits and relates expectation values of quantum dynamical observables to the mathematical theory of *C-numerical ranges* and their restriction to *relative C-numerical ranges* for subgroups of the unitary group. In (2.4) *examples of constrained optimisation* on relative C-numerical ranges are illustrated; (2.5) sketches concepts of *gradient flows on groups generated by systems algebras* solving these optimisations.

Sec. 3 is devoted to elements of system theory of open systems, where (3.1) draws a connection between *Lie semigroups* and *Markovian quantum channels* and (3.2) gives an outlook on how to address *reachable sets in open systems*.

In Sec. 4 the relation between abstract optimisation tasks on the reachble sets and *numerical optimal control* are outlined. The framework is matched to bilinear control systems as used in recent developments of quantum engineering.

Disclaimer: Unfortunately, here we cannot resort to a new category of mathematical proof that an Oberwolfach Meeting organised by Uwe Helmke in 2005 diagnosed to be a privilege to the Helmke group. It goes

Proof: Gunther Dirr could not find a counter example within five minutes. □

When we told Uwe, he took a breath, a pause and another breath to reply: '*Ja, das stimmt.*' — Uwe and Gunther, not only for this reason it has been a pleasure to share ideas with you. Thanks a lot and *ad multos annos!*

2 Quantum systems theory of closed systems

Hence here we will sketch a particularly simple and powerful alternative to assessing the controllability of quantum systems by way of easy-to-visualise *symmetry arguments*.

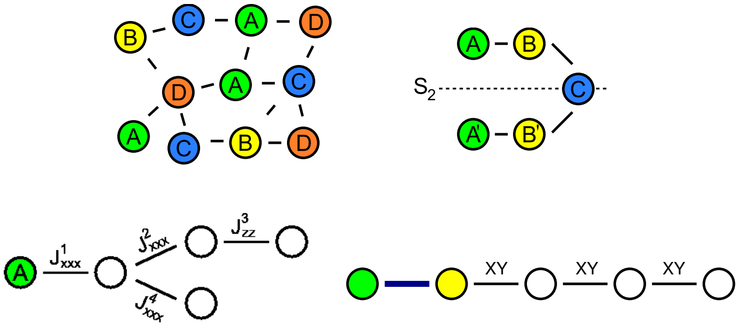


Figure 1: Graph representation of quantum dynamical control systems: vertices represent two-level systems (qubits), where common colour and letter code denotes joint local action, while the edges stand for pairwise coupling interactions. White vertices are qubits that are just coupled to the dynamic system without allowing to be controlled locally. The first and the last graph show no symmetries and their underlying control system is fully controllable. In contrast, the interior two graphs do exhibit symmetries: the left interior one has a mirror symmetry, while the right interior one leaves the Pauli operator σ_z on the upper terminal qubit invariant. These constants of the motion clearly preclude full controllability.

2.1 Symmetry conditions for controllability

To begin with, it pays to envisage the bilinear control systems by graphs in the way illustrated in Fig. 1: while the vertices represent *local* qubits as controlled by typical control Hamiltonians $B_j = iH_j$ (represented by Pauli matrices $\sigma_x, \sigma_y, \sigma_z$ acting on the qubit represented by the respective vertex), the edges stand for pair-wise *coupling* interactions as typically only occuring in the drift term $A = iH_0$ (represented by two-component tensor products of Pauli matrices as, e.g., $J_{zz} \cdot \sigma_z \otimes \sigma_z$ for the standard Ising interaction or $J_{XX} \cdot (\sigma_x \otimes \sigma_x + \sigma_y \otimes \sigma_y)$ for the so-called Heisenberg-XX interaction. Here the Pauli operators act on the two qubits connected by the respective edge).

As a central notion in the subsequent arguments, we characterise a quantum bilinear control system by its *system Lie algebra*, which results from the Lie closure of taking nested commutators (until no new linearly independent elements are generated)

$$\begin{aligned}\mathfrak{k} &:= \langle A, B_j \,|\, j = 1, 2, \dots, m\rangle_{\text{Lie}} \\ &= \langle iH_0, iH_j \,|\, j = 1, 2, \dots, m\rangle_{\text{Lie}} \subseteq \mathfrak{su}(N)\end{aligned} \tag{7}$$

as well as by its (potential) symmetries, i.e. the *centraliser* $\mathfrak{k}'$ in $\mathfrak{su}(N)$ to the system algebra $\mathfrak{k}$ collecting all terms that commute jointly with all Hamiltonian operators

$$\mathfrak{k}' := \{s \in \mathfrak{su}(N) | [s, H_\nu] = 0 \quad \forall \nu = 0; 1, 2, \dots, m\}. \tag{8}$$

If there are no symmetries, i.e. if the centraliser $\mathfrak{k}'$ is trivial (zero), then the system algebra $\mathfrak{k}$ is *irreducible*. This can easily be checked by determining the dimension of the nullspace (kernel) to the corresponding commutator superoperators (of dimension $N^2 \times N^2$)—so it boils down to solving a system of $m+1$ homogeneous equations in N^2 dimensions.

Lemma 1. *Let the system algebra $\mathfrak{k} \subseteq \mathfrak{su}(N)$ to a bilinear (qubit) control system Σ be a Lie subalgebra to the compact Lie algebra $\mathfrak{su}(N)$. Then one finds*

(1) if the centraliser $\mathfrak{k}'$ of $\mathfrak{k}$ in $\mathfrak{su}(N)$ is trivial, then $\mathfrak{k}$ is simple or semi-simple,

(2) if $\mathfrak{k}'$ is trivial and the coupling graph of the control system Σ is connected, then $\mathfrak{k}$ is simple.

Proof. (1) By compactness $\mathfrak{k}$ has a decomposition into its centre and a semi-simple part $\mathfrak{k} = \mathfrak{z}_\mathfrak{k} \oplus \mathfrak{ss}$ (see, e.g., [34] Corollary IV.4.25). As the centre $\mathfrak{z}_\mathfrak{k} = \mathfrak{k}' \cap \mathfrak{k}$ is trivial (zero), $\mathfrak{k}$ itself can only be semi-simple or simple. (2) Since $\mathfrak{k}$ must contain the Kronecker sum of local components, $\mathfrak{k} \supset su(2)_1 \widehat{\oplus} \mathfrak{su}(2)_2 \widehat{\oplus} \cdots \widehat{\oplus} \mathfrak{su}(2)_n$ where $A \widehat{\oplus} B := A \otimes \mathbb{1} + \mathbb{1} \otimes B$, and none of the partial sums is normalised whenever the pair interactions $iH_{jk} \in \mathfrak{su}(2)_j \otimes \mathfrak{su}(2)_k$ form a *connected* graph, the only ideals are trivial, hence $\mathfrak{k}$ has to be simple. (Details and generalisations to qu*d*its can be found in the Appendix to Ref. [65].) □

So a trivial centraliser plus a connected graph imply that the corresponding system algebra is *simple*. Since the largest possible Lie closure is $\mathfrak{su}(N)$, the system algebra $\mathfrak{k}$ of an irreducible connected qubit system has to be a (proper or improper) irreducible *simple subalgebra to* $\mathfrak{su}(N)$. By making heavy use of computer algebra, in Ref. [65] we have classified all these simple subalgebras of $\mathfrak{su}(N)$ for $N = 2^n$ with $n \leq 15$ qubits as summarised by the branching diagrams in Fig. 2 (see next page) thus extending the known results from $\mathfrak{su}(9)$ [43, 46] to $\mathfrak{su}(32768)$.

This figure also illustrates that every $\mathfrak{su}(N)$ with $N = 2^n$ has two canonical branches, a symplectic branch (shown in red) starting with $\mathfrak{sp}(N/2)$ and an orthogonal branch (blue) commencing with $\mathfrak{so}(N)$. Actually, for *odd* $n \leq 15$, these are the only ones (and we conjecture that this holds true even beyond 15 qubits). In contrast, for *even* n there are always subalgebras $\mathfrak{so}(2n+2)$ of unitary (spinor) type shown in black plus potential others (observe the instances of $\mathfrak{su}(4)$). — Clearly, if the (non-trivial) system algebra $\mathfrak{k}$ of a dynamic system in question can be ruled out to be on any of these three branches, then the corresponding control system is indeed *fully controllable* as will be shown next.

To this end, it is convenient to exclude the symplectic and orthogonal subalgebras in the first place. It is a task that can again be readily accomplished (after having made sure $\mathfrak{k}$ is irreducible) by determining the dimension of the joint null space (over S) to the following equations for each H_ν with $\nu = 0; 1, 2, \ldots, m$

$$SH_\nu^t + H_\nu S = 0 \tag{9}$$

or in its superoperator form

$$(H_\nu \otimes \mathbb{1} + \mathbb{1} \otimes H_\nu)\,\mathrm{vec}(S) = 0\,, \tag{10}$$

where by Schur's Lemma one must have $S\bar{S} = \pm \mathbb{1}$ [44]. If there is a non-trivial solution for the (+)-variant, then $\mathfrak{k} \subseteq \mathfrak{so}(N)$ is of orthogonal type, if there is for the (-)-variant, then $\mathfrak{k} \subseteq \mathfrak{sp}(N/2)$ is of symplectic type. So if the solution space for both cases (±) is

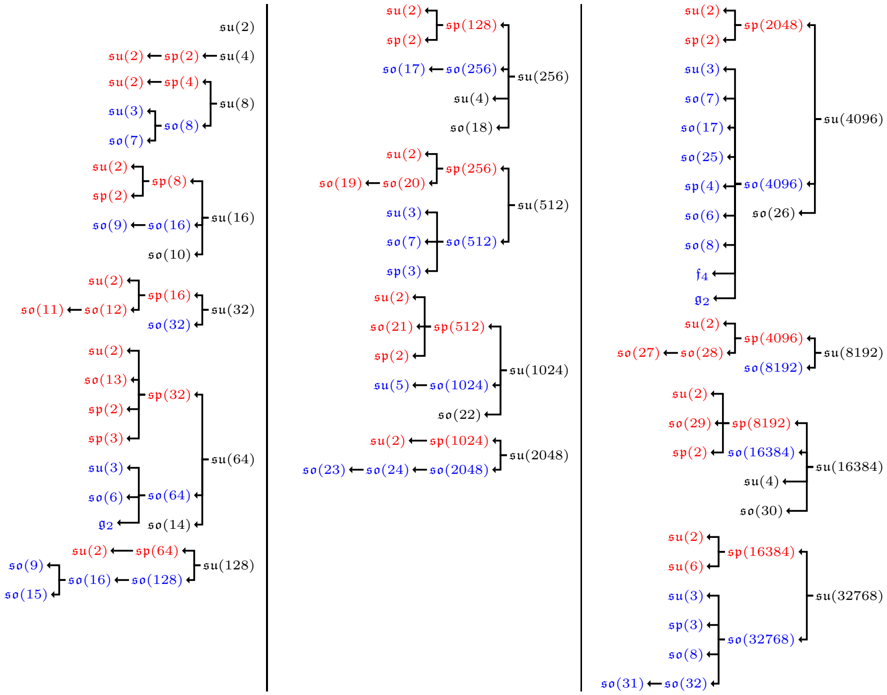


Figure 2: (Colour online) Branching diagrams showing all the irreducible simple subalgebras of $\mathfrak{su}(N)$ with $N := 2^n$ for n-qubit systems with $n \leq 15$ as given in [65]. Note that for *odd n* only the two canonical branches with orthogonal (blue) and symplectic (red) subalgebras occur. In contrast, for *even n* there are always unitary spinor-type subalgebras $\mathfrak{so}(2n+2)$ and in some instances $\mathfrak{su}(4)$. The orthogonal subalgebras are related to fermionic quantum systems, while the symplectic ones relate to compact versions of bosonic ones as described in the text and shown in Tabs. 1 and 2 below.

zero-dimensional (corresponding to the only solution being trivial) then $\mathfrak{k}$ is neither of orthogonal nor symplectic type. This can conveniently be decided by solving a homogeneous system of linear equations as done in Algorithm 3 of Ref. [65].

For odd $n \leq 15$, this does in fact already ensure full controllability, since only n even allows for unitary (spinor-type) simple subalgebras. Yet we conjecture that these findings also hold for all $n > 15$. Finally, for n even the spinor-type subalgebras may be excluded by the subsequent theorem of Ref. [65]. To prepare for it, observe that for $|S\rangle := \operatorname{vec}(S) \in \ker(H_\nu \otimes \mathbb{1} + \mathbb{1} \otimes H_\nu)$ hermiticity of $\{H_\nu\}$ and $|S\rangle\langle S|$ entails

$$\begin{aligned} (H_\nu \otimes \mathbb{1} + \mathbb{1} \otimes H_\nu)|S\rangle &= 0 \\ \Leftrightarrow (H_\nu \otimes \mathbb{1} + \mathbb{1} \otimes H_\nu)|S\rangle\langle S| &= 0 \\ \Leftrightarrow |S\rangle\langle S|(H_\nu \otimes \mathbb{1} + \mathbb{1} \otimes H_\nu) &= 0 \end{aligned}$$

hence the projector on $|S\rangle$ is in the commutant of the tensor square representation, i.e. $|S\rangle\langle S| \in (H_\nu \otimes \mathbb{1} + \mathbb{1} \otimes H_\nu)'$.

This motivates a closer analysis of the commutant of the tensor square representation

$$\Phi_{AB} := \{(iH_\nu \otimes \mathbb{1}_A + \mathbb{1}_B \otimes iH_\nu) \,|\, \nu = 0,1,\dots,m\}\,, \tag{11}$$

which provides a powerful single necessary and sufficient symmetry condition for full controllability:

Theorem 2 ([65])**.** *A bilinear control system governed by $\{iH_\nu \,|\, \nu = 0;1,2,\dots,m\}$ with system algebra $\mathfrak{k} := \langle iH_0, iH_j \,|\, j = 1,2,\dots,m\rangle_{\mathrm{Lie}}$ is fully controllable if and only if the joint centraliser to $\{(iH_\nu)\otimes\mathbb{1} + \mathbb{1}\otimes(iH_\nu) \,|\, \nu = 0;1,2,\dots,m\}$ in all complex matrices has dimension two.*

Proof. By Thm. 21 in Ref. [65], where we made use of Thm. 4.7 and Tab. 6 in the work of Dynkin [20]. □

To sum up, a bilinear n-qubit control system as in Eqn. (1) is fully controllable if and only if all of the following conditions are satisfied

(1) the system has no symmetries, i.e. $\mathfrak{k}'$ is trivial;

(2) the system has a connected coupling graph;

(3) the system algebra $\mathfrak{k}$ is neither of orthogonal nor of symplectic type, and finally

(4) the system algebra is not of any other type, in particular not of unitary spinor-type or of exceptional type ($\mathfrak{e}_6$).

While we gave a rigorous proof in Ref. [65] as already mentioned, the key arguments can easily be made intuitive as follows:

(1) symmetries would entail conserved entities (invariant one-parameter groups) thus precluding full controllability;

(2) coupling graphs with several connected components preclude that these components can be coherently coupled, which, however, is necessary for full controllability;

(3) orthogonal or symplectic subalgebras are proper subalgebras to $\mathfrak{su}(N)$ (for $N > 2$) and do not explore all dynamic degrees of freedom of $\mathfrak{su}(N)$, finally

(4) the same holds for unitary spinor-type or exceptional subalgebras ($\mathfrak{e}_6$) of $\mathfrak{su}(N)$.

By the branching diagrams in Fig. 2 it is immediately obvious: establishing full controllability boils down to ensuring the dynamic system is governed by a system algebra that is irreducible (no symmetries), and simple (connected coupling graph) and *top of the branch*. This shifts the paradigm from the *Lie-algebra rank-condition* to easily verifiable *symmetry conditions*, which can be checked using only the Hamiltonian generators.

2.2 Cross links to quantum simulation

Recall that fermionic quantum systems relate to orthogonal system algebras, while compact versions of bosonic ones (henceforth written as ‘bosonic’ for short) relate to symplectic system algebras. Then the link from controlled quantum systems to quantum simulation becomes obvious: the branching diagrams of Fig. 2 also illustrate that an (irreducible and connected) n-qubit quantum system is fully controllable if and only if it can simulate *both* ‘*bosonic*’ as well as *fermionic* systems.

This is because–clearly–a controlled bilinear dynamic system $\mathcal{A}$ can simulate another system $\mathcal{B}$ if and only if for the system algebras one has $\mathfrak{k}_A \supseteq \mathfrak{k}_B$. Moreover, given a fixed Hilbert space $\mathcal{H}$, $\mathcal{A}$ simulates $\mathcal{B}$ *efficiently* (i.e. with least state-space overhead in $\mathcal{H}$) if for any interlacing system $\mathcal{I}$ with system algebra $\mathfrak{k}_I$ satisfying $\mathfrak{k}_A \supseteq \mathfrak{k}_I \supseteq \mathfrak{k}_B$ one must have either $\mathfrak{k}_I = \mathfrak{k}_A$ or $\mathfrak{k}_I = \mathfrak{k}_B$ or (trivially) both.

For illustration, consider an n-qubit nearest-neighbour coupled Heisenberg-XX spin chain with single local controls. Then Tab. 1 shows that a single controllable qubit at one end suffices to simulate a fermionic system with quadratic interactions on n levels (governed by $\mathfrak{so}(2n+1)$), while local controls on both ends are required to simulate quadratic fermionic systems on $n+1$ levels with system algebra $\mathfrak{so}(2n+2)$. Most remarkably, if the controllable qubit is shifted to the second position, one gets dynamic degrees of freedom scaling *exponentially* in the number of qubits in the chain. This is by virtue of the system algebras $\mathfrak{so}(2^n)$ or $\mathfrak{sp}(2^{n-1})$, which most noticeably result depending on the length of the n-qubit chain: if $n \pmod 4 \in \{0,1\}$ the system is fermionic ($\mathfrak{so}(2^n)$), while for $n \pmod 4 \in \{2,3\}$ the system is ‘bosonic’ ($\mathfrak{sp}(2^{n-1})$) [65]. It is not until *two adjacent* qubits can be coherently controlled (as $\mathfrak{su}(4)$) that the Heisenberg-XX spin chains become fully controllable [11].

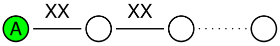
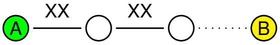
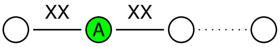
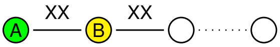
system type		fermionic	‘bosonic’	system algebra
n-spins-$\frac{1}{2}$	# levels	— coupling order —		
	n	2	–	$\mathfrak{so}(2n+1)$
	$n+1$	2	–	$\mathfrak{so}(2n+2)$
				
for n mod $4 \in \{0,1\}$	n	up to n	–	$\mathfrak{so}(2^n)$
for n mod $4 \in \{2,3\}$	n	–	up to n	$\mathfrak{sp}(2^{n-1})$
	n	up to n	up to n	$\mathfrak{su}(2^n)$

Table 1: Heisenberg-XX spin chains with a single control on one end (or both) can simulate either fermionic or ‘bosonic’ systems depending on the chain length as summarised in [65]. Local control over two adjacent qubits is required to make the system fully controllable (last row).

Moreover, Tab. 2 illustrates the power of classifying dynamic systems by symmetries and thereby in terms of their system Lie algebras: it turns out that joint controls on all the local qubits simultaneously suffice to even simulate effective three-body interactions (usually never occuring naturally), provided the Ising-ZZ coupling constants J_{zz} in odd-membered spin chains can be designed to have opposite signs on the two branches reaching out from the central spin.

Next we will illustrate how the system algebras $\mathfrak{k}$ obtained here by symmetry characterisation translate into reachable sets taking the form of group orbits $\mathcal{O}_{\mathbf{K}}(\rho_0)$ of initial states ρ_0. The orbits in turn can be projected onto detection operators to give the respective expectation values.

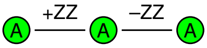
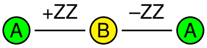
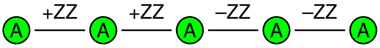
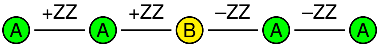
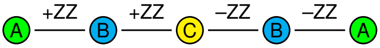
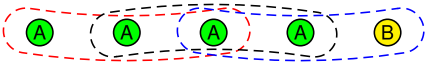
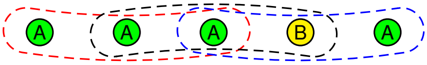
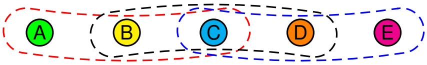
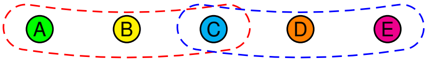
system type $n = 2k+1$ spins-$\frac{1}{2}$	# levels	'bosonic' coupling order	system algebra $\mathfrak{sp}(2^{n-1})$
	$n = 3$	up to 3	$\mathfrak{sp}(8/2)$
	—”—	—”—	—”—
	$n = 5$	up to 5	$\mathfrak{sp}(32/2)$
	—”—	—”—	—”—
	—”—	—”—	—”—
	—”—	—”—	—”—
	—”—	—”—	—”—
	—”—	—”—	—”—
	—”—	—”—	—”—

Table 2: Ising-ZZ spin chains with joint controls on all the qubits locally can simulate bosonic systems provided the coupling constants of the right and left branches leaving the central qubit have opposite signs as is also summarised in [65]. Note that even physically unavailable three-body interactions can be simulated by such systems. The system algebras given on the right specify that for a given chain length all systems are dynamically equivalent, which otherwise would be extremely difficult to analyse.

2.3 Reachable sets and expectation values of closed quantum systems: cross link to relative C-numerical ranges

Once the system algebra $\mathfrak{k} \subseteq \mathfrak{su}(N)$ of a bilinear control system Σ is determined, e.g., by symmetry characterisation as in the previous section, then the time evolution is brought about by the corresponding group $\mathbf{K} := \langle \exp \mathfrak{k} \rangle \subseteq SU(N)$. So the reachable set of an initial state ρ_0 is given by the corresponding group orbit $\mathcal{O}_{\mathbf{K}}(\rho_0)$

$$\text{Reach}(\rho_0) = \mathcal{O}_{\mathbf{K}}(\rho_0) := \{K\rho_0 K^\dagger \,|\, K \in \mathbf{K} \subseteq SU(N)\}\,. \tag{12}$$

In other words, the time evolution of the state ρ_0 is confined to $\rho(t) \in \mathcal{O}_{\mathbf{K}}(\rho_0)$ in the sense $\rho(t)$ solves the equation of motion (4) under Hamiltonian drift H_0 and controls H_j (in the absence of relaxation $\Gamma_L = 0$).

In quantum dynamics, the *expectation value* of a hermitian *observable* B, or more generally a detection operator C, is defined as projection of $\rho(t)$ onto B (or C) by way of the Hilbert-Schmidt scalar product

$$\langle C\rangle(t) := \text{tr}\{(C^\dagger \rho(t)\} = \text{tr}\{C^\dagger\, U(t)\rho(0)U(t)^\dagger\} \text{ where } U(t) \in \mathbf{K}\,. \tag{13}$$

Recall that the *field of values* of C is $W(C) := \{\langle u|Cu\rangle \,\big|\, u \in \mathbb{C}^N, \|u\| = 1\}$, while for $A,C \in \mathbb{C}^{N\times N}$ the *C-numerical range* of A is $W(C,A) := \{\text{tr}(C^\dagger\, UAU^\dagger)|U \in SU(N)\}$. So if $\rho(t)$ is a rank-1 projector (i.e. a pure state), the expectation value is an element of the field of values $\langle C\rangle(t) \in W(C)$, whereas for general $\rho(t)$ it is an element of the C-numerical range of $A \equiv \rho_0$, i.e. $\langle C\rangle(t) \in W(C,A)$. The latter is a star-shaped subset of the complex plane [12, 39] and it specialises to the form of a real line segment in case A and C are both hermitian.

As illustrated in the previous section, different quantum dynamical scenarios come with specific dynamical subgroups $\mathbf{K} \subsetneq SU(N)$ generated by the specific system algebras $\mathfrak{k}$. Typical examples for $\mathbf{K}$ include $SO(N)$ or $USp(N/2)$ or the subgroup of *local unitary operations* $SU_{\text{loc}}(2^n) = SU(2)^{\otimes n} := SU(2) \otimes SU(2) \otimes \cdots \otimes SU(2)$.

Consequently, in the instances of $\mathbf{K} \subsetneq SU(N)$, the admissible expectation values typically fill but a subset of $W(C,A)$, which hence motivates our definition of a restricted or *relative C-numerical range* [16, 52] as subgroup orbit $\mathcal{O}_{\mathbf{K}}(A)$ projected onto C

$$W_{\mathbf{K}}(C,A) := \{\text{tr}\,(C^\dagger KAK^\dagger)\,|\, K \in \mathbf{K} \subseteq SU(N)\} \subseteq W(C,A). \tag{14}$$

The particular case of $\mathbf{K} = SU(2)^{\otimes n}$ leads to what we call the *local C-numerical range*. If $\mathbf{K}$ is compact and connected, this obviously extends to $W(C,A)_{\mathbf{K}}$. However, note that although being connected, $W_{\mathbf{K}}(C,A)$ turns out to be in general neither star-shaped nor simply connected [16] in contrast to the usual C-numerical range [39].

The largest absolute value of the relative C-numerical range is defined as the *relative C-numerical radius*

$$r_{\mathbf{K}}(C,A) := \max_{K \in \mathbf{K}} |\,\text{tr}\{C^\dagger KAK^\dagger\}\,|\,; \tag{15}$$

it obviously plays a significant role for optimisations aiming at maximal expectation values.

With these stipulations, we will discuss recent applications of the local C-numerical range in quantum control.

2.4 Constrained optimisation and relative C-numerical ranges

In quantum control, one may face the problem to maximise the unitary transfer from matrices A to C subject to suppressing the transfer from A to D, or subject to leaving another state E invariant. For tackling those types of problems, in ref. [51] we asked for a 'constrained C-numerical range of A'

$$W(C,A)\big|_{\text{constraint}} := \big\{ \operatorname{tr}(UAU^\dagger C^\dagger) \,\big|\, \text{constraint} \big\} \subseteq W(C,A) \tag{16}$$

which form it takes and—in view of numerical optimisation—whether it is a connected set with a well-defined boundary. Connectedness is central to any numerical optimisation approach, because otherwise one would have to rely on initial conditions in the connected component of the (global) optimum.

Now the constrained C-numerical range of A is a compact and connected set in the complex plane, if the constraint can be fulfilled by restricting the full unitary group $SU(N)$ to a compact and connected subgroup $\mathbf{K} \subseteq SU(N)$. In this case, the constrained C-numerical range $W(C,A)|_{\text{constraint}}$ is identical to the relative C-numerical range $W_{\mathbf{K}}(C,A)$ and hence the constrained optimisation problem is solved within it, e.g., by the corresponding relative C-numerical radius $r_{\mathbf{K}}(C,A)$.

Example 3 (Constraint by Invariance)**.** The problem of maximising the transfer from A to C while leaving $E := \mu \mathbb{1} + \Omega$ with $\mu \in \mathbb{C}$ and $i\Omega \in \mathfrak{su}(N)$ invariant

$$\max_U |\operatorname{tr}\{UAU^\dagger C^\dagger\}| \quad \text{subject to} \quad UEU^\dagger = E \tag{17}$$

is straightforward, since the stabiliser group of E

$$\mathbf{K}_E := \{K \in U(N) \,|\, KEK^\dagger = E\} \tag{18}$$

is easy to come by: it is generated by the centraliser of Ω in $\mathfrak{u}(N)$

$$\mathfrak{k}_E := \{k \in \mathfrak{u}(N) \,|\, \operatorname{ad}_k(E) \equiv [k,E] = 0\} \,, \tag{19}$$

which (by Jacobi's identity) is easily seen to be Lie subalgebra $\mathfrak{k}_E \in \mathfrak{u}(N)$.

Remark 4. The stabiliser group of any $E \in \operatorname{Mat}_N(\mathbb{C}\,)$ in $U(N)$ is connected. This is in general not the case in $SU(N)$ as easily seen for $E := \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$. However, one can restrict the above optimisation to the connected component of the identity matrix in $SU(N)$ due to the invariance properties of the function $U \mapsto \operatorname{tr}\{UAU^\dagger C^\dagger\}$.

Hence a set of generators of $\mathfrak{k}_E$ may constructively be found via the kernel of the commutator map by solving a homogeneous linear system

$$\mathfrak{k}_E = \ker \operatorname{ad}_E \cap \mathfrak{su}(N) = \{k \in \mathfrak{su}(N) | (\mathbb{1} \otimes E - E^t \otimes \mathbb{1}) \operatorname{vec}(k) = 0\} \,. \tag{20}$$

So the optimisation problem of Eqn. (17) proceeds indeed over a constrained C-numerical range that is connected as it takes the form of a relative C-numerical range $W_{\mathbf{K}_E}(C,A)$ and the optimisation problem is solved by the relative C-numerical radius $r_{\mathbf{K}_E}(C,A)$. In Hermitian E, $\mathbf{K}_E$ includes a maximal torus group $\mathbf{T}$ of $U(N)$ since every Hermitian E can be chosen diagonal. Hence $\mathfrak{k}_E$ includes a maximal torus algebra $\mathfrak{t}$ with $\mathfrak{t} \subset \mathfrak{k}_E \subset \mathfrak{u}(N)$.

Obviously, the constraint of leaving E invariant while maximising the transfer from A to C only makes sense, if A and E do not share the same stabiliser group.

Example 5 (Pure-State Entanglement)**.** In terms of Euclidean geometry, maximising the real part in $W_{\text{loc}}(C,A)$ minimises the distance from C to the local unitary orbit $\mathcal{O}_{\text{loc}}(A)$. In Quantum Information Theory, the minimal distance has an interesting interpretation in the following setting: let A be an arbitrary rank-1 projector and let $C = \operatorname{diag}(1,0,\dots,0) \in \mathbb{C}^{N\times N}$. Thus in this case $W_{\text{loc}}(C,A)$ reduces to the *local field of values* $W_{\text{loc}}(A) = W_{SU(2)^{\otimes n}}(A)$. Then the minimial Euclidean distance

$$\Delta := \min_{K\in SU(2)^{\otimes n}} ||KAK^{\dagger} - C||_2 \tag{21}$$

is a measure of pure-state entanglement because it quantifies how far A is from the equivalence class of pure product states. It relates to the maximum real part of the local numerical range $W_{\text{loc}}(A)$ via

$$\begin{aligned} ||C - KAK^{\dagger}||_2^2 &= ||A||_2^2 + ||C||_2^2 - 2\operatorname{Re}\operatorname{tr}\{C^{\dagger}\, KAK^{\dagger}\} \\ &= 2 - 2\operatorname{Re}\operatorname{tr}\{C^{\dagger}\, KAK^{\dagger}\} \quad , \end{aligned} \tag{22}$$

where the last equality holds if also A is normalised to $||A||_2 = 1$. Note that the restriction to *local* unitaries is essential: when taken over the entire unitary group, the minimum distance would always vanish as soon as $||A||_2 = ||C||_2 = 1$.

The new concept of the relative (or restricted) C-numerical range has meanwhile become a popular tool, e.g., for analysing entanglement properties, see [24, 47] (and references therein).

2.5 Optimisation by gradient flows

First encounters with Uwe Helmke and his group were triggered by a pioneering paper [9], where Brockett introduced the idea of exploiting gradient flows on the orthogonal group for diagonalising real symmetric matrices and for sorting lists of eigenvalues. Soon these techniques were generalised to Riemannian manifolds, their mathematical and numerical details were worked out most prominently in the book by Helmke and Moore [28], where they turned out to be applicable to a broad range of optimisation tasks including eigenvalue and singular-value problems, principal component analysis, matrix least-squares matching problems, balanced matrix factorisations, and combinatorial optimisation, see also [6]. Our early application was to look for the corresponding gradient flow on the unitary orbit of quantum states [25].

Implementing a gradient method for optimisation on a smooth constrained manifold, such as an unitary orbit, via the Riemannian exponential map, inherently ensures that the discretised flow remains within the manifold. In this sense, gradient flows on manifolds are *intrinsic* optimisation methods [13], whereas *extrinsic* optimisations on an embedding space require in general non-linear projective techniques in order to stay on the constrained manifold. In particular, using the differential geometry of matrix manifolds has thus become a field of active research. For recent developments, however, without exploiting the Lie structure to the full extent, see, e.g., [2, 14].

Following joint work [53], here we sketch an overview how to treat various optimisation tasks for quantum dynamical systems in the common framework of gradient flows on smooth manifolds. Let M denote a smooth manifold, e.g., the unitary orbit of all quantum states relating to an initial state X_0. Then a *flow* is a smooth map

$$\Phi : \mathbb{R} \times M \to M \tag{23}$$

such that for all states $X \in M$ and times $t, \tau \in \mathbb{R}$ one has

$$\begin{aligned} \Phi(0,X) &= X \\ \Phi(\tau,\Phi(t,X)) &= \Phi(t+\tau,X) \\ \Phi_\tau \circ \Phi_t &= \Phi_{t+\tau} \; ; \end{aligned} \tag{24}$$

hence the flow acts as a one-parameter *group*, and for positive times $t, \tau \geq 0$ as a one-parameter *semigroup of diffeomorphisms* on M.

Now, let $f : M \to \mathbb{R}$ be a smooth quality function on M. Recall that the differential of $f : M \to \mathbb{R}$ is a mapping $Df : M \to T^*M$ of the manifold to its cotangent bundle T^*M, while the gradient vector field is a mapping $\operatorname{grad} f : M \to TM$ to its tangent bundle TM. So the scalar product $\langle \cdot | \cdot \rangle_X$ plays a central role as it allows for identifying $T^*_X M$ with $T_X M$ so the pair $\big(M, \langle \cdot | \cdot \rangle\big)$ is a *Riemannian manifold* with *Riemannian metric* $\langle \cdot | \cdot \rangle$. Thus one arrives at the *gradient flow* $\Phi : \mathbb{R} \times M \to M$ determined by

$$\dot{X} = \operatorname{grad} f(X) \; . \tag{25}$$

Formally, its solutions are obtained by integrating Eqn. (25) to give

$$\Phi(t,X) = \Phi(t,\Phi(0,X)) = X(t) \; , \tag{26}$$

where $X(t)$ denotes the unique solution of Eqn. (25) with initial value X(0) = X. Observe this ensures that f does increase along trajectories Φ of the system by virtue of following the gradient direction of f. — In generic problems, gradient flows typically run into some *local* extremum as sketched in Fig. 3 on the next page. Therefore a sufficiently large set of independent initial conditions may be needed to provide confidence into numerical results. However, in some pertinent applications, local extrema can be ruled out; prominent examples of this type will be discussed in detail in [53] in the context of Brockett's double bracket flow [9, 28].

Background: discretised gradient flows

In the simplest case, gradient flows may be solved by moving along the gradient $\operatorname{grad} f \in \mathbb{R}^m$ in the sense of a *Steepest Ascent Method*

$$X_{k+1} = X_k + \alpha_k \operatorname{grad} f(X_k), \tag{27}$$

where $\alpha_k \geq 0$ is an appropriate step size. Here, the manifold $M = \mathbb{R}^m$ coincides with its tangent space $T_X M = \mathbb{R}^m$ containing $\operatorname{grad} f(X)$. Clearly, a generalisation is required as soon as M and $T_X M$ are no longer identifiable. This gap is filled by the *Riemannian exponential map*

$$\exp_X : T_X M \to M, \quad \xi \mapsto \exp_X(\xi) \tag{28}$$

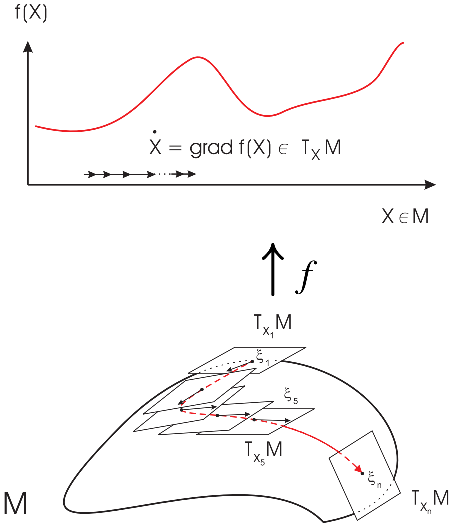


Figure 3: Abstract optimisation task: the quality function $f : M \to \mathbb{R}, X \mapsto f(X)$ (top trace) is driven into a (local) maximum by following the gradient flow $\dot{X} = \operatorname{grad} f(X)$ on the manifold M (lower trace).

such as to arrive at an *intrinsic Euler step method*. It is performed by the Riemannian exponential map, so straight line segments used in the standard method are replaced by geodesics on M in the *Riemannian Gradient Method*

$$X_{k+1} := \exp_{X_k}\big(\alpha_k \operatorname{grad} f(X_k)\big) \tag{29}$$

where $\alpha_k \geq 0$ is a step size ensuring convergence. For matrix Lie groups $\mathbf{G}$ with bi-invariant metric, Eqn. (29) simplifies to the *Gradient Method on a Lie Group*

$$X_{k+1} := \exp\big(\alpha_k \operatorname{grad} f(X_k)\, X_k^{-1}\big) X_k\,, \tag{30}$$

where $\exp : \mathfrak{g} \to \mathbf{G}$ is the usual exponential map.

In either case, the iterative procedure can be pictured as follows: at each point $X_k \in M$ one evaluates $\operatorname{grad} f(X_k)$ in the tangent space $T_{X_k}M$. Then one moves via the Riemannian exponential map in direction $\operatorname{grad} f(X_k)$ to the next point X_{k+1} on the manifold so that the quality function f improves, $f(X_{k+1}) \geq f(X_k)$, as shown in Fig. 3.

Extension: gradient flows on homogeneous spaces and subgroups

Let $\mathcal{O}(A)$ denote the unitary orbit of some $A \in \mathbb{C}^{N\times N}$and let $C \in \mathbb{C}^{N\times N}$ be another complex matrix. For minimising the (squared) Euclidean distance $\|X - C\|_2^2$ between C and the unitary orbit of A we derive a gradient flow maximising the target function

$$\widehat{f}(X) := \operatorname{Re}\operatorname{tr}\{C^\dagger X\} \tag{31}$$

over $X \in \mathcal{O}(A)$. Note the equivalence

$$\max_{X \in \mathcal{O}(A)} \widehat{f}(X) = \max_{U \in SU(N)} f(U) \tag{32}$$

for $f(U) := \operatorname{Re}\operatorname{tr}\{C^\dagger U A U^\dagger\}$. We have the following: $\mathcal{O}(A)$ constitutes a compact and connected naturally reductive homogeneous space isomorphic to $SU(N)/\mathbf{H}$, where now

$$\mathbf{H} := \{U \in SU(N) | \operatorname{Ad}_U A = A\} \tag{33}$$

denotes the stabiliser group of A. We obtain for the tangent space of $\mathcal{O}(A)$ at $X = \operatorname{Ad}_U A$ the form

$$T_X \mathcal{O}(A) = \left\{\operatorname{ad}_X \Omega \,\middle|\, \Omega \in \mathfrak{su}(N)\right\} \tag{34}$$

with $\operatorname{ad}_X \Omega := [X, \Omega]$.

Moreover, the kernel of $\operatorname{ad}_A : \mathfrak{su}(N) \to \mathfrak{g}$ reads $\mathfrak{h} = \{\Omega \in \mathfrak{su}(N) \,|\, [A, \Omega] = 0\}$. and forms the Lie subalgebra to $\mathbf{H}$. Define the ortho-complement to the above kernel as $\mathfrak{p} := \mathfrak{h}^\perp$. This induces a unique decomposition of any skew-Hermitian matrix $\Omega = \Omega^{\mathfrak{h}} + \Omega^{\mathfrak{p}}$ with $\Omega^{\mathfrak{h}} \in \mathfrak{h}$ and $\Omega^{\mathfrak{p}} \in \mathfrak{p}$ and an $\operatorname{Ad}_{SU(N)}$-invariant Riemannian metric on $\mathcal{O}(A)$ via $\operatorname{tr}\{\Omega_1^{\mathfrak{p}\,\dagger} \Omega_2^{\mathfrak{p}}\}$. Now, the main result on double-bracket flows reads:

Theorem 6 ([53]). *Set* $\widehat{f} : \mathcal{O}(A) \to \mathbb{R}$, $\widehat{f}(X) := \operatorname{Re}\operatorname{tr}\{C^\dagger X\}$. *Then one finds*

(a) *The gradient of* $\widehat{f}$ *with respect to the Riemannian metric defined above is given by*

$$\operatorname{grad} \widehat{f}(X) = [X, [X, C^\dagger]_S], \tag{35}$$

where $[X, C^\dagger]_S$ *denotes the skew-Hermitian part of* $[X, C^\dagger]$.

(b) *The gradient flow*

$$\dot{X} = \operatorname{grad} \widehat{f}(X) = [X, [X, C^\dagger]_S] \tag{36}$$

defines an isospectral flow on $\mathcal{O}(A) \subset \mathfrak{g}$. *The solutions exist for all* $t \geq 0$ *and converge to a critical point* X_∞ *of* $\widehat{f}(X)$ *characterised by* $[X_\infty, C^\dagger]_S = 0$.

Proof. (A detailed proof for the real case can be found in [28]; for an abstract Lie algebraic version see also [10].) □

In order to obtain a numerical algorithm for maximising $\widehat{f}$ one can discretise the continuous-time gradient flow (35) as

$$X_{k+1} = e^{-\alpha_k [X_k, C^\dagger]_S}\, X_k\, e^{\alpha_k [X_k, C^\dagger]_S} \tag{37}$$

with appropriate step sizes $\alpha_k > 0$. Note that Eqn. (37) heavily exploits the fact that the adjoint orbit $\mathcal{O}(A)$ constitutes a *naturally reductive homogeneous space* and thus the knowledge on its geodesics.

For A, C complex Hermitian (real symmetric) and the full unitary (or orthogonal) group or its respective orbit the gradient flow (35) is well understood. However, for non-Hermitian A and C, the nature of the flow and in particular the critical points have not been analysed in depth, because the Hessian at critical points is difficult to come by. Even for A,C Hermitian, a full critical point analysis becomes non-trivial as soon as the flow is restricted to a closed and connected *subgroup* $\mathbf{K} \subset SU(N)$. Nevertheless, the techniques from Theorem 6 can be taken over to establish a gradient flow and a respective gradient algorithm on the orbit $\mathcal{O}_\mathbf{K}$ in a straightforward manner.

Corollary 7. *The gradient flow of Eqn.* (35) *restricts to the subgroup orbit* $\mathcal{O}_\mathbf{K}(A) := \{KAK^\dagger \mid K \in \mathbf{K} \subset SU(N)\}$ *by taking the respective orthogonal projection* $P_\mathfrak{k}$ *onto the subalgebra* $\mathfrak{k} \subset \mathfrak{su}(N)$ *of* $\mathbf{K}$ *instead of projecting onto the skew-Hermitian part, i.e.* $\dot{X} = [X, P_\mathfrak{k}[X, C^\dagger]]$. □

With step sizes $\alpha_k > 0$ the corresponding discrete integration scheme reads

$$X_{k+1} = e^{-\alpha_k P_\mathfrak{k}[X_k, C^\dagger]} \, X_k \, e^{\alpha_k P_\mathfrak{k}[X_k, C^\dagger]} \; . \tag{38}$$

In view of unifying the interpretation of unitary networks, e.g., for the task of computing ground states of quantum mechanical Hamiltonians $H \equiv A$, the double-bracket flows for complex Hermitian A,C on the full unitary orbit $\mathcal{O}_u(A)$ as well as on the subgroup orbits $\mathcal{O}_\mathbf{K}(A)$ for different partitionings brought about by $\mathbf{K} := \{K \in SU(N_1) \otimes SU(N_2) \otimes \cdots \otimes SU(N_r) | \prod_{j=1}^r N_j = 2^n\}$ have shifted into focus. Therefore, we have given the foundations for the recursive schemes of Eqns. (37) and (38), which are listed with many more worked examples in the comprehensive joint work of Ref. [53].

In particular, in [53] we addressed gradient flows for constrained optimisation problems. The *intrinsic constraints* can be accomodated by restricting the dynamic group to proper subgroups $\mathbf{K} \subsetneq SU(N)$ of the unitary group. Beyond that, we also devised gradient flows combining intrinsic constraints by restrictions to proper subgroups with *extrinsic constraints* that were taken care of by Lagrange-type penalty parameters. So the work in [53] provides a full toolkit of gradient-flow based optimisations alongside [2, 14]. It has been very powerful when applied to best approximations by sums of compact group orbits [40].

3 Elements of a systems theory for open systems

While in closed systems there is a particularly simple characterisation of reachable sets in terms of the system algebra $\mathfrak{k}$ generating the Lie group $\mathbf{K} := \langle \exp(\mathfrak{k}) \rangle$ and the corresponding group orbit $\operatorname{Reach} \rho_0 = \mathcal{O}_\mathbf{K}(\rho_0) := \{K\rho_0 K^\dagger | K \in \mathbf{K} \subseteq SU(N)\}$, in open quantum systems it is considerably more intricate to estimate the reachable sets. Recall that in open systems (as for the rest of this section), we consider bilinear control systems of open quantum systems which are quantum maps following the master equation

$$\begin{aligned} \dot{F}(t) &= -(i\,\mathrm{ad}_{H_d} + i\sum_j u_j\,\mathrm{ad}_{H_j} + \Gamma_L) F(t) \quad \text{with} \quad F(0) = \mathbb{1} \\ &=: -\mathcal{L}(t) \circ F(t) \; . \end{aligned} \tag{39}$$

Just for unital systems (i.e. those with fixed point proportional to $\mathbb{1}$) which are further simplified by the (from the point-of-view of physics hopelessly idealising) assumption that all coherent controls are infinitely fast in the sense of

$$\langle iH_j \,|\, j = 1,2,\dots,m\rangle_{\text{Lie}} = \mathfrak{su}(N) \tag{40}$$

one finds by the seminal work of [62] and [3] on majorisation that

$$\text{Reach}\,\rho_0 \subseteq \{\rho \in \mathfrak{pos}_1 \,|\, \rho \prec \rho_0\} \tag{41}$$

as recently pointed out more explicitly in [64]. However, this simple characterisation becomes totally inaccurate in all physically more realistic scenarios (as longs as the noise itself is not switchable, see Sec. (3.2)), where the drift Hamiltonian H_0 is necessary to ensure full controllability in the sense of

$$\langle iH_0, iH_j \,|\, j = 1,2,\dots,m\rangle_{\text{Lie}} = \mathfrak{su}(N)\;. \tag{42}$$

In these experimentally more realistic and hence highly relevant cases, we have recently characterised the dynamic system in terms of the underlying Lie wedge $\mathfrak{w}$, i.e. the generating set of the dynamic system *Lie semigroup* $\mathbf{S}$ of irreversible (Markovian) time evolution in Refs. [17, 45]. Here the reachable sets can be conveniently and more accurately be approximated by

$$\begin{aligned} \text{Reach}\,\rho_0 &= \mathbf{S}\,\text{vec}\,\rho_0 \quad \text{where} \\ \mathbf{S} &\simeq e^{A_1} e^{A_2} \cdots e^{A_\ell} \end{aligned} \tag{43}$$

with $A_1, A_2, \dots, A_\ell \in \mathfrak{w}$ and where usually few factors suffice to give a good estimate. Suffice this to motivate the sketch of just some basic features of Lie semigroups.

3.1 Markovian quantum maps as Lie semigroups

Let us start with the following distinction: A (completely positive) trace-preserving quantum map is *(infinitely) divisible*, if $\forall r \in \mathbb{N}$ there is a S with $F = S^r$, while it is *infinitesimally divisible* if $\forall \varepsilon > 0$ there is a sequence $\prod_{j=1}^r S_j = F$ with $\|S_j - \text{id}\| \leq \varepsilon$. Moreover, a quantum map F is termed *time-(in)dependent* if it is the solution of a *time-(in)dependent* master eqn. $\dot{F} = -\mathcal{L}(t) \circ F$ with $\mathcal{L}(t)$ being time-(in)dependent. Now one finds

Theorem 8 (Wolf, Cirac [63]). *(1) The set of all time-independent Markovian quantum maps coincides with the set of all (infinitely) divisible quantum maps.*

(2) The set of all time-dependent Markovian quantum maps coincides with the closure of the set of all infinitesimally divisible quantum maps.

To sketch the relation to Lie semigroups, the basic vocabulary can be captured in the following definitions along the lines of Ref. [17]:

Definition 9. (1) A *subsemigroup* $\mathbf{S} \subset \mathbf{G}$ of a Lie group $\mathbf{G}$ with algebra $\mathfrak{g}$ contains $\mathbb{1}$ and follows $\mathbf{S} \circ \mathbf{S} \subseteq \mathbf{S}$. Its largest subgroup is denoted $E(\mathbf{S}) := \mathbf{S} \cap \mathbf{S}^{-1}$.

(2) Its *tangent cone* is defined by $\text{L}(\mathbf{S}) := \{\dot{\gamma}(0) \,|\, \gamma(0) = \mathbb{1},\, \gamma(t) \in \mathbf{S},\, t \geq 0\} \subset \mathfrak{g}$, for any $\gamma : [0,\infty) \to \mathbf{G}$ being a smooth curve in $\mathbf{S}$.

Definition 10 (Lie Wedge and Lie Semialgebra). (1) A *wedge* $\mathfrak{w}$ is a closed convex cone of a finite-dimensional real vector space.

(2) Its *edge* $E(\mathfrak{w}) := \mathfrak{w} \cap \text{-}\mathfrak{w}$ is the largest subspace in $\mathfrak{w}$.

(3) It is a *Lie wedge* if it is invariant under conjugation
$e^{\mathrm{ad}_g}(\mathfrak{w}) \equiv e^g \mathfrak{w} e^{-g} = \mathfrak{w}$
for all edge elements $g \in E(\mathfrak{w})$.

(4) A *Lie semialgebra* is a Lie wedge compatible with BCH multiplication
$X * Y := X + Y + \frac{1}{2}[X,Y] + \dots$ so that for a BCH neighbourhood B of $0 \in \mathfrak{g}$
$(\mathfrak{w} \cap B) * (\mathfrak{w} \cap B) \in \mathfrak{w}$.

Definition 11. (1) A subsemigroup is a *Lie subsemigroup*, if it is closed and fulfills $\mathbf{S} = \overline{\langle \exp \mathbf{L}(\mathbf{S}) \rangle}_{\mathbf{S}}$.

