Sparse structures and convex optimization for dynamical systems

Corbinian Schlosser

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le 26.04.2023 *2023* **par :**

Corbinian Schlosser

Sparse structures and convex optimization for dynamical systems

---

JURY

Maria Infusino	Assistant Professor	Membre du Jury
Colin Jones	Professor associé	Rapporteur
Milan Korda	Chargé de Recherche	Co-directeur de thèse
Dirk Lebiedz	Professor d'Université	Président du Jury
Christophe Prieur	Directeur de Recherche	Rapporteur
Pierre Weiss	Chargé de Recherche	Directeur de thèse

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :
Laboratoire d'Analyse et d'Architecture des Systèmes

Directeur(s) de Thèse :
Pierre Weiss et *Milan Korda*

Rapporteurs :
Colin Jones et *Christophe Prieur*

Acknowledgements

I deeply thank my two supervisors Milan and Pierre. One of their first acts as my bosses, when I just had started, proved already that they are more than supervisors but real mentors to me! Milan, thank you, for your supervision, your accessibility (either in the office or in the mountains), your rich ideas, our discussions, and your guidance throughout my three years at LAAS. You sent it, boss! Pierre, "climbing" the Mont Blanc without ever having been there was possible because you propose such curious team events. I hope we will "meet" at Mount Everest someday and maybe even add chess & cheese and wheelie practice there! And finally, I want to thank you two as a friend!

I want to express my thanks towards Colin Jones, Christophe Prieur, Dirk Lebiedz, and Maria Infusino. I am glad and honered that you accepted to take part in my defense as referees and jury members.

I am very grateful to the POEMA network for all its support, care, the workshops as well as the interesting people, projects, and topics I got in contact with thanks to the POEMA network!

My time at the office was always made lively through my colleagues and my companions on this journey. Particularly: Mathieu, voir ton attitude amicale et ouverte a été un plaisir. Je ne goûterai plus jamais le vin comme avant - même si je ne sais pas ce que je goûte exactement - et c'est à l'occasion de ton anniversaire que je me suis lancé à parler français ! Vitek, you fueled our tries of friendly trickery whenever we saw the chance (even though we missed our potentially biggest coup!). Val, pendant ces trois années, j'ai été heureux d'être ton petit frère de thèse, de nobosaurus et de la crêperie ! Tao, thanks for our friendship formed through french classes via "cake" baking, parties, movies with Viti, and dinner at Sakuraame to maybe couchsurfing someday.

I am grateful for the very versatile, interesting, and welcoming time in and off the office thanks to all of you: Adrien Doris, Adrien le Franc, Alban, Alberto, Alexey, Andries, Antonio, Baptiste, Felipe, Florent, Hoang, Isabel, Jared, Loi, Manon, Matteo de la Rosa, Matteo Tacchi, Mauro, Nicola, Nicolas, Olga, Philipp, Quentin, Rodolfo, Thong, Yoni, the whole POP and MAC team!, and Arthur and Arthur, Danish, Lucas, Ruben, Sander, Sophie and Sophie, Sven, Willem, and Guido, Daniel and Monique from CWI, and Sergiy, Jonathan, and Martin from IBM!

During my stays in Amsterdam and Dublin, I had the lasting joy of meeting Felipe, Samarth, and Christos! You contributed in an important way to how I was living in Amsterdam in Dublin, "Vamos, Chalo, Oriste!"

Isao and Masahiro, I honor it a lot that we worked together after we have met shortly in Santa Barbara several years ago. I hope we will meet again soon!

Not just during my last three years, I always felt unconditioned support and

enthusiasm from my parents. Ein unbedingtes Danke an euch!

Enfin, je tiens à remercier mes amis de l'ultimate frisbee à Toulouse – a profusion ca concerne : Alex, Elie, Samia, Steeve, Toto, Remi, Rob, Ernest, Elsa et Laurie. Vivre en France a été un voyage pour moi (et tant qu'il y a braise, c'est pas fini, et il y a braise !). Grâce à vous, Toulouse est devenue un chez moi, merci !

Abstract:
In this thesis, we describe and analyze an interplay between dynamical systems, sparse structures, convex analysis, and functional analysis. We approach global attractors through an infinite dimensional linear programming problem (LP), investigate the Koopman and Perron-Frobenius semigroups of linear operators associated with a dynamical system, and show how a certain type of sparsity induces decompositions of several objects related to the dynamical systems; this includes the global attractor as well as the Koopman and Perron-Frobenius semigroups.

The first part of this work focuses on sparsity for dynamical systems. We define a notion of subsystems of a dynamical system and present how the system can be decomposed into its subsystems. This decomposition carries over to many important objects for the dynamical system, such as the maximum invariant set, the global attractor, or the stable manifold. We present the theoretical and practical limitations of our approach. Where those limitations do not apply, we show that sparsity can be exploited for computational tasks. One example is the computation of global attractors via the two infinite dimensional LPs that we propose. For polynomial dynamical systems, we solve these LPs in an established line of reasoning via techniques from polynomial optimization resulting in a sequence of semidefinite programs. This gives rise to a sequence of outer approximations of the global attractor which converges to the global attractor with respect to Lebesgue measure discrepancy.

For the Koopman and Perron-Frobenius semigroup, sparsity induces a certain block structure of these operators. This implies a decomposition of corresponding spectral objects such as eigenfunctions and invariant measures. A direct consequence is that subsystems induce eigenfunctions for the whole system and invariant measures for the dynamical system induce invariant measures of the subsystems. However, reversing this result is less straightforward. We show that for invariant measures this problem can be answered positively under necessary compatibility assumptions and for eigenfunctions we restrict to principal eigenfunctions and assume additional regularity.

We complement the sparse investigation of Koopman and Perron-Frobenius operators with their analysis on reproducing kernel Banach spaces (RKBS). This follows and extends a path of current research that investigates reproducing kernel Hilbert spaces (RKHS) as domains for Koopman and Perron-Frobenius operators. We provide a general framework for analysis of these operators on RKBS including their basic properties concerning closedness and boundedness. More precisely, we extend basic known properties of these operators from RKHSs to RKBSs and state new results, including symmetry and sparsity concepts, on these operators on RKBS for discrete and continuous time systems.

Keywords: Dynamical system, sparsity, global attractors, polynomial optimization, semidefinite programming, Koopman operator, reproducing kernel Banach space

Résumé:
Nous décrivons et analysons une interaction entre systèmes dynamiques, structures parcimonieuses, analyse convexe et analyse fonctionnelle. Nous abordons les attracteurs globaux à travers un problème d'optimisation linéaire (OL) de dimension infinie, nous étudions les semigroupes de Koopman et de Perron-Frobenius d'opérateurs linéaires associés à un système dynamique, et nous montrons comment un certain type de parcimonie induit des décompositions de plusieurs objets liés aux systèmes dynamiques ; ceci inclut l'attracteur global ainsi que les semigroupes de Koopman et de Perron-Frobenius.

La première partie de ce travail se concentre sur la parcimonie pour les systèmes dynamiques. Nous définissons une notion de sous-systèmes d'un système dynamique et présentons comment le système peut être décomposé en ses sous-systèmes. Cette décomposition s'applique à de nombreux objets importants pour le système dynamique, tels que l'ensemble invariant maximal, l'attracteur global, ou la varieté stable stable et instable. Nous présentons les limites de notre approche d'un point de vue théorique et pratique.

Nous montrons que la parcimonie peut être exploitée pour des tâches de calcul algorithmique. Un exemple est le calcul des attracteurs globaux via les deux OL de dimension infinie que nous proposons. Pour les systèmes dynamiques polynomiaux, nous résolvons ces OLs selon un raisonnement établi via des techniques d'optimisation polynomiale, ce qui donne lieu à une séquence de programmes semi-définis. Cela occasionne une séquence d'approximations externes de l'attracteur global qui converge vers l'attracteur global en ce qui concerne la divergence de la mesure de Lebesgue.

Pour le semigroupe de Koopman et de Perron-Frobenius, la parcimonie induit une certaine structure en blocs de ces opérateurs. Cela implique une décomposition des objets spectraux correspondants tels que les fonctions propres et les mesures invariantes. Une conséquence directe est que les sous-systèmes induisent des fonctions propres pour le système entier et que les mesures invariantes pour le système dynamique induisent des mesures invariantes des sous-systèmes. Cependant, l'inversion de ce résultat est moins évidente. Nous montrons que pour les mesures invariantes, ce problème peut être résolu positivement sous les hypothèses de compatibilité nécessaires et pour les fonctions propres, nous nous limitons aux fonctions propres principales et supposons une régularité supplémentaire.

Nous complétons l'étude de parcimonie des opérateurs de Koopman et de Perron-Frobenius par leur analyse sur des espaces de Banach à noyau reproducteur (RKBS). Cela suit et étend une voie de recherche actuelle qui étudie les espaces de Hilbert à noyau reproducteur (RKHS) comme domaines pour les opérateurs de Koopman et de Perron-Frobenius. Nous fournissons un cadre général pour l'analyse de ces opérateurs sur les RKBS, y compris leurs propriétés de base concernant la continuité et la fermeture. Plus précisément,

nous étendons les propriétés de base connues de ces opérateurs des RKHS aux RKBS et nous énonçons de nouveaux résultats, y compris les concepts de symétrie et de parcimonie, sur ces opérateurs sur les RKBS pour les systèmes à temps discret et continu.

Mots clés : Systèmes dynamiques, parcimonie, attracteurs globaux, optimisation polynomiale, programmation semi-définie, opérateur de Koopman, espace de Banach à noyau reproduisant.

Contents

List of Figures

List of Algorithms

List of Acronyms

SOS	sum-of-squares
LP	linear programming problem
ILP	integer linear programming
SDP	semidefinite program
GA	global attractor
ROA	region of attraction
MPI	maximum positive invariant set
DMD	Dynamic mode decomposition
EDMD	Extended dynamic mode decomposition
s.t.	subject to

Chapter 1

Introduction

In nonlinear optimization one is often challenged with of certifying global optimality. This task is difficult and failing to answer it well can lead to getting stuck within a *local* optimum for the optimization problem. One way of circumventing that problem is constraining to instances where local information around a critical point is sufficient to certify optimality – such as in convex optimization problems. A famous subclass of convex optimization problems is the class of *linear programming problems* (LP) where even optimality bounds can be obtained via the *dual* LP. Since its invention, linear programming has enjoyed great success, a vast variety of applications and received numerous awards (such as the Nobel prize in economics for Leonid Kantorovich and Tjalling Koopmans in 1975). The access to fast solvers such as interior point methods and the simplex method, as well as the growing number of applications made LPs a desirable problem formulation. This raises the question

Which optimization problems can be formulated as LPs?

There are several answers to this question because there are as many subtleties hidden in the meaning of this question. One surprising answer is even "All!" and the quotation marks are well needed for this response. One such generic linearization technique is the following: Consider the optimization problem

$$\begin{aligned} f^* := \quad & \inf_x \quad f(x) \\ & \text{s.t.} \quad x \in K \end{aligned} \tag{1.1}$$

where K is a topological space and $f : K \to \mathbb{R}$ is a bounded Borel-measurable map. We formulate the following linear (or conic) program

$$\begin{aligned} l^* := \quad & \inf_\mu \quad \int_X f \, d\mu \\ & \text{s.t.} \quad \mu \text{ is a non-negative Borel measure on } K \\ & \qquad \mu(K) = 1. \end{aligned} \tag{1.2}$$

The condition that μ is a non-negative Borel measure on K is (convex) conic and the constraint that μ has mass 1 is an affine one. Thus, the optimization problem (1.2) is indeed a linear programming problem – but *infinite dimensional* (if K is non-finite), even if K itself is finite dimensional. And finally, both optimization problems (1.1) and (1.2) have the same optimal value, i.e.

$$l^* = f^*. \tag{1.3}$$

The claim (1.3) is easily verified. Let $x \in K$ then $f(x) = \int_K f \, d\delta_x$ where δ_x denotes the dirac measure in x. The Borel measure δ_x is feasible for (1.2) and hence $l^* \leq f^*$. On the other hand, it holds $f(x) \geq f^*$ for all $x \in K$, and by monotony of the integral we get for any non-negative measure μ on K with $\mu(K) = 1$ that

$$\int_K f \, d\mu \geq \int_K f^* \, d\mu = f^* \int_K d\mu = f^*,$$

i.e. $f^* \leq l^*$.

This approach is generic and does not take any specific characteristics of the optimization problem (1.1) into account. Therefore the reformulation (1.2) does not provide strong insight into the problem, but if any, its strength lies in formulating the problem in a different language where different techniques, i.e. linear ones, are applicable.

In order to overcome the limitations that come along with a generic approach, an LP formulation of the problem should be adapted to problem-intrinsic properties, such as regularity and its geometry. The advantages and obstacles of choosing a linear formulation adapted to the problem can be well distinguished through some historical turning and crossing points of linear programming and complexity theory. Linear programming arose in the 1940s and 1950s through work of Kantorovich and Koopmans on combinatorial transport problems. This was only the starting point of reformulating about all combinatorial optimization problems as integer linear programming (ILP) problems. Relaxing the ILPs, by removing the constraint of the decision variables being integers, leads to LP. This approach was further vitalized when Dantzig developed the simplex[1] algorithm in 1947. The simplex algorithm enjoys strong practical success but does not have polynomial time worst-case complexity bounds. This situation changed with the invention of the ellipsoid method in the 1970s. The ellipsoid method provides a polynomial time algorithm for solving LPs and we might be tempted to ask if applying the ellipsoid method to LP relaxations of the ILP formulations of hard combinatorial problems results in P = NP (where we omit explanation of the complexity classes P and NP, due to the great popularity the problem). There are two reasons (and both of them will reappear in different outfits later in the text) why we need to be careful

1. Typically the LP relaxations for (hard) combinatorial optimization problems are strict relaxations of the combinatorial problem and do not solve them exactly.

2. The LP relaxation can be of exponential size compared to the original problem (for example the Dantzig–Fulkerson–Johnson formulation for the travelling salesman problem).

The second point can be interpreted geometrically, namely, an LP formulation represents a polyhedral embedding of the combinatorial problem and this polyhedral

[1]The name "simplex algorithm" for Dantzig's algorithm was suggested by Theodore Samuel Motzkin who gave the first explicit example of a non-negative polynomial that is not a sum-of-squares polynomial. We state his example in (2.32).

embedding might describe a very complex geometric object. From this perspective, the first point reads that the "combinatorial quality" of the constraints in the LP relaxation is not strong enough to represent the combinatorial problem exactly. The question of whether an LP relaxation does or does not induce a strict relaxation of the original problem is present in many parts of this thesis – most vividly in Chapter 5.

As an illustration of an LP relaxation, we borrow the example of matchings in graphs. Let $G = (V, E)$ be a graph with nodes V and edges $E \subset \{\{v_1, v_2\} : v_1, v_2 \in V\}$ and $w : E \to \mathbb{R}_+$ be a weight function for the edges. A matching $M \subset E$ in G is a collection of disjoint edges in E, i.e. the edges in M do not share nodes. The weight of a matching $M \subset E$ is defined as

$$w(M) := \sum_{e \in M} w(e).$$

A maximal matching M^* is a matching M^* that has maximal weight among all matchings, i.e.

$$w(M^*) = \max\{w(M) : M \text{ matching in } G\}.$$

In Figure 1.1 we illustrate a simple example of a graph with weighted edges and a maximal matching.

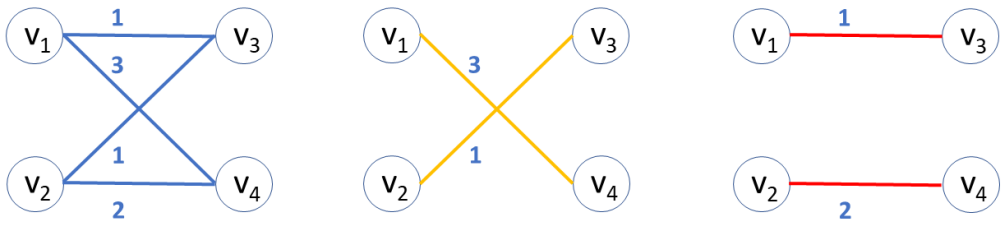

Figure 1.1: Example of a graph with weighted edges (left), a maximal matching (middle) and a non-maximal matching (right).

For the ILP formulation for the problem of finding a maximal matching, we associate a matching M with a vector $x^M = (x_e^M)_{e \in E} \in \{0, 1\}^E$ with $x_e^M = 1$ if $e \in M$ and $x_e^M = 0$ if $e \notin M$. If we write $\omega := (w(e))_{e \in E}$ we can express the weight $w(M)$ of a matching M by

$$w(M) = \sum_{e \in M} w(e) = \sum_{e \in E} x_e^M w(e).$$

The condition that M is a matching can be rephrased for x^M by the condition

$$\sum_{e = \{v, w\} \in M} x_e \leq 1 \text{ for all nodes } v \in V.$$

Thus, the ILP for the maximal matching problem reads

$$w(M^*) = \begin{array}{ll} \max\limits_{x=(x_e)_{e\in E}\in\{0,1\}^E} & \sum\limits_{e\in E} x_e^M w(e) \\ \text{s.t.} & \sum\limits_{e\in M: v\in e} x_e \leq 1 \quad \text{for all nodes } v \in V. \end{array} \tag{1.4}$$

For the LP relaxation of the ILP (1.4) we replace the constraint $x \in \{0,1\}^E$ be $x \in [0,1]^E$. The LP relaxation reads

$$m^* := \begin{array}{ll} \max\limits_{x\in\mathbb{R}^E} & \langle \omega, x^M \rangle \\ \text{s.t.} & 0 \leq x \leq 1 \\ & Ax \leq b, \end{array} \tag{1.5}$$

where $\langle \cdot, \cdot \rangle$ denotes the euclidean inner product on $\mathbb{R}^n$, the inequalities are understood componentwise, $A \in \mathbb{R}^{V\times E}$ denotes the incidence matrix of the graph G and $b \in \mathbb{R}^v$ the vector with all components equal to 1.

Among the observations and questions concerning the LP (1.5) which catch the eye are the following:

- Dimensionality: The space for the decision variables is $\mathbb{R}^E$, i.e. one dimension for each of the edges that we could choose in the matching problem.
- Relaxation: The LP (1.5) is a relaxation of the matching problem, i.e. m^* is an upper bound for the optimal value of the matching problem.
- Exploiting linearity: How do we make use of the linear structure?
- Duality: How does the dual problem to (1.5) look like and how should it be interpreted?
- Reconstruction: Given a solution x of (1.5) how do we reconstruct a (corresponding) matching?
- Structure exploitation: There are structures that simplify formulating and solving the matching problem. How do they look like and can they be recognized from the LP?

The above remarks indicate a guideline through this thesis and can be recognized, clothed in different notions, at many points in this text. To use this guideline we first have to adapt our viewpoint to dynamical systems. So far we have only discussed static optimization problems and their LP (relaxation) while this thesis focuses on dynamical systems. Indeed, LP formulations for certain problems from dynamical systems – or more generally a linear representation of the dynamical system itself – are possible. Our work related to linear reformulations for dynamical systems is based on two methods, namely the use of so-called occupation measures and Koopman theory. I would like to express the surprise and joy I had when I realized some of the many parallels between the LP approach towards combinatorial optimization problems and Koopman (and occupation measure) analysis for dynamical systems. It is an amusing fact that this relation seemingly continues

even to the names of some central creators of the corresponding fields – Tjalling Koopmans and Bernard Osgood Koopman.

For a brief comparison between the LP relaxation approach for combinatorial optimization and Koopman analysis for dynamical systems, we should first specify the latter. A dynamical system is a pair $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ consisting of a set X and a family $(\varphi_t)_{t\in\mathbb{R}_+}$ of maps $\varphi_t : X \to X$ that satisfy the semiflow property

$$\varphi_{t+s} = \varphi_t \circ \varphi_s \text{ for all } t, s \in \mathbb{R}_+. \tag{1.6}$$

For a dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ we can define a corresponding family $(T_t)_{t\in\mathbb{R}_+}$ of Koopman operators T_t for $t \in \mathbb{R}_+$. The Koopman operators mimic what we have seen for the matching problem: In the matching problem, the vector $x^M \in \mathbb{R}^E$ observes whether an edge e is chosen or not in a matching. The Koopman operators also act indirectly via observables. For each function $g : X \to \mathbb{R}$ the Koopman operator T_t at time $t \in \mathbb{R}_+$ is defined as

$$T_t g := g \circ \varphi_t \tag{1.7}$$

and hence is *linear* and represents the evolution of the observation g. Sometimes this procedure is referred to as lifting and the space where the Koopman operator acts is called the lifting space. Through the lifting the semiflow property (1.6) of φ translates into the following semigroup law for the Koopman operators

$$T_{t+s} = T_t T_s \text{ for all } t, s \in \mathbb{R}_+. \tag{1.8}$$

The semigroup property (1.8) reminds strongly of exponential maps e^{tA} arising from solutions of linear differential equations

$$\dot{x} = Ax,\ x(0) = x_0 \in \mathbb{R}^n \tag{1.9}$$

for a given matrix $A \in \mathbb{R}^{n\times n}$. In some sense, see [Engel 2006], such a relation is true also for the Koopman semigroup $(T_t)_{t\in\mathbb{R}_+}$, namely it holds for $t \in \mathbb{R}_+$

$$\frac{\mathrm{d}}{\mathrm{dt}} T_t g = \mathscr{A}(T_t g) \tag{1.10}$$

for a linear operator $\mathscr{A}$ with

$$\mathscr{A} g(\cdot) := \frac{\mathrm{d}}{\mathrm{dt}}\bigg|_{t=0} g(\varphi_t(\cdot)) \tag{1.11}$$

for functions $g \in \mathcal{C}(X)$ for which the limit (1.11) exists. The relation (1.10) unveils the full linear nature of the Koopman perspective on dynamical systems.

As promised, we will now emphasize striking parallels between linearizations for static combinatorial problems, at the example of the matching problem, and the "Koopman linearization" for dynamical systems. Such a comparison should begin with the simplest characteristic – the dimension of the lifting space.

Dimension of the lifting space: The idea of lifting is to shift the complexity of the problem into the lifting space, therefore it should be expected that the lifting space is of higher dimension than the original one.

Graph matching

For the graph matching problem we lifted the problem into $\mathbb{R}^E$.

Koopman operator

The space of all functions $g : X \to \mathbb{R}$ is infinite dimensional whenever X is not finite.

Relaxation error: This is a central question and relevant throughout this thesis. The answer is not simple and depends on the formulation and the problem at hand.

For the graph matching problem, the LP (1.5) is indeed a relaxation, see for instance [Schrijver 2003, Section 18.1]. For certain classes of graphs, there is no relaxation gap, i.e. the LP relaxation solves the original problem exactly. Whether or not there is a relaxation is determined by the geometry of the feasible set $\{x \in \mathbb{R}^E : 0 \leq x \leq , Ax \leq b\}$ for (1.5). For instance, if the graph is bipartite then all extremal points of that set are integer, i.e. correspond to a matching and thus there is no relaxation gap! The one who showed this first was Birkhoff who also proved influential results on Koopman theory!

For the Koopman operator from (1.7) on the space of all functions from X to $\mathbb{R}$ the semiflow $(\varphi_t)_{t\in\mathbb{R}_+}$ is uniquely determined by the family $(T_t)_{t\in\mathbb{R}_+}$. As we will address later in Chapter 6 the space of all functions from X to $\mathbb{R}$ is not always (computationally) practical and more appropriate choices should be made. For some of those choices, it has to be carefully investigated if the Koopman operator is well-defined on a large enough set of observables.

How do we make use of the linear structure? One way is by using algorithms or concepts particularly adapted to linear problems.

An alogrithm particularly suited for LPs is the simplex algorithm. Another important role for LPs is played by the dual LP – which is addressed next.

For the Koopman operator exploiting the linear structure manifests in employing semigroup and spectral theory (see for instance Theorem 2.47).

What is the dual? Duality will play an important role in several parts of this thesis. For linear programs, the duality is expressed through the dual LP and for operators we consider the adjoint operator.

In case of the matching problem, duality allows a short proof of König's celebrated matching theorem. Königs matching theorem states that *the maximum size of a matching in a bipartite graph is the minimum size of a vertex cover*[1]. The argument for König's matching theorem can be performed using duality [Schrijver 2003]: Let all $\omega = \mathbf{1}$, i.e. all weights $w(e)$ equal one. By Birkhoff's Theorem the optimal value m^* of LP (1.5) coincides with the maximum size of a matching. By strong duality for (finite dimensional) linear programming we obtain

$$m^* = \min\{\langle y, \mathbf{1}\rangle : y \geq 0,\ y^T A \geq \mathbf{1}\}$$

where $\mathbf{1} \in \mathbb{R}^V$ is the vector with all entries equal to one and $A^T \in \mathbb{R}^{E\times V}$ the adjoint of A. A feasible point $y = (y_v)_{v\in V} \in \mathbb{R}^V$ for the dual is interpreted as choosing a node v if $y_v > 0$. As for the primal problem, the extremal points for the feasible set of the dual problem are integer and we get $y \in \{0,1\}^V$ for any optimal y. The constraint $y^T A \geq \mathbf{1}$ represents that the set of nodes $\{v : y_v = 1\}$ is a node cover and thus it follows König's matching theorem.

The adjoint of the Koopman operator is called the Perron-Frobenius operator. This operator is the central object in Section 6.1 and describes the evolution of measures driven by the flow. This can be interpreted as the evolution of particle distributions. Depending on whether the evolution of observables or particle distributions is accessible from the data it is the Koopman operator or the Perron-Frobenius operator that is more suited. In many applications, Koopman and Perron-Frobenius theory is used to estimate the future state of the system based on sample trajectories. The most popular method for the estimation is the so-called dynamic mode decomposition (DMD) for the Koopman operator, which exists also for the Perron-Frobenius operator, where it is sometimes called kernel DMD.

How do we relate solutions of the lifted problem/system to objects of the original problem/system? This question essentially asks how to undo the lifting or what can be inferred for the original problem/system from its lift. Due to the nature of the lifts that we consider here, the interpretations are as natural as the lift itself.

From the way we motivated the lift a solution $x \in \mathbb{R}^E$ of the LP (1.5) is interpreted as a matching using the edges $M := \{e \in E : x_e = 1\}$. For bipartite graphs, we have seen that the dual

One way of reconstructing $\varphi_t(x)$ from T_t is choosing a set of observables $g_1, \ldots, g_n : X \to \mathbb{R}$ such that $g := (g_1, \ldots, g_n) : X \to \mathbb{R}^n$ is bijective (with inverse g^{-1}) because then $\varphi_t(x)$ is given

[1] A vertex cover in a graph is a set of vertices C such that each edge in the graph is adjacent to at least one of the nodes in C.

problem can be related to vertex covers.

by
$g^{-1}(T_t g_1(x), \ldots, T_t g_n(x))$.
This is the underlying idea of the celebrated extended dynamic mode decomposition (EDMD).

How do we exploit inherent structure, such as sparsity or symmetry? How do they translate to the lifted formulation? This question is motivated by the computation of solutions for our problems. Despite the linear structure of the lifted problem/system, computations can get intractable for large or even medium-sized problem instances, simply due to the high dimensional nature of the linear reformulation/relaxation. Inherent structures such as symmetries or sparsity should not be overseen in the process of solving, neither in the original nor the linear problem formulation. In this thesis the focus is on sparse structures – in our context sparsity means that there are independent substructures inherent in the problem instance.

For graphs it is clear that the matching problem can be decomposed into matching problems on each of the connected components of the graph. From the linearization (1.5) this can be inferred via a block structure in the incidence matrix A and vice versa.

In the dynamical setting, such decompositions exist as well. Based on a similar idea of disconnected parts of the dynamical system we will first define the notion of subsystems for the dynamical systems based. For graphs, our notion of subsystems corresponds to a refined notion of connected components.

In contrast to the order in which we presented a comparison (or rather the analogies) between the LP formulation procedure for combinatorial problems and Koopman lifting for dynamical systems, we will begin this thesis by presenting decompositions of dynamical systems based on certain sparse structures. In the second chapter, we apply lifting via the so-called occupation measures to the problem of computing the global attractor for a dynamical system and we show that this approach benefits from the sparse decomposition described in Chapter 4 . Sparsity also translates to Koopman and Perron-Frobenius operators. This is included in Chapter 6 in which we also investigate reproducing kernel Banach spaces as underlying function space for Koopman and Perron-Frobenius operators, hinting at a specific approach towards adapted choices of observables.

Chapter 2

Preliminaries

This chapter will cover important definitions, essential objects, and helpful results that we will work with in the remainder of the text. At the same time, I want to give a flavor of the re-appearing concept of embracing linear representations of non-linear tasks. Those small detours in this chapter are not always fundamental to this thesis, but interesting on their own and hopefully can highlight that concept from different angles.

In this thesis, the problems at hand arise from dynamical systems and the presented *linear* representations are achieved via a natural lift into an infinite dimensional framework. This thesis includes two approaches: Occupation measure formulations for set approximation, and Perron-Frobenius respectively Koopman lifts of the dynamical system.

In the case of the occupation measure approach, a trajectory is associated with a certain measure – its corresponding occupation measure. Therefore, we treat the space $\mathcal{M}(X)$ of Borel measures X and its pre-dual, the space of continuous functions, in Section 2.2 where we also illustrate their duality. Further on that line, methods from polynomial optimization, in the language of real algebraic geometry, are borrowed. Those concepts allow one to replace positivity constraints for polynomials by sum-of-squares-certificates and we state some of the related results in Section 2.4.

In Section 2.6 we provide the necessary material for the investigation of the Perron-Frobenius respectively Koopman operator on reproducing kernel Banach spaces in Chapter 6. This includes giving a definition of a reproducing kernel Banach space and stating important duality concepts in that setting.

The other central part of this thesis, the treatment of sparsity for dynamical systems does not require prior concepts. Therefore, no such section is included in the preliminaries.

2.1 Dynamical systems

Dynamical systems theory builds the groundwork for most topics in this thesis. The presented material on dynamical systems is standard and can be found in any classical text on dynamical systems and/or differential equations. Among the numerous books on dynamical systems that treat the concepts that are mentioned in this section in more detail are [Perko 2013, Meiss 2007, Anosov 1988, Teschl 2012, Bhatia 2006].

We work with topological dynamical systems, that is, the underlying state space X is always a topological space. In most of the cases, X will be a subset of $\mathbb{R}^n$ equipped with the relative topology inherited from the euclidean topology on $\mathbb{R}^n$.

Definition 2.1 (Dynamical system)**.** *A dynamical system is a pair $(X, (\varphi_t)_{t\in G})$ of a topological space X based and a topological (semi)group G, such that the following properties are satisfied*

i. φ_t is a function from X to X for all $t \in G$.

ii. Normalization

$$\varphi_0 = \mathrm{Id}, \text{ i.e. } \varphi_0(x) = x \text{ for all } x \in X. \tag{2.1}$$

iii. (Semi) flow property

$$\varphi_{t+s} = \varphi_t \circ \varphi_s \text{ for all } t, s \in G. \tag{2.2}$$

iv. Continuity

$$\varphi : G \times X \to X \text{ is continuous.} \tag{2.3}$$

We call a dynamical system $(X, (\varphi_t)_{t\in G})$ discrete if G is discrete.

In this work G represents time and therefore we treat only the case of $G = \mathbb{R}$ being the real numbers or $G = \mathbb{R}_+$ being the non-negative real numbers, respectively, their discrete analogs $G = \mathbb{Z}$ or $G = \mathbb{N}$. In case of discrete dynamical systems, the time-one map φ_1 contains all the information about the system and is typically denoted by $f : X \to X$, i.e. we consider the system

$$x_{k+1} = f(x_k),\ x_0 \in X.$$

The reason that it's sufficient to know the time-one map f in order to know the whole dynamical system $(X, (\varphi_n)_{n\in\mathbb{N}})$ is simply that $1 \in \mathbb{N}$ (additively) generates $\mathbb{N}$ respectively $\mathbb{Z}$ by $n = \underbrace{1 + \ldots + 1}_{n \text{ times}}$. For a dynamical system that means

$$\varphi_n(x) = \underbrace{\varphi_1 \circ \cdots \circ \varphi_1}_{n \text{ times}}(x) = (f \circ \cdots \circ f)(x) =: f^n(x).$$

In the case of continuous time dynamics, the situation is different, there is no point $r \in \mathbb{R}_+$ with the property $rN := \{n \cdot r : n \in \mathbb{N}\} = \mathbb{R}_+$. The semigroup $\mathbb{R}_+$ is not generated by a single element. Nevertheless, the smaller $r \in (0, \infty)$ the "denser" is its generated set $r\mathbb{N}$ in $\mathbb{R}_+$. That motivates the idea of defining a generating element for the dynamical system via a limit object and leads to an intimate connection between dynamical systems and ordinary differential equations.

Consider the ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{dt}}x(t) = f(x(t)) \ \ , x(0) = x_0 \in X, \tag{2.4}$$

where X is a compact subset of $\mathbb{R}^n$ and $f : X \to \mathbb{R}^n$ a smooth vector field. In the following, we will mostly write $\dot{x}(t)$ for $\frac{\mathrm{d}}{\mathrm{dt}}x(t)$ and omit the explicit time dependence whenever this is clear from the context. If X is positively invariant for the differential equation (2.4), we get a dynamical system on X via the map φ where

for $t \in \mathbb{R}_+$ the map $\varphi_t : X \to X$ is given by

$$\varphi_t(x_0) := \text{Solution of (2.4) with initial value } x_0 \text{ at time } t. \tag{2.5}$$

That φ satisfies the properties demanded in Definition 2.1 follows from f being a vector field (this implies *i.*), uniqueness of solutions by the Picard-Lindelöf theorem (this implies *ii.* and *iii.*) and the Gronwall Lemma (this implies *iv.*). This can be found in [Meiss 2007, Robinson 2003, Lee 2003, Perko 2013] or any classical text on dynamical systems and manifold theory. For dynamical systems on sets $X \subset \mathbb{R}^n$, where the map $t \mapsto \varphi_t(x)$ is differentiable for all x, we can define an underlying differential equation to the dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$. For $x_0 \in X$ we set $f(x_0)$ to

$$f(x_0) := \frac{\mathrm{d}}{\mathrm{dt}}\bigg|_{t=0} \varphi_t(x_0) \tag{2.6}$$

for $x_0 \in X$. Then $t \mapsto \varphi_t(x_0)$ is the unique solution to the initial value problem $\frac{\mathrm{d}}{\mathrm{dt}}x(t) = f(x(t))$ with $x(0) = x_0$.

Along with the definition of a dynamical system come very natural objects and notions concerning the orbit or trajectory of a solution and invariance.

Definition 2.2. *Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system. A set $A \subset X$ is called*

i. The orbit (or trajectory) of a point $x_0 \in X$ if $A = \{\varphi_t(x_0) : t \in \mathbb{R}_+\}$.

ii. Positively invariant if $\varphi_t(A) \subset A$ for all $t \in \mathbb{R}_+$.

iii. Invariant if $\varphi_t(A) = A$ for all $t \in \mathbb{R}_+$.

Invariance of a set A states that the properties specifying the set A are maintained throughout the evolution of the system. Without any doubt, invariance is therefore a fundamental question for dynamical systems and is of particular importance when investigating the asymptotic behavior of the dynamics (see Theorem 2.7). Longtime behavior of dynamical systems treats the evolution of the dynamical system for large $t \in \mathbb{R}$, more precisely, the limit case where t tends to infinity. The classical notions of (Lyapunov) stability and attractiveness are presented now.

Definition 2.3. *Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system. A set $A \subset X$ is called*

1. *stable if for each neighbourhood U of A there exists another neighbourhood V of A such that*
$$\varphi_t(V) \subset U \text{ for all } t \in \mathbb{R}_+, \tag{2.7}$$

2. *attractive if for all $x \in X$ and all neighbourhoods U of A there exists a time $T = T(x) \in \mathbb{R}_+$ such that $\varphi_t(x) \in U$ for all $t \geq T$. The set A is called uniformly attractive if T can be chosen uniformly in x,*

3. *asymptotically stable, if M is stable and attractive,*

4. *the basin of attraction of a set $B \subset X$ if*
$$A = \{x \in X : \varphi_t(x) \to B \text{ as } t \to \infty\}, \tag{2.8}$$

where $\varphi_t(x) \to B$ as $t \to \infty$ means that for any neighbourhood U of B there exists $T \in \mathbb{R}_+$ such that for all $t \geq T$ it holds $\varphi_t(x) \in U$.

There are several different notions for attractiveness in dynamical systems dating back to Lyapunov (whose notion we use in this work), Birkhoff and Milnor (see for instance [Kühner 2021]), among others. In all cases, an outstanding role is played the smallest among all attractive sets – the attractor – because it gives the best comprehension of how the system behaves asymptotically.

Definition 2.4 (Global attractor)**.** *A compact set $\mathcal{A} \subset X$ is called a global attractor if it is the minimal uniformly attracting set, i.e., it is the smallest compact set $\mathcal{A}$ that is uniformly attractive.*

Remark 2.5. *Despite a related definition, global attractors should not be confused with another common concept of attractors – the weak attractor. A weak attractor is the smallest closed set that attracts each trajectory (but not uniformly). Both, weak and global attractors can have striking differences, for instance, the global attractor for the differential equation $\dot{x} = -x$ on $\mathbb{R}^n$ is empty, while the weak attractor is the origin. Another example that illustrates many differences is a dynamical system that is given by a heteroclinic orbit, connecting an unstable equilibrium point x_0, with a stable equilibrium point x_1. As an example, consider the following differential equation*

$$\dot{x} = (x+1)(1-x), \quad x \in X := [-2, 2]. \tag{2.9}$$

The weak attractor $\mathcal{A}_\mathrm{w}$ is given by $\mathcal{A}_\mathrm{w} = \{-1, 1\}$ while the global attractor $\mathcal{A}$ is given by $\mathcal{A} = [-1, 1]$ (the choice $X = [-2, 2]$ is rather arbitrary, the only importance for this example is that X contains the interval $[-1, 1]$).

The following remark shows why interest in the global attractor is highly justified.

Remark 2.6. *The global attractor has the following properties.*

1. *Stability: The global attractor is stable [Robinson 2003] and therefore also asympotically stable. The weak attractor not necessarily stable. An example of an unstable weak attractor is a heteroclinic orbit as in (2.9).*

2. *"Attractors approximate trajectories" [Robinson 2003, p. 276]: Let X be a compact metric space with metric d. Then for all $x \in X$, $T > 0$ and $\varepsilon > 0$. For a global attractor $\mathcal{A}$, there exists $t_0 = t_0(T, \varepsilon) \in \mathbb{R}_+$ (independent of x!) and $x_0 = x_0(x, T, \varepsilon) \in \mathcal{A}$ such that*

   $$d(\varphi_t(x), \varphi_t(x_0)) < \varepsilon \text{ for all } t \in [t_0, t_0 + T].$$

   For weak attractors the time t_0 can not be chosen uniformly in x. Again, a simple example is given by a heteroclinic orbit as in (2.9).