(2) A Lie wedge is *global* in $\mathbf{G}$, if there is a subsemigroup $\mathbf{S} \subset \mathbf{G}$ with tangent cone $L(\mathbf{S}) = \mathfrak{w}$ so that $\mathbf{S} = \overline{\langle \exp(\mathfrak{w}) \rangle}_{\mathbf{S}}$.

In a joint paper [17] it turned out that the seminal work of Kossakowski and Lindblad on quantum maps can now be put into the context of Lie semigroups as follows:

Theorem 12 (Kossakowski, Lindblad [26, 35, 41]). *The* Lie wedge *to the connected component of the unity of the semigroup of all invertible (completely positive and trace-preserving) maps* $\mathbf{P}_0^{cp}$ *is given by the set of* all linear operators of GKS-Lindblad form*:*

$$L(\mathbf{P}_0^{cp}) = \{-\mathcal{L} | \mathcal{L} = -(i\,\mathrm{ad}_H + \Gamma_L)\} \quad \textit{with} \tag{44}$$

$$\Gamma_L(\rho) := \tfrac{1}{2} \sum_k \{V_k^\dagger V_k, \rho\}_+ - 2 V_k \rho V_k^\dagger \tag{45}$$

Theorem 13 ([17]). *The semigroup* $\mathbf{F} := \overline{\left\langle \exp\left(\mathrm{L}(\mathbf{P}_0^{\mathrm{cp}})\right)\right\rangle}_S \subsetneq \mathbf{P}_0^{\mathrm{cp}}$ *generated by* $\mathrm{L}(\mathbf{P}_0^{\mathrm{cp}})$ *is a Lie subsemigroup with global Lie wedge* $\mathrm{L}(\mathbf{F}) = \mathrm{L}(\mathbf{P}_0^{\mathrm{cp}})$*, where* $\mathbf{F} \neq \mathbf{P}_0^{\mathrm{cp}}$.

There are indeed elements in the connected component $\mathbf{P}_0^{\mathrm{cp}}$ that cannot be exponentially generated and hence fail to be within the Lie semigroup $\mathbf{F}$. Most noteworthy, they are exactly the *non-Markovian* quantum maps in $\mathbf{P}_0^{\mathrm{cp}}$. Thus in this sense, the Markov properties and the Lie properties of quantum maps are 1 : 1.

Finally, one finds:

Corollary 14 ([17]). *Let* $F = \prod_{j=1}^r S_j$ *be a time dependent Markovian channel with* $S_1 = e^{-\mathcal{L}_1}, S_2 = e^{-\mathcal{L}_2}, \dots, S_r = e^{-\mathcal{L}_r}$ *and let* $\mathfrak{w}_r$ *denote the smallest global Lie wedge generated by* $\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_r$*. Then*

(1) F boils down to a time in*dependent Markovian channel, if it is sufficiently close to the unity and if there is a representation so that the associated Lie wedge* $\mathfrak{w}_r$ *specialises to a Lie semialgebra.*

(2) conversely, if F is a time in*dependent Markovian channel, a representation with* $\mathfrak{w}_r$ *being a Lie semialgebra trivially exists.*

So in summary, the borderline between Markovian and non-Markovian quantum maps is drawn by the Lie-semigroup property, while the separation between time-dependent and time-independent Markovian quantum maps is marked by the generating Lie wedge and its specialisation to the form of a Lie semialgebra [17].

3.2 Outlook on ongoing work: reachable sets in dissipatively controlled open systems

As stated in the introductory part, we have recently characterised *coherently* controlled bilinear open systems (of n spins-$\frac{1}{2}$) of the form

$$\dot{F} = -(i\,\mathrm{ad}_{H_0} + i\sum_j u_j(t)\,\mathrm{ad}_{H_j} + \gamma\Gamma_L)\,F(t) \tag{46}$$

(here $\gamma > 0$ constant with Γ_L of the form of Eqn. (45)) by their respective Lie wedges $\mathfrak{w}$ generating the dynamic system *Lie semigroup* $\mathbf{S}$ of irreversible (Markovian) time evolution in Ref. [45]. As stated already, this promises that the reachable sets can conveniently be approximated by $\mathrm{Reach}\,\rho_0 = \mathbf{S}\,\mathrm{vec}\,\rho_0$ where $\mathbf{S} \simeq \mathrm{e}^{A_1}\mathrm{e}^{A_2}\cdots\mathrm{e}^{A_\ell}$ with $A_1, A_2, \ldots, A_\ell \in \mathfrak{w}$ and where usually few factors suffice to give a good estimate. — For the sequel, suppose the unitary part of the above system is fully controllable in the sense

$$\langle iH_0, iH_j \,|\, j = 1,2,\ldots,m\rangle_{\mathrm{Lie}} = \mathfrak{su}(N)\ . \tag{47}$$

We have currently gone a step further such as to include into a coupled network of two-level (spin-$\frac{1}{2}$) systems *a single qubit the relaxion amplitude of which shall be switchable in a bang-bang fashion* between the two values $\{0, \gamma_*\}$ with $\gamma_* > 0$. The situation corresponds to Eqn. (46), where $\gamma \in \{0, \gamma_*\}$ and the relaxation term acts locally on a single qubit

$$\Gamma_L := V^t \otimes V - \tfrac{1}{2}(\mathbb{1} \otimes V^\dagger V\ + V^t\bar{V} \otimes \mathbb{1})\,, \tag{48}$$

while all the remaining qubits undergo no relaxation. This paves they way to entirely new domains, since the reachable sets enlarge dramatically: if in addition to unitary control there is non-unital switchable (amplitude damping) noise on a single spin ($V := \sigma_x + i\sigma_y$ for Γ_L of the form of Eqn. (45)) one finds that the controlled system can act *transitively* on the entire set of density operators, while for unital (bit-flip) switchable noise on a single spin ($V := \sigma_x/2$), the reachable set fills all density operators that are majorised by the initial state.

More precisely, one gets the following:

Theorem 15. *Let Σ_n be an n-qubit bilinear control system as in Eqn.* (46) *satisfying Eqn.* (47)*. Suppose the n^{th} qubit undergoes (non-unital) amplitude-damping relaxation the noise amplitude of which can be switched in time between the two values $\gamma(t) \in \{0, \gamma_*\}$. If qubit n is coupled to the system by (possibly several) Ising ZZ-interactions, and if there are no further sources of relaxation, then in the limit $t \cdot \gamma_* \to \infty$ the system Σ_n acts transitively on the set of all density operators $\mathfrak{pos}_1$, i.e.*

$$\overline{\mathrm{Reach}_{\Sigma_n}(\rho_0)} = \mathfrak{pos}_1 \quad \forall \rho_0 \in \mathfrak{pos}_1\ . \tag{49}$$

Theorem 16. *Let Σ_u be an n-qubit bilinear control system as in Eqn.* (46) *satisfying Eqn.* (47) *now with the n^{th} qubit undergoing (unital) bit-flip relaxation with switchable noise amplitude $\gamma(t) \in \{0, \gamma_*\}$. If qubit n is coupled to the system by Ising interactions, and if there are no further sources of relaxation, then in the limit $t \cdot \gamma_* \to \infty$ the system Σ_u acts on the set of all density operators $\mathfrak{pos}_1$ according to*

$$\overline{\text{Reach}_{\Sigma_u}(\rho_0)} = \{\rho \in \mathfrak{pos}_1 \,|\, \rho \prec \rho_0\} \quad \textit{for any } \rho_0 \in \mathfrak{pos}_1 \,. \tag{50}$$

Proof. The proofs will be presented in [5]. In both cases, the key idea is to treat the relaxative action on a diagonally chosen representation of the initial density operator ρ_0. Then it is easy to show that the relaxative action may be limited successively to arbitrary single pairs of eigenvalues, where in the non-unital case one has actions resulting in density operators of the type

$$\rho(t) \simeq \operatorname{diag}\big(\cdots, [\rho_{jj} + \rho_{kk} \cdot (1 - e^{-t\gamma_*})]_{jj}, \cdots, [\rho_{kk} \cdot e^{-t\gamma_*}]_{kk}, \cdots\big)\,,$$

while in the unital variant one finds

$$\rho(t) \simeq$$
$$\tfrac{1}{2}\operatorname{diag}\big(\cdots, [\rho_{jj} + \rho_{kk} + (\rho_{jj} - \rho_{kk}) \cdot e^{-\frac{t}{2}\gamma_*}]_{jj}, [\rho_{jj} + \rho_{kk} + (\rho_{kk} - \rho_{jj}) \cdot e^{-\frac{t\gamma_*}{2}}]_{kk}, \cdots\big)$$

so in the latter case all T-transforms can be generated thus establishing majorisation on the diagonal vectors. The rest readily follows by unitary controllability. □

Needless to say, these physically mild extensions by *bang-bang dissipative control* on a single qubit on top of *unitary control* will have a significant impact on numerical optimal control of open quantum systems. This is already apparent after a first implementation into our numerical package DYNAMO [42]. Although not the focus here, we will finally draw the distinction between abstract optimisations on (possibly constrained) reachable sets and dynamic optimal control via experimentally accessible control amplitudes in a given parameterisation.

4 Relation to numerical optimal control

While in Secs. (2.4) and (2.5) optimisations are treated in an abstract fashion, i.e. over the dynamic group or over the specific state-space manifold given by the reachable set (as illustrated in Fig. 3 on page 380), quantum engineering takes the optimisation problems into the concrete parameterisation of the actual experimental setup. More precisely, the parameterisation is made in terms of the (discretised) control amplitudes, which then steer the quantum system on the state-space manifold as an intermediate level. This is illustrated in Fig. 4 in order to show the distinction from Fig. 3.

Building upon [33, 56], which is work initially also triggered by Uwe Helmke's contact to Roger Brockett, recently we have lined up all the principle numerical algorithms into a unified programming framework DYNAMO [42] matched to solve the underlying bilinear control problems: subject to the equation of motion (1) a target function $f(X_{\text{target}}, X_0) := \operatorname{Re}\operatorname{tr}\{X_t^\dagger X_0\}$ is maximised over all admissible piece-wise constant control vectors $u_j(t) := (u_j(0), u_j(\tau), u_j(2\tau), \ldots, u_j(M\tau = T))$. This turns

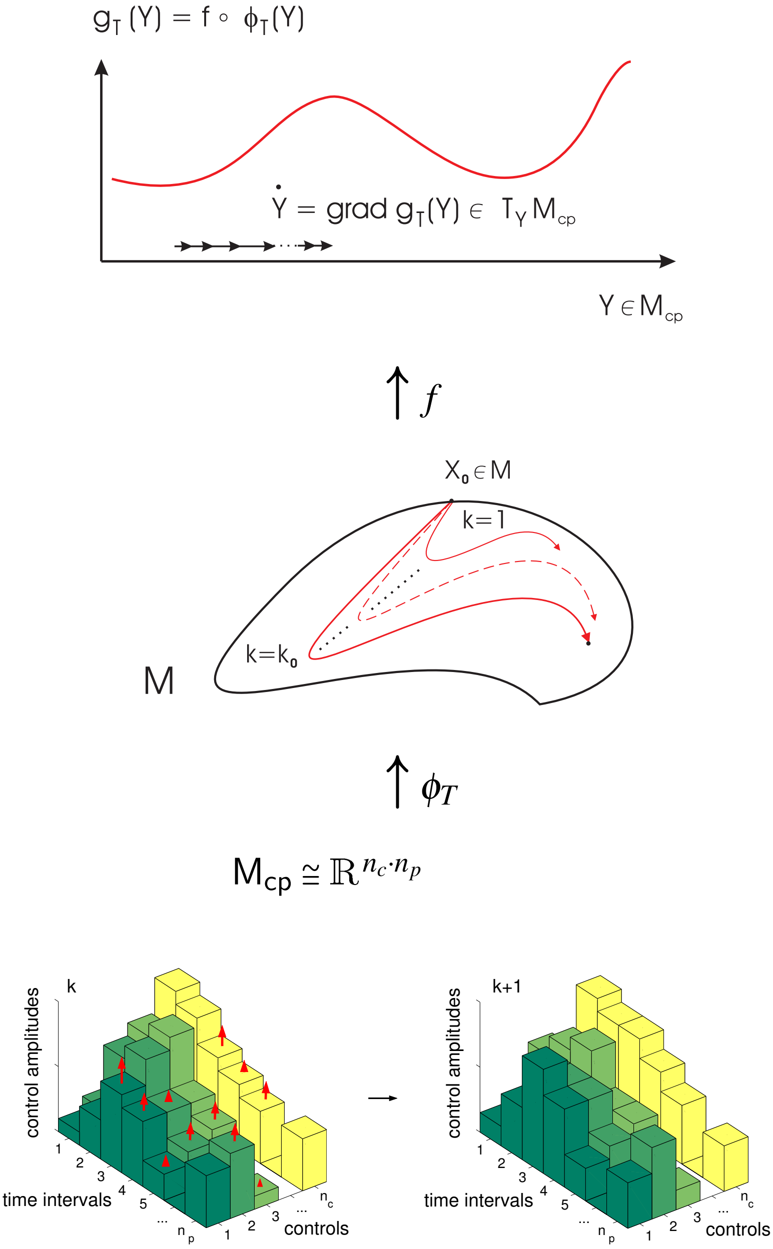
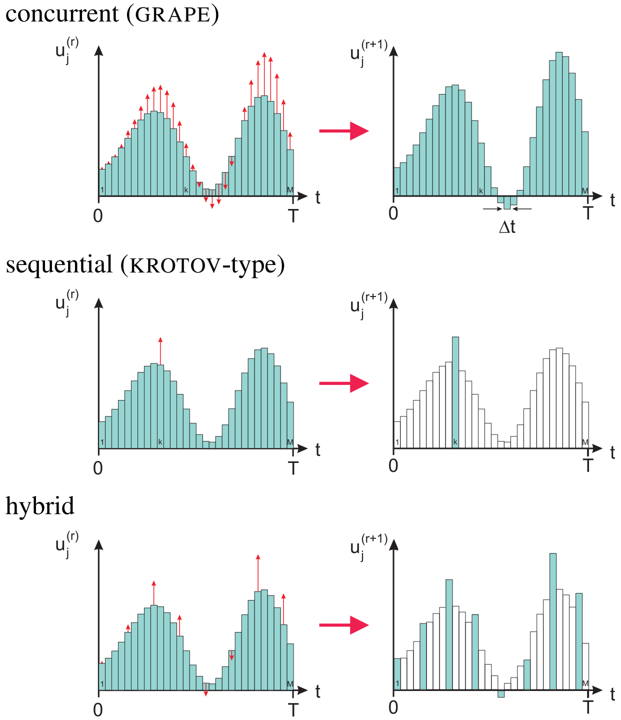


Figure 4: Optimal control task: the quality function $f : M \to \mathbb{R}, X \mapsto f(X)$ is driven into a (local) maximum on the reachable set $\text{Reach}(X_0) \subseteq M$ by following an implicit procedure (intermediate panel). It is brought about by a gradient flow $\dot{Y} = \text{grad}\, g_T(Y)$ on the level of experimental control amplitudes $Y \in M_{cp}$ (lower traces) where standard gradient-assisted methods apply.


concurrent (GRAPE)

sequential (KROTOV-type)

hybrid

Figure 5: Numerical optimal control schemes turn an initial guess of a control vector (left panels) into optimised control vectors by gradient-based first or second-order updates. This may be done concurrently, in a hybrid fashion, or sequentially. Our new DYNAMO programme package [42] offers all these options in a unified modular way.

a control vector (pulse sequence) from an initial guess into an optimised shape by following first-order gradients (or second-order increments) to all the time slices of the control vector as shown in Fig. 5, which may be done sequentially [36, 37, 57, 58], or concurrently [33, 56] or in the newly unified version DYNAMO allowing hybrids as well as switches on-the-fly from one scheme to another one [42].

These numerical schemes have been put to good use for steering quantum systems (in the explicit experimental parameter setting) such as to optimise

(1) the transfer between quantum states (pure or non-pure) [33],

(2) the fidelity of a unitary quantum gate to be synthesised in closed systems [56, 60],

(3) the gate fidelity in the presence of Markovian relaxation [55], and also

(4) the gate fidelity in the presence of non-Markovian relaxation [48]

In recent years, examples for spin systems [56, 60] as well as Josephson elements [60] have been illustrated in all detail. For optimising quantum maps in open systems, time-optimal controls have been compared to relaxation-optimised controls [55] in the light of an algebraic interpretation [17].

5 Conclusion

We have put a number of results emerging over the years in collaboration with the Helmke group into context with results obtained independently. In particular, the unifying frame comes for bilinear control systems of closed and open systems. This is of eminent importance also for engineering and steering quantum dynamical systems with high precision. In doing so, we have shown how a quantum systems theory emerges, which immediately links to many applications in quantum simulation and control without sacrificing mathematical rigour. Beyond addressing optimisation tasks on reachble sets and state-space manifolds, we have pointed out how to optimise the explicit steerings (control amplitudes) for manipulating closed and open (Markovian and non-Markovian) systems in finite dimensions.

Figure 6: Würzburg in August 1999. Gunther Dirr, Jochen Trumpf (rehearsing Einstein's posture), Thomas Schulte-Herbrüggen, and Eric Verriest during some aftermath to the *Workshop on Lie Theory and Applications* organised by Uwe Helmke and Knut Hüper. Though not quite generic, the scenary captured by Knut Hüper shows the prolific atmosphere that fostered early discussions on geodesics and time-optimality in quantum control.

Acknowledgments

It is a pleasure to thank Uwe Helmke and his group for the most fruitful atmosphere of sharing ideas (as in Fig. 6) ever since the first encounter in 1997. Quite early on, the contacts have also become transitive and triggered the exchange of Profs. Roger Brockett and Steffen Glaser via Navin Khaneja. After my visits in the Helmke group in 1999 with valuable contacts to Velimir Jurdjevic and Dirk Mittenhuber and many others, first joint projects emerged with Knut Hüper [27, 54] followed by a still ongoing series of collaborations including Martin Kleinsteuber and Indra Kurniawan all of them centered around Gunther Dirr. I would also like to acknowledge my coworkers Robert Zeier, Ville Bergholm, Uwe Sander, and Corey O'Meara for their contributions to the quoted papers [65] and [42, 50] as well as Shai Machnes, Pierre de Fouqières and Sophie Schirmer for their parts in [42].

This work has been supported in part by the EU programme Q-ESSENCE, the exchange with COQUIT, moreover by the Bavarian excellence network ENB via the International Doctorate Programme of Excellence *Quantum Computing, Control, and Communication* (QCCC) as well as by the *Deutsche Forschungsgemeinschaft* (DFG) in the collaborative research centre SFB 631 as well as the international research group supported via the grant SCHU 1374/2-1.

Bibliography

[1] D. S. Abrams and S. Lloyd. Simulation of many-body Fermi systems on a quantum computer. *Phys. Rev. Lett.*, 79:2586–2589, 1997. Cited p. 367.

[2] P. A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. Cited pp. 378 and 382.

[3] T. Ando. Majorization, doubly stochastic matrices, and comparison of eigenvalues. *Lin. Multilin. Alg.*, 118:163–248, 1989. Cited p. 383.

[4] C. H. Bennett, I. Cirac, M. S. Leifer, D. W. Leung, N. Linden, S. Popescu, and G. Vidal. Optimal simulation of two-qubit Hamiltonians using general local operations. *Phys. Rev. A*, 66:012305, 2002. Cited p. 367.

[5] V. Bergholm and T. Schulte-Herbrüggen. How to transfer between arbitrary *n*-qubit quantum states by coherent control and simplest switchable noise on a single qubit. e-print: `http://arXiv.org/pdf/1206.4945`, 2012. Cited p. 386.

[6] A. Bloch, editor. *Hamiltonian and Gradient Flows, Algorithms and Control*. Fields Institute Communications. AMS, 1994. Cited p. 378.

[7] R. W. Brockett. System theory on group manifolds and coset spaces. *SIAM J. Control*, 10:265–284, 1972. Cited p. 368.

[8] R. W. Brockett. Lie theory and control systems defined on spheres. *SIAM J. Appl. Math.*, 25:213–225, 1973. Cited p. 368.

[9] R. W. Brockett. Dynamical systems that sort lists, diagonalise matrices, and solve linear programming problems. In *Proc. IEEE Decision Control*, pages 779–803, 1988. See also: Lin. Alg. Appl., 146:79–91, 1991. Cited pp. 378 and 379.

[10] R. W. Brockett. Differential geometry and the design of gradient algorithms. *Proc. Symp. Pure Math.*, 54:69–91, 1993. Cited p. 381.

[11] D. Burgarth, K. Maruyama, S. Montangero, T. Calarco, F. Noi, and M. Plenio. Scalable quantum computation via local control of only two qubits. *Phys. Rev. A*, 81:040303, 2009. Cited p. 374.

[12] W.-S. Cheung and N.-K. Tsing. The C-numerical range of matrices is star-shaped. *Lin. Multilin. Alg.*, 41:245–250, 1996. Cited p. 376.

[13] M. T. Chou and K. R. Driessel. The projected gradient method for least-squares matrix approximations with spectral constraints. *SIAM J. Numer. Anal.*, 27:1050–1060, 1990. Cited p. 378.

[14] M. T. Chu. Linear algebra algorithms as dynamical systems. *Acta Numer.*, 17:1–86, 2008. Cited pp. 378 and 382.

[15] G. Dirr and U. Helmke. Lie theory for quantum control. *GAMM-Mitteilungen*, 31:59–93, 2008. Cited p. 368.

[16] G. Dirr, U. Helmke, M. Kleinsteuber, and T. Schulte-Herbrüggen. Relative C-numerical ranges for applications in quantum control and quantum information. *Lin. Multin. Alg.*, 56:27–51, 2008. Cited p. 376.

[17] G. Dirr, U. Helmke, I. Kurniawan, and T. Schulte-Herbrüggen. Lie semigroup structures for reachability and control of open quantum systems. *Rep. Math. Phys.*, 64:93–121, 2009. Cited pp. 383, 384, 385, and 389.

[18] J. L. Dodd, M. A. Nielsen, M. J. Bremner, and R. T. Thew. Universal quantum computation and simulation using any entangling Hamiltonian and local unitaries. *Phys. Rev. A*, 65:040301(R), 2002. Cited p. 367.

[19] J. P. Dowling and G. Milburn. Quantum technology: The second quantum revolution. *Phil. Trans. R. Soc. Lond. A*, 361:1655–1674, 2003. Cited p. 367.

[20] E. B. Dynkin. Maximal subgroups of the classical groups. *Amer. Math. Soc. Transl. Ser. 2*, 6:245–378, 1957. Reprinted in [21], pp. 37–170. Cited p. 373.

[21] E. B. Dynkin. *Selected Papers of E. B. Dynkin with Commentary*. American Mathematical Society and International Press, 2000. Cited p. 391.

[22] D. Elliott. *Bilinear Control Systems: Matrices in Action*. Springer, 2009. Cited p. 368.

[23] R. P. Feynman. Simulating physics with computers. *Int. J. Theo. Phys.*, 21:467–488, 1982. Cited p. 367.

[24] P. Gawron, Z. Puchała, J. A. Miszczak, Ł. Skowronek, and K. Życzkowski. Restricted numerical range: A versatile tool in the theory of quantum information. *J. Math. Phys.*, 51:102204, 2010. Cited p. 378.

[25] S. J. Glaser, T. Schulte-Herbrüggen, M. Sieveking, O. Schedletzky, N. C. Nielsen, O. W. Sørensen, and C. Griesinger. Unitary control in quantum ensembles: Maximising signal intensity in coherent spectroscopy. *Science*, 280:421–424, 1998. Cited p. 378.

[26] V. Gorini, A. Kossakowski, and E. C. G. Sudarshan. Completely positive dynamical semigroups of N-level systems. *J. Math. Phys.*, 17:821–825, 1976. Cited p. 384.

[27] U. Helmke, K. Hüper, J. B. Moore, and T. Schulte-Herbrüggen. Gradient flows computing the C-numerical range with applications in NMR spectroscopy. *J. Global Optim.*, 23:283–308, 2002. Cited p. 389.

[28] U. Helmke and J. B. Moore. *Optimisation and Dynamical Systems*. Springer, 1994. Cited pp. 378, 379, and 381.

[29] E. Jané, G. Vidal, W. Dür, P. Zoller, and J. I. Cirac. Simulation of quantum dynamics with quantum optical systems. *Quant. Inf. Computation*, 3:15–37, 2003. Cited p. 367.

[30] V. Jurdjevic. *Geometric Control Theory*. Cambridge University Press, 1997. Cited p. 368.

[31] V. Jurdjevic and H. Sussmann. Control systems on Lie groups. *J. Diff. Equat.*, 12:313–329, 1972. Cited p. 368.

[32] R. Kalman, P. L. Falb, and M. A. Arbib. *Topics in Mathematical System Theory*. McGraw-Hill, 1969. Cited p. 368.

[33] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser. Optimal control of coupled spin dynamics: Design of NMR pulse sequences by gradient ascent algorithms. *J. Magn. Reson.*, 172:296–305, 2005. Cited pp. 386 and 388.

[34] A. W. Knapp. *Lie Groups Beyond an Introduction*. Birkhäuser, 2nd edition, 2002. Cited p. 371.

[35] A. Kossakowski. On quantum statistical mechanics of non-Hamiltonian systems. *Rep. Math. Phys.*, 3:247–274, 1972. Cited p. 384.

[36] V. F. Krotov. *Global Methods in Optimal Control*. Marcel Dekker, 1996. Cited p. 388.

[37] V. F. Krotov and I. N. Feldman. Iteration method of solving the problems of optimal control. *Eng. Cybern.*, 21:123–130, 1983. Cited p. 388.

[38] W. S. Levine, editor. *The Control Handbook*. CRC Press in cooperation with IEEE Press, 1996. Cited p. 368.

[39] C.-K. Li. C-numerical ranges and C-numerical radii. *Lin. Multilin. Alg.*, 37:51–82, 1994. Cited p. 376.

[40] C. K. Li, Y. T. Poon, and T. Schulte-Herbrüggen. Least-squares approximation by elements from matrix orbits achieved by gradient flows on compact Lie groups. *Math. Computation*, 275:1601–1621, 2011. Cited p. 382.

[41] G. Lindblad. On quantum statistical mechanics of non-Hamiltonian systems. *Commun. Math. Phys.*, 48:119–130, 1976. Cited p. 384.

[42] S. Machnes, U. Sander, S. J. Glaser, P. de Fouquières, A. Gruslys, S. Schirmer, and T. Schulte-Herbrüggen. Comparing, optimising and benchmarking quantum control algorithms in a unifying programming framework. *Phys. Rev. A*, 84:022305, 2011. Cited pp. 386, 388, and 390.

[43] W. G. MacKay and J. Patera. *Tables of Dimensions, Indices, and Branching Rules for Representations of Simple Lie Algebras*. Marcel Dekker, 1981. Cited p. 371.

[44] M. Obata. On subgroups of the orthogonal group. *Trans. Amer. Math. Soc.*, 87:347–358, 1958. Cited p. 371.

[45] C. O'Meara, G. Dirr, and T. Schulte-Herbrüggen. Illustrating the geometry of coherently controlled unital quantum channels. In press, see extended e-print: `http://arXiv.org/pdf/1103.2703`, 2012. Cited pp. 383 and 385.

[46] T. Polack, H. Suchowski, and D. J. Tannor. Uncontrollable quantum systems. *Phys. Rev. A*, 79:053403, 2009. Cited p. 371.

[47] Z. Puchała, J. A. Miszczak, P. Gawron, C. F. Dunk, J. A. Holbrook, and K. Życzkowski. Restricted numerical shadow and geometry of quantum entanglement. e-print: `http://arXiv.org/pdf/1201.2524`, 2012. Cited p. 378.

[48] P. Rebentrost, I. Serban, T. Schulte-Herbrüggen, and F. K. Wilhelm. Optimal control of a qubit coupled to a non-Markovian environment. *Phys. Rev. Lett.*, 102:090401, 2009. Cited p. 388.

[49] S. Sachdev. *Quantum Phase Transitions*. Cambridge University Press, 1999. Cited p. 367.

[50] U. Sander and T. Schulte-Herbrüggen. Symmetry in quantum system theory of multi-qubit systems: Rules for quantum architecture design. e-print: `http://arXiv.org/pdf/0904.4654`, 2009. Cited p. 390.

[51] T. Schulte-Herbrüggen. *Aspects and Prospects of High-Resolution NMR*. PhD thesis, ETH, Zürich, 1998. Diss-ETH 12752. Cited p. 377.

[52] T. Schulte-Herbrüggen, G. Dirr, U. Helmke, M. Kleinsteuber, and S. J. Glaser. The significance of the C-numerical range and the local C-numerical range in quantum control and quantum information. *Lin. Multin. Alg.*, 56:3–26, 2008. Cited p. 376.

[53] T. Schulte-Herbrüggen, S. J. Glaser, G. Dirr, and U. Helmke. Gradient flows for optimization in quantum information and quantum dynamics: Foundations and applications. *Rev. Math. Phys.*, 22:597–667, 2010. Cited pp. 378, 379, 381, and 382.

[54] T. Schulte-Herbrüggen, K. Hüper, U. Helmke, and S. J. Glaser. *Applications of Geometric Algebra in Computer Science and Engineering*, chapter Geometry of Quantum Computing by Hamiltonian Dynamics of Spin Ensembles, pages 271–283. Birkhäuser, 2002. Cited p. 389.

[55] T. Schulte-Herbrüggen, A. Spörl, N. Khaneja, and S. J. Glaser. Optimal control for generating quantum gates in open dissipative systems. *J. Phys. B*, 44:154013, 2011. Cited pp. 388 and 389.

[56] T. Schulte-Herbrüggen, A. K. Spörl, N. Khaneja, and S. J. Glaser. Optimal control-based efficient synthesis of building blocks of quantum algorithms: A perspective from network complexity towards time complexity. *Phys. Rev. A*, 72:042331, 2005. Cited pp. 386, 388, and 389.

[57] K. Singer, U. Poschinger, M. Murphy, P. Ivanov, F. Ziesel, T. Calarco, and F. Schmidt-Kaler. Trapped ions as quantum bits: Essential numerical tools. *Rev. Mod. Phys.*, 82:2609, 2010. Cited p. 388.

[58] S. E. Sklarz and D. J. Tannor. Quantum computation via local control theory: Direct sum vs. direct product Hilbert spaces. *Chem. Phys.*, 322:87–97, 2006. Cited p. 388.

[59] E. Sontag. *Mathematical Control Theory*. Springer, 1998. Cited p. 368.

[60] A. K. Spörl, T. Schulte-Herbrüggen, S. J. Glaser, V. Bergholm, M. J. Storcz, J. Ferber, and F. K. Wilhelm. Optimal control of coupled Josephson qubits. *Phys. Rev. A*, 75:012302, 2007. Cited pp. 388 and 389.

[61] H. Sussmann and V. Jurdjevic. Controllability of nonlinear systems. *J. Diff. Equat.*, 12:95–116, 1972. Cited p. 368.

[62] A. Uhlmann. Sätze über Dichtematrizen. *Wiss. Z. Karl-Marx-Univ. Leipzig, Math. Nat. R.*, 20:633–637, 1971. Cited p. 383.

[63] M. M. Wolf and J. I. Cirac. Dividing quantum channels. *Commun. Math. Phys.*, 279:147–168, 2008. Cited p. 383.

[64] H. Yuan. Characterization of majorization monotone quantum dynamics. *IEEE. Trans. Autom. Contr.*, 55:955–959, 2010. Cited p. 383.

[65] R. Zeier and T. Schulte-Herbrüggen. Symmetry principles in quantum system theory. *J. Math. Phys.*, 52:113510, 2011. Cited pp. 371, 372, 373, 374, 375, and 390.

Properties of the BFGS method on Riemannian manifolds

Matthias Seibert
Technische Universität München
München, Germany
m.seibert@tum.de

Martin Kleinsteuber
Technische Universität München
München, Germany
kleinsteuber@tum.de

Knut Hüper
Julius-Maximilians-Universität
Würzburg, Germany
hueper@mathematik.uni-wuerzburg.de

Abstract. We discuss the BFGS method on Riemannian manifolds and put a special focus on the degrees of freedom which are offered by this generalization. Furthermore, we give an analysis of the BFGS method on Riemannian manifolds that are isometric to $\mathbb{R}^n$.

1 Introduction

Optimization problems can be found in a variety of forms, and there are countless different optimization algorithms that attempt to solve these problems. Newton's method represents one of the most famous of these algorithms, though it poses some computational bottlenecks. Quasi-Newton methods are variations developed to avoid these intricacies. The most successful of them turned out to be the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. It has many favorable properties and is therefore often in the main focus. The classical, unconstrained BFGS method on Euclidean spaces has been discussed extensively, see for instance the monographs [8, 11, 13].

However, Euclidean spaces are not the only spaces in which optimization algorithms are employed. There are many applications of optimization on Riemannian manifolds, especially in the field of engineering. As long as the considered manifolds are embedded in $\mathbb{R}^n$, the powerful tools of constrained optimization, that are, for example, examined in [13], can be applied. Still, in many cases such an embedding is *a priori* not at hand, and optimization methods explicitly designed for Riemannian manifolds have to be utilized. These concepts were to the authors' knowledge first introduced by Gabay in [6, 7] where he developed a steepest decent, a Newton, and a quasi-Newton algorithm. Some of these were further expanded by Udrişte in [16] where he discussed steepest descent and generalized Newton methods. In [5] the authors developed conjugate gradient and Newton algorithms for Stiefel and Graßmann manifolds. The common denominator of these approaches is that instead of conducting a linear step during the line search procedure, they define the step along a geodesic via the use of the exponential mapping. An alternative approach was presented in [2] using the concept of retractions, oftentimes a computationally cheaper way of mapping a tangent vector onto the manifold. It was used to formulate a Newton's method that maintains the convergence properties while being significantly cheaper to compute. This idea was picked up again in [1] and expanded with the notion of vector transport which no longer restricts us to the use of parallel transport to connect different tangent spaces. The purpose of vector transport is similar to that of retractions, i.e., less computational cost while maintaining the convergence properties. The authors used the concepts of retraction and vector transport to develop several optimization algorithms, namely a Newton's and a conjugate gradient descent algorithm as well as a quasi-Newton algorithm. This quasi-Newton algorithm was discussed further in [14]. In [3] a similar quasi-Newton algorithm, defined to work on a Graßmann manifold, was presented which also applies the notion of vector transport. In their recent work [15] Ring and Wirth proved the superlinear convergence of BFGS methods on Riemannian manifolds, however, under strong conditions to the manifold structure.

In all the works mentioned above, the implementation of the quasi-Newton algorithm on Riemannian manifolds is based on the method Gabay first introduced. However, the process of defining the algorithm on manifolds already offers a whole bunch of degrees of freedom. In this paper we will propose several different approaches to define the algorithm by diversifying the application of the vector transport. The other focus of this paper is to proof that the BFGS method on Riemannian manifolds that are isometric to $\mathbb{R}^n$ are equivalent to a classical BFGS method on $\mathbb{R}^n$.

2 Notation and basic concepts

In the following we introduce several important concepts, which will come of use later on. First, we recall the algorithm for the classical local BFGS method. Then we introduce some differential geometric concepts necessary to generalize the BFGS method on Riemannian manifolds based on [1]. Exponential mapping and retractions are used to associate tangent vectors with points on the manifold, while parallel and vector transport are used to identify different tangent spaces.

Algorithm 4: Local BFGS method

```
Given starting point $x_0$, convergence tolerance $\varepsilon > 0$, and an inverse
Hessian approximation $H_0$. Set $k = 0$.
repeat
    Solve the equation

        $H_k p_k = -\nabla f(x_k)$                                        (1)

    for $p_k$, and define the new iterate as $x_{k+1} := x_k + p_k$.
    Set $s_k := x_{k+1} - x_k$, $y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$, and compute a new
    approximation of the inverse Hessian matrix according to the update rule

        $H_{k+1} := H_k + \frac{y_k y_k^\top}{y_k^\top s_k} - \frac{H_k s_k s_k^\top H_k}{s_k^\top H_k s_k}.$     (2)

    Set $k \leftarrow k + 1$.
until $\|\nabla f(x_k)\| < \varepsilon$
```

2.1 The BFGS method

As mentioned before, Quasi-Newton (QN) methods were developed to circumvent problems that arise with Newton's method, such as the necessity to evaluate the Hessian matrix in each iteration. There certainly exist several prominent QN methods which all possess their advantages and disadvantages. In general, the BFGS method is proved to be reliable and robust. It is therefore recalled here. It should also be mentioned that instead of approximating the Hessian one can also approximate the inverse of the Hessian. The update for the inverse Hessian approximation has the form

$$B_{k+1} := B_k + \frac{(s_k - B_k y_k)s_k^\top + s_k(s_k - B_k y_k)^\top}{y_k^\top s_k} - \frac{(s_k - B_k y_k)^\top y_k}{(y_k^\top s_k)^2} s_k s_k^\top. \tag{3}$$

This yields the advantage that it is no longer necessary to solve a system of equations. Instead, only a matrix vector product has to be calculated.

The BFGS method maintains the excellent convergence properties of Newton's method. It is shown *e.g.* in [12] that it converges at a superlinear rate under certain conditions.

2.2 Affine connections

Affine connections are an important concept in differential geometry. They identify nearby tangent spaces with each other and thereby offer a possibility to differentiate tangent vector fields. Affine connections allow us to infinitesimally view a manifold as an Euclidean space. While any manifold admits an infinite amount of affine connections, there are some that offer unique and therefore often more attractive properties.

Definition 1 (Affine connection)**.** An *affine connection* ∇ on a manifold M is a mapping

$$\begin{aligned} \nabla \colon \Gamma(TM) \times \Gamma(TM) &\to \Gamma(TM), \\ (\eta, \xi) &\mapsto \nabla_\eta \xi \end{aligned} \tag{4}$$

that satisfies the following properties.

1. $C(M)$-linearity in the first variable: $\nabla_{f\eta + g\chi}\xi = f\nabla_\eta \xi + g\nabla_\chi \xi$.
2. $\mathbb{R}$-linearity in the second variable: $\nabla_\eta(a\xi + b\zeta) = a\nabla_\eta \xi + b\nabla_\eta \zeta$.
3. Product rule (Leibniz' law): $\nabla_\eta(f\xi) = (\eta f)\xi + f\nabla_\eta \xi$,

in which $\eta, \chi, \xi, \zeta \in \Gamma(TM)$, $f, g \in C(M)$, and $a, b \in \mathbb{R}$. $\nabla_\eta \xi$ is also called the *covariant derivative of* ξ *in the direction of* η.
Here, $\Gamma(TM)$ denotes the set of smooth vector fields on M.

The notation ηf that appears here is to be understood as the application of a vector field to a function which is defined as

$$\eta f(x) := \eta_x(f) = \dot{\gamma}(0)f := \left.\frac{d(f(\gamma(t)))}{dt}\right|_{t=0}. \tag{5}$$

An important affine connection with additional properties is the

Definition 2 (Levi-Civita connection)**.** On a Riemannian manifold (M,g), there exists a unique affine connection ∇ that satisfies the following two conditions.

1. $\nabla_\eta \xi - \nabla_\xi \eta = [\eta, \xi]$, i. e., it is *torsion free*.
2. $\chi\langle\eta,\xi\rangle = \langle\nabla_\chi \eta,\xi\rangle + \langle\eta,\nabla_\chi \xi\rangle$, i. e., it is *compatible with the Riemannian metric*

for all $\chi, \eta, \xi \in \Gamma(TM)$. This affine connection ∇ is called the *Levi-Civita connection* or the *Riemannian connection* of M.

Assuming this connection exists, it is uniquely defined.

2.3 Exponential mapping and retractions

A retraction R is a mapping from the tangent bundle TM onto the manifold M. At a specific point x the retraction is denoted by R_x and it is a mapping from T_xM onto M.

Definition 3 (Retraction)**.** A *retraction* on a manifold M is a smooth mapping R from the tangent bundle TM onto M with the following properties. Let R_x denote the restriction of R to T_xM.

1. $R_x(0_x) = x$ where 0_x denotes the zero element of T_xM.
2. With the canonical identification $T_{0_x}T_xM \simeq T_xM$, R_x satisfies $\mathrm{D}R_x(0_x) = \mathrm{id}_{T_xM}$ where id_{T_xM} denotes the identity mapping on T_xM.

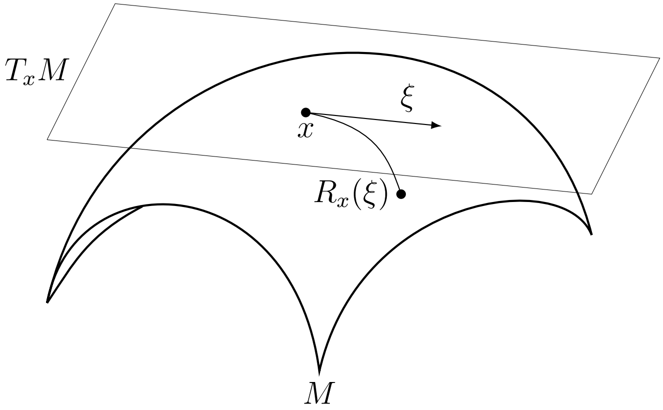


Figure 1: Illustration of a retraction.

In [1] it has been shown that the classical exponential map fulfills the requirements of a retraction. And while, in a geometric sense, the exponential mapping is the most natural retraction, it faces the problem that in order to find a solution, one is asked to solve a nonlinear ordinary differential equation. Retractions offer a way to approximate the exponential map at less computational cost while not adversely influencing the behavior of an optimization algorithm.

2.4 Parallel transport and vector transport

Given a Riemannian manifold (M,g) endowed with an affine connection and a smooth curve $\gamma{:}I \to M$. A vector field ξ on γ is called *parallel* if the equation

$$\nabla_{\dot{\gamma}(t)}\xi = 0 \text{ for all } t \in I \tag{6}$$

is satisfied. Now, let η_x be a tangent vector at x, and let γ be a smooth curve in M with $\gamma(a) = x$ and $\gamma(b) = y$. Then the parallel transport (or parallel translation) of η_x along γ is given by the vector field in the tangent bundle TM that fulfills the initial value problem

$$\begin{aligned} \nabla_{\dot{\gamma}(t)}\xi &= 0, \\ \xi_{\gamma(a)} &= \eta_x. \end{aligned} \tag{7}$$

The concept of parallel transport is used to identify tangent spaces.

In the same way that retractions avoid the high cost of evaluating exponential maps, there is a concept to avoid the necessity of solving a differential equation in order to identify tangent spaces. It is introduced in the next definition.

Definition 4 (Vector transport)**.** A *vector transport* on a manifold M is defined as a smooth mapping $\Gamma{:}M \times M \times TM \to TM$, $(x,y,\eta_x) \mapsto \Gamma_x^y(\eta_x)$, $x,y \in M$ that fulfills the properties:

1. There exists an associated retraction R and a tangent vector ξ_x satisfying $\Gamma_x^y(\eta_x) \in T_{R_x(\xi_x)}M$ for all $\eta_x \in T_xM$.

2. $\Gamma_x^x(\eta_x) = \eta_x$ for all $\eta_x \in T_xM$.

3. The mapping $\Gamma_x^y{:}T_xM \to T_yM$ is linear.

3 The Riemannian BFGS algorithm

The classical BFGS algorithm is defined on Euclidean spaces and is easily extended to submanifolds of $\mathbb{R}^n$ that are given by equality constraints, i. e., to the case where the regular value theorem applies. In order to define a BFGS algorithm on generic manifolds, further structure is necessary. The manifold has to be equipped with a Riemannian structure, so that we are able to calculate gradients and inner products necessary to define a quasi-Newton algorithm. The following algorithm introduces the seemingly most natural way to define a Riemannian BFGS (in short: RBFGS) algorithm. The changes to the original algorithm and possible variations will be discussed in detail later on. In [14] the following algorithm is presented:

Algorithm 5: BFGS on Riemannian manifolds

```
Given a Riemannian manifold $M$ with Riemannian metric $\langle\cdot,\cdot\rangle$, a vector transport $\Gamma$ with associated retraction $R$, the real valued function $f\colon M \to \mathbb{R}$, an initial iterate $x_0 \in M$, and an initial approximation to the Hessian $H_0$. Set $k := 0$.
repeat
    Solve $H_k p_k = -\operatorname{grad} f(x_k)$ for $p_k \in T_{x_k}M$.
    Obtain the step length $\alpha$ through an appropriate line search algorithm.
    Define
        $s_k := \Gamma_{x_k}^{x_{k+1}}(\alpha p_k)$,
        $y_k := \operatorname{grad} f(x_{k+1}) - \Gamma_{x_k}^{x_{k+1}}(\operatorname{grad} f(x_k))$,
        $H_{k+1}\eta := \tilde{H}_k\eta + \frac{\langle y_k,\eta\rangle}{\langle y_k,s_k\rangle} y_k - \frac{\langle s_k,\tilde{H}_k\eta\rangle}{\langle s_k,\tilde{H}_k s_k\rangle}\tilde{H}_k s_k$        (8)
        with $\tilde{H}_k := \Gamma_{x_k}^{x_{k+1}} \circ H_k \circ \Gamma_{x_{k+1}}^{x_k}$
    where $H_{k+1}\colon T_{x_{k+1}}M \to T_{x_{k+1}}M$ is a linear operator.
    Set $k \leftarrow k+1$.
until $\|\operatorname{grad} f(x_{k+1})\| < \varepsilon$
```

The first thing to note in Algorithm 5 is the change in the update formula for the approximated Hessian H_k. The classical update formulas for the approximation of the Hessian which are used in Euclidean space have no meaning in a Riemannian manifold setting. First, at instances where the transpose of a column vector is multiplied by another vector the standard inner product of vectors in Euclidean space is meant. When operating on Riemannian manifolds as above, s_k and y_k are vectors in the tangent space $T_{x_{k+1}}M$. The inner product on tangent spaces is then given by the chosen Riemannian metric. Furthermore, the dyadic product of a vector with the transpose of another vector, which results in a matrix in the Euclidean space, is not a naturally defined operation on a Riemannian manifold. And finally, while in Euclidean space the Hessian can be expressed as a symmetric matrix, on Riemannian manifolds it can be defined as a symmetric, bilinear form. However, due to the Lax-Milgram Lemma, [17], there exists a linear function $H\colon T_xM \to T_xM$ with

$$\mathrm{D}^2 f(x)(\eta,\xi) = \langle \eta, H\xi\rangle, \quad \eta,\xi \in T_xM. \tag{9}$$

This Lemma can be applied since the Hessian at x is a bilinear form on the tangent space T_xM which is actually a Hilbert space. It is this linear function H that will be updated during the BFGS algorithm instead of the Hessian matrix. Together with the use of the Riemannian metric, this leads to the update

$$H_{k+1}\eta = \tilde{H}_k\eta + \frac{\langle y_k,\eta\rangle}{\langle y_k,s_k\rangle} y_k - \frac{\langle s_k,\tilde{H}_k\eta\rangle}{\langle s_k,\tilde{H}_k s_k\rangle}\tilde{H}_k s_k \tag{10}$$

which was used in the algorithm above. As a result the search direction p_k is the vector in the tangent space $T_{x_{k+1}}M$ that satisfies $H_{k+1}p_k = -\operatorname{grad} f(x_{k+1})$. Instead of approximating the Hessian it is also possible to define the BFGS update for the inverse of the Hessian. This approach offers the advantage that it is not necessary to solve a system of equations. On Riemannian manifolds this update approach has the form

$$\begin{aligned} B_{k+1}\eta = \tilde{B}_k\eta + \frac{\langle s_k, \eta\rangle}{\langle y_k, s_k\rangle}s_k - \frac{\langle s_k, \eta\rangle}{\langle y_k, s_k\rangle}\tilde{B}_k y_k \\ - \frac{\langle y_k, \tilde{B}_k\eta\rangle}{\langle y_k, s_k\rangle}s_k + \frac{\langle y_k, \tilde{B}_k y_k\rangle\langle s_k, \eta\rangle}{\langle y_k, s_k\rangle^2}s_k. \end{aligned} \tag{11}$$

As before, it has the advantage that there is no need to solve a linear equation. Finding the solution to a linear equation has the additional difficulty that the solution has to be an element of a certain tangent space. Instead, this update only requires the evaluation of $B_{k+1}\,\mathrm{grad}f(x_{k+1})$.

Remember the notation $\tilde{H}_k$ and $\tilde{B}_k$ that has been introduced in these update formulas. The operators H_k and B_k can only be applied to elements of $T_{x_k}M$ by definition, whereas the search direction p_k is an element of $T_{x_{k+1}}M$. Thus, the tangent vector p_k has to be transported to $T_{x_k}M$. Consequently, one of the operators H_k or B_k can be applied, and the resulting tangent vector in x_k is moved back to $T_{x_{k+1}}M$ via vector transport in order for the final result to be a tangent vector at x_{k+1}. Succinctly, this means we have $\tilde{H}_k := \Gamma_{x_k}^{x_{k+1}} \circ H_k \circ (\Gamma_{x_k}^{x_{k+1}})^{-1}$ and $\tilde{B}_k = \Gamma_{x_k}^{x_{k+1}} \circ B_k \circ (\Gamma_{x_k}^{x_{k+1}})^{-1}$, respectively.

The definition of the values s_k and y_k has also been remodelled. In Algorithm 5 we chose to compute the new search direction as an element of $T_{x_{k+1}}M$. Thus, the natural choice to define s_k is αp_k, which is the closest relation we have to the difference between the last two iteration points, and transport it to $T_{x_{k+1}}M$ via vector transport. Furthermore, we define $y_k := \mathrm{grad}f(x_{k+1}) - \Gamma_{x_k}^{x_{k+1}}(\mathrm{grad}f(x_k))$. This concludes the definition of a BFGS algorithm on Riemannian manifolds.

However, this is not the only possible way to transfer the BFGS algorithm to Riemannian manifolds. There are several further degrees of freedom, which are as follows.

- The choice of the Riemannian metric.
- The choice of retraction and parallel transport.
- The choice in which tangent space the update is calculated.

For the second point, the natural choice is to use the parallel transport that is induced by the Levi-Civita connection as the vector transport and exponential mapping as the retraction. These two choices describe the “exact” operations and are the ones classically used in differential geometry. They have the drawback of high computational burden, as we already mentioned in Section 2.

An alternative way to address the third point is to move all the required variables to $T_{x_k}M$ and formulate the update there. Then the update step in Algorithm 5 has the following form:

Algorithm 6: Modified update step

5: Define

$$\begin{aligned} s_k &:= \alpha p_k, \\ y_k &:= (\Gamma_{x_k}^{x_{k+1}})^{-1}(\mathrm{grad} f(x_{k+1})) - \mathrm{grad} f(x_k), \\ \tilde{H}_{k+1}\tilde{\eta} &:= H_k\tilde{\eta} + \frac{\langle y_k, \tilde{\eta}\rangle}{\langle y_k, s_k\rangle} y_k - \frac{\langle s_k, H_k\tilde{\eta}\rangle}{\langle s_k, H_k s_k\rangle} H_k s_k, \quad \tilde{\eta} \in T_{x_k}M, \\ H_{k+1}\eta &:= \left(\Gamma_{x_k}^{x_{k+1}} \circ \tilde{H}_{k+1} \circ (\Gamma_{x_k}^{x_{k+1}})^{-1}\right)(\eta), \quad \eta \in T_{x_{k+1}}M \end{aligned} \tag{12}$$

where $\tilde{H}_{k+1}: T_{x_k}M \to T_{x_k}M$ is a linear operator.

It is a well known fact that the parallel transport induced by the Levi-Civita connection along geodesics leaves the inner product on the two connected tangent spaces invariant. So if the vector transport is chosen to be this parallel transport, it can easily be shown that the algorithm with update rule (12) is identical to the first algorithm in the sense that the point sequences they produce coincide. However, any other choice of vector transport will lead to a different point sequence, and it is not obvious whether there is an algorithm with superior convergence properties. Another fact to note is that while update rule (8) requires five instances of vector transport per iteration, the algorithm defined with (12) only takes three and is therefore computationally cheaper if only a single iteration is considered.

We now have considered moving all required variables either to $T_{x_k}M$ or to $T_{x_{k+1}}M$, and subsequently calculating the formula for the updated, approximated Hessian in the respective tangent space. While these are the options that first come to mind, it is also possible to move all variables to a third, completely unrelated tangent space. For the arbitrary but fixed point z on the manifold M the update step then assumes the following form:

Algorithm 7: Modified update step

5: Define

$$\begin{aligned} s_k &:= \Gamma_{x_k}^{z}(\alpha p_k), \\ y_k &:= \Gamma_{x_{k+1}}^{z}(\mathrm{grad} f(x_{k+1})) - \Gamma_{x_k}^{z}(\mathrm{grad} f(x_k)), \\ \hat{H}_{k+1}\hat{\eta} &:= \hat{H}_k\hat{\eta} + \frac{\langle y_k, \hat{\eta}\rangle}{\langle y_k, s_k\rangle} y_k - \frac{\langle s_k, \hat{H}_k\hat{\eta}\rangle}{\langle s_k, \hat{H}_k s_k\rangle} \hat{H}_k s_k, \quad \hat{\eta} \in T_z M, \\ H_{k+1}\eta &:= \left(\Gamma_{z}^{x_{k+1}} \circ \hat{H}_{k+1} \circ \Gamma_{x_{k+1}}^{z}\right)(\eta), \quad \eta \in T_{x_{k+1}}M \end{aligned} \tag{13}$$

where $\hat{H}_{k+1}: T_z M \to T_z M$ is a linear operator.

While this will most likely not have any computational advantages, it is still a possible way to define a working BFGS algorithm.

The diagrams in Figures 2 and 3 visualize the three different methods to employ the vector transport that were discussed.

$$\begin{tikzcd}
 & \tilde{p}_k \arrow[r, "\Gamma_{x_k}^{x_{k+1}}"] & p_k \\
\tilde{g}_{k+1} \arrow[ur, "\tilde{H}_{k+1}" description] & & \hat{p}_k \arrow[u, "\Gamma_{z}^{x_{k+1}}"'] \\
g_{k+1} \arrow[u, "\Gamma_{x_{k+1}}^{x_k}"] \arrow[uurr, "H_{k+1}" description] \arrow[r, "\Gamma_{x_{k+1}}^{z}"'] & \hat{g}_{k+1} \arrow[ur, "\hat{H}_{k+1}" description] &
\end{tikzcd}$$

Figure 2: Obtaining the search direction.

$$\begin{tikzcd}
T_{x_k}M \arrow[rr, "\Gamma_{x_k}^{x_{k+1}}" description] & & T_{x_{k+1}}M \\
 & T_{x_{k+1}}M \arrow[ur, "\Gamma_{x_{k+1}}^{x_{k+1}}" description] & \\
T_{x_{k+1}}M \arrow[uu, "\Gamma_{x_{k+1}}^{x_k}" description] \arrow[ur, "\Gamma_{x_{k+1}}^{x_{k+1}}" description] \arrow[rr, "\Gamma_{x_{k+1}}^{z}" description] & & T_zM \arrow[uu, "\Gamma_{z}^{x_{k+1}}" description]
\end{tikzcd}$$

Figure 3: Relations between the involved tangent spaces.

The diagram in Figure 2 describes the process of obtaining the search direction p_k starting from $g_{k+1} := \mathrm{grad} f(x_{k+1})$ while the diagram in Figure 3 pictures the tangent spaces that are passed through in the course of this operation. It is important to keep in mind that in general each method of obtaining the search direction produces a different result. Hence, the diagrams are not commutative.

While these alternative ways to incorporate the vector transport were only discussed for the direct BFGS update, it is obvious that the inverse update can be adjusted in the same way. The respective update formulas B_k are obtained analogously to the ones for the direct update.

Another aspect to mention is that both the retractions and vector transports obviously are specific to the manifold on which the algorithm operates. This means that for each new manifold that is considered, these operations have to be adjusted individually.

Finally, in the case that the manifold M is simply $\mathbb{R}^n$, it is obvious that Algorithm 5, as well as its alterations Algorithms 6 and 7 reduce to the classical BFGS method in Euclidean space.

3.1 Line search along geodesics

Although this will not be the main focus of this paper, we will give a brief introduction to line search procedures on Riemannian manifolds that are required for globalizing Algorithm 5.

There are several changes that have to be made to the well known line search algorithms that are used in Euclidean space in order to use them for optimization algorithms on Riemannian manifolds. The first method that is introduced is called a backtracking procedure and represents a very basic form of line search. The implementation used here is similar to the one in [3] and is defined as follows.

Algorithm 8: Step length calculation via backtracking

```
Set $\alpha = 1$ and define $c := \langle \mathrm{grad} f(x_k), p_k \rangle$.
While $f(R_{x_k}(2\alpha p_k)) - f(x_k) < \alpha c$, set $\alpha := 2\alpha$.
While $f(R_{x_k}(\alpha p_k)) - f(x_k) \geq 0.5\alpha c$, set $\alpha := 0.5\alpha$.
```

Another possibility is to customize classical line search conditions which are used for optimization problems in $\mathbb{R}^n$. For a Riemannian BFGS problem, the *Armijo condition* assumes the form

$$f(R_x(\alpha p_k)) \leq f(x_k) + c_1 \alpha \langle \mathrm{grad} f(x_k), p_k \rangle. \tag{14}$$

As in Euclidean space we can add a curvature condition, and obtain

$$\begin{aligned} f(R_x(\alpha p_k)) &\leq f(x_k) + c_1 \alpha \langle \mathrm{grad} f(x_k), p_k \rangle, \\ \langle \mathrm{grad} f(R_x(\alpha p_k), p_k \rangle &\geq c_2 \langle \mathrm{grad} f(x_k), p_k \rangle \end{aligned} \tag{15}$$

with $0 < c_1 < c_2 < 1$. This is the *Wolfe-Powell condition* in analogy to the original Wolfe-Powell condition found in [12]. Just as in $\mathbb{R}^n$ an even more constrictive condition can be derived from this by tightening the second condition to

$$|\langle \mathrm{grad} f(R_x(\alpha p_k), p_k \rangle| \geq c_2 |\langle \mathrm{grad} f(x_k), p_k \rangle| \tag{16}$$

which results in the *strong Wolfe-Powell condition* for Riemannian manifolds.

4 Relation of BFGS algorithms on isometric manifolds

In this section, we will analyze whether RBFGS algorithms that operate on manifolds which are linked by an isometry can be related to one another.

In the following let (M,g) and (N,h) be two Riemannian manifolds, and let the function $\Phi : M \to N$ be a smooth map between these two smooth manifolds.

Definition 5 (Pushforward)**.** The differential $\mathrm{D}\Phi_x$ of Φ at $x \in M$ is a linear mapping

$$\begin{aligned} \mathrm{D}\Phi_x : T_x M &\to T_{\Phi(x)} M \\ \xi &\mapsto \mathrm{D}\Phi_x[\xi]. \end{aligned} \tag{17}$$

It is called the *pushforward by* Φ. In future, we will often denote the linear operator that represents this mapping by $\Phi_*[x]$, i. e., $\mathrm{D}\Phi_x[\xi] = \Phi_*[x]\xi$. If the point at which Φ is evaluated is evident from the context, it is omitted to simplify notation.

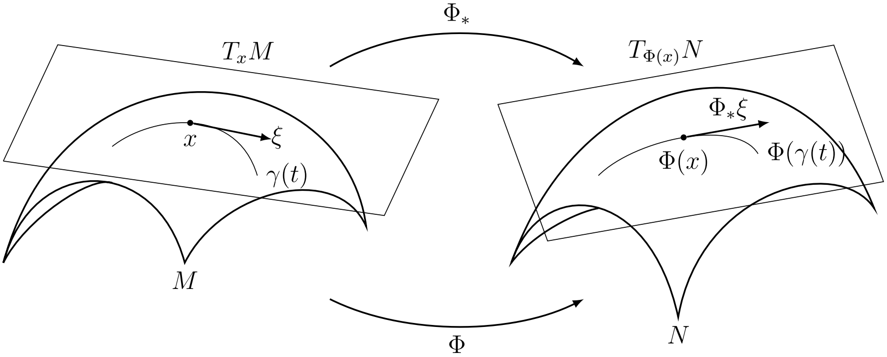


Figure 4: Illustration of the Pushforward operation.

One thing to note is that Φ is an immersion if and only if Φ_* is injective for all $x \in M$ or a submersion if and only if Φ_* is surjective for all $x \in M$.

Definition 6 (Pullback)**.** There is a linear map from the space of 1-forms on M to the space of 1-forms on N. This map is called the *pullback by* Φ and is denoted by

$$\Phi^* : T^*_{\Phi(x)}N \to T^*_x M. \tag{18}$$

Under the condition that Φ is a diffeomorphism, the pullback and the pushforward can be used to map any vector (and even tensor) field from M to N and vice versa. Of interest to us is the application of the pullback to two kinds of functions.

First, we consider the pullback of smooth cost functions. Let $f : N \to \mathbb{R}$ be a smooth function. Then the pullback of f by Φ is defined as

$$(\Phi^* f)(x) = f(\Phi(x)). \tag{19}$$

Secondly, the pullback of covariant tensor fields is examined. Let S be a $(0,s)$-tensor field on N of the form $S : T_yN \times T_yN \times ... \times T_yN \to \mathbb{R}$. Then the pullback of S by Φ to M is defined as

$$(\Phi^* S)_x(\xi_1, ..., \xi_s) := S_{\Phi(x)}(\Phi_* \xi_1, ..., \Phi_* \xi_s). \tag{20}$$

Note that a Riemannian metric represents a $(0,2)$-tensor field with certain additional properties.

Definition 7 (Isometry)**.** Let $\Phi : M \to N$ be a smooth map. Then Φ is called an *isometry* if Φ is a diffeomorphism with

$$g = \Phi^* h \tag{21}$$

where $\Phi^* h$ is the pullback of the Riemannian metric by Φ. For two arbitrary tangent vectors ξ, $\zeta \in T_xM$ the pullback is defined as

$$g(\xi, \zeta) = \Phi^* h(\xi, \zeta) = h(\Phi_* \xi, \Phi_* \zeta) \tag{22}$$

which means that Φ_* is an isometry of the Euclidean vector spaces T_xM and $T_{\Phi(x)}N$.

One important aspect of isometries is given in the following proposition.

Proposition 8. *Given the isometry $\Phi: M \to N$ and the geodesic γ on M. Then $\Phi \circ \gamma$ is a geodesic on N.*

Proof. Let x_1 be an arbitrary point in M with a neighborhood $U_r(x_1)$ such that for any $x_2 \in U_r(x_1)$ there exists a unique geodesic γ in $U_r(x_1)$ that connects x_1 and x_2. Without loss of generality we can assume that γ is parametrized by unit length. Since γ is a geodesic, we have equality in the triangle inequality, namely

$$d(x_1,x_2) = d(x_1,\gamma(t)) + d(\gamma(t),x_2) \tag{23}$$

where d is the distance function on M induced by the Riemannian metric g, i. e., the length of the shortest curve connecting the two points. Due to the fact that g is the pullback of h by Φ, we have $d(x_1,x_2) = d'(\Phi(x_1),\Phi(x_2))$ for the distance function d' on N which is induced by h. Hence, the equation

$$d'(\Phi(x_1),\Phi(x_2)) = d'(\Phi(x_1),\Phi(\gamma(t))) + d'(\Phi(\gamma(t)),\Phi(x_2)) \tag{24}$$

holds which means that we again have equality in the triangle inequality, and thus $\Phi \circ \gamma$ is a geodesic as well. □

For more information on pushforward, pullback, and isometries see [10] and [9].

Proposition 9. *Given two Riemannian manifolds (M,g), (N,h), the isometry $\Phi : M \to N$ with $v, w \in M$, $x := \Phi(v)$, $y := \Phi(w) \in N$, and the tangent vectors $p \in T_vM$, $q := \Phi_* p \in T_xN$. Let $\gamma : [a,b] \to M$ be the geodesic connecting v and w with $\gamma(a) = v$, $\gamma(b) = w$, and ∇ the Levi-Civita connection on N. Then the operations of parallel transport and pushforward commute, i. e.,*

$$\Phi_* \Gamma_v^w(p) = \Gamma_x^y(\Phi_* p). \tag{25}$$

The notation used for the parallel transport which we used here describes parallel transport along the geodesics γ and $\Phi \circ \gamma$ connecting the two respective points. An alternative notation that involves the path along which the parallel transport occurs is $\Gamma(\gamma)_a^b(p) = \Gamma_v^w(p)$.

Proof. The parallel transport of p along γ is defined as the solution the the initial value problem

$$\begin{aligned} \tilde{\nabla}_{\dot{\gamma}(t)} \xi &= 0, \\ \xi_{\gamma(a)} &= p \end{aligned} \tag{26}$$

with $\tilde{\nabla}$ being a connection on M. Note that in a local trivialization this is a system of linear differential equations. Since the connection on N is required to be the Levi-Civita connection, the connection on M has to be chosen to suit our purposes. The pullback connection $\Phi^* \nabla$ defined by $(\Phi^* \nabla)_X Y := \nabla_{\Phi_* X} \Phi_* Y$ appears to be an adequate choice for $\tilde{\nabla}$ because it involves the isometry that connects the two manifolds. With this connection the first equation in (26) can be written and transformed as follows

$$\begin{aligned} & \tilde{\nabla}_{\dot{\gamma}(t)}\xi = 0 \\ \Leftrightarrow \quad & (\Phi^*\nabla)_{\dot{\gamma}(t)}\xi = 0 \\ \Leftrightarrow \quad & \nabla_{\Phi_*\dot{\gamma}(t)}\Phi_*\xi = 0 \\ \Leftrightarrow \quad & \nabla_{\dot{\tilde{\gamma}}(t)}\zeta = 0. \end{aligned} \tag{27}$$

In the last equivalence we introduced substitutions for the smooth curve $\tilde{\gamma} := \Phi \circ \gamma$ and the vector field $\zeta := \Phi_*\xi$. Note that $\tilde{\gamma}$ is the geodesic connecting x and y in N since isometries preserve geodesics as seen in Proposition 8, and that the pushforward of a vector field is only defined because Φ is an isometry and therefore, by definition, a diffeomorphism.

This shows that the system of differential equations (26) is equivalent to the problem

$$\begin{aligned} \nabla_{\dot{\tilde{\gamma}}(t)}\zeta &= 0, \\ \zeta_{\tilde{\gamma}(a)} &= q \end{aligned} \tag{28}$$

in the sense that if ξ is a solution to (26), then $\zeta := \Phi_*\xi$ is a solution to (28). With these two solutions of the differential equations this, together with the definition of the parallel transport of a tangent vector from one tangent space to another in Section 2.4, leads to

$$\Phi_*\Gamma_v^w(p) = \Phi_*\Gamma(\gamma)_a^b(p) = \Phi_*\xi_{\gamma(b)} = \zeta_{\tilde{\gamma}(b)} = \Gamma(\tilde{\gamma})_a^b(\Phi_* p) = \Gamma_x^y(q) \tag{29}$$

which proves the proposition. □

Proposition 10. *Let (M,g), (N,h) be two Riemannian, isometric manifolds, i. e., there exists an isometry $\Phi : M \to N$. Let f be a smooth cost function defined on N and $\tilde{f} := f \circ \Phi$ the composition of the cost function and the isometry. Then the equation*

$$\Phi_* \operatorname{grad}_M \tilde{f} = \operatorname{grad}_N f \tag{30}$$

holds.

Proof. To define the gradient of f at $y := \Phi(x)$ for an $x \in M$ the directional derivative of f at y in direction $\xi \in T_yN$ is needed. To obtain it, a smooth curve with the properties $\alpha(t) \in N$, $\alpha(0) = y$, $\dot{\alpha}(0) = \xi$ is required. The directional derivative then is

$$\mathrm{D}f(y)[\xi] = \tfrac{\mathrm{d}}{\mathrm{d}t} f(\alpha(t))|_{t=0} = \mathrm{D}f(y)[\dot{\alpha}(0)] = \mathrm{D}f(y)[\xi]. \tag{31}$$

The directional derivative of the function $\tilde{f}$ can be obtained by using the chain rule

$$\begin{aligned} \mathrm{D}\tilde{f}(x)[\zeta] = \mathrm{D}f(\Phi(x))[\zeta] &= \tfrac{\mathrm{d}}{\mathrm{d}t} f(\Phi(\beta(t)))|_{t=0} \\ &= \mathrm{D}f(\Phi(\beta(0)))[\mathrm{D}\Phi(\beta(0))[\dot{\beta}(0)]] \\ &= \mathrm{D}f(\Phi(x))[\mathrm{D}\Phi(x)[\zeta]] = \mathrm{D}f(y)[\mathrm{D}\Phi(x)[\zeta]]. \end{aligned} \tag{32}$$

The function $\beta(t)$ is a smooth curve in M with $\beta(0) = x$, $\dot{\beta}(0) = \zeta$. According to the Riesz representation theorem, the gradient of the function f at y is the unique element $\operatorname{grad}_N f(y)$ in T_yN that satisfies

$$\langle \xi, \operatorname{grad}_N f(y)\rangle_N = \mathrm{D}f(y)[\xi] \quad \forall \xi \in T_yN. \tag{33}$$

Analogously, for $\tilde{f}$ the gradient at x is defined as the unique element $\operatorname{grad}_M \tilde{f}(x)$ in T_xM with

$$\langle \zeta, \operatorname{grad}_M \tilde{f}(x)\rangle_M = \mathrm{D}f(y)[\mathrm{D}\Phi(x)[\zeta]] \quad \forall \zeta \in T_xM. \tag{34}$$

As we know from Definition 5, $\mathrm{D}\Phi(x)[\zeta] = \Phi_* \zeta$ is the pushforward of ζ by Φ. Thus, it is an element of T_yN. Since both manifolds have the same dimension and Φ is an isometry, Φ_* describes a vector space homomorphism from T_xM to $T_{\Phi(x)}N$. Furthermore, because we have $g = \Phi^* h$, the expression on the right side of the equation can be written as

$$\langle \zeta, \operatorname{grad}_M \tilde{f}(x)\rangle_M = \langle \Phi_* \zeta, \Phi_* \operatorname{grad}_M \tilde{f}(x)\rangle_N. \tag{35}$$

In combination this yields

$$\mathrm{D}f(y)[\xi] = \langle \xi, \Phi_* \operatorname{grad}_M \tilde{f}(x)\rangle_N \quad \forall \xi \in T_yN, \tag{36}$$

and as a result we have $\operatorname{grad}_N f(y) = \Phi_* \operatorname{grad}_M \tilde{f}(x)$ which concludes the proof. □

In the following, the relation of the BFGS algorithm on two isometric manifolds is analyzed. Therefore, we need to introduce the necessary parameters. Let (M,g) and (N,h) be Riemannian submanifolds related by the isometry $\Phi : M \to N$ with the associated pushforward operation Φ_*. Furthermore, the parallel transport along the shortest geodesic connecting the two points x and y in either manifold is denoted by Γ_x^y. The considered cost functions are $f : N \to \mathbb{R}$ and $\tilde{f} : M \to \mathbb{R}$.

Given $x_k \in M$, $\alpha_M \in \mathbb{R}$, and $\tilde{\eta} \in T_{x_k}M$ where $\tilde{\eta}$ is the search direction established in the previous iteration. The point $x_{k+1} \in M$ is then defined as $x_{k+1} = \exp_M(\alpha_M \tilde{\eta}_k)$. Furthermore, we have $p \in T_{x_{k+1}}M$, $\tilde{B}_k{:}T_{x_k}M \to T_{x_k}M$, $\hat{\tilde{B}}_k := \Gamma_{x_k}^{x_{k+1}} \circ \tilde{B}_k \circ \Gamma_{x_{k+1}}^{x_k}$. The BFGS update on M is then defined as

$$\begin{aligned} \tilde{B}_{k+1} p = \hat{\tilde{B}}_k p &+ \frac{\langle s_k, p\rangle_M}{\langle s_k, v_k\rangle_M} s_k + \frac{\langle v_k, \hat{\tilde{B}}_k v_k\rangle_M \langle s_k, p\rangle_M}{\langle s_k, v_k\rangle_M^2} s_k \\ &- \frac{\langle s_k, p\rangle_M}{\langle s_k, v_k\rangle_M} \hat{\tilde{B}}_k v_k - \frac{\langle v_k, \hat{\tilde{B}}_k p\rangle_M}{\langle s_k, v_k\rangle_M} s_k \end{aligned} \tag{37}$$

with $s_k = \Gamma_{x_k}^{x_{k+1}}(\alpha_M \tilde{\eta}_k)$ and $v_k = \operatorname{grad}_M \tilde{f}(x_{k+1}) - \Gamma_{x_k}^{x_{k+1}} \operatorname{grad}_M \tilde{f}(x_k)$.

To define the analogous algorithm on N, all the variables need to be transported to N and its tangent bundle TN via the isometry. Doing this leads to $y_k := \Phi(x_k) \in N$, $\alpha_N := \alpha_M$, and $\eta := \Phi_* \tilde{\eta}$ which is assumed for now but will be shown as a byproduct of the following proof. Due to the fact that Φ is an isometry the point that is defined as

$y_{k+1} = \Phi(x_{k+1})$ is the same point as $y_{k+1} := \exp(\alpha_N \eta)$ produces. Given in addition are $q \in T_{y_{k+1}}N$, $B_k\colon T_{y_k}N \to T_{y_k}N$, and $\hat{B}_k := \Gamma_{y_k}^{y_{k+1}} \circ B_k \circ \Gamma_{y_{k+1}}^{y_k}$. To simplify the notation we define $\Phi_* := \Phi_*[x_{k+1}]$. Then the BFGS update on N has the form

$$\begin{aligned} B_{k+1}q = \hat{B}_k q + \frac{\langle \sigma_k, q\rangle_N}{\langle \sigma_k, \gamma_k\rangle_N}\sigma_k + \frac{\langle \gamma_k, \hat{B}_k\gamma_k\rangle_N \langle \sigma_k, q\rangle_N}{\langle \sigma_k, \gamma_k\rangle_N^2}\sigma_k \\ - \frac{\langle \sigma_k, q\rangle_N}{\langle \sigma_k, \gamma_k\rangle_N}\hat{B}_k\gamma_k - \frac{\langle \gamma_k, \hat{B}_k q\rangle_N}{\langle \sigma_k, \gamma_k\rangle_N}\sigma_k \end{aligned} \tag{38}$$

with $\sigma_k := \Gamma_{y_k}^{y_{k+1}}(\alpha_N \eta)$ and $\gamma_k := \operatorname{grad}_N f(y_{k+1}) - \Gamma_{y_k}^{y_{k+1}} \operatorname{grad}_N f(y_k)$. Mind that the equation

$$\begin{aligned} \gamma_k &= \operatorname{grad}_N f(y_{k+1}) - \Gamma_{y_k}^{y_{k+1}} \operatorname{grad}_N f(y_k) \\ &= \Phi_*(\operatorname{grad}_M \tilde{f}(x_{k+1}) - \Gamma_{x_k}^{x_{k+1}} \operatorname{grad}_M \tilde{f}(x_k)) \\ &= \Phi_* v_k \end{aligned} \tag{39}$$

follows from Proposition 9 and 10.

These two updates already look pretty similar, especially if $\sigma_k = \Phi_* s_k$ holds. However, this is not immediately obvious, but will be shown in the course of the proof of the following proposition.

Proposition 11. *Given the two isometric Riemannian submanifolds (M, g), (N, h) which are connected by the isometry $\Phi : M \to N$ with the associated pushforward operation $\Phi_* : TM \to TN$. Let $\tilde{B}_k$ and B_k be the approximations to the inverse Hessians on the two manifolds with the update rule defined as above and $p \in T_{x_k}M$, $q := \Phi_* p \in T_{y_k}N$. Then the equation*

$$B_k q = B_k \Phi_* p = \Phi_* \tilde{B}_k p \tag{40}$$

holds. That is, the two BFGS updates are related via the pushforward operation.

Proof. By induction:
Before we start, we have to show that $\sigma_0 = \Phi_* s_0$. Since $\tilde{B}_0$ and B_0 are both the identity on the respective tangent space they are defined on, the first search direction on M is $\tilde{\eta}_0 = -\operatorname{grad}_M \tilde{f}(x_0)$ while on N it is $\eta_0 = -\operatorname{grad}_M f(y_0)$. Because of Proposition 10 and the linearity of the pushforward, this means we have $\eta_0 = \Phi_* \tilde{\eta}_0$. It can be assumed that the line search produces the same step length α for both algorithms due to the way the functions are defined. Then the equation

$$\sigma_0 = \Gamma_{y_0}^{y_1}(\alpha\eta_0) = \Gamma_{y_0}^{y_1}(\alpha\Phi_*\tilde{\eta}_0) = \Phi_*\Gamma_{x_0}^{x_1}(\alpha\tilde{\eta}) = \Phi_* s_0 \tag{41}$$

holds per definition of σ_k, s_k and because of Proposition 9.

Now the induction can be started. For $k = 1$ we have $\hat{\tilde{B}}_1 p = p$ and $\hat{B}_1 q = q$ with $p \in T_{x_1}M$ and $q := \Phi_* p \in T_{y_1}N$ since $B_0 = id_{T_{x_0}M}$, $\tilde{B}_0 = id_{T_{y_0}N}$. In combination with the fact that we are using a pullback metric we get

$$B_1 q = \Phi_*(\tilde{B}_1 p). \tag{42}$$

The notation $\Phi_* := \Phi_*[x_1]$ is used to simplify the expression. This means that the proposition holds for $k = 1$.

Assume that the proposition is true for an arbitrary $k \in \mathbb{N},\ k \geq 1$. Then for the search directions

$$\eta_k = -B_k \operatorname{grad}_N f(y_k) = -B_k \Phi_* \operatorname{grad}_M \tilde{f}(x_k) = -\Phi_* \tilde{B}_k \operatorname{grad}_M \tilde{f}(x_k) = \Phi_* \tilde{\eta}_k \tag{43}$$

holds. With identical step lengths this leads to $\sigma_k = \Phi_* s_k$ in the same way as for the case $k = 0$. Hence, for $k+1$ we get the following equation

$$B_{k+1} q = \Phi_*(\tilde{B}_{k+1} p). \tag{44}$$

with $\Phi_* := \Phi_*[x_{k+1}]$, $p \in T_{x_{k+1}}$, and $q := \Phi_* p \in T_{y_{k+1}}$
In order for this equation to hold, the equation

$$\hat{B}_k q = \Phi_*[x_{k+1}](\hat{\tilde{B}}_k p) \tag{45}$$

has to hold. To show this, let us examine what is done in detail:

$$\begin{aligned} \hat{B}_k q &= \left(\Gamma_{y_k}^{y_{k+1}} \circ B_k \circ \Gamma_{y_{k+1}}^{y_k} \circ \Phi_*[x_{k+1}]\right)(p) \\ &= \left(\Gamma_{y_k}^{y_{k+1}} \circ B_k \circ \Phi_*[x_k] \circ \Gamma_{x_{k+1}}^{x_k}\right)(p) \\ &= \left(\Gamma_{y_k}^{y_{k+1}} \circ \Phi_*[x_k] \circ \tilde{B}_k \circ \Gamma_{x_{k+1}}^{x_k}\right)(p) \\ &= \left(\Phi_*[x_{k+1}] \circ \Gamma_{x_k}^{x_{k+1}} \circ \tilde{B}_k \circ \Gamma_{x_{k+1}}^{x_k}\right)(p) \\ &= \Phi_*[x_{k+1}](\hat{\tilde{B}}_k p) \end{aligned} \tag{46}$$

The other transformations hold because of Proposition 9 and the assumption made for the induction. This concludes the proof. □

We can now show that the BFGS method is invariant under isometries, i. e., given two isometric Riemannian manifolds (M,g), (N,h) with the isometry $\Phi : M \to N$ and the cost functions $f : N \to \mathbb{R}$, $\tilde{f} : M \to \mathbb{R}$ with $\tilde{f} := \Phi^* f$. Starting from the two points $x_0 \in M$, $y_0 := \Phi(x_0) \in N$ the BFGS methods on the two manifolds converge to $x_*, y_* = \Phi(x_*)$, respectively.

To show this, we consider the BFGS step from the starting point $x_0 \in M$ and, respectively, the BFGS step on N starting from $y_0 := \Phi(x_0)$. As seen in Proposition 10 we have

$$\Phi_* \operatorname{grad}_M \tilde{f}(x_0) = \operatorname{grad}_N f(y_0). \tag{47}$$

Furthermore, we know from Proposition 11 that the equation

$$\begin{gathered} B_0 \operatorname{grad}_N f(y_0) = B_0 \Phi_* \operatorname{grad}_M \tilde{f}(x_0) = \Phi_* \tilde{B}_0 \operatorname{grad}_M \tilde{f}(x_0) \\ \text{with } B_0 = id_{T_{y_0}N} \text{ and } \tilde{B}_0 = id_{T_{x_0}M} \end{gathered} \tag{48}$$

holds. This means that the search directions are equivalent under pushforward by Φ. According to the RBFGS algorithm the next iteration point in M then lies on the geodesic that emanates from x_0 in direction $\eta := -\tilde{B}\,\mathrm{grad}_M \tilde{f}$ while in N it is placed on the one geodesic starting in y_0 in direction $\Phi_* \eta$. Its exact location is determined by using a line search algorithm which yields the step lengths α_M and α_N. Since the function $\tilde{f}$ is defined as $\Phi^* f$ and with the backtracking algorithm 8 chosen as the line search, the two step lengths are equal to one another. Therefore, we will omit the subscript indices and simply write α for the step length. Since Φ is an isometry, it maps geodesics on M onto geodesics on N, and in conclusion we have $y_1 = \Phi(x_1)$. Thus, the first iterate is invariant.

Now, assume that $y_k = \Phi(x_k)$ for $k \geq 1 \in \mathbb{N}$. Then, analogous to the case where $k = 0$, we have

$$\begin{aligned} \Phi_* \,\mathrm{grad}_M \tilde{f}(x_k) &= \mathrm{grad}_N f(y_k), \\ B_k \,\mathrm{grad}_N f(y_k) = B_k \Phi_* \,\mathrm{grad}_M \tilde{f}(x_k) &= \Phi_* \tilde{B}_k \,\mathrm{grad}_M \tilde{f}(x_k) \end{aligned} \tag{49}$$

according to Propositions 10 and 11. As before, this means that the search direction for the BFGS algorithm on N is the pushforward by Φ of the search direction for the BFGS algorithm on M. Hence, the geodesic describing the search direction on N is the image of Φ of the geodesic on M, and they lead (with step length $\alpha_M = \alpha_N$) to the iteration points $x_{k+1} \in M$ and $y_{k+1} \in N$ with $y_{k+1} = \Phi(x_{k+1})$. By induction this means that the BFGS algorithms on the two isometric Riemannian manifolds converge to the points x^*, y^* with $y^* = \Phi(x^*)$. Thus, the BFGS algorithm is invariant under isometries. Figure 5 illustrates this.

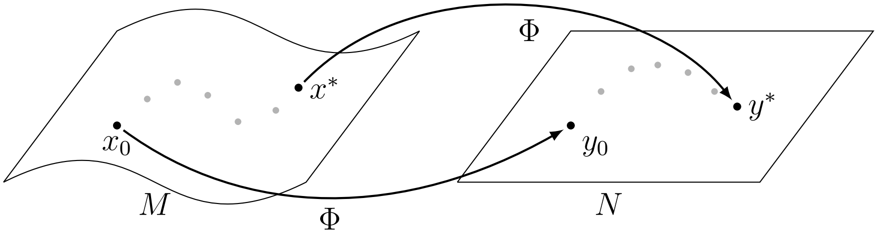


Figure 5: The RBFGS algorithm on isometric manifolds. The gray points represent the iteration points of the RBFGS algorithm on the respective manifolds.

In summary these propositions yield the following result:

Corollary 12. *For every BFGS algorithm on Riemannian manifolds that are isometric to $\mathbb{R}^n$ there is an equivalent BFGS algorithm on $\mathbb{R}^n$. This means that the convergence proof that is given in [4] for the unrestricted BFGS method in $\mathbb{R}^n$ can be applied to these algorithms, and the global superlinear convergence rate carries over to these Riemannian manifolds.*

Bibliography

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008. Cited pp. 396 and 398.

[2] R. L. Adler, J.-P. Dedieu, J. Y. Margulies, M. Martens, and M. Shub. Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA J. of Numerical Analysis*, 22:359–390, 2002. Cited p. 395.

[3] I. Brace and J. H. Manton. An improved BFGS-on-manifold algorithm for computing weighted low rank approximations. In *Proceedings of the MTNS*, pages 1735–1738, 2006. Cited pp. 396 and 404.

[4] J. E. Dennis and J. J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Math. Comp.*, 28:549–560, 1974. Cited p. 411.

[5] A. Edelmann, T. Arias, and S. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix. Anal. Appl.*, 20(2):303–353, 1998. Cited p. 395.

[6] D. Gabay. Minimizing a differentiable function over a differential manifold. Research Report RR-0009, INRIA, 1980. Cited p. 395.

[7] D. Gabay. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37:177–219, 1982. Cited p. 395.

[8] P. E. Gill, W. Murray, and M. H. Wright. *Practical optimization*. Academic Press, 1981. Cited p. 395.

[9] J. Jost. *Riemannian geometry and geometric analysis*. Springer, 5th edition, 2008. Cited p. 406.

[10] J. M. Lee. *Introduction to Smooth Manifolds*. Springer, 2002. Cited p. 406.

[11] D. G. Luenberger and Y. Ye. *Linear and nonlinear programming*. Springer, 3rd edition, 2008. Cited p. 395.

[12] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999. Cited pp. 397 and 404.

[13] E. Polak. *Optimization. Algorithms and consistent approximations*. Springer, 1997. Cited p. 395.

[14] C. Qi, K. A. Gallivan, and P.-A. Absil. Riemannian BFGS algorithm with applications. In M. Diehl, F. Glineur, E. Jarlebring, and W. Michiels, editors, *Recent Advances in Optimization and its Applications in Engineering*, pages 183–192. Springer, 2010. Cited pp. 396 and 399.

[15] W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 2012. To appear. Cited p. 396.

[16] C. Udrişte. *Convex Functions and Optimization Methods on Riemannian Manifolds*. Kluwer Academic Publishers, 1994. Cited p. 395.

[17] K. Yosida. *Functional analysis*. Springer, 6th edition, 1980. Cited p. 400.

On compact invariant sets of some systems arising in cosmology

Konstantin E. Starkov
Instituto Politecnico Nacional
CITEDI
Tijuana Baja California, Mexico
konst@citedi.mx

Abstract. This paper contains a review of some results of the author relating to studies of a location of all compact invariant sets of the Bianchi VIII / Bianchi IX Hamiltonian systems and the static spherically symmetric Einstein-Yang-Mills equations. In particular, we describe some invariant domains which do not contain any periodic orbits; any homoclinic orbits; any heteroclinic orbits. In addition, for the Bianchi VIII / Bianchi IX Hamiltonian systems we have established nonchaoticity of dynamics in some invariant domains with help of a localization of their omega-limit sets.

1 Introduction

Spatially homogeneous cosmological models arisen as solutions of the Einstein field equations with a perfect fluid as a source and under some additional conditions attract attention of many specialists in nonlinear dynamics because these models can be efficiently analysed by powerful methods of modern qualitative theory of ordinary differential equations and dynamical system theory. Nowadays it is well-recognized, see e.g. [2, 12], that major problems respecting dynamical systems of the cosmological origin are related to a description of their asymptotic states. Therefore it is not surprising that during the last two decades there has been demonstrated an active interest to studies of a long-time behavior of dynamical systems of the cosmological origin near the big bang, and at late positive times. It is well-known that the long-time dynamics of a system may be investigated via analysis of $\omega-$limit sets and $\alpha-$limit sets of its trajectories. In this case problems of the existence/nonexistence of non-empty $\omega-$limit sets or/and $\alpha-$limit sets for trajectories in global or in some invariant sets are naturally appeared. These problems may be tractable as problems of a localization of compact invariant sets in the chosen invariant set (domain in some cases) U of the state space of a dynamical system. Here a localization means a description of the location of all compact invariant sets in U by means of equations and inequalities depending on parameters of the system. Finding a localization domain, i.e. a domain which contains all compact invariant sets is of a substantial interest because of the potential application of computer-based methods for its search narrowed in the localization domain. The localization analysis may give information about important qualitative features of a long-time behaviour of the system, for example, nonchaoticity. We notice that the existence or the nonexistence of periodic orbits expresses the fact of the presence or the lack correspondingly of repeatable behavior. Any globally bounded motion of the system is contained in one of compact invariant sets. Studies regarding to compact invariant sets of cosmological systems have been appeared more than 20 years ago, see e.g. papers [5, 13]. In this work a review of recent results of the author connecting to two cosmological systems is presented; complete versions are contained in papers [10, 11]. We examine global dynamics relating to the localization problem of compact invariant sets of the Bianchi VIII and Bianchi IX Hamiltonian systems which are also known as Mixmaster universe models, [3, 9], and the static spherically symmetric Einstein- Yang- Mills (EYM) equations, see [1]. In essence, our approach is based on exploiting the localization method of all compact invariant sets proposed earlier by Krishchenko and the author in [6]. Here the principal idea is to study extrema of some differentiable functions called localizing which are restricted on trajectories taken from compact invariant sets; this idea is realized with help of using the first order extremum conditions and the high order extremum conditions.

2 Some helpful results

We consider a nonlinear system

$$\dot{x} = F(x) \tag{1}$$

where $x \in \mathbb{R}^n$, $F(x) = (F_1(x), \dots, F_n(x))^T$ is a differentiable vector field. Our basic tool consists in using the following assertions, [6]. Let U be some domain in $\mathbb{R}^n$.

Proposition 1. *1. For any $h(x) \in C^\infty(\mathbb{R}^n)$ all compact invariant sets of the system (1) located in U are contained in the set defined by the formula*

$$K(U;h) := \{x \in U \mid h_{\inf}(U) \le h(x) \le h_{\sup}(U)\}$$

as well. 2. If $S(h) \cap U = \varnothing$ then (1) has no compact invariant sets in U.

Proposition 2. *Let $h_m(x), m = 1, 2, \dots$ be a sequence of functions from $C^\infty(\mathbb{R}^n)$. Sets*

$$K_1 = K(U;h_1), \quad K_m = K_{m-1} \cap K_{m-1,m}, \quad m > 1,$$

with

$$K_{m-1,m} = \{x : h_{m,\inf} \le h_m(x) \le h_{m,\sup}\},$$
$$h_{m,.\sup} = \sup_{S_{h_m} \cap K_{m-1}} h_m(x),$$
$$h_{m,\inf} = \inf_{S_{h_m} \cap K_{m-1}} h_m(x).$$

contain all compact invariant sets of the system (1) and $K_1 \supseteq K_2 \supseteq \dots \supseteq K_m \supseteq \dots$.

These assertions in a combination with the LaSalle theorem are useful in a derivation of our results; in addition, we apply computations in level sets of Hamiltonians/ first integrals in order to fulfill localization analysis.

Below by $C\{M\}$ we denote the complement to the set $M \subset \mathbb{R}^n$.

3 On compact invariant sets of Bianchi VIII and Bianchi IX Hamiltonian systems

In this section we examine the Bianchi VIII and Bianchi IX Hamiltonian systems. The first system is defined by the following equations:

$$\begin{aligned}\dot{y}_1 &= \frac{1}{2}y_1(z_2-z_1),\\ \dot{y}_2 &= \frac{1}{2}y_2(z_1+z_2),\\ \dot{y}_3 &= y_3z_3,\\ \dot{z}_1 &= 2(y_1+y_2)(y_1-y_2-y_3),\\ \dot{z}_2 &= 2y_3(-y_1+y_2+y_3),\\ \dot{z}_3 &= (y_1+y_2)^2-y_3^2.\end{aligned} \tag{2}$$

with the Hamiltonian

$$H=(y_1+y_2)^2+y_3(2y_2-2y_1+y_3)-z_2z_3+\frac{1}{4}(z_1^2-z_2^2).$$

The system (2) is given in the form obtained from the standard equations, [11], with help of applying some linear change of coordinates. The Bianchi IX Hamiltonian system is written as

$$\begin{aligned}\dot{y}_1 &= y_1(z_1-z_2-z_3),\\ \dot{y}_2 &= y_2(-z_1+z_2-z_3),\\ \dot{y}_3 &= y_3(-z_1-z_2+z_3),\\ \dot{z}_1 &= -y_1(y_1-y_2-y_3),\\ \dot{z}_2 &= -y_2(-y_1+y_2-y_3),\\ \dot{z}_3 &= -y_3(-y_1-y_2+y_3).\end{aligned} \tag{3}$$

see in [9]. The system (3) has the Hamiltonian $H=G(z_1,z_2,z_3)+G(y_1,y_2,y_3)$, with $G(z_1,z_2,z_3)=z_1^2+z_2^2+z_3^2-2z_1z_2-2z_3z_2-2z_1z_3$.

Let $\Pi_i(y)=\{y_i=0\}$, $i=1,2,3$. We introduce the notation $M=\mathbb{R}^6-\{\cup_{i=1}^3\Pi_i(y)\}$. Let $\alpha_j\in\{+;0;-\}$, $j=1,2,3$. We define the partition of $\mathbb{R}^6$ into 27 sets

$$M_{\alpha_1\alpha_2\alpha_3}:=\cap_{j=1}^3\left\{\begin{array}{l}\alpha_j=+,y_j>0;\\ \alpha_j=0,y_j=0;\\ \alpha_j=-,y_j<0.\end{array}\right\}.$$

By N we denote the subset of M containing all $M_{\alpha_1\alpha_2\alpha_3}$ with nonzero indices. It is clear that for the both of systems (2) and (3) each of sets $M_{\alpha_1\alpha_2\alpha_3}$ is an invariant set; N is an invariant set too. It is easy to see that the set M_{000} consists of equilibrium points for the both of systems. Below it is convenient to describe our assertions concerning the location of compact invariant sets in each of sets $M_{\alpha_1\alpha_2\alpha_3}$ separately, with nonzero $(\alpha_1,\alpha_2,\alpha_3)$.

Firstly, we present results of analysis for the Bianchi VIII Hamiltonian system:

Theorem 3. *1. All compact invariant sets of the system (2) contained in N are located in the set $M_{+-+} \cup M_{-+-}$. 2. Let $(\alpha_1,\alpha_2,\alpha_3) \in \{(+-+),(-+-)\}$. Then the ω−limit set for trajectories bounded with $t > 0$ and laying in the set $M_{\alpha_1\alpha_2\alpha_3} \cap H^{-1}(0)$ is contained in the plane $z_2 + z_3 = 0$. Further, there are no periodic orbits and neither homoclinic, nor heteroclinic orbits of the system (2) contained in $(M_{+-+} \cup M_{-+-}) \cap H^{-1}(0)$.*

Theorem 4. *1. There are no compact invariant sets contained in any of sets M_{0++}; M_{0--}; M_{++0}; M_{--0};M_{+0-}; M_{-0+}; $M_{00\pm}$;$M_{0\pm0}$;$M_{\pm00}$; 2-1). Let $(\alpha_1,\alpha_2,\alpha_3) \in \{(+0+),(-0-)\}$. Then the ω−limit set for trajectories bounded with $t > 0$ and laying in the set $M_{\alpha_1\alpha_2\alpha_3} \cap H^{-1}(0)$ is contained in the plane $z_2 - z_1 + 2z_3 = 0$. 2-2). Let $(\alpha_1,\alpha_2,\alpha_3) \in \{(+-0),(-+0)\}$. Then the ω−limit set for trajectories bounded with $t > 0$ and laying in the set $M_{\alpha_1\alpha_2\alpha_3} \cap H^{-1}(0)$ is contained in the plane $z_2 = 0$. 2-3). Let $(\alpha_1,\alpha_2,\alpha_3) \in \{(0+-),(0-+)\}$. Then the ω−limit set for trajectories bounded with $t > 0$ and laying in the set $M_{\alpha_1\alpha_2\alpha_3} \cap H^{-1}(0)$ is contained in the plane $z_1 + z_2 + 2z_3 = 0$. Besides, in each of these cases there are no periodic orbits and neither homoclinic, nor heteroclinic orbits contained in the set $M_{\alpha_1\alpha_2\alpha_3} \cap H^{-1}(0)$.*

Now we formulate results for the Bianchi IX Hamiltonian system:

Theorem 5. *1. All compact invariant sets of the system (3) contained in N are located in the set $M_{---} \cup M_{+++}$. 2. All compact invariant sets in $(M_{---} \cup M_{+++}) \cap H^{-1}(0)$ are contained in the plane $z_1 + z_2 + z_3 = 0$ as well. Let $(\alpha_1,\alpha_2,\alpha_3) \in \{(+++),(---)\}$. Then the ω−limit set for trajectories bounded with $t > 0$ and laying in the set $M_{\alpha_1\alpha_2\alpha_3} \cap H^{-1}(0)$ is contained in the plane $z_1 = z_2 = z_3 = 0$. Further, there are no periodic/ homoclinic/ heteroclinic orbits of the system (3) contained in the set $(M_{---} \cup M_{+++}) \cap H^{-1}(0)$.*

Proposition 6. *There are no compact invariant sets contained in any of sets M_{0+-}; M_{0-+}; M_{+-0};M_{-+0}; M_{+0-};M_{-0+};$M_{00\pm}$;$M_{0\pm0}$;$M_{\pm00}$.*

4 On compact invariant sets of the static spherically symmetric EYM equations

Now we consider the localization problem of compact invariant sets for the static spherically symmetric EYM equations

$$\begin{aligned}
\dot r &= rN,\\
\dot W &= rU,\\
\dot N &= (k-N)N - 2U^2,\\
\dot k &= s(1-2ar^2) + 2U^2 - k^2,\\
\dot U &= sWT + (N-k)U,\\
\dot T &= 2UW - NT.
\end{aligned} \tag{4}$$

defined on $\mathbb{R}^6$, with a cosmological real constant $a, s = 1$ or -1, [1]. The sign $s = 1$ corresponds the case when the physical time is considered as a temporal variable, while $s = -1$ corresponds the case when the physical time is considered as a spatial variable.

It is easy to see that the hypersurface $2kN - N^2 - 2U^2 - s(1 - T^2 - ar^2) = 0$ is invariant for these equations. Similarly to [7], we solve the last equality respecting U^2 and substitute the corresponding expression into the 4th equation of (4). As a result, we come to the system

$$\begin{aligned}
\dot{x}_1 &= X_1(x_1,...,x_6) = x_1x_3,\\
\dot{x}_2 &= X_2(x_1,...,x_6) = x_1x_5,\\
\dot{x}_3 &= X_3(x_1,...,x_6) = (x_4 - x_3)x_3 - 2x_5^2,\\
\dot{x}_4 &= X_4(x_1,...,x_6) = -(x_4 - x_3)^2 - asx_1^2 + sx_6^2,\\
\dot{x}_5 &= X_5(x_1,...,x_6) = sx_2x_6 - (x_4 - x_3)x_5,\\
\dot{x}_6 &= X_6(x_1,...,x_6) = 2x_2x_5 - x_3x_6.
\end{aligned} \tag{5}$$

which is equivalent to (4) on this hypersurface; here and below we use notations $x_1 = r; x_2 = W; x_3 = N; x_4 = k; x_5 = U; x_6 = T$. Next, by $\varphi(x,t)$ we denote the solution of (5), with initial condition $\varphi(x,0) = x$. The system (5) possesses two first integrals: $G_1(x) = 2x_3x_4 - x_3^2 + s(ax_1^2 + x_6^2) - 2x_5^2$, $G_2(x) = x_2^2 - x_1x_6$. In what follows, let $N_1(0) := \{G_1(x) = 0\}; N_2(1) = \{G_2(x) = 1\}$. Besides, by Π_j we denote the hyperplane $\{x_j = 0\}$; $j = 1,\dots,6$. It has been shown that invariant surfaces $N_1(0)$ and $N_2(1)$ appear as constraints in the process of desingularization of the field equations and, thus, dynamical analysis of the system (5) restricted on the invariant set $N_1(0) \cap N_2(1)$ has a physical meaning, see details in the paper [1]. Below we present results concerning the location of compact invariant sets of the system (5) in the inside the invariant set $N_1(0) \cap N_2(1)$.

Theorem 7. *1. Let $s = 1$ and $a < 0$. Then all compact invariant sets contained in the set $C\{\Pi_1\}$ are located in the set*

$$K_{1a}(C\{\Pi_1\}; h_1) = \{-\sqrt{-\frac{a}{3}} \le \frac{x_3}{x_1} \le \sqrt{-\frac{a}{3}}\}$$

as well. 2. Let $s = 1$ and $a > 0$. Then there are no compact invariant sets contained in the set $C\{\Pi_1\}$.