3. *Continuity: The global attractor is upper semicontinuous (see [Robinson 2003, p. 278] for the result and related definitions). That means small changes in the vector field can not cause a drastic increase in the global attractor. A*

system where a spiraling trajectory turns into a periodic orbit under a small perturbation is described in [Robinson 2003, p. 267]. Thus, weak attractors do not enjoy upper semicontinuity.

A natural question that remains is whether the global attractor exists. In the case where X is compact, the answer is yes. And we can say even more.

Theorem 2.7 ([Robinson 2003, Chapter 10])**.** *Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system and X be compact. Then*

1. *The global attractor exists and is non-empty if X is non-empty.*
2. *The global attractor is the largest compact invariant set $A \subset X$.*
3. *If φ_t is injective for all $t \in \mathbb{R}_+$ then the dynamical system is invertible on the global attractor $\mathcal{A}$, i.e. for all $x \in \mathcal{A}$ we can define $\varphi_t(x)$ for all $t \in \mathbb{R}$ such that the flow property (2.2) is satisfied on $\mathbb{R}$. Furthermore, φ is continuous in (t, x) for all $(t, x) \in \mathbb{R} \times \mathcal{A}$.*

Next, we connect the attractor with the classical concept of stability via Lyapunov functions. We will view it from a lifting procedure and we will meet that perspective again in Chapter 5.

Definition 2.8. *For a dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ we call a continuous function $g : X \to \mathbb{R}$*

1. *a (strict) Lyapunov function if $g(x) \geq 0$ for all $x \in X$ and $g\circ\varphi_t(x)$ is (strictly) decreasing in t whenever $g(x) \neq 0$.*
2. *a Hamilton function if $g(\varphi_t(x)) = g(x)$ for all $x \in X$ for all $t \in \mathbb{R}_+$.*

Level sets of Hamilton functions provide a partition of the state space X into positively invariant disjoint sets. Because the intersection of two positively invariant sets is again an invariant set, level sets of different Hamilton functions can be intersected to gain finer partitions of X. It was shown in [Mezić 1999, Mezić 2005, Küster 2021] that for measure preserving systems an ergodic partition, i.e. the finest partition into invariant sets, can be obtained in this way. To do so, the authors in [Mezić 1999, Mezić 2005, Küster 2021] started from the reformulation that Hamilton functions are eigenvectors of the Koopman operator (see Chapter 6) with eigenvalue 1. That makes available functional analytic decomposition of the Koopman operator [Küster 2021] and arguments from ergodic theory [Mezić 1999, Mezić 2005]. This allows for finding a rich set of Hamilton functions in this linear setting.

A similar observation and strategy appeals for Lyapunov functions. First, it follows from the definition of a Lyapunov function that its sublevel sets are positively invariant and, further, that the global attractor is contained in its zero-level set. The second claim is stated in the following known result.

Lemma 2.9. *Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system with global attractor $\mathcal{A}$ and V a strict Lyapunov function. Then $\mathcal{A} \subset V^{-1}(\{0\})$.*

Proof. Let $c := \max_{x \in \mathcal{A}} g(x)$. By compactness of $\mathcal{A}$ we can find $z \in \mathcal{A}$ with $V(z) = c$. By Theorem 2.7 2. there exists $y \in \mathcal{A}$ with $\varphi_1(y) = z$. Since V is a Lyapunov function it holds

$$0 \leq V(z) = V(\varphi_1(y)) \leq V(y). \tag{2.10}$$

Because V is a strict Lyapunov function, the inequality in (2.10) is strict if $V(y) > 0$. In that case we get $V(y) > V(z) = c = \max_{x \in \mathcal{A}} V(x)$, which is in conflict with $y \in \mathcal{A}$. We conclude $0 = V(z) = \max_{x \in \mathcal{A}} V(x)$. Because V is non-negative the statement follows. □

The connection between Lyapunov functions and global attractors is even mutual, as the following theorem shows.

Theorem 2.10 ([Bhatia 2006, Theorem 2.7.1 and Remark 2.7.22])**.** *Let $(X, (\varphi_t)_{t \in \mathbb{R}_+})$ be a dynamical system with global attractor $\mathcal{A}$. A closed subset $M \subset X$ is asympotically stable if and only if there exists a strict Lyapunov function V with*

$$V(x) = 0 \text{ if and only if } x \in M. \tag{2.11}$$

If desired, V can be chosen such that

$$V(\varphi_t(x)) \leq e^{-t} V(x) \text{ for all } x \in X. \tag{2.12}$$

In particular, the global attractor $\mathcal{A}$ is the smallest compact set M for which there exists a strict Lyapunov function V satisfying $V^{-1}(\{0\}) = M$.

This makes it possible to search for attractors through searching for Lyapunov functions. To have a handier way of searching for a Lyapunov function we turn back to the case where X is a smooth submanifold of $\mathbb{R}^n$ and the dynamical system induced by a vector field $f : \mathbb{R}^n \to \mathbb{R}^n$. If the function V is smooth then condition (2.12) is equivalent to

$$V(x) \geq 0 \text{ and } \nabla V(x) \cdot f(x) \leq -\beta V(x) \quad \text{for all } x \in X \tag{2.13}$$

where ∇V denotes the gradient of V and $\nabla V \cdot f$ its pointwise (euclidean) inner product with the vector field f. That we can always find smooth Lyapunov functions is central for the arguments in Section 5.2 and guaranteed by the following inverse Lyapunov theorem.

Theorem 2.11 ([Teel 2000])**.** *Let $X \subset \mathbb{R}^n$ be open and bounded and $(X, (\varphi_t)_{t \in \mathbb{R}_+})$ be a dynamical system induced by a vector field f on $\mathbb{R}^n$. Let $\mathcal{A}$ be the global attractor for that system. Then there exists a Lyapunov function $V \in \mathcal{C}^\infty(X)$ with $V^{-1}(\{0\}) = \mathcal{A}$ that satisfies (2.13).*

We want to introduce the following optimization perspective on Lyapunov functions which we will encounter again in Chapter 5. After rearranging (2.13) reads

$$V \geq 0 \text{ and } -\nabla V \cdot f - \beta V \geq 0 \quad \text{on } X. \tag{2.14}$$

This is now a conic constraint in V in the space of continuously differentiable functions on X! We consider the order cone induced by the natural order $h_1 \leq h_2$ for functions $h_1, h_2 \in \mathcal{C}(X)$ if $h_1(x) \leq h_2(x)$ for all $x \in X$. This observation was the starting point for many important applications in control theory and flourished further through the influential thesis of Parrilo [Parrilo 2000]. Parrilo considered sum-of-squares certificates for verifying the non-negativity constraints in (2.14) – an approach that will appear at many points in this thesis.

2.2 The space of continuous functions and its dual

For a compact set X, the space of continuous functions is denoted by $C(X)$ and equipped with the supremum norm

$$\|g\|_\infty := \sup_{x \in X} |g(x)| \text{ for } g \in \mathcal{C}(X)$$

where we omit the dependence on X if the set X is clear from the context. This turns $(\mathcal{C}(X), \|\cdot\|)$ into a Banach space [Rudin 2006]. Further, this space carries additional algebraic and order structure. The space $\mathcal{C}(X)$ inherits multiplication (pointwise) from the multiplicativity of $\mathbb{R}$. Similarly, for the order, for $h_1, h_2 \in \mathcal{C}(X)$ we say $h_1 \leq h_2$ in $\mathcal{C}(X)$ if $h_1(x) \leq h_2(x)$ for all $x \in X$. With these additional structures the space $\mathcal{C}(X)$ becomes a commutative Banach algebra [Rudin 1991] and a Banach lattice [Schaefer 1974] with positivity cone $\mathcal{C}(X)_+$ consisting of the non-negative continuous function. These notions mean that for any $h_1, h_2 \in \mathcal{C}(X)$ the multiplication satisfies the following compatibility with the norm

$$\|h_1 \cdot h_2\|_\infty \leq \|h_1\|_\infty \|h_2\|_\infty$$

and for the order structure, it holds

$$0 \leq h_1 \leq h_2 \text{ implies } \|h_1\|_\infty \leq \|h_2\|_\infty.$$

The dual space of $(\mathcal{C}(X), \|\cdot\|_\infty)$ can be identified with the space of signed Borel measures $\mathcal{M}(X)$ by the Riesz-Markov representation theorem [Rudin 2006].

Theorem 2.12 (Riesz-Markov representation theorem)**.** *Any linear form $L : \mathcal{C}(X) \to \mathbb{R}$ with the property $L(h) \geq 0$ for any $h \in \mathcal{C}(X)_+$ can be uniquely represented by a non-negative Borel measure $\mu \in \mathcal{M}(X)_+$ in the form*

$$L(h) = \int_X h \, d\mu. \tag{2.15}$$

The Riesz-Markov representation theorem lets us indicate a path toward polynomial optimization on which we want to characterize linear functionals on polynomials and their geometric support. This path starts with the close relation between the Riesz-Markov representation theorem and Haviland's moment problem. Haviland's moment problem asks when is a linear functional L on the ring of polynomials $\mathbb{R}[x]$ induced by a Borel measure as in (2.15)? The answer seems not surprising

after presenting the Riesz-Markov theorem. Namely, L is represented by a measure supported on a set K if and only if $L(p) \geq 0$ for all polynomials $p \in \mathbb{R}[x]$ that are non-negative on K, see for instance [Marshall 2008]. The support $\operatorname{supp}(\mu)$ of a Borel measure $\mu \in \mathcal{M}(X)_+$ is defined, see for instance [Elstrodt 1996], as

$$\operatorname{supp}(\mu) := \{x \in X : \mu(U) > 0 \text{ for all open neighbourhoods } U \text{ of } x\} \tag{2.16}$$

and thus contains all the region that essentially carries mass. It holds the following property, see [Elstrodt 1996],

$$\int_X h \, d\mu = \int_{\operatorname{supp}(\mu)} h \, d\mu. \tag{2.17}$$

The history on the moment problem will lead to Putinar's and Schmüdgen's Positivstellensätze (Theorems 2.34 and 2.30) in Section 2.4. These Positivstellensätze are essential for the computational application of the sum-of-squares techniques that are used in this thesis. Before we concentrate on polynomial optimization in Section 2.4, we state the missing connecting piece between the answer to Haviland's moment problem and the Riesz-Markov representation theorem: The Stone-Weierstraß theorem (see for instance [Rudin 1991] or any other book on approximation theory or function spaces).

Theorem 2.13 (Weierstraß approximation theorem)**.** *Let $X \subset \mathbb{R}^n$ be compact. The set $\mathbb{R}[x_1, \ldots, x_n]$ of polynomials is dense in $\mathcal{C}(X)$ with respect to $\|\cdot\|_\infty$, i.e. for each $g \in \mathcal{C}(X)$ there exists a sequence $(p_m)_{m\in\mathbb{N}} \subset \mathbb{R}[x_1, \ldots, x_n]$ such that*

$$\|g - p_m\|_\infty \to 0 \text{ as } m \to \infty.$$

We end this section by defining the set of continuously differentiable functions on a compact subset of $\mathbb{R}^n$. Let $U \subset \mathbb{R}^n$ be open and $X = \overline{U}$ the closure of U. For $k \in \mathbb{N}$, by $\mathcal{C}^k(X)$ we denote the set of k-times continuously differentiable functions on U whose derivatives up to order k, can be continuously extended on X, that is

$$\mathcal{C}^k(X) := \{g \in \mathcal{C}^k(X) : g^{(i)} \text{ extends continuously to } X \text{ for } 0 = 1, \ldots, k\},$$

where $g^{(i)}$ denotes the i-th derivative of g, with $g^{(0)} := g$. The space $(C^k(X), \|\cdot\|_{\mathcal{C}^k})$, with the norm

$$\|g\|_{C^k} := \|g\|_\infty + \sum_{i=1}^{k} \|g^{(i)}\|_\infty$$

for

$$\|g^{(i)}\|_\infty := \sup_{x\in X} \|g^{(i)}(x)\|, \tag{2.18}$$

where $\|g^{(i)}(x)\|$ is the operator norm of $g^{(i)}(x)$, is also a Banach algebra. The Weierstraß approximation theorem holds true as well for $(C^k(X), \|\cdot\|_{\mathcal{C}^k})$, which we state in the following theorem. When X is not open we define $\mathcal{C}^k(X)$ as follows: A function g belongs to $C^k(X)$ if and only if there exists an open set $U \supset X$, such that $f \in C^k(U)$.

Theorem 2.14 (Stone-Weierstraß approximation theorem)**.** *Let $U \subset \mathbb{R}^n$ be open and bounded and $X := \overline{U}$ its closure. Let $k \in \mathbb{N}$. The set $\mathbb{R}[x_1, \ldots, x_n]$ of polynomials is dense in $(C^k(X), \|\cdot\|_{\mathcal{C}^k})$.*

2.3 Adjoint operators and (dual) conic programs

Duality plays an important role in many parts of this thesis and appears through dual linear programming problems or adjoint operators. In this section, we provide the necessary notions and follow [Barvinok 2002].

We begin with the dual space.

Definition 2.15. *Let $(V, \|\cdot\|)$ be a normed real vector space. Its dual space V^* consists of all bounded linear forms on V, i.e.*

$$V^* = \{l : V \to \mathbb{R} : l \text{ is linear}, |l(v)| \leq c\,\|v\| \text{ for some } c \geq 0 \text{ for all } v \in V\}.$$

The dual space V^* is equipped with the norm

$$\|l\|_{V^*} := \sup_{\|v\|=1} |l(v)|,$$

which makes $(V^*, \|\cdot\|_{V^*})$ a Banach space.

Duality carries over to operators. This leads to the notion of the adjoint operator.

Definition 2.16. *Let V, W be two normed vector spaces and $T : V \to W$ be a bounded linear operator. The adjoint operator of T is denoted by T^* and given by*

$$T^* : W^* \to V^*, \qquad T^* l := l \circ T.$$

Remark 2.17. *The adjoint T^* of a bounded linear operator T is well-defined, linear, and bounded as well with the same operator norm as T.*

Finally, we present a well-established link between duality and optimization at the example of conic programs. Therefore, we extend the notion of dual spaces to convex cones.

Definition 2.18. *Let V be a normed vector space. A convex cone $C \subset V$ is a convex subset of V such that for all $\lambda \geq 0$ and $c \in C$ it holds $\lambda c \in C$. The dual cone C^* is defined by all linear forms in V^* that are non-negative on C, i.e.*

$$C^* := \{l \in V^* : l(c) \geq 0 \text{ for all } c \in C\}.$$

Remark 2.19. *The Riesz-Markov representation theorem, see Theorem 2.12, states that for compact sets X the space of non-negative Borel measures $\mathcal{M}(X)_+$ is the dual cone of the cone of non-negative continuous functions on X.*

A special case of the observation from Remark 2.19 is that the dual cone of the non-negative orthant $\mathbb{R}^n_+ := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_i \geq 0, i = 1, \ldots, n\}$ is again

$\mathbb{R}^n_+$. The cone $\mathbb{R}^n_+$ plays a central role in finite dimensional linear programming. The standard form of finite dimensional LPs is given by

$$\begin{array}{ll} \inf\limits_{x\in\mathbb{R}^n} & \langle x, a\rangle \\ \text{s.t.} & x \geq 0 \\ & Ax = b \end{array} \tag{2.19}$$

where $a \in \mathbb{R}^n$ is the cost vector, $\langle\cdot,\cdot\rangle$ the euclidean inner product, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The LP (2.19) generalizes in the following way

$$\begin{array}{lll} p^* := & \inf\limits_{v\in V} & l(v) \\ & \text{s.t.} & v \in C \\ & & Tv = w, \end{array} \tag{2.20}$$

where V is a normed vector space, C is a convex cone, $l(v) := \langle v, a\rangle$ a bounded linear form , $T : V \to W$ a bounded linear operator and w an element of W. An optimization problem as (2.20) is called a conic program.

Definition 2.20. *Let V, W be normed vector spaces, $C \subset V$ a convex cone, $l \in V^*$, $T : V \to W$ be a bounded linear operator, and $w \in W$. Then we call the optimization problem (2.20) a conic optimization problem. To a conic program (2.20) we associate its dual program*

$$\begin{array}{lll} d^* := & \sup\limits_{w^*\in W^*, s\in C^*} & w^*(w) \\ & s.t. & T^*w^* + s = l. \end{array} \tag{2.21}$$

One of the fundamental features of the dual program is that it provides lower bounds on the primal problem.

Theorem 2.21 (Weak duality)**.** *Assume that for both the optimization problems (2.20) and (2.21) the feasible set is non-empty. Then it holds*

$$d^* \leq p^*.$$

For this thesis, we content ourselves with this basic introduction to duality and conic optimization problems. For a presentation of the rich theory and applications of linear programming problems, we refer to [Barvinok 2002, Boyd 1997].

2.4 A glimpse at polynomial optimization – real algebra, sum of squares, Positivstellensätze and the Lasserre hierarchy

In this section, we switch the notation for the underlying set from X to K in order not to confuse the set with the variable x in the ring of polynomials $\mathbb{R}[x]$.

We start with the following optimization problem

$$\begin{aligned} f^* := \inf_x \quad & f(x) \\ \text{s.t.} \quad & x \in K. \end{aligned} \tag{2.22}$$

for a (compact) set K and a continuous function $f : K \to \mathbb{R}$. We will again use the following related optimization problem that we have encountered in the introduction section

$$\begin{aligned} \inf_\mu \quad & \int_K f \, d\mu \\ \text{s.t.} \quad & \mu \in \mathcal{M}(K)_+ \\ & \mu(K) = 1. \end{aligned} \tag{2.23}$$

where $\mathcal{M}(K)_+$ is the set of (non-negative) Borel measures on K. The great advantage of (2.23) is that it is a *linear (or conic) programming problem*! This simple but interesting fact is a direct consequence of the following three facts:

1. Integrating a fixed function f against a (signed) measure μ is linear in μ.
2. The set $\mathcal{M}(K)_+$ of non-negative Borel measures is a cone in $\mathcal{M}(X)$. Hence, the condition that μ belongs to $\mathcal{M}(X)_+$ is a conic constraint in the vector space of signed Borel measures on K.
3. The condition $\mu(K) = 1$ can be formulated as the following affine constraint

$$\int_K \mathbf{1} \, d\mu = 1 \tag{2.24}$$

where $\mathbf{1}$ denotes the constant one function $\mathbf{1}(x) := 1$ for all $x \in K$.

A quick verification, as we have done in the introduction section, reveals that both problems (2.22) and (2.23) are equivalent. That is, they both have the same optimal value and there the support $\operatorname{supp}(\mu)$ for any minimizer of the problem (2.23) is contained in the set of minimizers for (2.22). Because (2.23) is a conic programming problem it is useful to state its dual program. By the Riesz-Markov representation theorem, $\mathcal{M}(K)_+$ is the dual cone to the non-negative continuous functions $C(K)_+$, and hence the dual program has the following form

$$\begin{aligned} \sup_{s,g} \quad & s \\ \text{s.t.} \quad & s \in \mathbb{R}, \;\; g \in \mathcal{C}(K)_+ \\ & f - s\mathbf{1} = g \end{aligned} \tag{2.25}$$

Problem (2.25) searches for the largest lower bound s of f on K and is, therefore, also equivalent to our original problem (2.22).

But the obtained linear structure came at a price: The problems (2.25) and (2.23) are infinite dimensional. The complexity of the original problem (2.22) is now expressed in the infinite dimensional non-negativity constraints $\mu \in M(K)_+$

respectively $f - s \in \mathcal{C}(K)_+$. This is where real algebra enters and allows to flourish the certification of the infinite dimensional non-negativity constraints by the application of the Positivstellensätze. In the following we will guide through this main idea – we will briefly visit non-negative polynomials, Hilbert's 17th problem and sum-of-squares polynomials, positive polynomials and finally the celebrated Positivstellensätze by Schmüdgen and Putinar that build the pillar of Lasserre's hierarchy for reformulating an polynomial optimization problem as a hierarchy of semidefinite programs. For detailed and inspiring texts on polynomial optimization, consult [Lasserre 2001, Lasserre 2009, Marshall 2008] among others.

Towards the Lasserre hierarchy

Before introducing the algebraic structure of polynomial optimization problems, we want to fix the following simplification in notation: We denote the space of polynomials $\mathbb{R}[x_1, \dots, x_n]$ in n variables by $\mathbb{R}[x]$. We hope that confusion with a scalar variable x is avoided by the context of its appearance.

The final goal will be to solve the infinite dimensional conic problems (2.23), (2.25) via a hierarchy finite dimensional problems, while benefiting from the linear structure of (2.23), (2.25)! The story begins by trying to efficiently certify membership of a signed measure μ to $\mathcal{M}(K)_+$ in (2.23) respectively to validate the constraint $f - s\mathbf{1} \in \mathcal{C}(K)_+$ from (2.25). In [Lasserre 2001] in 2001, Lasserre realized the that this task can be efficiently achieved using Positivstellensätze from real algebraic geometry.

In order to apply results from real algebraic geometry we need to assume additional algebraic structure to the minimization problem (2.22). Here that means that

1. f is a polynomial
2. K is described by polynomials.

The second condition is made more precise by the notion of semialgebraic sets.

Definition 2.22. *A subset $K \subset \mathbb{R}^n$ is called closed basic semialgebraic if there exists $m \in \mathbb{N}$ and polynomials $p_1, \dots, p_m \in \mathbb{R}[x_1, \dots, x_n]$ such that K has the representation*

$$K = \mathcal{K}(p_1, \dots, p_m) := \{x \in \mathbb{R}^n : p_1(x) \geq 0, \dots, p_m(x) \geq 0\}. \tag{2.26}$$

The set K is called closed semialgebraic if it is a finite union of closed basic semialgebraic sets.

For the minimization problem (2.22) this leads to the notion of polynomial optimization.

Definition 2.23 (Polynomial optimization problem)**.** *The problem*

$$\begin{aligned} f^* := \inf_x \quad & f(x) \\ s.t. \quad & x \in K. \end{aligned} \tag{2.27}$$

with a polynomial f and $K \subset \mathbb{R}^n$ a closed basic semialgebraic set is called a polynomial optimization problem.

An informal formulation of Lasserre's influential result in [Lasserre 2001] is the following.

Theorem 2.24 (Lasserre hierarchy [Lasserre 2001]; informal)**.** *Consider the polynomial optimization problem (2.27) and assume K is compact. There is an efficient algorithm using (convex) semidefinite programming for approximating the global minimum f^*.*

Remark 2.25. *In the context of complexity theory, the word "efficient" often refers to polynomial running time. In the informal formulation of the Lasserre hierarchy in Theorem 2.24, the term "efficient" needs to be treated carefully. Many NP hard problems can be stated as polynomial optimization problems [Boyd 1997, Boyd 2004] and hence polynomial optimization is NP hard. Therefore, we should not expect a polynomial running time for an algorithm that solves polynomial optimization problems exactly. We will address complexity of the Lasserre-hierarchy in Theorem 2.39.*

The algorithm referred to in Theorem 2.24 is often called *Lasserre-hierarchy* or moment-sums-of-squares and the rest of this section is devoted to giving some insides into this procedure.

The algebraic linear programming problem Consider the polynomial optimization problem from Definition 2.23 and its linear reformulation (2.25)

$$\begin{aligned} \sup_{s,g} \quad & s \\ \text{s.t.} \quad & s \in \mathbb{R}, \ \ g \in \mathcal{C}(K)_+ \\ & f - s\mathbf{1} = g \end{aligned} \tag{2.28}$$

It seems inopportune that (2.28) does not see the additional algebraic structure provided by (2.27). We first note that, because f is a polynomial, the term $f - s\mathbf{1}$ is also a polynomial, and we can equivalently replace (2.28) by

$$\begin{aligned} \sup_{s,g} \quad & s \\ \text{s.t.} \quad & s \in \mathbb{R}, \ \ g \in \mathbb{R}[x] \\ & f - s\mathbf{1} = g \\ & g \geq 0 \text{ on } K \end{aligned} \tag{2.29}$$

This formulation emphasizes the need for an answer to the fundamental question about which form non-negative polynomials can take.

How can the non-negative polynomials on K be characterized? This famous question underwent an impressive development. The starting point is built by natural candidates for non-negative polynomials – the sum-of-squares (SOS)

polynomial. A polynomial p is SOS if there are $m \in \mathbb{N}$ and polynomials $q_1, \ldots, q_m \in \mathbb{R}[x]$ such that p can be written

$$p = \sum_{i=1}^{m} q_i^2. \tag{2.30}$$

The set of all SOS polynomials is denoted by

$$\Sigma := \left\{ \sum_{i=1}^{m} q_i^2 : m \in \mathbb{N},\ q_1, \ldots, q_m \in \mathbb{R}[x] \right\}. \tag{2.31}$$

In 1888 Hilbert [Hilbert 1888] presented a non-explicit construction of a non-negative polynomial that is not a sum-of-squares. It took until 1965 for an explicit example of a non-negative polynomial that is not SOS [Motzkin 1967]. Motzkin presented the following polynomial M

$$M(x, y) := 1 + x^4y^2 + x^2y^4 - 2x^2y^2 \tag{2.32}$$

which is non-negative but not an SOS polynomial. By Hilbert's abstract construction, it was clear that Motzkin's polynomial is not the only non-negative polynomial that is not SOS. Indeed, SOS polynomials are rare among the non-negative polynomials – in 2006 (and an earlier preprint in 2003) Blekherman [Blekherman 2006] showed that there are "significantly more non-negative polynomials than SOS".

These results show in a striking fashion that SOS polynomials are not rich enough to describe all non-negative polynomials. In his celebrated list of 23 problems, in 1900, Hilbert asked the following question

> Hilbert's 17th problem: *Is every globally non-negative polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ a sum of squares of rational functions?*

This question was very adequate, as was demonstrated by Artin in [Artin 1927] in 1927 by showing that the answer is *"yes!"* and the powerful field of real algebraic geometry emerged. Among the very fruitful results from real algebraic geometry are the celebrated Nichtnegativstellensätze and Positivstellensätze by Krivine, Stengle, Schmüdgen and Putinar [Marshall 2008], see Theorems 2.29, 2.30 and 2.34. Those Nichtnegativstellensätze and Positivstellensätze concern non-negativity respectively positivity on semialgebraic sets K. Thus, the algebraic nature of the set K needs to be taken into account. One way is to generalize the set of SOS polynomials to the quadratic module $\mathcal{Q}(p_1, \ldots, p_m)$ generated by $p_1, \ldots, p_m$, see the following Definition 2.26. And for computations, we are interested in a truncated variant of $\mathcal{Q}(p_1, \ldots, p_m)$ where only polynomials up to a certain degree are considered. For $d \in \mathbb{N}$ we denote the set of polynomials of degree at most d by $\mathbb{R}[x]_d$.

Definition 2.26. *For $p_1, \ldots, p_m \in \mathbb{R}[x]$ the quadratic module $\mathcal{Q}(p_1, \ldots, p_m)$ is defined by*

$$\mathcal{Q}(p_1, \ldots, p_m) := \left\{ \sigma_0 + \sum_{i=1}^{m} \sigma_i p_i : \sigma_0, \sigma_1, \ldots, \sigma_m \in \Sigma \right\}. \tag{2.33}$$

We denote by $\mathcal{Q}_d(p_1,\dots,p_m)$ *the following part of* $\mathcal{Q}(p_1,\dots,p_m)$ *that is obviously contained in* $\mathbb{R}[x]_d$ *and given by*

$$\mathcal{Q}_d(p_1,\dots,p_m) := \Big\{\sigma_0+\sum_{i=1}^m \sigma_i p_i \quad : \quad \sigma_0,\sigma_1,\dots,\sigma_m \in \Sigma, \quad \deg(\sigma_0),\deg(\sigma_1 p_1),\dots,\deg(\sigma_m p_m) \le d\Big\}. \tag{2.34}$$

If the closed basic semialgebraic set K is given by $K = \mathcal{K}(p_1,\dots,p_m)$ as in (2.26) then all polynomials in $\mathcal{Q}(p_1,\dots,p_m)$ are non-negative on K. But, not surprisingly, since we have seen the global case – the set $\mathcal{Q}(p_1,\dots,p_m)$ does not cover all non-negative polynomials on K, even if K is compact.

Example 2.27. *As an example let* $p_1(x) = (1-x^2)^3$ *in one scalar variable* $x \in \mathbb{R}$ *defining the set* $K = [-1,1]$. *Consider the, on* K, *non-negative polynomial* $f(x) := 1-x^2$. *A representation*

$$f = \sigma_0 + \sigma_1(1-x^2)^3 \tag{2.35}$$

for $\sigma_0,\sigma_1 \in \Sigma$ *would imply* $\sigma_0(-1) = \sigma_0(1) = f(1) = 0$ *and even* $\sigma_0'(-1) = \sigma_0'(1) = 0$ *because* σ_0 *is non-negative and thus* $-1, 1$ *are global minimizers of* σ_0. *That means* $(1-x^2)^2$ *is a factor of* σ_0, *in particular, the right-hand side of (2.35) factors by the polynomial* $(1-x^2)^2$ *but the left-hand side does not. Arguably, one might say that the representation of* K *by* $p_1(x) = (1-x^2)^3$ *is not appropriate and a better choice would be to represent* K *by* $\hat{p}_1(x) := 1-x^2$. *In this case we would have* $f = \mathbf{1}\cdot\hat{p}_1$ *with the SOS polynomial* $\mathbf{1}(x) = 1 = \mathbf{1}(x)^2$ *for all* x. *This is true but trivially sidesteps the question about representation of non-negative polynomials in the following way: Let* f *be non-negative on* $K = \{x \in \mathbb{R}^n : p_1(x) \ge 0,\dots,p_m(x) \ge 0\}$, *then also*

$$K = \mathcal{K}(p_1,\dots,p_m,f) = \{x \in \mathbb{R}^n : p_1(x) \ge 0,\dots,p_m(x) \ge 0, f(x) \ge 0\},$$

and there is the trivial representation $f = 0 + 0\cdot p_1 + \ldots + 0\cdot p_m + \mathbf{1}\cdot f$. *The underlying flaw that happened here was that instead of verifying* ***if*** f *is non-negative, the knowledge* ***that*** f *is non-negative was already used.*

Because the quadratic module $\mathcal{Q}(p_1,\dots,p_m)$ does not contain enough non-negative polynomials the search for better classes of candidate non-negative polynomials continued. The set $\mathrm{Pos}(K)$ of non-negative polynomials on a set K naturally carries the following properties

1. for any $q_1, q_2 \in \mathrm{Pos}(K)$ also q_1+q_2 and $q_1\cdot q_2$ are non-negative on K, i.e.

$$\mathrm{Pos}(K)+\mathrm{Pos}(K), \mathrm{Pos}(K)\cdot\mathrm{Pos}(K) \subset \mathrm{Pos}(K) \tag{2.36}$$

2. for any $p \in \mathbb{R}[x]$ the polynomial p^2 is non-negative on K, i.e.

$$\Sigma \subset \mathrm{Pos}(K) \tag{2.37}$$

3. the polynomial $p = -\mathbf{1}$ is *not* non-negative on K, i.e.

$$-\mathbf{1} \notin \mathrm{Pos}(K). \tag{2.38}$$

The properties (2.36), (2.37), and (2.38) should be satisfied for any set that reflects non-negativity. In real algebraic geometry, a set that satisfies (2.36), (2.37), and (2.38) is called a preordering. The smallest preordering containing polynomials $p_1, \ldots, p_m$ is denoted by

$$\mathrm{Pre}(p_1, \ldots, p_m) := \left\{ \sum_{e=(e_1,\ldots,e_m)\in\{0,1\}^m} \sigma_e p_1^{e_1} \ldots p_m^{e_m} : \sigma_e \in \Sigma \; \forall e \in \{0,1\}^m \right\}. \tag{2.39}$$

The preordering $\mathrm{Pre}(p_1, \ldots, p_m)$ contains the quadratic module $\mathcal{Q}(p_1, \ldots, p_m)$ and all polynomials in $\mathrm{Pre}(p_1, \ldots, p_m)$ are non-negative polynomials on $\mathcal{K}(p_1, \ldots, p_m)$ and that this observation can be partially reversed is the statement of Schmüdgen's Positivstellensatz, see Theorem 2.30.

Before moving to a seemingly small but effectively strong restriction to positive polynomials we want to emphasize that neither the quadratic module $\mathcal{Q}(p_1, \ldots, p_m)$ nor the preordering $\mathrm{Pre}(p_1, \ldots, p_m)$ reflect purely geometric properties of the corresponding set $K = \{x \in \mathbb{R}^n : p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$. We specify this in the following remark.

Remark 2.28. *The $\mathcal{Q}(p_1, \ldots, p_m)$ and $\mathrm{Pre}(p_1, \ldots, p_m)$ are not geometric invariants of the set and $K = \mathcal{K}(p_1, \ldots, p_m)$. That is for different representations*

$$\mathcal{K}(p_1, \ldots, p_m) = K = \mathcal{K}(q_1, \ldots, q_l)$$

for $m, l \in \mathbb{N}$ and polynomials $p_1, \ldots, p_m, q_1, \ldots, q_l \in \mathbb{R}[x]$ the sets $\mathcal{Q}(p_1, \ldots, p_m)$ and $\mathcal{Q}(q_1, \ldots, q_l)$ respectively $\mathrm{Pre}(p_1, \ldots, p_m)$ and $\mathrm{Pre}(q_1, \ldots, q_l)$ differ in general.

A little bit about positive polynomials As Artin [Artin 1927] showed in his answer to Hilbert's 17th problem it is possible to represent every globally non-negative polynomial as sum-of-squares of rational functions. This was further generalized to non-negative polynomials on closed basic semialgebraic sets in the Krivine-Stengle Nichtnegativstellensatz.

Theorem 2.29 (Krivine-Stengle Stellensätze; [Marshall 2008])**.** *Let $p_1, \ldots, p_m \in \mathbb{R}[x]$, $K = \mathcal{K}(p_1, \ldots, p_m)$, $\mathcal{P} = \mathrm{Pre}(p_1, \ldots, p_m)$ and $f \in \mathbb{R}[x]$. Then*

1. *$f > 0$ on K is equivalent to $pf = 1 + q$ for some $p, q \in \mathcal{P}$.*
2. *$f \geq 0$ on K is equivalent to $pf = f^{2m} + q$ for some $m \in \mathbb{N}$ and $p, q \in \mathcal{P}$.*
3. *$f = 0$ on K is equivalent to the existence of $m \in \mathbb{N}$ such that $-f^{2m} \in \mathcal{P}$.*
4. *$K = \emptyset$ is equivalent to $-1 \in \mathcal{P}$.*

The Stellensätze 1. and 2. in Theorem 2.29 contain denominators and the question about removing the need for denominators remained unanswered until

Schmüdgen's Positivstellensatz appeared in [Schmüdgen 1991] in 1991. It is clear from Hilbert's construction [Hilbert 1888] that non-negativity alone is not enough to obtain a denominator-free representation. As Schmüdgen showed, sufficient conditions to remove denominators are compactness and strict positivity.

Theorem 2.30 (Schmüdgen's Positivstellensatz [Schmüdgen 1991])**.** *Let polynomials $p_1, \dots, p_m \in \mathbb{R}[x]$ such that $K = \mathcal{K}(p_1, \dots, p_m)$ is compact. Then for any polynomial $f > 0$ on K it holds $f \in \mathrm{Pre}(p_1, \dots, p_m)$.*

There are at least two perspectives from which it becomes clearer why strictly positive polynomials are treated significantly easier:

- From an algebraic perspective the positivity of a polynomial f allows dividing by $f(x)$ in a well-defined way (as in [Marshall 2008] in the proof the Krivine-Stengle Positivstellensatz) or related objects (as in [Prestel 2013] for the generalized abstract Positivstellensatz).

- A functional analytic perspective realizes the set of non-negative polynomials with zeros as the boundary in $\mathbb{R}[x]$ (with respect to the supremum norm topology) of the cone of the non-negative polynomials and is, therefore, harder to distinguish from the complement of this cone via linear separation.

We can also extract the need for compactness from the functional analytic perspective: If K is not compact the supremum norm $\|p\|_\infty$ is not well defined for some $p \in \mathbb{R}[x]$. Indeed, Schmüdgen's Positivstellensatz does not hold without the compactness assumption, as we present in the following example.

Example 2.31 (Schmüdgen's Positivstellensazt does not hold without the compactness assumption)**.** *Let $p_1(x) = x^3$ and $f(x) = x + 1$ then $K = \mathcal{K}(p_1) = [0, \infty)$ and f is strictly positive on K. A representation $f = \sigma_0 + x^3\sigma_1$ of f with $\sigma_0, \sigma_1 \in \Sigma$ as suggested Schmüdgen's Positivstellensatz is not possible because σ_0 has even degree. Therefore, the leading term of $x^3\sigma_1$ would have to equal the leading term x of f.*

Schmüdgen's original idea of the proof of his Positivstellensatz it embraces the same guiding principle that flavors several parts of this thesis:

Translating a nonlinear problem into a linear one and borrowing tools from functional analysis.

Because this guiding principle is beautifully expressed in his proof we want to sketch the ideas of the proof.

Proof. Sketch of Schmüdgen's proof of Schmüdgen's Positivstellensatz. The proof is based on the combination of two results: The first is a refined Hahn-Banach argument for separating a polynomial from the preordering $\mathrm{Pre}(p_1, \dots, p_m)$. The second is that such a separating functional can be represented by a non-negative measure.

Let us assume that f is positive on K but does not belong to $\mathrm{Pre}(p_1, \dots, p_m)$. The separation argument [Schmüdgen 1991, Proof of Corollary 3] states that there

exists a non-trivial linear functional $L : \mathbb{R}[x] \to \mathbb{R}$ (continuous in the finest locally convex topology on $\mathbb{R}[x]$) with the property

$$L(f) \leq 0 \text{ and } L(p) \geq 0 \text{ for all } p \in \mathrm{Pre}(p_1, \ldots, p_m). \tag{2.40}$$

By [Schmüdgen 1991, Theorem 1] (see the following Theorem 2.32), this implies that there exists a non-trivial non-negative measure $\mu \in \mathcal{M}(K)_+$ with

$$L(g) = \int_K g \, d\mu \quad \text{for all } g \in \mathbb{R}[x]. \tag{2.41}$$

Inserting f for g in (2.41) gives

$$L(f) = \int_K f \, d\mu > 0, \tag{2.42}$$

where the positivity of the right hand side follows from strict positivity of f and μ being non-negative and non-trivial. Finally, (2.42) contradicts the separation property (2.40) of L and we conclude the statement. □

The main argument, [Schmüdgen 1991, Theorem 1], that concluded Schmüdgen's Positivstellensatz is the following theorem.