Proposition 8. *Let $s = 1$; $a < 0$ and we consider the location of compact invariant sets in $C\{\Pi_1\}$. 1. If some compact invariant set $\Gamma \subset \{x_4x_1^{-1} > 0\}$ then $\Gamma \subset \{x_4x_1^{-1} \ge 2\sqrt{-\frac{a}{3}}\}$. 2. Further, if $\Gamma \subset \{x_4x_1^{-1} < 0\}$ then $\Gamma \subset \{x_4x_1^{-1} \le -2\sqrt{-\frac{a}{3}}\}$.*

Now we take the case $s = 1; a > 0; x \in \Pi_1$. Here we formulate

Theorem 9. *All compact invariant sets contained in $N_1(0)$ are located in the plane $\Pi_3 \cap \Pi_4$ as well.*

The further analysis leads to

Proposition 10. *All compact invariant sets contained in $N_1(0) \cap N_2(1)$ are located in the set $K_{1a}(N_1(0) \cap N_2(1), h_4) = \{x_1x_2 = 0\}$ as well.*

Examining the location of compact invariant sets in the invariant set

$$\Pi_1 \cap N_1(0) \cap N_2(1) = \Pi_1 \cap N_1(0) \cap \{x_2 = \pm 1\}$$

we come to the following conclusion:

Theorem 11. *The unique compact invariant set contained in* $\Pi_1 \cap N_1(0) \cap N_2(1)$ *is a pair of equilibrium points* $O_1 := (0,1,0,0,0,0)^T$ *and* $O_2 = (0,-1,0,0,0,0)^T$.

Further, in case $s = -1; a < 0$ the following assertion holds:

Theorem 12. *The set of equilibrium points consists of only two points* O_1 *and* O_2*. Further, there are neither periodic orbits nor homoclinic and heteroclinic orbits for the system* (5)*.*

Finally, we consider the case $s = -1; a > 0$. Here we present the following assertion:

Theorem 13. *The set of equilibrium points consists of only two points* O_1 *and* O_2*. Further, there are neither periodic orbits nor homoclinic orbits and heteroclinic orbits for the system* (5)*.*

5 Concluding remarks about nonchaoticity of some cosmological systems

In cases when we can demonstrate that all $\omega-$limit sets of trajectories of the system considered in the invariant domain U are located in some plane as well we conclude that the phase flow of this system has no recurrence property in U in the sense of Birkhoff. In particular, we have established this feature of dynamics for phase flows of the Bianchi VIII and IX Hamiltonian systems for some planes of dimension 2. Since recurrence property is the standard ingredient in the definition of the deterministic chaos one may deduce that dynamics of the Bianchi VIII and IX Hamiltonian systems does not exhibit the deterministic chaos inside U. This fact corresponds to results obtained in [4] for the system (3).

During a preparation of this text the author has discovered that the cosmological system formed by the minimally coupled field

$$\begin{aligned}
\dot{x}_1 &= -y_1, \\
\dot{x}_2 &= \frac{1}{x_1^2} y_2, \\
\dot{y}_1 &= 2kx_1 - 4x_1^3(\Lambda + m^2 x_2^2) + \frac{1}{x_1^3} y_2^2 + \frac{2\kappa^2}{x_1^3 x_2^2}, \\
\dot{y}_2 &= -2m^2 x_1^4 x_2 + \frac{2\kappa^2}{x_1^2 x_2^3}.
\end{aligned}$$

see in [8], with a positive cosmological constant Λ and parameters $k = 1; \kappa^2 > 0; m^2 > 0$, has the same property. Namely, we have found some level sets of the Hamiltonian

$$H = \frac{1}{2}(-y_1^2 + \frac{1}{x_1^2} y_2^2) - kx_1^2 + \Lambda x_1^4 + m^2 x_1^4 x_2^2 + \frac{\kappa^2}{x_1^2 x_2^2}$$

in which all $\omega-$ limit sets are contained in some plane of dimension 2; this fact implies nonchaoticity of corresponding dynamics.

Bibliography

[1] P. Breitenlohner, P. Forgács, and D. Maison. Static spherically symmetric solutions of the Einstein-Yang-Mills equations. *Communications in Mathematical Physics*, 163:141–172, 1994. Cited pp. 414, 416, and 417.

[2] A. A. Coley. *Dynamical Systems and Cosmology*. Kluwer, 2003. Cited p. 413.

[3] G. Contopoulos, B. Grammaticos, and A. Ramani. The Mixmaster universe model, revisited. *Journal of Physics A*, 27:5795–5799, 1994. Cited p. 414.

[4] R. Cushman and J. Sniatyski. Local integrability of the Mixmaster model. *Reports on Mathematical Physics*, 36:75–89, 1995. Cited p. 418.

[5] C. Hewitt and J. Wainwright. A dynamical systems approach to Bianchi cosmologies: Orthogonal models of class B. *Classical and Quantum Gravity*, 10:99–124, 1993. Cited p. 414.

[6] A. P. Krishchenko and K. E. Starkov. Localization of compact invariant sets of the Lorenz system. *Physics Letters A*, 353:383–388, 2006. Cited p. 414.

[7] J. Llibre and C. Valls. On the integrability of the Einstein-Yang-Mills equations. *Journal of Mathematical Analysis and Applications*, 336:1203–1230, 2007. Cited p. 417.

[8] A. J. Maciejewski, M. Przybylska, T. Stachowiak, and M. Szydlowski. Global integrability of cosmological scalar fields. *Journal of Physics A*, 41:465101 (26 pages), 2008. Cited p. 418.

[9] A. J. Maciejewski, J.-M. Strelcyn, and M. Szydlowski. Nonintegrability of Bianchi VIII Hamiltonian system. *Journal of Mathematical Physics*, 42:1728–1741, 2001. Cited pp. 414 and 415.

[10] K. E. Starkov. Compact invariant sets of the static spherically symmetric Einstein-Yang-Mills equations. *Physics Letters A*, 374:1728–1731, 2010. Cited p. 414.

[11] K. E. Starkov. Compact invariant sets of the Bianchi VIII and Bianchi IX Hamiltonian systems. *Physics Letters A*, 375:3184–3187, 2011. Cited pp. 414 and 415.

[12] J. Wainwright and G. F. R. Ellis. *Dynamical Systems in Cosmology*. Cambridge University Press, 1997. Cited p. 413.

[13] J. Wainwright and L. Hsu. A dynamical systems approach to Bianchi cosmologies: Orthogonal models of class A. *Classical and Quantum Gravity*, 6:1409–1431, 1989. Cited p. 414.

On state observers

Jochen Trumpf
Australian National University
Canberra, Australia
Jochen.Trumpf@anu.edu.au

Abstract. This paper contains a detailed proof that Luenberger's Sylvester equations characterize asymptotic (functional) state observers in the category of linear time-invariant finite-dimensional systems in input / state / output form. The proof is given entirely by state space and transfer function methods and is based on a simple necessary condition for output stability of certain cascading systems. Two discussions of how these results relate to prior results in the literature and of some implications for the existence of state observers with an invertible output map conclude the paper.

1 Introduction

The problem of observing the state or, more generally, a linear function of the state of a linear time-invariant finite-dimensional system in input / state / output form has been researched for close to five decades and many results have been obtained during that time. Arguably the most important of these results is the recognition that if the state is at all observable from the input and the output then it is also observable through the use of an auxiliary system, the *observer*, that takes the input and the output of the observed system as its input and produces an estimate for the required function of the state of the observed system as its output. Moreover, this observer can itself be taken as a linear time-invariant finite-dimensional system in input / state / output form, allowing to treat the whole problem within a simple and nice categorical setting. Much of the research on this problem has hence focussed on existence conditions for such observers with particular additional traits (such as minimal dynamic order, or invertible output map) and their "design" (meaning construction).

A related but somewhat less thoroughly researched topic is that of *characterizing* such observers, i.e. answering the question: "Exactly when is a *given* such system an observer for the *given* function of the state of the *given* system?" If a good answer to this question can be provided in terms of a set of necessary and sufficient equations and inequalities, complete solutions to the existence (resp. design) problem are obtained via solvability (resp. solutions) of these (in-)equalities.

In this paper I provide a detailed proof that Luenberger's Sylvester equations [15, Equation (5.5)] characterize asymptotic (functional) state observers in the category of linear time-invariant finite-dimensional systems in input / state / output form, where the observed system is assumed to have no stable uncontrollable modes. I explain why this latter assumption is necessary. I also provide a detailed discussion of related results in the literature in Section 3.1, and highlight Uwe Helmke's contributions in this area. These results are preceded by a discussion of output stability in Section 2 and followed by some results on existence of observers with invertible output map in Section 4.

I would like to take this opportunity to thank Uwe Helmke for his advice and friendship over the years. Much of what I know about mathematical system theory I have learned from Uwe, first as his student and later as a colleague. It is an absolute pleasure to dedicate this paper to him on the occasion of his 60th birthday.

2 Output stability for a simple cascading system

This section constains a review of some basic results on output stability, including for a certain type of simple cascading system that typically appears in observer error analysis.

Consider the following linear, time-invariant, finite-dimensional state space system,

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad x(0) = x_0, \\ y &= Cx + Du, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$, and where we are interested in either C^∞ or locally integrable solutions (u,x,y).[1]

It is a classical result that system (1) will respond to an exponential input u with an exponential response (output) y. In fact, the *exponential behavior* of the complexification of system (1), i.e. the collection of these exponential inputs and the corresponding outputs, uniquely determines the transfer function and hence the controllable part of the behavior of system (1). For a general discussion of this fact see Section 8.2 in [19]. Here we only need the response of system (1) to real exponential inputs.

Lemma 1. *Let $J_A = \{s \in \mathbb{R} \,|\, (sI - A) \text{ invertible}\}$, then the response of system* (1) *to the exponential input $u_{s,v_s}(t) = v_s \mathrm{e}^{st}$, where $s \in J_A$ and $v_s \in \mathbb{R}^m$, is*

$$y_{s,v_s}(t) = C\mathrm{e}^{At}\left(x_0 - (sI-A)^{-1}Bv_s\right) + G(s)v_s\mathrm{e}^{st},$$

where $G(s) = C(sI-A)^{-1}B + D$ is the transfer function of system (1).

Proof. For a proof see, e.g., Section 8.2 in [2] or Section 8.2 in [19]. □

We will also need the unforced exponential response of system 1. For the real case, the full level of detail is rarely worked out in the literature, hence it is included here.

Lemma 2. *Consider the linear system*

$$\begin{aligned} \dot{x} &= Ax, \qquad x(0) = x_0, \\ y &= Cx, \end{aligned} \tag{2}$$

where $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{p\times n}$ and where we are interested in either C^∞ or locally integrable solutions (x,y). Let $C \neq 0$ and denote the rows of C by $c_l^\top$, $l = 1,\ldots,p$. Let

[1] See [19] for a discussion of the technical differences between these choices. They are immaterial for the results reported here.

$c_k^\top$ be the highest numbered non-zero row of C then there exists an initial value x_0 such that $y(t)$ has the form

$$y(t) = \mathrm{e}^{\alpha t}\begin{bmatrix} \sum_{m=0}^{M} \frac{t^m}{m!}(a_{1m}\cos(\beta t) + b_{1m}\sin(\beta t)) \\ \vdots \\ \sum_{m=0}^{M} \frac{t^m}{m!}(a_{(k-1)m}\cos(\beta t) + b_{(k-1)m}\sin(\beta t)) \\ a_{k0}\cos(\beta t) + b_{k0}\sin(\beta t) \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{3}$$

where $\alpha + j\cdot\beta \in \mathbb{C}$ is an eigenvalue of A, M is a number less than the size of the largest Jordan block associated with the eigenvalue $\alpha + j\cdot\beta$, $a_{lm}, b_{lm} \in \mathbb{R}$ for $l = 1,\dots,k-1$ and $m = 0,\dots,M$, and $a_{k0}, b_{k0} \in \mathbb{R}$ with $a_{k0} \neq 0$.

Proof. As a consequence of the real Jordan normal form theorem applied to A, there exists an invertible $T \in \mathbb{R}^{n\times n}$ such that $T\mathrm{e}^{At}T^{-1}$ is block diagonal where the blocks are either of the form

$$\mathrm{e}^{\alpha t}\begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{m-1}}{(m-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{m-2}}{(m-2)!} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & t \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix},$$

corresponding to real eigenvalues α of A, or of the form

$$\mathrm{e}^{\alpha t}\begin{bmatrix} D & tD & \frac{t^2}{2!}D & \cdots & \frac{t^{m-1}}{(m-1)!}D \\ 0 & D & tD & \cdots & \frac{t^{m-2}}{(m-2)!}D \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & tD \\ 0 & \cdots & \cdots & 0 & D \end{bmatrix}, \quad D = \begin{bmatrix} \cos(\beta t) & -\sin(\beta t) \\ \sin(\beta t) & \cos(\beta t) \end{bmatrix},$$

corresponding to complex eigenvalues $\alpha + j\cdot\beta$ of A, see e.g. Section 56.1 in [11]. Consider the highest numbered non-zero row $c_k^\top$ of C and denote the entries of $c_l^\top T^{-1}$ by c_{lm} for $l = 1,\dots,k$ and $m = 1,\dots,n$. Look at the i-th column of $T\mathrm{e}^{At}T^{-1}$, where i is the lowest index m for which $c_{km} \neq 0$.

In the case where this index falls into a Jordan block corresponding to a real eigenvalue α, it follows that the i-th entry of the vector $c_k^\top \mathrm{e}^{At}T^{-1}$ is $c_{ki}\mathrm{e}^{\alpha t}$ and we set $a_{k0} := c_{ki} \neq 0$, $b_{k0} := 0$ and $\beta := 0$. Similarly, it follows that the i-th entry of the vector

$c_l^\top \mathrm{e}^{At} T^{-1}$ is $\sum_{m=0}^{M} c_{l(i-m)} \mathrm{e}^{\alpha t} \frac{t^m}{m!}$, where M is the number of entries above the diagonal in the given column of the Jordan block. Set $a_{lm} := c_{l(i-m)}$ and $b_{lm} := 0$ for $l = 1,\dots,k-1$ and $m = 0,\dots,M$.

In the case where i matches an even numbered column of a Jordan block corresponding to a complex eigenvalue $\alpha + j \cdot \beta$, it follows that the i-th entry of the vector $c_k^\top \mathrm{e}^{At} T^{-1}$ is $c_{ki} \mathrm{e}^{\alpha t} \cos(\beta t)$ and we set $a_{k0} := c_{ki} \neq 0$ and $b_{k0} := 0$. Similarly, it follows that the i-th entry of the vector $c_l^\top \mathrm{e}^{At} T^{-1}$ is

$$\sum_{m=0}^{M} \left(c_{l(i-2m)} \mathrm{e}^{\alpha t} \frac{t^m}{m!} \cos(\beta t) - c_{l(i-2m-1)} \mathrm{e}^{\alpha t} \frac{t^m}{m!} \sin(\beta t) \right),$$

where M is the number of 2×2 blocks above the diagonal in the given block column of the Jordan block. Set $a_{lm} := c_{l(i-2m)}$ and $b_{lm} := -c_{l(i-2m-1)}$ for $l = 1,\dots,k-1$ and $m = 0,\dots,M$.

In the remaining case, the i-th entry of the vector $c_k^\top \mathrm{e}^{At} T^{-1}$ is $c_{ki} \mathrm{e}^{\alpha t} \cos(\beta t) + c_{k(i+1)} \mathrm{e}^{\alpha t} \sin(\beta t)$ and we set $a_{k0} := c_{ki} \neq 0$ and $b_{k0} := c_{k(i+1)}$. Similarly, it follows that the i-th entry of the vector $c_l^\top \mathrm{e}^{At} T^{-1}$ is

$$\sum_{m=0}^{M} \left(c_{l(i-2m)} \mathrm{e}^{\alpha t} \frac{t^m}{m!} \cos(\beta t) + c_{l(i-2m+1)} \mathrm{e}^{\alpha t} \frac{t^m}{m!} \sin(\beta t) \right),$$

where M is the number of 2×2 blocks above the diagonal in the given block column of the Jordan block. Set $a_{lm} := c_{l(i-2m)}$ and $b_{lm} := c_{l(i-2m+1)}$ for $l = 1,\dots,k-1$ and $m = 0,\dots,M$.

This completes the proof, since $x(t) = \mathrm{e}^{At} T^{-1} e_i$, $y(t) = Cx(t)$ is the solution of system (2) corresponding to the initial condition $x_0 := T^{-1} e_i$. □

The following lemma is a simple consequence of the above results.

Lemma 3. *Let all uncontrollable modes of system* (1) *be unstable. Then, for every* $Q \in \mathbb{R}^{q \times n}$ *with* $Q \neq 0$ *there exists an initial condition* x_0 *and an input* u *such that* $Qx(t) \not\to 0$ *as* $t \to \infty$.

Proof. Given that all uncontrollable modes of system (7) are unstable, there exists an invertible $S \in \mathbb{R}^{n \times n}$ such that

$$SAS^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \text{ and } SB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

where (A_{11}, B_1) is controllable and all eigenvalues of A_{22} have non-negative real parts (Kalman decomposition). Define

$$\begin{bmatrix} Q_1 & Q_2 \end{bmatrix} := QS^{-1} \quad \text{and} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} := Sx,$$

where the block sizes are as for SAS^{-1} above. Applying Lemma 1 to the system

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad x(0) = x_0, \\ y &= Qx, \end{aligned} \tag{4}$$

we obtain $y(t) = Q\mathrm{e}^{At}\left(x_0 - (sI - A)^{-1}Bv_s\right) + Q_1(sI - A_{11})^{-1}B_1 v_s \mathrm{e}^{st}$ when applying the input $u(t) := v_s \mathrm{e}^{st}$ with $s \in J_A$ and $v_s \in \mathbb{R}^m$ arbitrary.

Assume now that $Q_1 \neq 0$ and pick a non-zero row $q^\top$ of Q_1. Assume, to arrive at a contradiction, that $q^\top(sI - A_{11})^{-1}B_1 = 0$ for all $s \in J_A$ with $s > 0$. Expanding the strictly proper rational matrix function $(sI - A_{11})^{-1}B_1$ in terms of its Markov parameters this implies $q^\top A_{11}^{i-1}B_1 = 0$ for all i, a contradiction to the controllability of (A_{11}, B_1). It follows that there exists $s \in J_A$, $s > 0$ such that $q^\top(sI - A_{11})^{-1}B_1 \neq 0$. We can hence choose $v_s \in \mathbb{R}^m$ such that $Q_1(sI - A_{11})^{-1}B_1 v_s \mathrm{e}^{st} \not\to 0$ as $t \to \infty$. Setting $x_0 := (sI - A)^{-1}Bv_s$ and $u(t) := v_s \mathrm{e}^{st}$ this completes the proof in the case where $Q_1 \neq 0$.

Assume now that $Q_1 = 0$ then $Q_2 \neq 0$ and system (4) is equivalent to

$$\begin{aligned} \dot{x_1} &= A_{11}x_1 + A_{12}x_2 + B_1 u, \\ \dot{x_2} &= A_{22}x_2, \\ y &= Q_2 x_2. \end{aligned}$$

Focussing on the last two equations, the result now follows from Lemma 2 and the fact that all eigenvalues of A_{22} have non-negative real parts. □

Remark 4. The previous result is most general in the sense that it is clearly false if system (1) has stable uncontrollable modes. It is easy to see (using the Kalman decomposition and a stable-unstable decomposition of the unforced dynamics) that then certain non-trivial linear functions of the state will always go to zero as time goes to infinity.

The following proposition now gives a simple necessary condition for output stability.

Proposition 5. *If* $\lim_{t\to\infty} y(t) = 0$ *for all choices of* x_0 *and* u *in system* (1) *then its transfer function* $G = 0$ *and* $D = 0$*. If, moreover, all uncontrollable modes of system* (1) *are unstable then* $C = 0$.

Proof. Assume, to arrive at a contradiction, that $G \neq 0$. With the same notation as in Lemma 1, we can choose $s \in J_A$ such that $s > 0$ and $G(s) \neq 0$ since $G(s) = 0$ only at a finite number of points $s \in \mathbb{C}$. We can then choose $v_s \in \mathbb{R}^m$ such that $G(s)v_s \neq 0$. But then the choices $x_0 := (sI - A)^{-1}Bv_s$ and $u(t) := v_s \mathrm{e}^{st}$ yield $y(t) = G(s)v_s \mathrm{e}^{st} \not\to 0$ as $t \to \infty$, a contradiction. Hence $G = 0$. Since $G(s) = C(sI - A)^{-1}B + D$ where D is constant and $C(sI - A)^{-1}B$ is strictly proper, it follows that $D = 0$. The remaining statement now follows from an application of Lemma 3 with $Q := C$. □

Next we explore what happens if we pass a vector of the form (3) through a linear system.

Lemma 6. *Let* $K \in \mathbb{R}^{p\times p}$ *and let*

$$\dot{e} = Ke + y, \qquad e(0) = 0.$$

If $y(t)$ *is of the form* (3) *with* $\alpha \geq 0$ *and* $a_{k0} \neq 0$*, and if* $\alpha \pm j\cdot\beta$ *are* not *eigenvalues of* K*, then* $e(t) \not\to 0$ *for* $t \to \infty$.

Proof. The Laplace transform of a typical entry

$$y_l(t) = \mathrm{e}^{\alpha t} \sum_{m=0}^{M} \frac{t^m}{m!} (a_{lm} \cos(\beta t) + b_{lm} \sin(\beta t))$$

of $y(t)$ is

$$Y_l(s) = \sum_{m=0}^{M} \frac{(-1)^m}{m!} \frac{\partial^m}{\partial s^m} \left(\frac{a_{lm}(s-\alpha) + b_{lm}\beta}{(s-\alpha)^2 + \beta^2} \right),$$

which is a strictly proper function with poles only at $\alpha \pm j \cdot \beta$. At least $Y_k(s)$ is non-trivial since $a_{k0} \neq 0$. We want to show that $E(s) = (sI-K)^{-1} Y(s)$ has poles at $\alpha \pm j \cdot \beta$ as well, i.e. that these poles do not get cancelled by multiplication with $(sI-K)^{-1}$. To this end let $Y(s) = D_l(s)^{-1} N_l(s) = N_r(s) D_r(s)^{-1}$ be left (resp. right) coprime factorizations. By Theorem 2 (iii) in [1] there is no pole-zero cancellation at $s_0 \in \mathbb{C}$ if and only if both

$$\begin{bmatrix} I \\ D_l(s_0) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} s_0 I - K & N_r(s_0) \end{bmatrix}$$

have full rank. This is clearly the case if $s_0 I - K$ is invertible. Since by assumption $s_0 = \alpha \pm j \cdot \beta$ are *not* eigenvalues of K, the required poles are still present in $E(s)$. But then $e(t) \not\to 0$ for $t \to \infty$ since $\alpha \geq 0$. This is most easily seen by partial fraction expansion followed by the inverse Laplace transform. □

The following proposition is the main result of this section. It provides a necessary condition for output stability of a simple type of cascading system that appears as an error system in the study of observers.

Proposition 7. *Consider the composite system*

$$\begin{aligned} \dot{x} &= Ax + Bu, & x(0) &= x_0, \\ \dot{e} &= Ke + Rx + Su, & e(0) &= e_0, \end{aligned} \tag{5}$$

and assume that $\lim_{t\to\infty} e(t) = 0$ *for all choices of* x_0*,* e_0 *and* u*. If all uncontrollable modes of* $\dot{x} = Ax + Bu$ *are unstable then* $R = 0$ *and* $S = 0$*.*

Proof. Apply Proposition 5 to system (5) augmented with an output,

$$\dot{\begin{bmatrix} x \\ e \end{bmatrix}} = \begin{bmatrix} A & 0 \\ R & K \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix} + \begin{bmatrix} B \\ S \end{bmatrix} u, \qquad \begin{bmatrix} x \\ e \end{bmatrix}(0) = \begin{bmatrix} x_0 \\ e_0 \end{bmatrix},$$

$$e = \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix},$$

to obtain $(sI-K)^{-1} \left[-R(sI-A)^{-1}B + S \right] \equiv 0$ and hence $-R(sI-A)^{-1}B + S \equiv 0$. Since S is constant and $R(sI-A)^{-1}B$ is strictly proper, it follows that $S = 0$ and $R(sI-A)^{-1}B \equiv 0$. If $\dot{x} = Ax + Bu$ was controllable, we would be done at this point, since then $R(sI-A)^{-1}B \equiv 0$ would imply $R = 0$. With the help of Lemmas 2 and 6 we can, however, treat the more general case of this proposition.

Given that all uncontrollable modes of $\dot{x} = Ax + Bu$ are unstable, there exists an invertible $S \in \mathbb{R}^{n \times n}$ such that

$$SAS^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \text{ and } SB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

where (A_{11}, B_1) is controllable and all eigenvalues of A_{22} have non-negative real parts (Kalman decomposition). Define

$$\begin{bmatrix} R_1 & R_2 \end{bmatrix} := RS^{-1} \quad \text{and} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} := Sx,$$

where the block sizes are as for SAS^{-1} above. Then $R(sI - A)^{-1}B \equiv R_1(sI - A_{11})^{-1}B_1 \equiv 0$ and hence $R_1 = 0$ because (A_{11}, B_1) is controllable. It follows that

$$\begin{aligned} \dot{x}_2 &= A_{22}x_2, & x_2(0) &= x_{02}, \\ \dot{e} &= Ke + R_2x_2, & e(0) &= e_0. \end{aligned} \tag{6}$$

Choose $x_0 = 0$ and $u = 0$ in system (5) to see that K must be stable. This follows from Lemma 2 with a simple proof by contradiction. Since all eigenvalues of A_{22} have non-negative real parts, this means that no eigenvalue of A_{22} is also an eigenvalue of K.

Assume, to arrive at a contradiction, that $R_2 \neq 0$. By choosing $e_0 := 0$ in (6) and applying Lemmas 2 and 6, we conclude the existence of x_{02} such that $e(t) \not\to 0$ for $t \to \infty$, a contradiction. It follows that $R_2 = 0$ and hence $R = 0$. □

Remark 8. In the literature on observers, results equivalent to the above proposition are often considered to be obvious with justifications along the lines of "if $S \neq 0$, we can find a u to make e not go to zero and if $R \neq 0$ we can find a u to generate an x which makes e not go to zero", see e.g. the proof of Theorem 7-9 in [5] (for the controllable case). Since u and x are not independent signals (in fact, they are jointly constrained by the equation $\dot{x} = Ax + Bu$), this argument is hand-waving at best. However, the underlying intuition is confirmed by the more thorough analysis in this section. In essence, the result depends on the fact that the transfer from u to x is strictly proper and hence u and x would be distinguishable at the output e. Furthermore, the spectral separation between K and the uncontrollable dynamics of x ensures that the latter would be observable at the output e as well.

3 Characterization of functional state observers

Consider the linear time-invariant finite-dimensional system in state space form given by

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx, \\ z &= V_0x, \end{aligned} \tag{7}$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$ and $V_0 \in \mathbb{R}^{r \times n}$.

We will be interested in the characterization of asymptotic observers for z given y and u. In particular, we will be interested in observers of the following type usually considered in the geometric control literature:

$$\begin{aligned} \dot{v} &= Kv + Ly + Mu, \\ \hat{z} &= Pv + Qy, \end{aligned} \tag{8}$$

where $K \in \mathbb{R}^{s \times s}$, $P \in \mathbb{R}^{r \times s}$ and the other matrices are real and appropriately sized. Note that P can be rectangular (tall or wide) and/or not of full rank. The asymptotic condition for this type of observer is

$$\lim_{t \to \infty} \hat{z}(t) - z(t) = 0 \tag{9}$$

for every choice of input u and initial conditions $x(0)$ and $v(0)$. The observer (8) is called *observable* if it is an observable system in the usual sense, i.e. if the pair (P, K) is observable.

The following theorem provides a necessary and sufficient condition for an observable system of type (8) to be an asymptotic observer for z given u and y. See the discussion in Section 3.1 to what extent this is a known result.

Theorem 9. *Let all uncontrollable modes of system* (7) *be unstable. Then system* (8) *is an observable asymptotic observer for z given u and y if and only if there exists a matrix $U \in \mathbb{R}^{s \times n}$ such that*

$$\begin{aligned} UA - KU - LC &= 0, \\ M - UB &= 0, \\ V_0 - PU - QC &= 0, \end{aligned} \tag{10}$$

K is Hurwitz and (P, K) is observable.

Proof. Let system (8) be an observable asymptotic observer for z given u and y and define $e := \hat{z} - z$. Then

$$e = Pv + Qy - V_0 x = Pv - (V_0 - QC)x.$$

Assume, to arrive at a contradiction, that $\operatorname{Im}(V_0 - QC) \not\subset \operatorname{Im}(P)$. Then there exists an invertible $S \in \mathbb{R}^{r \times r}$ such that

$$SP = \begin{bmatrix} P_1 \\ 0 \end{bmatrix} \text{ and } S(V_0 - QC) = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

with $V_2 \neq 0$. Now $\lim_{t \to \infty} Se(t) = 0$ implies $\lim_{t \to \infty} V_2 x(t) = 0$ for all initial conditions $x(0)$ and all inputs u, a contradiction to Lemma 3. We conclude that $\operatorname{Im}(V_0 - QC) \subset \operatorname{Im}(P)$ and hence there exists a matrix $U \in \mathbb{R}^{s \times n}$ such that $V_0 - QC = PU$. This implies the third equation in (10).

Define $d := v - Ux$ then

$$\begin{aligned} \dot{d} &= \dot{v} - U\dot{x} \\ &= Kv + Ly + Mu - UAx - UBu \\ &= Kv - KUx + KUx + LCx + Mu - UAx - UBu, \end{aligned}$$

and hence the observation error $e = Pv - (V_0 - QC)x = Pv - PUx$ is governed by the error system

$$\begin{aligned} \dot{d} &= Kd - (UA - KU - LC)x + (M - UB)u, \\ e &= Pd. \end{aligned} \tag{11}$$

Using the notation $R := -(UA - KU - LC)$ and $S := M - UB$, it remains to show that $R = 0$ and $S = 0$, but this follows immediately from Proposition 7 and the fact that (P, K) is observable (hence $e(t) \to 0$ implies $d(t) \to 0$).

Conversely, assume that systems (7) and (8) fulfill Equation (10) with K Hurwitz, then $\lim_{t\to\infty} e(t) = 0$ follows immediately from the form of the error system (11). In its derivation we have only used the third line of Equation (10). □

3.1 Related results from the literature

The early literature on state observers is exclusively concerned with the case where the observed system is controllable. This is owing to the fact that at the time observers were only studied as a building block in (observer-based) closed loop control.

A simplified version of the Sylvester equation, the first equation in (10), appears already in Luenberger's first observer paper published in 1964 [13, Equation (4)]. It appears there in the context of full state observation for uncontrolled systems, while the main focus of the paper is on observer-based controller design for single-input single-output systems. It has later been claimed that Luenberger's paper already established necessity of the Sylvester equation for observer design (it did discuss sufficiency as well as single functional observers and observer order reduction), but a close reading of the paper reveals that it only shows necessity for an observer that has the *tracking property* and not the more important *asymptotic property* that we discuss here. An observer has the tracking property if it keeps producing zero estimation error whenever the initial estimation error happened to be zero (due to a lucky choice of the initial value). Proving that the asymptotic property implies the tracking property is equally hard as proving necessity of the Sylvester equation from first principles (and is generally false). The same restrictions apply to Luenberger's subsequent papers [14] and [15], although the 1971 paper [15] contains the fully general equation. No claim of necessity is made by Luenberger for the asymptotic case.

It appears that the subtle difference between tracking observers and asymptotic observers started to be overlooked as early as 1973, see e.g. [17, page 309] where Luenberger's necessity result is referenced as if it applied to asymptotic observers.

A more refined result was presented by Bongiorno and Youla in 1968 [3], see also [4]. The paper demonstrated sufficiency of the Sylvester equation in the case of full state observation [3, page 223] and provided a necessary and sufficient condition (not based on the Sylvester equation) for the single output case [3, Corollary 2], again considering only full state observation.

The first paper that established a general necessity result appears to be a 1972 paper by Fortmann and Williamson [9] in which necessity of the Sylvester equation is shown for the controllable case. However, the proof is based on a different definition of asymptotic observers, namely one that includes asymptotically matching derivatives of all orders. This condition enters the proof in a decisive way, see Equation (4) in the proof of Lemma 2 in the paper appendix [9]. Again, proving that a zero-th order asymptotic condition implies all the higher order asymptotic conditions (for C^∞ trajectories), is equally hard as directly proving necessity of the Sylvester equation.

Moore and Ledwich attempted a general proof of necessity and sufficiency in the controllable case in 1975 [16]. Unfortunately, their proof makes use of a misshaped "reachability matrix", see [16, Equation (2.7)]. With a little bit more effort the proof idea can in fact be rescued, replacing the factoring out of that matrix with a transfer function argument similar to our Proposition 5, although the second equation in the stated necessary and sufficient condition is still wrong and should correctly read $EH' - K' = 0$ (in the notation of [16]). Interestingly, just a few pages earlier, in the same issue of the Transactions, Roman and Bullock claimed that Luenberger proved necessity and sufficiency of the Sylvester equation citing his three above papers [20, page 615].

It appears that by the late 1970s it had become somewhat of a "folk theorem" that the Sylvester equation is necessary and sufficient for asymptotic observers, at least in the controllable case, although there was no correct proof in the literature at the time. For example, Sirisena used the same asymptotic condition as Fortmann and Williamson but cited Moore and Ledwich [22]. Kawaji simply asserted necessity without reference [12]. The claim subsequently found its way into textbooks, most prominently into the 1983 edition of the book by O'Reilly [18, Theorem 3.2] and the 1984 edition of the book by Chen [5, Theorem 7-9]. While O'Reilly correctly proved sufficiency, for the necessity part he partly (incorrectly) appealed to Luenberger's work, and partly re-hashed the "proof" by Moore and Ledwich. Chen provided a hand-waving argument as discussed in Remark 4.

In much of the subsequent literature on this topic the necessity result was assumed or "proved" with direct or indirect reference to the above literature, e.g. [6, 8]. Darouach [6] neglects to mention conditions on the observed system (making his theorem wrong as stated) and refers to Chen's book, while Fernando et al. [8] give the correct, most general condition of no stable uncontrollable modes and give an argument similar to Chen's. Note that both these references treat the special case $P = I$, cf. Section 4.

To the best of my knowledge, the first complete proof of necessity of the Sylvester equation in the controllable case was given by Paul Fuhrmann and Uwe Helmke in 2001 [10, Theorem 5.4], making observer characterization a very fitting topic for Uwe's birthday volume. In my PhD thesis, supervised by Uwe and Paul, I slightly generalized their proof idea to also allow for direct feedthrough of the output measurement in the observer (the same setting that is discussed in this paper) [23]. The result for observed systems with no stable uncontrollable modes (but without direct feedthrough in the observer) was derived as a corollary to a much more general behavioral result in [24]. The proof given there makes use of a behavioral internal model principle for observers and is not easily translated into state space and transfer function thinking.

It should be pointed out that the above list is by no means exhaustive, many more papers on this topic have appeared over the years, but I tried to concentrate on those papers that are most relevant to this discussion. I do, however, wish to mention Schumacher's 1980 paper [21] that contains a comprehensive discussion of asymptotic observers from the perspective of geometric control theory but does not tackle the characterization problem.

4 Existence of functional state observers

Consider system (7). In this section we will be interested in conditions for the existence of asymptotic observers for z given y and u. In particular, we will be interested in observers of type (8) as well as of the following type:

$$\begin{aligned} \dot{v} &= Kv + Ly + Mu, \\ w &= v + Qy, \end{aligned} \tag{12}$$

where $K \in \mathbb{R}^{s \times s}$, with $s \geq r$, and the other matrices are real and appropriately sized. The asymptotic condition for this type of observer requires the existence of $V_1 \in \mathbb{R}^{(s-r)\times n}$ such that

$$\lim_{t\to\infty} w(t) - \begin{bmatrix} z(t) \\ V_1 x(t) \end{bmatrix} = 0$$

for every choice of input u and initial conditions $x(0)$ and $v(0)$. In other words, the first r components of w provide an asymptotic estimate for z. Observers of type (12) are automatically observable.

The reader is referred to the discussion in [7] for a detailed explanation of why this observer structure is of particular relevance, and indeed can be assumed without loss of generality when discussing practical observer design, but not for the observer characterization problem as we shall see in the example given below.

Clearly, the existence of an asymptotic observer of type (12) implies the existence of an asymptotic observer of type (8) of the same dynamic order s. The following theorem gives conditions under which the opposite conclusion holds.

Theorem 10. *Let all uncontrollable modes of system* (7) *be unstable and assume that* $\left[\begin{smallmatrix} V_0 \\ C \end{smallmatrix}\right]$ *has full row rank. If there exists an observable asymptotic observer of type* (8) *then there exists an asymptotic observer of type* (12) *of the same dynamic order s.*

Proof. Assume that we are given an observable asymptotic observer of type (8). By Theorem 9 there exists a matrix $U \in \mathbb{R}^{s\times n}$ such that

$$\begin{aligned} UA - KU &= LC, \\ M &= UB, \\ V_0 &= PU + QC, \end{aligned} \tag{13}$$

where K is Hurwitz. Assume now, to arrive at a contradiction, that P does not have full row rank. Then there exists a vector $x \in \mathbb{R}^r$, $x \neq 0$ such that $x^\top P = 0$. Using Eq. (13) it follows that $x^\top V_0 = x^\top QC$ and hence that

$$\begin{bmatrix} x^\top & -x^\top Q \end{bmatrix} \begin{bmatrix} V_0 \\ C \end{bmatrix} = 0,$$

a contradiction to the assumption that $\left[\begin{smallmatrix} V_0 \\ C \end{smallmatrix}\right]$ has full row rank. We conclude that $P \in \mathbb{R}^{r\times s}$ has full row rank, so in particular $s \geq r$.

Since P has full row rank, there exists an invertible matrix $S \in \mathbb{R}^{s\times s}$ such that

$$P = \begin{bmatrix} I & 0 \end{bmatrix} S.$$

Decompose

$$SU = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$$

with $U_1 \in \mathbb{R}^{r\times n}$ then Eq. (13) implies

$$\begin{aligned} (SU)A - (SKS^{-1})(SU) &= (SL)C, \\ (SM) &= (SU)B, \\ U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} &= \begin{bmatrix} V_0 \\ V_1 \end{bmatrix} - \begin{bmatrix} Q \\ 0 \end{bmatrix} C, \end{aligned}$$

where $V_1 := U_2 \in \mathbb{R}^{(s-r)\times n}$. Note that SKS^{-1} is similar to K and hence also Hurwitz. Applying Theorem 9 now yields an asymptotic observer of type (12) of dynamic order s given by

$$\begin{aligned} \dot{v} &= (SKS^{-1})v + (SL)y + (SM)u, \\ w &= v + \begin{bmatrix} Q \\ 0 \end{bmatrix} y. \end{aligned}$$

This completes the proof. □

Corollary 11. *Let all uncontrollable modes of system* (7) *be unstable and assume that* $\left[\begin{smallmatrix} V_0 \\ C \end{smallmatrix}\right]$ *has full row rank. The minimal dynamic orders for asymptotic observers of both types coincide.*

Proof. A minimal order observer of type (8) is necessarily observable. This follows from a simple application of the Kalman decomposition for non-observable systems, see e.g. [23, Proposition 3.69]. The statement now follows directly from Theorem 10 and the remark in the paragraph preceding it. □

The following example shows that the rank condition in Theorem 10 can not be dispensed with.

Example 12. Consider system (7) with

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \ B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \ C = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \text{ and } V_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This system is controllable but does not fulfill the rank condition of Theorem 10. By Theorem 9 we need to find a solution to the following two matrix equations in order to construct an asymptotic observer of type (12) of dynamic order $s = 2$.

$$U = V_0 - QC = \begin{bmatrix} 1 & 0 & -q_1 \\ 0 & 0 & 1-q_2 \end{bmatrix},$$

$$0 = UA - KU - LC = \begin{bmatrix} 0 & -q_1 & 0 \\ 0 & 1-q_2 & 0 \end{bmatrix} - \begin{bmatrix} k_{11} & 0 & * \\ k_{21} & 0 & * \end{bmatrix} - \begin{bmatrix} 0 & 0 & -l_1 \\ 0 & 0 & -l_2 \end{bmatrix}.$$

Clearly, it follows that $k_{11} = k_{21} = 0$ and hence that K is necessarily singular, so in particular not Hurwitz. This means that there is no asymptotic observer of type (12) of dynamic order $s = 2$ for this system. Note that $r = 2$ in this example, so $s = 2$ would be the minimal possible dynamic order for such an observer.

However, an observable asymptotic observer of type (8) of dynamic order $s = 2$ is given by

$$\begin{aligned} \dot{v} &= \begin{bmatrix} 0 & -1 \\ 1 & -2 \end{bmatrix} v + \begin{bmatrix} -2 \\ -3 \end{bmatrix} y + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u, \\ \hat{z} &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} v + \begin{bmatrix} -1 \\ -1 \end{bmatrix} y. \end{aligned} \tag{14}$$

The easiest way of checking this is via Theorem 9.

From a practical perspective, the rank condition in Theorem 10 can of course be assumed w.l.o.g. in the following sense: A different strategy for observer design in this example would be to first remove the second row of V_0, then construct an observer of type (12) for the reduced V_0,[2] and finally add an algebraic equation estimating the "missing" functional directly from y. The *overall* observer would then of course not be of type (12) but of type (8)! In fact, modulo different choices for the eigenvalues of K, it would be just a different realization of system (14).

5 Conclusion

We have given a detailed proof for the fact that Luenberger's Sylvester equations (10) characterize asymptotic (functional) state observers in the category of linear time-invariant finite-dimensional systems in input / state / output form, where the observed system is assumed to have no stable uncontrollable modes. From Remark 4 it is clear that the case of stable uncontrollable modes can be treated using a Kalman decomposition and imposes no further equality constraints on the observer. It is merely required that the observer modes estimating these stable uncontrollable observed system modes are themselves stable, i.e. that the state transition matrix associated with these observer modes is Hurwitz.

The discussion in the last section shows that transformations that entail no loss of generality for practical observer design may well do so for the observer characterization problem.

[2] This is possible with dynamic order $s = 2$ by adding $V_1 := \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$.

Bibliography

[1] B. D. O. Anderson and M. R. Gevers. On multivariable pole-zero cancellations and the stability of feedback systems. *IEEE Transactions on Circuits and Systems*, 28(8):830–833, 1981. Cited p. 426.

[2] K. J. Åström and R. M. Murray. *Feedback systems*. Princeton, 2008. Cited p. 422.

[3] J. J. Bongiorno, Jr. and D. C. Youla. On observers in multi-variable control systems. *International Journal of Control*, 8(3):221–243, 1968. Cited p. 429.

[4] J. J. Bongiorno, Jr. and D. C. Youla. Discussion of "On observers in multivariable control systems". *International Journal of Control*, 12(1):183–190, 1970. Cited p. 429.

[5] C.-T. Chen. *Linear system theory and design*. Harcourt Brace College Publishers, 1984. Cited pp. 427 and 430.

[6] M. Darouach. Existence and design of functional observers for linear systems. *IEEE Transactions on Automatic Control*, 45(5):940–943, 2000. Cited p. 430.

[7] T. L. Fernando, L. S. Jennings, and H. M. Trinh. Generality of functional observer structures. In *Proceedings of the 50th IEEE Conference on Decision and Control*, pages 4000–4004, 2011. Cited p. 431.

[8] T. L. Fernando, H. M. Trinh, and L. Jennings. Functional observability and the design of minimum order linear functional observers. *IEEE Transactions on Automatic Control*, 55(5):1268–1273, 2010. Cited p. 430.

[9] T. E. Fortmann and D. Williamson. Design of low-order observers for linear feedback control laws. *IEEE Transactions on Automatic Control*, 17(3):301–308, 1972. Cited pp. 429 and 430.

[10] P. A. Fuhrmann and U. Helmke. On the parametrization of conditioned invariant subspaces and observer theory. *Linear Algebra and its Applications*, 332–334:265–353, 2001. Cited p. 430.

[11] L. Hogben, editor. *Handbook of Linear Algebra*. Chapman&Hall/CRC, 2007. Cited p. 423.

[12] S. Kawaji. Design procedure of observer for the linear functions of the state. *International Journal of Control*, 32(3):381–394, 1980. Cited p. 430.

[13] D. G. Luenberger. Observing the state of a linear system. *IEEE Transactions on Military Electronics*, 8(2):74–80, 1964. Cited p. 429.

[14] D. G. Luenberger. Observers for multivariable systems. *IEEE Transactions on Automatic Control*, 11(2):190–197, 1966. Cited p. 429.

[15] D. G. Luenberger. An introduction to observers. *IEEE Transactions on Automatic Control*, 16(6):596–602, 1971. Cited pp. 421 and 429.

[16] J. B. Moore and G. F. Ledwich. Minimal order observers for estimating linear functions of a state vector. *IEEE Transactions on Automatic Control*, 20(5):623–632, 1975. Cited p. 430.

[17] P. Murdoch. Observer design for a linear functional of the state vector. *IEEE Transactions on Automatic Control*, 18(3):308–310, 1973. Cited p. 429.

[18] J. O'Reilly. *Observers for linear systems*. Academic Press, 1983. Cited p. 430.

[19] J. W. Polderman and J. C. Willems. *Introduction to mathematical systems theory: A behavioral approach*. Springer, 1998. Cited p. 422.

[20] J. R. Roman and T. E. Bullock. Design of minimal order stable observers for linear functions of the state via realization theory. *IEEE Transactions on Automatic Control*, 20(5):613–622, 1975. Cited p. 430.

[21] J. M. Schumacher. On the minimal stable observer problem. *International Journal of Control*, 32(1):17–30, 1980. Cited p. 431.

[22] H. R. Sirisena. Minimal-order observers for linear functions of a state vector. *International Journal of Control*, 29(2):235–254, 1979. Cited p. 430.

[23] J. Trumpf. *On the geometry and parametrization of almost invariant subspaces and observer theory*. PhD thesis, Universität Würzburg, Germany, 2002. Cited pp. 430 and 432.

[24] J. Trumpf, H. L. Trentelman, and J. C. Willems. An internal model principle for observers. In *Proceedings of the 50th IEEE Conference on Decision and Control*, pages 3992–3999, 2011. Cited p. 430.

From integration by parts to state and boundary variables of linear differential and partial differential systems

Arjan J. van der Schaft
Johann Bernoulli Institute for
Mathematics & Computer Science
University of Groningen
The Netherlands
a.j.van.der.schaft@rug.nl

Paolo Rapisarda
School of Electronics and Computer
Science
University of Southampton
United Kingdom
pr3@ecs.soton.ac.uk

Abstract. We elaborate on an idea originally expressed in [13]: the remainders resulting from repeated integration by parts of a set of linear higher-order ordinary differential equations define *state vectors*. Furthermore, these remainders and the corresponding state maps can be easily computed by factorization of a certain two-variable polynomial matrix, which is directly derived from the one-variable polynomial matrix describing the set of higher-order differential equations. Recently [7] we have extended this same idea to the construction of state maps for systems of linear partial differential equations involving, apart from the time variable, also spatial variables. In the current paper we take a next step by considering partial differential equations on a bounded spatial domain, and we show how integration by parts yields, next to the construction of state maps, also a recipe to define boundary variables in a natural manner.

It is a great pleasure for the first author to congratulate Uwe Helmke on his sixtieth birthday. Starting from my first close encounters with Uwe, probably at the famous Edzell meetings in Scotland, connecting the Systems & Control groups of Warwick, Bremen and Groningen in the early 1980s, it was a continuing joy to meet him and to discuss with him on topics of common interest.

1 Recall of state maps for finite-dimensional linear systems

In [13] we have shown how the notion of *'state'* for linear systems described by higher-order differential equations is intimately related to the procedure of *integration by parts*, and how the articulation of this relation yields an insightful and direct way of computing state maps.

In particular, consider a linear system

$$P\left(\frac{d}{dt}\right)y(t) = Q\left(\frac{d}{dt}\right)u(t), \quad y(t) \in \mathcal{Y} := \mathbb{R}^p,\, u(t) \in \mathcal{U} := \mathbb{R}^m, \tag{1}$$

or more generally, without distinguishing between inputs u and outputs y and letting $w := \begin{bmatrix} y \\ u \end{bmatrix}, q := p+m$, consider $R\left(\frac{d}{dt}\right)w(t) = 0, w(t) \in \mathcal{W} := \mathbb{R}^q$.

In all these equations, $P\left(\frac{d}{dt}\right), Q\left(\frac{d}{dt}\right)$, and $R\left(\frac{d}{dt}\right)$ describe linear (higher-order) differential operators, or, equivalently, $P(\xi), Q(\xi)$, and $R(\xi)$ are polynomial matrices of appropriate dimensions in the indeterminate ξ.

It is well-known [4] that for an observable *input-state-output system*

$$\begin{aligned} \tfrac{d}{dt}x &= Ax+Bu, \quad x(t)\in\mathbb{R}^n, u(t)\in\mathbb{R}^m \\ y &= Cx+Du, \quad y(t)\in\mathbb{R}^p \end{aligned} \tag{2}$$

the state x can be written as a linear combination of the outputs and inputs and their derivatives, i.e., $x = X_y\left(\frac{d}{dt}\right)y + X_u\left(\frac{d}{dt}\right)u$ for certain linear differential operators $X_y\left(\frac{d}{dt}\right), X_u\left(\frac{d}{dt}\right)$, or more compactly

$$x = X\left(\frac{d}{dt}\right)w, \tag{3}$$

for some $n\times q$ polynomial matrix $X(\xi)$. We will call (3) a *state map*.

Conversely, consider the system of linear higher-order differential equations

$$R\left(\frac{d}{dt}\right)w(t) = 0, \quad w(t)\in\mathcal{W} = \mathbb{R}^q, \tag{4}$$

where $R(\xi) = R_0 + R_1\xi^1 + \ldots + R_N\xi^N \in \mathbb{R}^{p\times q}[\xi]$. How do we construct state maps $x = X\left(\frac{d}{dt}\right)w$, which also allow to represent the system (4) into state space form ?

Before answering this question we need to formalize the space of solutions of (4), as well as the notion of state. An ordinary N-times differentiable solution of (4) will be called a *strong solution*. Denote the space of locally integrable trajectories from $\mathbb{R}$ to $\mathbb{R}^q$ by $\mathcal{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$. Recall that $w\in\mathcal{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q)$ is a *weak solution* of (4) if

$$\int_{-\infty}^{\infty} w^T(t)R^T\left(-\frac{d}{dt}\right)\varphi(t)dt = 0 \tag{5}$$

for all $\mathcal{C}^\infty$ test functions $\varphi:\mathbb{R}\to\mathbb{R}^p$ with compact support. The set of all weak solutions of (4), called the *behavior* $\mathcal{B}$, is denoted as

$$\mathcal{B} := \{w:\mathbb{R}\to\mathbb{R}^q \mid w\in\mathcal{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q) \text{ and } (4) \text{ is satisfied weakly}\} \tag{6}$$

Consider now two solutions $w_1, w_2\in\mathcal{B}$, and define the *concatenation* of w_1 and w_2 at time 0 as the time-function

$$(w_1\wedge_0 w_2)(t) := \begin{cases} w_1(t) & , \quad t<0 \\ w_2(t) & , \quad t\geq 0 \end{cases}, \quad t\in\mathbb{R}. \tag{7}$$

We say that $w_1, w_2\in\mathcal{B}$ are *equivalent at time* 0, denoted as $w_1\sim_0 w_2$, if for all $w\in\mathcal{B}$:

$$w_1\wedge_0 w\in\mathcal{B} \Leftrightarrow w_2\wedge_0 w\in\mathcal{B}. \tag{8}$$

Thus equivalent trajectories admit the same continuations starting from time $t=0$.

Let $X(\xi)\in\mathbb{R}^{n\times q}[\xi]$. Then the differential operator

$$\begin{aligned} X\left(\frac{d}{dt}\right): \mathcal{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^q) &\to \mathcal{L}_1^{\mathrm{loc}}(\mathbb{R},\mathbb{R}^n) \\ w &\mapsto x := X\left(\frac{d}{dt}\right)w \end{aligned}$$

is said to be a *state map* [8] for the system (4), with set of solutions $\mathcal{B}$ defined in (6), if for all $w_1, w_2 \in \mathcal{B}$ and corresponding $x_i := X\left(\frac{d}{dt}\right) w_i$, $i = 1,2$, the following property (the *state property*) holds:

$$x_1(0) = x_2(0) \text{ and } x_1, x_2 \text{ continuous at } t = 0 \implies w_1 \sim_0 w_2 \,. \tag{9}$$

If (9) holds, then the vector x contains all the information necessary to conclude whether any two trajectories in $\mathcal{B}$ admit the same continuation at time $t = 0$. For this reason the vector $x(0) = X\left(\frac{d}{dt}\right) w(0)$ is called a *state* of the system at time 0 corresponding to the time-function w, and $\mathcal{X} = \mathbb{R}^n$ is called a *state space* for the system.

Remark 1. In the context of *linear* systems (as in this paper) equation (7) is equivalent to requiring that $w_1 \wedge_0 w$ and $w_2 \wedge_0 w \in \mathcal{B}$ for *some* $w \in \mathcal{B}$. Furthermore in this case, since $w_2 \wedge_0 w_2 \in \mathcal{B}$, it follows that $w_1 \sim_0 w_2$ if and only $w_1 \wedge_0 w_2 \in \mathcal{B}$. Because of the symmetry of this last condition, it also means that equivalence of $w_1, w_2 \in \mathcal{B}$ at $t = 0$ amounts to w_1 and w_2 having the same *precursors*. (Note that for *nonlinear* systems these equivalences in general do not hold; see [11] for some initial ideas about the construction of state maps in this case.)

The basic idea of [13] is to show how state maps can be obtained from the *integration by parts* formula. Take any N-times differentiable functions $w : \mathbb{R} \to \mathbb{R}^q$ and $\varphi : \mathbb{R} \to \mathbb{R}^p$, and denote $w^{(i)} := \frac{d^i}{dt^i} w$, $i \in \mathbb{N}$, and analogously for φ. For each pair of time instants $t_1 \le t_2$ repeated integration by parts yields

$$\int_{t_1}^{t_2} w^T(t) R^T\left(-\frac{d}{dt}\right) \varphi(t) dt = \int_{t_1}^{t_2} \varphi^T(t) R\left(\frac{d}{dt}\right) w(t) dt + B_\Pi(\varphi, w)\Big|_{t_1}^{t_2} \,, \tag{10}$$

where we call the expression $B_\Pi(\varphi, w)(t)$ the *remainder*, which has the form

$$B_\Pi(\varphi, w)(t) = \begin{bmatrix} \varphi^T(t) & \varphi^{(1)T}(t) & \dots & \varphi^{(N-1)T}(t) \end{bmatrix} \tilde{\Pi} \begin{bmatrix} w(t) \\ w^{(1)}(t) \\ \vdots \\ w^{(N-1)}(t) \end{bmatrix} , \tag{11}$$

for some constant matrix $\tilde{\Pi}$ of dimension $Np \times Nq$.

The *differential version* of the integration by parts formula (10) (obtained by dividing (10) by $t_2 - t_1$ and letting t_1 tend to $t_2 = t$) is

$$w^T(t) R^T\left(-\frac{d}{dt}\right) \varphi(t) - \varphi^T(t) R\left(\frac{d}{dt}\right) w(t) = \frac{d}{dt} B_\Pi(\varphi, w)(t) \,, \tag{12}$$

Both sides of this equality define a bilinear differential operator form, or briefly a *bilinear differential form* (BDF), i.e., a bilinear functional of two trajectories and of a finite number of their derivatives. Formally, a bilinear differential form B_Φ as defined in [15] is a bilinear map $B_\Phi : \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^p) \times \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q) \to \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ involving two vector-valued functions and a finite set of their time-derivatives, that is, at any time t

$$B_\Phi(\varphi, w)(t) = \sum_{k,l=0}^{M-1} \left[\frac{d^k}{dt^k} \varphi(t) \right]^T \Phi_{k,l} \frac{d^l}{dt^l} w(t) \tag{13}$$

for certain constant $p \times q$ matrices $\Phi_{k,l}, k,l = 0,\cdots,M-1$. The matrix $\tilde{\Phi}$ whose (k,l)-th block is the matrix $\Phi_{k,l}$ for $k,l = 0,\ldots,M-1$, is called the *coefficient matrix* of the bilinear differential form B_Φ. It follows that the coefficient matrix of the bilinear differential form B_Π corresponding to the remainder is precisely the matrix $\tilde{\Pi}$ as defined in (11).

Remark 2. For a scalar polynomial or a square polynomial matrix $R(\xi)$ the formula's (10) and (12) are classically referred to as *Green's*, respectively *Lagrange's, identity*, while the matrix $\tilde{\Pi}$ for a scalar $R(\xi)$ is called the *bilinear concomitant*, see [3].

There is a useful one-to-one correspondence between the bilinear differential form B_Φ in (13) and the two-variable polynomial matrix $\Phi(\zeta,\eta)$ defined as

$$\Phi(\zeta,\eta) := \sum_{k,l=0}^{M-1} \Phi_{k,l}\zeta^k \eta^l \,. \tag{14}$$

The crucial observation, see [1, 15], is that for any bilinear differential form B_Φ the bilinear differential form corresponding to its *time-derivative*, defined as

$$\begin{aligned} B_\Psi(\varphi,w)(t) &:= \tfrac{d}{dt}\left(B_\Phi(\varphi,w)\right)(t) \\ &= \textstyle\sum_{k,l=0}^{M-1} \left[\tfrac{d^{k+1}}{dt^{k+1}}\varphi(t)\right]^T \Phi_{k,l}\tfrac{d^l}{dt^l}w(t) + \left[\tfrac{d^k}{dt^k}\varphi(t)\right]^T \Phi_{k,l}\tfrac{d^{l+1}}{dt^{l+1}}w(t), \end{aligned} \tag{15}$$

corresponds, by the product rule of differentiation, to the two-variable polynomial matrix

$$\Psi(\zeta,\eta) = (\zeta+\eta)\Phi(\zeta,\eta)\,. \tag{16}$$

As a consequence, the differential version of the integration by parts formula (12) has associated to it the *two-variable polynomial matrix equality*

$$R(-\zeta) - R(\eta) = (\zeta+\eta)\Pi(\zeta,\eta) \tag{17}$$

From this formula it follows how the two-variable polynomial matrix $\Pi(\zeta,\eta)$ and its coefficient matrix $\tilde{\Pi}$ (corresponding to the remainder) can be easily computed: since the two-variable polynomial matrix $R(-\zeta)-R(\eta)$ is zero for $\zeta+\eta=0$, it directly follows that $R(-\zeta)-R(\eta)$ contains a factor $\zeta+\eta$, and thus we can *define* the two-variable polynomial matrix $\Pi(\zeta,\eta)$ as

$$\Pi(\zeta,\eta) := \frac{R(-\zeta)-R(\eta)}{\zeta+\eta}\,. \tag{18}$$

It now turns out that state maps for a system $R(\frac{d}{dt})w = 0$ can be computed from a *factorization* of the two-variable polynomial matrix $\Pi(\zeta,\eta)$ into a product of single-variable polynomial matrices. Indeed, any factorization $\Pi(\zeta,\eta) = Y^T(\zeta)X(\eta)$ of the two-variable polynomial matrix $\Pi(\zeta,\eta)$ leads from (17) to the matrix polynomial equality

$$R(-\zeta) - R(\eta) = (\zeta+\eta)Y^T(\zeta)X(\eta)\,, \tag{19}$$

and to the corresponding bilinear differential form equality, expanding (12)

$$\begin{aligned} w^T(t)R^T\left(-\tfrac{d}{dt}\right)\varphi(t) \quad - \quad & \varphi^T(t)R\left(\tfrac{d}{dt}\right)w(t) = \\ = \quad & \tfrac{d}{dt}\left[\left(Y\left(\tfrac{d}{dt}\right)\varphi(t)\right)^T X\left(\tfrac{d}{dt}\right)w(t)\right], \end{aligned} \tag{20}$$

which immediately yields (see [6, 13] for further developments)

Theorem 3. *For any factorization* $\Pi(\zeta,\eta) = Y^T(\zeta)X(\eta)$ *the map*

$$w \mapsto x := X\left(\frac{d}{dt}\right)w$$

is a state map.

Remark 4. Furthermore [13], $Y(\xi)$ can be seen to define a state map for the *adjoint system* of (4).

2 State maps for linear systems of partial differential equations on an unbounded spatial domain

In [7] the approach of the previous section has been extended to the case of systems described by linear *partial differential* equations, involving a time variable t, and spatial variables $z_1,\cdots,z_k$. In particular for $k = 1$ (single spatial variable) we consider systems described by linear PDEs

$$R(\frac{\partial}{\partial t},\frac{\partial}{\partial z})w = 0\,, \tag{21}$$

where $R(\xi,\delta) = \sum_{i,j=0}^{L} R_{ij}\xi^i\delta^j$ with ξ and δ the indeterminates, $R_{ij} \in \mathbb{R}^{\mathrm{p}\times q}$ for $i,j = 0,\ldots,N$. An N-times differentiable (both in t and in z) solution of (21) will be called a *strong solution*. Furthermore, denote by $\mathcal{L}_1^{\mathrm{loc}}(\mathbb{R}^2,\mathbb{R}^q)$ the space of locally integrable functions from $\mathbb{R}^2$ to $\mathbb{R}^q$. Then $w \in \mathcal{L}_1^{\mathrm{loc}}(\mathbb{R}^2,\mathbb{R}^q)$ is a *weak solution* of (21) if

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w^{\mathrm{T}}(t,z)\left[R\left(-\frac{\partial}{\partial t},-\frac{\partial}{\partial z}\right)^{\mathrm{T}}\varphi(t,z)\right]dt\,dz = 0 \tag{22}$$

for all infinitely-differentiable test functions $\varphi : \mathbb{R}^2 \to \mathbb{R}^p$ with compact support. The *behavior* $\mathcal{B}$ is defined as the set of weak solutions of (21) , i.e.,

$$\mathcal{B} := \{w : \mathbb{R}^2 \to \mathbb{R}^q \mid w \in \mathcal{L}_1^{\mathrm{loc}}(\mathbb{R}^2,\mathbb{R}^q) \text{ and } (21) \text{ is satisfied weakly}\} \tag{23}$$

In order to define state maps, we consider partitions $(\mathcal{S}_-,\mathcal{S}_c,\mathcal{S}_+)$ of $\mathbb{R}^2$ induced by vertical lines $t = c$, with $c \in \mathbb{R}$, as depicted in Figure 1 on the next page;

$$\begin{aligned} \mathcal{S}_- &:= \{(t,z) \in \mathbb{R}^2 \mid t < c\}, \\ \mathcal{S}_c &:= \{(t,z) \in \mathbb{R}^2 \mid t = c\}, \\ \mathcal{S}_+ &:= \{(t,z) \in \mathbb{R}^2 \mid t > c\}. \end{aligned}$$

Since the behavior described by (21) is invariant with regard to shifts in t a special

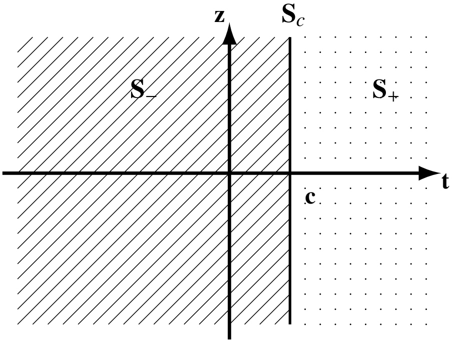


Figure 1: A partition of the plane induced by a vertical line.

role will be played by the partition $(\mathcal{S}_-,\mathcal{S}_0,\mathcal{S}_+)$ of $\mathbb{R}^2$ induced by the vertical line $t=0$.

Let $w_1,w_2 \in \mathcal{B}$; we define the *concatenation of* w_1 *and* w_2 *along* $\mathcal{S}_0$ *as*

$$\left(w_1 \wedge_{\mathcal{S}_0} w_2\right)(t,z) := \begin{cases} w_1(t,z) & , \quad (t,z) \in \mathcal{S}_- \\ w_2(t,z) & , \quad (t,z) \in \mathcal{S}_0 \cup \mathcal{S}_+ \end{cases} .$$

We may again define an equivalence on the space of solutions.

Definition 5. $w_1,w_2 \in \mathcal{B}$ are *equivalent along* $\mathcal{S}_0$, denoted by $w_1 \sim_{\mathcal{S}_0} w_2$, if for all $w \in \mathcal{B}$:

$$\left[w_1 \wedge_{\mathcal{S}_0} w \in \mathcal{B}\right] \Leftrightarrow \left[w_2 \wedge_{\mathcal{S}_0} w \in \mathcal{B}\right] .$$

If we interpret the partition $(\mathcal{S}_-,\mathcal{S}_0,\mathcal{S}_+)$ as imposing a distinction between "past" $\mathcal{S}_-$, "present" $\mathcal{S}_0$ and "future" $\mathcal{S}_+$, the equivalence of trajectories corresponds to w_1 and w_2 admitting the same future continuations. In the $1D$ case $\mathcal{S}_- = (-\infty,0)$, $\mathcal{S}_0 = \{0\}$ and $\mathcal{S}_+ = (0,+\infty)$, and consequently equivalence of trajectories corresponds to w_1 and w_2 bringing the system to the same state at time $t=0$ (see [8, 13]). For the $2D$ case, a similar property to our definition of equivalence is the notion of Markovianity, see [9, 10]. Note however that in our case there is a clear distinction between the time variable t and the spatial variable z.

Under which conditions are two weak solutions w_1 and w_2 equivalent along $\mathcal{S}_0$? We will first consider this question for *strong* solutions w_1 and w_2; the general answer will then follow from the fact that the strong solutions are dense in the set of weak solutions, cf. [5] and a similar argument in [13]. Write

$$R(\xi,\delta) = \sum_{i=0}^{L} R_i(\delta)\xi^i ,$$

where $R_i \in \mathbb{R}^{p\times q}[\delta]$, and $R_L(\delta) \neq 0$. Observe that $w_i \wedge_{\mathcal{S}_0} w \in \mathcal{B}$, $i=1,2$, if and only if

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left(w_i \wedge_{\mathcal{S}_0} w\right)^{\top}(t,z)\left[R\left(-\frac{\partial}{\partial t},-\frac{\partial}{\partial z}\right)^{\top} \varphi(t,z)\right] dt\, dz = 0 , \qquad (24)$$

for all test functions φ. Now integrate (24) by parts with respect to t and z repeatedly till all derivatives of the function φ have disappeared. Recalling that φ is of compact support (and thus equal to zero for t and z equal to $-\infty$ and ∞), and that $R(\frac{\partial}{\partial t},\frac{\partial}{\partial z})w_i = 0$ in $(-\infty,0]\times\mathbb{R}$, $i=1,2$, it follows that $w_1 \sim_{\mathcal{S}_0} w_2$ if and only if for all compact support infinitely-differentiable test functions φ and for $i=1,2$ it holds that

$$\int_{-\infty}^{+\infty}\begin{bmatrix}\varphi(0,z)\\ \frac{\partial\varphi}{\partial t}(0,z)\\ \vdots\\ \frac{\partial^{L-1}\varphi}{\partial t^{L-1}}(0,z)\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\Pi_{00}(\frac{\partial}{\partial z}) & \dots & \Pi_{0,L-1}(\frac{\partial}{\partial z})\\ \Pi_{10}(\frac{\partial}{\partial z}) & \dots & \Pi_{1,L-1}(\frac{\partial}{\partial z})\\ \vdots & \dots & \vdots\\ \Pi_{L-1,0}(\frac{\partial}{\partial z}) & \dots & \Pi_{L-1,L-1}(\frac{\partial}{\partial z})\end{bmatrix}\begin{bmatrix}w_1(0,z)\\ \frac{\partial w_1}{\partial t}(0,z)\\ \vdots\\ \frac{\partial^{L-1}w_1}{\partial t^{L-1}}(0,z)\end{bmatrix}dz$$

$$=\int_{-\infty}^{+\infty}\begin{bmatrix}\varphi(0,z)\\ \frac{\partial\varphi}{\partial t}(0,z)\\ \vdots\\ \frac{\partial^{L-1}\varphi}{\partial t^{L-1}}(0,z)\end{bmatrix}^{\mathsf{T}}\begin{bmatrix}\Pi_{00}(\frac{\partial}{\partial z}) & \dots & \Pi_{0,L-1}(\frac{\partial}{\partial z})\\ \Pi_{10}(\frac{\partial}{\partial z}) & \dots & \Pi_{1,L-1}(\frac{\partial}{\partial z})\\ \vdots & \dots & \vdots\\ \Pi_{L-1,0}(\frac{\partial}{\partial z}) & \dots & \Pi_{L-1,L-1}(\frac{\partial}{\partial z})\end{bmatrix}\begin{bmatrix}w_2(0,z)\\ \frac{\partial w_2}{\partial t}(0,z)\\ \vdots\\ \frac{\partial^{L-1}w_2}{\partial t^{L-1}}(0,z)\end{bmatrix}dz\,, \tag{25}$$

where $\Pi_{i,j}(\frac{\partial}{\partial z})\in\mathbb{R}^{\mathrm{p}\times q}[\frac{\partial}{\partial z}]$ for $i,j=0,\dots,L$, are certain matrix differential operators (in the spatial variable z) summarizing the remainders at $t=0$ in the repeated integration by parts procedure. (Note that since w_1,w_2 are strong solutions the remainders arising from repeated integration by parts with respect to the spatial variable z are at $-\infty$ and at ∞, and are thus equal to zero.)

Furthermore, the polynomial matrices $\Pi_{i,j}\in\mathbb{R}^{\mathrm{p}\times q}[\delta]$ can be easily obtained from a 2*D bilinear differential form* (see [14]) obtained from $R(\xi,\delta)$. In fact, since the three-variable polynomial matrix $R(-\zeta,\delta)-R(\eta,\delta)$ is zero whenever $\zeta+\eta=0$, we can factorize

$$R(-\zeta,\delta)-R(\eta,\delta)=(\zeta+\eta)\Pi(\zeta,\eta,\delta)\,, \tag{26}$$

for some three-variable polynomial matrix $\Pi(\zeta,\eta,\delta)$. It turns out that

$$\Pi(\zeta,\eta,\delta)=\begin{bmatrix}I_p & \dots & I_p\zeta^{L-1}\end{bmatrix}\begin{bmatrix}\Pi_{00}(\delta) & \dots & \Pi_{0,L-1}(\delta)\\ \vdots & \dots & \vdots\\ \Pi_{L-1,0}(\delta) & \dots & \Pi_{L-1,L-1}(\delta)\end{bmatrix}\begin{bmatrix}I_q\\ \vdots\\ I_q\eta^{L-1}\end{bmatrix}, \tag{27}$$

where $\Pi_{ij}(\frac{\partial}{\partial z})\in\mathbb{R}^{p\times q}[\frac{\partial}{\partial z}]$ are the matrix differential operators as obtained in the integration by parts procedure.

Remark 6. Thus the remainder at $t=0$ given by the polynomial matrices $\Pi_{i,j}\in\mathbb{R}^{p\times q}[\delta]$ is obtained from $R(\xi,\delta)$ by performing the same procedure as in the previous section (for ordinary differential equations) *only with respect to* the indeterminate ξ corresponding to the time variable t.

Due to the arbitrariness of the test function φ, the following result follows [7].

Proposition 7. *Let $R \in \mathbb{R}^{p\times q}[\xi,\delta]$, and define $\mathcal{B}$ as in (23). Let $w_1, w_2 \in \mathcal{B}$; then $w_1 \sim_{\mathcal{S}_0} w_2$ if and only if*

$$\begin{bmatrix} \Pi_{00}(\frac{\partial}{\partial z}) & \dots & \Pi_{0,L-1}(\frac{\partial}{\partial z}) \\ \Pi_{10}(\frac{\partial}{\partial z}) & \dots & \Pi_{1,L-1}(\frac{\partial}{\partial z}) \\ \vdots & \dots & \vdots \\ \Pi_{L-1,0}(\frac{\partial}{\partial z}) & \dots & \Pi_{L-1,L-1}(\frac{\partial}{\partial z}) \end{bmatrix} \begin{bmatrix} w_1(0,z) \\ \frac{\partial w_1}{\partial t}(0,z) \\ \vdots \\ \frac{\partial^{L-1} w_1}{\partial t^{L-1}}(0,z) \end{bmatrix}$$
$$= \begin{bmatrix} \Pi_{00}(\frac{\partial}{\partial z}) & \dots & \Pi_{0,L-1}(\frac{\partial}{\partial z}) \\ \Pi_{10}(\frac{\partial}{\partial z}) & \dots & \Pi_{1,L-1}(\frac{\partial}{\partial z}) \\ \vdots & \dots & \vdots \\ \Pi_{L-1,0}(\frac{\partial}{\partial z}) & \dots & \Pi_{L-1,L-1}(\frac{\partial}{\partial z}) \end{bmatrix} \begin{bmatrix} w_2(0,z) \\ \frac{\partial w_2}{\partial t}(0,z) \\ \vdots \\ \frac{\partial^{L-1} w_2}{\partial t^{L-1}}(0,z) \end{bmatrix}, \tag{28}$$

where $\Pi_{ij} \in \mathbb{R}^{p\times q}[\delta]$, $i,j = 0,\dots,L$, are defined from (26)–(27).

Furthermore, the condition stated in Proposition 7 amounts to first-order representation with respect to only the time variable. Recall the definition of $\Pi(\zeta,\eta,\delta)$, and define

$$x := \begin{bmatrix} \Pi_{00}(\frac{\partial}{\partial z}) & \dots & \Pi_{0,L-1}(\frac{\partial}{\partial z}) \\ \vdots & \dots & \vdots \\ \Pi_{L-1,0}(\frac{\partial}{\partial z}) & \dots & \Pi_{L-1,L-1}(\frac{\partial}{\partial z}) \end{bmatrix} \begin{bmatrix} I_q \\ \vdots \\ \frac{\partial^{L-1}}{\partial t^{L-1}} I_q \end{bmatrix} w\,. \tag{29}$$

Proposition 8. *Let $R \in \mathbb{R}^{p\times q}[\xi,\delta]$, and define $\mathcal{B}$ as in (23). Let $w_1, w_2 \in \mathcal{B}$, and define correspondingly x_1, x_2 as in (29). Then*

$$w_1 \sim_{\mathcal{S}_0} w_2 \Longleftrightarrow x_1(0,z) = x_2(0,z) \textit{ for all } z \in \mathbb{R}\,.$$

Thus, x contains all information necessary to determine whether two trajectories in $\mathcal{B}$ admit the same continuation; for this reason we call x a *state* for $\mathcal{B}$, and we call the polynomial differential operator acting on w on the right of (29) a *state map* for the system of linear PDEs.

Finally, the state x defined in (29) corresponds to a description of $\mathcal{B}$ involving first-order (in time) equations in x, and zeroth-order (in time) equations in w. Observe that from (26), for every $w \in \mathcal{B}$ and corresponding x defined by (29), and every test function φ it holds that

$$\begin{bmatrix} \varphi & \dots & \frac{\partial^{L-1}\varphi}{\partial t^{L-1}} \end{bmatrix} \frac{\partial}{\partial t} x + \begin{bmatrix} \frac{\partial}{\partial t}\varphi & \dots & \frac{\partial^{L}\varphi}{\partial t^{L}} \end{bmatrix} x = \begin{bmatrix} \varphi & \dots & \frac{\partial^{L}\varphi}{\partial t^{L}} \end{bmatrix} \begin{bmatrix} R_0(\frac{\partial}{\partial z}) \\ \vdots \\ (-1)^L R_L(\frac{\partial}{\partial z}) \end{bmatrix} w\,.$$

Denoting with n the number of variables of the state x the above equation can be rewritten as

$$\begin{bmatrix} \frac{\partial}{\partial t}\varphi & \dots & \frac{\partial^{L}\varphi}{\partial t^{L}} \end{bmatrix} \left(\begin{bmatrix} I_n \\ 0_{p\times n} \end{bmatrix} \frac{\partial}{\partial t} x + \begin{bmatrix} 0_{p\times n} \\ I_n \end{bmatrix} x + \begin{bmatrix} R_0(\frac{\partial}{\partial z}) \\ \vdots \\ -(-1)^L R_L(\frac{\partial}{\partial z}) \end{bmatrix} w \right) = 0\,.$$

From the arbitrariness of φ we thus conclude that

$$\begin{bmatrix} I_n \\ 0_{p\times n} \end{bmatrix} \frac{\partial}{\partial t} x + \begin{bmatrix} 0_{p\times n} \\ I_n \end{bmatrix} x + \begin{bmatrix} -R_0(\frac{\partial}{\partial z}) \\ \vdots \\ -(-1)^L R_L(\frac{\partial}{\partial z}) \end{bmatrix} w = 0\,. \tag{30}$$

It is a matter of straightforward verification to check that by *eliminating* x in (30), the set of w-trajectories for which there exists x such that (30) holds is precisely equal to the solutions of the PDEs $R(\frac{\partial}{\partial t}, \frac{\partial}{\partial z})w = 0$; consequently we call (30) a *state representation* of $\mathcal{B}$.

3 From integration by parts to the definition of boundary variables

In the preceding section we considered linear partial differential equations on an unbounded spatial domain $z \in (-\infty, \infty)$. In many cases of interest the spatial domain is *bounded*, and there is an essential role to be played by *boundary variables*. These boundary variables are either *prescribed*, giving rise to partial differential equations with *boundary conditions*, or are the variables through which the system interacts with its environment, leading to *boundary control systems*. Note that in fact the first case (boundary conditions) can be seen to be a special case of the second case (interaction with the environment) in the sense that boundary conditions may be interpreted as corresponding to interaction with a *static* environment (e.g., an ideal constraint or a source system). In this section we will show how integration by parts leads to a natural definition of boundary variables for systems of linear partial differential equations.

Consider a set of linear partial differential equations as before, but now on a *bounded* spatial domain $[a,b]$, that is

$$R(\frac{\partial}{\partial t}, \frac{\partial}{\partial z})w = 0, \quad z \in [a,b] \tag{31}$$

We now perform the same integration by parts procedure as in the previous section, however interchanging the t and z variable, and replacing the line $t = 0$ by the two lines $z = a$ and $z = b$. Dually to the situation considered in the previous section this will correspond to the factorization

$$R(\xi, -\gamma) - R(\xi, \varepsilon) = (\gamma + \varepsilon)\Sigma(\xi, \gamma, \varepsilon)\,, \tag{32}$$

for some three-variable polynomial matrix $\Sigma(\xi, \gamma, \varepsilon)$ (i.e., we do the factorization with respect to the indeterminate δ corresponding to the spatial variable z). As before in the case of factorization with respect to ξ we thus obtain

$$\Sigma(\xi, \gamma, \varepsilon) = \begin{bmatrix} I_{\mathrm{p}} & \dots & I_{\mathrm{p}}\gamma^{N-1} \end{bmatrix} \begin{bmatrix} \Sigma_{00}(\xi) & \dots & \Sigma_{0,N-1}(\xi) \\ \vdots & \dots & \vdots \\ \Sigma_{N-1,0}(\xi) & \dots & \kappa_{N-1,N-1}(\xi) \end{bmatrix} \begin{bmatrix} I_q \\ \vdots \\ I_q \varepsilon^{N-1} \end{bmatrix}, \tag{33}$$

where $\Sigma_{ij}(\frac{\partial}{\partial t}) \in \mathbb{R}^{p\times q}[\frac{\partial}{\partial t}]$ equal the matrix differential operators obtained in integration by parts with respect to the spatial variable z.

Then define the vectors $w_\partial(a)(t), w_\partial(b)(t)$ (functions of time t) as

$$\begin{aligned} w_\partial(a)(t) &:= \begin{bmatrix} \Sigma_{00}(\frac{\partial}{\partial t}) & \cdots & \Sigma_{0,N-1}(\frac{\partial}{\partial t}) \\ \vdots & \cdots & \vdots \\ \Sigma_{N-1,0}(\frac{\partial}{\partial t}) & \cdots & \Sigma_{N-1,N-1}(\frac{\partial}{\partial t}) \end{bmatrix} \begin{bmatrix} I_q \\ \vdots \\ \frac{\partial^{L-1}}{\partial z^{N-1}} I_q \end{bmatrix} w(t,a) \\ w_\partial(b)(t) &:= \begin{bmatrix} \Sigma_{00}(\frac{\partial}{\partial t}) & \cdots & \Sigma_{0,N-1}(\frac{\partial}{\partial t}) \\ \vdots & \cdots & \vdots \\ \Sigma_{N-1,0}(\frac{\partial}{\partial t}) & \cdots & \Sigma_{N-1,N-1}(\frac{\partial}{\partial t}) \end{bmatrix} \begin{bmatrix} I_q \\ \vdots \\ \frac{\partial^{L-1}}{\partial z^{N-1}} I_q \end{bmatrix} w(t,b) \end{aligned} \tag{34}$$

We claim that the variables $w_\partial(a), w_\partial(b)$ qualify as a natural set of *boundary variables*. Indeed, they provide just enough information to *extend* a solution on the spatial domain $[a,b]$ to a *weak solution* of the same set of partial differential equations on a *larger* spatial domain $[c,d]$, with $c \leq a, b \leq d$. Indeed, as in the previous section for the case of the computation of the state at $t = 0$, the vector $w_\partial(a)$ provides just enough information to extend a solution $w(t,z)$ to a weak solution for values of the spatial variable z to the left of a; while the same holds for $w_\partial(b)$ with regard to extension of the solution to a weak solution for values of z to the right of b.

Example 9. Consider a system of linear conservation laws

$$\frac{\partial w_1}{\partial t}(t,z) = -\frac{\partial}{\partial z}\frac{\partial \mathcal{H}}{\partial w_2}(w_1(t,z), w_2(t,z))$$

$$\frac{\partial w_2}{\partial t}(t,z) = -\frac{\partial}{\partial z}\frac{\partial \mathcal{H}}{\partial w_1}(w_1(t,z), w_2(t,z))$$

for a quadratic Hamiltonian density

$$\mathcal{H}(w_1, w_2) = \frac{1}{2}\begin{bmatrix} w_1 & w_2 \end{bmatrix} Q \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

with Q a symmetric 2×2 matrix, on a spatial domain $z \in [a,b]$. Note that many physical systems, including the telegrapher's equations of the dynamics of an ideal (lossless) transmission line and the equations of a linear vibrating string, are of this form, with $\int_a^b \mathcal{H}(w_1, w_2) dz$ denoting the total energy stored in the system, see [12]. Computing the boundary vectors $w_\partial(a), w_\partial(b)$ amounts to

$$w_\partial(a)(t) = \begin{bmatrix} \frac{\partial \mathcal{H}}{\partial w_2}(t,a) \\ \frac{\partial \mathcal{H}}{\partial w_1}(t,a) \end{bmatrix} \quad w_\partial(b)(t) = \begin{bmatrix} \frac{\partial \mathcal{H}}{\partial w_2}(t,b) \\ \frac{\partial \mathcal{H}}{\partial w_1}(t,b) \end{bmatrix}$$

These are exactly the boundary variables as defined in [12] based on physical considerations. For example, in the case of the telegrapher's equations, the variables w_1, w_2 will be the charge, respectively, flux density, while the boundary vectors

$w_\partial(a), w_\partial(b)$ will be the vector of current and voltage at $z = a$, respectively $z = b$. Clearly, these are the natural boundary variables.

Similarly, in the case of a vibrating string the vector of boundary variables at $z = a, b$ will consist of the velocity and force at these boundary points.

4 Conclusions and outlook

Although we have restricted ourselves in this paper to PDEs involving a single spatial variable z the construction of state maps given immediately extends to systems of partial differential equations involving multiple spatial variables, of the general form

$$R(\frac{\partial}{\partial t}, \frac{\partial}{\partial z_1}, \frac{\partial}{\partial z_2}, \cdots, \frac{\partial}{\partial z_k})w = 0 \,. \tag{35}$$

Indeed, by factorizing

$$R(-\zeta, \delta_1, \cdots, \delta_k) - R(\eta, \delta_1, \cdots, \delta_k) = (\zeta + \eta)\Pi(\zeta, \eta, \delta_1, \cdots, \delta_k) \,, \tag{36}$$

the polynomial matrix $\Pi(\zeta, \eta, \delta_1, \cdots, \delta_k)$, written out as

$$\Pi(\zeta, \eta, \delta_1, \cdots, \delta_k) = \\ \begin{bmatrix} I_p & \dots & I_p\zeta^{L-1} \end{bmatrix} \begin{bmatrix} \Pi_{00}(\delta_1, \cdots, \delta_k) & \dots & \Pi_{0,L-1}(\delta_1, \cdots, \delta_k) \\ \vdots & \dots & \vdots \\ \Pi_{L-1,0}(\delta_1, \cdots, \delta_k) & \dots & \Pi_{L-1,L-1}(\delta_1, \cdots, \delta_k) \end{bmatrix} \begin{bmatrix} I_q \\ \vdots \\ I_q\eta^{L-1} \end{bmatrix}, \tag{37}$$

defines the state map

$$x := \begin{bmatrix} \Pi_{00}(\frac{\partial}{\partial z_1}, \cdots, \frac{\partial}{\partial z_k}) & \dots & \Pi_{0,L-1}(\frac{\partial}{\partial z_1}, \cdots, \frac{\partial}{\partial z_k}) \\ \vdots & \dots & \vdots \\ \Pi_{L-1,0}(\frac{\partial}{\partial z_1}, \cdots, \frac{\partial}{\partial z_k}) & \dots & \Pi_{L-1,L-1}(\frac{\partial}{\partial z_1}, \cdots, \frac{\partial}{\partial z_k}) \end{bmatrix} \begin{bmatrix} I_q \\ \vdots \\ \frac{\partial^{L-1}}{\partial t^{L-1}} I_q \end{bmatrix} w \,. \tag{38}$$

In a similar fashion the construction of boundary variables can be extended to higher-dimensional spatial domains.

A very much challenging avenue for further research concerns the extension of the ideas put forward in this paper to *nonlinear* higher-order ordinary or partial differential equations. Some initial ideas for doing this, based on considering the variational (i.e., linearized) systems, have been proposed in [11], also drawing inspiration from some results in [2].

Bibliography

[1] R. W. Brockett. Path integrals, Lyapunov functions, and quadratic minimization. In *Proceedings of the 4th Allerton Conference on Circuit and System Theory*, pages 685–698, 1966. Cited p. 440.

[2] P. E. Crouch and A. J. van der Schaft. *Variational and Hamiltonian control systems*. Springer, 1987. Cited p. 447.

[3] E. L. Ince. *Ordinary differential equations*. Dover, 1956. Cited p. 440.

[4] T. Kailath. *Linear Systems*. Prentice-Hall, 1980. Cited p. 438.

[5] J. W. Polderman and J. C. Willems. *Introduction to mathematical system theory: A behavioral approach*. Springer, 1997. Cited p. 442.

[6] P. Rapisarda and A. J. van der Schaft. Canonical realizations by factorization of constant matrices. *Systems and Control Letters*, 61(8):827–833, 2012. Cited p. 441.

[7] P. Rapisarda and A. J. van der Schaft. Trajectory concatenability for systems described by partial differential equations. In *Proceedings of the 20th International Symposium on Mathematical Theory of Networks and Systems*, 2012. Paper no. 011 (no pagination). Cited pp. 437, 441, and 443.

[8] P. Rapisarda and J. C. Willems. State maps for linear systems. *SIAM Journal on Control and Optimization*, 35(3):1053–1091, 1997. Cited pp. 439 and 442.

[9] P. Rocha and J. C. Willems. Markov properties for systems described by PDEs and first-order representations. *Systems and Control Letters*, 55(7):538–542, 2006. Cited p. 442.

[10] P. Rocha and J. C. Willems. Markovian properties for 2D behavioral systems described by PDEs: The scalar case. *Multidimensional Systems and Signal Processing*, 22(1–3):45–53, 2011. Cited p. 442.

[11] A. J. van der Schaft. Representing a nonlinear input-output differential equation as an input-state-output system. In V. Blondel, E. D. Sontag, M. Vidyasagar, and J. C. Willems, editors, *Open problems in systems theory*, pages 171–176. Springer, 1998. Cited pp. 439 and 447.

[12] A. J. van der Schaft and B. M. Maschke. Hamiltonian formulation of distributed-parameter systems with boundary energy flow. *Journal of Geometry and Physics*, 42:166–194, 2002. Cited p. 446.

[13] A. J. van der Schaft and P. Rapisarda. State maps from integration by parts. *SIAM Journal on Control and Optimization*, 49(6):2145–2439, 2011. Cited pp. 437, 439, 441, and 442.

[14] J. C. Willems and H. K. Pillai. Lossless and dissipative distributed systems. *SIAM Journal on Control and Optimization*, 40(5):1406–1430, 2002. Cited p. 443.

[15] J. C. Willems and H. L. Trentelman. On quadratic differential forms. *SIAM Journal on Control and Optimization*, 36(5):1703–1749, 1998. Cited pp. 439 and 440.

Canonical forms for pseudo-continuous multi-mode multi-dimensional systems (M^3D) with conservation laws

Erik I. Verriest
Georgia Institute of Technology
Atlanta, GA, USA
erik.verriest@ece.gatech.edu

Abstract. A class of multi-mode multi-dimensional (M^3D) systems is described where the modes may have dynamics of different dimension. It is assumed that the timing of the switches is fully controllable. Pseudo-continuity (p) is introduced as a desirable property from a modeling and control point of view: At the mode switchings the state transitions are constrained to be such that at any sequential instantaneous switching cycle starting at the mode of lowest dimension in this cycle and returning to it does not change the state. A precise definition of state space is given (a sheaf), and canonical forms are derived for pM^3D systems. In addition we consider the class of auto-hybrid systems (i.e., systems where the state transitions are triggered by conditions on partial states) for which the switching maps are governed by certain conservation principles (of linear and quadratic forms). It is shown that canonical forms of the above form exist if certain additional symmetries hold.

1 Introduction

After being introduced by Witsenhausen [12] and perhaps revitalized by Brockett [2], hybrid systems of all types and forms have been described extensively. See for instance [1, 4–6]. In a series of papers, [7, 8, 10] a class of switched systems was introduced, where the different modes do not necessarily have the same dimension. We called these systems multi-mode multi-dimensional (M^3D) systems. Let Ω denote the set of modes (also called *locations* in [6]). We shall assume that Ω is a countable set, and label the modes by the integers 1 to N (if Ω is finite) and $\mathbb{Z}_+$ otherwise.

For modes with linear dynamics, $\dot{x}_i = A_i x_i + b_i u_i$; $i \in \Omega$, the x_i is called the *partial state* (of mode i) with $\dim x_i = n_i$. The switches between the modes can be governed by an external signal (control), in which case we call the system an *exo-hybrid* system (also called time-controlled in [6]), or the transitions can be triggered by conditions on the partial state in the mode (an *auto-hybrid* system or state-controlled in [6]). We shall assume that for an exo-hybrid system, the timing of the mode switches is completely free, but that if the system is in mode i, only the modes in a subset $\Omega_i \subset \Omega$ can be switched to. To complete the hybrid system description, we must also specify the transitions (aka resets) of the partial state from one mode to the next. With the obvious notation $\tau^{\pm}$ referring to the infinitesimal time before or after the mode switching time τ, we define for all $j \in \Omega_i$ the reset maps $\mathcal{S}_{ji} : \mathbb{R}^{n_i} \to \mathbb{R}^{n_j}$

$$x_j(\tau^+) = \mathcal{S}_{ji}(x_i(\tau^-)). \tag{1}$$

It follows that the pair (i,x_i) characterizes the future behavior of the system, provided the continuous controls, u_k, timing of the switches and sequencing of the modes is known.

In Section 2, we focus on the exo-pM^3D systems, and derive a precise notion of their structure, including the notion of a state space. This will set the stage for Section 3, where canonical forms are derived. A specific class of auto-pM^3D systems is discussed in Section 4. The structure implied by sequential switching (if multiple switches can indeed occur) is discussed. Finally Section 5 gives an example of systems restricted by conservation laws. For a linear mode, the set of all linear and quadratic invariants is characterized. There, we also show that pseudo-continuity can still be assumed if an additional symmetry is introduce (the exchange operator).

It is my pleasure and an honor to contribute this article to this Festschrift in honor of Uwe Helmke on the occasion of his 60th birthday. I had the joy and the privilege of collaborating with him in a study of the structure of periodic systems [3]. In this work, discrete periodic systems were studied where the transition from state x_k to state x_{k+1} may involve different dimensions. In the present work our interest is to explore such state transitions in a non-periodic setting.

2 Exo-pM^3D systems

This section summarizes some of the results shown in [10], and proofs are omitted. The idea of defining the state space as a sheaf is new.

2.1 Sheaf as state space

It is not directly clear how a state space should be defined for a M^3D-system[1]. While the pair (i,x_i) defines a *partial state* in the i-th mode, this pair does not qualify as a state of the hybrid system, for the reason that it has a varying dimension. The notion of a state space should be an invariant construct, otherwise one cannot reasonably talk about trajectories as paths in the state space. To embed these partial states in a stationary structure, let Ω be equipped with the discrete topology (all subsets of Ω are open sets). Assign to each nonempty subset, $U = \{q_1, q_2, \ldots, q_\nu\}$, of Ω, where $q_1 > q_2 > \ldots > q_\nu$, the set $\mathcal{F}(U) = \mathbb{R}^{n_{q_1}} \oplus \cdots \oplus \mathbb{R}^{n_{q_\nu}}$, where n_i is the dimension of the i-th mode. For each pair of open sets U and V with $V = \{q_{i_1}, q_{i_2}, \ldots, q_{i_\mu}\} \subset U = \{q_1, \ldots, q_\nu\}$, define the restriction homomorphisms $r_V^U : \mathcal{F}(U) \to \mathcal{F}(V)$ by

$$r_V^U(x_{q_1}, \ldots, x_{q_\nu}) = (x_{q_{i_1}}, \ldots, x_{q_{i_\nu}}).$$

This defines $\mathcal{F}$ as a presheaf (see [11]) over the topological space Ω. In addition, for every collection U_i of subsets of Ω with $U = \cup U_i$, the following properties hold:

(i) If $x,y \in \mathcal{F}(U)$, and $r_{U_i}^U(x) = r_{U_i}^U(y)$, then $x = y$.

(ii) If $x^{(i)} \in \mathcal{F}(U_i)$ and for $U_i \cap U_j \neq \varnothing$, then $r_{U_i \cap U_j}^{U_i}(x^{(i)}) = r_{U_i \cap U_j}^{U_j}(x^{(j)})$ for all i, then there exists $x \in \mathcal{F}(U)$ such that $r_{U_i}^U(x) = x^{(i)}$ for all i.

[1] This section corrects erroneous statements in [7–10].

Properties (i) and (ii) make the presheaf $\mathcal{F}$ into a sheaf. Condition (i) expresses that data defined on large open sets are uniquely determined locally (restriction), and (ii) expresses that local data can be pieced together to give the global picture (overlaps). Roughly speaking, a sheaf over a space Ω is a mapping $\mathcal{F}$:{open sets in Ω}→ {algebraic objects}. The algebraic objects are here direct sum vector spaces. Clearly this gives some kind of product structure for the state space. Note however that it is not possible to characterize the state space as a fibre bundle since it cannot look the same at any point of the base space Ω: A single fibre F to which the fibres $\mathbb{R}^{n_i}$ over q_i in the base space Ω are homeomorphic does not exist in the multi-dimension case. Represent the state by $\mathbf{x} = (q;x_1,x_2,\ldots,x_N) \in \mathbf{X} = \Omega\times(\mathbb{R}^{n_1}\oplus\cdots\oplus\mathbb{R}^{n_N})$ equipped with the *equivalence*: $\forall i = 1,\ldots,N; q = i$ implies

$$(q;x_1,\cdots,x_{i-1},x_i,x_{i+1}\cdots x_N) \sim (q;y_1,\cdots,y_{i-1},x_i,y_{i+1}\cdots y_N)$$

for all $(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_N) \in (\mathbb{R}^{n_1}\oplus\cdots\oplus\mathbb{R}^{n_{i-1}}\oplus\mathbb{R}^{n_{i+1}}\oplus\cdots\oplus\mathbb{R}^{n_N})$. The equivalence expresses he redundancy of the information from any past modes. For ease of notation, we will continue to denote the above state by (i,x_i), or even x_i.

2.2 Pseudo-continuity of the partial state

In [10] we singled out the class of hybrid systems with the property that the state cannot be altered by instantaneously switching through a set of modes and returning to the original mode when none of these modes in the cycle have lower dimension than the initial one. Note that switching from a given mode to a lower dimensional one and back necessarily involves a loss of information. The rationale of this is that there is no free ride (cost-free control) possible by instantaneous switching. This led us to define the notion of *pseudo-continuity*, which was stated more generally:

Definition 1. The M^3D exo-system is *pseudo-continuous* if for any sequence of modes $q(1) \to q(2) \to \cdots \to q(k)$ where $q(i) \in \Omega$, and which does does not contain any discrete state (mode) with dimension less than the minimum of the dimension in the initial and final mode, (min$(n_{q(1)},n_{q(k)})$), the autonomous transitions satisfy

$$\mathcal{S}_{n_{q(k)}n_{q(k-1)}} \circ\cdots\circ \mathcal{S}_{n_{q(2)}n_{q(1)}} = \mathcal{S}_{n_{q(k)}n_{q(1)}}. \qquad (2)$$

Costfree control via instantaneous switching only requires (2) for $n_{q_1} = n_{q_2}$. The set of transition maps for a pseudo-continuous M^3D system fails to generate a semigroup under concatenation, but still inherits a nice structure. A complete study of the structure of the dynamics involves first a discussion of the allowed mode transitions (the “possible"), as well as the finer structure of the transitions. This finer structure will be studied for linear reset maps: $\mathcal{S}_{ji}(x_i) = S_{ji}x_i$ with $S_{ji} \in \mathbb{R}^{n_j\times n_i}$.

2.3 Isochronous structure of exo-qM^3D systems

We studied the switching behavior of the exo-hybrid system in [10]. Borrowing from the theory of Markov chains, a mode q_1 is said to *connect* to mode q_2 if a transition from q_1 to q_2 is allowed ($q_2 \in \Omega_{q_1}$). This led us to define a mode transition possibility matrix, T. Its ij-entry is 1 if the mode transition from $i \in \Omega$ to $j \in \Omega$ is allowed, and zero else. By introducing a *Booleanization map*, $\mathcal{B}: (\mathbb{Z}_+)^{N\times N} \to \{0,1\}^{N\times N}$, with $(\mathcal{B}(M))_{ij} = 1 - \delta_{M_{ij},0}$, ($\delta_{m0}$ is the Kronecker delta), and a *possibility product*, $A \odot B = \mathcal{B}(AB)$, the k-transition possibility matrix $T^{(k)}$ is expressible as the power $T^{\odot k}$.

The isochronous switching behavior is the behavior induced by instantaneous sequential switching through several modes. It is characterized by the fine structure matrix, $\mathbb{S}$, a block matrix whose (ij)-block is the reset matrix S_{ij}, for $i \neq j$. Without loss of generality, the diagonal blocks of $\mathbb{S}$ may be chosen as identity matrices (S_{jj} was not defined as a reset matrix).

The particular structure of $\mathbb{S}$ induced by the pseudo-continuity was investigated in [10], where the following result is shown:

Theorem 2. *Let $F_{p,q}(\mathbb{R})$ denote the subset of $\mathbb{R}^{p\times q}$ containing all matrices of full rank. The fine structure matrix $\mathbb{S}$ of a pseudo-continuous M^3D system is parameterized by $(G_1, G_2, \ldots, G_{N-1}) \in F_{n_1,n_2}(\mathbb{R}) \times \ldots \times F_{n_{N-1},n_N}(\mathbb{R})$, and $(\overline{G}_1, \overline{G}_2, \ldots, \overline{G}_{N-1}) \in F_{n_2,n_1}(\mathbb{R}) \times \ldots \times F_{n_N,n_{N-1}}(\mathbb{R})$. The (ij)-th block is*

$$S_{ij} = G_i G_{i+1} \cdots G_{j-1} \quad \text{for} \quad j > i \tag{3}$$

$$S_{ij} = \overline{G}_{i-1} \overline{G}_{i-2} \cdots \overline{G}_j \quad \text{for} \quad j < i. \tag{4}$$

but with $\overline{G}_i G_i = I_{n_{i+1}}$. (Note that by Sylvester's rank theorem, the conditions $\overline{G}_i G_i = I_{n_{i+1}}$ already imply $G_i \in F_{n_i,n_{i+1}}(\mathbb{R})$ and $\overline{G}_i \in F_{n_{i+1},n_i}(\mathbb{R})$).