Theorem 2.32 ([Schmüdgen 1991, Theorem 1])**.** *Let polynomials $p_1, \ldots, p_m \in \mathbb{R}[x]$ such that $K = \mathcal{K}(p_1, \ldots, p_m)$ is compact and $L : \mathbb{R}[x] \to \mathbb{R}$ be a linear map. The map L is representable by a Borel measure $\mu \in M(K)_+$ if and only if L is non-negative on* $\mathrm{Pre}(p_1, \ldots, p_m)$.

We will only sketch its proof in order to emphasize the beautiful interplay between algebraic and functional analytic arguments and refer to [Schmüdgen 1991] for the details.

Proof. Sketch. The proof consists of the following steps

1. Define a complex Hilbert space $\mathcal{H}$ of (equivalence classes) of polynomials with an inner product $\ell : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ associated to the linear functional L.

2. For each variable x_j for $j = 1, \ldots, n$ introduce the multiplication operators $M_j : \mathcal{H} \to \mathcal{H}$ with $p \mapsto x_j \cdot p$, and show that they are well-defined, bounded, symmetric and commuting.

3. Apply the spectral theorem for the family operators $M_1, \ldots, M_n$ to obtain a spectral measure μ.

4. Show that the measure μ represents L.

Before we start, we pass to the complexification $\mathbb{C}[x]$ of $\mathbb{R}[x]$ because we want to use spectral theory. Therefore, we extend L to $\mathbb{C}[x] = \mathbb{R}[x] + i\mathbb{R}[x]$ in the natural way by $L(p + iq) := L(p) + iL(q)$ where i denotes the imaginary unit.

For the first step, we define a bilinear map l from L in the following way

$$l : \mathbb{C}[x] \times \mathbb{C}[x] \to \mathbb{C}, l(p, q) := L(p \cdot \overline{q}) \tag{2.43}$$

where $\overline{q}$ denotes the complex conjugate of q. The map l is symmetric and for all $q_1, \ldots, q_k \in \mathbb{C}[x]$ and $a_1, \ldots, a_n \in \mathbb{C}$ satisfies

$$\sum_{r,s=1}^{k} a_r \overline{a_s} l(q_r, q_s) = \sum_{r,s=1} a_r \overline{a_s} L(q_r \cdot \overline{q_s}) = L\left(\left(\sum_{r=1}^{n} a_r q_r\right) \cdot \overline{\left(\sum_{r=1}^{n} a_r q_r\right)}\right) \geq 0, \tag{2.44}$$

The last inequality in (2.44) follows from the fact that $q \cdot \overline{q}$ is sum of the squares of the real part imaginary part of q, and that L is non-negative on real sum-of-squares. Statement (2.44) means that L is a positive semidefinite kernel on the set $\mathbb{C}[x]$. Thus, we can associate a reproducing kernel Hilbert space with it (see Section 2.6.1 for the definition of kernels and reproducing kernel Hilbert spaces). In this case, it is advantageous to recall the procedure of building a reproducing kernel Hilbert space from a positive semidefinite kernel. It starts by setting

$$N := \{p \in \mathbb{R}[\mathbf{X}] : L(p^2) = 0\} \tag{2.45}$$

to be the set where l fails to have definiteness. By the Cauchy-Schwarz inequality for the positive semidefinite bilinear form l, the set N is an ideal and we can consider the inner product space

$$(H' := {}^{\mathbb{C}[x]}/_{N}, l')$$

where $l' : H' \times H' \to \mathbb{C}$ is the induced and well defined bilinear map on H' given by

$$l'(p + N, q + N) := l(p, q) = L(p \cdot \overline{q}).$$

The desired complex Hilbert space $(\mathcal{H}, \ell)$ is obtained by completion of (H', l').

For the second step, the goal is to define the operators M_j that multiply by the monomials x_j for $j = 1, \ldots, n$. We begin with defining the multiplication operators $M_j : H' \to H'$ for $j = 1, \ldots, n$ by

$$M_j(f + N) := x_j \cdot f + N. \tag{2.46}$$

Since N is an ideal it follows that M_j is well defined for all $1 \leq j \leq n$. Because $\mathbb{C}[x]$ is a commutative algebra, the operators $M_1, \ldots, M_n$ are pairwise commuting. Further, they are self adjoint because

$$\begin{aligned} l'(M_j(f + N), g + N) &= l(x_j \cdot f, g) = L(x_j \cdot f \cdot \overline{g}) = L(f \cdot \overline{x_j \cdot g}) \\ &= l(f, x_j \cdot g) = l'(f + N, M_j(g + N)). \end{aligned}$$

The crucial part, which we will only address in Remark 2.33, is showing that the operators M_j are bounded on H' – here the compactness of K becomes essential. Once it is guaranteed that the operators M_j are bounded (on H'), they can be extended to bounded selfadjoint pairwise commuting operators $\hat{M}_j$ on the Hilbert space $\mathcal{H}$ and the spectral theorem [Rudin 1991, Theorem 12.22] can be applied.

The spectral theorem provides a measure μ supported on a compact set $\mathscr{K} := \sigma(\hat{M}_1) \times \ldots \times \sigma(\hat{M}_n) \subset \mathbb{R}^n$, where $\sigma(\hat{M}_j)$ denotes the spectrum of the operator $\hat{M}_j$ for each $j = 1, \ldots, n$, such that for all $\alpha_1, \ldots, \alpha_n \in \mathbb{N}_0$

$$l'(\hat{M}_1^{\alpha_1} \cdots \hat{M}_n^{\alpha_n}(\mathbf{1}+N), (\mathbf{1}+N)) = \int_{\mathscr{K}} x_1^{\alpha_1} \cdot x_n^{\alpha_n} \, d\mu.$$

But the left-hand side is nothing else than $L(x_1^{\alpha_1} \cdots x_n^{\alpha_n})$, which shows the desired representation

$$L(p) = \int_{\mathscr{K}} p \, d\mu \quad \text{for all } p \in \mathbb{C}[x].$$

Hence μ is our desired measure once we have checked that the support $\mathscr{K}$ of μ is contained in K. Therefore, note that for all $h \in \mathbb{R}[x]$ and $1 \leq j \leq m$ it holds

$$\int_{\mathscr{K}} h^2 p_j \, d\mu = L(h^2 p_i) \geq 0.$$

Since $\mathscr{K} \subset \mathbb{R}^n$ is compact it follows from the Stone-Weierstraß theorem that for all $h \in \mathcal{C}(\mathscr{K})$ it holds

$$\int_{\mathscr{K}} h^2 p_j \, d\mu \geq 0.$$

Since each non-negative continuous function has a continuous root it follows that for all $h \in \mathcal{C}(\mathscr{K})_+$

$$\int_{\mathscr{K}} h p_i \, d\mu \geq 0,$$

and hence $\operatorname{supp}(\mu) \subset p_i^{-1}([0,\infty))$, in particular $\operatorname{supp}(\mu) \subset \bigcap_{i=1}^{m} p_i^{-1}([0,\infty)) = K$. □

The part of Schmüdgen's proof of his Positivstellensatz that we did not present in the above proof is the part that shows that the multiplication operators M_j are bounded. This is addressed in the following remark.

Remark 2.33. *In his proof, Schmüdgen had to perform a clever reduction to the one dimensional moment problem to obtain the desired result that the multiplication operators M_j from (2.46) are bounded. The arguments become significantly simpler using the so-called Wörmann's trick [Marshall 2008] which appeared only in 1998 [Wörmann 1998]. Wörmann showed that the set $\mathcal{K}(p_1, \ldots, p_m)$ is compact if and only if there exists a $r \in [0, \infty)$ such that*

$$r - \sum_{j=1}^{n} x_j^2 \in \operatorname{Pre}(p_1, \ldots, p_m). \tag{2.47}$$

The interesting part in this statement is that compactness implies (2.47). Evoking Wörmann's result, the proof of boundedness of the multiplication operators M_j can be simplified and we will state it here. First note that (2.47) implies that for the

linear map L from Theorem 2.32 it holds

$$L\left(\left(\sum_{j=1}^{n} x_j^2\right)\cdot p\right) \leq rL(p) \text{ for all } p \in \mathrm{Pre}(p_1,\ldots,p_m). \tag{2.48}$$

Further, because L is non-negative on $\mathrm{Pre}(p_1,\ldots,p_m)$ it holds

$$0 \leq L(x_j^2\cdot p) \leq L\left(\left(\sum_{j=1}^{n} x_j^2\right)\cdot p\right) \text{ for all } p \in \mathrm{Pre}(p_1,\ldots,p_m). \tag{2.49}$$

Now, using the notation from the above proof of Schmüdgen's Positivstellensatz it holds for all $p \in \mathbb{C}[x]$

$$\begin{aligned} l'(M_j(p+N), M_j(p+N)) &= l'(x_j\cdot p+N, x_j\cdot p+N) = L(x_j^2\cdot p\cdot \overline{p}) \\ &\overset{(2.49)}{\leq} L\left(\left(\sum_{j=1}^{n} x_j^2\right)\cdot p\cdot\overline{p}\right) \overset{(2.48)}{\leq} rL(p\overline{p}) \\ &= rl'(p+N, p+N). \end{aligned}$$

This shows that the operators M_j are bounded (with operator norm $\|M_j\|_{\leq}\sqrt{r}$), which was to be shown.

The condition (2.47) on the preordering $\mathrm{Pre}(p_1,\ldots,p_m)$ is called Archimedean condition in real algebraic geometry. A closer look at the presented proof of Theorem 2.32 reveals that the Archimedean property is the only property of the set $\mathrm{Pre}(p_1,\ldots,p_m)$ that was used but which is not shared by the quadratic module $\mathcal{Q}(p_1,\ldots,p_m)$. Thus, enforcing the Archimedean property by design gives the celebrated Putinar's Positivstellensatz.

Theorem 2.34 (Putinar's Positivstellensatz [Putinar 1993])**.** *Let $p_1,\ldots,p_m \in \mathbb{R}[x]$ such that one of the sets $\{x\in\mathbb{R}^n : p_j(x)\geq 0\}$ for $j=1,\ldots,n$ is compact. Then any $f\in\mathbb{R}[x]$ with $f>0$ on $\mathcal{K}(p_1,\ldots,p_m)$ belongs to $\mathcal{Q}(p_1,\ldots,p_m)$.*

The condition that one of the sets $\{x\in\mathbb{R}^n : p_j(x)\geq 0\}$ for $j=1,\ldots,n$ is compact guarantees that the Archimedean condition is satisfied for the quadratic module $\mathcal{Q}(p_1,\ldots,p_m)$ [Lasserre 2009]. In practice this is often trivially enforced by adding a redundant constraint $p_{m+1}(x) := r^2 - \sum_{j=1}^{n} x_j^2$ for $r\in\mathbb{R}$ for which it is a priori known that K is contained in the closed ball $\bar{B}_r(0)$ of radius r centered at the origin.

Convex formulation for being SOS In the previous paragraphs, we rephrased the polynomial optimization problem (2.27) into a problem about membership to the quadratic module $\mathcal{Q}(p_1,\ldots,p_m)$. Now we address how we can computationally verify membership to $\mathcal{Q}(p_1,\ldots,p_m)$ and we will do so in a convex fashion.

Let us begin with a sum-of-squares polynomial $\sigma\in\Sigma$. Let the degree $\deg(\sigma)$

be at most $2d \in \mathbb{N}$ and we write

$$\sigma = \sum_{l=1}^{L} q_l^2$$

for polynomials

$$q_l = \sum_{\alpha \in \mathbb{N}_0^n, |\alpha| \leq d} q_{l,\alpha} x^\alpha \ \in \mathbb{R}[x] \ \text{ for } l = 1, \ldots, L$$

where $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ denotes a multi-index, $|\alpha| := \sum_{j=1}^{n} \alpha_i$ its degree and $x^\alpha := (x_1^{\alpha_1}, \ldots, x_n^{\alpha_n})$. For convenience let us denote the vector of monomials up to degree d by $v_d = (x_\alpha)_{|\alpha| \leq d} \in \mathbb{R}^{s(d)}$ where $s(d) := \binom{n+d}{d}$ is the dimension of $\mathbb{R}[x]_d$. For $l = 1, \ldots, L$ let us write

$$q_l = \sum_{|\alpha| \leq d} c_{l,\alpha} x^\alpha = c^T v_d,$$

with coefficients $c_l := (c_{l,\alpha})_{|\alpha| \leq d} \in \mathbb{R}^{s(d)}$. Then q_l^2 is given by

$$q_l^2 = \sum_{|\alpha|,|\beta| \leq d} c_{l,\alpha} c_{l,\beta} x^{\alpha+\beta} = v_d^T c_l c_l^T v_d.$$

In other words, we can write

$$q_l^2 = v_d^T C_l v_d$$

with a matrix $C_l = c_l c_l^T \in \mathbb{R}^{s(d) \times s(d)}$. The matrix C_l is positive semidefinite, which we denote by $C_l \succeq 0$. And for the SOS polynomial $\sum_{l=1}^{L} q_l^2$ we get

$$\sum_{l=1}^{L} q_l^2 = \sum_{l=1}^{L} v_d^T C_l v_d = v_d^T \sum_{l=1}^{L} C_l v_d = v_d^T C v_d$$

for the positive semidefinite matrix $C = \sum_{l=1}^{L} C_l$. On the other hand any polynomial q, that can be written as $q = v_d^T C v_d$ with C being a positive semidefinite matrix, is an SOS polynomial and can be found in any text on polynomial optimization. The argument is writing $C = HH^T$ with $H \in \mathbb{R}^{s(d) \times \operatorname{rank}(C)}$ which gives

$$q = v_d^T C v_d = v_d^T H H^T v_d = \sum_{l=1}^{\operatorname{rank}(C)} \left((H^T v_d)_l \right)^2.$$

Thus, we have the following equivalence

$$\sigma \in \Sigma \text{ if and only there exists } 0 \preceq C \in \mathbb{R}^{s(d) \times s(d)} \text{ with } \sigma = v_d^T C v_d.$$

Simply by comparing the coefficients of σ and $v_d^T C v_d$ we get

$$\sigma = \sum_{|\gamma| \leq 2d} \sigma_\gamma x^\gamma \in \Sigma \quad \text{if and only there exists} \quad C = (c_{\alpha,\beta})_{|\alpha|,|\beta| \leq d} \in \mathbb{R}^{s(d) \times s(d)}$$
$$\text{s.t. } 0 \preceq C \text{ and } \sigma_\gamma = \sum_{\alpha+\beta=\gamma} C_{\alpha,\beta}.$$

This extends to polynomials q of the form $q = \sigma p$ for $\sigma \in \Sigma$ and $p \in \mathbb{R}[x]$ fixed: We represent σ by a positive semidefinite matrix C as $\sigma = v_d^T C v_d$ and comparing coefficients of q and the polynomial $\sigma \cdot p$ leads to a linear constraint on the matrix C. For $d \in \mathbb{N}$, we arrive at verifying membership of a polynomial q $\mathcal{Q}_d(p_1, \ldots, p_m)$ via the following condition:

A polynomial q satisfies $q \in \mathcal{Q}_d(p_1, \ldots, p_m)$ if and only if there exist positive semidefinite matrices $C_0, C_1, \ldots C_m$ of size $s\left(\left\lfloor \frac{d}{2} \right\rfloor\right)$ respectively $s\left(\left\lfloor \frac{d_1}{2} \right\rfloor\right), \ldots, s\left(\left\lfloor \frac{d_m}{2} \right\rfloor\right)$ that (jointly) satisfy the linear constraint

$$v_{d_0}^T C_0 v_{d_0} + \sum_{j=1}^{m} v_{d_j}^T C_j v_{d_j} p_j = q. \tag{2.50}$$

By merging the matrices together into a block-diagonal matrix C, by

$$C := \begin{pmatrix} C_0 & 0 & \cdots & 0 \\ 0 & C_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_n \end{pmatrix},$$

the condition that $C_j \succeq 0$ for all $j = 0, \ldots, m$ is represented by $C \succeq 0$ and the linear constraint (2.50) translates to a linear constraint on C. That rephrases the question about membership to $\mathcal{Q}_d(p_1, \ldots, p_m)$ as the task of finding a positive semidefinite matrix satisfying linear constraints. This means that we search for an affine slice in the convex cone of positive semidefinite matrices – in particular, we search for a *convex* set.

The procedure that we have outlined in this paragraph is standard in polynomial optimization and can be found in any related text, for example [Lasserre 2001, Lasserre 2009, Lasserre 2015].

Semidefinite programs The penultimate step towards the Lasserre hierarchy is to extend the feasibility question for the matrix $C \succeq 0$ to an optimization problem. We will consider costs linear in C. The reason is that – because the set of symmetric positive semidefinite matrices forms a convex cone – this gives a conic optimization

problem in the sense of Definition 2.20. This special class of conic programs carries a name – semidefinite program (SDP).

Definition 2.35 (Semidefinite program)**.** *Let $n, r \in \mathbb{N}$, $F_0, F_1, \ldots, F_r \in \mathbb{R}^{n\times n}$ and $b_1, \ldots, b_r \in \mathbb{R}$. The following optimization problem is called a semidefinite program*

$$\begin{aligned} P^* := \quad & \sup_{C\in\mathbb{R}^{n\times n}} && \langle F_0, C\rangle \\ & s.t. && \langle F_i, C\rangle = b_j, \quad j = 1, \ldots, r \\ & && C \succeq 0 \end{aligned}$$

where $\langle A, B\rangle := \mathrm{Tr}(AB) = \sum_{i,j=1}^{n} a_{ij}b_{ij}$ for $A, B \in \mathbb{R}^{n\times n}$ denotes the trace inner product on $\mathbb{R}^{n\times n}$. Its dual program (in the sense of conic programs from Section 2.21) has the following form

$$\begin{aligned} D^* := \quad & \inf_{x=(x_1,\ldots,x_r)\in\mathbb{R}^r} && b^T x \\ & s.t. && \sum_{j=1}^{r} x_j F_j - F_0 \succeq 0. \end{aligned}$$

Since the set of positive semidefinite matrices is a convex cone and the cost is linear in the matrix C, a semidefinite program is convex. By weak duality, it always holds $P^* \geq D^*$. For more insights into semidefinite optimization, such as criteria on strong duality, applications and algorithms, see for example [Blekherman 2012, Boyd 1997].

The Lasserre hierarchy Now we have all the machinery to state Lasserre's hierarchy for polynomial optimization. For a polynomial $f \in \mathbb{R}[x]$ and a set $K = \mathcal{K}(p_1, \ldots, p_m) = \{x \in \mathbb{R}^n : p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$ for given $p_1, \ldots, p_m \in \mathbb{R}[x]$ we want to solve the following polynomial optimization problem

$$\begin{aligned} f^* := \inf_x \quad & f(x) \\ \text{s.t.} \quad & x \in K. \end{aligned} \tag{2.51}$$

As we have stated in (2.28) the corresponding conic program

$$\begin{aligned} \sup_{s,g} \quad & s \\ \text{s.t.} \quad & s \in \mathbb{R}, \quad g \in \mathbb{R}[x] \\ & f - s\mathbf{1} = g \\ & g \geq 0 \text{ on } K. \end{aligned}$$

Motivated by Putinar's Positivstellensatz, this leads to the following optimization problem

$$\begin{aligned} s^* := \sup_s \quad & s \\ \text{s.t.} \quad & s \in \mathbb{R}, g \in \mathcal{Q}(p_1, \dots, p_m) \\ & f - s\mathbf{1} = g \end{aligned} \tag{2.52}$$

and indeed, by Putinar's Positivstellensatz, Theorem 2.34, it holds

$$s^* = f^*.$$

Putting all the previous paragraphs together, we finally arrive at a computationally tractable algorithm for polynomial optimization problems – this algorithm is called the Lasserre hierarchy.

Definition and Theorem 2.36 (Lasserre hierarchy for polynomial optimization)**.** *Let $p_1, \dots, p_m \in \mathbb{R}[x]$ such that the condition from Putinar's theorem is satisfied. Let $f \in \mathbb{R}[x]$. The Lasserre hierarchy associated with the polynomial optimization*

$$\begin{aligned} f^* := \inf_{x \in \mathbb{R}^n} \quad & f(x) \\ s.t. \quad & x \in \mathcal{K}(p_1, \dots, p_m) \end{aligned}$$

is given by the following sequence of sum-of-squares programs: For each $d \in \mathbb{N}$ consider the optimization problem

$$\begin{aligned} s_d^* := \sup_{s \in \mathbb{R}} \quad & s \\ s.t. \quad & f - s \in \mathcal{Q}_d(p_1, \dots, p_m). \end{aligned} \tag{2.53}$$

It holds

$$s_d^* \leq s_{d+1}^* \text{ for all } d \in \mathbb{N} \text{ and } s_d \text{ converges to the } \textbf{\textit{global}} \text{ optimum } f^* \text{ as } d \to \infty. \tag{2.54}$$

For each $d \in \mathbb{N}$ the sum-of-squares program (2.53) can be expressed as a semidefinite program (2.35).

The statement (2.54) contains the essence why the Lasserre hierarchy so valuable – namely it provides a hierarchy of convex problems whose optimal values monotonously convergent to the global minimum of f on K! The first statement from (2.54) follows from inclusion $\mathcal{Q}_d(p_1, \dots, p_m) \subset \mathcal{Q}_{d+1}(p_1, \dots, p_m)$ and the convergence is a direct consequence of Putinar's Positivstellensatz.

In the following remark, we outline how the SDP representing the SOS program (2.53) is formed.

Remark 2.37. *For each $d \in \mathbb{N}$ the optimization problem (2.53) can be formulated as an SDP. To do so we use the semidefinite description around (2.50) of membership to $\mathcal{Q}_d(p_1, \dots, p_m)$. We express the condition $f - s\mathbf{1} \in \mathcal{Q}_d(p_1, \dots, p_m)$ by the existence of positive semidefinite matrices $C_0, \dots, C_m$ (of corresponding size, see (2.50)), such*

that

$$v_{d_0}^T C_0 v_{d_0} + \sum_{j=1}^{m} v_{d_j}^T C_j v_{d_j} p_j = f - s\mathbf{1}. \tag{2.55}$$

Let f_0 denote the constant term in f, then this infers in particular that the constant term of the left-hand side of (2.55) has to equal $f_0 - s$. The constant term of the polynomial on the left-hand side is given by

$$(C_0)_{11} + \sum_{j=1}^{m} (C_j)_{11} p_{j,0}$$

where $(C_j)_{11}$ denotes the first entry of the matrix C_j for $j = 0, \dots, m$ and $p_{j,0}$ denotes the constant term in the polynomial p_j. Thus, we can replace the cost term s in (2.53) by

$$f_0 - (C_0)_{11} + \sum_{j=1}^{m} (C_j)_{11} p_{j,0}. \tag{2.56}$$

The term (2.56) is linear in the positive semidefinite matrix $C := \operatorname{diag}(C_0, \dots, C_m)$, as well as the constraint (2.55) and the technical constraint that C is block diagonal with blocks of the corresponding size. All in all, this gives a semidefinite programming representation of the SOS formulation (2.53).

Remark 2.38. *Instead of $\mathcal{Q}_d(p_1, \dots, p_m)$ also Schmüdgen's Positivstellensatz and a degree d truncation of the preordering $\operatorname{Pre}(p_1, \dots, p_m)$ can be used. The computational advantage of using Putinar's Positivstellensatz is that fewer sum-of-squares polynomials σ_j are needed and therefore the overall SDP is much smaller.*

Complexity of the Lasserre hierarchy Because the Lasserre hierarchy is based on SDPs its computational complexity is determined by the computational complexity of solving SDPs, which can be done in polynomial time – but the true complexity is hidden in the size of the SDPs!

Remark 2.39 (Complexity of the Lasserre hierarchy)**.** *Solving an SDP with an error less than ε is polynomial in the size of the appearing matrix blocks and $\log(\varepsilon^{-1})$, see for instance [Lasserre 2009]. This is not in contrast with the fact that polynomial optimization covers many NP hard problems. The reason is twofold: First, the hierarchy only provides asymptotic convergence, and second, and more importantly, the complexity is expressed in the size of the matrices appearing in the SDPs in the Lasserre hierarchy. The reason is simple, the blocks of the matrices for the degree d level of the hierarchy are of size*

$$s\left(\frac{d}{2}\right) = \binom{n + \left\lfloor \frac{d}{2} \right\rfloor}{n}$$

and hence grow combinatorially in (d, n).

Even though finite convergence of the Lasserre hierarchy is generic [Lasserre 2015, Theorem 6.5], the practical limitations for the Lasserre hierarchy arise from the

truncation level $d \in \mathbb{N}$ being potentially large. Therefore, obtaining degree bounds for d is a very active and important topic in polynomial optimization. Recently the available bound for such the needed degree $d \in \mathbb{N}$ to guarantee $f \in \mathcal{Q}(p_1, \ldots, p_m)$ was strongly improved [Baldi 2022] from being exponential to being polynomial in the degree of f and its positivity $\frac{f^*}{\sup\limits_{x \in \mathcal{K}(p_1,\ldots,p_m)} f(x)}$.

Moment approach and duality for the semidefinite program It is only natural that the duality between the optimization problem (2.23) and (2.25) relates to the duality between the primal and the dual SDP. The dual SDP to the SOS program has a direct interpretation as a moment problem. Therefore, we remind of the measure formulation for the polynomial optimization problem. It holds

$$\begin{aligned} f^* = \inf_{\mu} \quad & \int_K f \, d\mu \\ \text{s.t.} \quad & \mu \in \mathcal{M}(K)_+ \\ & \mu(K) = 1. \end{aligned} \tag{2.57}$$

As motivated by Theorems 2.32 and Theorem 2.12 we want to re-view the set $M(K)_+$ by linear forms satisfying certain positivity constraints. As usual we assume that K is given by $\mathcal{K}(p_1, \ldots, p_m)$ and that the condition from Putinar's Positivstellensatz is satisfied. The idea is that, by Putinar's Positivstellensatz, a linear form $L : \mathbb{R}[x] \to \mathbb{R}$ is representable by a measure μ if and only if L is non-negative on $\mathcal{Q}(p_1, \ldots, p_m)$.

In the following, we will introduce the needed notation. We begin with representing a linear form $L : \mathbb{R}[x] \to \mathbb{R}$ by the sequence $\mathbf{y} : \mathbb{N}_0^n \to \mathbb{R}$ given by $(\mathbf{y}_\alpha)_{\alpha \in \mathbb{N}_0^n}$ via identifying $L(x^\alpha)$ with $\mathbf{y}_\alpha$. This is motivated by moment sequences of measures, i.e. for a measure $\mu \in M(K)_+$ and $\alpha \in \mathbb{N}_0^n$ we denote by

$$\mathbf{y}_\alpha := \int_K x^\alpha \, d\mu$$

its α-moment. Reversely, for a given sequence $\mathbf{y} = (\mathbf{y}_\alpha)_{\alpha \in \mathbb{N}_0^n}$ we call the corresponding function

$$\ell_{\mathbf{y}} : \mathbb{R}[x] \to \mathbb{R} \text{ given by } \ell_{\mathbf{y}}(x^\alpha) := \mathbf{y}_\alpha \tag{2.58}$$

the Riesz-functional for $\mathbf{y}$. In analogy to Theorem 2.32, the dual version of Putinar's Positivstellensatz [Putinar 1993] states that $\mathbf{y}$ is the moment sequence of a measure $\mu \in M(K)_+$ if and only if its corresponding Riesz-functional $\ell_{\mathbf{y}}$ is non-negative on $\mathcal{Q}(p_1, \ldots, p_m)$. As we liked to do before, we want to express this property in a semidefinite fashion. This will be done using the moment-localizing matrices $M(p_j \mathbf{y})$ for $j = 0, \ldots, m$ defined as follows

$$(M(p_j \mathbf{y})_{\alpha,\beta} := \ell_{\mathbf{y}}^d(g_i j x^\alpha x^\beta) \quad , \quad \alpha, \beta \in \mathbb{N}_0^n \tag{2.59}$$

where we denote $p_0 := \mathbf{1}$ for simpler notation. The infinite matrix $M(g_i \mathbf{y})$ is symmetric and for any vector of coefficients $\mathbf{c} := (c_\alpha)_{\alpha \in \mathbb{N}_0^n}$ with only finitely many

$c_\alpha \neq 0$ it holds

$$\mathbf{c}^T M(p_j\mathbf{y})\mathbf{c} = \ell_{\mathbf{y}}\Big(p_j \underbrace{\big(\sum_{\alpha\in\mathbb{N}_0^n} c_\alpha x^\alpha\big)^2}_{\in\Sigma}\Big) \geq 0. \tag{2.60}$$

We denote the condition (2.60) by $M(p_j\mathbf{y}) \succeq 0$ and the dual version of Putinar's Positivstellensatz reads

$$\left\{\left(\int_K x^\alpha \, d\mu\right)_{\alpha\in\mathbb{N}_0^n} : \mu \in M(K)_+\right\} := \{\mathbf{y} \in \mathbb{R}^{\mathbb{N}_0^n} : M(p_j\mathbf{y}) \succeq 0, j = 0, 1, \ldots, m\}.$$

This motivates the following finite dimensional truncation: For $d \in \mathbb{N}$ we consider truncated moment sequences $\mathbf{y} = (y_\alpha)_{|\alpha|\leq d}$ and define the

$$\mathcal{M}_d(p_1, \ldots, p_m) := \{\mathbf{y} \in \mathbb{R}^{\{\alpha\in\mathbb{N}_0^n : |\alpha|\leq d\}} : M_d(p_j\mathbf{y}) \succeq 0, j = 0, 1, \ldots, m\} \tag{2.61}$$

where the matrices $M_d(p_j\mathbf{y})$ for $j = 0, \ldots, m$ are the truncated moment-localizing matrices, that is

$$M_d(g_j\mathbf{y}) := \ell_{\mathbf{y}}^d(g_j v_{d_j} v_{d_j}^T) \tag{2.62}$$

where $d_j := \lfloor \frac{d-\deg g_j}{2} \rfloor$ and $\ell_{\mathbf{y}}^d : \mathbb{R}[x]_d \to \mathbb{R}$ is defined as the Riesz functional in (2.58) but only for sequences up to order d. Equipped with these notations we can state the moment Lasserre hierarchy [Lasserre 2001].

Definition and Theorem 2.40 (Moment hierarchy for polynomial optimization)**.** *Let $p_1, \ldots, p_m \in \mathbb{R}[x]$ such that the condition from Putinar's theorem is satisfied. Let $f \in \mathbb{R}[x]$. The moment hierarchy associated with the polynomial optimization*

$$\begin{aligned} f^* := \inf_{x\in\mathbb{R}^n} \quad & f(x) \\ s.t. \quad & x \in \mathcal{K}(p_1, \ldots, p_m) \end{aligned}$$

is given by the following sequence programs: For each $d \in \mathbb{N}$ consider the optimization problem

$$\begin{aligned} m_d^* := \inf_{\mathbf{y}} \quad & \ell_{\mathbf{y}}^d(f) \\ s.t. \quad & \mathbf{y} \in \mathcal{M}_d(\mathbf{1}, p_1, \ldots, p_m). \end{aligned} \tag{2.63}$$

Analog to (2.53), the problems (2.63) can be formulated as SDPs and the resulting SDPs are dual to the SDPs for (2.53). By weak duality it holds

$$s_d^* \leq r_d^* \leq f^* \text{ for all } d \in \mathbb{N} \text{ and } r_d^* \nearrow f^* \text{ as } d \to \infty.$$

2.5 The Koopman and Perron-Frobenius operators

What follows is a typical ergodic theoretic construction that represents a dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ by a semigroup of linear operators, and is yet another example

of a lift of a nonlinear problem to a linear formulation. Particularly, with respect to the flavor of transferring notions and properties of the dynamical system to functional analytic ones, the book [Eisner 2015] is fantastic.

The idea of *Koopman semigroup* describes the flow indirectly – via the evolution of its observables, i.e functions $f : X \to \mathbb{R}$. The idea dates back to Bernard Osgood Koopman in 1931 [Koopman 1931] where Koopman made it possible for operator theory to enter the domain of ergodic theory. This immensely fruitful idea had strong impact on related topics varying from number theory, such as the Green-Tao theorem [Tao 2008], over Billiards [Sinai 1989], statistical physics [Sinai 1989], decompositions of dynamical systems (as the Jacobs-de Leeuw-Glicksberg decomposition [Eisner 2015], or the ergodic partition [Mezić 1999]), weather prediction [Froyland 2021] and more recently in data analysis for dynamical systems [Budisic 2012], among many other applications of ergodic theory (see for instance [Çömez 2021] for more (interdisciplinary) examples).

Composition and Koopman operators The Koopman operator is a composition operator, that means, for a given function $\varphi : X \to X$ and a real or complex-valued function $h : X \to \mathbb{C}$ the Koopman operator T of φ acts on h by

$$Th := h \circ \varphi.$$

The mapping T has the following striking properties:

1. *Linearity:* T is linear in h! (But not in φ!)

2. *Contravariance:* Let $\phi : X \to X$ be another map and let $T(\varphi)$ denote the Koopman operator for φ and $T(\phi)$ denote the Koopman operator for ϕ. Then the Koopman operator $T(\phi \circ \varphi)$ for $\phi \circ \varphi$ is given by

$$T(\phi \circ \varphi) = T(\varphi) \circ T(\phi).$$

3. *No loss of information:* Let F be a function space of complex valued functions on X such that F separates points, i.e. for all $x \neq y \in X$ there exists $h \in F$ with $h(x) \neq h(y)$. Then the operator T restricted to F uniquely determines the map φ.

The last point indicates already that the domain $D(T)$ on which we define the Koopman operator T, i.e. the choice of function space F, plays an important role in the study of the Koopman operator. To be more precise, the underlying function space F determines which questions about the map φ we can answer through its Koopman operator T with domain F. For a dynamical system $(X, (\varphi_g)_{g\in G})$ we define the Koopman semigroup for on a function space F as follows.

Definition 2.41 (Koopman semigroup)**.** *Let $(X, (\varphi_g)_{g\in G})$ be a dynamical system and F be a subset of all functions from X to $\mathbb{C}$. We distinguish the two cases of discrete time dynamical systems, i.e. $G = \mathbb{N}$, and continuous dynamical systems, i.e. $G = \mathbb{R}_+$.*

1. *Discrete time systems: Let* $f := \varphi_1 : X \to X$. *The Koopman operator* $T : D(T) \to F$ *is defined as*

$$Th := h \circ f \text{ with } D(T) := \{h \in F : h \circ f \in F\}. \tag{2.64}$$

The Koopman semigroup for discrete time systems is the family of operators $(T^n)_{n\in\mathbb{N}_0}$ *with corresponding domains* $D(T^n) = \{h \in F : h \circ f^n \in F\}$.

2. *Continuous systems: The Koopman semigroup* $(T_t)_{t\in\mathbb{R}_+}$ *is the family of operators* $T_t : D(T_t) \to F$ *for* $t \in \mathbb{R}_+$, *with*

$$T_t h := h \circ \varphi_t \text{ with } D(T) := \{h \in F : h \circ \varphi_t \in F\}. \tag{2.65}$$

Furthermore, it holds $T_0 = \mathrm{Id}$ *and for* $t, s \in \mathbb{R}_+$ *that* $T_{t+s}g = T_t T_s g$ *for all* $g \in \{g \in F : g \in D(T_s) \text{ and } T_s g \in D(T_t)\}$

The function space F should be able to reflect the question we might ask on the dynamical system, such as: Is the flow function φ continuous/smooth, is the system stable, is it energy preserving, do periodic orbits exist, how do invariant sets look like, is the system chaotic? In Examples 2.42 and 2.43 we present two natural choices of function spaces aiming to give functional analytic representations of the above question. From an application perspective, the space F should be easily accessible. At the same time, it should be rich enough to contain all important characteristics of the systems while being adapted enough to clearly single out desirable properties. Some of those demands on the function space F are working against each other and that is why finding a good choice of space F is a subtle task. We address a certain class of candidates for F in Section 6.1.

Two important examples In this thesis, we are mostly concerned with boundedness of the Koopman operators. There are two eminent classical examples of choices for observables, respectively function spaces, for which the Koopman semigroup consists of bounded operators. One is ergodic theory (see Example 2.42), which investigates measure preserving systems - therefore spaces of integrable functions are the right choice for the function space F in Definition 2.41. The other example, Example 2.43, focuses on topological properties. That is why the space of all continuous functions is a natural and good choice for F.

Example 2.42. *Let* X *be a topological space,* $\mathcal{B}$ *the Borel sigma algebra, and* $f : X \to X$ *is assumed to be Borel measurable. We choose* $F = \mathrm{L}^2(X, \mathcal{B}, \mu)$ *where* μ *is an invariant measure, that is* $\mu(f^{-1}(B)) = \mu(B)$ *for all* $B \in \mathcal{B}$. *Then* $T : F \to F$ *is well defined and if* f *is essentially invertible, unitary. This describes the classical ergodic theory setting [Eisner 2015, Sinai 1989] from where many directions can be explored. A central application is spectral decompositions which are obtained through the spectral theorem for normal operators on Hilbert spaces and can be interpreted as diagonalizing the dynamics. The application of the spectral theorem only gives a glimpse of the rich theory that can be developed from this perspective [Eisner 2015].*

Example 2.43. *Here we assume that X is compact and that $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ is a dynamical system. The choice $F = \mathcal{C}(X)$ gives that the Koopman operator T_t is a bounded linear operator for all $t \in \mathbb{R}_+$ (and they carry even more structure as we will see in Theorems 2.45 and 2.47!).*

Both, Examples 2.42 and 2.43 highlight the essential property that the Koopman operator T respectively T_t for $t \in \mathbb{R}_+$ is *linear* despite that we did not enforce any linear structure on the dynamics nor the space itself! This is a consequence of the linearity of composition when the output space Z (here $Z = \mathbb{R}$ or $Z = \mathbb{C}$) is a vector space. Nevertheless, this immediately raises questions about related linear objects or concepts connected to the Koopman operator. We begin with boundedness – for continuous time systems that leads to one-parameter semigroups of bounded operators.