Summary: The mode transition possibility rate matrix of a reducible pseudo-continuous M^3D system with N nontransient modes is a direct sum of ν blocks $\mathbf{1}_{k_i}\mathbf{1}_{k_i}^\top$, $i = 1, \ldots, \nu$ with $\sum_{i=1}^{\nu} k_i = N$, and where $\mathbf{1}_k$ is the vector of dimension k with all entries one. Each such block represents an irreducible component comprised of $k_i \leq N$ modes. The associated transition possibility matrix decomposes also as the direct sum of ν irreducible ones with corresponding dimensions k_i, $i = 1 \ldots, \nu$. The fine structure matrix blocks are generated by the elements G_i directly above the diagonal, and elements $\overline{G}_i$ directly underneath, but constrained by $\overline{G}_i G_i = I$.

3 Canonical forms

Each mode realization can be transformed under similarity, leaving the external behavior invariant. Invoking a different base change in each stalk (mode) of an irreducible set transforms the fine structure matrices. This can be exploited to obtain canonical forms of the instantaneous mode change switching behavior of the system. See [10] for the proof.

Theorem 3. *The canonical fine structure transition matrix of an irreducible set of modes in a pseudo-continuous M^3D system is given by Figure* 1*, where the solid diagonal lines have entries 1.*

It should be noted that the canonical form for the pseudo-continuous M^3D system described in Theorem 3 is obtained by applying the similarity T_i to mode i (of dimension n_i). Even if $n_{i+1} = n_i$, the similarity applied to the two modes will be different in general. The pseudo-continuity is what makes this simple canonical form

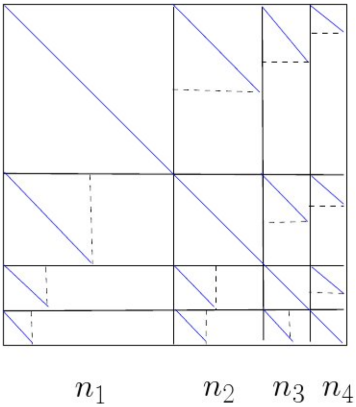


Figure 1: Canonical Form of the Fine Structure Matrix for a pM^3D system.

of the fine structure matrix possible. It is completely determined by the (ordered) stalk dimensions, n_i, $i = 1,\dots,N$, and contains no other free parameters. It follows that for an irreducible mode set in canonical form, the mode transitions form a nested set of either *pure projections*, $[I,0]$, when mapping to lower dimensional fibers, or *pure embeddings*, $[I,0]^\top$, into higher dimensional fibers.

Consider now an irreducible mode set, given in its canonical form (this is solely parameterized by the (ordered) N-tuple of fibre dimensions, $(n_1 \geq n_2 \geq \cdots \geq n_N)$. Let $\mathbb{S}^*$ be the canonical fine structure transition matrix. The group of all state space transformations is $\mathcal{G} = G\ell_{n_1}(\mathbb{R}) \times G\ell_{n_2}(\mathbb{R}) \times \cdots \times G\ell_{n_N}(\mathbb{R})$. The *stabilizer subgroup* (aka isotropy subgroup) of $\mathbb{S}^*$ under the action of $\mathcal{G}$ is the subgroup of $\mathcal{G}$ that leaves $\mathbb{S}^*$ invariant. It is easily verified that this stabilizer subgroup for the 2-tuple, $(p \geq q)$, is $G\ell_{p-q}(\mathbb{R}) \times G\ell_q(\mathbb{R})$ and its elements, T_{stab}, have the form

$$T_{\text{stab}} = \left[\begin{array}{cc|c} T & & \\ & T_c & \\ \hline & & T \end{array}\right], \quad T \in G\ell_q(\mathbb{R}),\ T_c \in G\ell_{p-q}(\mathbb{R}).$$

There are therefore $q^2 + (p-q)^2$ degrees of freedom, q^2 for the low dimensional fibre, and $(p-q)^2$ for the high dimensional fibre. This freedom can be exploited to constrain the structural matrices of the continuous time dynamics on each fibre. For instance, the general bi-modal pseudo continuous M^3D system with fibre dimensions $(2,1)$ has the canonical input-to-state canonical forms

$$\left(\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \begin{bmatrix} b \\ 1 \end{bmatrix}\right), (a_1, 1)\right), \left(\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \begin{bmatrix} b \\ 0 \end{bmatrix}\right), (a_1, 1)\right),$$

(with the a and b free parameters) if mode 2 is reachable; and if not,

$$\left(\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, (a_1, 0)\right)\right),$$

with $[\beta_1,\beta_2] \in \{[1,1],[1,0],[0,1],[0,0]\}$. The latter case is excluded for sure if the overall bi-modal system is reachable. In general, elements of the stabilizer group of $\mathbb{S}^*$ have a block diagonal structure, consistent with the partitioning of $\mathbb{S}^*$. It is readily verified that the previous discussion extends to

$$\mathcal{G}_{\text{stab}}(n_1 \geq n_2 \geq \cdots \geq n_N) = G\ell_{n_N}(\mathbb{R}) \times G\ell_{n_{N-1}-n_N}(\mathbb{R}) \times \cdots \times G\ell_{n_1-n_2}(\mathbb{R}),$$

and has dimension (setting $n_{N+1} = 0$)

$$\dim \mathcal{G}_{\text{stab}}(n_1 \geq n_2 \geq \cdots \geq n_N) = \sum_{i=1}^{N} (n_i - n_{i+1})^2.$$

4 Auto-pM^3D systems

Many of the ideas of exo-hybrid systems carry over to the auto-hybrid case. However we make a distinction between between *autonomous switching* auto-hybrid systems, where the mode switch is triggered by a state condition of the form $h(x_q) = 0$, where x_q is the partial state in the stalk above the current mode $q \in \Omega$, and *autonomous hybrid automata* as described in [6, p. 9]. In the latter, the modes (called locations) are determined by the location invariants. This class of systems can only be defined for modes of the same dimension, as the location invariants partition the fixed state space. Dichotomies of the form $h(x) > 0$ and $h(x) < 0$ are typical for bimodal systems. The dynamics on $h(x) = 0$ is left unspecified, but sliding mode behavior on $h(x) = 0$ is possible if on either side of the separating manifold $h(x) = 0$ the vector fields point towards the other region. Typically the transition matrices are then also the identity matrix. Hence, h must have codimension one, while for autonomous switching systems, this is not required. We also note that it is possible to bounce off the manifold $h(x) = 0$ and resume the motion in the *same* half space but with different dynamics. That, of course is impossible behavior for autonomous hybrid automata.

4.1 Mode insertion

The remainder part of the section will focus on autonomous switching systems. Consider first a scalar system without external input, given by the dynamics $\dot{x}_1 = x_1$ for mode (1), and $\dot{x_2} = -x_2$ for mode (2). Let the trigger for both modes be given by $h(x) \equiv x - 1 = 0$. Clearly, when the partial state in mode (1) hits $x = 1$, which only happens for an initial state in the interval (0,1), then the dynamics switch to $\dot{x} = -x$. However the initial state for this mode is $x = 1$, and hence this directly prescribes another switch, and so ad infinitum. The problem is not well-posed. As a way out of this impasse, we suggest inserting a transition mode, (0), with state $[\xi,\eta,\tau]^\top \in \mathbb{R}^3$ and dynamics

$$\begin{aligned}\dot{\xi} &= \eta\\ \dot{\eta} &= 0\\ \dot{\tau} &= 1\end{aligned}$$

Define two trigger conditions for this mode: The trigger for the transition from (0) to (2) is the two-dimensional

$$h_{20}(\xi,\eta,\tau) \equiv \begin{bmatrix} \tau-\varepsilon \\ \eta+1 \end{bmatrix} = 0,$$

for some $\varepsilon > 0$. The trigger for the transition from (0) to (1) is

$$h_{10}(\xi,\eta,\tau) \equiv \begin{bmatrix} \tau-\varepsilon \\ \eta-1 \end{bmatrix} = 0.$$

Define the transitions $S_{01}x = [x,-x,0]^\top$, $S_{02}x = [x,x,0]^\top$, and $S_{i0}[\xi,\eta,\tau]^\top = \xi$ for $i = 1,2$. Then for $x_0 \in (0,1)$, at time $t_1 = -\log x_0$, the trigger state $x = 1$ is reached. The mode switches to the inserted (transition) mode (0) with entrance state $[1,-1,0]$, The (0)-dynamics yields the exit state $[1-\varepsilon,-1,\varepsilon]$, at which point h_{20} triggers the transition to mode (2), with the entrance state $x_2 = 1-\varepsilon$. Now the subsequent dynamics evolves in mode (2), and this regardless how small we let $\varepsilon > 0$ be. Likewise, if $x = 1$ is reached from an initial state $x_0 > 1$ in mode (2), the condition $h_{02}(x) = 0$ triggers a switch to mode (0) with entrance state $[1,1,0]^\top$. Dynamics in (0) yields the exit state $[1+\varepsilon,1,\varepsilon]$, where h_{10} triggers the subsequent switch to mode (1) with entrance state $1+\varepsilon$. There are no further switches. A complete flow is shown in Figure 2. Note that in this figure, mode (0) has been represented symbolically, and only the relevant part of the flow is shown. In the limit for $\varepsilon \to 0$, only small neighborhoods of two points, $[1,\pm1,0]^\top$, are needed. We note that the $\dot{\tau} = 1$ is added to the dynamics in mode (0) in order to keep a causal relation between the switchings. On the other hand, the choice of the resets contains some arbitrariness. Had we chosen $S_{01}x = [x,x,0]^\top$, $S_{02}x = [x,-x,0]^\top$, then fast oscillation about $x = 1$ with fast mode switching (chattering). In the limit, this gives a "sliding mode" (in fact here simply a dynamic equilibrium, $x = 1$. Hence mode (0) may be identified with a single point (zero-dimensional mode). The flow is shown in Figure 3 on the next page. The basin of attraction of the equilibrium $x = 1$ is $((1)\times(0,1))\cup((2)\times(1,\infty))$. Perhaps

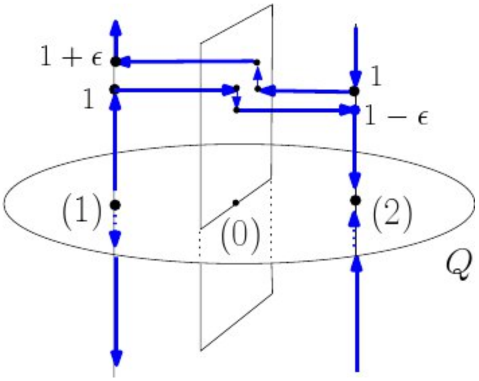


Figure 2: Flow of hybrid system.

more important is the fact that here the solution is uniquely defined for the partial state initial condition, $x = 1$, in either mode. The combined transition matrices are $S_{20}S_{01} = S_{10}S_{02} = 1$, and although neither S_{12} nor S_{21} are defined in this case, this property has the allure of pseudo-continuity. Which model should be used depends on the context, which is not given in the original hybrid description.

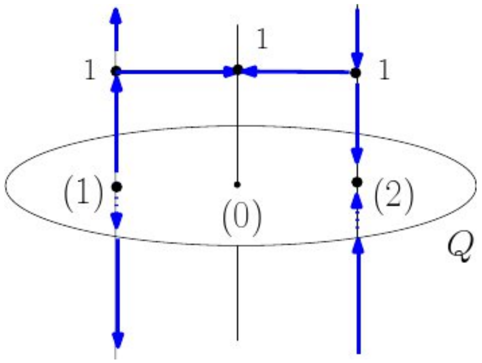


Figure 3: Flow of hybrid system with 0-dimensional mode.

More generally, let for all $i \in \Omega$, the trigger conditions h_{ji} and the state transitions matrices S_{ji} be given for all $j \in \Omega_i$. Assume in addition that the realizations (A_i, B_i) in each mode are reachable. Then the reachability condition for the switched system is governed by the switching structure, i.e. the interplay of the h_{ji} and the S_{ji} for all pairs $(i, j) \in \Omega^2$ where these are defined. If for $i \in \Omega$, and $j \in \Omega_i$ it holds that the condition $h_{ji}(x_i) = 0$ implies for all $k \in \Omega_j$ that $h_{kj}(S_{ji}x_i) \neq 0$, then a jump from mode (i) to mode (j) cannot be followed instantaneously by another jump. Hence if the above condition holds for all $(i) \in \Omega$, all jumps must be isolated in time. The switched system is then asynchro-sequential. On the other hand, if for some x_i it holds that $h_{j_1 i}(x_i) = 0$, $h_{j_2 j_1}(S_{j_1 i}x_i) = 0, \ldots, h_{j_k, j_{k-1}}(S_{j_{k-1}, j_{k-2}} \cdots S_{j_2, j_1} S_{j_1, i} x_i) = 0$, then an instantaneous path $(i, j_1, j_2, \ldots, j_k)$ through Ω is possible, and in fact mandatory for initial partial state x_i. The system is synchro-sequential. Moreover if for all ℓ, $h_{\ell, j_k}(S_{j_k, j_{k-1}} \cdots S_{j_2, j_1} S_{j_1, i} x_i) \neq 0$, then this path of k instantaneous consecutive switches cannot be extended further. This allows the reduction of the switching behavior of the system as follows. If the state (i, x_i) leads after its maximal number, $k > 1$, of instantaneous but consecutive switches to (ℓ, x_ℓ), then erase all intermediate modes and consider only the one step transition from (i, x_i) to (ℓ, x_ℓ). Obviously the trigger condition for this transition is the union of the conditions $h_{j_1 i} = 0$, $h_{j_2 j_1} \circ S_{j_1 i} = 0, \ldots$, while the entrance state in mode ℓ is $S_{\ell \times} \cdots S_{\times i} x_i$. Hence we define the latter as $S_{\ell i} x_i$.

4.2 Well posed hybrid systems

The conditions $h_{ji}(x_i) = 0, S_{ji}x_i = x_j, h_{ki}(x_i) = 0, S_{ki}(x_i) = x_k$ imply resets to two different states from the same state, hence nonuniqueness of the solution. Likewise, $h_{ji}(x_i) = 0, S_{ji}x_i = x_j, h_{ij}(x_j) = 0, S_{ij}(x_j) = x_i$ lead to an impasse as illustrated above. The auto-hybrid system is said to be well-posed if its solutions are unique and no deadlocks occur.

4.2.1 Graph representation

The *synchro-sequential* switching structure of the switching circuit with linear reset maps is determined by the set of possible instantaneous jumps between modes, and can be represented by a digraph. Let Ω be the (discrete) set of modes, and let $\mathbf{X}^{(q)} = \mathcal{F}(q)$ denote the partial state-space for mode q. Define $\Sigma_{ji} = \{x_i \in \Omega \,|\, h_{ji}(x_i) = 0\}$, i.e., the trigger set for jumps from mode (i) to mode (j). Let Ω_i be the set of modes that are accessible (via the resets) from mode (i). Let also $\Sigma_i = \cup_{j \in \Omega_i} \Sigma_{ji}$, be the set of all exit points of $\mathbf{X}^{(i)}$. The allowed instantaneous transition from $x_i \in \mathbf{X}^{(i)}$ to $x_j \in \mathbf{X}^{(j)}$ will be represented by a directed edge: x_i, an exit point for $\mathbf{X}^{(i)}$ is the initial point of the directed edge, characterized by $h_{ji}(x_i) = 0$, and x_j, the entrance point in $\mathbf{X}^{(j)}$, characterized by $x_j = S_{ji}x_i$, is the final point of the directed edge. A final edge is called terminal for the digraph, if it is not the initial point of another edge, (equivalently, its outdegree is 0).

Note that a digraph with vertices $\{x_i, x_j\}$ and edges $\{(x_i, x_j), (x_j, x_i)\}$ implies an impasse. The system switches infinitely fast between the two states. Time does not even advance in the ideal case. While the previous discussion showed a way out of this, but not without a certain ambiguity, we shall preclude such cases. The digraph with vertices $\{x_i, x_j, x_k\}$ and edges $\{(x_i, x_j), (x_i, x_k)\}$ with $x_j \neq x'_k$ both terminal, implies nonuniqueness of the solution. We call the auto-hybrid system system well-posed if there are no deadlocks and solutions are uniquely defined. In contrast, the digraph with vertices $\{x_i, x'_i x_j, x'_j\}$ and edges $\{(x_i, x_j), (x'_j, x'_i)\}$ with $x_i \neq x'_i$ and $x_j \neq x'_j$. poses no problem as long as x_i and x'_i are separated and x_j and x'_j are too. Separation of x_i and x'_i implies that a transit from x_i to x'_i must involve a nonzero lapse of time. The digraph with vertices $\{x_i, x'_i, x_j\}$ and edges $\{(x_i, x_j), (x_j, x'_i)\}$ with $x_i \neq x'_i$, is not pseudo-continuous, as an instantaneous transition from x_i to x'_i without any other control would exist.

4.2.2 Sequential switching

The digraph associated with vertices $\{x_i, x_j, x_k\}$ and edges $\{(x_i, x_j), (x_j, x_k)\}$ means $\{h_{ij}(x_i) = 0, x_j = S_{ji}x_i\}$ together with $\{h_{kj}(x_j) = 0, S_{kj}(x_j) = x_k\}$, which implies $x_k = S_{kj}S_{ji}x_i$. In turn this implies the edge (x_i, x_k). and thus a reset to mode (k), with trigger condition the combined $\{h_{ji}(x_i) = 0, h_{kj}(S_{ji}x_i) = 0\} \equiv h_{ki}(x_i) = 0$. Since a reset from x_i to mode (k) must already be accounted for in the complete description, pseudo-continuity requires that But $S_{ki}x_i = S_{kj}S_{ji}x_i$. Unlike the exo-hybrid system case, this does not imply that $S_{ki} = S_{kj}S_{ji}$. It needs only to hold on the set $\{x_i \in \Sigma_{ji} \,|\, S_{ji}x_i \in \Sigma_{kj}\}$. The description is then completed with the given edge (x_j, x_k). More generally, the sequential structure represented by the digraph with vertices $V = \{x_i, x_{k_1}, x_{k_2}, \ldots, x_{k_m}, x_j\}$ and edges $E = \{(x_i, x_{k_1}), \ldots, (x_{k_{m-1}}, x_{k_m}), (x_{k_m}, x_j)\}$ is reducible to the digraph as shown in Figure 4 on the next page, with the same vertex set, V, and edges $E_r = \{(x_i, x_j), \ldots, (x_{k_{m-1}}, x_j), (x_{k_m}, x_j)\}$.

While two edges with the same initial vertex but different final vertices implies nonuniqueness if the final vertices are terminal, the case of distinct paths $(x_i, x_{i_1}), \ldots,$ (x_{i_m}, x_j) and $(x_i, x_{j_1}, \ldots, x_{j_n}, x_j)$ is allowed. In view of the previous reduction, it is equivalent to the digraph with the same vertices, and edges $\{(x_i, x_j), (x_{i_1}, x_j), \ldots, (x_{i_m}, x_j), (x_{j_1}, x_j), \ldots, (x_{j_n}, x_j)\}$. This can be lifted to the sets Σ_{ji}. If $\Sigma_{ji} \cap \Sigma_{ki} \neq \varnothing$, then Σ_{ji}

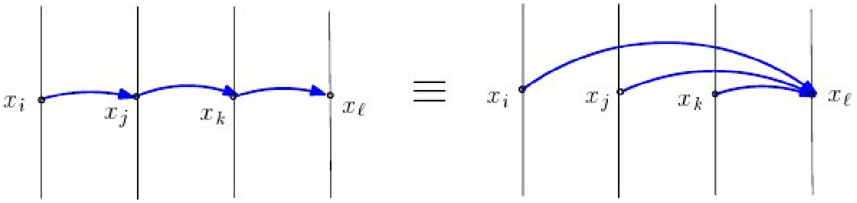


Figure 4: Reduction of the sequential switching digraph.

and Σ_{ki} can be partitioned (superscript c denoting complementation) in $\Sigma'_{ji} = \Sigma_{ji} \smallsetminus \Sigma^c_{ki}$, $\Sigma_{ji} \cap \Sigma_{ki}$, and $\Sigma'_{ki} = \Sigma_{ki} \smallsetminus \Sigma^c_{ji}$. Well-posedness (uniqueness) requires then the existence of a set of modes (α), none equal to $(i),(j)$,or (k), such that $\cup_\alpha \Sigma_{\ell_\alpha i} = \Sigma_{ji} \cap \Sigma_{ki}$ with paths with digraph containing the partial graphs $\{(\Sigma'_{ji}, S_{ji}\Sigma'_{ji}), (\Sigma'_{ki}, S_{ki}\Sigma'_{ki}), \cup_\alpha(\Sigma_{\ell_\alpha,i}, S_{\ell_\alpha i}\Sigma_{\ell_\alpha,i})\}$.

Assuming that the smooth dynamics in each mode is completely reachable, the reachability properties of the auto-hybrid system are determined by the collection of sets Σ_{ji} (nullsets of the maps h_{ji}). In particular Σ_{ji} may be a proper union of connected components, and have components of different dimension. Obviously, partial states in Σ_i are themselves not reachable by the smooth dynamics in the mode (as that cannot be maintained). Let $\Sigma^{(k)}_{ji}$ be a connected component of Σ_{ji}, then this set is separating if $\mathbf{X}^{(i)} \smallsetminus \Sigma^{(k)}_{ji}$ is not connected. Obviously, in this case the separated parts of $\mathbf{X}^{(i)}$ cannot be reached from each other by the dynamics in the mode (i) only. Reachability implies a path, necessarily involving smooth dynamics, passing through different modes. If no smooth dynamics were involved, pseudo-continuity would be violated. It may be advantageous to consider each separated components in mode (i) as a set of individual modes. Hence the stalk $\mathcal{F}(\{i\})$ degenerates into the multistalk $\cup_\alpha \mathcal{F}(i_\alpha)$, with each $\mathcal{F}(i_\alpha) \smallsetminus \Sigma_i$ connected.

5 Impact systems

Consider the simple example of the elastic collision of two point masses. let's assume that the masses are equal. Then the total momentum before collision at time t,

$$p = m\dot{r}_1(t_-) + m\dot{r}_2(t_-)$$

is conserved, as well as the total energy

$$E = \frac{m}{2}\|\dot{r}_1(t_-)\|^2 + \frac{m}{2}\|\dot{r}_2(t_-)\|^2$$

and thus after collision we have similar expressions, with t_- replaced by t_+. There are only two solutions:

$$\begin{aligned}\dot{r}_1(t_+) &= \dot{r}_1(t_-)\\ \dot{r}_2(t_+) &= \dot{r}_2(t_-),\end{aligned}$$

and

$$\begin{aligned}\dot{r}_1(t_+) &= \dot{r}_2(t_-)\\ \dot{r}_2(t_+) &= \dot{r}_1(t_-),\end{aligned}$$

The first is physically impossible as the two particles should pass through each other. This leaves only the second possibility, which means that the particles have exchanged their momentum.
This example illustrates a natural mode switching system, where the two modes have identical dynamics, but the state of the entire system jumps as state variables between the subsystems are exchanged.

This leads us to pose a general problem for a composite system where the individual parts all have dynamics of the same form, say

$$\dot{x}_i = f(x_i;\theta_i), \quad i = 1,\ldots,N \tag{5}$$

where $x \in X$, the state space, and θ_i is a parameter vector characterizing the i-th subsystem. In the above example, θ_i would be the mass of the i-th particle.
Let us assume that the trigger for the switching event between subsystems i and j is given by the submanifold

$$h(x_i) = h(x_j), \tag{6}$$

with $h : X \to \mathbb{R}^r$, the *trigger-function*, and that the effect of the switch (occurring at time t) is the exchange

$$\left[\begin{array}{c} x_i(t_+)\\ x_j(t_+)\end{array}\right] = \left[\begin{array}{cc} S_1(\theta_i,\theta_j) & S_2(\theta_i,\theta_j)\\ S_2(\theta_j,\theta_i) & S_1(\theta_j,\theta_i)\end{array}\right]\left[\begin{array}{c} x_i(t_-)\\ x_j(t_-)\end{array}\right] \tag{7}$$

for some *interaction matrices* $S_1(\theta_i,\theta_j)$ and $S_2(\theta_i,\theta_j)$. We assume that the states x_k, $(k \notin \{i,j\})$ of the noninteracting subsystems remain the same during the ij interaction. That way, it suffices to consider only systems consisting of two interacting subsystems. We shall further assume that the trigger condition remains satisfied immediately after the interaction, thus perhaps enabling multiple sequential interactions.

$$h(x_i^-) = h(x_j^-) \Rightarrow h(x_i^+) = h(x_j^+). \tag{8}$$

For the example at the introduction, each particle is characterized by its mass ($\theta_i = m_i$), and elementary physics tells us that for the one dimensional motion, the elastic collision equations give $S_1(\theta_i,\theta_j) = \frac{m_i - m_j}{m_i + m_j}$, while $S_2(\theta_i,\theta_j) = \frac{2m_j}{m_i + m_j}$. We note that while S_1 relates the velocity of the *same* particle before and after collision, it still depends on the parameters of *both* interacting particles. The trigger condition is $x_i = x_j$.
To simplify notation, the parameter arguments will be omitted, and we introduce the *exchange operator* (resulting in a permutation of the arguments) σ for $(\theta_i,\theta_j) \to (\theta_j,\theta_i)$. This implies a product rule for functions of the parameters: $(fg)\sigma =$

$(f\sigma)(g\sigma)$. We place σ on the right of f as it acts not on f but rather its arguments. Thus the matrix S in (7) is

$$S = \begin{bmatrix} S_1 & S_2 \\ S_2\sigma & S_1\sigma \end{bmatrix},$$

and note that $S\sigma = JSJ$, where

$$J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}.$$

Analogous to the momentum and energy equations, assume that the dynamics is such that for each subsystem the vector $V(x_i(t);\theta_i)$ is constant along the noninteracting trajectories, and that these vectors are additive in the sense that during every interaction $\sum_{i=1}^N V(x_i(t);\theta_i)$ is conserved. This requirement restricts the interaction matrices S_1 and S_2. Thus

$$\begin{aligned} &V(x_i,\theta_i) + V(x_j,\theta_j) \\ &\quad = V(S_1(\theta_i,\theta_j)x_i + S_2(\theta_i,\theta_j)x_j,\theta_i) + V(S_2(\theta_j,\theta_i)x_i + S_1(\theta_j,\theta_i)x_j,\theta_j), \end{aligned} \tag{9}$$

Finally, since $V(x,\theta)$ represents constants of the (autonomous) motion, we have

$$\frac{\partial V}{\partial x}(x;\theta)f(x;\theta) = 0. \tag{10}$$

5.1 Linear-quadratic invariants

Let us now turn to a system mode with linear dynamics, $\dot{x} = Ax$ in the autonomous case ($u \equiv 0$). Under some conditions the system will possess linear and quadratic invariants, by which we mean that for some matrices $P \in \mathbb{R}^{p\times n}$ for some $p < n$ and $Q = Q^\top \in \mathbb{R}^{n\times n}$ the linear and quadratic forms $L(x) = P^\top x$ and $K(x) = x^\top Q x$ will be constants of the motion.

Invariance of the dynamics (10) imposes for a linear mode, $\dot{x} = Ax$, the restrictions $P^T A = 0$, and $A^T Q + QA = 0$. Clearly, If A has full rank, no linear invariants of the motion can exist, and if no two eigenvalues of A add to zero, no quadratic invariant can exist. For the existence of a quadratic invariant, it is not necessary that A is singular.

The following is an obvious result in invariant subspaces:

Theorem 4. *The n-dimensional system mode $\dot{x} = Ax$ possesses p independent linear invariants of the motion if* $\operatorname{rank} A = n - p$.

To get a characterization of existence of quadratic invariants, we need to find all solutions $Q = Q^\top$ of $A^\top Q + QA = 0$, and note that these need not be definite. The first step towards the solution follows from the existence of a similarity $T \in G\ell_n(\mathbb{C})$ such that TAT^{-1} has a block-diagonal form with Jordan blocks $J_{\alpha_i}(\lambda_i) = \lambda_i I_{\alpha_i} + J_{\alpha_i}(0)$. Here α_i is the size of the block, and the $\lambda_i \in \operatorname{Spec}(A)$ are the eigenvalues (possibly repeated) of A. The matrix $J_m = J_m(0)$ has ones at its $(i,i+1)$ entries for $i = 1,\ldots,m-1$ and zeros elsewhere. The effect of the similarity transformation is

$$J^*P + PJ = 0$$

where $J = TAT^{-1}$ and $P = P^* = T^{-*}QT^{-1}$. Let

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1m} \\ \vdots & & \vdots \\ P_{m1} & \cdots & P_{mm} \end{bmatrix}$$

consistently with the partitioning of the Jordan blocks. Noting that $P^*_{ii} = P_{ii}$ and $P^*_{ij} = P_{ji}$, we get for all (i,j)

$$J^*_{\alpha_i}(\lambda_i)P_{ij} + P_{ij}J_{\alpha_j}(\lambda_j). \tag{11}$$

The equation for the unknown blocks are effectively decoupled. Using Kronecker product properties and the vectorizing operator, vec, which represents a matrix by its columns stacked on top of each other, this becomes

$$\left[J^*_{\alpha_i}(\lambda_i) \otimes I_{n_j} + I_{n_i} \otimes J^{\top}_{\alpha_j}(\lambda_j)\right] \mathrm{vec}(P_{ij}) = 0.$$

Finally, $J^*_{\alpha_i}(\lambda_i) = J^{\top}_{\alpha_i}(\overline{\lambda}_i)$ yields

$$\left[J^{\top}_{\alpha_i}(\overline{\lambda}_i) \otimes I_{n_j} + I_{n_i} \otimes J^{\top}_{\alpha_j}(\lambda_j)\right] \mathrm{vec}\,(P_{ij}) = 0.$$

Using the decomposition of the Jordan block, this gives

$$(\overline{\lambda}_i + \lambda_j)I_{\alpha_i\alpha_j} + \left[J^{\top}_{\alpha_i} \otimes I_{n_j} + I_{n_i} \otimes J^{\top}_{\alpha_j}\right] \mathrm{vec}\,(P_{ij}) = 0. \tag{12}$$

The matrix in the square brackets has all its eigenvalues equal to zero. Hence if A possesses eigenvalues such that $\lambda_i + \lambda_j = 0$, then quadratic invariants will exists. If no such pair of eigenvalues exist, no quadratic invariants can exist. Let us now study $M_{\alpha_i,\alpha_j} = \left[J^{\top}_{\alpha_i} \otimes I_{n_j} + I_{n_i} \otimes J^{\top}_{\alpha_j}\right]$ in more detail. Without loss of generality let $\alpha_i \geq \alpha_j$ (use transposition in the other case). Partition the matrix in α_i^2 blocks of size $\alpha_j \times \alpha_j$, and each block has Jordan structure. Moreover all entries on the α_j-th parallel above the diagonal contains ones, as in the illustrated in the (3,2) case below.

$$\left[\begin{array}{cc|cc|cc} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

By performing alternating block row and block column operations (use block col 1 to reduce block col 2, reduce block row $\alpha_j - 1$ with block row α_j). This turns the second column into a zero column. Now repeat the procedure on the inner submatrix containing $(\alpha_j - 2)^2$ blocks. Each step in this reduction produces an additional zero column. There are two ways in which the procedure terminates depending on whether α_j is even or odd. In each case only zero columns and nonidentical columns of the identity matrix result.

If α_j is odd, say $2\nu + 1$, then ν zero columns are created, and noting that the first column was already a zero column, it follows that the big matrix has nullity $\nu + 1$. Likewise, when α_j is even, 2ν say, then it is seen that only $\nu - 1$ zero columns are created and the nullity is ν. Combining, we get a concise formula ($n(M)$ is the nullity of M)

$$n\left(M_{\alpha_i,\alpha_j}\right) = \left\lceil \frac{\alpha_j}{2} \right\rceil + 1.$$

This counts the number of linear independent solutions (or the number of degrees of freedom in P_{ij}). However, when $i = j$ there is an additional constraint: The solution matrix P_{ii} needs to be Hermitian. For this case, the number of degrees of freedom of $M(\alpha_i, \alpha_i)$ reduces to

$$\left\lfloor \frac{\alpha_i + 1}{2} \right\rfloor.$$

This is perhaps easiest to see directly from Equation (11), which involves only a single Jordan block.

Combining, we see that the number of degrees of freedom for $P = P^*$ for the solution to

$$J^* P + PJ = 0$$

where J has all zero eigenvalues, depends strongly on the Jordan structure of J, and equals

$$\sum_i \left\lfloor \frac{\alpha_i + 1}{2} \right\rfloor + \sum_{i<j} \left(\left\lceil \frac{\min(\alpha_i, \alpha_j)}{2} \right\rceil + 1 \right). \tag{13}$$

For instance, if $J = J_4 \oplus J_3$ one finds $\lfloor (4+1)/2 \rfloor + \lfloor (3+1)/2 \rfloor + (\lceil 3/2 \rceil + 1) = 2 + 2 + 3 = 7$ degrees of freedom. On the other hand,$J_3 \oplus J_2 \oplus J_2$ gives 10, and $J_3 \oplus J_2$ gives 5 degrees of freedom. The structure of the matrix of the quadratic form follows directly from the Kronecker structure of the blocks $M(\alpha_1, \alpha_j)$. The index of the zero columns resulting from the reduction determine the “free" elements in vecP_{ij}. Since the other columns are linearly independent, all with a single nonzero element (which is 1), the corresponding solution for the component in vecP_{ij} is zero. The degrees of freedom are reduced for the diagonal blocks since they must be Hermitian.

Theorem 5. *If all eigenvalues of A are at the origin, then the matrix Q possesses $d(A)$ degrees of freedom, where*

$$d(A) = \frac{\alpha(\alpha - 1)}{2} + \sum_{i=1}^{\alpha} \left(\left\lfloor \frac{\alpha_i + 1}{2} \right\rfloor + (\alpha - i) \left\lceil \frac{\alpha_i}{2} \right\rceil \right), \tag{14}$$

and α is the number of Jordan blocks defining the structure of A, and α_i is the size of the i-th block (thus, $\sum_{i=1}^{\alpha} \alpha_i = n$.)

Proof. As discussed above, a similarity can be found (complex in general) that reduces A to $J = \text{Blockdiag}\,(J_{\alpha_i}(0); i = 1\ldots,\alpha)$. We may also assume that the blocks are ordered by size, $\alpha_i \leq \alpha_j$ if $i < j$. Then Equation (13) yields

$$\begin{aligned} d(J) &= \sum_{i<j} 1 + \sum_i \left\lfloor \frac{\alpha_i + 1}{2} \right\rfloor + \sum_i \sum_{j=i+1}^{\alpha} \left(\left\lceil \frac{\alpha_i}{2} \right\rceil + 1 \right) \\ &= \frac{\alpha(\alpha-1)}{2} + \sum_{i=1}^{\alpha} \left(\left\lfloor \frac{\alpha_i + 1}{2} \right\rfloor + \sum_{j=i+1}^{\alpha} \left(\left\lceil \frac{\alpha_i}{2} \right\rceil + 1 \right) \right) \\ &= \frac{\alpha(\alpha-1)}{2} + \sum_{i=1}^{\alpha} \left(\left\lfloor \frac{\alpha_i + 1}{2} \right\rfloor + (\alpha - i) \left(\left\lceil \frac{\alpha_i}{2} \right\rceil + 1 \right) \right) \end{aligned}$$

The number of degrees of freedom is not changed by similarity, hence (14) follows. □

We can now state the general result for the maximal invariant (invariant with the largest number of free parameters)

Theorem 6. *The system $\dot{x} = Ax$ has a maximal quadratic invariant with $d(A)$ degrees of freedom given by*

$$d(A) = \sum_{\{i|\lambda_i=0\}} \left\lfloor \frac{\alpha_i + 1}{2} \right\rfloor + \sum_{\{i<j|\overline{\lambda}_i+\lambda_j=0\}} \left(\left\lceil \frac{\min(\alpha_i,\alpha_j)}{2} \right\rceil + 1 \right). \tag{15}$$

Proof. This follows directly in view of Equation 12. The terms (diagonal and off diagonal) contribute to the degrees of freedom only if the diagonal block has a eigenvalue 0, and the off diagonal (ij) block only contributes if $\overline{\lambda}_i + \lambda_j = 0$. □

Because of the linear character of the solution, Q, of $A^\top Q + QA = 0$, we can write $Q = \sum_{i=1}^{d(A)} q_1 S_i$, where the $q_i, i = 1,\ldots,d(A)$ are the free parameters, and the S_i are the *structure matrices*. The quadratic forms $V_i(x) = \frac{1}{2}x^\top S_i x$ are the elementary quadratic invariants. Note also that if $L(x)$ is a linear invariant, then $L(x)^2$ is obviously a quadratic invariant, so that the elementary quadratic invariants and linear invariants are not necessarily independent.

Example 7. Let

$$A = \left[\begin{array}{cc|cc} \sigma & -\omega & & \\ \omega & \sigma & & \\ \hline & & -\sigma & \omega \\ & & -\omega & -\sigma \end{array}\right]$$

The general solution is

$$Q = \left[\begin{array}{cc|cc} & & q_2 & q_1 \\ & & q_1 & -q_2 \\ \hline q_2 & q_1 & & \\ q_1 & -q_2 & & \end{array}\right]$$

The elementary quadratic invariants are

$$\begin{aligned} V_1(x) &= x_1x_4 + x_2x_4 \\ V_2(x) &= x_1x_3 - x_2x_4. \end{aligned}$$

Theorem 6 implies that A may be brought by similarity (not uniquely) to a real block triangular form

$$A \to \begin{bmatrix} H & 0 \\ C & N \end{bmatrix},$$

where the decoupled block, H, corresponds to a Hamiltonian system with modes satisfying $\overline{\lambda}_i = \lambda_j$, i.e., the condition for existence of quadratic invariants. In contrast, the block N contains the remaining modes The block C denotes the coupling between these subsystems. Let the state vector in this realization be partitioned into $x^\top = [x_H^\top, x_N^\top]$. The dimension of x_H is necessarily even. Then the block H may be interpreted as the solution of an LQ problem.

Example 8. Let the H-block be

$$H = \begin{bmatrix} a & -b^2/q \\ -p & -a \end{bmatrix}.$$

Then it is readily seen that $H^\top P + PH = 0$ is solved by

$$P = \begin{bmatrix} p & a \\ a & -b^2/q \end{bmatrix}.$$

However, if we consider the LQ-problem for the system $\dot{x} = ax + bu$, with performance index $\frac{1}{2}\int (px^2 + qu^2)\,dt$, then the stationarity condition is $u = -b\lambda/q$ where λ is the costate, satisfying the Euler-Lagrange equation $\dot{\lambda} = -px - a\lambda$. If we let $x_H^\top = [x, \lambda]$, and substitue the solution for u, then the resulting equations are precisely the above Hamiltonian subsystem. The quadratic invariant is

$$\frac{1}{2}[x,\lambda]\begin{bmatrix} p & a \\ a & -b^2/q \end{bmatrix}\begin{bmatrix} x \\ \lambda \end{bmatrix} = \frac{1}{2}(px^2 - \frac{b^2}{q}\lambda^2 + 2a\lambda x) = \frac{1}{2}(px^2 - qu^2 - \frac{2qa}{b}ux).$$

This is precisely the stationary value of the Hamiltonian $\frac{1}{2}(px^2 + qu^2) + \lambda(ax + bu)$ of the optimally controlled system. Indeed since the dynamics and the cost rate do not depend explicitly on time, it is well known that the Hamiltonian is a constant of the motion.

5.2 Exchange operator

In what follows, we shall assume that we have an aggregate of $N \geq 2$ subsystems, all of the same linear form. Let A have a nontrivial null space, so that p linear and a quadratic constant of the motion exist. Let a mode of the full system consist of $N \geq 2$ copies of the subsystem $\dot{x} = Ax$. The state of the i-th copy will now be denoted by x_i.

Let the trigger condition be the linear form $h^\top x_i = h^\top x_j$, at which point the system changes to another mode, but again consisting of the same number of subsystems. It is easily shown that linear and quadratic invariants cannot coexist if the number of subsystems changes from one mode to the other

Let thus $V(x,\theta)$ consists of the p linear forms $P^\top(\theta)x$ and of a quadratic form $x^\top Q(\theta)x$, think momenta and energy. We will assume that $Q(\theta)$ is symmetric and positive semi-definite.

At the interaction we must have

$$\begin{aligned} V_{lin} &= [P^\top, P^\top\sigma]\begin{bmatrix} S_1 & S_2 \\ S_2\sigma & S_1\sigma \end{bmatrix}\begin{bmatrix} x_i \\ x_j \end{bmatrix} = [P^\top, P^\top\sigma]\begin{bmatrix} x_i \\ x_j \end{bmatrix}, \\ V_{quad} &= \left[x_i^\top, x_j^\top\right]\begin{bmatrix} S_1^\top & S_2^\top\sigma \\ S_2^\top & S_1^\top\sigma \end{bmatrix}\begin{bmatrix} Q & \\ & Q\sigma \end{bmatrix}\begin{bmatrix} S_1 & S_2 \\ S_2\sigma & S_1\sigma \end{bmatrix}\begin{bmatrix} x_i \\ x_j \end{bmatrix} \\ &= \left[x_i^\top, x_j^\top\right]\begin{bmatrix} Q & \\ & Q\sigma \end{bmatrix}\begin{bmatrix} x_i \\ x_j \end{bmatrix}, \end{aligned}$$

for all $(x_i, x_j) \in \mathcal{M}$, where $\mathcal{M} = \{(x_i, x_j) \in X^2 \,|\, h(x_i) = h(x_j)\}$. These equations imply

$$P^\top S_1^\top + (P^\top S_2^\top)\sigma = P^\top \tag{16}$$

$$S_1^\top Q S_1 + (S_2^\top Q S_2)\sigma = Q \tag{17}$$

$$S_1^\top Q S_2 + (S_2^\top Q S_1)\sigma = 0. \tag{18}$$

We shall first solve the general (unstructured) problem, thus setting $\overline{P}, \overline{Q}$, and T for

$$\begin{bmatrix} P \\ P\sigma \end{bmatrix}, \begin{bmatrix} Q & \\ & Q\sigma \end{bmatrix} \text{and} \begin{bmatrix} S_1 & S_2 \\ S_2\sigma & S_1\sigma \end{bmatrix},$$

respectively.

Problem 1: Determine $\mathcal{R}_{P,Q} = \{T \in G\ell_{2n}(\mathbb{R}) \,|\, \overline{P}^\top T = \overline{P}^\top, T^\top \overline{Q} T = \overline{Q}\}$.

Solution: We first solve the problem of determining $\mathcal{C}_Q = \{T \in G\ell_{2n}(\mathbb{R}) \,|\, T^\top \overline{Q} T = \overline{Q}\}$. Let the eigen decomposition of $\overline{Q}$ be $U^\top \Lambda U$, where $\Lambda \geq 0$ diagonal, and $U \in O_{2n}(\mathbb{R})$. Then

$$\begin{aligned} \mathcal{C}_{\overline{Q}} &= \{T \in G\ell_{2n}(\mathbb{R}) \,|\, T^\top U^\top \Lambda U T = U^\top \Lambda U\} \\ &= \{T \in G\ell_{2n}(\mathbb{R}) \,|\, U T^\top U^\top \Lambda U T U^\top = \Lambda\} \\ &= U \mathcal{C}_\Lambda U^\top. \end{aligned}$$

The problem is now reduced to a simpler one. Let $\Lambda^{1/2} \geq 0$ be the diagonal square root of Λ.

$$\begin{aligned} \mathcal{C}_\Lambda &= \{T \in G\ell_{2n}(\mathbb{R}) \,|\, T^\top \Lambda T = \Lambda\} \\ &= \{T \in G\ell_{2n}(\mathbb{R}) \,|\, T^\top \Lambda^{1/2} = \Lambda^{1/2} W,\, W \in O_{2n}(\mathbb{R})\} \\ &= \Lambda^{-1/2} O_{2n}(\mathbb{R}) \Lambda^{1/2}. \end{aligned}$$

Then

$$\begin{aligned}\mathcal{R}_{\overline{P},\overline{Q}} &= \{T \in U\Lambda^{-1/2}O_{2n}(\mathbb{R})\Lambda^{1/2}U^\top \mid T^\top\overline{P} = \overline{P}\}\\ &= U\Lambda^{-1/2}\{W \in O_{2n}(\mathbb{R}) \mid U\Lambda^{1/2}W\Lambda^{-1/2}U\overline{P} = \overline{P}\}\Lambda^{1/2}U^\top\\ &= U\Lambda^{-1/2}\{W \in O_{2n}(\mathbb{R}) \mid W\Lambda^{-1/2}U\overline{P} = \Lambda^{-1/2}U\overline{P}\}\Lambda^{1/2}U^\top\end{aligned}$$

At this point we go back to the specific symmetry in the matrices $\overline{P}$ and $\overline{Q}$. Careful consideration of the above preliminary transformations shows that generically both U and Λ are blockdiagonal, where the lower block is the permutation symmetric (σ) of the upper block. This implies that $\widehat{P} = \Lambda^{-1/2}U\overline{P}$ also has the same symmetric structure, $\widehat{P}^\top = [\hat{P}^\top, \hat{P}^\top\sigma]$. Let $H^\top KG$ be a singular value decomposition (SVD) of $\hat{P}$, i.e., $H \in O_n$, $G \in O_p$ and K a $n \times p$ diagonal matrix (full rank). then $\{W \in O_{2n}(\mathbb{R}) \mid W\widehat{P} = \widehat{P}\}$ consists of matrices of the form

$$\left[\begin{array}{c|c} H & \\ \hline & H\sigma \end{array}\right]\left[\begin{array}{cc|cc} \alpha I_p & & \beta I_p & \\ & \hat{W}_{11} & & \hat{W}_{12} \\ \hline \gamma I_p & & \delta I_p & \\ & \hat{W}_{21} & & \hat{W}_{22} \end{array}\right]\left[\begin{array}{c|c} H^\top & \\ \hline & H^\top\sigma \end{array}\right].$$

where the submatrix

$$\begin{bmatrix} \hat{W}_{11} & \hat{W}_{12} \\ \hat{W}_{21} & \hat{W}_{22} \end{bmatrix} \in O_{2(n-p)}(\mathbb{R}), \text{ and } \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \in O_2(\mathbb{R}).$$

This leaves $(n-p)(2(n-p)+1)+1$ degrees of freedom (Euler angles). Canonical forms are obtained by a nice selection, $|\alpha| = |\beta| = |\gamma| = |\delta|$ forming an improper rotation and of the submatrix $\hat{W}$, for instance enforcing $\hat{W}_{22} = \hat{W}_{11}\sigma$, and $\hat{W}_{21} = \hat{W}_{12}\sigma$. If, in addition, the transformation S is such that two successive transformations return the original state, then $S^2 = I$, and in partitioned form this leads to

$$S_1^2 + S_2(S_2\sigma) = I \tag{19}$$

$$S_1S_2 + S_2(S_1\sigma) = 0. \tag{20}$$

5.3 Insertion of intermediate modes

Reconsider now the collision of two freely moving masses in one dimension. Let the masses be m_1 and m_2. Conservation of momentum and kinetic energy dictates that right after the impact the momenta will be ($i \neq j$)

$$p_i' = \frac{m_i - m_j}{m_i + m_j}p_i + \frac{2m_j}{m_i + m_j}p_j \overset{\text{def}}{=} S_1(m_i, m_j)p_i + S_2(m_i, m_j)p_j.$$

In terms of the state variables $[x_o, p_o, \omega, \delta]$, where x_o is the center of mass, p_o the total momentum of the masses, $\delta = x_1 - x_2$, and $\omega = p_1/m_1 - p_2/m_2$, then the dynamics of the 4-dimensional (two particle) system is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} x_o \\ p_o \\ \omega \\ \delta \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{(m_1+m_2)} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} x_o \\ p_o \\ \omega \\ \delta \end{bmatrix}. \tag{21}$$

The individual spatial coordinates are read out from this state as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -\frac{m_2}{m_1+m_2} \\ 1 & 0 & 0 & \frac{m_1}{m_1+m_2} \end{bmatrix}. \tag{22}$$

The trigger is $\delta = 0$ at which instant the 4-dimensional state maps to the 3-dimensional mode with state $[x_o, p_o, \omega]^\top$ just before the collision. The transition to the post collision gives $[x_o, p_o, -\omega]^\top$, and from there back to the 4-state $[x_o, p_o, \omega, \delta = 0]^\top$. The state therefore does not return to the original one before impact, and pseudo-continuity does not hold. But it does not need to, if the post collision mode is considered as a new mode, different from the pre-collision one. But another option exists: If we let the post-collision 3-dimensional mode correspond to the system with the identities of the subsystems (including their masses) *permuted* (i.e., a symmetry involving the exchange operator) $\sigma : (m_1, m_2) \to (m_2, m_1)$, then the full transition from the 4-system before and after collision is given by

$$\begin{bmatrix} I_3 & 0 \\ 0 & 1 \end{bmatrix}.$$

The (4,4) entry is actually immaterial because of the trigger condition. Moreover the dynamics in this mode is the same as in the original 4-th order mode. Expressed another way: If the total momentum is P and the total kinetic energy V, with $P^2 < 2(m_1 + m_2)V$, then there are exactly two solutions to the momenta of the individual particles (Intersection of the line $p_1 + p_2 = P$ with the ellipse $\frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} = V$.) One of these two solutions (say p_i is associated with mass m_i) corresponds to the momentum before the collision, the other (p_i') with the momentum of the same mass after the collision. However, if at the collision the masses are permuted, then the system labeled “1" has mass m_2 and therefore its momentum $p_1'' = p_2'$. See Figure 5. The use of the exchange operator in multi-particle quantum theory is standard. Unlike classical mechanics where a particle may be identified by its trajectory, this is not possible in quantum theory. Particles have no identifiable attributes once they are represented by overlapping wave functions. The solid line corresponds to subsystem labeled “1": No mass permutation on the left, permutation on the right.

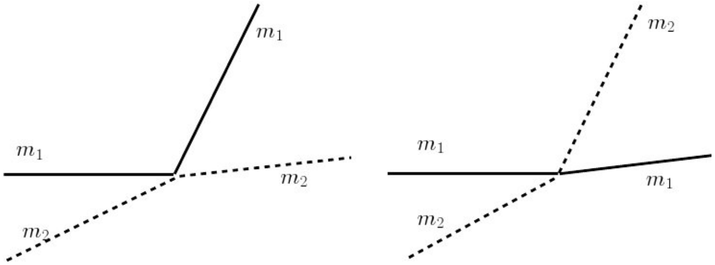


Figure 5: Two viewpoints: Left: the usual; right: permutation of masses.