Definition 2.44 (One-parameter Semigroup of bounded operators)**.** *Let V be a normed vector space. A family of operators $(U_t)_{t\in\mathbb{R}_+}$ is called a one-parameter semigroup of bounded operators if*

1. *U_t is a bounded linear operator for all $t \in \mathbb{R}_+$*
2. *The family of operators $(U_t)_{t\in\mathbb{R}_+}$ satisfies the semigroup property*

$$U_0 = \mathrm{Id} \quad \text{and } U_{t+s} = U_t U_s \text{ for all } t, s \in \mathbb{R}_+. \tag{2.66}$$

A one-parameter semigroup $(U_t)_{t\in\mathbb{R}_+}$ of bounded linear operators is called strongly continuous if for all $g \in V$

$$U_t g \to U_0 g = g \text{ as } t \to 0.$$

The rest of this chapter is devoted to presenting some results on dynamical systems from a semigroup theory perspective. For detailed semigroup theory for linear operators, we refer to [Engel 2006].

Motivated by Example 2.43 we focus on the Koopman operator on the space $F = \mathcal{C}(X)$ to further illustrate the "Koopman perspective" on dynamical systems. We begin with stating that the Koopman semigroup on $\mathcal{C}(X)$ for compact sets X and continuous dynamics is indeed a one-parameter semigroup of bounded linear operators and it preserves algebraic and order structures of $\mathcal{C}(X)$.

Theorem 2.45. *Let X be compact and $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system. Then the Koopman semigroup $(T_t)_{t\in\mathbb{R}_+}$ on $\mathcal{C}(X)$ has the following properties*

1. *$T_t : \mathcal{C}(X) \to \mathcal{C}(X)$ is well defined for all $t \in \mathbb{R}_+$ and $(T_t)_{t\in\mathbb{R}_+}$ is a strongly continuous one-parameter semigroup of bounded operators. Furthermore, T_t is contractive with $\|T_t\| = 1$ for all $t \in \mathbb{R}_+$.*
2. *$(T_t)_{t\in\mathbb{R}_+}$ determines φ uniquely, that is for another family $(\phi_t)_{t\in\mathbb{R}_+}$ with Koopman operators $(S_t)_{t\in\mathbb{R}_+}$ it holds $S_t = T_t$ if and only if $\varphi_t = \phi_t$ for any $t \in \mathbb{R}_+$.*
3. *T_t is an algebra homomorphism for all $t \in \mathbb{R}_+$, that is*

$$T_t \mathbf{1} = \mathbf{1} \text{ and } T_t(g \cdot h) = T_t g \cdot T_t h \quad \text{for all } g, h \in \mathcal{C}(X).$$

4. T_t is a Markovian lattice homomorphism for all $t \in \mathbb{R}_+$, that is

$$T_t\mathbf{1} = \mathbf{1} \text{ and } |T_t g| = T_t|g| \quad \text{for all } g \in \mathcal{C}(X).$$

Proof. Properties 1., 3. and 4. are verified by direct calculation. The continuity of φ is only needed to conclude that T_t is well defined and that the Koopman semigroup is strongly continuous. The second statement follows from Urysohn's lemma. □

It's a surprising and beautiful result that parts of Theorem 2.45 can be reversed.

Theorem 2.46 (Koopman operators are exactly the Markov operators and algebra homomorphism; [Eisner 2015, Theorem 4.13], [Arendt 1986, Theorem 3.5])**.** *Let X be compact and $U : \mathcal{C}(X) \to \mathcal{C}(X)$ be a linear operator. Then the following are equivalent*

1. *There exists a map $\varphi : X \to X$ such that $Ug = g \circ \varphi$ holds for all $g \in \mathcal{C}(X)$.*
2. *U is an algebra homomorphism.*
3. *U is a Markovian lattice homomorphism.*

Additional to its elegance the Koopman lifting gives rise to an interesting application, see Theorem 2.47. It concerns continuity of the map φ and states that for semiflows φ joint and separate continuity coincide. On a topological space X a map $\phi : \mathbb{R}_+ \times X \to X$ is called jointly continuous if it's continuous with respect to the product topology on $\mathbb{R}_+ \times X$, the map ϕ is called separately continuous if the map $\phi_t : X \to X$ is continuous for each $t \in \mathbb{R}_+$ and the map $\phi.(x) : \mathbb{R}_+ \to X$ is continuous for $x \in X$. As promised, for semiflows both concepts coincide.

Theorem 2.47 ([Arendt 1986, Lemma 3.2])**.** *Let X be compact and $(\varphi_t)_{t\in\mathbb{R}_+}$ be a semiflow on X, i.e. φ satisfies i., ii., iii. from Definition 2.1. Then φ is jointly continuous if and only if φ is separately continuous.*

It is interesting that, in Theorem 2.47, trying to check the implication, that separate continuity implies joint continuity, is not straightforward. This difficulty is supported by the proof of the above Theorem 2.47 in [Arendt 1986]. The idea in the proof is that joint continuity of φ is equivalent to strong continuity of the Koopman semigroup and separate continuity of φ is equivalent to weak continuity (see [Engel 2006, Theorem 1.6] for the notion of weak continuity of operator semigroups) of the Koopman semigroup. These two properties are still easy to verify, but the strong argument from operator semigroup theory that is used and that immediately concludes Theorem 2.47 is the following [Engel 2006]: A semigroup of bounded linear operators on a Banach space is strongly continuous if and only if it is weakly continuous.

The generator of the Koopman semigroup We want to end this section by introducing the generator of the Koopman semigroup for continuous time systems. The generator of a semigroup is a fundamental object in semigroup theory and its central role is motivated by the interest of studying the whole semigroup by only a single operator.

Definition 2.48 (Generator of a one-parameter semigroup)**.** *Let $(U_t)_{t\in\mathbb{R}_+}$ be a strongly continuous one-parameter semigroup of bounded linear operators on V. The generator $A : D(A) \to V$ is defined as*

$$Av := \lim_{t\to 0} \frac{U_t v - v}{t} \quad \textit{for } v \in D(A) := \{v \in V : \lim_{t\to 0} \frac{U_t v - v}{t} \textit{ exists}\}. \tag{2.67}$$

The generator A of a one-parameter semigroup $(U_t)_{t\in\mathbb{R}_+}$ generates the semigroup $(U_t)_{t\in\mathbb{R}_+}$ in the sense that $U_t = e^{tA}$ in an appropriate meaning and determines the semigroup uniquely [Engel 2006, Chapter II]. In particular, we can interpret a one-parameter semigroup on V as the flow map for the *linear* differential equation

$$\dot{x} = Ax \ \ , x(0) = v \in V.$$

We use this as a good transition to return back to nonlinear differential equations, and only refer to [Engel 2006] for the rich theory on the intimate relation between generators and their corresponding one-parameter semigroups of bounded linear operators.

Let us consider a dynamical system given as the solution of a differential equation $\dot{x} = f(x)$ with $x(0) = x_0 \in \mathbb{R}^n$. Then f is given by

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}t}\varphi_t(x)\big|_{t=0}. \tag{2.68}$$

This analogy to the concept of generator carries over to a relation between the vector field f generating a dynamical system and the generator of the corresponding Koopman semigroup. To convince ourselves of such a relation let $X \subset \mathbb{R}^n$ be compact and positively invariant for (2.68) and $(T_t)_{t\geq 0}$ be the corresponding Koopman semigroup on $\mathcal{C}(X)$. Then the generator A of the Koopman semigroup on $\mathcal{C}(X)$ acts on function $g \in \mathcal{C}^1(X)$ in the following way

$$\begin{aligned} Ag(x) &= \lim_{t\to 0} T_t g\big|_{t=0}(x) = \frac{\mathrm{d}}{\mathrm{d}t}(g \circ \varphi_t)(x)\big|_{t=0} \\ &= \nabla g(x) \cdot \frac{\mathrm{d}}{\mathrm{d}t}\varphi_t(x)\big|_{t=0} = \nabla g(x) \cdot f(x) \end{aligned} \tag{2.69}$$

i.e. A acts on g by applying the vector field f to g. This shows that $\mathcal{C}^1(X)$ is contained in the domain $D(A)$ of the generator A. Because the space $\mathcal{C}^1(X)$ is invariant with respect to the Koopman semigroup the generator is A of the Koopman semigroup is given by the closure of the operator $Bg := \nabla g \cdot f$ for function $g \in \mathcal{C}(X)$ with bounded derivative (see [Engel 2006, Section 3.28] for details and the notion of closure of an operator).

The Perron-Frobenius semigroup – the adjoint semigroup to the Koopman semigroup The adjoint semigroup of the Koopman semigroup is called the Perron-Frobenius semigroup. While a Koopman operator is defined by composition with the flow for any function space, its dual, the Perron-Frobenius operator, is affected by the geometry of the underlying function space. To get acquainted

with the Perron-Frobenius semigroup we revisit the Examples 2.42 and 2.43 of the Koopman semigroup $\mathcal{C}(X)$ and $\mathrm{L}(X)$ that we have discussed earlier.

Definition 2.49 (Perron-Frobenius operator on $\mathrm{L}^2(X)$)**.** *Let X be a topological space, $\mathcal{B}$ the Borel sigma algebra, μ a Borel measure and $f : X \to X$ is assumed to be measurable and essentially invertible. Assume μ is an invariant measure for the discrete dynamical system given by the discrete evolution f. Let T be the Koopman operator on $\mathrm{L}^2(X, \mathcal{B}, \mu)$ from Example 2.42. The Perron-Frobenius operator $P : \mathrm{L}^2(X, \mathcal{B}, \mu) \to \mathrm{L}^2(X, \mathcal{B}, \mu)$ is defined as the adjoint of T, i.e. for all $g, h \in \mathrm{L}^2(X, \mathcal{B}, \mu)$ it holds*

$$\int_X Tg(x)h(x)\, d\mu(x) = \int_X g(x)Ph(x)\, d\mu(x).$$

In the case when X is an open subset of $\mathbb{R}^n$, the invariant measure μ is the Lebesgue measure and $f : X \to X$ is a Diffeomorphism, then the Perron-Frobenius operator can be expressed explicitly by

$$Ph(x) = h(f^{-1}))\det(Df(x)^{-1}).$$

For the Perron-Frobenius operator respectively semigroup on $\mathcal{M}(X)$, i.e. the adjoint to the Koopman operator on $\mathcal{C}(X)$, we can express the semigroup explicitly, too. For doing so we recall the pushforward measure.

Definition 2.50 (Pushforward)**.** *Let (X, Σ_Y, μ) be a measure space and (Y, Σ_Y) be a measurable space. Let $\phi : X \to Y$ be measurable with respect to Σ_X and Σ_Y. The pushforward measure $\phi_\# \mu$ is defined by*

$$\phi_\# \mu(A) := \mu(\phi^{-1}(A))$$

for all $A \in \sigma_Y$.

The duality between composition and pushforward is expressed in the dual relation between the Koopman and Perron-Frobenius operators on $\mathcal{C}(X)$ respectively $\mathcal{M}(X)$.

Definition 2.51 (Perron-Frobenius semigroup on $\mathcal{M}(X)$)**.** *Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a topological dynamical system. The Perron-Frobenius semigroup is the semigroup of linear operators $(P_t)_{t\in\mathbb{R}_+}$ given by*

$$P_t : \mathcal{M}(X) \to M(X), \quad P_f\mu := (\varphi_t)_\# \mu.$$

In contrast to the Koopman semigroup on $\mathcal{C}(X)$ the Perron-Frobenius semigroup on $\mathcal{M}(X)$ is not strongly continuous. The strong continuity is inherited to the adjoint semigroup in reflexive spaces. But in the case of $\mathcal{M}(X)$ we check easily that strong continuity of the Perron-Frobenius semigroup does not hold in general. Assume for the dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ there exists a $x \in X$ such that $\varphi_t(x) \neq x$ for all $0 \leq t \leq \epsilon$ for some $\epsilon > 0$. Then $\|P_f\delta_x - \delta\| = \|\delta_{\varphi_t(x)} - \delta_x\| = 2$ for all $0 \leq t \leq \epsilon$.

As it turns out, see Section 6.1, the class of function spaces that we treat in the following Section 2.6, provides another class of function spaces on which the Perron-Frobenius semigroup can be expressed explicitly as well.

2.6 Reproducing kernel Banach spaces

We begin this section with the much better-known concept of reproducing kernel Hilbert spaces (RKHS). Reproducing kernel Banach spaces (RKBS) is a more recent generalization of RKHS to a Banach space setting.

2.6.1 Reproducing kernel Hilbert spaces

The results and concepts on reproducing kernel Hilbert spaces present in this section can all be found in [Saitoh 2016, Paulsen 2016] or any other book on RKHS theory and we refer to these books for the details and more inspiring properties and applications of RKHS.

There are two main perspectives on RKHS: The first one reflects its name:

> A reproducing kernel Hilbert space is a space of functions for which there exists a reproducing function.

The second one is the reason why RKHS has been so popular and powerful in many machine learning tasks:

> The geometry of the Hilbert space can directly be accessed via a single kernel function $k : X \times X \to \mathbb{C}^n$.

We being with the definition of an RKHS via its reproducing property.

Definition 2.52 (Reproducing kernel Hilbert space)**.** *Let X be a set. A reproducing kernel Hilbert space $\mathcal{H}$ on X is a Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ of complex valued functions on X, such that for any $h \in \mathcal{H}$ the point evaluation $h \mapsto h(x)$ is continuous in $\mathcal{H}$.*

Because for each $x \in X$ the point evaluation is bounded in an RKHS $\mathcal{H}$ on X we can use the Riesz-Fréchet representation theorem to find an element k_x such that

$$h(x) = \langle h, k_x \rangle \text{ for all } h \in \mathcal{H}. \tag{2.70}$$

This is where the name reproducing originates from. Since the element $k_x \in \mathcal{H}$ is a function on X it gives rise to the kernel function of the RKHS.

Definition 2.53 (Kernel of an RKHS)**.** *Let $\mathcal{H}$ be an RKHS on X. The function k defined*

$$k : X \times X \to \mathbb{C},\ k(x, y) := k_x(y) \tag{2.71}$$

is called the kernel for $\mathcal{H}$. We will use the notation $k(x,\cdot)$ for the representing function k_x in the following.

From the reproducing property, we infer the following properties of the kernel k [Saitoh 2016].

Proposition 2.54. *Let k be the kernel for the RKHS $\mathcal{H}$. Let $x, y \in X$. It holds*

1. $k(x,x) = \|k_x\|^2_{\mathcal{H}} \geq 0$.
2. *The operator norm of the point evaluation in x on $\mathcal{H}$ is given by $\sqrt{k(x,x)}$.*
3. $k(x,y) = \overline{k(y,x)}$.
4. *For all $n \in \mathbb{N}$ and $a_1, \ldots, a_n \in \mathcal{C}$ and $x_1, \ldots, x_n \in X$ it holds*

$$\sum_{i,j=1}^{n} a_i \overline{a_j} k(x_i, x_j) \geq 0. \tag{2.72}$$

5. $\mathrm{Span}\{k(x,\cdot) : x \in X\}$ *is dense in $\mathcal{H}$.*
6. *k determines $\mathcal{H}$ uniquely: Let H be another RKHS with the same kernel k, then $H = \mathcal{H}$.*

Even though Proposition 2.54 is basic and is found in any literature on RKHS we state its proof because it is a first illustration of the interaction of kernels and the inner product in $\mathcal{H}$.

Proof. Because $k(x,\cdot)$ belongs to $\mathcal{H}$ we can apply the reproducing property to $k(x,\cdot)$ and get

$$k(x,x) = \langle k(x,\cdot), k(x,\cdot)\rangle = \|k(x,\cdot)\|^2_{\mathcal{H}}.$$

This gives the first statement. The second statement follows from the first and the fact that the point evaluation is represented by $k(x,\cdot)$. The third follows from the symmetry of the inner product because, via the reproducing property, we get

$$k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle = \overline{\langle k(y,\cdot), k(x,\cdot)\rangle} = \overline{k(y,x)}. \tag{2.73}$$

The trick in (2.72) is that the left hand side in (2.72) is the norm squared of the element $\sum_{i=1}^{n} a_i k(x_i,\cdot) \in \mathcal{H}$, thus non-negative. To verify this we compute

$$\begin{aligned}\left\|\sum_{i=1}^{n} a_i k(x_i,\cdot)\right\|^2_{\mathcal{H}} &= \langle \sum_{i=1}^{n} a_i k(x_i,\cdot), \sum_{j=1}^{n} a_j k(x_j,\cdot)\rangle = \sum_{i,j=1}^{n} a_i \overline{a_j} \langle k(x_i,\cdot), k(x_j,\cdot)\rangle \\ &= \sum_{i,j=1}^{n} a_i \overline{a_j} k(x_i, x_j).\end{aligned}$$

For the 5th statement, we use a Hahn-Banach argument. Let $h \in \mathcal{H}$ be in the orthogonal complement of $\mathrm{Span}\{k(x,\cdot) : x \in X\}$. That implies in particular that $0 = \langle h, k(x,\cdot)\rangle = h(x)$ for all $x \in X$. Hence, $h = 0$ in $\mathcal{H}$ and it follows that $\mathrm{Span}\{k(x,\cdot) : x \in X\}$ is dense in $\mathcal{H}$. The last statement follows now because $\mathrm{Span}\{k(x,\cdot) : x \in X\}$ is contained in both Hilbert spaces $\mathcal{H}$ and H and is dense in

both. Further, on $\mathrm{Span}\{k(x,\cdot) : x \in X\}$ the inner products $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ of $\mathcal{H}$ and $\langle\cdot,\cdot\rangle_H$ of H coincide because they can be expressed by

$$\langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}} = k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle_H. \tag{2.74}$$

Thus, $\mathcal{H}$ and H coincide on the dense subset $\mathrm{Span}\{k(x,\cdot) : x \in X\}$ and thus coincide completely. □

An important result in RKHS theory is the second mentioned perspective on RKHS, namely that functions $k : X \times X \to \mathbb{C}$ that satisfy properties 3. and 4. in Proposition 2.54 themselves give rise to an RKHS, see Theorem 2.56.

Definition 2.55. *Let X be a set. We call a function $k : X \times X \to \mathbb{C}^n$ a positive definite kernel if*

i. k is symmetric, i.e.

$$k(x,y) = k(y,x) \text{ for all } x, y \in X.$$

ii. k is positive definite, i.e. for all $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{C}$ and $x_1, \dots, x_n \in X$,

$$\sum_{i,j=1}^{n} a_i \overline{a_j} k(x_i, x_j) \geq 0.$$

We call a kernel strictly positive definite if the left-hand side of the above inequality is strictly positive whenever $(a_1, \dots, a_n) \neq 0$.

The Moore–Aronszajn theorem [Saitoh 2016, Theorem 2.2.], [Paulsen 2016, Theorem 2.14] states that a kernel function in the sense of Definition 2.55 induces a unique RKHS.

Definition and Theorem 2.56 (Moore-Aronszajn)**.** *Each kernel function $k : X \times X \to \mathbb{C}$ induces a unique RKHS $\mathcal{H}_k$ on X such that k is the kernel for $\mathcal{H}_k$. The RKHS $\mathcal{H}_k$ is given by the completion of*

$$\left\{ \sum_{n=1}^{m} a_n k(x_n, \cdot) : m \in \mathbb{N}, x_1, \dots, x_m \in X,\ a_1, \dots, a_m \in \mathbb{R} \right\} \tag{2.75}$$

with respect to the following (well-defined) inner product given by

$$\Big\langle \sum_i a_i k(x_i,\cdot), \sum_j b_j k(x_j,\cdot) \Big\rangle := \sum_{i,j} a_i \overline{b_j} k(x_n, x_m).$$

We denote by $\mathcal{H}_k$ the RKHS corresponding to k.

The close relation between kernels and RKHS is further outlined by the fact that the kernel functions inherit regularity (such as continuity and smoothness) to its RKHS and vice versa [Saitoh 2016, Section 2.1.3]. Another important question concerns the richness of functions in the RKHS. One way to address this question is via embeddings into different well-understood spaces. For continuous kernels that leads to the notion of universal property.

Definition 2.57. *Let X be compact. A reproducing kernel Hilbert space $\mathcal{H}$ on X has the universal property if the embedding $i : \mathcal{H} \hookrightarrow \mathcal{C}(X)$ with $i(g) := g$ is well defined, bounded, and has a dense range.*

Before turning to examples we mention the following short corollary concerning the kernel being strictly positive definite.

Corollary 2.58. *If $\mathcal{H}$ has the universal property then the kernel is strictly positive definite.*

Proof. Assume there exist $x_1, \dots, x_n \in X$ and $a_1, \dots, a_n \in \mathbb{C} \setminus \{0\}$ such that $\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) = 0$. Then it follows

$$0 = \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) = \| \sum_{i=1}^{n} a_j k(x_i, \cdot) \|. \tag{2.76}$$

Hence we have for all $f \in \mathcal{H}$ the functions values $f(x_1), \dots, f(x_n)$ satisfy the following relation

$$\sum_{i=1}^{n} a_i f(x_i) = \langle f, \sum_{i=1}^{n} a_i k(x_i, \cdot) \rangle = 0.$$

Hence a function $g \in \mathcal{C}(X)$ with $\sum_{i=1}^{n} a_i g(x_i) \neq 0$ can not be approximated well by functions in $\mathcal{H}$. That is a contradiction to the universal property. □

Now, we present some examples of RKHS and encounter familiar spaces from an RKHS point of view.

Example 2.59. *The sobolev space $H_0^1(0,1)$ of square integrable functions on $[0,1]$ that vanish in the points $0, 1$ and have square integrable derivatives is an example of an RKHS [Paulsen 2016]. The kernel is given by [Paulsen 2016]*

$$k(x,y) = \begin{cases} (1-y)x, & x \leq y \\ (1-x)y, & x \geq y. \end{cases} \tag{2.77}$$

This example is a special case of RKHSs arising from Sobolev spaces from Example 2.63.

Motivated by Section 2.4 on polynomial optimization, we turn to some RKHS that (densely) contain polynomials.

Example 2.60. *The easiest example of an RKHS containing polynomials is the space $\mathbb{R}[x]_d$ of polynomials up to a fixed degree $d \in \mathbb{N}$ (with an arbitrary inner product). Because this space is a finite dimensional vector space, we can easily turn it into RKHS. A kernel can be chosen as*

$$k(x,y) := (1 + x^T y)^d.$$

Other examples of RKHS that contain all (complex) polynomials are

1. *the Bragmann-Fock space [Rosenfeld 2022] on $X = \mathbb{C}^n$ with kernel*

$$k(x, y) := e^{\bar{x}^T y},$$

where $\bar{x}$ denotes the complex conjugate of $x \in \mathbb{C}$, consisting of the holomorphic functions g on $\mathbb{C}^n$ with finite integrals

$$\int_{\mathbb{C}^n} g(z) e^{-\|z\|^2} \, dz$$

2. *the Bergman space $A^2(G)$ of square-integrable holomorphic functions on a domain $G \subset \mathbb{C}^n$. In the case where $G \subset \mathbb{C}$ is the unit disc, the kernel is given by*

$$k(z, w) = \frac{1}{(1 - \bar{z}w)^2}.$$

3. *Sobolev spaces, which we treat in Example 2.63.*

In the above-mentioned examples, the set of polynomials is even dense with respect to the corresponding topologies; on the contrary, the RKHSs corresponding to the Gaussian kernel treated in Example 2.62, does not contain any non-zero polynomial [Dette 2021].

The connection between RKHS and complex analysis touched in the previous examples is rich [Saitoh 2016, Paulsen 2016]. This story is continued with the following example of Hardy spaces. The Hardy space $H^2(D^n)$, where $n \in \mathbb{N}$ and D is the unit disc $D := \{z \in \mathbb{C} : |z| < 1\}$ in $\mathbb{C}$, consists of all analytic functions on D^n for which the following norm is finite

$$\|g\|_{H^2} := \sup_{0 \le r < 1} \left(\int_0^{2\pi} |g(re^{i\theta})|^2 \, d\theta \right)^{\frac{1}{2}}. \tag{2.78}$$

Example 2.61. *The kernel for the Hardy space $H^2(D^n)$ with inner product*

$$\langle g, h \rangle := \lim_{r \to 1} \int_0^{2\pi} g(re^{i\theta}) \overline{h(re^{i\theta})} \, d\theta$$

is called the Szegö kernel and given by

$$k(z, w) := \prod_{i=1}^{n} \frac{1}{1 - z_i \overline{w_i}}$$

and turns $H^2(D^n)$ into an RKHS. Because functions in $H^2(D^n)$ can be unbounded on D^n, an embedding $H^2(D^n)$ into $\mathcal{C}(\overline{D}^n)$ is not possible. Nevertheless, $H^2(D^n)$ contains the set of polynomials, which is dense in $\mathcal{C}(\overline{D}^n)$.

Another famous example of RKHS, the Gaussian kernel RKHS, is covered by the class of RKHS induced by positive definite functions.

Example 2.62. *A map $u : \mathbb{R}^d \to \mathbb{C}$ is called* positive definite function *if $k(x, y) := u(x-y)$ is a positive definite kernel. Thanks to Bochner's theorem[Katznelson 2004, p. 150], a positive definite function on $\mathbb{R}^d$ can be realized as a Fourier transform of a finite Borel measure. Namely, in the case where u is continuous, u is a positive definite function if and only if there exists a finite Borel measure μ on $\mathbb{R}^d$ such that*

$$u(x) = \widehat{\mu}(x) := \int_{\mathbb{R}^d} \mathrm{e}^{-2\pi i x\cdot\xi}\, d\mu(\xi).$$

Let us consider the case $\mu = w(x)\mathrm{d}x$ where $w \in L^1 \cap L^\infty \setminus \{0\}$ and $w \geq 0$ almost everywhere. Then, if $k(x, y) = \widehat{\mu}(x - y)$, we have

$$\mathcal{H} = \left\{ h \in C^0 \cap L^2 : \widehat{h} \in L^p(w^{-1}) \right\}.$$

A very popular example is the Gaussian kernel $k(x, y) := \exp\left(\frac{\|x-y\|^2}{\sigma^2}\right)$.

Finally, we mention Sobolev spaces, which not only demonstrate the interplay between regularity of the kernel function and the functions in the RKHS but also the need for enough regularity. This example of RKHS recently found strong applications in optimization [Rudi 2020].

Example 2.63 (Sobolev spaces)**.** *For $\Omega \subset \mathbb{R}^n$ open and bounded with C^1 boundary. For $k \in \mathbb{N}$ we denote by $W^{k,2}(\Omega)$ the Sobolev space of square integrable functions with square-integrable weak derivatives up to order k. The space $W^{k,2}(\Omega)$ is a Hilbert space with inner product for $g, h \in W^{k,2}(\Omega)$*

$$\langle g, h\rangle := \sum_{j=0}^{k} \int_\Omega g^{(j)}(x) h^{(j)}(x)\, dx$$

where $g^{(j)}$ respectively $h^{(j)}$ denotes the j-th weak derivative of g respectively h. If $k > \frac{n}{2}$, the Sobolev embedding [Brézis 2011, Section 9.3] tells that $\mathrm{W}^{k,2}(\Omega)$ is a subspace of $\mathcal{C}(\overline{\Omega})$ and there is a constant C with $\|g\|_\infty \leq C\|g\|_{W^{k,2}}$. From this, it follows from the Definition 2.52 that $W^{k,2}(\Omega)$ is an RKHS.

In the previous example, we only mentioned the Sobolev spaces $W^{k,2}(\Omega)$ because they are Hilbert spaces but clearly the spaces $W^{k,p}(\Omega)$ for any $p \geq 1$ are of great importance! This can be seen as motivation for extending the notion of reproducing kernel Hilbert spaces to Banach spaces. We address such an extension to reproducing kernel Banach spaces in the next section.

2.6.2 Reproducing kernel Banach spaces

There are several notions of reproducing kernel Banach spaces, but we follow [Lin 2022] because they provide a unified and general formulation that we find suited for the context of this thesis. The rest of this section follows, up to small deviations, the text [Lin 2022].

The concept of RKBS follows the idea of reproducing kernel Hilbert spaces but aims to extend this concept to (pairs of) Banach spaces. That means we want to

keep the property of continuous point evaluation but, at the same time, we allow different geometries than the ones that arise from an inner product on a Hilbert space.

Definition 2.64 (Reproducing Banach space [Lin 2022])**.** *Let X be a set and $\mathcal{B}$ be a Banach space of complex (or real) valued functions on X. We call $\mathcal{B}$ a reproducing Banach (RBS) space if the point evaluation $\mathcal{B} \ni g \mapsto g(x)$ is continuous for all $x \in X$.*

One of the favorite spaces in this text is $\mathcal{C}(X)$ and it is nice to see that $\mathcal{C}(X)$ naturally is an RBS.

Example 2.65 (A familiar example of an RBS)**.** *Let X be compact. The space $(\mathcal{C}(X), \|\cdot\|_\infty)$ enjoys bounded point evaluation and hence is an RBS. But $\mathcal{C}(X)$ is not an RKHS and it is not obvious how a reproducing kernel function should be defined – but it is possible as we will see in Example 2.69.*

Related to the above example is the universal property that we have stated in Definition 2.57 for RKHS.

Definition 2.66. *Let X be compact. An RBS $\mathcal{H}$ on X has the universal property if the embedding $i : \mathcal{B} \hookrightarrow \mathcal{C}(X)$ with $i(g) := g$ is well-defined, bounded, and has a dense range.*

As mentioned in the above example, we should address how a reproducing kernel function can be incorporated into an RBS. Simply replacing the Hilbert space $\mathcal{H}$ by a Banach space $\mathcal{B}$ in the definition of RKHS does not lead to the "reproducing property" via a kernel yet. The reason is that in Banach spaces the Riesz-Fréchet representation theorem does not apply and the existence of a kernel function is not granted. Therefore, the kernel function is incorporated into the definition of RKBS.

Definition 2.67 (RKBS with Kernels [Lin 2022])**.** *A quadruple $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ is called an RKBS with kernel k if $\mathcal{B}$ is an RBS on a set X, $\mathcal{B}'$ a Banach space of complex (or real) valued functions on a set Y, $\langle\cdot,\cdot\rangle : \mathcal{B} \times \mathcal{B}' \to \mathbb{C}$ a continuous bilinear form and $k : X \times Y \to \mathbb{C}$ is such that for all $x \in X$ we have $k(x,\cdot) \in \mathcal{B}'$ and*

$$g(x) = \langle g, k(x,\cdot)\rangle \quad \text{for all } g \in \mathcal{B}. \tag{2.79}$$

If additionally, $\mathcal{B}'$ is also an RBS and for all $y \in Y$ we have $k(\cdot, y) \in \mathcal{B}$ and

$$h(y) = \langle k(\cdot, y), h\rangle \quad \text{for all } h \in \mathcal{B}' \tag{2.80}$$

then we call $\mathcal{B}'$ an adjoint RKBS. If $Y = X$ we call $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ an RKBS on X with kernel k.

Now, that we have given a definition of an RKBS, we have to show that it indeed generalizes the notion of RKHS. This is done in the following remark.

Remark 2.68. *If $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ is an RKBS with $\mathcal{B} = \mathcal{B}' = \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with scalar product $\langle\cdot,\cdot\rangle$, then $\mathcal{B}$ is an RKHS with kernel k. Vice versa, an RKHS $\mathcal{H}$ with scalar product $\langle\cdot,\cdot\rangle$ with kernel k induces naturally the RKBS $(\mathcal{H}, \mathcal{H}, \langle\cdot,\cdot\rangle, k)$.*

We follow [Lin 2022] and are now able to present the space $\mathcal{C}(X)$ as an RKBS. Interestingly, the kernel is not unique as we will also see in the example of $\mathcal{C}(X)$.

Example 2.69 ($\mathcal{C}(X)$ as an RKBS.)**.** *Let X be compact and $\mathcal{C}(X)$ equipped with the supremum norm $\|\cdot\|_\infty$. By Example 2.65, the space $\mathcal{C}(X)$ is an RBS but there is freedom in the choice of $\mathcal{B}'$ and the kernel. The main idea for inducing kernels is the use of kernel mean embeddings [Lin 2022]: Let $k : X \times X \to \mathbb{R}$ continuous such that $\mathrm{Span}\{k(\cdot, x) : x \in X\}$ is a dense subset of $\mathcal{C}(X)$. We define the RKBS in the following way: Let $\mathcal{B} = \mathcal{C}(X)$ and $\mathcal{B}'$ be the space of kernel mean embeddings, i.e.*

$$\mathcal{B}' = \left\{ g_\mu : \mu \in M(X),\ g_\mu(x) := \int_X k(y, x)\, d\mu(y) \right\} \tag{2.81}$$

and the bilinear form $\langle \cdot, \cdot \rangle : \mathcal{B} \times \mathcal{B}' \to \mathbb{R}$ is given by

$$\langle h, g_\mu \rangle := \int_X h\, d\mu. \tag{2.82}$$

The condition that $\mathrm{Span}\{k(x, \cdot) : x \in X\}$ is dense in $\mathcal{C}(X)$ guarantees that the bilinear form (2.82) is well defined. Then $(\mathcal{B}, \mathcal{B}', \langle \cdot, \cdot \rangle, k)$ is an RKBS with kernel k [Lin 2022]. To verify that k is a kernel let $x \in X$. Because for $y \in X$ we have $k(x, \cdot)(y) = k(x, y) = \int_X k(x, z)\, d\delta_y(z)$, i.e. $k(x, \cdot) = g_{\delta_x}$ (with the notion from (2.81)). It follows for all $h \in \mathcal{C}(X)$

$$\langle h, k(x, \cdot) \rangle = \langle h, g_{\delta_x} \rangle = \int_X h\, d\delta_x = h(x).$$

Further, k is an adjoint kernel as well. To check this let $\mu \in M(X)$. For g_μ we have

$$g_\mu(x) = \int_X k(y, x)\, d\mu(y) = \int_X k(\cdot, x)\, d\mu = \langle k(\cdot, x), g_\mu \rangle.$$

Examples of such kernels k for $X = [0, 1]$ are

$$k(x, y) = 1 - |x - y|,\ k(x, y) = e^{xy} \text{ or } k(x, y) = (1 + y)^x.$$

In the above example, we observe a close relation between $\mathcal{B}'$ and the dual space $M(X)$ of $\mathcal{B} = \mathcal{C}(X)$, where it holds $\mathcal{B}' \cong \mathcal{M}(X)$ through identifying g_μ with μ. In general, this relation is less strong and we only get an embedding of $\mathcal{B}'$ into $\mathcal{B}^*$. This is discussed in the following remark.

Remark 2.70. *The continuous bilinear form $\langle \cdot, \cdot \rangle$ induces a map*

$$\phi : \mathcal{B}' \to \mathcal{B}^*, h \mapsto \langle \cdot, h \rangle \tag{2.83}$$

where $\mathcal{B}^*$ denotes the dual space of $\mathcal{B}$. The map ϕ is continuous due to the continuity of $\langle \cdot, \cdot \rangle$ and represents how much $\langle \cdot, \cdot \rangle$ differs from the natural pairing of $\mathcal{B}$ and its dual $\mathcal{B}^*$. For RKHS, we have $\mathcal{H} = \mathcal{H}'$ and the bilinear form is given by the inner

product and hence the map ϕ is the natural isomorphism between $\mathcal{H}$ and its dual $\mathcal{H}^*$. In contrast, for RKBS with kernel the map ϕ from (2.83) does not need to be isomorphic.

Remark 2.70 raises the question if the canonical isomorphism between a Hilbert space and its dual space has its analog for RKBS with kernel (2.83). The class of RKBS, which is closer to the Hilbert spaces in that sense, is the class of reflexive RKBS.

Definition 2.71 (Reflexive RKBS)**.** *Let $\mathcal{B}$ and $\mathcal{B}'$ be Banach spaces and $\langle\cdot,\cdot\rangle$ be a continuous bilinear form on $\mathcal{B}\times\mathcal{B}'$. We call $(\mathcal{B},\mathcal{B}',\langle\cdot,\cdot\rangle)$ a dual pairing if $\mathcal{B}'$ is isomorphic to $\mathcal{B}^*$ via the map ϕ from (2.83). We call $(\mathcal{B},\mathcal{B}',\langle\cdot,\cdot\rangle)$ reflexive if it is a dual pairing and $\mathcal{B}$ is reflexive.*

Pullback kernel Via a pullback, we want to transfer RKBS structure. This will play an important role when we investigate conjugated dynamical systems.

Lemma 2.72 (Pullback kernel)**.** *Let $(\mathcal{B},\mathcal{B}',\langle\cdot,\cdot\rangle,k)$ be an RKBS on X with kernel k and $\phi: Y\to X$ be a bijective map. Then $(\mathcal{B}_\phi,\mathcal{B}'_\phi,\langle\cdot,\cdot\rangle_\phi,k_\phi)$ is an RKBS on Y with kernel for*

$$\mathcal{B}_\phi := \{g\circ\phi : g\in\mathcal{B}\} \text{ with norm } \|h\|_{\mathcal{B}_\phi} := \|h\circ\phi^{-1}\|_{\mathcal{B}} \tag{2.84}$$

and

$$\mathcal{B}'_\phi := \{g\circ\phi : g\in\mathcal{B}'\} \text{ with norm } \|h\|_{\mathcal{B}'_\phi} := \|h\circ\phi^{-1}\|_{\mathcal{B}'}, \tag{2.85}$$

with bilinear form

$$\langle h,h'\rangle_\phi := \langle h\circ\phi^{-1}, h'\circ\phi^{-1}\rangle \tag{2.86}$$

and kernel

$$k_\phi : Y\times Y\to\mathbb{K},\quad k_\phi(y_1,y_2) := k(\phi(y_1),\phi(y_2)), \tag{2.87}$$

where $\mathbb{K}$ denotes $\mathbb{R}$ or $\mathbb{C}$. Further, the composition operator T_ϕ with $T_\phi g := g\circ\phi$ defines isometric isomorphisms between $\mathcal{B}$ and $\mathcal{B}_\phi$ and $\mathcal{B}'$ and $\mathcal{B}'_\phi$ and preservers the bilinear forms, i.e. $\langle T_\phi g, T_\phi h\rangle_\phi = \langle g,h\rangle$ for all $g\in\mathcal{B}$ and $h\in\mathcal{B}'$.

Proof. By definition of $\mathcal{B}_\phi$ and $\mathcal{B}'_\phi$ it follows that T_ϕ induces isometric isomorphisms from $\mathcal{B}$ to $\mathcal{B}_\phi$ and from $\mathcal{B}'$ to $\mathcal{B}'_\phi$. Hence, $\mathcal{B}_\phi$ and $\mathcal{B}'_\phi$ are Banach spaces (of functions on Y). Similarly, we see that $\langle T_\phi g, T_\phi h\rangle_\phi = \langle g,h\rangle$, and in particular $\langle\cdot,\cdot\rangle_\phi$ is continuous on $\mathcal{B}_\phi\times\mathcal{B}'_\phi$. It remains to check the reproducing property, this as well follows from the (pull back) definition, namely, we have for all $h=g\circ\phi\in\mathcal{B}_\phi$ and $y\in Y$

$$\begin{aligned} h(y) &= g(\phi(y)) = \langle g, k(\phi(y),\cdot)\rangle = \langle T_\phi g, T_\phi k(\phi(y),\cdot)\rangle_\phi \\ &= \langle h, k_\phi(y,\cdot)\rangle. \end{aligned}$$

□

Revisiting the RKHS examples As we have see in Remark 2.68, RKBS with kernels is a generalization of an RKHS and the example of viewing $(\mathcal{C}(X),\|\cdot\|_\infty)$ as an RKBS with kernel while $(\mathcal{C}(X),\|\cdot\|_\infty)$ is not even a Hilbert space shows that

the extension is strict. Further generalizations are also applicable to some of the Examples 2.60, 2.61 and 2.63.