6 Conclusions

We made an excursion into some aspects of hybrid systems, both for externally and autonomously controlled switching. But we were primarily interested in doing so for the case where the modes of the system may have different dimensions. The notion of pseudo-continuity was introduced as a means to avoid the nonphysical situation where the state could be altered simply by fast switching through a mode cycle. Critique was given that an interesting class of systems, the one modeling collisions, does not satisfy this property. Here we have shown that pseudo-continuity may be restored in several ways: First we can use a degenerate representation, where one mode splits into several modes. For instance two consecutive impacts for systems moving in a ring, indeed return to the original state. Alternatively, the exchange operator, albeit somewhat far-fetched in the classical case, may be introduced to restore the pseudo-continuity. But perhaps more interestingly, we have given what we believe to be a reasonable definition of what the concept of state space (and trajectory) for such a multi-mode multi-dimensional system may be: a sheaf. We have also shown some results (proved elsewhere) on the canonical structure of externally switched systems (exo-hybrid systems), and looked at similar aspects for the autonomous switching systems, with switches triggered by the partial states in the modes. We also characterized the quadratic invariants a system may have.

Bibliography

[1] M. S. Branicky, V. S. Borkar, and S. Mitter. A unified framework for hybrid control theory: Model and optimal control theory. *IEEE Transactions on Automatic Control*, 43(1):31–45, 1998. Cited p. 449.

[2] R. W. Brockett. Hybrid models for motion description control systems. In H. L. Trentelman and J. C. Willems, editors, *Essays on Control: Perspectives in the Theory and its Applications*. Birkhäuser, 1993. Cited p. 449.

[3] U. Helmke and E. I. Verriest. Structure and parametrization of periodic systems. *Mathematics of Control, Signals and Systems*, 23(1–3):67–99, 2011. Cited p. 450.

[4] M. Petreczky and J. H. van Schuppen. Realization theory for hybrid systems. *IEEE Transactions on Automatic Control*, 55(10):2282–2297, 2010. Cited p. 449.

[5] J. W. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory. A Behavioral Approch.* Springer, 1998. Cited p. 449.

[6] A. J. van der Schaft and J. M. Schumacher. *An Introduction to Hybrid Dynamical Systems*. Springer, 2000. Cited pp. 449 and 454.

[7] E. I. Verriest. Multi-mode multi-dimensional systems. In *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems*, pages 1268–1274, 2006. Cited pp. 449 and 450.

[8] E. I. Verriest. Multi-mode multi-dimensional systems with poissonian sequencing. *Communications in Information and Systems*, 9(1):77–102, 2009. Cited pp. 449 and 450.

[9] E. I. Verriest. Multi-mode multi-dimensional systems. In *Proceedings of the 49th IEEE Conference on Decision and Control*, pages 7021–7026, 2010. Cited p. 450.

[10] E. I. Verriest. Pseudo-continuous multi-dimensional multi-mode systems: Behavior, structure and optimimal control. *Discrete Event Dynamical Systems*, 22(1):27–59, 2012. Cited pp. 449, 450, 451, and 452.

[11] R. O. Wells. *Differential Analysis on Complex Manifolds*. Springer, 1980. Cited p. 450.

[12] H. S. Witsenhausen. A class of hybrid-state continuous-time dynamic systems. *IEEE Transactions on Automatic Control*, 11(2):161–167, 1966. Cited p. 449.

Legendre dualities between matrix subspace flows

Shintaro Yoshizawa
Gotemba Theoretical Science
Research, Japan
yzw2000@netscape.net

Abstract. We show that the principal subspace flow introduced by Oja and the minor subspace flow introduced by Manton, Helmke and Mareels are Legendre duals, and give an application of the duality approach that is related to MacMahon's Master Theorem.

1 Introduction

Various algorithms for principal component and principal subspace analysis have been proposed based on the ordinary differential equation (ODE) method [1, 4]. Similarly, minor component and minor subspace analysis have also been studied. An example of a principal component flow for a time-constant and positive definite symmetric matrix $A \in \mathbb{R}^{n\times n}$ is

$$X' = AXB - XBX^{\top}AX, \tag{1}$$

where X' denotes the derivative of $X = X(t) \in \mathbb{R}^{n\times k}$ $(k \le n)$ with respect to time t, the subscript $\top$ denotes the matrix transpose, and $B \in \mathbb{R}^{k\times k}$ is a time-constant and diagonal matrix with distinct positive eigenvalues. This flow was introduced and partially studied in [7, 8, 10]. In [11] it was shown that Equation (1) is given as a negative gradient flow

$$X' = -\mathrm{grad} F_P(X), \tag{2}$$

where the function F_P is

$$F_P(X) = -\tfrac{1}{2}\mathrm{tr}(A^2XB^2X^{\top}) + \tfrac{1}{4}\mathrm{tr}\left\{(AXBX^{\top})^2\right\} \tag{3}$$

with the Riemannian metric g

$$\langle \Omega_1, \Omega_2 \rangle_g = \mathrm{tr}(A\Omega_1 B{\Omega_2}^{\top}) \tag{4}$$

for any tangent vectors $\Omega_1, \Omega_2 \in T_X\mathbb{R}^{n\times k} \cong \mathbb{R}^{n\times k}$. See [11] for the details. An example of a minor component flow for a time-constant and symmetric matrix $C \in \mathbb{R}^{n\times n}$ is

$$Z' = -CZB + \mu Z(B - Z^{\top}Z), \tag{5}$$

where $Z = Z(t) \in \mathbb{R}^{n\times k}$, and $B \in \mathbb{R}^{k\times k}$ is a time-constant and diagonal matrix with distinct positive eigenvalues. Equation (5) was introduced and analyzed in [6] for appropriate choices of the constant $\mu \in \mathbb{R}$ and $B \in \mathbb{R}^{k\times k}$, respectively. In [6] it was also shown that the negative gradient flow of the cost function

$$F_M(Z) = \tfrac{1}{2}\mathrm{tr}(CZBZ^{\top}) + \tfrac{\mu}{4}\mathrm{tr}\left\{(B - Z^{\top}Z)^2\right\} \tag{6}$$

with respect to the Euclidean metric is given by Equation (5). Moreover, it was shown in [6] that the coordinate transformation

$$Z = \mu^{-\frac{1}{2}} A^{\frac{1}{2}} X B^{\frac{1}{2}} \quad \text{with} \quad A = \mu I - C \tag{7}$$

converts the principal component flow (1) into the minor component flow (5) for a sufficiently large μ satisfying $0 < \mu$ and $0 < \mu I - C$, where I denotes the identity matrix. In [2, Problem 3.9.], a conjecture was stated as follows:

Conjecture 1. *If a principal subspace flow is a gradient flow for a cost function f, then the corresponding dual minor subspace flow is a gradient flow for the Legendre dual cost function f^* of f.*

However, the transformation (7) is not given as the Legendre transformation. The purpose of this article is two-fold: first to show the Legendre duality between the functions (3) and (6) in the case where $B = I$; second to give an application of the duality theory that is related to MacMahon's Master Theorem in combinatory analysis.

I remember my fruitful stay at Würzburg university, where I did my postdoctoral research with Professor Uwe Helmke and his colleagues from 2000 to 2002. Throughout this stay, I learned about his philosophy and approach to optimization and dynamical systems, and I would like to take this opportunity to dedicate this article to Professor Uwe Helmke on the occasion of his 60th birthday.

2 The general concept of Legendre duality

In order to understand the duality between minor and principal component flows, this section introduces the Legendre transformation and basic duality concepts [9].

2.1 The Legendre transformation

Let V and V^* be two real vector spaces, and let $\langle\cdot,\cdot\rangle$ be a non degenerate bilinear form on the Cartesian product $V \times V^*$. Let F be a C^∞ real valued function defined on some subset of V, then the Legendre transformation is defined by

$$F^*(Z) = \{\langle X, Z\rangle - F(X) | \partial_X F(X) = Z\}, \tag{8}$$

where the derivative $\partial_X F$ denotes $\frac{\partial F}{\partial X}$ with respect to the non degenerate bilinear form, and, in general, F^* is a multivalued map from some subset of V^* to $\mathbb{R}$. If F is *convex*, then the function $X \mapsto \langle Z, X\rangle - F(X)$ is *concave*, and the gradient equation $Z = \partial_X F(X)$ means that this function attains its maximum at X. Equation (8) then becomes

$$F^*(Z) = \max_X \{\langle X, Z\rangle - F(X)\}, \tag{9}$$

where F^* is a real valued function defined on some subset of V^*. The following observation is used in this and the next sections, and so we designate it a proposition.

Proposition 2. *Let F be a C^∞ convex function, and let F^* be its Legendre transformation. Defining*

$$D(F(X),F^*(Z)) = F(X) + F^*(Z) - \langle X,Z\rangle, \tag{10}$$

then

$$0 \leq D(F(X),F^*(Z)) \tag{11}$$

for any $X \in \mathrm{Dom}(F)$ and $Z \in \mathrm{Dom}(F^*)$, where $\mathrm{Dom}(F)$ denotes the domain of the function F. Equality in (11) *holds if and only if $Z = \partial_X F(X)$.*

The quantity $D(\cdot,\cdot)$ is usually called *relative entropy*, or *Bregman divergence*, but is also sometimes known under other names.

2.2 The duality between the subspace flows

The cost functions (3) and (6) are not convex, so, in general, the gradient equation $Z = \partial_X F(X)$ can not be solved uniquely in $X \in \mathbb{R}^{n\times k}$. This means that, in general, there is no explicit expression for the dual cost function in the dual variable Z. However, changing a Riemannian metric suitably, there is a possibility to solve the gradient equation in X.

Let F_M be the cost function (6) with $B = I$. For notational convenience, we write this function also as F_M^0.

Theorem 3. *Defining the Riemannian metric g_2 by*

$$\langle \Omega_1, \Omega_2\rangle_{g_2} = \mathrm{tr}\left[\{(C-\mu I) + ZZ^\top\}\Omega_1\Omega_2^\top\right] \tag{12}$$

and the domain of F_M^0 by

$$\mathrm{Dom}(F_M^0) = \{Z \in \mathbb{R}^{n\times k} | 0 < (C-\mu I) + ZZ^\top\}, \tag{13}$$

then we have
(i) The Legendre dual function to F_M^0 is

$$(F_M^0)^*(X) = -\frac{1}{2}\mathrm{tr}(AXX^\top) + \frac{3\mu}{4}\mathrm{tr}\left\{(XX^\top)^2\right\} - \frac{\mu k}{4}, \tag{14}$$

where $A \in \mathbb{R}^{n\times k}$ is defined by $A = \mu I - C$, and

$$\mathrm{Dom}(F_M^0)^* = \{X \in \mathbb{R}^{n\times k} | 0 < XX^\top - A\}$$

with $0 < A$.
(ii) By $\hat{X} = (3\mu)^{\frac{1}{2}} A^{-\frac{1}{2}} X$, the dual function (14) *is changed to*

$$(\hat{F}_M^0)^*(\hat{X}) = -\frac{1}{2}\mathrm{tr}(\tilde{A}^2\hat{X}\hat{X}^\top) + \frac{1}{4}\mathrm{tr}\left\{(\tilde{A}\hat{X}\hat{X}^\top)^2\right\} - \frac{\mu k}{4}, \tag{15}$$

which is defined on

$$\mathrm{Dom}(\hat{F}_M^0)^* = \left\{\hat{X} \in \mathbb{R}^{n\times k} | 0 < \hat{X}\hat{X}^\top - 3\mu I\right\}. \tag{16}$$

The gradient flow of (15) *is the principal subspace flow for $\tilde{A}$;*

$$\hat{X}' = (I - \hat{X}\hat{X}^\top)\tilde{A}\hat{X}$$

with respect to the Riemannian metric g_3;

$$\langle \Omega_1, \Omega_2 \rangle_{g_3} = \mathrm{tr}(\tilde{\mathrm{A}}\Omega_1\Omega_2^\top), \tag{17}$$

where $\tilde{A} = (3\mu)^{-\frac{1}{2}}A$.

Proof. (i) The gradient of F_M^0 with respect to the Riemannian metric (12) is

$$\begin{aligned} \mathrm{grad}F_M^0(Z) &= \{(C - \mu I) + ZZ^\top\}^{-1}\{CZ - \mu Z(I - Z^\top Z)\} \\ &= Z. \end{aligned} \tag{18}$$

Since the dual cost function $(F_M^0)^*$ is given as

$$(F_M^0)^*(X) = \langle Z, X \rangle_{g_2} - F_M^0(Z), \tag{19}$$

and substituting $Z = X$ into (19), we obtain

$$(F_M^0)^*(X) = -\frac{1}{2}\mathrm{tr}(AXX^\top) + \frac{3\mu}{4}\mathrm{tr}\{(XX^\top)^2\} - \frac{\mu k}{4}.$$

(ii) By $\tilde{Z} = (3\mu)^{\frac{1}{4}}X$, the dual function (15) is changed to

$$(\tilde{F}_M^0)^*(\tilde{X}) = -\frac{1}{2}(\tilde{A}\tilde{X}\tilde{X}^\top) + \frac{1}{4}\mathrm{tr}\{(\tilde{\mathrm{X}}\tilde{\mathrm{X}}^\top)^2\} - \frac{\mu \mathrm{k}}{4}, \tag{20}$$

where $\tilde{A} = (3\mu)^{-\frac{1}{2}}A$. By $\hat{X} = \tilde{A}^{-\frac{1}{2}}\tilde{X}$, the function (20) is changed to (15). Thus, we obtain the result. □

2.3 Duality for an Oja like flow

In [5], the basic convergence properties of the flow

$$X' = AX - XX^\top X, \quad X = X(t) \in \mathbb{R}^{n \times k} \tag{21}$$

were investigated for a time-constant and symmetric matrix A. If A is positive definite, then by $\tilde{X} = A^{-\frac{1}{2}}X$, the flow (21) is equivalent to the Oja flow (Equation (1) with $B = I$) ,

$$\tilde{X}' = A\tilde{X} - \tilde{X}\tilde{X}^\top A\tilde{X}.$$

So we call (21) an Oja like flow.

Proposition 4. *For a time-constant and positive definite symmetric matrix $A \in \mathbb{R}^{n \times n}$, define*

$$\tilde{F}_P(X) = -\log\det(A - XX^\top) \tag{22}$$

with

$$\mathrm{Dom}(\tilde{\mathrm{F}}_{\mathrm{P}}) = \{\mathrm{X} \in \mathbb{R}^{\mathrm{n}\times\mathrm{k}} | 0 < \det(\mathrm{A} - \mathrm{XX}^\top)\}. \tag{23}$$

Then the negative gradient flow with respect to the Riemannian metric $\tilde{g}_1$

$$\langle \Omega_1, \Omega_2 \rangle_{\tilde{g}_1} = 2\mathrm{tr}\{(\mathrm{A} - \mathrm{XX}^\top)^{-2}\Omega_1\Omega_2^\top\} \tag{24}$$

for $\Omega_1, \Omega_2 \in T_X\mathrm{Dom}(\tilde{\mathrm{F}}_{\mathrm{P}}) \cong \mathbb{R}^{\mathrm{n}\times\mathrm{k}}$ *is the Oja like flow* (21).

Proof. The directional derivative of $\tilde{F}_P(X)$ in the direction $\Omega \in \mathbb{R}^{n\times k}$ is calculated to be $\mathcal{D}\tilde{F}_P(X)(\Omega) = 2\mathrm{tr}\{(\mathrm{A} - \mathrm{XX}^\top)^{-1}\mathrm{X}\Omega^\top)$, from which we obtain the result. □

Here, we give the dual cost function to $\tilde{F}_P$.

Proposition 5. *The Legendre dual cost function* $\tilde{F}_P^*$ *to the function* (22) *is*

$$\tilde{F}_P^*(Z) = 2\mathrm{tr}(\mathrm{ZZ}^\top) - \log\det(\mathrm{I} + \mathrm{ZZ}^\top) + \log(\det \mathrm{A}) \tag{25}$$

with respect to the Riemannian metric $\tilde{g}_2$*;*

$$\langle \Omega_1, \Omega_2 \rangle_{\tilde{g}_2} = 2\mathrm{tr}\{(\mathrm{A} - \mathrm{XX}^\top)^{-\frac{1}{2}}\Omega_1\Omega_2^\top\} \tag{26}$$

for $\Omega_i \in T_Z\mathrm{Dom}(\tilde{\mathrm{F}}_{\mathrm{P}}) \cong \mathbb{R}^{\mathrm{n}\times\mathrm{k}}$. *The domain of* $\tilde{F}_P^*$ *is* $\{Z \in \mathbb{R}^{n\times k}\}$.

Proof. The directional derivative of $\tilde{F}_P$ in the direction $\Omega \in \mathbb{R}^{n\times k}$ is calculated to be

$$\mathcal{D}F(X)\Omega = 2\mathrm{tr}\{(\mathrm{A} - \mathrm{XX}^\top)^{-1}\mathrm{X}\Omega^\top\}. \tag{27}$$

Hence, we obtain the gradient;

$$\mathrm{grad}\tilde{F}_P(X) = (A - XX^\top)^{-\frac{1}{2}}X$$

with respect to the Riemannian metric (26). Defining $Z = (A - XX^\top)^{-\frac{1}{2}}X$, we solve this in X as follows.
From

$$\begin{aligned} ZZ^\top &= (A - XX^\top)^{-\frac{1}{2}}XX^\top(A - XX^\top)^{-\frac{1}{2}} \\ &= (A - XX^\top)^{-\frac{1}{2}}A(A - XX^\top)^{-\frac{1}{2}} - I, \end{aligned}$$

we get

$$(A - XX^\top)^{-\frac{1}{2}} = A^{-\frac{1}{2}}\{A^{\frac{1}{2}}(ZZ^\top + I)A^{\frac{1}{2}}\}^{\frac{1}{2}}A^{-\frac{1}{2}}. \tag{28}$$

Substituting Equation (28) into $Z = (A - XX^\top)^{-\frac{1}{2}}X$, we have

$$X = A^{\frac{1}{2}}\{A^{\frac{1}{2}}(ZZ^\top + I)A^{\frac{1}{2}}\}^{-\frac{1}{2}}A^{\frac{1}{2}}Z. \tag{29}$$

Furthermore, substituting (28) and (29) into the equation;

$$\tilde{F}_P^*(Z) = \langle X, Z \rangle_{\tilde{g}_2} - \tilde{F}_P(X), \tag{30}$$

we have the Legendre dual cost function (25). □

The equilibrium point of the dual gradient flow of (22) is given by

$$Z' = \mathrm{grad}\tilde{F}_P^*(Z) \Leftrightarrow Z' = Z + 2ZZ^\top Z \tag{31}$$

with respect to the Riemannian metric $\tilde{g}_3$;

$$\langle \Omega_1, \Omega_2 \rangle_{\tilde{g}_3} = 2\mathrm{tr}\{(\mathrm{I} + \mathrm{Z}\mathrm{Z}^\top)^{-1}\Omega_1\Omega_2^\top\} \tag{32}$$

for $\Omega_i \in T_X\mathbb{R}^{n\times k} \cong \mathbb{R}^{n\times k}$. Thus, we easily see that the equilibrium point Z_∞ of (31) is 0. In order to characterize the relation between the equilibrium points of (22) and (31), we refer the following Lemma due to [5].

Lemma 6 ([5]). *Let* $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_\mathrm{n})$ *with* $\lambda_n < \ldots < \lambda_1$. *An equilibrium point* X_∞ *of* (21) *is characterized by*

$$X_\infty X_\infty^\top = \mathrm{diag}(\varepsilon_1\lambda_1, \ldots, \varepsilon_\mathrm{n}\lambda_\mathrm{n}), \quad \varepsilon_\mathrm{i} \in \{0, 1\}. \tag{33}$$

The following proposition gives a characterization for the equilibrium points of the primal flow (Oja like flow) and its dual flow.

Proposition 7. *Let* $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_\mathrm{n})$ *with* $0 < \lambda_n < \ldots < \lambda_1$. *The primal equilibrium point* X_∞ *is characterized by*

$$X_\infty X_\infty^\top = 0. \tag{34}$$

Proof. By $Z = (A - XX^\top)^{-\frac{1}{2}}X$, we have

$$(A - XX^\top)^{\frac{1}{2}}(ZZ^\top + I)(A - XX^\top)^{\frac{1}{2}} = A \tag{35}$$

for any $X \in Dom(\tilde{F}_P)$ and $Z \in \mathbb{R}^{n\times k}$. The primal equilibrium point X_∞ and its dual equilibrium point Z_∞ should satisfy (35). Therefore, by (35) and $Z_\infty = 0$, we conclude that the primal equilibrium point X_∞ is specified by $X_\infty X_\infty^T = 0$, which is a specific case in Lemma 6. □

3 An application

Considering the logarithmic determinant cost function

$$\mathcal{F}(X, Y) = \log\det(I + XY^\top) \tag{36}$$

for any $X, Y \in \mathbb{R}^{n\times k}$ satisfying $0 < \det(I + XY^\top)$, we will show a simple inequality related to the following MacMahon's Master Theorem in combinatory analysis.

Theorem 8 (MacMahon's Master Theorem). *Given an n by n matrix* $A = (a_{ij})$ *over some commutative ring R and commuting indeterminates* $x_1, \ldots, x_n$ *over R. Let* $\boldsymbol{m} = (m_1, \ldots, m_n) \in \mathbb{Z}^n$ *be a multi-index with* $0 \le m_i$, *and let* $C_A(\boldsymbol{m})$ *be the R-coefficient of* $\boldsymbol{x^m} = x_1{}^{m_1}x_2{}^{m_2}\ldots x_n{}^{m_n}$ *in* $\prod_{i=1}^n(\sum_{j=1}^n a_{ij}x_j)^{m_i} \in R[x_1, \ldots, x_n]$.
Then the following identity holds;

$$1 = \det(I_n - A\cdot\mathrm{diag}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}))\cdot\sum_{\boldsymbol{m}} C_A(\boldsymbol{m})\boldsymbol{x^m}. \tag{37}$$

Proof. See [3] for example. □

Theorem 9. *Let $\mathcal{F}(X,Y)$ be the function* (36) *defined on*

$$\mathrm{Dom}(\mathcal{F}) = \{(\mathrm{X},\mathrm{Y}) \in \mathbb{R}^{\mathrm{n}\times\mathrm{k}} \times \mathbb{R}^{\mathrm{n}\times\mathrm{k}} | \quad 0 < \det(\mathrm{I} + \mathrm{XY}^\top)\}. \tag{38}$$

If we define the Riemannian metric **g** *on* $\mathrm{Dom}(\mathcal{F})$ *as*

$$\langle\Omega_1,\Omega_2\rangle_{\mathbf{g}} = \mathrm{tr}\left\{(\mathrm{I_n} + \mathrm{YX}^\top)^{-\frac{1}{2}}\Omega_{11}\Omega_{21}^\top + (\mathrm{I_n} + \mathrm{XY}^\top)^{-\frac{1}{2}}\Omega_{12}\Omega_{22}^\top\right\} \tag{39}$$

for tangent vectors $\Omega_1 = (\Omega_{11},\Omega_{12})$*,* $\Omega_2 = (\Omega_{21},\Omega_{22}) \in T_{(X,Y)}\mathrm{Dom}(\mathcal{F}) \cong \mathbb{R}^{\mathrm{n}\times\mathrm{k}} \times \mathbb{R}^{\mathrm{n}\times\mathrm{k}}$ *then the dual function* $\mathcal{F}^*$ *to* (36) *is*

$$\mathcal{F}^*(Z,W) = 2\mathrm{tr}(\mathrm{WZ}^\top) + \log\det(\mathrm{I_n} - \mathrm{WZ}^\top) \tag{40}$$

defined on

$$\mathrm{Dom}(\mathcal{F}^*) = \{(\mathrm{Z},\mathrm{W}) \in \mathbb{R}^{\mathrm{n}\times\mathrm{k}} \times \mathbb{R}^{\mathrm{n}\times\mathrm{k}} | (\mathrm{Z},\mathrm{W}) \in (\mathrm{Im}(\Phi_1),\mathrm{Im}(\Phi_2))\}, \tag{41}$$

where $\mathrm{Im}(\Phi)$ *denotes the image of* Φ*, and*

$$\Phi_1(X,Y) = (I_n + YX^\top)^{-\frac{1}{2}}Y, \quad \Phi_2(X,Y) = (I_n + XY^\top)^{-\frac{1}{2}}X. \tag{42}$$

The relative entropy (10) *is*

$$D(\mathcal{F}(X,Y),\mathcal{F}^*(Z,W)) = \log\det(I_n + XY^\top) + \log\det(I_n - WZ^\top) \geq 0 \tag{43}$$

for any $(X,Y) \in \mathrm{Dom}(\mathcal{F})$ *and* $(Z,W) \in \mathrm{Dom}(\mathcal{F}^*)$*. Equality in* (43) *holds if and only if*

$$Z = (I_n + YX^\top)^{-\frac{1}{2}}Y, \quad W = (I_n + XY^\top)^{-\frac{1}{2}}X.$$

Proof. In order to define the dual variables, we consider the gradient of (36) with respect to the Riemannian metric **g** as follows;

$$\begin{aligned} \partial_X\mathcal{F}(X,Y) &= \Phi_1(X,Y) = (I_n + YX^\top)^{-\frac{1}{2}}Y = Z, \\ \partial_Y\mathcal{F}(X,Y) &= \Phi_2(X,Y) = (I_n + XY^\top)^{-\frac{1}{2}}X = W. \end{aligned} \tag{44}$$

From

$$\begin{aligned} WZ^\top &= (I + XY^\top)^{-\frac{1}{2}}XY(I + XY^\top)^{-\frac{1}{2}} \\ &= I - (I + XY^\top), \end{aligned}$$

we get

$$I + XY^\top = (I - WZ^\top)^{-1}. \tag{45}$$

Substituting (45) and its transpose into (44), X and Y are of the form

$$X = (I - WZ^\top)^{-\frac{1}{2}}W, \quad Y = (I - ZW^\top)^{-\frac{1}{2}}Z. \tag{46}$$

Furthermore, substituting (46) into the equation

$$\mathcal{F}^*(Z,W) = \langle(X,Y),(Z,W)\rangle_{\mathbf{g}} - \mathcal{F}(X,Y),$$

we get the dual function. By the definition of the relative entropy (10), we obtain the inequality (43) . □

The following corollary is an immediate consequence of Theorem 9.

Corollary 10. *For any* $(X,Y) \in \mathrm{Dom}(\mathcal{F})$ *and* $(Z,W) \in \mathrm{Dom}(\mathcal{F}^*)$*, we have*

$$1 \le \det(I_n + XY^\top) \cdot \det(I_n - WZ^\top) \tag{47}$$

and equality holds if and only if

$$Z = (I_n + YX^\top)^{-\frac{1}{2}} Y, \quad W = (I_n + XY^\top)^{-\frac{1}{2}} X.$$

Acknowledgments

Thanks to Dr Jochen Trumpf for pointing out an error in Proposition 7.

Bibliography

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. *Princeton University Press*, 2007. Cited p. 471.

[2] V. Blondel and A. Megretski. Unsolved problems in mathematical systems and control theory. *Princeton University Press*, 2004. Cited p. 472.

[3] I. J. Good. A short proof of MacMahon's Master Theorem. *Proc. Cambridge Philos. Soc.*, 58:160, 1962. Cited p. 477.

[4] U. Helmke and J. B. Moore. Optimization and dynamical systems. *Springer*, 1994. Cited p. 471.

[5] U. Helmke, M. Prechtel, and M. A. Shayman. Riccati-like flows and matrix approximations. *Kybernetika*, 29:563–582, 1993. Cited pp. 474 and 476.

[6] J. Manton, U. Helmke, and I. M. Y. Mareels. A dual purpose principal and minor component flow. *Systems and Control Letters*, 54:759–769, 2005. Cited pp. 471 and 472.

[7] E. Oja, H. Ogawa, and J. Wangviwattana. Principal component analysis by homogeneous neural networks, Part 1: The weighted subspace criterion. *IEICE Trans. Inform. Systems*, 3:366–375, 1992. Cited p. 471.

[8] E. Oja, H. Ogawa, and J. Wangviwattana. Principal component analysis by homogeneous neural networks, Part 2: Analysis and extension of the learning algorithms. *IEICE Trans. Inform. Systems*, 3:376–382, 1992. Cited p. 471.

[9] R. T. Rockafellar. Convex analysis. *Princeton University Press*, 1970. Cited p. 472.

[10] L. Xu. Least mean square error recognition principle for self organizing neural nets. *Neural Networks*, 6:627–648, 1993. Cited p. 471.

[11] S. Yoshizawa, U. Helmke, and K. Starkov. Convergence analysis for principal component flows. *Int. J. Appl. Math. Comput. Sci.*, 11:223–236, 2001. Corrections in ibid. 12:299, 2002. Cited p. 471.

Dynamic negotiation under switching communication

Daniel Zelazo
University of Stuttgart
Stuttgart, Germany
daniel.zelazo@ist.uni-stuttgart.de

Mathias Bürger
University of Stuttgart
Stuttgart, Germany
buerger@ist.uni-stuttgart.de

Frank Allgöwer
University of Stuttgart
Stuttgart, Germany
frank.allgower@ist.uni-stuttgart.de

Abstract. This work considers the problem of a dynamic negotiation process amongst selfish agents under a switching communication scheme. We study a negotiation problem between dynamical agents with discrete-time integrator dynamics. Each agent desires to minimize its own quadratic objective function with the additional requirement that the ensemble must collectively agree on a terminal state in finite time. The trajectories of each agent are generated in real-time and the negotiation process to determine the terminal state occurs over a switching communication network. We present what we term a “shrinking horizon" property to enforce the terminal constraint. A first result shows that the algorithm is equivalent to a switched linear system and the performance of the system is studied in the context of certain error signals relating the algorithm to a centralized optimization problem. The performance of the algorithm is shown to depend on certain assumptions in the switching signal, namely joint connectedness, and certain spectral properties of the switched graphs.

1 Introduction

An important feature of multi-agent systems is their ability to perform complex tasks in a distributed manner. Central to many of the tasks performed by these systems is the ability for the team to distributedly reach an agreement on a certain parameter. This can include mundane objectives such as distributedly computing the average of a set of numbers[1, 14], or more complex behaviors including agreeing on a desired heading for a team of unmanned vehicles[10, 12]. In a broader context, the task of reaching agreement on a parameter can be viewed as a certain optimization problem. This optimization problem should be solved in a distributed fashion according to the constraints of the system [11].

It is very common in many multi-agent system applications to make certain assumptions on the tasks to be solved and the constraints of the system. For instance, many distributed optimization algorithms will assume a fixed communication graph over which their coordination occurs [11]. In other systems, the goal of reaching an agreement on a parameter is specified only as an asymptotic behavior [2]. In many settings, however, such assumptions can not be justified, and this motivates the present work.

We consider a team of dynamic agents modeled by a discrete-time integrator dynamics. The team is tasked with the objective of agreeing on a common value for their state. In fact, the agents must *physically* move to this final state from their initial condition along some trajectory to be computed. This task is complicated by a number of additional important constraints. Each agent in the system is considered “selfish"; that is each agent has a local quadratic objective function penalizing its distance to a “preference" state and its control energy. The negotiation of the terminal state must be performed distributedly over a dynamic communication graph. We assume that the communication graph switches as time progresses. The switching process can be used to model many real-world constraints, such as packet losses in a network, power and bandwidth restrictions on communication, or state-dependent sensor measurements. Finally, the team must arrive at the terminal state within a specified time horizon. This last constraint emphasizes that communication between agents to coordinate takes time and it is therefore in the best interest of each agent to move along a trajectory it believes to be optimal at each communication round. This is in contrast to methods that might require the agents to first agree on a terminal state and then compute their trajectories and move [6]. In this way, we consider the time horizon of the problem as a hard deadline for the agents for determining trajectories, negotiating the terminal state, and physically moving along these trajectories.

This problem builds on a previous work with a similar setup, the main difference being the switched communication scheme presented here [15, 16]. The main result of that work was the presentation of a distributed algorithm, termed the *shrinking horizon preference agreement* (SHPA) algorithm, and an analysis of its convergence properties. Mirroring the outline of that work, we explore how the presence of a switching communication scheme affects the algorithm. In this direction, we first provide a convergence analysis for a distributed sub-gradient algorithm with switched communication that asymptotically computes the solution of the preference agreement problem. The convergence of the algorithm depends on an assumption on the switching signal requiring the communication graph to be jointly connected over a finite interval of time; this is in the spirit of similar results for switched consensus protocols and distributed optimization over random graphs [5, 8, 9]. We proceed to show that the SHPA algorithm is equivalent to a switched linear dynamical system. As this is a finite horizon problem, we use a contraction-based argument to analyze its performance. Our analysis concludes that in addition to the joint connectedness assumption, we require additional assumptions on the spectral properties of the switching graphs to guarantee an overall contraction of the system over the entire horizon. When the spectral constraints can not be met, we are able to show contraction over a specified interval.

The organization of this paper is as follows. The next subsection introduces the main notations of this work. The general problem set-up is given in §2. This section also presents a distributed algorithm that can asymptotically solve the preference agreement problem under a switching communication scheme. In §3 the SHPA algorithm and its associated switched linear system is presented. The performance of the algorithm is given in §4. Finally, some simulation examples are provided in §5 and concluding remarks offered in §6.

Notation

The notation we employ is standard. The set of real numbers is denoted $\mathbb{R}$, and $\mathbb{R}_{>}$ ($\mathbb{R}_{\geq}$) is the set of positive (non-negative) numbers. For a vector $x \in \mathbb{R}^n$, we denote its transpose by $x^\top$, and its ith component by $x(i)$; the ijth element of the matrix A is given as $[A]_{ij}$. The all ones vector of length n is denoted $\mathbb{1}_n$ and I_n is the $n \times n$ identity matrix. The inner-product of two vectors is denoted $\langle x,y \rangle = x^\top y$; the Euclidean norm of a vector x is denoted $\|x\|_2 = \langle x,x \rangle^{1/2}$. The communication structure between agents is captured by a graph $\mathcal{G}$ with node set $\mathcal{V} = \{v_1,\ldots,v_n\}$ and edge set $\mathcal{E}$. The *complete graph*, denoted K_n, is the graph with each pair of distinct vertices connected by an edge. The *node-edge incidence matrix* of the graph $\mathcal{G}$, denoted $E(\mathcal{G}) \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$ is defined in the usual way [3]. The union of k graphs $\mathcal{G}_i = (\mathcal{V},\mathcal{E}_i)$, for $i = 1,\ldots,k$ is a graph on the node-set $\mathcal{V}$ with edge-set $\cup_{i=1}^k \mathcal{E}_i$; this is denoted as $\mathcal{G} = \cup_{i=1}^k \mathcal{G}_i$.

> *This manuscript is dedicated to Uwe Helmke, a gifted researcher and educator, and an inspiring research companion and friend over many years.*

2 The finite-time agreement problem

We consider a group of n self-interested dynamical agents that must agree upon a common state at the end of a given time horizon. Each agent is modeled as a single integrator,

$$x_i(t+1) = x_i(t) + u_i(t),\; x_i(0) = x_{i0}, \tag{1}$$

with $i = 1,\ldots,n$ and $x_i(t) \in \mathbb{R}$. The state and control vector for all n agents are denoted as $x(t) = [x_1(t),\ldots,x_n(t)]^\top$ and $u(t) = [u_1(t),\ldots,u_n(t)]^\top$.

The self-interest of each agent is modeled as a quadratic objective, attaining its minimum at a specific individual preference value ξ_i. Each agent aims to minimize the objective

$$J_i(t_0,T,x_i,u_i) = \frac{1}{2}\left(\sum_{t=t_0}^{T-1}(x_i(t+1)-\xi_i)^2 + u(t)^2\right). \tag{2}$$

The individual agents are coupled by a requirement to achieve agreement on their state at the end of the time horizon T; that is there is a terminal time state constraint,

$$x_1(T) = x_2(T) = \cdots = x_n(T). \tag{3}$$

This constraint can be compactly written using the incidence matrix for the complete graph as $E(K_n)^\top x(T) = 0$.

From a centralized perspective, the preference-based agreement problem can be stated as the optimal control problem with terminal state constraint

$$OCP(t_0,T,x_0)\colon \min_{x,u} \sum_{i=1}^n J_i(t_0,T,x_i,u_i) \tag{4}$$

$$\text{s.t.}\quad x(t+1) = x(t) + u(t),\; x(t_0) = x_0 \tag{5}$$

$$E(K_n)^\top x(T) = 0. \tag{6}$$

We collect the entire state and control trajectories of each agent into the row vectors $\mathbf{x}_i = \begin{bmatrix} x_i(t_0+1) & \cdots & x_i(T) \end{bmatrix}$ and $\mathbf{u}_i = \begin{bmatrix} u_i(t_0) & \cdots & u_i(T-1) \end{bmatrix}$. As we are considering a team of n agents, we introduce further notation to streamline the presentation. The bold-face vectors

$$\mathbf{x} = \begin{bmatrix} (\mathbf{x}_1)^\top & \cdots & (\mathbf{x}_n)^\top \end{bmatrix}^\top \in \mathbb{R}^{n\times T} \text{ and } \mathbf{u} = \begin{bmatrix} (\mathbf{u}_1)^\top & \cdots & (\mathbf{u}_n)^\top \end{bmatrix}^\top \in \mathbb{R}^{n\times T}$$

denote the complete trajectories for the state and control of the entire ensemble of agents, respectively, and $(\overline{\mathbf{x}},\overline{\mathbf{u}})$ denotes the *optimal* trajectory generated by the solution of $OCP(t_0,T,x_0)$. At times, we will be interested in the state or control trajectory value for all agents at a particular time τ; we will denote this by $\mathbf{x}(\tau) \in \mathbb{R}^{n\times 1}$ and $\mathbf{u}(\tau) \in \mathbb{R}^{n\times 1}$.

The problem $OCP(t_0,T,x_0)$ can be reformulated as a static quadratic program. Using the introduced notation, the objective for each agent can be stated as

$$J_i(t_0,T,\mathbf{x}_i,\mathbf{u}_i) = \frac{1}{2}(\|\mathbf{x}_i - \mathbb{1}_T^\top \xi_i\|_2^2 + \|\mathbf{u}_i\|_2^2),$$

and the dynamic constraint as the linear equation

$$\mathbf{x}_i = \mathbb{1}_T^\top x_{i0} + \mathbf{u}_i B_T^\top. \tag{7}$$

Here, $B_T \in \mathbb{R}^{T\times T}$ is defined such that $[B_T]_{kl} = 1$ for $k \geq l$ and zero otherwise.

The algorithmic theme of this work builds upon the framework of dual sub-gradient methods for non-linear optimization [13]. In this direction, we will rely on the formulation of the corresponding *dual problem* of (5). The dual problem is obtained by relaxing the coupling constraint with a multiplier μ into the objective to obtain the Lagrangian,

$$\mathcal{L}(\mathbf{x},\mathbf{u},\mu) = \sum_{i=1}^{n} J_i(t_0,T,\mathbf{x}_i,\mathbf{u}_i) + \mu^\top E(K_n)^\top \mathbf{x}(T). \tag{8}$$

The dual function is obtained by minimizing (8) subject to the dynamic constraint (7), $g(\mu) = \min_{\mathbf{x},\mathbf{u}} \mathcal{L}(\mathbf{x},\mathbf{u},\mu)$. We denote the optimal solution of the primal and dual problems as $(\overline{\mathbf{x}}^{t_0},\overline{\mathbf{u}}^{t_0},\overline{\mu}^{t_0})$; the superscript notation is used to explicitly specify the initial condition time used for $OCP(t_0,T,x_0)$. It is important to point out that in fact, there will not be a *unique* multiplier $\overline{\mu}^{t_0}$ corresponding to $OCP(t_0,T,x_0)$. This is a consequence of the redundant constraints encoded by the matrix $E(K_n)$; its kernel is not trivial. The multiplier $\overline{\mu}^{t_0}$ belongs to a set of optimal multipliers, characterized by the first-order optimality conditions for $OCP(t_0,T,x_0)$ and the kernel of the matrix $E(K_n)$. In particular, if $\overline{\mu}^{t_0}$ is one optimal dual multiplier, then all multipliers in the set

$$\mathbb{M} = \left\{\mu \in \mathbb{R}^{|\mathcal{E}(K_n)|} \,|\, \mu = \overline{\mu}^{t_0} + \nu,\ \nu \in \mathcal{N}(E(K_n))\right\} \tag{9}$$

are also solutions of the dual problem. As $OCP(t_0,T,x_0)$ is a strictly convex problem (a quadratic program with linear constraints), we have strong duality which implies that $g(\overline{\mu}^{t_0}) = J(t_0,T,\overline{\mathbf{x}}^{t_0},\overline{\mathbf{u}}^{t_0})$ [13].

In the sequel, we discuss how $OCP(t_0,T,x_0)$ can be solved in a distributed fashion even in the presence of a switching communication network. This will provide the necessary framework to present the main result of this work, the *shrinking horizon preference agreement problem*.

2.1 A dual algorithm for OCP with all-to-all communication

The first question that must be addressed is how the centralized problem $OCP(t_0,T,x_0)$ can be solved in a distributed manner. Indeed, if the communication graph is fixed, a standard approach to solve the problem $OCP(t_0,T,x_0)$ is by a *dual sub-gradient algorithm* [13]. We will briefly summarize the sub-gradient algorithm.

Observe that the Lagrangian function (8) is separable across each agent in the ensemble. The local terminal state constraint of a single agent is penalized in the Lagrangian by a corresponding Lagrange multiplier on the edges incident to that agent. In a similar manner, we can also consider a variable associated with each agent instead of each edge by defining

$$\gamma := E(K_n)\mu \in \mathbb{R}^n. \tag{10}$$

In this setting, the Lagrangian can be written as the separable function

$$\mathcal{L}(\mathbf{x},\mathbf{u},\gamma) = \sum_{i=1}^{n} J_i(t_0,T,\mathbf{x}_i,\mathbf{u}_i) + \gamma_i \mathbf{x}_i(T). \tag{11}$$

The dual sub-gradient algorithm proceeds now as follows. At each iteration step k, the dual function is computed for a fixed value of $\hat{\gamma}^{[k]}$. That is, each agent solves the following quadratic program, $QP_i(k)$,

$$(\hat{\mathbf{x}}_i^{[k+1]},\hat{\mathbf{u}}_i^{[k+1]}) = \underset{\hat{\mathbf{x}}_i^{[k]},\hat{\mathbf{u}}_i^{[k]}}{\arg\min} J_i(t_0,T,\hat{\mathbf{x}}_i^{[k]},\hat{\mathbf{u}}_i^{[k]}) + \hat{\gamma}_i^{[k]}\hat{\mathbf{x}}_i^{[k]}(T) \tag{12}$$

$$\text{s.t.} \quad \hat{\mathbf{x}}_i^{[k]} = \mathbb{1}_{\tilde{T}}^{\top} x_{i0} + \hat{\mathbf{u}}_i^{[k]} B_{\tilde{T}}^{\top}. \tag{13}$$

Here we have temporarily abused our notation to facilitate this discussion. The superscript, as in $\gamma^{[k]}$, denotes the iteration count for the sub-gradient algorithm, and the notation $(\hat{\mathbf{x}}_i^{[k]},\hat{\mathbf{u}}_i^{[k]})$ denotes the optimization variables for $QP_i(k)$. While ensuring that the initial values of the dual variables satisfy $\gamma^{[0]} = E(\mathcal{G})\mu^{[0]}$, the next step is then to update the multiplier using the sub-gradient as

$$\hat{\gamma}^{[k+1]} = \hat{\gamma}^{[k]} + \alpha^{[k]} E(K_n)E(K_n)^{\top}\hat{\mathbf{x}}^{[k+1]}(T). \tag{14}$$

The sub-gradient for the edge multiplier μ is precisely $E(K_n)^{\top}\hat{\mathbf{x}}^{[k]}(T)$, and using (10) leads to (14). The matrix $E(K_n)E(K_n)^{\top}$ is the *graph Laplacian* of K_n, $L(K_n)$[3]. Note that owing to the particular structure of the optimization problem $QP_i(k)$, an *analytic* solution for the terminal state $\hat{\mathbf{x}}_i^{[k+1]}$ can be obtained from the first-order optimality conditions of the problem. In particular, we have that

$$\hat{\mathbf{x}}_i^{[k+1]} = k(x_i(t_0) - \xi_i) + \xi_i - p\hat{\gamma}_i^{[k]}; \tag{15}$$

the constants k and p appear from solving the first-order optimality conditions, and are identical for all agents. In fact, they will play an important role later in this work, and we will revisit their derivation and interpretation again. The key point is the analytic solution can be used explicitly in the update of the multipliers to obtain the following iteration,

$$\hat{\gamma}^{[k+1]} = (I - \alpha^{[k]} pL(K_n))\hat{\gamma}^{[k]} + \alpha^{[k]} L(K_n)\left(k(x(t_0) - \xi) + \xi\right). \tag{16}$$

When formulated in this way, it becomes clear that the choice of the step-size $\alpha^{[k]}$ becomes a critical parameter for the convergence of the algorithm. For a suitable choice of the step-size, the sub-gradient algorithm will converge to the optimal solution of $OCP(t_0, T, x_0)$,

$$\lim_{k\to\infty} (\hat{\mathbf{x}}^{[k]}, \hat{\mathbf{u}}^{[k]}, \hat{\gamma}^{[k]}) = (\overline{\mathbf{x}}^{(t_0,x_0)}, \overline{\mathbf{u}}^{(t_0,x_0)}, E(K_n)\overline{\mu}^{(t_0,x_0)}).$$

For a more detailed discussion of appropriate step-size rules and sub-gradient methods the reader is referred to [13].

The appeal of this method is that the update rule (14) is inherently distributed. That is, each agent can compute the value $\gamma_i^{[k+1]}$ to use in the next iteration step solely through communication with its neighbors, as defined by the communication graph. In particular, agent i must only send the value $\hat{\mathbf{x}}_i^{[k]}(T)$ to all neighboring agents. However, with this formulation we have assumed a complete communication graph. This solution method will work for any connected graph $\mathcal{G}$, reducing the overall communication requirement for the graph. In the next sub-section, we examine this algorithm when the communication graph switches at each iteration of the algorithm.

2.2 A dual algorithm for OCP with switching communication

We now consider how this algorithm performs if at each iteration step k, the communication graph changes. We assume in the following that agents can communicate with each other synchronously according to a time-varying communication graph. The set of all graphs on the node-set $\mathcal{V} = \{v_1, \ldots, v_n\}$ is denoted as $\mathbf{G}$. To model the time-varying nature of the communication, we introduce the *switching signal* $\sigma : \{0, 1, \ldots\} \to \mathbf{G}$ and denote the communication graph available to the agents at time k (here equivalent to an iteration of the algorithm) as $\mathcal{G}_{\sigma(k)} \in \mathbf{G}$. Often we will interpret the graph $\mathcal{G}_{\sigma(k)}$ as the complete graph with $\{0,1\}$−weights on each edge such that the weight on edge $e \in \mathcal{E}(K_n)$ is 1 if and only if that edge is present in the graph $\mathcal{G}_{\sigma(k)}$, and is zero otherwise; this is captured by the diagonal weight matrix $W_{\sigma(k)} \in \mathbb{R}^{|\mathcal{E}(K_n)| \times |\mathcal{E}(K_n)|}$ and denoted $\mathcal{G}_{\sigma(k)} = (\mathcal{V}, \mathcal{E}(K_n), W_{\sigma(k)})$. In this way, the incidence matrix for the graph $\mathcal{G}_{\sigma(t)}$ can be written as $E(\mathcal{G}_{\sigma(k)}) = E(K_n)W_{\sigma(k)}$. Similarly, the graph Laplacian matrix for the graph $\mathcal{G}_{\sigma(k)}$ can be expressed as $L(\mathcal{G}_{\sigma(k)}) = E(K_n)W_{\sigma(k)}E(K_n)^T$ [3]. We make the following assumption on the sequence of graphs generated by the switching signal.

Assumption 1 (Uniformly Jointly Connected)**.** There exists a finite and positive integer Δ such that for all $k_0 \geq 0$, the graph

$$\hat{\mathcal{G}} = \cup_{i=0}^{\Delta-1} \mathcal{G}_{\sigma(k_0+i)}$$

is connected.

We study now the behavior of the dual sub-gradient algorithm performed by a network of agents communicating according to a switching topology. In this case, the multiplier update stage can only communicate with neighbors specified by the graph $\mathcal{G}_{\sigma(k)}$. This leads to a modified iteration for the multiplier update (16), given as

$$\hat{\gamma}^{[k+1]} = (I - \alpha^{[k]} p L(\mathcal{G}_{\sigma(k)}))\hat{\gamma}^{[k]} + \alpha^{[k]} L(\mathcal{G}_{\sigma(k)})\,(k(x(t_0) - \xi) + \xi)\,, \tag{17}$$

The convergence analysis of (17) now falls under the realm of *switched linear systems* [7]. We now provide some basic results that will aid in the stability analysis of the system in (17).

Lemma 2. *The signal $\mathbb{1}^\top \hat{\gamma}^{[k]}$ is invariant under the dynamics (17). In particular, when $\hat{\gamma}^{[0]}$ is initialized as $\hat{\gamma}^{[0]} = E(K_n)W_{\sigma(0)}\mu^{[0]}$ for an arbitrary vector $\mu^{[0]} \in \mathbb{R}^{|\mathcal{E}(K_n)|}$, then $\mathbb{1}^\top \hat{\gamma}^{[k]} = 0$ for all $k = 0, 1, 2, \ldots$.*

Proof. The invariance of $\mathbb{1}^\top \hat{\gamma}^{[k]}$ under the dynamics (17) is verified by recalling that $\mathbb{1}^\top L(\mathcal{G}_{\sigma(k)}) = 0$ for any undirected graph $\mathcal{G}_{\sigma(k)} \in \mathbf{G}$. When $\gamma^{[0]}$ is initialized as above, one has $\gamma^{[0]} = \mathbb{1}^\top E(K_n)W_{\sigma(k)}\mu^{[0]} = 0$. □

It is also useful to observe that there exists a constant step-size $\overline{\alpha}$ such that the matrix

$$A_{\sigma(k)} = (I - \overline{\alpha} p L(\mathcal{G}_{\sigma(k)})) \tag{18}$$

contains all its eigenvalues inside the closed unit-disc.

Lemma 3. *The matrix $A_{\sigma(k)}$ contains only real eigenvalues all within the interval $[-1, 1]$ for any value of $\overline{\alpha}$ satisfying*

$$0 < \overline{\alpha} \le \frac{1}{p(n-1)}.$$

Furthermore, there is at least one eigenvalue at one, and the multiplicity of that eigenvalue is equal to the number of connected components in the graph $\mathcal{G}_{\sigma(k)}$.