Remark 2.73. *The examples including square integrable functions, such as the Bergman space $A^2(G)$ from Example 2.60, the Hardy space $H^2(D^n)$ from Example 2.61, or Sobolev spaces from Example 2.63, can be generalized to classes of different regularity, by choosing L^p regularity instead of L^2 regularity. Keeping the same bilinear form and kernel combined with L^p duality theory provides an RKBS with kernel.*

We illustrate the above remark with the example of Sobolev spaces.

Example 2.74 (Sobolev spaces revisited)**.** *For $\Omega \subset \mathbb{R}^n$ open and bounded with C^1 boundary. For $k \in \mathbb{N}$ and $p \in [2, \infty)$ we denote by $W^{k,p}(\Omega)$ the Sobolev space of functions with p-integrable weak derivatives up to order k. The space $W^{k,p}(\Omega)$ is a Banach space with dual space $\mathrm{W}^{k,q}(\Omega)$ for $\frac{1}{p} + \frac{1}{q} = 1$. If $k > \frac{n}{p}$, the Sobolev embedding [Brézis 2011, Section 9.3] tells that $\mathrm{W}^{k,p}(\Omega)$ is a subspace of $\mathcal{C}(\overline{\Omega})$ and there is a constant C with $\|g\|_\infty \leq C\|g\|_{W^{k,p}}$. In this case, taking the kernel k obtained from the RKHS situation $W^{k,2}(\Omega)$, choosing $\mathcal{B} = W^{k,p}(\Omega)$, $\mathcal{B}' := W^{k,q}(\Omega)$ and the dual pairing*

$$\langle g, h\rangle := \sum_{j=0}^{k} \int_\Omega g^{(j)}(x) h^{(j)}(x)\, dx$$

gives an RKBS with kernel $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$.

Adjoint operators in RKBS We end this Section by adapting the notion of adjoint operators to a setting that covers RKBS. Let $\mathcal{B}$ and $\mathcal{B}'$ be Banach spaces on which we have a continuous bilinear form $\langle\cdot,\cdot\rangle : \mathcal{B} \times \mathcal{B}' \to \mathbb{C}$. We begin by defining when a set $W \subset \mathcal{B}$ is called dense with respect to $\langle\cdot,\cdot\rangle$. We say that $W \subset \mathcal{B}$ is dense with respect to $\langle\cdot,\cdot\rangle$ if

$$\langle w, g\rangle = 0 \text{ for all } w \in W \text{ implies } g = 0. \tag{2.88}$$

If $\mathcal{B}$ is a Hilbert space, $\mathcal{B}'$ its dual space and $\langle\cdot,\cdot\rangle$ the inner product, then (2.88) characterizes density in $\mathcal{B}$. Also parallel to Hilbert spaces, we define adjoint operators with respect to a bilinear as follows.

Definition 2.75 (Adjoint operator)**.** *Let B and B' be Banach spaces with a continuous bilinear form $\langle\cdot,\cdot\rangle : \mathcal{B} \times \mathcal{B}' \to \mathbb{C}$. Let T be a linear operator $T : D(T) \to B$, such that $D(T)$ is dense in $\mathcal{B}$ with respect to $\langle\cdot,\cdot\rangle$. We call $T' : D(T') \to B'$ the adjoint operator of T with respect to $\langle\cdot,\cdot\rangle$ if $D(T')$ is given by*

$$D(T') := \{y \in Y : \exists z \in Z \text{ with } \langle Tx, y\rangle = \langle x, z\rangle \text{ for all } x \in D(T)\}$$

and it holds

$$\langle Tx, y\rangle = \langle x, T'y\rangle \qquad \text{for all } x \in D(T) \text{ and } y \in D(T'). \tag{2.89}$$

In Section 6.1, we investigate continuity of the Koopman and Perron-Frobenius operator acting on RKBS. It turns out that we cannot expect continuity in general. Therefore, we turn to a weaker notion of continuity of an operator, namely, closedness, and relate it to the domain of the adjoint operator.

Definition 2.76 (Closed operator)**.** *Let X and Y be Banach spaces. A linear operator $A : D(A) \to Y$ with domain $D(A) \subset X$ is called closed if*

$$D(A) \ni x_n \to x \text{ and } Ax_n \to y \text{ as } n \to \infty \text{ implies } x \in D(A) \text{ and } Ax = y.$$

We call the operator $(A, D(A))$ closable if it has an extension $\bar{A} : D(\bar{A}) \to Y$, i.e. $D(A) \subset D(\bar{A})$ and $\bar{A} = A$ on $D(A)$, which is a closed operator.

For Hilbert spaces, the adjoint operator of a densely defined operator is uniquely defined and always closed [Rudin 1991, Theorem 3.9]. We want to transfer this result to RKBS and begin with the following lemma.

Lemma 2.77. *Let $T : B \supset D(T) \to B$ be a densely defined linear operator. Then T' is uniquely determined and a closed operator.*

Proof. For all $y \in D(T')$ we have for all $x \in D(T)$ that $\langle Tx, y\rangle = \langle x, T'y\rangle$. Since $D(T)$ is dense in $\mathcal{B}$ with respect to $\langle\cdot,\cdot\rangle$, the element $T'y$ is uniquely determined. To check closeness, let $x_n \in D(T')$ with $x_n \to x$ and $Tx_n \to y$ as $n \to \infty$. Then $\langle z_1, x_n\rangle \to \langle z_1, x\rangle$ and $\langle z_2, T'x_n\rangle \to \langle z_2, y\rangle$ for all $z_1, z_2 \in \mathcal{B}$. Thus, for all $v \in D(T)$ we have

$$\langle v, y\rangle \quad \leftarrow \quad \langle v, T'x_n\rangle = \langle Tv, x_n\rangle \to \langle Tv, x\rangle. \tag{2.90}$$

Since T is densely defined (with respect to $\langle\cdot,\cdot\rangle$), it follows $x \in D(T')$ and $T'x = y$. □

The following Lemma tells us how we can retrieve information about closedness of T from the domain of its adjoint.

Lemma 2.78. *Let $T : B \supset D(T) \to B$ be a densely defined operator. If T' is densely defined then T is closable.*

Proof. If T' is densely defined we can build its adjoint with respect to the bilinear form $\langle\cdot,\cdot\rangle' : \mathcal{B}' \times \mathcal{B} \to \mathbb{K}$ defined by $\langle b', b\rangle := \langle b, b'\rangle$. Then the adjoint T'' of T' is closed by Lemma 2.77. We claim that T'' is an extension of T. Let $x \in D(T)$, i.e. we have for each $y \in D(T')$ that $\langle Tx, y\rangle = \langle x, T'y\rangle = \langle T'y, x\rangle'$. But this exactly states that $x \in D(T'')$ with $T''x = Tx$. □

Chapter 3

Overview, contribution and embedding into existing work

In this chapter, an overview of the thesis is presented. We want to motivate the developed results and give insights into this thesis by outlining how it embeds into existing work and distinguishing its contributions. This section aims at taking a step back and explaining the ideas using only as much notation as needed. A rigorous treatment of the mentioned work in this thesis will be done in the corresponding chapters that follow. This thesis covers four subjects, which we relate in Figure 3.1,

- Sparsity for dynamical systems
- Linear programming problem formulations for computing the global attractor
- Koopman and Perron-Frobenius operators on reproducing kernel Banach spaces
- Sparsity exploitation for the Koopman and Perron-Frobenius operator and occupation measure linearization techniques.

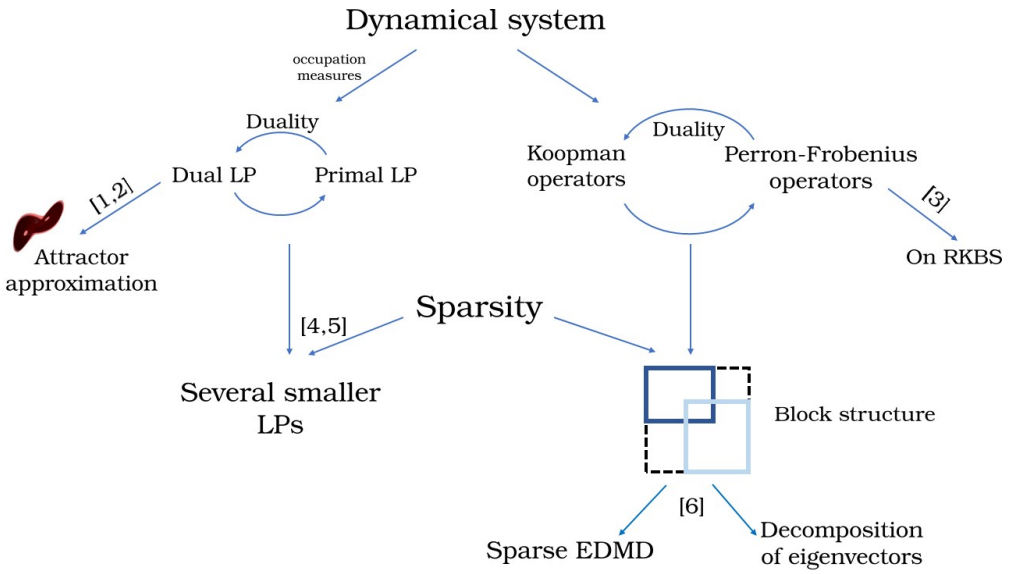

Figure 3.1: Relation between the several topics in this thesis. For the sake of space, the references [1,2] represent the references [Schlosser 2021, Schlosser 2022a], [3] represents [Ikeda 2022b], [4] represents [Schlosser 2020], [5] represents [Wang 2021b] and [6] represents [Schlosser 2022b].

3.1 Sparsity structures for dynamical systems

Main contribution: We define a notion of subsystems based on certain sparse structures. We show that many important objects for dynamical systems decompose according to such subsystems, see Theorem 4.23 in Chapter 4. Finally, we demonstrate that subsystems can be efficiently found via the sparsity graph of the dynamics.

The definitions, notations, concepts, and results concerning dynamical, which are important systems for this section, are provided in the preliminary Sections 2.1. A detailed presentation of the results stated in this section is provided in Chapter 4. The presented ideas are based on and extend the text [Schlosser 2020].

To get acquainted with the notion of sparsity that we develop, we will be guided by two examples displayed in Figures 3.2 and 3.3.

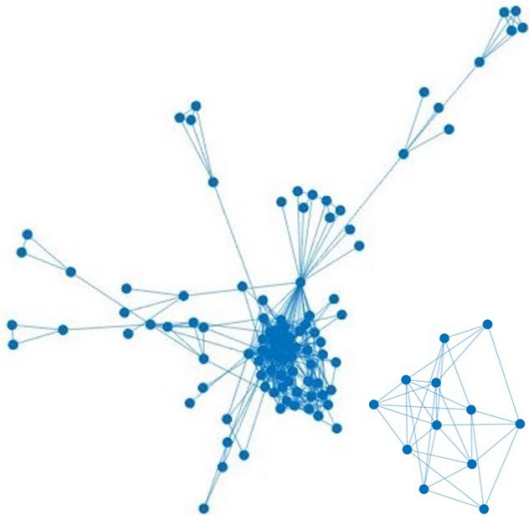

Figure 3.2: Example of a social network graph.

Figure 3.2 is an illustrative representation of a social graph. The nodes represent people and a connecting edge between two people states that their opinion or behavior (symmetrically or asymmetrically) influences the other's.

Applications of (social) networks include epidemiology, political influence, economy, and climate, to name only a few. Due to the complexity (including their size) of such networks, reduction techniques are necessary for their understanding and analysis.

In Figure 3.2 we can identify several distinguished structures of the network. For instance, the separate network on the right bottom corner, which do not have

any connecting edge to the rest of the network. Such an autonomous part is a first step towards what we will call subsystems and relates closely to causality for time series as in [Granger 1969], [Peters 2022]. We make the following informal definition of subsystems.

Definition 3.1 (Informal definition of subsystems)**.** *A subsystem of a dynamical system is an ensemble of states that evolve independently from the rest of the system.*

In contrast to the subsystem in the right bottom corner in Figure 3.2, another structure that catches the eye is the dense accumulation of points in the center of the figure. These highly connected nodes form a very non-sparse part of the network. One objective will be to characterize what hampers sparsity and the existence of subsystems. We will see that the antagonists to sparsity are cycles. Thus we should search for subsystems where there are no (or few) cycles. Graphs without cycles are trees and in the social network graph from Figure (3.2) we can recognize tree-like parts in the out-reaching "branches" of the network. Possible scenarios producing such branches are situations where information flows in a directed way from some nodes (such as "influencers", politicians, celebrities, etc.) to their "followers". Indeed, sparsity is inherent there because the followers' opinions do not actively influence each other. To allow such asymmetric scenarios where the action/opinion of person A affects person B but not vice versa requires the social network graph to be directed. In order to keep the illustration simpler we omitted to draw arrows instead of edges in Figure 3.2.

Figure 3.3 shows a directed graph, representing the energy distribution in an electrical power grid. An arrow from a node A to a node B indicates that A provides energy for B. We thank Edgar Fuentes for pointing out to us that radial distribution networks provide common examples of networks with tree structures, see [Chakravorty 2001].

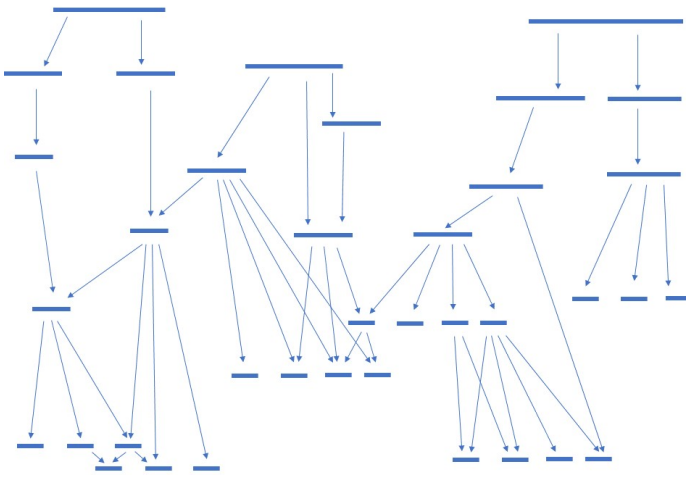

Figure 3.3: Example of a radial power grid model.

The situation in the power grid network in Figure 3.3 is less evident but autonomous structures can be found as well – the reason is that the graph in Figure

3.3 is directed and the power grid is radial, that is, there is no back-flow of energy, i.e. no cycles in the network graph. We can distinguish nodes that play outstanding roles concerning the sparse distribution of energy. These can be found in the top and bottom rows of the power grid network in Figure 3.3. The former nodes evolve independently from the rest – there is no incoming edge – and the latter ones affect no other node – there is no outgoing edge. Those nodes allow us to recognize certain subsystems. In Figure 3.4 two subsystems are presented – one consists of all orange nodes the other consists of the red nodes. Indeed, from the graph, we conclude that there is no other node outside the subsystem that influences the subsystem.

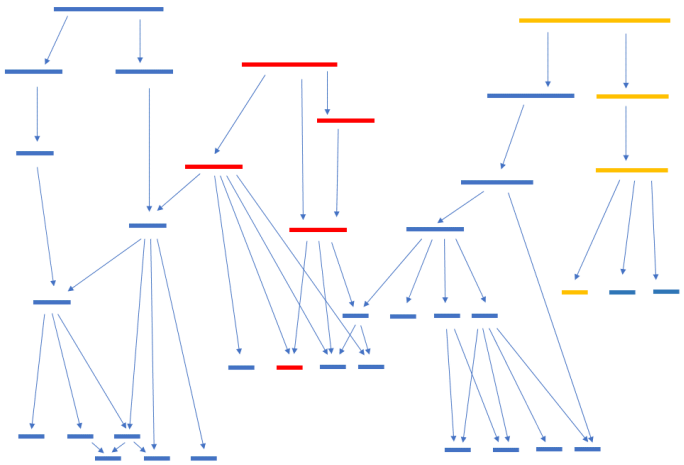

Figure 3.4: Examples (orange, red) of subsystems in the power grid from Figure 3.3.

Other large-scale networks that tend to exhibit subsystems are communication networks, interacting networks, hierarchical networks, citation networks, the internet, food web, and others. Some of the mentioned examples are discussed in [Strogatz 2001]. Another interesting class of systems where subsystems appear can be found in (distributed) multicellular programming [Regot 2011], [Tamsir 2011] or supply networks such as water networks, data routing, and logistic networks [Bullo 2019] and traffic networks [Li 2022], [Kwee 2018].

In Figure 3.5 we summarize our approach toward a practical sparse decomposition of dynamical systems. It begins with identifying inherent subsystems and decoupling the dynamics accordingly. In the next step each subsystem is analyzed separately before, in the final step, the analysis is merged into an analysis of the whole system.

To specify the approach to sparse dynamical systems indicated in Figure 3.5, we divide it into the following main categories:

1. *Definition of subsystems*

2. *Decomposition of the dynamical system according to subsystems and computational application*

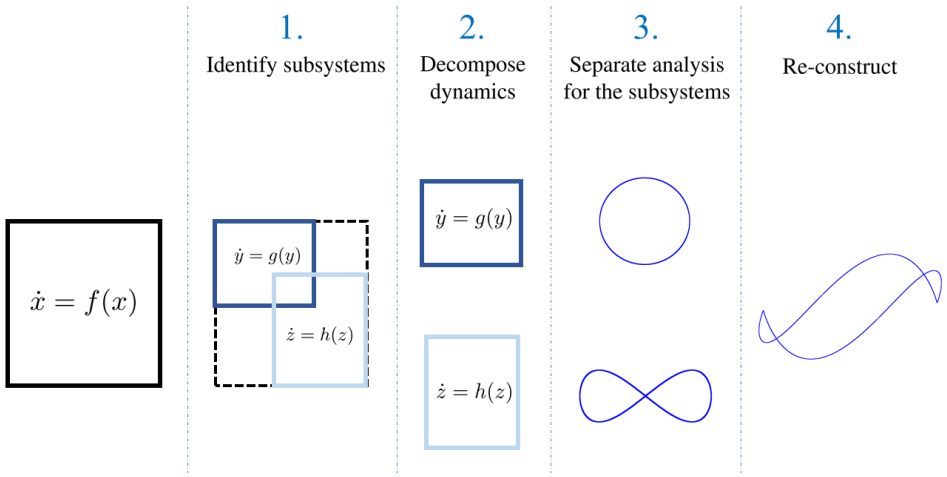

Figure 3.5: Illustration of a decomposition procedure for sparse dynamical systems.

3. *Identification of subsystem via the sparsity graph*

4. *Extension to other classes of systems*

5. *Limitations for sparse decompositions*

6. *Comparison with other existing decomposition methods*

In the first part we will introduce the notion of subsystems, building up on investigations by [Chen 2018]. We show how the interplay between subsystems and the whole system induces decomposition of the dynamical system. We base our investigation on important characteristics of the dynamical system, such as equilibrium points, invariant sets, stability analysis, etc., and continue via Lyapunov functions, stable manifolds, etc. One of our main results in this chapter is that many of those (but not all!) characteristics can be decomposed according to the sparse structure as well. Those decompositions give rise to decoupling computational procedures for the corresponding objects. We specify such decompositions in Section 5.3 for the LPs from Chapter 5 and in Section 6.3.4 for computational applications to the Koopman operator. The definition of subsystems that we give in Definition 3.3 is coordinate-dependent, but in Section 4.5 we present a coordinate-free formulation of subsystems. The reason for the coordinate dependent definition is that it makes available a tool from discrete maths, namely the so-called sparsity graph G_f of the dynamics f. Using the sparsity graph, the identification of subsystems is translated into the language of graphs and simple algorithms for identifying subsystems can be formulated. We continue with extensions to other classes of dynamical systems, such as control systems, time-delay systems, and stochastic ordinary differential equations. Finally, we give a comparison to existing decomposition techniques for dynamical systems and point out how they could be combined with the methods from this thesis.

Subsystems

In this section, we want to define subsystems of a given continuous time dynamical system

$$\dot{x} = f(x) \text{ on } \mathbb{R}^n \text{ for } f = (f_1, \dots, f_n) : \mathbb{R}^n \to \mathbb{R}^n \text{ locally Lipschitz continuous.}$$

We restrict to continuous time dynamical systems but discrete time dynamical systems $x_{k+1} = f(x_k)$ can be treated in the same way. We will often use the notation $[n]$ for $\{1, \dots, n\}$ for natural numbers $n \in \mathbb{N}$.

The most basic example of a system that decomposes into several systems is a product system. This will act as the guiding idea for our notion of subsystems for systems that are not product systems.

Example 3.2 (Product systems and subsystems)**.** *Let $n_1, n_2 \in \mathbb{N}$ and for $i = 1, 2$ let $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ be a Lipschitz continuous vector fields. Consider the dynamical systems with dynamics $\dot{x}_i = f_i(x_i)$ with flows $\varphi^{(i)}$. The product system is defined as the dynamical system on $\mathbb{R}^{n_1+n_2} \cong \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with dynamics $f = f_1 \otimes f_2$ defined by*

$$f =: \mathbb{R}^{n_1+n_2} \to \mathbb{R}^{n_1+n_2} \text{ with } f((x_1, x_2)) := (f_1(x_1), f_2(x_2)). \tag{3.1}$$

The corresponding flow φ is given by

$$\varphi = \varphi^{(1)} \otimes \varphi^{(2)}. \tag{3.2}$$

In the above example we recover the flows $\varphi^{(1)}$ and $\varphi^{(2)}$ from the global flow φ by projecting onto the corresponding coordinates. For a subset of indices $I \subset [n]$ we denote by $\mathbb{R}^I$ the space

$$\mathbb{R}^I := \{(x_i)_{i \in I} : x_i \in \mathbb{R}\}. \tag{3.3}$$

The corresponding natural projections of $\mathbb{R}^n$ onto the canonical coordinates indexed by I are denoted by

$$\Pi_I : \mathbb{R}^n \to \mathbb{R}^I,\ \Pi_I(x_1, \dots, x_n) = (x_i)_{i \in I}. \tag{3.4}$$

The essential observation from Example 3.2 is that the two components $\varphi^{(1)}$ and $\varphi^{(2)}$ evolve independently because there is no coupling between their dynamics f_1 and f_2. This property motivates our definition of systems.

Definition 3.3 (Induced (sub)system)**.** *Let $J \subset \mathbb{N}$ and $f : \mathbb{R}^J \to \mathbb{R}^J$. A subsystem of a dynamical system on $\mathbb{R}^J$ with dynamics f is a set of states $(x_i)_{i \in I}$ with $I \subset J$ such that $f_I := \Pi_I \circ f = (f_i)_{i \in I}$ only depends on the states $(x_i)_{i \in I}$ indexed by I. In that case, we say the pair (I, f_I), or just I when f is clear from the context, induces a subsystem of (J, f).*

Remark 3.4. *We specify, what it means for a function to depend only on certain states, in Section 4 around (4.2).*

Our investigation of subsystems was strongly motivated by the text [Chen 2018] where a decomposition of the reachability set is given for systems of the form as in

Example 3.5. This example shows the most basic case, in which subsystems strictly generalize product systems.

Example 3.5. *The system on $\mathbb{R}^{n_1+n_2+n_3}$, where we write $x = (x_1, x_2, x_3)$ with $x_i \in \mathbb{R}^{n_i}$ for $i = 1, 2, 3$, with dynamics*

$$\begin{aligned} \dot{x}_1 &= f_1(x_1) \\ \dot{x}_2 &= f_2(x_1, x_2) \\ \dot{x}_3 &= f_3(x_1, x_3) \end{aligned} \tag{3.5}$$

Then the pairs

$$(I_1, f_1),\ (I_2, (f_1, f_2)) \text{ and } (I_3, (f_1, f_3))$$

induce non-trivial subsystems, where $I_1 = \{1, \dots, n_1\}, I_2 = \{1, \dots, n_1 + n_2\}, I_3 = \{1, \dots, n_1, n_1+n_2+1, \dots, n_1+n_2+n_3\}$. Note that the whole system is not a product system.

Remark 3.6. *A first observation is that the sets I that induce a subsystem form a topology on $[n]$, see Lemma 4.4, that is intersections and unions of subsystems form again subsystems.*

Guided by product systems from Example 3.2, the idea of a subsystem is that we can treat it as a lower dimensional dynamical system.

Let (I, f_I) induce a subsystem. we view f_I as a vector field on $\mathbb{R}^I$ by identifying f_I with the map from $\mathbb{R}^I$ to $\mathbb{R}^I$ given by

$$(x_i)_{i\in I} \mapsto (f_i(x))_{i\in I} \text{ where } x = (x_1, \dots, x_n) \in \mathbb{R}^n \text{ satisfies } \Pi_I(x) = (x_i)_{i\in I}.$$

For instance we can choose $x = (x_1, \dots, x_n)$ with $x_j = 0$ whenever $j \notin I$.

The semiflow induced by $f_I : \mathbb{R}^I \to \mathbb{R}^I$ is denoted $\varphi^I_t : \mathbb{R}^I \to \mathbb{R}^I$ for $t \in \mathbb{R}_+$. The flows of the whole system and of a subsystems are connected as follows, see Corollary 4.6,

$$\varphi^I_t \circ \Pi_I = \Pi_I \circ \varphi_t. \tag{3.6}$$

This is rephrased by the commuting diagram in Figure 3.6.

In the next section we show how subsystems can be used to decompose many (but not all!) important objects in the analysis of dynamical systems.

Decomposition based on subsystems

In this section, we follow two goals: The first one is to show that many properties of the whole system are inherited by its subsystems. The second and more interesting one is to reverse this process. That is, we want to infer properties of the whole system from properties of its subsystems.

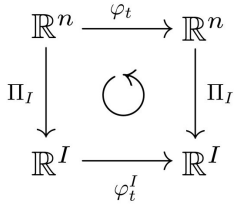

Figure 3.6: Subsystems induce the above commuting diagram.

The simplest objects to consider in that context are equilibrium points $x^* \in \mathbb{R}^n$ with $f(x^*) = 0$. Let $I_1, \ldots, I_k$ induce subsystems. It is obvious that $f_{I_l}(\Pi_{I_l}(x^*)) = 0$ for all $l = 1, \ldots, k$, i.e. we have for the set $\mathcal{E}$ of equilibrium points that

$$\mathcal{E} \subset \{x \in \mathbb{R}^n : f_{I_l}(\Pi_{I_l}(x)) = 0 \text{ for all } l = 1, \ldots, k\}. \tag{3.7}$$

To guarantee equality of the two sets in (3.7) we need to make sure that each state of the system is contained in at least one of the subsystems. This is expressed in the following condition, which is essential for the decomposition results presented in this thesis.

For a dynamical system on $\mathbb{R}^n$, we say that $(I_1, f_{I_1}), \ldots, (I_k, f_{I_k})$ induce a subsystem decomposition if each pair (I_l, f_{I_l}) induces a subsystem and

$$I_1 \cup \ldots \cup I_k = [n] \tag{3.8}$$

At the example of determining equilibria points, we illustrate how the condition (3.8) is applied to gain information about the whole system from its subsystems only. We show that the two sets in (3.7) coincide. If (3.8) is satisfied and $x^* \in \mathbb{R}^n$ is such that $\Pi_{I_l}(x^*)$ is an equilibrium point for subsystem induced by I_l, i.e.

$$f_{I_l}(\Pi_{I_l}(x^*)) = 0 \text{ for all } l = 1, \ldots, k,$$

then x^* is an equilibrium point for the whole system. In other words, we can identify all equilibrium points $\mathcal{E}$ by

$$\mathcal{E} = \{x \in \mathbb{R}^n : f_{I_l}(\Pi_{I_l}(x)) = 0 \text{ for all } l = 1, \ldots, k\}.$$

The right-hand side is computed only based on the equilibrium points for the subsystems. This line of reasoning can be generalized to less trivial cases. We demonstrate this at the example of stable manifolds. Let $x^* \in \mathbb{R}^n$ be an equilibrium point and

$$\mathcal{S}(x^*) := \left\{ x \in \mathbb{R}^n : \lim_{t \to \infty} (\varphi_t(x)) = x^* \right\}$$

the stable manifold for x. By (3.6), for each $l = 1, \ldots, k$ we have

$$\varphi_t^{I_l}(\Pi_{I_l}(x)) = \Pi_{I_l}(\varphi_t(x^*)) \to \Pi_{I_l}(x^*) \text{ as } t \to \infty.$$

That shows that the projection of the stable manifold onto the subsystem is contained in the stable manifold for the subsystem. In particular, we have

$$\mathcal{S}(x^*) \subset \{x \in \mathbb{R}^n : \Pi_{I_l}(x) \in \mathcal{S}_l(\Pi_{I_l}(x^*)) \text{ for all } l = 1, \dots, k\} \tag{3.9}$$

where $\mathcal{S}_l(\Pi_{I_l}(x^*))$ denotes the stable manifold in $\Pi_{I_l}(x^*)$ for the subsystem induced by I_l. We show now that the two sets in (3.9) coincide. The argument is simple: Any point $x \in \mathbb{R}^n$ from the right-hand side of (3.9) satisfies for all $l = 1, \dots, k$, by (3.6),

$$\Pi_{I_l}(\varphi_t(x)) = \varphi_t^{I_l}(\Pi_{I_l}(x)) \to \Pi_{I_l}(x^*) \quad \text{as } t \to \infty.$$

This shows that $\varphi_t(x)$ converges in each coordinate (because of $\bigcup_{l=1}^{k} I_l = [n]$) to x^*, i.e $\varphi_t(x)$ converges to x^*. We conclude that the two sets in (3.9) are equal.

In Theorem 4.23 we show that not only equilibrium points and stable manifolds but also reachable sets (this was originally shown by [Chen 2018] for systems of a specific form), maximal invariant sets and attractors, among others, decompose in this fashion. The procedure is always as follows

1: *Find subsystems:* Identify subsystems $I_1, \dots, I_k$ with $\bigcup_{l=1}^{k} I_l = [n]$.

2: *Compute the object of interest in the subsystem:* Let M be the set of interest for the whole system and $M_1, \dots, M_k$ its analogue for each of the subsystems induced by $I_1, \dots, I_k$.

3: *"Gluing":* As candidate representation of M based on $M_1, \dots, M_k$ we "glue together" the sets $M_1, \dots, M_k$, as in (3.9), by

$$\tilde{M} := \{x \in \mathbb{R}^n : \Pi_{I_l}(x) \in M_l \text{ for all } l = 1, \dots, k\}. \tag{3.10}$$

4: *Verification:* In the last step we need to very that $\tilde{M} = M$.

In Proposition 4.10 we show that such decomposition also carries over to some important functional characteristics of dynamical systems, such as Lyapunov functions.

Remark 3.7. *The above procedure provides a decoupling scheme that can be applied to numerical algorithms for computing (decomposable) objects of interest. In Algorithm 1 we formulate this procedure and Theorem 4.25 provides corresponding convergence properties.*

Coordinate-free formulation

The concept of subsystems from Definition 3.3 is not intrinsic – it depends on the coordinates the dynamical system is written in. Consider for instance a linear dynamical system $\dot{x} = Ax$ where A has only non-zero entries but is diagonalizable. It does not allow any non-trivial subsystem but a change of coordinates, that diagonalizes A, transforming the system into a dynamical system where all states are independent.

For a coordinate-free formulation we assume that M is a (compact) smooth manifold (with boundary) of dimension n and φ a semiflow on M such that M is positively invariant.

Definition 3.8. *We call $(\mathcal{N}, P)$ a subsystem of $(M, (\varphi_t)_{t\in\mathbb{R}_+})$ if $\mathcal{N}$ is a compact smooth manifold and $P : M \to \mathcal{N}$ a smooth submersion map, i.e. the derivative of P has always full rank, such that there exists a flow $\varphi_t^{\mathcal{N}}$ on $\mathcal{N}$ such that*

$$\varphi_t^{\mathcal{N}} \circ P = P \circ \varphi_t.$$

Definition 3.8 is motivated by (3.6) and the concept is known as factor systems in the case where it is not demanded that P is a submersion. We include the submersion condition on P in order to make sure that $\mathcal{N}$ is of a lower dimension than M, see Remark 4.27. That N has indeed at most the dimension of M follows because P is a submersion.

For subsystems induced by $I_1, \dots, I_k$ according to Definition 3.3 we used the condition

$$I_1 \cup \ldots \cup I_k = [n] \tag{3.11}$$

in order to obtain full information about the system from its subsystems. The reason why condition (3.11) is helpful, is that for each $x \in \mathbb{R}^n$

$$\{y \in \mathbb{R}^n : \Pi_{I_l}(y) = \Pi_{I_l}(x), l = 1, \dots, k\} = \{x\}. \tag{3.12}$$

The statement (3.12) says that there is only one global object y that coincides with x on each subsystem – and this global object is x itself. Without any extra effort, we can generalize (3.12) to the coordinate-free definition of subsystems by replacing the maps Π_{I_l} by P_l for subsystems $(\mathcal{N}_1, P_1), \dots, (\mathcal{N}_k, P_k)$, see Definition 4.28. Corresponding decompositions for the dynamical system hold true under this concept, but the trade-off for the flexibility in the choice of P is that the treatment of the nonlinear map P is computationally more challenging because the map P can not be inferred anymore solely from the sparsity graph of the dynamics.

The sparsity graph of the dynamics

As we have seen in the case of social networks in Figure 3.2 and power grid networks in Figure 3.3, a representation through a graph is a helpful tool for identifying subsystems.

The definition of subsystems does not involve any quantitative information about the dynamics, but only causal dependence on other states in the dynam-

ics. Thus, it is sufficient to draw a graph that represents only which states are dynamically connected.

Definition 3.9 (Sparsity graph)**.** *Let $f = (f_1, \ldots, f_n) : \mathbb{R}^n \to \mathbb{R}^n$ be a function. The sparsity graph G_f associated to f is defined by:*

1. *The set of nodes is $\{x_1, \ldots, x_n\}$.*
2. *For $i \neq j$ there pair (x_i, x_j) is an edge if f_j depends explicitly on x_i .*

As an example consider dynamics $f : \mathbb{R}^n \to \mathbb{R}^n$ that are of the form

$$f = (f_1(x_1), f_2(x_1, x_2), f_3(x_1, x_3, x_4), f_4(x_1, x_4), f_5(x_1, x_4, x_5)), \tag{3.13}$$

The sparsity graph G_f of f is given in Figure 3.7.

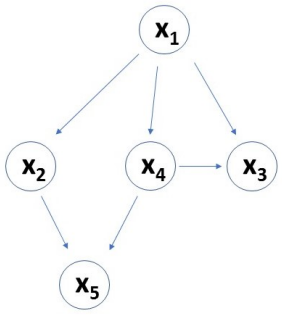

Figure 3.7: The sparsity graph of the function from (3.13)

The subsystems for a dynamical system with dynamics f given by (3.13) are shown in Figure 3.8.

From Figure 3.8 we observe that subsystems containing a state x_i have to contain all nodes x_j for which there exists a directed path from x_j to x_i. This leads to a purely graph-theoretic characterization of subsystems. This is illustrated in Proposition 3.10, which we will revisit in Chapter 4, as Proposition 4.39, where we state its proof. To simplify the formulation of Proposition 3.10, we introduce the following notion: For a sparsity graph G_f of a function $f : \mathbb{R}^n \to \mathbb{R}^n$ and an index $1 \leq i \leq n$, the *past* of i refers to all indices $1 \leq j \leq n$ for which there exists a directed path from x_j to x_i in G_f.

Proposition 3.10 (Characterization of subsystems)**.** *(I, f_I) induces a subsystem if and only if for all $i \in I$ the past of i is also contained in I.*

From this graph-theoretic characterization of subsystems we can derive several hierarchies of subsystems based on simple graph algorithms.

Let us review the example from (3.13). Consider the subsystems $I_1 := \{1, 2, 4, 5\}$ and $I_2 := \{1, 3, 4\}$. It holds $I_1 \cup I_2 = \{1, 2, 3, 4, 5\}$, as needed in the decomposition result Theorem 4.23. Further, for any other family of sets $J_1, \ldots, J_k$ that induce subsystems with $J_1 \cup \ldots \cup J_k = \{1, 2, 3, 4, 5\}$, there are sets J_{l_1} and J_{l_2} with $1 \leq$

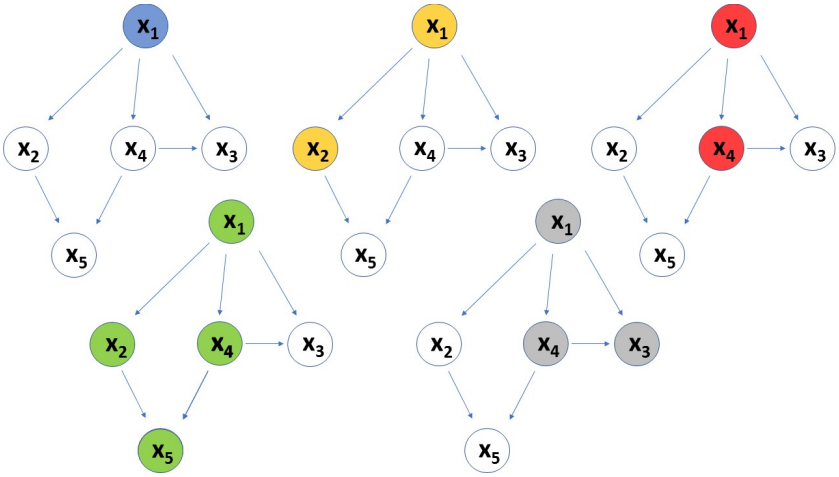

Figure 3.8: Sparsity graph G_f for f from (3.13). Nodes colored with the same color represent a subsystem. The remaining subsystems that are not presented are induced by $(\emptyset, f_\emptyset), (\{1,2,4\}, f_{\{1,2,4\}}), (\{1,2,3,4\}, f_{\{1,2,4\}})$ and $(\{1,2,3,4,5\}, f)$. Any of these can be written as a union of the subsystems colored in blue, yellow, red, green, and grey respectively.

$l_1, l_2, \leq k$ such that $I_1 \subset J_{l_1}$ and $I_2 \subset J_{l_2}$. In other words, I_1 and I_2 provide a minimal subsystem-covering of the whole system. Importantly we can characterize I_1 and I_2 as the pasts of single nodes! It holds

$$I_1 := \{j : \text{there exists a path in } G_f \text{ from } x_j \text{ to } x_5\} \tag{3.14}$$

and

$$I_2 := \{j : \text{there exists a path in } G_f \text{ from } x_j \text{ to } x_3\}. \tag{3.15}$$

The elements x_5 in (3.14) and x_3 in (3.15) are characterized by being leaves in the sparsity graph G_f, i.e. they do not have outgoing edges. Such a result is true for any sparsity graph (after a potential condensation of the sparsity graph, see Section 4.8 and Proposition 4.47).

State constraints

So far we have considered subsystems for systems defined on $\mathbb{R}^n$. An important extension is to allow for constraint sets $X \subset \mathbb{R}^n$. This can either mean we are considering the whole system only on an invariant set M or we want to enforce that the trajectories satisfy certain properties, specified by X, for all positive times in which case we restrict our attention to the set

$$M_+ := \{x_0 \in X : \text{ the trajectory starting from } x_0 \text{ stays in } X \text{ for all } t \in \mathbb{R}_+\}.$$

When a constraint set X is present, a sparse decomposition similar to the unconstrained case is possible, but needs to take the sparse structure of X into account as well. Typically, for a sparse dynamical system arising from applications, the same sparse structure of the dynamics is inherent in the constraint set X as well, due to

the sparse nature of the problem. When this is not the case, a sparse decomposition of the state constraint dynamical system is obtained by an interplay between the sparse structure of X and the dynamics, see Section 4.4. A graphical illustration via the sparsity graph $G_{f,X}$ representing the sparsity of f and X is possible This carries over to the sparsity graph

Extensions to other classes of dynamical systems

Because of its functional definition, the notion of subsystems carries over to time-dependent systems, differential inclusions, time-delay systems, hybrid systems, control systems, and stochastic ordinary differential equations. This is treated in detail in Section 4.9. To emphasize that the approach for those classes of subsystems is indeed parallel to how we treated ordinary differential equations, we consider control systems as an explanatory example. Let

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = x_0 \in \mathbb{R}^n \tag{3.16}$$

for a vector field $f = (f_1, \dots, f_n) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and controls $u = (u_1, \dots, u_m) : \mathbb{R}_+ \to \mathbb{R}^m$. The idea for a subsystem here is that we want to decouple (4.55) into smaller systems of the form

$$(\dot{x}_i)_{i\in I} = (f_i\left((x_i)_{i\in I}, (u_k)_{k\in K}\right))_{i\in I} \tag{3.17}$$

where $I \subset [n]$ and $K \subset [m]$. Whenever f_I only depends on $(x_i)_{i\in I}$ and $(u_k)_{k\in K}$ it follows that any solution $(x(\cdot), u(\cdot))$ of (3.16) induces indeed a solution $(\Pi_I(x(\cdot)), \Pi_K(u(\cdot))$ for (3.17), since

$$\frac{\mathrm{d}}{\mathrm{dt}}\Pi_I(x(t)) \quad = \quad \Pi_I\dot{x}(t) = \Pi_I(f(x(t), u(t)) = f_I\left((x_i(t))_{i\in I}, (u_k(t))_{k\in K}\right)) \tag{3.18}$$

But we need to be careful! The control u should be the same function for all subsystems! A decomposition into subsystems as in (3.18) would tempt us to use different controls u_k in each subsystem depending on the task at hand. To take this into account we have to add an extra decomposition condition on the control u for the notion of a control subsystem.

We call $I \times K \subset [n] \times [m]$ a subsystem if

1. the dynamics of $(x_i)_{i\in I}$ only depend on the sates $(x_i)_{i\in I}$ and the controls $(u_k)_{k\in K}$,
2. none of the controls $(u_k)_{k\in K}$ directly affect a state x_j for $j \notin I$.

With this notion of subsystems for control systems analogous decomposition results as for the uncontrolled case can be obtained, see Section 4.9.

Limitations

We want to point out the limitations of decompositions based on subsystems. We discuss two types of limitations. One is "practical limitations", which discusses the practical relevance of subsystems, the other is that many *but not all* objects concerning dynamical systems decompose according to subsystems.

Practical limitations A class of seemingly very sparse systems is given by cascaded systems, where information is flowing only downstream. They are of practical relevance as they appear in power flows, see [Chakravorty 2001, Mohr 2020a] and Figure 3.3, but also in water-energy cascade reservoir systems [Liu 2019], KRAS pathways in cancer analysis [Zhu 2014]. Another example is given by chemical systems where products of reactions act as reactants or enzymes for further reactions downstream, such as the Heinrich-Model and Huang-Ferrell model [Young 2017] shown in Figure 3.9.

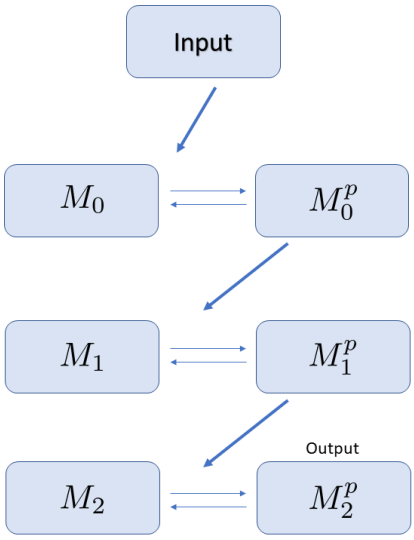

Figure 3.9: Example of a chemical cascade reaction where the product of one reaction acts as enzyme for the next reaction

It is possible to find small subsystems in cascade systems. They are even of simple nature. For instance in cascade systems without branching, such as the Heinrich-Model, the sparsity graph is just a straight path $x_1 \to x_2 \to \ldots \to x_n$ and the subsystems are all of the form $I = \{x_1, \ldots, x_k\}$ for some $1 \leq k \leq n$, as illustrated in Figure 3.10.

Unfortunately, from a purely computational perspective, the smallest subsystem containing the leaf node x_n in that graph is the whole system itself. Thus, for any family of set $I_1, \ldots, I_k \subset [n]$ that induces subsystems which has the property $I_1 \cup \ldots \cup I_k = [n]$, it already holds $I_l = [n]$ for some $1 \leq l \leq k$, i.e. we have to consider the whole system anyway! In such cases subsystems still give more insight into the dynamics and allow refined analysis but computationally we are not able

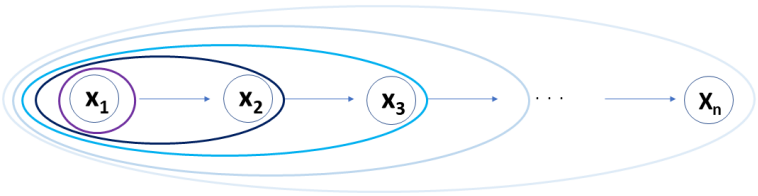

Figure 3.10: Sparsity graph of for the chemical cascade from Figure 3.9. The subsystems are all of the form $I = \{x_1, \dots, x_k\}$ for some $1 \leq k \leq n$, this is indicated by a circle around the nodes $x_1, \dots, x_k$.

to reduce the dimension. Sparse decomposition in such cases is an active research direction and different ideas have to be used, see for instance [Tacchi 2020a].

Despite the restrictiveness of our notion of subsystems we find that there is a large class of systems where sparsity in the sense of subsystems can be beneficially exploited. Among these examples are the ones mentioned at the beginning of this Section 3.1 (social networks, radial networks, etc.).

Decomposition limitations Theorem 4.23 provides a list of objects that decompose according to subsystems but there are objects for which this is not true. Among these is the weak attractor (Definition 2.4). The weak attractor is the smallest compact set that attracts each trajectory of the dynamical system. To point out why it does not decompose according to the subsystems as in (2.4) let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system and X be compact. The weak attractor $\mathcal{A}_{\mathrm{w}}$ can be characterized as the set of all accumulations points

$$\mathcal{A}_{\mathrm{w}} = \bigcup_{x\in X} \{y \in X : \text{ there exists } t_n \nearrow \infty \text{ such that } \varphi_{t_n}(x) \to y\}. \tag{3.19}$$

For a fixed $x \in X$ and subsystems induced by two index sets I and J let

$$z = \lim_{n\to\infty} (\varphi^I_{t_n}(\Pi_I(x)) \text{ and } \tilde{z} = \lim_{n\to\infty} (\varphi^J_{\tilde{t}_n}(\Pi_J(x))$$

be accumulation points of the trajectory of the subsystem starting in $\Pi_I(x)$ and $\Pi_J(x)$ respectively with corresponding sequences of times $(t_n)_{n\in\mathbb{N}}, (\tilde{t}_n)_{n\in\mathbb{N}} \subset \mathbb{R}_+$. The reason why the weak attractor does not decouple in general is that the sequences $(t_n)_{n\in\mathbb{N}}$ and $(\tilde{t}_n)_{n\in\mathbb{N}}$ possibly differ. This is the idea of the Example 4.34 where we state an explicit system for which the weak attractor does not decompose according to its subsystem. The principle that hampers a decomposition is that there is a "choice" (of subsequence $(t_n)_{n\in\mathbb{N}}$) inherent in the object of interest (here the weak attractor $\mathcal{A}_{\mathrm{w}}$). This possibly allows having incompatible choices and incompatible corresponding objects z and $\tilde{z}$ for different subsystems.

Comparison with other decomposition techniques for dynamical systems

The increasing interest in large-scale systems has driven research on reducing dynamical systems to lower dimensional ones. Many different concepts have emerged. In this section, we want to compare our approach to some of these methods. We base our comparison on the following five classifications

1. *Exactness:* This concerns the question of whether an error is induced by the decomposition procedure.
2. *Additional (stability, regularity, etc.) assumptions:* That comparison point is based on whether the proposed methods require additional assumptions on the dynamical systems.
3. *Generality:* Here we want to distinguish decomposition techniques that are designed for a specific task from decomposition techniques that can be applied to a variety of problems concerning the dynamical system.
4. *Different notion of sparsity:* Clearly the underlying notion of sparsity determines the conditions and objectives of the decomposition. Thus we will differentiate methods based on the underlying concept of sparsity.
5. *Performance:* This point addresses the computational advantages of the different methods and our measure for reduction is the size of the largest resulting subsystem.

We will not much discuss the last point, performance, because general statements of how much smaller the considered systems are, are rarely given and such a statement would heavily depend on the assumed underlying structure. Nevertheless, we want to mention a rule of thumb: Allowing a perturbation or assuming additional conditions on the dynamical system allow more reduction, as can be expected. Therefore, our sparse decomposition procedure may give rather conservative reductions compared to some approaches that allow perturbations; see for instance [Anderson 2012a].

What characterizes our method is that it provides exact decompositions into subsystems according to Theorem 4.23 and we require only minimal assumptions on f (locally Lipschitz) and the constraint set (compact). Thus, our method applies to a broad class of problems from dynamical systems (Theorem 4.23) and it transfers to different classes of dynamical systems (Section 3.1 and 4.9).To find subsystems we use the sparsity graph of the dynamics. This allows for simple and quick methods for identifying subsystems. However, our notion of sparsity is restrictive in the sense that for some arguably sparse systems such as cascade systems without branching, illustrated in Figures 3.9 and 3.10, only a refined qualitative analysis is possible but computational savings are small.

In [Anderson 2011a, Anderson 2011b, Al Maruf 2018] reduction techniques are presented that search for lower dimensional systems, from which approximate solutions of the original system can be constructed or in which stability of the system is aimed to be inferred from the stability of the lower dimensional systems.

The decomposition of the dynamical system into smaller components is motivated by partitioning the states such that the interaction between the different components is minimized. This differs from our approach and can be interpreted as removing (certain, well-chosen) edges from the sparsity graph. Methods as in [Anderson 2011a, Anderson 2011b, Al Maruf 2018] potentially allow to consider a dynamical system only on the sates $(x_i)_{i\in I}$ for *any* subset $I \subset [n]$ (whether it induces a subsystem or not). This difference in such techniques compared to ours becomes evident, for instance, in the Heinrich model illustrated in Figures 3.9 and 3.10 where our approach does not lead to a reduction in dimension, but a splitting into the states $x_1, \dots, x_k$ and $x_{k+1}, \dots, x_n$ is feasible within the methods from [Anderson 2011a, Anderson 2011b, Al Maruf 2018]. The trade-off for partitioning into smaller systems is that an error is induced when trajectories of the whole system are approximated by the perturbed partitioned systems. A treatment for sparse exploitation specifically for systems with sparsity graphs as in Figure 3.10, was investigated for the region of attraction problem in [Tacchi 2020a].

The analysis of decomposing a system into smaller ones can be refined using additional assumptions on the system such as dissipation inequalities and comparison systems can be found in [Anderson 2011b, Anderson 2010, Anderson 2012b, Al Maruf 2018, Dashkovskiy 2011]. Under these additional assumptions, stability can be verified based on potentially smaller decompositions than obtained with our approach, see [Anderson 2011b, Dashkovskiy 2011, Al Maruf 2018].

In [Anderson 2012a, Chapter 4 and 6] and [Al Maruf 2018], reduction techniques for certifying the asymptotic stability of the origin are presented. They are based on comparison systems or dissipation inequalities utilizing additional assumptions. From the perspective of our approach, via the sparsity graph, that allows including and taking into account weights on edges in the sparsity graph and uses concepts particularly related to the task at hand – in the case of Chapters 4 and 6 in [Anderson 2012a] and [Al Maruf 2018], the authors do so with focus on stability.

Another field concerning a representation of the system by a lower dimensional one is model reduction. By nature, this approach aims at a low dimensional approximation of the system by one single other system and therefore carries a more quantitative character than our approach. In contrast to such quantitative perspectives are topological decompositions of dynamical systems, see for instance [Mischaikow 2002]. There, the system is decomposed according to its attracting, repelling, or (quasi)periodic parts. Sparsity in this context refers to simple structures of those objects. Computationally, these structures are typically not easily accessible just from the vector field and cannot be inferred solely from the sparsity graph.

Another example of a reduction technique that relates to a different notion of sparsity is [Elkin 2012, Elkin 2008]. In contrast to our notion of subsystems, which describes independence of states, the notion of subsystems in [Elkin 2012, Elkin 2008] treats explicit functional dependence between different states and relates to symmetries in the system.

3.2 Linear programming problem formulations for attractors

Main contribution: We state two infinite dimensional linear programming problems that give rise to tight outer approximations of the global attractor, see Theorems 3.12 and 3.16. We show that, asymptotically, these LPs can be solved via the moment-SOS-hierarchy, Theorem 3.20.

The definitions, notations, concepts and results concerning dynamical, which are important systems for this section, are provided in the preliminary Sections 2.1 and 2.4. A detailed presentation of the results stated in this section is provided in Chapter 5 which presents the results from the texts [Schlosser 2021, Schlosser 2022a, Schlosser 2020, Wang 2021b].

We present two computational methods for approximating the global attractor of a dynamical system based on the works [Schlosser 2021, Schlosser 2022a]. Both methods follow the idea of translating the problem into an infinite dimensional linear setting and give rise to approximations of the global attractors via hierarchies of convex optimization problems. The first method, [Schlosser 2021], achieves a linear formulation through the use of so-called occupation measures. The second approach, [Schlosser 2022a], is motivated by classical Lyapunov theory and strongly builds up on [Jones 2021a].

We consider dynamical systems on $\mathbb{R}^n$, induced by the following ordinary differential equation

$$\dot{x} = f(x),\ x(0) = x_0 \in \mathbb{R}^n \tag{3.20}$$

for Lipschitz continuous vector fields f and equip them with a constraint set $X \subset \mathbb{R}^n$. The constraint set X focuses our interest on the maximum positively invariant set M_+, i.e. all the points $x_0 \in X$ whose solution of (3.20) stays in X for all positive times. We define the global attractor with respect to the dynamical system $(M_+, \varphi_t\big|_{M_+})$.

Definition 3.11 (Global attractor)**.** *For a given compact set $X \subset \mathbb{R}^n$ we call a compact set $\mathcal{A} \subset X$ the global attractor if it is the minimal compact set that uniformly attracts all solutions of (3.20) that stay in X for all positive times.*

Attractors play a fundamental role in the analysis of longtime behavior of dynamical systems. In control theory they show their importance when a certain feedback control law is implemented to stabilize the origin – the attractor $\mathcal{A}$ provides a certificate, via $\mathcal{A} = \{0\}$ or $\mathcal{A} \neq \{0\}$, of correctness of that control law. Another example related to this thesis arises from optimization problems

$$\min_{x \in X} \phi(x) \tag{3.21}$$

where $X \subset \mathbb{R}^n$ is a given set. First-order optimization methods often can be interpreted as discretizations of the gradient flow

$$\dot{x} = -\nabla\phi(x). \tag{3.22}$$

Convergence of such first-order methods then naturally relates to attractors. These are just two examples of a large variety of applications of attractors. Therefore, many methods for computing the attractor have been created. The classical approach is via Lyapunov functions, and the second method that we propose falls into that category. Even in that category, there are many different approaches toward computations of Lyapunov functions, see [Giesl 2015] for a survey. Among the methods closest related to ours are [Prajna 2004] and [Parrilo 2000, Chapter 7] where polynomial Lyapunov functions are computed, with the difference that [Prajna 2004] and [Parrilo 2000, Chapter 7] aim at verifying the stability of a given set while we want to localize the attractor. A similar, but dual perspective is investigated in [Rantzer 2001] and relates to the occupation measures approach we will present first. Other existing approaches to approximating global attractors are, for example, finite time truncations, spatial and temporal discretization or set-oriented methods [Dellnitz 2002, Dellnitz 2001].

Computing the attractor via occupation measures

In this section, we present the method proposed in [Schlosser 2021]. We follow an established line of reasoning [Rubio 1975, Vinter 1978, Korda 2014, Jones 2021a]. This approach has two central ingredients: First, point 2 in Theorem 2.7, which characterizes the global attractor as the largest invariant set in X. And secondly, the method from [Korda 2014] for computing the maximum positively invariant set is founded on an infinite dimensional linear programming problem on the space of measures. The result from [Schlosser 2021] is obtained by merging the LPs from [Korda 2014] for maximal invariant sets in forward and backward time direction into one LP. As in [Korda 2014], applying the machinery of the moment-sum-of-squares hierarchy leads to a hierarchy of finite dimensional semidefinite programs whose solutions converge to the solution to the infinite dimensional LP.

Historically, the idea of transforming various problems from nonlinear dynamical systems and control into infinite-dimensional LPs dates back at least to the work [Rubio 1975, Vinter 1978] dedicated to optimal control. Solving these problems by a hierarchy of semidefinite problems (SDPs) with proven convergence was proposed in [Lasserre 2008], although SDP approximations to infinite-dimensional LPs were used already in [Prajna 2004] for the problem of global stabilization. Since then, this approach was used to tackle a number of problems, including the region of attraction [Henrion 2013], maximum (control) invariant and reachable sets [Korda 2014, Magron 2019b] or, more recently, analysis and control of nonlinear partial differential equations [Marx 2018, Korda 2022, Goluskin 2019] or safety analysis [Miller 2021], to name just a few. Closest to our work from this line of research is [Goluskin 2018, Goluskin 2020], treating the problem of estimating the maximum of a given function on the attractor.

In [Schlosser 2021] we proposed the following LP for computing the volume of the global attractor

$$\begin{aligned} p^* := \quad & \sup_{\mu_0,\hat{\mu}_0,\mu_+,\mu_-} && \mu_0(X) \\ & \text{s.t.} && \mu_0, \hat{\mu}_0, \mu_+, \mu_- \in M(X)_+ \\ & && \int_X v^1 - \nabla v^1 \cdot f \, d\mu_+ = \int_X v^1 \, d\mu_0 \quad \forall v^1 \in \mathcal{C}^1(\mathbb{R}^n) \\ & && \int_X v^2 + \nabla v^2 \cdot f \, d\mu_- = \int_X v^2 \, d\mu_0 \quad \forall v^2 \in \mathcal{C}^1(\mathbb{R}^n) \\ & && \mu_0 + \hat{\mu}_0 \quad = \lambda|_X \end{aligned} \tag{3.23}$$

where $M(X)_+$ denotes the set of non-negative Borel measures on X and $\lambda|_X$ is the restriction of the Lebesgue measure to X. The primal problem (3.23) can be naturally motivated via occupation measures based on [Korda 2014], see Section 5.1. Unfortunately, it contains less information about the attractor itself, as is drastically demonstrated in the case when $\lambda(\mathcal{A}) = 0$. Then the trivial solution $(\mu_0, \hat{\mu}_0, \mu_+, \mu_-) = (0, 0, 0, 0)$ is optimal and we cannot infer more than vanishing Lebesgue volume from this minimizer. On the contrary, we will show that the dual problem gives more insight into the form of the attractor. The dual problem is given by

$$\begin{aligned} d^* := \quad & \inf_{w,v^1,v^2} && \int_X w(x) \, dx \\ & \text{s.t.} && (w, v^1, v^2) \in \mathcal{C}(\mathbb{R}^n) \times \mathcal{C}^1(\mathbb{R}^n) \times \mathcal{C}^1(\mathbb{R}^n) \\ & && w \geq v^1 + v^2 + \mathbf{1} \\ & && w \geq 0 \\ & && v^1 - \nabla v^1 \cdot f \geq 0 \\ & && v^2 + \nabla v^2 \cdot f \geq 0. \end{aligned} \tag{3.24}$$

In [Schlosser 2021] we showed the following

Theorem 3.12. *Let X be compact. Then*

$$p^* = d^* = \lambda(\mathcal{A})$$

where $\lambda(\mathcal{A})$ denotes the Lebesgue volume of the attractor $\mathcal{A}$. Further, for each feasible point (w, v^1, v^2) for the dual LP (3.24) it holds

$$\mathcal{A} \subset w^{-1}([1, \infty)). \tag{3.25}$$

We present Theorem 3.12 and it's proof in Chapter 5 in Theorem 5.4.

A closer look at the LP (3.24) and the set in (3.25) reveals, now only using non-negativity of w, the following bound

$$\lambda(w^{-1}([1, \infty)) \setminus \mathcal{A})) = \lambda(w^{-1}([1, \infty)) - \lambda(\mathcal{A}) \leq \int_X w(x) \, dx - \lambda(\mathcal{A}). \tag{3.26}$$

In particular, we see that the set $w^{-1}([1, \infty))$ gives a convergent outer approximation of the global attractor when (w, v^1, v^2) gets optimal. We formulate this in the

following proposition.

Proposition 3.13. *With the notation from Theorem 3.12, let (w_k, v_k^1, v_k^2) be a minimizing sequence for the LP (3.24) then the sets $A_k := w_k^{-1}([1,\infty))$ are outer approximations of the global attractor $\mathcal{A}$ and it holds*

$$\lambda(A_k \setminus \mathcal{A}) \to 0 \text{ as } k \to \infty.$$

The LPs (3.23) and (3.24) can be interpreted in the following way: For the primal problem (3.23) we already know an optimal solution, by construction, namely $\mu_0 = \lambda|_{\mathcal{A}}$ and $\hat{\mu}_0 := \lambda|_{X\setminus\mathcal{A}}$ and μ_+, μ_- the corresponding occupation measure for the flow in forward respectively backward in time, see (5.2) for the definition of occupation measures. For the dual problem the optimal solution, if feasible, would be of the form

$$w = \chi_{\mathcal{A}}, v^1 = \chi_{\mathbb{R}^n\setminus M_+} \text{ and } v^2 = \chi_{\mathbb{R}^n\setminus M_-} \tag{3.27}$$

where χ denotes the indicator function and M_+ respectively M_- the maximum positively invariant set for X respectively the maximum positively invariant set for X in backward time direction. The functions from (3.27) are typically not continuous and therefore not feasible but from $d^* = \lambda(\mathcal{A})$ we conclude that for any minimizing sequence $(w_l, v_l^1, v_l^2)_{l\in\mathbb{N}}$ for (3.24) it holds $w_l \to \chi_{\mathcal{A}}$ pointwise, $v^1 \geq 0$ on $\mathbb{R}^n \setminus M_+$ and $v^2 \geq 0$ on $\mathbb{R}^n \setminus M_-$, see Theorem 5.4 and Lemma 5.3. This insight on minimizing sequences unveils expected numerical behavior such as oscillations around the boundary of the attractor and motivates research to reduce such limitations, as done in [Tacchi 2020b] for volume computation of semialgebraic sets based on infinite dimensional linear programming.

Remark 3.14. *We only guarantee convergent approximations of the attractor with respect to Lebesgue measure discrepancy. Therefore, we do not have control about topological properties of the approximations.*

Computing the attractor via almost Lyapunov functions

The way we approximate global attractors in [Schlosser 2022a] is based on the intimate relation between attractors and Lyapunov functions, which is vividly illustrated in Theorem 2.11.

We call a function $V : U \to \mathbb{R}$ for an open set $U \subset \mathbb{R}^n$ a Lyapunov function for the system, induced by $\dot{x} = f(x)$ for $x(0) \in U$ if $V \in \mathcal{C}^1(U)$ and

$$V \geq 0 \text{ and } \nabla V \cdot f \leq -V \tag{3.28}$$

Theorem 2.11 states the global attractor can be characterized as the smallest set $\mathcal{A}$ for which there exists a Lyapunov function V with $\mathcal{A} = V^{-1}(\{0\})$. Thus, we can transfer the search for attractors to the search for Lyapunov functions. Unfortunately, a polynomial approach is limiting when it comes to finding Lyapunov functions. The reason is that there are polynomial systems where no polynomial Lyapunov function V exists that satisfies $V^{-1}(\{0\}) = \mathcal{A}$ for the attractor [Ahmadi 2018]. In [Jones 2021a] this problem is overcome by relaxing the notion of a Lyapunov function while maintaining desirable properties for computing

the attractor. Our extension of [Jones 2021a] addresses the missing convergence guarantee of the proposed computational method. We do so by incorporating ideas from moment formulations, as in (3.24), resulting in a moment-SOS hierarchy for approximating global attractors based on the concept of Lyapunov functions from [Jones 2021a].

We briefly outline the approach presented in [Jones 2021a] and [Schlosser 2022a]. Let us assume that $X \subset \mathbb{R}^n$ is open, bounded, and positively invariant for (3.20). We begin with the observation that, by Theorem 2.11, we can search for the attractor through sublevel sets of Lyapunov functions, therefore it holds

$$\begin{aligned} \mathcal{A} = \quad & \inf_{V} \quad V^{-1}(\{0\}) \\ & \text{s.t.} \quad V \in \mathcal{C}^1(X) \\ & \qquad V \geq 0 \\ & \qquad \nabla V \cdot f \leq -V \end{aligned}$$

where $\mathcal{A}$ is the global attractor for the dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ induced by (3.20). As a first step, we modify the cost function, because it is set-valued and therefore unhandy to optimize. From the monotony of the Lebesgue measure λ we conclude

$$\begin{aligned} \lambda(\mathcal{A}) = \quad & \inf_{V} \quad \lambda(V^{-1}(\{0\})) \\ & \text{s.t.} \quad V \in \mathcal{C}^1(X) \\ & \qquad V \geq 0 \\ & \qquad \nabla v \cdot f \leq -V. \end{aligned} \tag{3.29}$$

As a next step we would like to replace the search space $\mathcal{C}^1(X)$ by $\mathbb{R}[x]$, but as mentioned before the class of polynomials is too restrictive for Lyapunov functions. For this reason, in [Jones 2021a] the notion of Lyapunov function was relaxed to functions J satisfying

$$J \geq 0 \quad \text{and} \quad \nabla J \cdot f \leq -J + \varepsilon \tag{3.30}$$

for $\varepsilon > 0$. We call a function that satisfys (3.30) an almost Lyapunov function[1].

Searching for functions that satisfy (3.30) is helpful for two reasons. First, (3.30) is a relaxed Lyapunov condition since any Lyapunov function solves (3.30) for $\varepsilon = 0$, and hence for all $\varepsilon > 0$. Secondly, by the Weierstraß approximation theorem, there always exist polynomials satisfying (3.30), see [Jones 2021a]. And third, (3.30) preserves the following two desirable properties in Lemma 3.15, that we are used to from Lyapunov functions [Jones 2021a].

Lemma 3.15. *Let $J \in \mathcal{C}^1(X)$ satisfy (3.30) then $J^{-1}([0, \varepsilon])$ contains the attractor and is positively invariant.*

Proof. The argument is based directly on Lyapunov analysis, namely, for any function J satisfying (3.30) we choose a smooth function $\rho : \mathbb{R} \to \mathbb{R}$ with $\rho = 0$ on $(-\infty, \varepsilon]$ and ρ is strictly increasing on (ε, ∞). Then $V := \rho \circ J$ is a Lyapunov function in the classical sense and thus $J^{-1}([0, \varepsilon]) = V^{-1}(\{0\})$ contains the global attractor and is positively invariant. □

[1] This is not the same notion of almost Lyapunov functions differs as in [Liu 2020].

And finally, in [Jones 2021a] the authors showed that the global attractor can be approximated arbitrarily well by sets $J^{-1}([0,\varepsilon])$ for *polynomials* $J \in \mathbb{R}[x]$ satisfying (3.30). The argument for this result is very instructive for the treatment of the relaxed constraint (3.30). Namely, for $\frac{1}{2} > \varepsilon > 0$ choose a Lyapunov function $V \in \mathcal{C}^1(X)$ with $V(x) \geq 1$ for all $x \in X$ with $\operatorname{dist}(x, \mathcal{A}) > \varepsilon$ (that such Lyapunov functions exist is an easy consequence of Theorem 2.11 and the compactness of X). Approximating the function $V + \frac{\varepsilon}{2}$ closely enough (in the $\mathcal{C}^1(X)$ metric) by a polynomial J results in J satisfying (3.30) and

$$\mathcal{A} \subset J^{-1}([0,\varepsilon]) \subset \{x \in X : \operatorname{dist}(x, \mathcal{A}) < \varepsilon\}.$$

Following up on this results in the following optimization problem[Jones 2021a] for computing the global attractor and obtaining outer approximations

$$\begin{array}{rlll} \lambda(\mathcal{A}) = & \inf\limits_{J} & \lambda(J^{-1}(\{0\})) & \\ & \text{s.t.} & J \in \mathbb{R}[x] & \\ & & J \geq 0 & \text{on } X \\ & & \nabla J \cdot f \leq -J + 1 & \text{on } X. \end{array} \tag{3.31}$$

Unfortunately, the moment-SOS hierarchy is not applicable yet because the cost term $\lambda(J^{-1}(\{0\}))$ is not linear in J, and not even convex. This makes the (3.31) difficult to optimize. In [Jones 2021a] a heuristic was used to relate the coefficients of J and the volume $\lambda(J^{-1}(\{0\}))$. The use of the heuristic prevents convergence guarantees, which we overcame in [Schlosser 2022a] by linearizing the optimization problem (3.31) in the fashion of the previous paragraph. We introduce the new decisions variable w that shall play the role of a smooth approximation of $w^* := \chi_K$ for $K := J^{-1}([0,\varepsilon])$. Adding the constraints

$$w \geq 0 \text{ and } w + J \geq \mathbf{1} \text{ on } X \tag{3.32}$$

does the job for $\varepsilon = 0$, i.e.

$$\begin{array}{rlll} \lambda(\mathcal{A}) = & \inf\limits_{w,J} & \int\limits_X w(x)\, dx & \\ & \text{s.t.} & w \in \mathcal{C}(X), J \in \mathcal{C}^1(X) & \\ & & w \geq 0 & \text{on } X \\ & & w + J \geq \mathbf{1} & \text{on } X \\ & & J \geq 0 & \text{on } X \\ & & \nabla J \cdot f + J \leq 0 & \text{on } X. \end{array} \tag{3.33}$$

But the feasible functions J for the LP (3.33) are exactly the Lyapunov functions for the system. To verify that the optimal value in the LP (3.33) is indeed $\lambda(\mathcal{A})$ let V be a Lyapunov function with $V^{-1}(\{0\}) = \mathcal{A}$. For $m \in \mathbb{N}$, the pair (w_m, V_m) given by

$$w_m := \max\{0, 1 - V_m\} \quad \text{and} \quad V_m := m \cdot V \tag{3.34}$$

is feasible for (3.33) and $w_m \searrow \chi_{\mathcal{A}}$ as $m \to \infty$. By the monotone convergence

theorem, it follows

$$\int_X w_m(x)\,dx \to \int_X \chi_{\mathcal{A}}(x)\,dx = \lambda(\mathcal{A}).$$

Thus, when allowing $\varepsilon > 0$, we relax the problem and should add a penalty to ε. It turns out that the volume of X is an exact penalty and we can propose the following infinite dimensional linear programming problem

$$\begin{aligned} s^* = \inf_{w,J,\varepsilon} \quad & \int_X w(x)\,dx + \varepsilon\lambda(X) \\ \text{s.t.} \quad & w \in \mathcal{C}(X), J \in \mathcal{C}^1(X), \varepsilon \geq 0 \\ & w \geq 0 && \text{on } X \\ & w + J \geq \mathbf{1} && \text{on } X \\ & J \geq 0 && \text{on } X \\ & \nabla J \cdot f + J \leq \varepsilon && \text{on } X. \end{aligned} \tag{3.35}$$

Theorem 3.16. *Let X be compact and positively invariant, $\mathcal{A}$ be the global attractor for X and assume that the basin of attraction $B_f(\mathcal{A})$ is open. Then it holds for s^* from (3.35)*

$$s^* = \lambda(\mathcal{A}).$$

Furthermore, for feasible (w, J, ε) for (3.35), the set $J^{-1}([0, \varepsilon])$ provides an outer approximation of the GA that gets tight when (w, J, ε) gets optimal.

The proof of Theorem 3.16 is based on the LP (3.33) where we have showed already that the optimal value is given by $\lambda(\mathcal{A})$. It only remains to show that adding the slack variable ε does not lead to a relaxation. We show this rigorously in Theorem 5.11 in Chapter 5.

Remark 3.17. *The condition that X is positively invariant in Theorem 3.16 can be removed, see Theorem 5.11. This is done by adding an additional decision variable v with the constraint $v - \nabla v \cdot f \geq 0$, as we have used in the LP (3.24).*

We want to remind why we added the ε slack variable, even though we already computed the volume of the global attractor in the LP (3.29).

Remark 3.18. *In contrast to the LP (3.29), for the LP (3.35) there always exists a minimizing sequence consisting $(w_k, J_k, \varepsilon_k)_{k\in\mathbb{N}}$ of where w_k and J_k can be chosen polynomial for each $k \in \mathbb{N}$ such that the inequality constraints in (3.35) are strictly satisfied.*

Now that we have stated the infinite dimensional linear programming problems (3.24) and (3.35), we want to solve them numerically. We approached this task using the moment-SOS hierarchy because of several preferable properties. Those properties as well as how the moment-SOS hierarchy is applied is discussed in the following paragraph.

Solving the infinite dimensional LP

We solved the LPs (3.24) and (3.35) via the established moment-SOS hierarchy [Lasserre 2009, Lasserre 2015] just as in [Korda 2014, Henrion 2013] for similar

problems. Therefore, we need to make the following assumption on the polynomial structure of the problem.

Assumption 3.19. *The vector field f is polynomial and $X = \mathcal{K}(p_1, \dots, p_j)$ is a compact basic semi-algebraic set, that is, there exist polynomials $p_1, \dots, p_j \in \mathbb{R}[x_1, \dots, x_n]$ such that $X = \{x \in \mathbb{R}^n : p_i(x) \geq 0$ for $i = 1, \dots, j\}$. Further we assume that one of the p_i is given by $p_i(x) = R_X^2 - \|x\|_2^2$ for some large enough $R_X \in \mathbb{R}$.*

Conceptually the approach is as elegant as simple: First, we replace the search space $\mathcal{C}(X)$ respectively $\mathcal{C}^1(\mathbb{R}^n)$ by the space of polynomials $\mathbb{R}[x]$. For the dual problems (3.24) and (3.35) we tighten the non-negativity constraints to an SOS constraint. Leveraging the Weierstraß approximation theorem and Putinar's Positivstellensatz this tightening maintains the optimal value of the problem, see Theorem 5.7 and Section 5.2. In other words, we can approximate the attractor via polynomials and SOS constraints. As in the discussion around the Lasserre Hierarchy in Section 2.4 we obtain a hierarchy of semidefinite programs by substituting the infinite dimensional space $\mathbb{R}[x]$ by $\mathbb{R}[x]_k$, the space of polynomials of degree at most k, for $k \in \mathbb{N}$. For the LP (3.24) this translates to the following hierarchy of sums-of-squares problem

$$\begin{aligned} d_k := \quad & \inf_{w, v^1, v^2, \{q_i\}, \{t_i\}, \{r_i\}, \{s_i\}} \quad \mathbf{w}'\mathbf{l} \\ \text{s.t.} \quad & -v^1 - v^2 + w - 1 = q_0 + \sum_{i=1}^{j} q_i p_i \\ & w = t_0 + \sum_{i=1}^{j} t_i p_i \\ & \beta v^1 - \nabla v^1 \cdot f = r_0 + \sum_{i=1}^{j} r_i p_i \\ & \beta v^2 + \nabla v^2 \cdot f = s_0 + \sum_{i=1}^{j} s_i p_i \end{aligned} \tag{3.36}$$

where $\mathbf{w}'$ is the vector of coefficients of the polynomial w and $\mathbf{l}$ is the vector of the moments of the Lebesgue measure over X (i.e., $\mathbf{l}_\alpha = \int_X x^\alpha \, d\lambda(x)$, $\alpha \in \mathbb{N}^n$, $\sum_i \alpha_i \leq k$), both indexed in the same basis of $\mathbb{R}[x]_k$; hence $\mathbf{w}'\mathbf{l} = \int_X w(x) \, d\lambda(x)$. The decision variables v^1, v^2, w are polynomials in $\mathbb{R}[x]_k$ and all the polynomials $q_0, \dots, q_j$, $r_0, \dots, r_j$, $s_0, \dots, s_j$, $t_0, \dots, t_j$ are SOS polynomials such that q_0, t_0, r_0, s_0, $q_i p_i$, $t_i p_i$, $r_i p_i$, $s_i p_i$ are all in $\mathbb{R}[x]_k$ for all $i = 1, \dots, j$.

In the following theorem, we state that this leads to a convergent hierarchy of SDPs which preserves the property of the LP (3.24) that the attractor can be approximated from feasible points.