Proof. The matrix $A_{\sigma(k)}$ is symmetric and thus has only real eigenvalues. The bound on $\overline{\alpha}$ is obtained by a straight-forward application of Gershgorin's Circle Theorem [4]. The multiplicity of the eigenvalues at one is derived from the result that the rank of the Laplacian matrix is equal to $n - c$, where c is the number of connected components in the graph [3]. □

For the remainder of this analysis, we will assume a constant step-size $\overline{\alpha}$ from Lemma 3 for the dynamics in (17). To study the stability of this switched system, we must first characterize its fixed points, described by the set

$$\mathcal{A} = \{\gamma \in \mathbb{R}^n \,|\, \gamma = p^{-1}\,(k(x(t_0) - \xi) + \xi) - cp^{-1}\mathbb{1}, c \in \mathbb{R}\}. \tag{19}$$

Note that in fact, the equilibrium points are *independent* of the switching signal σ. While Lemma 3 suggests there may be additional equilibrium points due to multiple eigenvalues at unity, we note that this can not be an equilibrium under arbitrary switching. This can be seen explicitly from the following result.

Lemma 4.

$$\cap_{i=1}^{|\mathbf{G}|}\mathcal{N}(L(\mathcal{G}_i)) = \mathbf{span}\{\mathbb{1}\}.$$

Proof. The kernel of any graph Laplacian contains $\mathbf{span}\{\mathbb{1}\}$, and the dimension of the kernel is equal to $n-c$, where c is the number of connected components in the graph [3]. □

We now present two corollaries of Lemma 4 characterizing the spectrum of $A_{\sigma(k)}$ and certain related products.

Corollary 5. *Let $\mathcal{G}_{\sigma(k)}$ have c connected components. Then the matrix $A_{\sigma(k)} = I-\overline{\alpha}pL(\mathcal{G}_{\sigma(k)})$ has c eigenvalues at unity.*

Corollary 6. *Assume that the switching signal σ satisfies Assumption* 1*. Then the matrix product*

$$\tilde{A} = (I-\overline{\alpha}A_{\sigma(k+\Delta-1)})(I-\overline{\alpha}A_{\sigma(k+\Delta-2)})\cdots(I-\overline{\alpha}A_{\sigma(k)})$$

has only one eigenvalue at unity.

Proof. Observe that $cp^{-1}\mathbb{1}$ for any $c\in\mathbb{R}$, is a fixed point of the matrix $\tilde{A}$; this follows from Lemma 4. To show that there can be no other fixed points, assume the contrary; that is that there exists a vector x such that $\tilde{A}x = x$. Expanding the matrix product then reveals that for x to be a fixed point, it must satisfy

$$x\in\cap_{i=k}^{k+\Delta-1}\mathcal{N}(L(\mathcal{G}_{\sigma(i)}) = \mathcal{N}\left(L\left(\cup_{i=1}^{k+\Delta-1}\mathcal{G}_{\sigma(i)}\right)\right) = \mathbf{span}\{\mathbb{1}\},$$

contradicting the assumption. □

An important observation, however, is that if the system in (17) is initialized as $\gamma^{[0]} = E(K_n)W_{\sigma(k)}\mu^{[0]}$ for any $\mu^{[0]}$, then by Lemma 2, the set of equilibrium points is the unique point

$$\gamma^{*} = p^{-1}\left(k(x(t_0)-\xi)+\xi\right) - \left(\frac{\mathbb{1}^{\mathsf{T}}\left(k(x(t_0)-\xi)+\xi\right)}{n}\right)p^{-1}\mathbb{1}\in\mathcal{A} \qquad (20)$$

This last point is important when contrasted with the continuum of multiplier solutions for the problem $OCP(t_0,T,x_0)$ described in (9).

We are now prepared to present the main result on the stability and convergence of the system (17).

Theorem 7. *Consider the switched dynamical system in (*17*) with $\alpha^{[k]} = \overline{\alpha}$ for $k = 0,1,\ldots$ satisfying the condition of Lemma* 3*. Then the system asymptotically converges to the point*

$$\gamma^{*} = p^{-1}\left(k(x(t_0)-\xi)+\xi\right) - \left(\frac{\mathbb{1}^{\mathsf{T}}\left(k(x(t_0)-\xi)+\xi\right)}{n}\right)p^{-1}\mathbb{1}$$

for any switching signal σ satisfying Assumption 1 *and for all initial conditions in the set*

$$\Gamma_0 = \{\gamma\in\mathbb{R}^n \mid \gamma = E(K_n)W_{\sigma(0)}\mu,\, \mu\in\mathbb{R}^{|\mathcal{E}(K_n)|}\}.$$

Proof. To simplify the analysis, we introduce the state transformation $z^{[k]} = p^{1/2}(\hat{\gamma}^{[k]} - \gamma^*)$ to obtain the *symmetric* and autonomous system

$$z^{[k+1]} = (I - \overline{\alpha} pL(\mathcal{G}_{\sigma(k)}))z^{[k]} = (I - \overline{\alpha}\tilde{A}_{\sigma(k)})z^{[k]}.$$

From Assumption 1 and Corollary 6, we can conclude that over any iteration interval $[k, k+K-1]$,

$$\|(I - \overline{\alpha}\tilde{A}_{\sigma(k)})(I - \overline{\alpha}\tilde{A}_{\sigma(k+1)})\cdots(I - \overline{\alpha}\tilde{A}_{\sigma(k+K-1)})\| \leq 1.$$

Furthermore, Corollary 6 also allows us to conclude that the above matrix product has only one eigenvalue at unity, and that is spanned by the vector $\mathbb{1}$. The quadratic Lyapunov function $V(z) = z^T z$ can then be used as a *common weak Lyapunov function* [7], and for any vector $z^{[k]} \notin \mathbf{span}\{\mathbb{1}\}$, one has

$$V(z^{[k+K-1]}) - V(z^{[k]}) < 0.$$

Invoking LaSalle's Invariance Principle, we can conclude that the state asymptotically converges to the largest invariant set, $\mathbf{span}\{\mathbb{1}\}$. Finally, we recall that the initial condition for $\gamma^{[0]}$ is restricted to the set Γ_0 and Lemma 2 requires $\mathbb{1}^{\top}\gamma^{[k]} = 0$ for all k. Therefore, the dynamics of $z^{[k]}$ initialized in the set

$$\{z^{[0]} \in \mathbb{R}^n | z^{[0]} = p^{1/2}(\gamma^{[0]} - \gamma^*), \gamma^{[0]} \in \Gamma_0\}$$

asymptotically converges to the origin, concluding the proof. □

The importance of Theorem 7 is that the distributed sub-gradient algorithm can solve the problem $OCP(t_0,T,x_0)$ even when the communication graph between agents is switching. On the other hand, this result represents only an asymptotic behavior of the system, and the ensemble of agents must execute this *before* they can actually begin to move along their optimal trajectories. Indeed, the convergence rate of this algorithm may be significantly longer than the desired horizon time T of the actual problem.

This then motivates the question whether it is possible to run the algorithm *on-line* and allow each agent to move along a trajectory it believes to be optimal at each iteration step of the algorithm. In other words, we seek to find an algorithm where each iteration step corresponds to the actual physical time of the process and agents “move" at each time step. This algorithm must also negotiate the final agreement value at time T while simultaneously minimizing the local performance index of each agent.

3 A shrinking horizon algorithm

The results of the previous section showed that the problem $OCP(t_0,T,x_0)$ can be solved distributedly even in the presence of a switching communication graph. On the other hand, the dual sub-gradient algorithm only asymptotically computes the solution and the time required to reach this optimum may be unacceptably long depending on the application. This point is further emphasized when considering that each communication round between agents takes some finite amount of time. If we consider the horizon time T as an absolute deadline, then an optimal strategy would require each agent to move towards their preference state in order to minimize their individual objectives before maneuvering to the consensus state.[1]

This motivates the need for a real-time algorithm that allows agents to dynamically negotiate the terminal agreement state while simultaneously attempting to minimize their local objectives as time progresses. The general strategy of this algorithm is to physically propagate the states forward at each iteration. The corresponding sub-problem to be solved at the next time step is the quadratic program $QP_i(t)$, described in (12), but with the horizon window reduced; instead of minimizing from $t = 0$ to the horizon T, we minimize from $t = 1$. It can be considered as a *shrinking-horizon sub-gradient algorithm.*

Here we recall that the state signal $x_i(t)$ corresponds to the true physical state of agent i at time t, and the vectors $\hat{\mathbf{x}}_i^t$ and $\hat{\mathbf{u}}_i^t$ correspond to the optimization variables associated with problem $QP_i(t)$. Note also that as time progresses, the window is shrinking, and $\hat{\mathbf{x}}_i^t, \hat{\mathbf{u}}_i^t \in \mathbb{R}^{\tilde{T}}$ with $\tilde{T} = T - t$. See Algorithm 9 for a description.

Algorithm 9: Shrinking horizon preference agreement (SHPA) algorithm

```
Data: Initial conditions $x_i(0) = x_{i0}$ and $\gamma(0) = E(K_n)\mu_0$; $t = 0$.
begin
    for $t := 0$ to $T\text{-}1$ do
        $\tilde{T} = T - t$
        Each agent solves the sub-problem $QP_i(t)$:

        $$\min_{\hat{\mathbf{x}}_i(t),\hat{\mathbf{u}}_i(t)} J_i(t,T,\hat{\mathbf{x}}_i^{(t)},\hat{\mathbf{u}}_i^{(t)}) + \gamma_i^t \hat{\mathbf{x}}_i^{(t)}(T) \text{ s.t. } \hat{\mathbf{x}}_i^t = \mathbb{1}_{\tilde{T}} x_i(t) + B_{\tilde{T}} \hat{\mathbf{u}}_i^t \qquad (21)$$

        The physical state and multipliers are propagated forward using the
        solution of $QP_i(t)$:

        $$x_i(t+1) = x_i(t) + \hat{\mathbf{u}}_i^t(t),\ i = 1,\dots,n \qquad (22)$$
        $$\gamma(t+1) = \gamma(t) + \alpha(t) L(\mathcal{G}_{\sigma(t)})^{\top} \hat{\mathbf{x}}^t(T) \qquad (23)$$

        where $\alpha(t)$ satisfies some step-size rule.
    end
end
```

In the algorithm, each iteration step corresponds to the true progression of time for the dynamic system. At each discrete time instant $t < T$, agent i solves its own optimal control problem over the horizon t to T using the current value of $\gamma_i(t)$. As discussed in the previous section, the solution of $QP_i(t)$ admits an analytic solution and that can be used to propagate the true *physical system state*, $x_i(t)$, forward. The analytic solution of the terminal state value can be used to propagate the multiplier value $\gamma_i(t)$.

[1] This reasoning assumes that T is sufficiently large. For a shorter horizon each agent might not have enough time to reach its preference.

The updated state and multiplier values are then used as the initial condition in the next iteration round. The key point in this algorithm is with each iteration step of the algorithm, the agents are physically moving along a trajectory they believe to be optimal based on the multiplier value they have.

The relation of the SHPA algorithm to the dual methods presented in §2 is clear from the update equation of the multiplier $\gamma(t)$. The main difference is that at each time step the physical state of the system is changing and the corresponding sub-problem $QP_i(t)$ is also modified. Furthermore, the algorithm terminates after $T-1$ steps. In this way, the SHPA algorithm can be interpreted as a *dynamic negotiation protocol* to determine the consensus value. The multipliers $\gamma_i(t)$ can then be considered as an estimate by each agent of the preferences of neighboring agents.

It remains to analyze the trajectories produced by Algorithm 9 and evaluate its performance as related to the asymptotic algorithms presented in §2. A first result in this direction is to show that the trajectories produced by Algorithm 9 are equivalent to the trajectories of a switched linear dynamical system. The following theorem summarizes this result.

Theorem 8. *Algorithm* 9 *is equivalent to the switched linear system*

$$\begin{bmatrix} x(t+1) \\ \gamma(t+1) \end{bmatrix} = \begin{bmatrix} (1-p(\tilde{T}))I & -k(\tilde{T})I \\ \alpha(t)k(\tilde{T})L(\mathcal{G}_{\sigma(t)}) & I-\alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)}) \end{bmatrix} \begin{bmatrix} x(t) \\ \gamma(t) \end{bmatrix} + \begin{bmatrix} p(\tilde{T})I \\ \alpha(t)(1-k(\tilde{T}))L(\mathcal{G}_{\sigma(t)}) \end{bmatrix} \xi \tag{24}$$

with $\tilde{T} = T-t$ and the gains $p(\tilde{T})$ and $k(\tilde{T})$ satisfy the recursion

$$p(\tilde{T}+1) = \frac{1+p(\tilde{T})}{2+p(\tilde{T})}, \qquad p(1) = \frac{1}{2} \tag{25}$$

$$k(\tilde{T}+1) = \frac{k(\tilde{T})}{2+p(\tilde{T})}, \qquad k(1) = \frac{1}{2}. \tag{26}$$

The main effort of the proof for Theorem 8 lies in the derivation of the gains $p(\cdot)$ and $k(\cdot)$. The details of their derivation can be found in a companion work [15, 16]. It turns out that the gains $p(\cdot)$ correspond to the time-varying finite-horizon LQR gains for each agent. Therefore they can be computed off-line and independently of the algorithm or even the communication graph. In this problem set-up, we have assumed all agents have the same state and control gain, and consequently the gains $p(\cdot)$ and $k(\cdot)$ are identical for each agent. We also make use of the analytic solutions for $\hat{\mathbf{u}}_i^t(t)$ and $\hat{\mathbf{x}}^t(T)$, given in (15), which can be derived directly from the quadratic program sub-problem in the SHPA algorithm; we present the expression for $\hat{\mathbf{u}}^{(t)}(t)$ and $\hat{\mathbf{x}}^{(t)}(T)$ here for completeness.

$$\hat{\mathbf{u}}^{(t)}(t) = -p(\tilde{T})(x(t)-\xi) - k(\tilde{T})\gamma(t) \tag{27}$$

$$\hat{\mathbf{x}}^{(t)}(T) = K(\tilde{T})(x(t)-\xi) + \xi - Q^{-1}P(\tilde{T})\gamma(t). \tag{28}$$

The utility of the switched linear system representation of Algorithm 9 is that stability and convergence issues can be analyzed directly using tools from switched linear systems theory. In particular, we observe that the only free parameter is the step-size $\alpha(t)$. Therefore, choosing an appropriate value for $\alpha(t)$ can now be cast as a stabilization and performance problem of the system. In the sequel, we examine the performance and convergence of the system.

4 Performance of the SHPA algorithm

The ultimate objective of the SHPA algorithm is for the collection of dynamic agents to negotiate in real-time a terminal state within a finite horizon. A natural measure of the performance of the algorithm, therefore, is the distance the agents are from an agreement state at the time T,

$$\|E(K_n)^\top x(T)\|_2. \tag{29}$$

Recall that the *optimal* state trajectories generated by the asymptotic algorithms in §2 (e.g., the problem $OCP(t,T,x(t))$) will satisfy the terminal constraint exactly, and $\|E(K_n)^\top x(T)\|_2 = 0$. Associated with the optimal trajectory and final state is also an optimal multiplier, $\overline{\gamma}^{(t,x(t))}$. The key observation is that we can not expect the SHPA algorithm to reach perfect agreement. Therefore, another measure of the performance of the system is considered here. Motivated by the proof of Theorem 7, we consider the error between the optimal multiplier $\overline{\gamma}^{(t,x(t))}$ and the multiplier generated by the algorithm $\gamma(t)$ as a performance measure. This is equivalent to the dual error in the static implementation of the algorithm.

For each state $x(t)$ and each remaining time horizon $\tilde{T}$, there is a unique multiplier value that corresponds to the optimal trajectory of the finite-horizon optimal control problem. As previously shown in (20), this optimal multiplier computes as

$$\begin{aligned}\overline{\gamma}^{(t,x(t))} &= p^{-1}(\tilde{T})\left(k(\tilde{T})(x(t)-\xi)+\xi\right)\\ &\quad -\underbrace{\left(\frac{\mathbb{1}^\top\left(k(\tilde{T})(x(t)-\xi)+\xi\right)}{n}\right)}_{=:c(\tilde{T},x(t))}\left(p^{-1}(\tilde{T})\mathbb{1}\right).\end{aligned} \tag{30}$$

The multiplier error at time t can then be expressed as

$$\varepsilon(t) := \gamma(t)-\overline{\gamma}^{(t,x(t))}.$$

We can now examine the evolution of the error dynamics.

Theorem 9. *The multiplier error $\varepsilon(t) := \gamma(t)-\overline{\gamma}^{(t,x(t))}$ evolves according to the switched linear dynamics*

$$\varepsilon(t+1) = \left(\frac{p(\tilde{T})}{p(\tilde{T}-1)}I-\alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)})\right)\varepsilon(t) \tag{31}$$

with initial condition $\varepsilon(0) = \gamma(0)-\overline{\gamma}^{(0,x(0))}$.

Proof. We have from the dynamics (24)

$$\gamma(t+1) = \gamma(t) - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)})\gamma(t) + \alpha(t)L(\mathcal{G}_{\sigma(t)})(k(\tilde{T})(x(t)-\xi)+\xi). \tag{32}$$

Additionally, it follows from (30) that

$$p(\tilde{T})\bar{\gamma}^{(t,x(t))} = k(\tilde{T})(x(t)-\xi)+\xi - c(\tilde{T},x(t))\mathbb{1}. \tag{33}$$

Note now that by the structure of the graph Laplacian $L(\mathcal{G}_{\sigma(t)})\mathbb{1}=0$, for any switching signal $\sigma(t)$. Therefore, we can add zero to (32), and rewrite the expression as

$$\begin{aligned}\gamma(t+1) = {}& \gamma(t) - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)})\gamma(t)\\ &+\alpha(t)L(\mathcal{G}_{\sigma(t)})\left(k(\tilde{T})(x(t)-\xi)+\xi - c(\tilde{T},x(t))\mathbb{1}\right).\end{aligned} \tag{34}$$

By inserting now (33) into (34), we obtain

$$\begin{aligned}\gamma(t+1) &= \gamma(t) - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)})(\gamma(t)-\bar{\gamma}^{t,x(t)})\\ &= \gamma(t) - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)})\varepsilon(t).\end{aligned} \tag{35}$$

From the *principle of optimality* follows that the optimal multiplier value does not change along the optimal trajectory. We refer the interested reader to [16] for an explicit discussion of this issue. That is

$$\bar{\gamma}^{(t,x(t))} = \bar{\gamma}^{(t+1,x(t)+\bar{u}^{(t,x(t))}(t))}$$

where $\bar{u}^{(t,x(t))}(t)$ is the optimal control input computed according to (27) using the optimal multiplier vector $\bar{\gamma}^{(t,x(t))}$. By adding again zero, we can rewrite (35) as

$$\begin{aligned}\bar{\gamma}^{(t+1,x(t+1))} ={}& p^{-1}(\tilde{T}-1)(k(\tilde{T}-1)(x(t)+\bar{u}^{(t,x(t))}(t)-\xi)+\xi)\\ &-p^{-1}(\tilde{T}-1)\mathbb{1}\cdot c(\tilde{T}-1,x(t)+\bar{u}^{(t,x(t))})\\ &+p^{-1}(\tilde{T}-1)k(\tilde{T}-1)((u(t)-\bar{u}^{(t,x(t))}(t))\\ &-\frac{p^{-1}(\tilde{T}-1)k(\tilde{T}-1)}{n}P^{-1}(\tilde{T}-1)\mathbb{1}\mathbb{1}^{\top}(u(t)-\bar{u}^{(t,x(t))}(t)).\end{aligned} \tag{36}$$

Thus, the evolution of the optimal multiplier can be expressed as

$$\bar{\gamma}^{(t+1,x(t+1))} = \bar{\gamma}^{(t,x(t))} + p^{-1}(\tilde{T}-1)k(\tilde{T}-1)\left(I-\frac{1}{n}\mathbb{1}\mathbb{1}^{\top}\right)(u(t)-\bar{u}^{(t,x(t))}(t)), \tag{37}$$

From the analytic expression of the optimal control input (27) follows directly that

$$u(t)-\bar{u}^{(t,x(t))}(t) = -k(\tilde{T})(\gamma(t)-\bar{\gamma}^{(t,x(t))}).$$

Thus, we can also express the evolution of the optimal multiplier in terms of the error $\varepsilon(t)$, i.e.,

$$\bar{\gamma}^{(t+1,x(t+1))} = \bar{\gamma}^{(t,x(t))} - p^{-1}(\tilde{T}-1)k(\tilde{T}-1)k(\tilde{T})\left(I-\frac{1}{n}\mathbb{1}\mathbb{1}^{\top}\right)\varepsilon(t). \tag{38}$$

With the two recursions (35) and (38), one can write the dynamics of the error as

$$\begin{aligned}\varepsilon(t+1) = \Big(&I - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)}) \\ &+ p^{-1}(\tilde{T}-1)k(\tilde{T}-1)k(\tilde{T})\Big(I - \frac{1}{n}\mathbb{1}\mathbb{1}^\top\Big)\Big)\varepsilon(t)\end{aligned} \tag{39}$$

It is straightforward to verify from the recursions given in (25) and (26) that

$$k(\tilde{T}-1)k(\tilde{T}) = (p(\tilde{T}) - p(\tilde{T}-1)). \tag{40}$$

Now, the error dynamics can be compactly expressed as

$$\varepsilon(t+1) = \left(I - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)}) + \left(\frac{p(\tilde{T})}{p(\tilde{T}-1)} - 1\right)\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}^\top\right)\right)\varepsilon(t) \tag{41}$$

Finally, a direct consequence of Lemma 2 is that $\mathbb{1}^\top\varepsilon(t) = 0$ for all time, leading to the desired result. □

An immediate observation from the error dynamics is its similarity to the switched system dynamics obtained from the asymptotic algorithm described in Theorem 7. However, the same analysis used there will not be sufficient to determine if a quadratic Lyapunov function for the dynamics in (31) uniformly decreases over a certain interval. Indeed, application of the Gersgorian Disc Theorem can not guarantee that for any graph in **G** that there exists an $\alpha(t)$ that places the eigenvalues inside the unit disc.

The first step for analyzing the trajectories of (31) is to characterize the eigenvalues of the state matrix

$$A_{\varepsilon,\sigma(t)} = \left(\frac{p(\tilde{T})}{p(\tilde{T}-1)}I - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)})\right).$$

It is interesting to note the similarity between the matrix $A_{\varepsilon,\sigma(t)}$ and the matrix $A_{\sigma(t)}$ introduced in (18). In fact, the two matrices only differ by the factor $p(\tilde{T})p(\tilde{T}-1)^{-1}$ in the identity matrix. While this represents only a subtle change on the surface, the implications of this factor are sever in terms of the performance and properties of the corresponding dynamic system. It is worth to note that the ratio $p(\tilde{T})p(\tilde{T}-1)^{-1}$ is always strictly greater than one, and increases as the time-horizon $\tilde{T}$ shrinks. We analyze now the dynamics in detail.

Lemma 10. *The eigenvalues of the matrix* $A_{\varepsilon,\sigma(t)} = \left(\frac{p(\tilde{T})}{p(\tilde{T}-1)}I - \alpha(t)p(\tilde{T})L(\mathcal{G}_{\sigma(t)})\right)$ *are*

$$\Lambda_{\varepsilon,\sigma(t)} = \left\{\frac{p(\tilde{T})}{p(\tilde{T}-1)}, \frac{p(\tilde{T})}{p(\tilde{T}-1)} - \alpha(t)p(\tilde{T})\lambda_i(\mathcal{G}_{\sigma(t)}), i = 2,\dots,n\right\}. \tag{42}$$

Furthermore, if the graph $\mathcal{G}_{\sigma(t)}$ *has c connected components, then* $A_{\varepsilon,\sigma(t)}$ *has precisely c eigenvalues at* $p(\tilde{T})p(\tilde{T}-1)^{-1} > 1$.

Proof. The eigenvalues are obtained by diagonlizing $A_{\varepsilon,\sigma(t)}$. The statement on the number of eigenvalues at $p(\tilde{T})p(\tilde{T}-1)^{-1}$ is a direct consequence of properties of the graph Laplacian [3]. Finally, it can be verified from the recursion (25) that $p(\tilde{T})p^{-1}(\tilde{T}-1) \in (1, 1.2]$. □

Consider as an example a time t such that $A_{\varepsilon,\sigma(t)}$ has c eigenvalues at $p(\tilde{T})p(\tilde{T}-1)^{-1}$. An important question is to determine if it is possible to ensure with a proper choice of step-size α that the remaining $n-c$ eigenvalues are contained inside the unit disc.

Lemma 11. *Let the matrix $A_{\varepsilon,\sigma(t)}$ have c eigenvalues at $p(\tilde{T})p^{-1}(\tilde{T}-1)$. Then for all graphs $\mathcal{G}_{\sigma(t)}$ that satisfy*

$$\frac{\underline{\lambda}(\mathcal{G}_{\sigma(t)})}{\lambda_n(\mathcal{G}_{\sigma(t)})} > \frac{\phi-1}{6},$$

where $\underline{\lambda}(\mathcal{G}_{\sigma(t)})$ is the smallest non-zero eigenvalue of $\mathcal{G}_{\sigma(t)}$ and $\phi \approx \frac{1+\sqrt{5}}{2}$ is the golden ratio *there exists a constant step-size $\overline{\alpha}$ in the interval*

$$\frac{1}{3\underline{\lambda}(\mathcal{G}_{\sigma(t)})} < \overline{\alpha} < \frac{2}{(\phi-1)\lambda_n(\mathcal{G}_{\sigma(t)})},$$

that will ensure that the remaining eigenvalues are inside the unit disc, i.e.,

$$-1 < \frac{p(\tilde{T})}{p(\tilde{T}-1)} - \overline{\alpha}p(\tilde{T})\lambda_i(\mathcal{G}_{\sigma(t)}) < 1,\ i = n+c+1,\dots,n. \tag{43}$$

Proof. The recursion (25) can be used to bound $p(\tilde{T})$ as $p(\tilde{T}) \in [1/2, \phi-1)$ where $p(\tilde{T})p^{-1}(\tilde{T}-1) \in (1, 1.2]$. The bound is then obtained by substituting these least upper-bounds and greatest lower-bounds in the inequality (43). The requirement on the ratio of the smallest and largest non-zero eigenvalues is needed to ensure the interval for $\overline{\alpha}$ has an interior. □

This result has several important implications and requires some discussion. First, note that the matrix $A_{\varepsilon,\sigma(t)}$ has for any possible graph one eigenvalue at $p(\tilde{T})/p(\tilde{T}-1) > 1$. However, it can be easily shown that this eigenvalue is associated to the eigenvector $\mathbb{1}$. Since, as a direct consequence of Lemma 2, the multiplier error is always orthogonal to the all-ones vector, $\mathbb{1}^{\mathsf{T}}\varepsilon(t) = 0$, this eigenvalue outside the unit disk does not affect the dynamic evolution of the error.

However, since $p(\tilde{T})/p(\tilde{T}-1) > 1$ there are, in fact, certain graphs for which there exists *no* step-size that will place all the eigenvalues inside the unit disk. In particular, if the graph is not connected some of the eigenvalues will be placed at $p(\tilde{T})/p(\tilde{T}-1) > 1$, independent of the step-size α. Thus, the multiplier error might grow at an iteration where the communication graph is not connected. Recall that the convergence proof used for the static problem provided in Theorem 7 heavily relied on the fact that the dual error was non-increasing, even if the communication graph is not connected at some time instants. This highlights directly an important difference between an implementation of the dual sub-gradient algorithm for the static problem and the real-time implementation as proposed with the SHPA algorithm. In the SHPA algorithm, the joint connectivity assumption on the communication graph is not sufficient to guarantee a non-increase of the multiplier error.

Up until now we have focused our discussion on the performance of the multiplier error system. We must also consider how this impacts the error of our terminal state for the agents, as described in (29). Recall that the SHPA algorithm computes at each time step a prediction of the terminal state value, $\hat{\mathbf{x}}^t(T)$, and then uses the next-step optimal control to propagate the state forward. Therefore, the terminal state $x(T)$ is exactly equal to the predicted state at the last step of the algorithm, i.e., $\hat{\mathbf{x}}^{(T-1)}(T)$. This then motivates the study of the "predicted disagreement" for the system,

$$\mathbf{e}(t) = E(K_n)^\top \hat{\mathbf{x}}^{(t)}(T).$$

Note that from the discussion it is clear that $\mathbf{e}(T-1) = E(K_n)^\top x(T)$.

Theorem 12. *The predicted disagreement* $\mathbf{e}(t) = E(K_n)^\top \hat{\mathbf{x}}^t(T)$ *evolves according to the switched linear system*

$$\mathbf{e}(t+1) = \left(\frac{p(\tilde{T})}{p(\tilde{T}-1)} I - \alpha(t) p(\tilde{T}) E(K_n)^\top E(K_n) W_{\sigma(t)} \right) \mathbf{e}(t).$$

Proof. We can express $\hat{\mathbf{x}}^t(T)$ using the analytic solution provided in (15) as

$$\begin{aligned}\mathbf{e}(t) &= E(K_n)^\top k(\tilde{T})(x(t)-\xi) - p(\tilde{T})E(K_n)^\top \gamma(t) + E(K_n)^\top \xi \\ &= E(K_n)^\top k(\tilde{T})(x(t)-\xi) - p(\tilde{T})E(K_n)^\top \overline{\gamma}^{(t,x(t))} \\ &\quad - p(\tilde{T})E(K_n)^\top \varepsilon(t) + E(K_n)^\top \xi;\end{aligned} \tag{44}$$

We have substituted $\gamma(t)$ with the error expression. Note also that if the optimal multiplier $\overline{\gamma}^{(t,x(t))}$ is used to compute the state trajectories at time t, the the final consensus error will be identically zero; i.e., it is solving the centralized problem $OCP(t,T,x(t))$. Therefore, all the terms except the $\varepsilon(t)$ term will vanish, leading to the expression

$$\mathbf{e}(t) = -p(\tilde{T})E(K_n)^\top \varepsilon(t). \tag{45}$$

Propagating the state forward, and recalling that $L(\mathcal{G}_{\sigma(t)}) = E(K_n)W_{\sigma(t)}E(K_n)^\top$ leads to the desired result. □

It is interesting to observe that the dynamics for the predicted disagreement are equivalent to the dynamics of the multiplier error. The matrix $E(K_n)^\top E(K_n)W_{\sigma(t)}$ has the same non-zero eigenvalues as $L(\mathcal{G}_{\sigma(t)})$ with the remaining eigenvalues at the origin. Consequently, the analysis for this system is identical to the multiplier system. This means that for all switching signals, we can not guarantee for a switching between arbitrary graphs that the predicted error will decrease.

5 Simulation example

We illustrate the behavior of the shrinking horizon algorithm in a simulation study.

We consider a problem set-up that contains $n = 100$ agents, each starting at a random initial condition in the interval $[-20, 20]$. The preference state of each agent is chosen randomly in the interval $[-10, 10]$. The group objective is to agree on an optimal state at time $T = 20$. To negotiate the optimal meeting point, the agents perform the novel SHPA-algorithm as described in Section 3 with at constant step size $\alpha = 0.2$. The agents communicate at each time instant with randomly chosen neighbors. That is, the communication graph $\mathcal{G}_{\sigma(t)}$ is at each time instant a random graph, where an edge between two nodes is formed with probability $\mathrm{p}_c \in (0, 1)$.

Figures 1 and 2 on the next page show the trajectories of the position $x(t)$ and the multiplier error $\varepsilon(t)$, respectively, with an edge probability $\mathrm{p}_c = 0.1$. Here, the SHPA algorithm performs fairly well as online negotiation mechanism. The multiplier error $\varepsilon(t)$ is uniformly decreasing over the time horizon such that the agents reach almost perfect agreement at time $T = 20$. Note that the agents cannot reach perfect agreement since the algorithm is only performing for a finite time. This simulation suggests that the SHPA algorithm can be an efficient method for real time negotiations between dynamical systems.

The communication structure is crucial for the performance of the algorithm. To illustrate this, we consider again the same problem set-up as before, but reduce the edge probability to $\mathrm{p}_c = 0.001$. The simulation results are shown in Figures 3 and 4 on page 497. Agents are now communicating significantly less often, but the joint connectivity assumption, Assumption 1, is still satisfied if the period Δ is chosen sufficiently large. However, since the SHPA algorithm performs only on a finite time horizon the communication is not sufficient to ensure a decrease of the multiplier error, leading to a significant disagreement between the agents at the final time $T = 20$.

However, the intuitive conclusion one might draw from the previous discussion, that more communication is better, is not true in the real-time setup considered here. Figures 5 and 6 on page 498 show the trajectories of $x(t)$ and $\varepsilon(t)$, respectively, again for the same problem configuration as before but now using a higher edge probability, $\mathrm{p}_c = 0.15$. This simply means that more agents are communicating with each other at each communication round. Surprisingly at first place, this addition of communication leads to a severe decrease of the system performance. It can be clearly seen in Figure 6 on page 498 that the overall system shows an unstable behavior as the end of the time horizon in approached. In fact, the multiplier error is increasing and, consequently, the agents are not reaching agreement at the end of the time horizon.

This behavior can be well understood considering Lemma 2. While for the first set-up with an edge probability of $\mathrm{p}_c = 0.1$ the step-size $\alpha = 0.2$ is sufficiently small to keep all eigenvalues of the multiplier error dynamics inside the unit disk, this is no longer true as the edge probability is increased. Allowing more communication between the agents moves the eigenvalues of the graph Laplacian matrix and can lead to instability.

The proposed real time implementation of the dual-based negotiation mechanism, as proposed with the SHPA algorithm, can produce a behavior, which approaches the

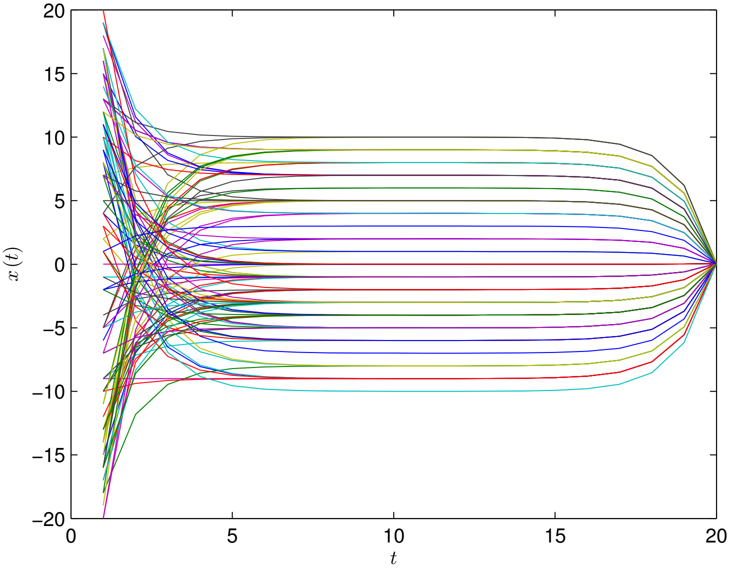


Figure 1: Trajectories of the position $x(t)$ for a network of $n = 100$ agents, constant step-size $\alpha = 0.2$, and random communication graphs with edge probability $p_c = 0.1$.

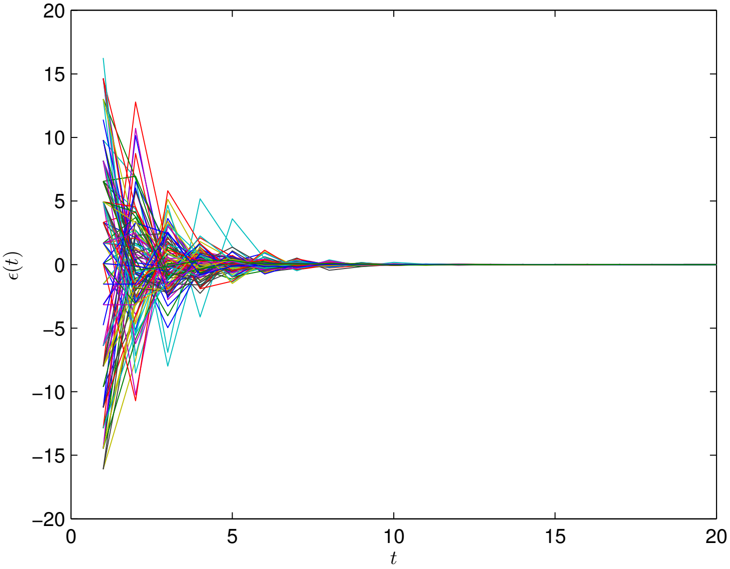


Figure 2: Trajectories of the multiplier error $\varepsilon(t)$ for a network of $n = 100$ agents, constant step-size $\alpha = 0.2$, and random communication graphs with edge probability $p_c = 0.1$ (cf. Figure 1).

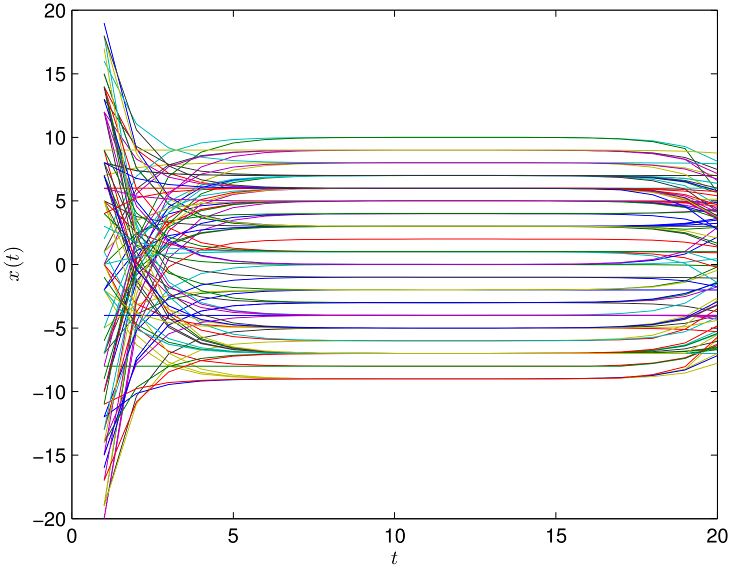


Figure 3: Trajectories of the position $x(t)$ for a network of $n = 100$ agents, constant step-size $\alpha = 0.2$, and random communication graphs with edge probability $\mathrm{p}_c = 0.001$.

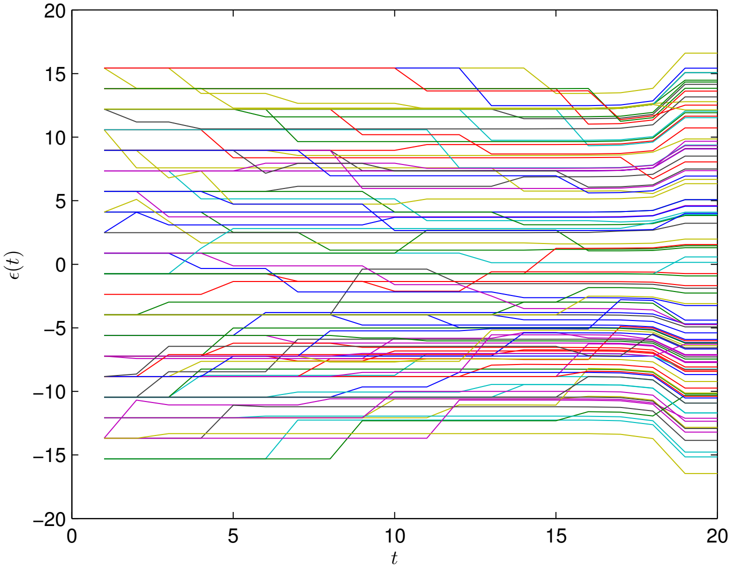


Figure 4: Trajectories of the multiplier error $\varepsilon(t)$ for a network of $n = 100$ agents, constant step-size $\alpha = 0.2$, and random communication graphs with edge probability $\mathrm{p}_c = 0.001$ (cf. Figure 3).

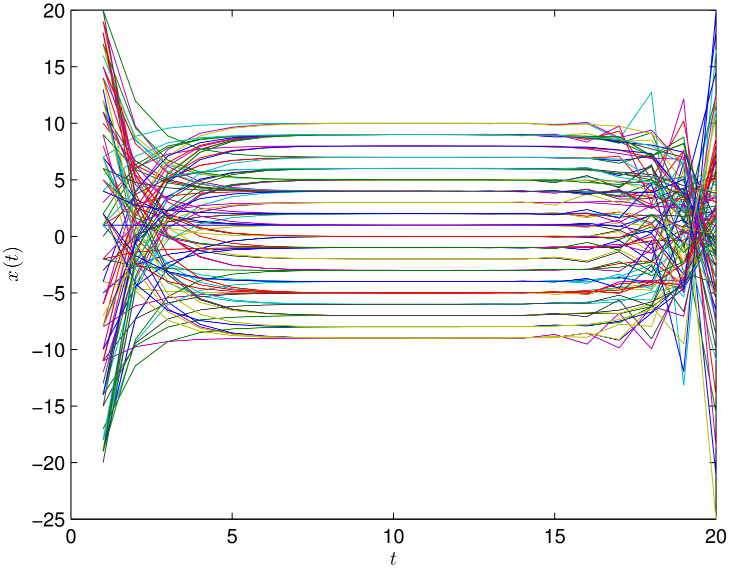


Figure 5: Trajectories of the position $x(t)$ for a network of $n = 100$ agents, constant step-size $\alpha = 0.2$, and random communication graphs with edge probability $p_c = 0.15$.

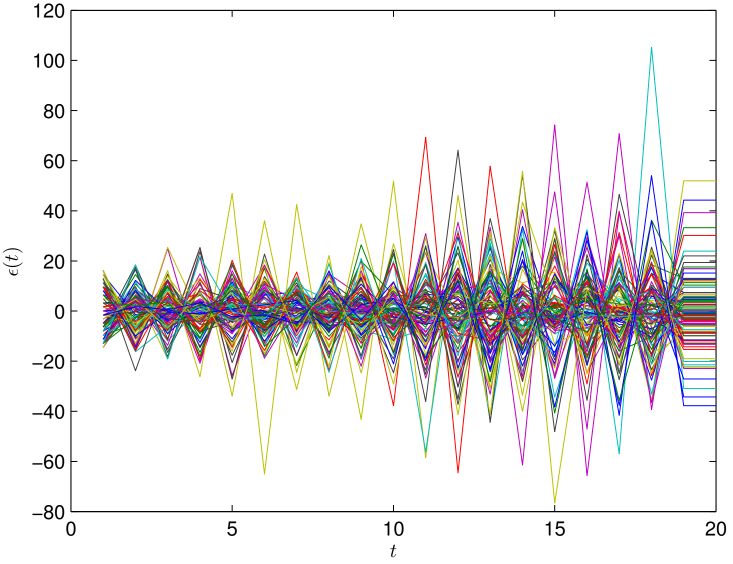


Figure 6: Trajectories of the position multiplier error $\varepsilon(t)$ for a network of $n = 100$ agents, constant step-size $\alpha = 0.2$, and random communication graphs with edge probability $p_c = 0.15$ (cf. Figure 5).

optimal solution of the original problem. However, such an implementation requires special attention and the system design has be done carefully, based on analytic considerations. In fact, already a small changes in the communication structure can lead to an unstable behavior.

6 Concluding remarks

We studied in this work the negotiation between agents in a dynamic environment. A preference agreement problem was considered, where a group of agents is required to agree on a common state exactly at a predefined time. We showed first that the dual gradient algorithm is a suitable negotiation mechanism, which allows the agents to solve the problem in a fully distributed manner. The algorithm works even if the communication is changing over time, as long as the resulting communication graph is “jointly connected" over a finite time interval.

Motivated by the observation that communication between agents requires significant time, we proposed the SHPA-algorithm as a real-time implementation of the dual sub-gradient algorithm. In this algorithm, the agents already act while they are negotiating. At each iteration an agent moves in the direction which it expects to be optimal and communicates with its neighbors to improve its estimate of the optimal solution.

We have shown that switching communication becomes a critical issue in such a dynamic realization of the negotiation mechanism. In fact, in the dual implementation the multiplier error might grow if the communication graph is not connected at a time instant. Additionally, the parameters of the algorithm must be chosen very carefully, since a badly chosen step-size can lead to instabilities of the physical process. The step-size has to be chosen, in particular, in accordance to the communication structure. We have shown that adding communication between agents can cause a former stable process to become unstable.

Bibliography

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation*. Prentice Hall, 1989. Cited p. 479.

[2] M. Bürger, D. Zelazo, and F. Allgöwer. Network clustering: A dynamical systems and saddle-point perspective. In *Proceedings of the 50th IEEE Conference on Decision and Control*, pages 7825–7830, 2011. Cited p. 479.

[3] C. D. Godsil and G. Royle. *Algebraic graph theory*. Springer, 2001. Cited pp. 481, 483, 484, 485, 486, and 493.

[4] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1991. Cited p. 485.

[5] A. Jadbabaie, J. Lin, and A. S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, 2003. Cited p. 480.

[6] B. Johansson, A. Speranzon, M. Johansson, and K. H. Johansson. On decentralized negotiation of optimal consensus. *Automatica*, 44(4):1175–1179, 2008. Cited p. 480.

[7] H. Lin and P. Antsaklis. Stability and stabilizability of switched linear systems: A survey of recent results. *IEEE Transactions on Automatic Control*, 54(2):308–322, 2009. Cited pp. 485 and 487.

[8] I. Lobel and A. Ozdaglar. Distributed subgradient methods for convex optimization over random networks. Technical report, MIT, 2009. Cited p. 480.

[9] J. Lu, C. Y. Tang, and P. R. Regier. Gossip algorithms for convex consensus optimization over networks. e-print: `http://arxiv.org/pdf/1002.2283.pdf`, 2011. Cited p. 480.

[10] M. Mesbahi and M. Egerstedt. *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010. Cited p. 479.

[11] A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009. Cited p. 479.

[12] R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007. Cited p. 479.

[13] A. Ruszczynski. *Nonlinear Optimization*. Princeton University Press, 2006. Cited pp. 482, 483, and 484.

[14] J. N. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, MIT, 1984. Cited p. 479.

[15] D. Zelazo, M. Bürger, and F. Allgöwer. A distributed real-time algorithm for preference-based agreement. In *2011 IFAC World Congress*, pages 8933–8938, 2011. Cited pp. 480 and 489.

[16] D. Zelazo, M. Bürger, and F. Allgöwer. A finite-time dual method for negotiation between dynamical systems. *SIAM Journal on Control and Optimization*, to appear. Cited pp. 480, 489, and 491.

Uwe asking a question during his Techfest.

Uwe Helmke TechFest @ MTNS 2012

A TechFest for Uwe Helmke was held as part of the 20th International Symposium on Mathematical Theory of Networks and Systems in Melbourne, Australia. The following is a list of the presentations given at the event. A reprint of Didi Hinrichsen's greeting address can be found on the next page. The TechFest concluded with a nice dinner at which Uwe was presented with a preliminary version of this Festschrift.

Grußwort (greeting address)
Didi Hinrichsen (read by Jochen Trumpf)

Subspace entropy and controlled invariant subspaces
Fritz Colonius

On the zero properties of tall linear systems with single-rate and multirate outputs
Brian Anderson, Mohsen Zamani, Giulio Bottegal

The separation principle, revisited
Tryphon Georgiou, *Anders Lindquist*

Parsimonious triggering: Lyapunov based triggering with fewer events
Fabian Wirth

Active noise control with sampled-data filtered-x adaptive algorithm
Masaaki Nagahara, Kenichi Hamaguchi, *Yutaka Yamamoto*

Lyapunov function based step size control for numerical ODE solvers
Lars Grüne, Iasson Karafyllis

Decoding of subspace codes, a problem of Schubert calculus
Joachim Rosenthal, Anna-Lena Trautmann

Double quotient structures for invariant computations
Robert Mahony, Rodolphe Sepulchre, Pierre-Antoine Absil

Optimisation geometry
Jonathan Manton

Optimization problems with matrix unknowns
Bill Helton

Linear switching systems and random products of matrices
Masaki Ogura, *Clyde Martin*

Canonical forms for pseudo-continuous multi-mode multi-dimensional systems with conservation laws
Erik Verriest

Detection of motion direction of targets using a turtle retinal patch model
Mervyn Ekanayake, *Bijoy Ghosh*

Didi Hinrichsen

Grußwort

I compliment the organizers on the idea of having a Uwe Helmke TechFest in the context of MTNS12. The MTNS12 in Melbourne provides an especially congenial setting for celebrating Uwe's 60th birthday. For many years Uwe Helmke has had a special relationship with Australia, the country, its people and its outstanding group of system theorists.

Moreover, since the organization of the MTNS93 in Regensburg (with R. Mennicken) he has been closely related to the MTNS symposia, which in my memory have always been the most pleasant and rewarding international conferences in mathematical systems theory.

Uwe Helmke is today one of the leading figures of mathematical systems theory in Germany. By his impressive mathematical culture, his wide interdisciplinary interests and scientific activities he has strongly influenced the development of dynamical systems and control theory in Germany. Under his guidance the mathematical department of Würzburg University possesses today one of the strongest research groups in mathematical systems theory and is one of the main centres of the field in the Federal Republic of Germany.

In the publications of Uwe Helmke mathematical systems theory and pure mathematics interact in a very elegant way. His results on the topology of moduli spaces of linear systems are amongst the deepest in the area. In his papers and his joint book on Optimization and Dynamical Systems with John Moore this is a characteristic feature: the brilliant interplay between mathematical systems theory and other branches of mathematics. He not only applies tools from topology, algebra, global analysis, differential and algebraic geometry to the solution of system theoretic problems, but also contributes to various fields of pure and applied/numerical mathematics making use of system theoretic concepts, ideas and results. By his work he has strongly promoted the standing of systems theory in the German mathematical community.

Altogether, Uwe Helmke's work is a beautiful illustration of what his co-author and friend Paul Fuhrmann expressed in the preface of his book on Linear Systems and Operators in Hilbert Space: "It seems to me that system theory – besides being intellectually exciting – is today one of the richest sources of ideas for the mathematician as well as a major area of application of mathematical knowledge."

Dear Uwe, I hope that you will continue to contribute to the development of mathematical systems theory with your wealth of ideas for many years to come. I wish you good health and that you remain as young in spirit as your friends know you. Enjoy this workshop in your honour, together with your many co-authors, colleagues, friends and former students! I wish all the participants a pleasant and exciting Uwe Helmke TechFest.

Didi Hinrichsen
Bremen, May 2012

List of Authors