Theorem 3.20. *Under Assumption 3.19 it holds*

$$d_k \geq d_{k+1} \text{ for all } k \in \mathbb{N} \text{ and } d_k \searrow \lambda(\mathcal{A}) \text{ as } k \to \infty.$$

Further, for $k \in \mathbb{N}$ and optimal points (w_k, v_k^1, v_k^2) for the SDP (3.36) (with corre-

sponding multipliers) it holds $w_k^{-1}([1,\infty)) \supset \mathcal{A}$ *and*

$$\lambda(w_k^{-1}([1,\infty)) \setminus \mathcal{A}) \leq d_k - \lambda(\mathcal{A}) \to 0 \text{ as } k \to \infty.$$

We present a formal proof of Theorem 3.20 in Chapter 5, see Section 5.1.3.

Remark 3.21. *Analogous to the SDPs (3.36) we can formulate a hierarchy of SDPs for the LP (3.35). Theorem (3.20) holds in a similar way, under the additional assumption needed in Theorem 3.16. The essential argument is that there exists a minimizing sequence for the LP ((3.35) which consists of polynomials that satisfy the inequality constraints in LP ((3.35) with strict inequality, see Remark 3.18 and Chapter 5 Section 5.2. This allows applying Putinar's Positivstellensatz and convergence of the SDP hierarchy follows.*

We illustrate this method based on the moment-SOS hierarchy in the following example of the Lorenz system on $X = [-10, 10]^3$.

$$\dot{x} = 10(y - x), \quad \dot{y} = x(28 - z) - y, \quad \dot{z} = xy - \frac{8}{3}z$$

Figure 3.11 shows the outer approximation (drawn in light red) of the global attractor (colored in black) for the Lorenz system.

Figure 3.11: Outer approximations for the Lorenz attractor obtained by degree 8 polynomials, projected from two angles. The time to solve the corresponding SDP was 0.67s with MOSEK 8.1 running on a machine with 4,2 GHz Intel Core i7 and 32 GB 2400 MHz DDR4 RAM.

Comparison between the different LPs

We want to briefly discuss the differences between the three closely related methods from [Schlosser 2021], [Jones 2021a] and [Schlosser 2022a].

As mentioned before, the work in [Schlosser 2022a] builds on [Jones 2021a] and addresses the missing convergence guarantee of the proposed computational method in [Jones 2021a]. In [Jones 2021a] an optimization problem is stated whose solution gives the volume of the GA and provides convergent outer approximations but the cost term in that optimization problem is difficult to treat and non-convex. For computations, in [Jones 2021a] the authors replaced that cost function by a convex cost function based on a heuristic, at the price of guaranteed convergence. Thus, our approach via the LP (3.35) should be interpreted as a practical extension of the computational method from [Jones 2021a].

Compared to the approach via occupation measures, the approach via almost Lyapunov functions has the qualitative advantage that it produces approximations of the attractor that are positively invariant. In our numerical examples, we observed that this qualitative improvement came at the cost of performance, i.e. the approximations obtained via the same level in the SOS hierarchy for the LP (3.24) were closer to the attractor than the ones obtained from the LP (3.35).

In terms of computational complexity the three methods, i.e. the SOS hierarchy for (3.36), the one for the LP (3.35) and the SOS hierarchy proposed in [Jones 2021a], are of the same order. This is because all three methods are based on the moment-SOS hierarchy and only differ in the number of decision variables in the corresponding LPs. Hence, their computational complexity is of the same order in each level of the hierarchy.

We summarize the main differences between the methods from [Jones 2021a] (with and without heuristic), [Schlosser 2021], and [Schlosser 2022a] in the following table.

	[Schlosser 2021]	[Jones 2021a]	[Jones 2021a] + heuristic	[Schlosser 2022a]
Convex	✓		✓	✓
Invariant sets		✓	✓	✓
Convergence	✓	✓		✓

The second line, convex problem, refers to the optimization problem being convex, the third line to the property that the obtained sets are positively invariant, and the fourth line to the guaranteed convergence of these sets towards the GA.

Computational complexity and exploiting sparsity and symmetry

As a trade-off for the beneficial properties of the moment-SOS hierarchy comes its computational complexity. We mentioned in Remark 2.39 that the computational cost for computing ε-optimal solutions of SDPs respectively sums-of-squares programs is polynomial in the input size. But as for the Lasserre-hierarchy for polynomial optimization described the size of the appearing SDPs grows combinatorially in the space dimension n and the level k of the hierarchy. More specifically, for the SOS program (3.36) the blocks in the corresponding SDP are of size

$$\mathcal{O}\left(\binom{n+\left\lfloor \frac{k}{2} \right\rfloor}{n}\right).$$

This poses practical limitations already for dynamical systems in a few variables. For the attractor of the Lorenz system, i.e. $n = 3$, we were able to compute the corresponding SDPs up to degree $k = 10$ polynomials on a usual laptop. For larger systems, the SDPs easily render intractable for current computing devices. Therefore, exploiting problem-inherent structure is necessary in order to make larger-sized problems computationally accessible via the moment-SOS hierarchy. The most popular concepts for doing so in polynomial optimization are symmetry [Riener 2013] and sparsity [Lasserre 2015, Chapter 8] and [Wang 2021a]. We cannot directly apply those techniques to the moment-SOS hierarchies for the optimization problems (3.24) and (3.35). The reason is that, different to static polynomial optimization, the polynomial for which we search a SOS for is a decision variable itself. Therefore, we need to show that we can inherit additional structure from the problem to the decision variables in (3.24) and (3.35). In our case that would follow the subsequent guiding principle

> *If the dynamical system is symmetric/sparse then there exists a minimizing sequence for (3.24) and (3.35) consisting of symmetric/sparse polynomials.*

This allows us to further reduce the search space to symmetric respectively sparse polynomials and therefore reduce the size of the appearing SDPs. Symmetry was exploited in such a fashion in [Fantuzzi 2020] for bounding extreme events for dynamical systems. In accordance to the decomposition procedure Algorithm 2, we showed in [Schlosser 2020] that the sparsity concept from Section 3.1 respectively Chapter 4 can be successfully applied to the SOS programs (3.36) and corresponding programs for computing the maximum positively invariant set and the reachable set. In [Wang 2021b] we investigated to pair the works [Korda 2014, Henrion 2013, Schlosser 2021] with a term sparsity concept which proved successful for static polynomial optimization [Wang 2021a].

Extension to computing asymptotic extreme values

The LP representations (3.23) and (3.24) of the attractor allow for further applications. One is to extend the computation of the attractor to computing extreme events on the attractor as initiated in [Goluskin 2018, Goluskin 2020]. The LP, for bounding a function $\phi : \mathbb{R}^n \to \mathbb{R}$ on the global attractor, from [Goluskin 2018] reads:

$$\begin{array}{lll} p^* = \inf & c & \\ \text{s.t.} & \beta > 0, c \in \mathbb{R}, V \in \mathcal{C}^1(\mathbb{R}^n) & \\ & V - \phi \geq 0 & \text{on } X \\ & c - V - \beta \nabla V \cdot f \geq 0 & \text{on } X \end{array} \tag{3.37}$$

It was shown in [Goluskin 2018] that

$$p^* \geq \max_{x \in \mathcal{A}} \phi(x). \tag{3.38}$$

If equality holds in (3.38) was an open question in the case when X is not already positively invariant. Using the LPs (3.23) and (3.24) we can extend the LP 3.37) to the following one, for which we will show that the optimal value coincides with

$\max_{x\in\mathcal{A}} \phi(x)$ even if X is not positively invariant,

$$\begin{array}{lll} q^* = \inf & c+\varepsilon & \\ \text{s.t.} & \varepsilon, c \in \mathbb{R}, J, v \in \mathcal{C}^1(\mathbb{R}^n) & \\ & c + J \geq \phi & \text{on } X \\ & J + \nabla J \cdot f + v \leq \varepsilon & \text{on } X \\ & J \geq 0 & \text{on } X \\ & v - \nabla v \cdot f \geq 0 & \text{on } X. \end{array} \tag{3.39}$$

The LPs (3.37) and (3.39) are related through

1. $\beta = 1$ in (3.39),
2. J from (3.39) and J from (3.37) can be transformed into each via $V := J + c$

the additional decision variables v and ε in (3.39) take care of positive invariance backward in time direction and ε is the slack variable for the Lyapunov equation, that we know by now from the LP (3.35).

Proposition 3.22. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be Lipschitz continuous, φ be the semiflow corresponding to the differential equation $\dot{x} = f(x)$, $X \subset \mathbb{R}^n$ be compact and $\phi : X \to \mathbb{R}$ be continuous. Let $\mathcal{A}$ be the global attractor for X and assume that the basin of attraction $B_f(\mathcal{A})$ of $\mathcal{A}$ is open. It holds*

$$q^* = \max_{x\in\mathcal{A}} \phi(x).$$

Proof. Note first that

$$q^* \geq \max_{x\in\mathcal{A}} \phi(x).$$

This follows because, as for the LP (3.33), we have $V \leq \varepsilon$ on the global attractor, and hence it holds for the cost $c + \varepsilon$

$$c + \varepsilon \geq c + V \geq \phi \quad \text{on } \mathcal{A}.$$

To show that q^* is also a lower bound on $\max_{x\in\mathcal{A}} \phi(x)$, we proceed as in (3.34). Choose a Lyapunov function $V : \mathbb{R}^n \to [0, \infty)$ with $V^{-1}(\{0\}) = \mathcal{A}$ and $\nabla V \cdot f = -V$ on $B_f(\mathcal{A})$, which contains the maximum positively invariant set M_+. For v we choose a function as in [Schlosser 2021, proof of Theorem 2] with

$$v - \nabla v \cdot f = 0 \text{ on } X, v = 0 \text{ on } M_+ \text{ and } v < 0 \text{ on } X \setminus M_+. \tag{3.40}$$

Let $c^* := \max_{x\in\mathcal{A}} \phi(x)$ then for any $\delta > 0$

$$(c^* + \delta, \delta, m \cdot V, m \cdot v) \tag{3.41}$$

is feasible for (3.39) for $m, \in \mathbb{N}$ large enough and the corresponding cost is given by $c^* + 2\delta$. Since $\delta > 0$ was arbitrary we also $p^* \leq \max_{x\in\mathcal{A}} \phi(x)$ and the statement. □

3.3 Koopman semigroup; reproducing kernel Banach spaces and sparsity

Main contribution: We define the Koopman and Perron-Frobenius operators on reproducing kernel Banach spaces and state elementary properties concerning continuity and boundedness, Theorem 6.7.
For the Koopman and Perron-Frobenius operator on $\mathcal{C}(X)$, we show that subsystems induce a decomposition of principal eigenvalues and invariant measures, see Theorems 6.44 and 3.35.

The definitions, notations, concepts and results concerning dynamical, which are important systems for this section, are provided in the preliminary Sections 2.5 and 2.6.2. A detailed presentation of the results stated in this section is provided in Chapter 6 which is based on the texts [Ikeda 2022b, Schlosser 2022b].

Koopman theory views dynamical systems $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ from a lifted perspective – instead of looking at the flow φ directly, the Koopman semigroup $(T_t)_{t\in\mathbb{R}_+}$ describes the evolution of observables g, i.e.

$$T_t g := g \circ \varphi_t,$$

see Definition 2.41. The resulting family of operators $(T_t)_{t\in\mathbb{R}_+}$ forms a semigroup of (bounded) linear operators, see Theorem 2.45. This gives access to a fruitful combination of dynamical systems theory and functional analysis. Among the central objects in the study of the Koopman semigroup is its spectrum $\sigma(T_t)$ for $t \in \mathbb{R}_+$. As much as the Koopman operator itself, does the spectrum depend on the underlying space of observables. Firstly, if the underlying space is an algebra, the spectrum inherits the algebraic nature of the Koopman operator. This is illustrated by the following theorem on the Koopman semigroup on $\mathcal{C}(X)$, the space of continuous functions on X.

Theorem 3.23 ([Schaefer 1974, Theorem 4.4])**.** *Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system with X compact and $(T_t)_{t\in\mathbb{R}_+}$ be a Koopman semigroup on $\mathcal{C}(X)$. For each $t \in \mathbb{R}_+$ the point spectrum $\sigma_p(T_t)$, defined by*

$$\sigma_p(T_t) := \{\lambda \in \mathbb{C} : \text{ there exists } 0 \neq g \in \mathcal{C}(X) \text{ with } T_t g = \lambda g\}, \tag{3.42}$$

is multiplicative and the spectrum $\sigma(T_t)$, defined by

$$\sigma(T_t) := \{\lambda \in \mathbb{C} : \lambda \mathrm{Id} - T_t \text{ is invertible}\}, \tag{3.43}$$

is cyclic, i.e. for all $re^{i\alpha} \in \sigma(T_t)$ also $r \cdot e^{ik\alpha} \in \sigma(T_t)$ for all $k \in \mathbb{Z}$. Furthermore, for the spectral radius $r(T_t)$ we have $r(T_t) = 1$ and, in particular, it holds

$$\sigma(T_t) \subset \overline{B_1(0)}.$$

Note that, the multiplicativity of the point spectrum $\sigma_p(T_t)$ follows because T_t is an algebra homomorphism, and thus, the product and powers of eigenfunctions are again eigenfunctions. On the other hand, verifying that the whole spectrum $\sigma(T_t)$ is cyclic, requires fine arguments [Schaefer 1974, Theorem 4.4].

Additionally, a decomposition of the spectrum tells if the flow is (eventually) invertible. This is stated in the following theorem.

Theorem 3.24 (Spectrum of *Koopman semigroups*; [Scheffold 1971])**.** *Under the assumptions of Theorem 3.23, one of the following cases holds*

i. $\sigma(T_t) \subset \mathbb{S}^1$ *for all* $t > 0$ *which is the case if and only if* φ_t *is a bijective map on* X *for all* $t > 0$*.*

ii. for all $t > 0$ *there exists a cyclic set* $M_t \subset \mathbb{S}^1$ *with* $\sigma(T_t) = \{0\} \cup M_t$*. This is the case, exactly when the global attractor* $\mathcal{A}$ *does not coincide with* X*, is given by* $\mathcal{A} = \varphi_t(X)$ *for some* $t \in \mathbb{R}_+$ *and* φ_s *is injective on* $\mathcal{A}$ *for all* $s \in \mathbb{R}_+$*.*

iii. $\sigma(T_t) = \overline{B_1(0)}$ *for all* $t > 0$*.*

Theorem 3.24 states that from the spectrum Koopman semigroup alone we can only infer limited statements about the dynamical systems. Fortunately, the investigation of the spectrum of the Koopman operator on $\mathcal{C}(X)$ (and related spaces such as $\mathcal{C}^k(X)$ or $\mathrm{L}^2(X, \mu)$) does not end here and has produced many desirable results – such as ergodic partitioning of the state space [Küster 2015, Mezić 1999, Mezić 2005] or the concept of principal eigenvalues [Kvalheim 2021], which we will encounter again in Section 6.3.3.

Another important class of observables for the Koopman semigroup is $\mathrm{L}^2(X, \mu)$ for some invariant measure μ, see Example 2.42. This setting is one of the pillars of ergodic theory. Analysis in this context focuses on the asymptotic behavior of the system but comes at the price that the point evaluation $\mathrm{L}^2(X, \mu) \ni g \mapsto g(x)$ for $x \in X$ may not even be well-defined. The feature of bounded point evaluation is important for practical applications because it ensures robustness in the measurement process. We view this "restriction" as good and exciting motivation for investigating different underlying function spaces for the Koopman operator. In particular, it encourages looking at reproducing kernel Hilbert spaces (or Banach spaces).

3.3.1 Reproducing kernel Banach space domains

Both Koopman theory and reproducing kernel Hilbert spaces (RKHS) aim to translate problems into a functional analytic setting. There are many texts that investigate this fruitful connection; the following list is only a very short collection of beautiful directions in this area [Budisic 2012, Eisner 2015, Williams 2015, Korda 2018a, Saitoh 2016, Rudi 2020].

Due to their conceptual resemblance, it is not surprising that there has been work that merged both fields of Koopman and RKHS theories. Among these are: [Kawahara 2016, Williams 2015, Rosenfeld 2022, Das 2018, Alexander 2020, Klus 2020] for dynamic mode decomposition on RKHS, [Ishikawa 2018] for metrics

on dynamical systems, [Ikeda 2022a, Abanin 2017, Cowen 1995] in complex analysis, [Rosenfeld 2019] for system identification, [Zagabe 2023] for stability analysis, [Das 2018] for spectral analysis of the Koopman operator, [Das 2021] for compactification of unitary Koopman semigroups and weather prediction [Froyland 2021].

Our work on Koopman theory on reproducing kernel Banach spaces (RKBS) aims at generalizing the treatment of Koopman theory on RKHS to RKBS and to lay a base of elementary properties of Koopman operators on RKBS to build on for future research in this direction. In our opinion, the properties that make RKBS interesting for Koopman analysis are the following

1. *Continuous point evaluation:* For all x the point evaluations $g \mapsto g(x)$ are continuous in g with respect to the Banach space norm. This assures robustness, with respect to the observable g, of measurements $g(x)$ in already single events x.

2. *Explicitness:* The kernel gives access to the point evaluations and the geometry of the space. This allows explicit computations of orthogonal projections, such as in kernel dynamic mode decomposition [Williams 2015].

For the notion of RKBS we follow [Lin 2022], because their notion of RKBS unifies several other existing concepts of of RKBS. Furthermore, it follows parallel ideas to the well-understood case of RKHS. This analogy becomes especially useful in situations where a technical treatment of RKBS runs the risk of concealing the underlying arguments.

Our analysis is based on the interplay between the Koopman operator and its adjoint, the Perron-Frobenius operator. Both operators enjoy different advantages which we try to transfer through their intimate dual relation.

Definition 3.25. *Let $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS on X with kernel k. For a continuous time dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ the Koopman semigroup consists of the operators T_t for $t \in \mathbb{R}_+$, which are, as usual, given by $T_t : D(T_t) \to \mathcal{B}$ where*

$$T_t g := g \circ \varphi_t \quad \text{for } g \in D(T_t) := \{h \in \mathcal{B} : h \circ \varphi_t \in \mathcal{B}\}. \tag{3.44}$$

For discrete time dynamics $f : X \to X$ we consider the Koopman operator $T : D(T) \to \mathcal{B}$ with

$$Tg := g \circ f \quad \text{for } g \in D(T) := \{h \in \mathcal{B} : h \circ f \in \mathcal{B}\}.$$

The explicit treatment of the domain $D(T_t)$ respectively $D(T)$ indicates that it is not always the case that the Koopman operators T_t will be defined on the whole space. It can even easily happen that this domain is trivial, as we show in the following simple example.

Example 3.26. *Consider the RKHS $(\mathbb{R}^n, \mathbb{R}^n, \langle\cdot,\cdot\rangle, k)$ on $X = \mathbb{R}^n$ where $\langle\cdot,\cdot\rangle$ denotes the euclidean inner product and k be corresponding kernel. For $x \in \mathbb{R}^n$ we interpret it as a function $\hat{x}$ on $\mathbb{R}^n$ given by*

$$\hat{x} : \mathbb{R}^n \to \mathbb{R}, \hat{x}(a) := \langle a, x\rangle.$$

That means the RKHS $(\mathbb{R}^n, \mathbb{R}^n, \langle\cdot,\cdot\rangle, k)$ consists of linear functionals on $\mathbb{R}^n$. Therefore, non-linear dynamics f cannot induce a Koopman operator on the RKHS with a full domain. For instance, let $f(x) := \|x\| \cdot x$ be a discrete time dynamics then $D(T) = \{0\}$.

The domain of the Koopman operator describes how well the RKBS is adapted to the dynamics and opens up the analysis of the system via the Koopman operator. Therefore, the interplay between the domain of T_t, boundedness of T_t, and its adjoint operator, the Perron-Frobenius operator, is the central aspect of our investigation.

For defining the Perron-Frobenius operator on RKBS we are led by the well-studied case of the Perron-Frobenius operator on the space of measures $M(X)$ for dynamical systems $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ on compact sets X. For dirac measures δ_x in $x \in X$, the Perron-Frobenius operator P_t on the space of measures $M(X)$ acts by $P_t\delta_x = \delta_{\varphi_t(x)}$. The Perron-Frobenius operator on RKBS acts in the same way.

Remark 3.27. *We will denote the Perron-Frobenius operator on RKBS by K_t following [Kawahara 2016]. We hope there will not be any confusion due to the nomenclature because the letter K refers to the operator acting on the kernel functions $k(x,\cdot)$ and not to Koopman.*

Definition 3.28 (Perron-Frobenius operator on RKBS; [Kawahara 2016])**.** *Let $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS on X with kernel k. For a continuous time dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$, we call the semigroup $(K_t)_{t\in\mathbb{R}_+}$ the Perron-Frobenius semigroup, where for $t \in \mathbb{R}_+$ the operator $K_t : \operatorname{Span}\{k(x,\cdot) : x \in X\} \to \operatorname{Span}\{k(x,\cdot) : x \in X\}$ is defined by*

$$K_t\left(\sum_{i=1}^m a_i k(x_i,\cdot)\right) := \sum_{i=1}^m a_i k(\varphi_t(x_i),\cdot), \tag{3.45}$$

for $a_1, \ldots, a_m \in \mathbb{R}$. For discrete time dynamical systems with dynamic $f : X \to X$ the Perron-Frobenius operator $K_f : \operatorname{Span}\{k(x,\cdot) : x \in X\} \to \operatorname{Span}\{k(x,\cdot) : x \in X\}$ is given by

$$K_f k(x,\cdot) := k(f(x),\cdot) \text{ for } x \in X \tag{3.46}$$

and extended linearly to $\operatorname{Span}\{k(x,\cdot) : x \in X\}$.

In order to make the definition (3.45) respectively (3.46) meaningful we need to make the assumption that $\operatorname{Span}\{k(x,\cdot) : x \in X\}$ is linearly independent.

Assumption 3.29. *$\operatorname{Span}\{k(x,\cdot) : x \in X\}$ is linearly independent.*

Remark 3.30. *Note that in order to extend the definition of K_t to the closure of $\operatorname{Span}\{k(x,\cdot) : x \in X\}$ we need continuity properties of K_t which we will address in the next section.*

We were imprecise when we called the Perron-Frobenius operator the adjoint of the Koopman operator – it's rather the other way around!

Proposition 3.31. *The Koopman operator T is the adjoint of the Perron-Frobenius operator K in the sense of Definition 2.75.*

We revisit this result in Section 6.1 in Chapter 6, in Lemma 6.6 where we give a proof, where the notion of adjoint operator in RKBS is treated with more care.

A frequent encounter for Koopman and Perron-Frobenius operator on RKBS is the following asymmetry in their properties: The functional description $T_t g := g \circ \varphi_t$ can be stated easily but it can be difficult to verify for which g in the function space $\mathcal{B}$ it holds that $T_t g$ is still an element of $\mathcal{B}$. On the other hand, the Perron-Frobenius operator K_t is defined on the (dense) set $\mathrm{Span}\{k(x,\cdot) : x \in X\}$, but at the same time establishing a functional expression for $K_t h$ for an arbitrary element $h \in \mathcal{B}'$ is not obvious. In Theorem 6.7 we illustrate that these trade-offs relate to the closability of the Perron-Frobenius operator. In order to avoid additional notations we present some of the statements from Theorem 6.7 informally and consider discrete time dynamical systems.

Let $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS on X with kernel and $f : X \to X$ be a map.

1. If $\hat{f} : X \to X$ is another map then
$$K_f K_{\hat{f}} = K_{f \circ \hat{f}}. \tag{3.47}$$
2. Under a certain density assumption, see (2.88), it holds
$$f = \hat{f} \qquad \text{if and only if} \qquad K_f = K_{\hat{f}}. \tag{3.48}$$
3. The Koopman operator T_f is a closed operator and is bounded if and only if
$$D(T_f) = \mathcal{B}. \tag{3.49}$$
4. If X is compact, the map $x \mapsto k(x,\cdot) \in \mathcal{B}'$ is continuous, $\mathcal{B}$ embeds continuously and densely into $\mathcal{C}(X)$ and X has infinitely many points then K_f is not closed.

We give a precise and extended statement and proof of the above results in Theorem 6.7 of Chapter 6. Here, we only want to emphasize the underlying concepts.

Property (3.47) states that applying the Perron-Frobenius operator is a covariant functor, a property that is inherited from its natural definition. The property (3.48) tells us that we maintain all the information about the dynamics when considering the Perron-Frobenius operator K_f instead of the dynamics f. The condition (3.49) tells us that the boundedness of T_f is decided by its domain. The last of the above statements is in a similar direction and shows that in many cases we cannot expect the Perron-Frobenius operator to be closed. This shows that boundedness of the Koopman operator is a subtle question whose importance we want to address in the following paragraph.

Boundedness of the Koopman and Perron-Frobenius operators From a practical and theoretical perspective, the boundedness of T_f and K_f is an appealing property. For applications, it represents robustness with respect to the observables $g \in \mathcal{B}$, respectively $k(x,\cdot) \in \mathcal{B}'$. For theoretical investigations boundedness lies at the core of spectral analysis or, as stated around (3.49), determines that the Koopman operator can be defined on any element $g \in \mathcal{B}$.

Remark 3.32. *One reason why boundedness of the Perron-Frobenius operator on RKBS is more subtle than for the Perron-Frobenius operator on $M(X)$, is that the point evaluations in $\mathcal{C}(X)$ are extremal points of the unit ball in the dual space $\mathcal{C}(X)^*$. This geometric characterization of the point evaluations does not have to be true in RKBSs.*

Remark 3.32 emphasizes the importance of the geometry of the RKBS for the study of the Koopman and Perron-Frobenius operators on such spaces, see for example Proposition 6.27 in which we give a geometrical condition for the Koopman semigroup to be contractive.

It is important to point out that our results are conservative. The reason is that the choice of RKBS should be adapted to the dynamical system. Our work in [Ikeda 2022b] respectively Section 6.1 aims at providing a framework, examples, and discussions of natural structures for Koopman and Perron-Frobenius operators on RKBS, which hopefully can be of help for future work on combining RKBS and Koopman theory. To bypass the limits of our general treatment we emphasize the need for specific investigations of adapted choices of RKBS to perform detailed analysis such as in [Saitoh 2016, Ikeda 2022a, Ishikawa 2021, Abanin 2017, Cowen 1995, Doan 2017, Chacon 2007, Carswell 2003], to name only a few.

Preservation of structure The lifting process from the dynamical system to the Koopman operator is natural and should preserve structural properties of the dynamical systems, such as symmetry, sparsity, and conjugacy. These three concepts translate into Koopman language in the following way

1. *Symmetry:* A bijective map $\phi : X \to X$ is called a symmetry for the discrete time dynamics $f : X \to X$, if it holds

$$f \circ \phi = \phi \circ f.$$

   For the Koopman operator this translates to the following commutation relation, see [Salova 2019],

$$T_\phi T_f = T_f T_\phi \tag{3.50}$$

   where $T_f, T_\phi : \mathbb{R}^X \to \mathbb{R}^X$ are the Koopman operators for f respectively ϕ and $\mathbb{R}^X$ denotes the set of all functions from X to $\mathbb{R}$.

2. *Sparsity:* Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system on a cube $X \subset \mathbb{R}^n$, induced by a differential equation $\dot{x} = f(x)$. Let (I, f_I) be subsystem with corresponding flow φ_t^I. We show in Section 6.3 that subsystems (I, f_I), see in

Section 3.1 and Chapter 4 for the notations, induce an intertwining relation

$$T_{\Pi_I} T_t = T_t^I T_{\Pi_I} \qquad \text{for all } t \in \mathbb{R}_+ \tag{3.51}$$

for the Koopman operators T_t for the whole system, the Koopman operator T_t^I for the subsystem and the composition operator T_{Π_I}. See Proposition 6.38 for the precise statement.

3. *Conjugacy:* We call two discrete time systems $f : X \to X$ on a set X and $g : Y \to Y$ on a set Y conjugated if there exists a bijective map $\psi : X \to Y$ such that
$$\psi \circ f = g \circ \psi.$$
Again, this functional relation between the two maps carries over to a similarity relation of their Koopman operators, namely
$$T_f T_\psi = T_\psi T_g. \tag{3.52}$$

In Proposition 6.9 and Section 6.2.2 we show that the relations (3.50), (3.51) and (3.52) translate in dual form to the Perron-Frobenius operator on RKBS. With the notations from (3.50), (3.51) and (3.52) it holds for RKBS $(\mathcal{B}, \mathcal{B}', \langle \cdot, \cdot \rangle, k)$ on X with kernel that

1. *Symmetry (see Proposition 6.31):* Let K_f respectively K_ϕ the Perron-Frobenius operators for $f, \phi : X \to X$. Then it holds
$$K_f K_\phi = K_\phi K_f$$

2. *Sparsity (see Proposition 6.33):* Let $(\mathcal{B}_I, \mathcal{B}_I', \langle \cdot, \cdot \rangle_I, k_I)$ be an RKBS on $\Pi_I(X)$ with kernel. Then
$$K_{\Pi_I} K_t = K_t^I K_{\Pi_I}$$
where $K_{\Pi_I} : \operatorname{Span}\{k(x, \cdot) : x \in X\} \to \operatorname{Span}\{k_I(y, \cdot) : y \in \Pi_I(X)\}$ with $K_{\Pi_I} k(x, \cdot) := k_I(\Pi_I(x), \cdot)$ and $(K_t)_{t \geq 0}$ the Perron-Frobenius operator for the whole system and $(K_t^I)_{t \in \mathbb{R}_+}$ the Perron-Frobenius semigroup for the subsystem.

3. *Conjugacy (see Proposition 6.9):* Let $(\mathcal{B}_\psi, \mathcal{B}_\psi', \langle \cdot, \cdot \rangle_\psi, k_\psi)$ be the pullback RKBS with respect to ψ, see RKBS on Y with kernel. Let K_f and K_g be the Perron-Frobenius operators for the two discrete time systems induced by $f :\to X$ and $g : Y \to Y$. Then
$$T_\psi K_f = K_g T_\psi$$
where $T_\psi : \mathcal{B} \to \mathcal{B}_\phi$ is the composition operator isomorphism from Lemma 2.72.

The second point, sparsity, provides a good transition to the following section, in which we discuss the spectral decompositions of the Koopman operator on $\mathcal{C}(X)$ for sparse dynamical systems.

3.3.2 Sparsity structures for Koopman operators

Our approach towards sparse structures of the Koopman semigroup has close connections to the one for symmetry exploitation for Koopman theory in [Salova 2019]. The text [Salova 2019] acted also as a very illustrative guideline for how we wanted to exploit sparse structure via the Koopman operator – namely through an intertwining relation between the Koopman operators for the whole system and the Koopman operators for the subsystems.

We address sparsity in the sense of Section 4. For the Koopman on $\mathcal{C}(X)$ and Perron-Frobenius operator on $\mathcal{M}(X)$, we are particularly interested in certain spectral objects such as eigenfunctions and eigenmeasures and their decoupling into corresponding objects for subsystems. A related approach in the special case of cascaded systems can be found in [Mohr 2020b].

Definition 3.33. *An element $g \in \mathcal{C}(X)$ (respectively $\mu \in \mathcal{M}(X)$) is an eigenfunction (respectively eigenmeasure) with eigenvalue $\lambda \in \mathbb{C}$ of the Koopman (respectively Perron-Frobenius) operator if $g \neq 0$ (respectively $\mu \neq 0$) and for all $t \in [0, \infty)$ we have*

$$T_t g = e^{\lambda t} g \qquad \text{(respectively} \qquad P_t \mu = e^{\lambda t} \mu). \tag{3.53}$$

We will show that certain eigenfunctions and eigenmeasures can be decomposed according to subsystems. Therefore, we work with the setting from Chapter 4. We consider the ordinary differential equation on $\mathbb{R}^n$ given by

$$\dot{x} = f(x), \quad x(0) = x_0 \in \mathbb{R}^n \tag{3.54}$$

for a Lipschitz continuous vector field $f : \mathbb{R}^n \to \mathbb{R}^n$. As usual in this text, we denote the corresponding semiflow by φ_t for $t \in \mathbb{R}_+$. We restrict the dynamics to a subset $X \subset \mathbb{R}^n$ which we assume to be positively invariant. We understand sparsity in the sense of subsystems from Section 3.1 and Chapter 4 and use the corresponding notion, i.e. a subsystem induced by $I \subset [n]$ with flow φ^I. The Koopman operator for the subsystem induced by I with corresponding constraint set $\Pi_I(X)$ is denoted by T_t^I. Note that by Theorem 4.23 the set $\Pi_I(X)$ is also positively invariant. For simplicity, we consider T_t^I to act on $\mathcal{C}(\Pi_I(X))$; the Perron-Frobenius operator for the subsystem is denoted by P_t^I acting on $M(\Pi_I(X))$.

Sparsity in the language of the Koopman operator The definition of sparsity via (3.6)

$$\varphi_t^I \circ \Pi_I = \Pi_I \circ \varphi_t$$

translates to

$$V_I T_t^I = T_t V_I \quad \text{where } V_I : \mathcal{C}(\Pi_I(X)) \to \mathcal{C}(X) \text{ with } V_I g := g \circ \Pi_I. \tag{3.55}$$

The relation (3.55)) builds the starting point for our investigations. By dualizing we immediately obtain a corresponding relation between the Perron-Frobenius operators P_t^I and P_t on $M(\Pi_I(X))$ respectively $M(X)$. Regarding eigenvectors of the two semigroups, as a direct consequence of (3.55) we get the following decomposition result.

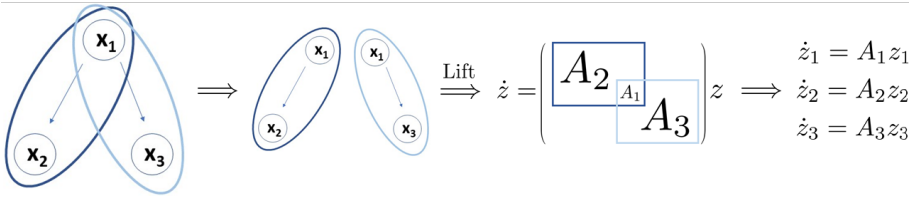

Figure 3.12: Illustration of sparse decomposition of the Koopman operator: 1. Identification of subsystems, 2. The subsystems induce a block structure for the Koopman operator (lift), 3. Exploitation of the block structure via decoupling of the lifted subsystems.

Proposition 3.34 ([Eisner 2015, page 233])**.** *Let X be positively invariant and I induce a subsystem. Then*

1. *If $g \in \mathcal{C}(\Pi_I(X))$ is an eigenfunction with eigenvalue λ for the Koopman operator T_t^I for the subsystem then $\hat{g} \in \mathcal{C}(X)$ defined by $\hat{g} := g \circ \Pi_I$ is an eigenfunction with eigenvalue λ of the Koopman operator T_t for the whole system.*

2. *If $\mu \in \mathcal{M}(X)$ is an eigenmeasure with eigenvalue λ of the Perron-Frobenius operator P_t, so is the push forward measure of μ by Π_I, i.e. $\mu_I := (\Pi_I)_{\#}\mu$, an eigenmeasure with eigenvalue λ for the Perron-Frobenius operator for the subsystem P_t^I.*

Decomposition of eigenvectors In our article [Schlosser 2022b] we intended to reverse the statement of Proposition 3.34, at least partially. This aims at computing eigenfunctions and eigenmeasures for the whole system based on computations of eigenfunctions and eigenmeasures for the subsystems, and vice versa. In Figure 3.12 we illustrate the idea via the simple sparse dynamical system from (3.5), i.e.

$$\begin{aligned} \dot{x}_1 &= f_1(x_1) \\ \dot{x}_2 &= f_2(x_1, x_2) \\ \dot{x}_3 &= f_3(x_1, x_3) \end{aligned}$$

For our decomposition result on the Perron-Frobenius operator, Theorem 6.39 in Chapter 6, we restrict to non-negative eigenmeasures of the Perron-Frobenius semigroup with eigenvalue $\lambda = 0$. These are exactly the invariant measures. Restricting to invariant probability measures allows us to use the crucial property that a certain set that we work with is compact. For eigenmeasures with different eigenvalues, the same construction does not guarantee compactness for the corresponding set.

In Theorem 6.39 we characterize exactly the invariant probability measures that decompose according to subsystems. Here we state a slightly informal version of Theorem 6.39.

Theorem 3.35 (Informal)**.** *Assume that $I_1, \ldots, I_N$ induce subsystems for (3.54) such that $\bigcup_{k=1}^{N} I_k = \{1, \ldots, n\}$ and that X is compact and invariant and decomposes*

accordingly (see Definition 4.17). For $k = 1, \dots, N$ *let* $\mu_k \in \mathcal{M}(\Pi_{I_k}(X))$ *be an invariant probability measure for the subsystem induced by* I_k*. Then there exists an invariant probability measure* $\mu \in \mathcal{M}(X)$ *such that*

$$(\Pi_{I_k})_{\#}\mu = \mu_k \text{ for all } k = 1, \dots, N$$

if and only if for all $k, l \in \{1, \dots, N\}$

$$(\Pi_{I_k \cap I_l})_{\#}\mu_k = (\Pi_{I_k \cap I_l})_{\#}\mu_l$$

This statement is motivated by the interpretation that the subsystem structure induces a block-structure of the Koopman respectively Perron-Frobenius operator, as illustrated in Figure 3.12.

In the case of eigenvectors of the Koopman operator, the situation is more subtle. The difference is that $M(X)$ as the dual space of $\mathcal{C}(X)$ enjoys more compactness properties than $\mathcal{C}(X)$ itself. An essential argument in the proof of Theorem 6.39 was the Markov-Kakutani Fixpoint theorem, which relies heavily on the weak-* compactness of the unit ball in $M(X)$. For the Koopman operator, we overcome this problem by restricting to a particular class of dynamical systems, namely systems with globally asymptotically stable fixed point, and further distinguished spectral objects – namely principal eigenvectors. Because these eigenvectors are uniquely determined [Kvalheim 2021] by certain smooth properties, see Definition 6.41, we show in Theorem 6.44 that the principal eigenvectors can be recovered from the subsystems.

Computational applications We follow the strategy for computational exploitation of sparsity, from Section 3.1, of decoupling the systems into their subsystems. In Section 6.3.4 we present computational applications of subsystems based on Theorems 6.39 and 6.44 to a sparse computation of extremal invariant measures and extended dynamic mode decomposition (EDMD). For EDMD we observed a lower computational cost and higher accuracy when we incorporate a priori knowledge on subsystems. This increase in performance is an effect of reducing the dimension of the problem and was observed as well in sparse computation for the global attractor in Section 5.3.

Sparsity and data Starting from [Budisic 2012], Koopman theory became a powerful tool in data analysis and thus, we should put our decomposition method into context with other sparse or sparsity promoting data-driven methods for Koopman theory. Data analysis got accessible to Koopman theory through the celebrated DMD. Therefore, we focus on (E)DMD for the following discussion on data-driven sparse methods for Koopman theory. The strength of our approach, compared to other sparsification techniques based on data, is that we make use of the inherited (exact) sparse structure of the dynamical system. Hence, for such cases, the sparse structure in the data comes naturally and allows lower dimensional treatments of the system, although it needs to be known a priori. In that respect, our approach is related to [Baddoo 2021] where physics-informed DMD is investigated and Chap-

ter 4.5 in [Baddoo 2021] already mentions how EDMD for specific cascade systems can be decoupled. In the case where prior information on the sparse structure of the underlying dynamical process, is not available, it is important to incorporate sparsity in the data. It is possible to directly implement sparsity in dynamic mode decomposition, which includes restricting to low-rank matrices [Kutz 2016, Chapter 9] and [Pan 2021, Balakrishnan 2021] or penalizing non-sparsity via ℓ^1 [Jovanović 2014, Kutz 2016].

Chapter 4

Sparsity structures and decompositions for dynamical systems

In this chapter, we present a decomposition of dynamical systems based on certain sparse structures. Based on ideas from [Chen 2018], we defined in [Schlosser 2020] sparse sub-structures, which we will call subsystems. Our goal is to *provide a global analysis of the dynamical system via an interplay of analyzing the subsystems and their interconnections*. The idea for our notion of subsystems is driven by the observation that in some systems there are families of states that evolve independently of others. Such families are the subsystems. We will give a precise definition of subsystems and show how they can be identified via the so-called sparsity graph of the dynamics. The main result in this section is that subsystems induce decouplings of the dynamical systems which give rise to decompositions of several classical objects from dynamical systems theory.

In terms of complexity, this approach proves beneficial whenever the dynamical system is sparse and the curse of dimensionality is inherent. Given a dynamical system and a related task at hand, we propose the following decoupled computational approach:

1. Identifying (a good class of) subsystems: This will be done Section 4.7 via simple graph algorithms.

2. Solving the task on each of the systems obtained from the first step: This depends on the problem at hand and a compatible method for solving the task on each subsystem.

3. Combining the results from the second step to a global object/statement/result: Here the interconnection between the subsystems plays an essential role.

Point 3. in this list, is the main part of the analysis in this chapter. We get back to point 2. in the next chapters when we merge this procedure with dynamic mode decomposition, or sum-of-squares technique for approximating the global attractor, maximum positively invariant sets, and the region of attraction for sparse systems.

4.1 Subsystems of dynamical systems

In this section, we want to define subsystems of a given continuous time dynamical system

$$\dot{x} = f(x),\ x_0 \in \mathbb{R}^n \text{ for } f = (f_1, \dots, f_n) : \mathbb{R}^n \to \mathbb{R}^n \text{ locally Lipschitz continuous.} \tag{4.1}$$

We further assume that the corresponding flow map φ_t exists for all $t \in \mathbb{R}_+$.

We restrict to continuous time dynamical systems but discrete time dynamical systems $x_{k+1} = f(x_k)$ can be treated in the same way. We will often use the notation $[n]$ for $\{1, \dots, n\}$ for natural numbers $n \in \mathbb{N}$.

A subsystem of (4.1) consists of certain subsets of the states of the systems. Therefore, we use the notation $\mathbb{R}^I := \prod_{i \in I}$ in the following definition of subsystems.

Definition 4.1 (Induced subsystem)**.** *Let $J \subset \mathbb{N}$ and $f : \mathbb{R}^J \to \mathbb{R}^J$. A subsystem of a dynamical system on $\mathbb{R}^J$ with dynamics f is a set of states $(x_i)_{i \in I}$ with $I \subset J$ such that $f_I := \Pi_I \circ f = (f_i)_{i \in I}$ only depends on the states $(x_i)_{i \in I}$ index by I. In that case, we say the pair (I, f_I), or sometimes just I, induces a subsystem of (J, f).*

To make it formally precise what we mean by "the function f_j depends on x_i" we define it as follows:

For $1 \le i \le n$, a function $g : \mathbb{R}^n \to \mathbb{R}^m$ does not depend on the coordinate x_i if for all $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in \mathbb{R}^{n-1}$ and any two $x_i, x_i' \in \mathbb{R}$ it holds

$$g(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) = g(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n). \tag{4.2}$$

For $I \subset [n]$ we say g depends on the set of variables $(x_i)_{i \in I}$ if g does not depend on (or is independent of) the remaining variables $(x_j)_{j \notin I}$. We say the function g depends explicitly on x_i if it is **not** independent of x_i.

Remark 4.2. *We typically consider the initial index set $J = [n]$, hence we talk about dynamical systems on $\mathbb{R}^n$. Up to relabelling, we can always assume that J is of this form. That is why in the following we will mostly look at subsystems induced by (I, f_I) for sets $I \subset [n]$.*

Example 4.3. *The trivial subsystems are the empty system, induced by $(\emptyset, f_\emptyset)$, and the full system is induced by $([n], f_{[n]} = f)$. For a less trivial example let $f : \mathbb{R}^5 \to \mathbb{R}^5$ be given by*

$$f(x_1, \dots, x_5) = \left(2x_1, \sin(x_1 x_2), e^{-x_3} \frac{1}{1 + x_1^2 x_4^2}, x_4 - x_1, \sin(x_2 x_4) x_5\right).$$

The subsystems are induced by the sets

$$\emptyset, \{1\}, \{1, 2\}, \{1, 3, 4\}, \{1, 4\}, \{1, 2, 4\}, \{1, 2, 4\}, \{1, 2, 3, 4\}, \{1, 2, 3, 4, 5\}. \tag{4.3}$$

In the above example, we can observe that the family of sets inducing subsystems is closed under taking unions and intersections.

Lemma 4.4. *Let (I, f_I) and (K, f_K) induce subsystems of a system induced by $([n], f)$. Then $(I \cap K, f_{I\cap K})$ and $(I \cup K, f_{I\cup K})$ induce subsystems. Further $(I \cap K, f_{I\cap K})$ is the largest set contained in (I, f_I) and (K, f_K) which still induces a subsystem. The pair $(I \cup K, f_{I\cup K})$ is the smallest set containing (I, f_I) and (K, f_K) which induces a subsystem.*

Proof. It suffices to show that $I \cap K$ and $I \cup K$ induce subsystems whenever I and K do. We begin by showing that this is true for $I \cap K$. Let I and K induce subsystems. Let $r \in [n]$ be an index for which $f_{I\cap K}$ depends explicitly on x_r. Because $f_{I\cap K}$ are components of f_I and f_K, also f_I and f_K depend explicitly on x_r. Because I and K induce subsystems, r belongs to I and K, thus $r \in I \cap K$. This shows that $I \cap K$ induces a subsystem. For $I \cup K$ the argument is even easier. The functions f_I respectively f_K are independent of the variables $[n] \setminus I$ respectively $[n] \setminus K$. Thus, $f_{I\cup K}$ is independent of the variables $[n] \setminus (I \cup K)$, i.e. $I \cup K$ induces a subsystem. □

Remark 4.5. *Lemma 4.4, together with the fact that $\emptyset$ and $[n]$ induce subsystems, states that the family of subsystems forms a topology on the discrete set $[n]$.*

Guided by product systems from Example 3.2, the idea of a subsystem is that we can treat it as a lower dimensional dynamical system.

Let (I, f_I) induce a subsystem. We view f_I as a vector field on $\mathbb{R}^I$ by identifying f_I with the map from $\mathbb{R}^I$ to $\mathbb{R}^I$ given by

$$(x_i)_{i\in I} \mapsto (f_i(x))_{i\in I} \text{ where } x = (x_1, \dots, x_n) \in \mathbb{R}^n \text{ satisfies } \Pi_I(x) = (x_i)_{i\in I}. \tag{4.4}$$

For instance we can choose $x = (x_1, \dots, x_n)$ with $x_j = 0$ whenever $j \notin I$.

Due to the definition of a subsystem, the map (4.4) is indeed independent of the specific choice of such x. We then can rephrase the definition of subsystems in terms of a functional relation between f, f_I, and Π_I. Namely, (I, f_I) induces a subsystem if and only if

$$f_I \circ \Pi_I = \Pi_I \circ f. \tag{4.5}$$

The semiflow induced by $f_I : \mathbb{R}^I \to \mathbb{R}^I$ is denoted $\varphi_t^I : \mathbb{R}^I \to \mathbb{R}^I$ for $t \in \mathbb{R}_+$. The statement of Corollary 4.6 is that by virtue of (4.5) the flows of the whole system and the subsystem are connected as follows

$$\varphi_t^I \circ \Pi_I = \Pi_I \circ \varphi_t. \tag{4.6}$$

The relation (4.6) states that subsystems are so-called factor systems, see for instance [Eisner 2015, Sinai 1989]. This additionally motivates our goal of transferring properties from one system to the other.

Corollary 4.6. *Let (I, f_I) induce a subsystem. Then $\dot{x}_I = f_I(x_I)$ induces a well defined dynamical system and (4.6) holds.*

Proof. Because f is locally Lipschitz continuous the same is true for f_I interpreted as a map from $\mathbb{R}^I$ to $\mathbb{R}^I$ as in (4.4). It follows that $\dot{x}_I = f_I(x_I)$ is well defined and induces a dynamical system on $\mathbb{R}^I$. For the second claim let $x_I \in \mathbb{R}^I$ and $x \in \mathbb{R}^n$ be any vector extending x_I to a vector in $\mathbb{R}^n$. Let us denote $x = (x_I, \hat{x}_I)$ after a possible relabelling of the components. We have for the curve $t \mapsto \Pi_I \varphi_t(x)$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Pi_I \circ \varphi_t)(x) = \Pi_I \frac{\mathrm{d}}{\mathrm{d}t}\varphi_t(x) = \Pi_I f(\varphi_t(x)) = f_I(\varphi_t(x)) = f_I(\Pi_I(\varphi_t(x)).$$

Hence for all $\hat{x}_I$ we have that $\Pi_I \varphi_t(x_I, \hat{x}_I)$ solves the differential equation $\dot{y} = f_I(y)$ with initial condition $y(0) = x_I$. Hence, it follows that $\Pi_I \varphi_t(x) = \varphi_t^I(x_I)$. □

By Corollary 4.6 the dynamics f_I induce a dynamical system on x_I. Hence, we may consider (I, f_I) a dynamical system and can extend the concept of subsystems also to subsystems themselves.

Lemma 4.7. *Let (I, f_I) be a subsystem of (J, f_J) and (J, f_J) be a subsystem of (K, f_K). Then (I, f_I) is a subsystem of (K, f_K).*

Proof. From the definition of subsystems we have $I \subset J \subset K$ and by (4.5)

$$f_I \circ \Pi_I = \Pi_I \circ f_J \quad , \quad f_J \circ \Pi_J = \Pi_J \circ f_K.$$

It follows

$$\Pi_I \circ f_K = \Pi_I \circ \Pi_J \circ f_K = \Pi_I \circ f_J \circ \Pi_J = f_I \circ \Pi_I \circ \Pi_J = f_I \circ \Pi_I$$

because for the canonical projections we have $\Pi_A \circ \Pi_B = \Pi_A$ whenever $A \subset B$. □

In the following sections, we study the interplay between properties of the whole system and properties of the subsystems.

4.2 Properties inherited by subsystems

In this section, we investigate some properties that are inherited from the whole system to the subsystem. This will include equilibrium points, periodic orbits, invariance, attraction, and stability.

Our main ingredient, now and in the following, is (4.6). Namely, let $I \subset [n]$ induce a subsystem, then

$$\varphi_t^I \circ \Pi_I = \Pi_I \circ \varphi_t \quad \text{for all } t \in \mathbb{R}_+.$$

Most of the results we mention in this section follow directly from this relation and some can be found in [Eisner 2015, Chapter 2] and [Sinai 1989, Chapter 1].

For our analysis of subsystems, we begin with the simplest objects – equilibrium points. A point $x^* \in \mathbb{R}^n$ is an equilibrium for the dynamical system (4.1) if and only

$$f(x^*) = 0.$$

By definition of a subsystem, we get

$$f_I(\Pi_I(x^*)) = \Pi_I(f(x^*)) = \Pi_I(0) = 0, \tag{4.7}$$

which means that $\Pi_I(x^*)$ is an equilibrium point for the subsystem induced by I. In Proposition 4.8, we list some more properties that are inherited by subsystems. The statements from Proposition 4.8 follow immediately from (4.6), see also [Eisner 2015, Sinai 1989]. The notion of positive invariance, attraction, and stability can be found in Definitions 2.2 and 2.3.

Proposition 4.8. *Let $I \subset [n]$ induce a subsystem for (4.1). Then,*

1. *If $x^* \in \mathbb{R}^n$ is equilibrium point for the whole system, then $\Pi_I(x^*) \in \mathbb{R}^I$ is an equilibrium point for the subsystem.*
2. *If $x \in \mathbb{R}^n$ has a periodic orbit with period $T > 0$ for the whole system, i.e. $\varphi_T(x) = x$, then $\Pi_I(x) \in \mathbb{R}^I$ has a periodic orbit with period $T > 0$ for the subsystem.*
3. *If $M \subset \mathbb{R}^n$ is (positively) invariant under the flow of the whole system then $\Pi_I(M) \subset \mathbb{R}^I$ is (positively) invariant under the flow for the subsystem.*
4. *If M is attractive under the flow of the whole system then $\Pi_I(M)$ is attractive under the flow for the subsystem.*
5. *If $M \subset \mathbb{R}^n$ is stable under the flow of the whole system then $\Pi_I(M)$ is stable under the flow for the subsystem.*
6. *If M is asymptotically stable under the flow of the whole system then $\Pi_I(M)$ is asymptotically stable under the flow for the subsystem.*

Proof. The first statement was shown in (4.7). For periodic orbits we use (4.6) and get for a point $x \in \mathbb{R}^n$ and $T \in \mathbb{R}_+$ with $\varphi_T(x) = x$ that

$$\varphi_T^I(\Pi_I(x)) = \Pi_I(\varphi_T(x)) = \Pi_I(x).$$

To show the third statement, we begin with positive invariance. Let M be positively invariant, i.e. $\varphi_t(x) \in M$ for all $x \in M$ and $t \in \mathbb{R}_+$. It follows for all $x \in M$

$$\varphi_t^I(\Pi_I(x)) = \Pi_I(\varphi_t(x)) \in \Pi_I(M),$$

i.e. that $\Pi_I(M)$ is positively invariant for the subsystem. For the statement concerning invariance, let $t \in \mathbb{R}_+$, $y \in \Pi_I(M)$ and $x \in M$ with $\Pi_I(x) = y$. By invariance of M there is $z \in M$ with $\varphi_t(z) = x$. Hence for $\Pi_I(z) \in M$ we get $\varphi_t^I(\Pi_I(z)) = \Pi_I(\varphi_t(z)) = \Pi_I(x) = y$. Since $y \in M$ was arbitrary it follows that $\Pi_I(M)$ is invariant. To show the fourth statement, let $y \in \mathbb{R}^I$ and $x \in \mathbb{R}^n$ with $\Pi_I(x) = y$. Because M is attractive, for all $t \in \mathbb{R}_+$ there exists $m_t \in M$ such that $\|\varphi_t(x) - m_t\|_2 \to 0$ as $t \to \infty$. For the points $\Pi_I(m_t) \in \Pi_I(M)$, we get

$$\begin{aligned} \left\|\varphi_t^I(y) - \Pi_I(m_t)\right\|_2 &= \left\|\varphi_t^I(\Pi_I(x)) - \Pi_I(m_t)\right\|_2 = \|\Pi_I(\varphi_t(x)) - \Pi_I(m_t)\|_2 \\ &= \|\Pi_I(\varphi_t(x) - m_t)\|_2 \leq \|\varphi_t(x) - m_t\|_2 \to 0 \text{ as } t \to \infty. \end{aligned}$$

Where the inequality in the above estimate holds because the projection Π_I is a contraction. This shows that $\Pi_I(M)$ is attractive for (I, f_I). For 5., let $U_I \subset \mathbb{R}^I$ be an open neighbourhood of $\Pi_I(M)$. We define

$$U := \{x \in \mathbb{R}^n : \Pi_I(x) \in U\} = \{x = (x_1, \dots, x_n) \in \mathbb{R}^n : (x_i)_{i\in I} \in U\}.$$

Then $U \subset \mathbb{R}^n$ is an open neighbourhood of M and $\Pi_I(U) = U_I$. By stability of M, we can find a neighborhood V of M such that $\varphi_t(V) \subset U$ for all $t \in \mathbb{R}_+$. We claim that $V_I := \Pi_I(V)$ is an open neighbourhood of $\Pi_I(M)$ such that $\varphi_t^I(V_I) \subset U_I$ for all $t \in \mathbb{R}_+$. From $M \subset V$ we obtain $\Pi_I(M) \subset \Pi_I(V) = V_I$. Since Π_I is an open map, i.e. it maps open sets to open sets, we get that V_I is open. Finally, we get, again by (4.6),

$$\varphi_t^I(V_I) = \varphi_t^I(\Pi_I(V)) = \Pi_I(\varphi_t(V)) \subset \Pi_I(U) = U_I.$$

This proves 5. The last statement is just a combination of 4. and 5. □

Now we show that also the existence of Lyapunov functions (see Definition 2.8) is inherited from the whole system to its subsystems.

Corollary 4.9. *Let I induce a subsystem for (4.1). If there exists a strict Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}$ for the whole system then there exists a strict Lyapunov function $V_I : \mathbb{R}^I \to \mathbb{R}_+$ for the subsystem (I, f_I) such that $V_I^{-1}(\{0\}) = \Pi_I(V^{-1}(\{0\}))$.*

Proof. We use the intimate connection between asymptotically stable sets and strict Lyapunov functions from Theorem 2.10. Because V is a strict Lyapunov function the set $V^{-1}(\{0\})$ is asymptotically stable. By Proposition 4.8 6., the set $\Pi_I(V^{-1}(\{0\}))$ is asymptotically stable for the subsystem. Again by Theorem 2.11, there exists a Lyapunov V_I for the subsystem with $V_I^{-1}(\{0\}) = \Pi_I(V^{-1}(\{0\}))$. □

Conversely to the above Corollary, we show in Proposition 4.15 that Lyapunov functions and Hamilton functions for the whole system can be obtained from subsystems.

Proposition 4.10. *Let I induce a subsystem of (4.1). Then*

1. *If a function $V : \mathbb{R}^I \to \mathbb{R}_+$ is a (strict) Lyapunov function for the subsystem induced by I then $\overline{V} := V \circ \Pi_I$ is a (strict) Lyapunov function for the whole system.*

2. *If g is a Hamilton function for the subsystem (I, f_I) then $\overline{g} := g \circ \Pi_I$ is a Hamilton function for the whole system.*

Proof. This result is again an immediate consequence of (4.6). Let V be a Lyapunov function for the subsystem induced by I. For $\overline{V}$ we get for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}_+$

$$\overline{V}(\varphi_t(x)) = V(\Pi_I(\varphi_t(x)) = V(\varphi_t^I(\Pi_I(x)) \le V(\Pi_I(x)) = \overline{V}(x).$$

This shows that $\overline{V}$ is a Lyapunov function and that $\overline{V}$ is a strict Lyapunov function whenever V is. An analog computation for $\overline{g}$ shows 2. □

We conclude this section by partially reversing points 4.-6. of Proposition 4.8.

Corollary 4.11. *Let I induce a subsystem of (4.1) and $M \subset \mathbb{R}^I$ be closed. Then,*

1. *If M is positively invariant under the flow of the subsystem induced by I then $\Pi_I^{-1}(M) \subset \mathbb{R}^n$ is positively invariant under the flow of the whole system.*

2. *If M is attractive under the flow of the subsystem induced by I then $\Pi_I^{-1}(M) \subset \mathbb{R}^n$ is attractive under the flow of the whole system.*

3. *If M is asymptotically stable under the flow of the subsystem induced by I then the set $\tilde{M} := \Pi_I^{-1}(M) \subset \mathbb{R}^n$ is asymptotically stable for the whole system.*

Proof. The first statement is easily verified: Let $x \in \Pi_I^{-1}(M_I)$ and $t \in \mathbb{R}_+$ then

$$\Pi_I(\varphi_t(x)) = \varphi_t^I(\Pi_I(x)) \in M_I$$

by positive invariance of M_I. Thus, $\Pi_I^{-1}(M_I)$ is invariant. For the second statement, let $x \in \mathbb{R}^n$. Because $M \subset \mathbb{R}^I$ is attractive, for each $t \in \mathbb{R}_+$ there exist $m_t = (m_{t,i})_{i \in I} \in M$ with

$$\left\| \varphi_t^I(\Pi_I(x)) - m_t \right\|_2 \to \infty \quad \text{as } t \to \infty. \tag{4.8}$$

For $t \in \mathbb{R}_+$ we define the points $\overline{m}_t = (\overline{m}_{t,1}, \dots, \overline{m}_{t,n}) \in \mathbb{R}^n$ by

$$\overline{m}_{t,j} := \begin{cases} m_{t,j}, & j \in I \\ (\varphi_t(x))_j, & j \notin I \end{cases}$$

where $(\varphi_t(x))_j$ denotes the j-th component of $\varphi_t(x) \in \mathbb{R}^n$. By construction we have $\overline{m}_t \in \Pi_I^{-1}(M)$ and we get by (4.8)

$$\|\varphi_t(x) - \overline{m}_t\|_2 = \left\| \varphi_t^I(\Pi_I(x)) - m_t \right\|_2 \to \infty \quad \text{as } t \to \infty.$$

This shows that $\Pi_I^{-1}(M)$ is attractive. For the third statement, by Theorem 2.10, we can find a strict Lyapunov function $V : \mathbb{R}^I \to [0, \infty)$ for the subsystem induced by I with $M = V^{-1}(\{0\})$. By Proposition 4.10 the function $\overline{V} := V \circ \Pi_I$ is a strict Lyapunov function for the whole system with

$$\Pi_I^{-1}(M) = \Pi_I^{-1}(V^{-1}(\{0\})) = \overline{V}^{-1}(\{0\}) \tag{4.9}$$

Therefore, by Theorem 2.10, the set $\Pi_I^{-1}(M)$ is asymptotically stable under the flow of the whole system. □

We use the above results from Proposition 4.10 and Corollary 4.11 as a transition to the next Chapter in which we will investigate more properties of the whole system that can be analyzed through (many) subsystems.

4.3 Subsystem based decompositions of dynamical systems

In this section, we reverse the question from the previous section. Now we are interested in properties of the whole system that can be inferred from the subsystems only. We begin with the most fundamental question about conserving information of the original dynamical system when considering only certain subsystems:

Can we recover the flow φ of the whole systems from flows of certain subsystems?

To answer this question, we will use again the relation (4.6). This relation states that the flow φ^I of a subsystem induced by $I \subset [n]$ tells us the components $\Pi_I \circ \varphi_t = \varphi_t^I$ of the flow φ of the whole system. Thus, we can recover the flow φ from a family of subsystems $(I_1, f_{I_1}), \dots, (I_k, f_{I_k})$ if the subsystems cover each state, i.e.

$$\bigcup_{l=1}^{k} I_l = [n]. \tag{4.10}$$

This leads to the following definition from [Schlosser 2020].

Definition 4.12 (Subsystem decomposition)**.** *We say that $I_1, \dots, I_k \subset [n]$ induces a subsystem decomposition if I_l induces a subsystem for $l = 1, \dots, k$ and (4.10) holds.*

Remark 4.13. *We sometimes also say that $(I_1, f_{I_1}), \dots, (I_k, f_{I_k})$ is a subsystem decomposition if $I_1, \dots, I_k$ induces a subsystem decomposition.*

As claimed, the next lemma shows that for subsystem decompositions we can reconstruct the flow for the whole system from the flow of the subsystems.

Lemma 4.14. *Let $I_1, \dots, I_k \subset [n]$ induce a subsystem decomposition with corresponding flows $\varphi^{I_1}, \dots, \varphi^{I_k}$. For $x_0 \in \mathbb{R}^n$ the function $x(t) := \varphi_t(x_0)$ is the unique map $x(\cdot)$ for which it holds*

$$\varphi_t^{I_l}(\Pi_{I_l}(x_0)) = \Pi_{I_l}(x(t)) \text{ for } l = 1, \dots, k \text{ for all } t \in \mathbb{R}_+. \tag{4.11}$$

Proof. That $x(\cdot)$ solves (4.11) follows from (4.6). That $I_1, \dots, I_k$ induces a subsystem decomposition implies that each coordinate of a function $x(\cdot)$ solving (4.11) is uniquely determined, i.e. the whole function $x(\cdot)$ is uniquely determined. □

Next, we analyze how we can decompose objects of the whole system into their counterparts in the subsystems. When decomposing sets according to subsystems we use a "gluing" procedure. For index sets $I_1, \dots, I_k \subset [n]$ and sets $M_l \in \mathbb{R}^{I_l}$ for $l = 1, \dots, k$ we glue together the sets $M_1, \dots, M_k$ to a set $S(M_1, \dots, M_k) \subset \mathbb{R}^n$ by

$$S(M_{I_1}, \dots, M_{I_k}) = \{x \in \mathbb{R}^n : \Pi_{I_l} x \in M_{I_l} \text{ for } l = 1, \dots, k\} \tag{4.12}$$

Our proposed decoupling procedure is illustrated in Figure 4.1 at an example of a dynamical system that decomposes into two subsystems.

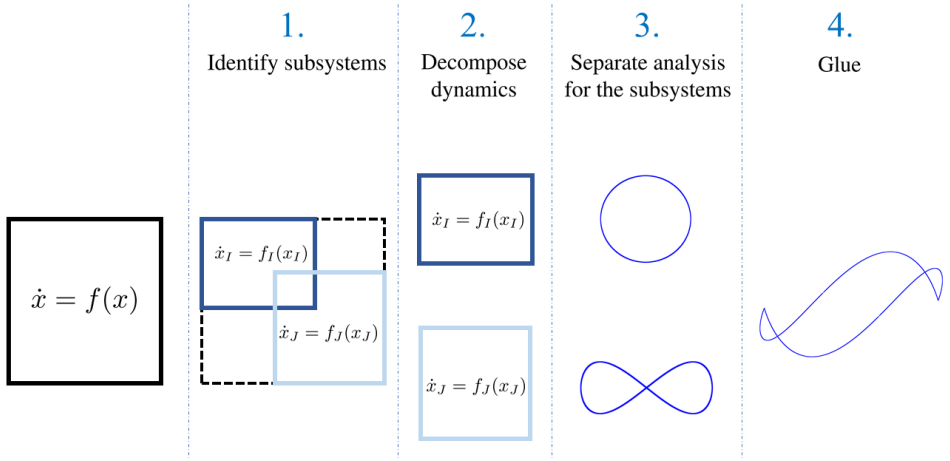

Figure 4.1: Illustration of a decomposition procedure for a sparse dynamical system with two subsystems induced by I and J.

Proposition 4.15 states that, for certain properties, the decomposition (4.12) is indeed the correct way of "gluing together" sets obtained from the subsystems.

Proposition 4.15. *Let $I_1, \dots, I_k \subset [n]$ induce a subsystem decomposition. Then,*

1. *$x \in \mathbb{R}^n$ is an equilibrium point (or has a periodic orbit with period T) for the whole system if and only if for all $l = 1, \dots, k$ the point $\Pi_{I_l}(x)$ is an equilibrium point (or has a periodic orbit with period T) for the subsystem induced by I_l.*

2. *If for $1 \leq l \leq k$ the set $M_l \subset \mathbb{R}^{I_l}$ is (positively) invariant under the flow of the subsystem induced by I_l then $S(M_1, \dots, M_k)$ is (positively) invariant under the flow of the whole system.*

3. *If for $1 \leq l \leq k$ the set $M_l \subset \mathbb{R}^{I_l}$ is compact and attractive under the flow of the subsystem induced by I_l then $S(M_1, \dots, M_k)$ is compact and attractive under the flow of the whole system.*

4. *If for $1 \leq l \leq k$ the set $M_l \subset \mathbb{R}^{I_l}$ is asymptotically stable under the flow of the subsystem induced by I_l then $S(M_1, \dots, M_k)$ is asymptotically stable under the flow of the whole system.*

Proof. 1. A point x being an equilibrium point means $f(x) = 0$. Because $I_1, \dots, I_k$ induces a subsystem decomposition, this is equivalent to $f_I(x) = 0$ for all $j = 1, \dots, k$. The argument for periodic orbits is similar.

2. If $M_1, \dots, M_l$ are positively invariant for the corresponding subsystems then for each $x \in S(M_1, \dots, M_k)$, i.e. it holds $\Pi_{I_l}(x) \in M_l$ for all $1 \leq l \leq k$, we get for all $t \in \mathbb{R}_+$

$$\Pi_{I_l}(\varphi_t(x)) = \varphi_t^{I_l}(\Pi_{I_l}(x)) \in M_l.$$

That shows $\varphi_t(x) \in S(M_1, \dots, M_l)$, i.e. positive invariance of $S(M_1, \dots, M_l)$. For the statement concerning invariance, it suffices to note that invariance is the same as positive invariance in forward and in backward time direction. Positive invariance in backward time direction can be treated via the time-reversing vector field $-f$. Note that the dynamical system induced by f and $-f$ have the same subsystems. Therefore, we can apply the first part of this statement in the backward time direction as well and conclude the claim about invariance.

3. Because the sets $M_1, \dots, M_k$ are compact, also the set $S(M_1, \dots, M_k)$ is compact. We use that a compact set is attractive if and only if for all $x \in \mathbb{R}^n$ it contains all accumulations points of the trajectory $(\varphi_t(x))_{t\in\mathbb{R}_+}$. Therefore, let $1 \leq l \leq k$, $x \in M_+$, $\mathbb{R} \ni t_m \nearrow \infty$ and $y \in X$ with $\varphi_{t_m}(x) \to y$. We have

$$\varphi^{I_l}_{t_m}(\Pi_{I_l}(x)) = \Pi_{I_l}\varphi_{t_m}(x) \to \Pi_{I_l}(y).$$

Hence $\Pi_{I_l}(y)$ is an accumulation point for the trajectory $(\varphi^{I_l}_t(\Pi_{I_l}(x)))_{t\in\mathbb{R}_+}$. Because M_l is compact and attractive, it follows $\Pi_{I_l}(y) \in M_l$. In other words, we have $y \in S(M_1, \dots, M_k)$, which was to be shown.

4. Because for $1 \leq l \leq k$, the set M_l is asymptotically stable for the subsystem (I_l, f_{I_l}), we can find, by Theorem 2.10, a strict Lyapunov function $V_l : \mathbb{R}^{I_l} \to [0, \infty)$ with $V_l^{-1}(\{0\}) = M_l$. Using the Lyapunov functions $V_1, \dots, V_k$ we construct another strict Lyapunov function V for the whole system with $V^{-1}(\{0\}) = S(M_1, \dots, M_k)$. We define by $V(x) := \sum_{l=1}^{k} V_l(\Pi_{I_l}(x))$. Indeed, we have $V(x) \geq 0$, $V(x) = 0$ if and only if $x \in S(M_1, \dots, M_k)$ and also $V(\varphi_t(x)) < V(x)$ for all $x \notin S(M_1, \dots, M_k)$. Again, from Theorem 2.10 we conclude that $S(M_1, \dots, M_k) = V^{-1}(\{0\})$ is asymptotically stable.

□

Combining Propositions 4.8 and 4.15, for the particular case of equilibrium points, leads to the following corollary.

Corollary 4.16. *Let $(\mathbb{R}^n, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system given by $\dot{x} = f(x)$ for a Lipschitz continuous map $f : \mathbb{R}^n \to \mathbb{R}^n$. Let $I_1, \dots, I_k \subset [n]$ induce a subsystem decomposition. An equilibrium point $x^* \in \mathbb{R}^n$ is*

1. *attractive if and only if for all $1 \leq l \leq k$ the point $\Pi_{I_l}(x^*)$ is attractive for the subsystem (I_l, f_{I_l}).*
2. *stable if and only if for all $1 \leq l \leq k$ the point $\Pi_{I_l}(x^*)$ is stable for the subsystem (I_l, f_{I_l}).*
3. *asymptotically stable if and only if for all $1 \leq l \leq k$ the point $\Pi_{I_l}(x^*)$ is asymptotically stable for the subsystem (I_l, f_{I_l}).*

Proof. The necessity part in the statements follows from Proposition 4.8 4.-6. The sufficiency in 1. and 3. is treated in Proposition 4.15 3. and 4. It remains to

show that stability of $\Pi_{I_l}(x^*)$ for all $l = 1, \dots, k$ is also sufficient for stability of x^*. Therefore, let U be an open neighbourhood of x^*, $\varepsilon > 0$ such that $B_\varepsilon(x^*) \subset U$, and $1 \leq l \leq k$. By stability of $\Pi_{I_l}(x^*)$ for the subsystem induced by I_l there exists $\delta_l > 0$ with

$$\varphi_t^{I_l}(B_\delta(\Pi_{I_l}(x^*)) \subset B_{\frac{\varepsilon}{n}}(\Pi_{I_l}(x^*)). \tag{4.13}$$

We choose $\delta := \min_{l=1,\dots,k} \delta_l > 0$. We show that it holds

$$\varphi_t(B_\delta(x^*)) \subset U \tag{4.14}$$

for all $t \in \mathbb{R}_+$, by showing that even $\varphi_t(B_\delta(x^*)) \subset B_\varepsilon(x^*)$ holds. Let $x^* = (x_1^*, \dots, x_n^*)$ and $\varphi_t(\cdot) = (\varphi_{1,t}(\cdot), \dots, \varphi_{n,t}(\cdot))$. For each component $i = 1, \dots, n$ let $1 \leq l(i) \leq k$ such that $i \in I_{l(i)}$. Let $x \in B_\delta(x^*)$. From (4.13) we get

$$|x_i^* - \varphi_{i,t}(x)| \leq \left\| \Pi_{I_l}(x^*) - \varphi_t^{I_l}(\Pi_{I_l}(x)) \right\|_2 < \frac{\varepsilon}{n}.$$

We conclude

$$\|x^* - \varphi_t(x)\|_2 \leq \sum_{i=1}^{n} |x_i^* - \varphi_{i,t}(x)| < n\frac{\varepsilon}{n} = \varepsilon.$$

This shows (4.14). Statement 3. is just the combination of 1. and 2. □

4.4 Systems with state constraints

So far we have only considered systems on the whole space $\mathbb{R}^n$. With regard to applications, this is too restrictive and this section aims to extend the notion of subsystems to state constrained dynamical systems. This means we equip the dynamical system (4.1) with a constraint set $X \subset \mathbb{R}^n$ and we consider only solutions which stay in X for all positive times. The guiding principle for the decomposition of the state constraints is:

Feasibility for the whole system should be determined exactly by feasibility for the subsystems. (4.15)

Our treatment including a constraint set is rather technical therefore we provide a simple intuition.

Intuition for state constrained subsystems Again we borrow intuition from product systems. Consider the systems on $\mathbb{R}^{n_1}$ respectively $\mathbb{R}^{n_2}$ respectively $\mathbb{R}^{n_1+n_2}$ with dynamics f_1 respectively f_2 respectively $f_1 \otimes f_2$ from Example 3.2. For $i = 1, 2$ we additionally equip the subsystems on $\mathbb{R}^{n_i}$ with dynamics f_i with constraint sets $X_i \subset \mathbb{R}^{n_i}$. It is natural to demand that both systems should be state constrained subsystems of their product system with state constraint $X := X_1 \times X_2 \subset \mathbb{R}^{n_1+n_2}$. The important property is that membership of (x_1, x_2) to the constraint set $X_1 \times X_2$ can be verified from simultaneous membership of x_1 to X_1 and x_2 to X_2. Following this idea leads to the notion of decomposition for X in Definition 4.17 and relates closely to the decomposition which we have already seen in (4.12).

Definition 4.17. *We say that the constraint set X decomposes according to a family of index sets $J_1, \dots, J_N \subset [n]$ if there exist $X_1 \subset \mathbb{R}^{|J_1|}, \dots, X_N \subset \mathbb{R}^{|J_N|}$ such that*

$$X = S(X_1, \dots, X_N) = \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in X_l \text{ for } l = 1, \dots, N\}. \tag{4.16}$$

If X is the constraint set for a dynamical system (4.1) and the sets $J_1, \dots, J_N$ induce a subsystem decomposition then a corresponding state constrained subsystem decomposition is given by the subsystems (J_l, f_{J_l}) equipped the constraint set X_l.

Remark 4.18. *The sets $X_1, \dots, X_N$ from Definition 4.17 seem to be a priori unknown. But, if we know that X decomposes according to $J_1, \dots, J_N$, we can choose*

$$X_l := \Pi_{J_l}(X) \text{ for } 1 \le l \le N,$$

in (4.16). Note that $X \subset \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in \Pi_{J_l}(X), 1 \le l \le N\}$ is trivially satisfied. On the other hand, if X decomposes according to $J_1, \dots, J_N$ via sets $X_1', \dots, X_N'$, then

$$X_l' \supset \Pi_{J_l}(X) \text{ for all } 1 \le l \le N \tag{4.17}$$

To verify (4.17), assume that there is $1 \le l \le N$ with $X_l' \not\supseteq \Pi_{J_l}(X)$, i.e. there is $x \in X$ with $\Pi_{J_l}(x) \notin X_l'$. This contradicts that X decomposes according to $J_1, \dots, J_N$ via $X_1', \dots, X_N'$. Now, from (4.17) we conclude

$$\begin{aligned} X &= \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in X_l' \text{ for } l = 1, \dots, N\} \\ &\supset \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in \Pi_{J_l}(X) \text{ for } l = 1, \dots, N\} \supset X. \end{aligned}$$

The notion of a set X decomposing accordingly covers the situation where the set X factors into a cartesian product, see Remark 4.19. Note that a factorization of X into a cartesian product is not required for X to decompose accordingly because the sets J_l in Definition 4.17 are allowed to overlap.

Remark 4.19. *In the case where X factors into a cartesian product*

$$X = X_1 \times \cdots \times X_N \text{ with } X_l \subset \mathbb{R}^{n_l} \text{ for } l = 1, \dots, N$$

the set X decomposes according to $J_1, \dots, J_N$ with $J_1 := \{1, \dots, n_1\}$ and $J_l = \{n_1 + \dots + n_{l-1} + 1, n_1 + \dots n_l\}$ for $l = 2, \dots, N$, namely $X = \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in X_l \text{ for } l = 1, \dots, N\}$.

4.4.1 Subsystem based decompositions for systems with state constraints

We return to the question from (4.15). We answer this question in the next lemma.

Lemma 4.20. *Let $(I_1, f_{I_1}), \dots, (I_k, f_{I_k})$ be a subsystem decomposition with corresponding flows $\varphi^{I_1}, \dots, \varphi^{I_k}$. If the system is equipped with a constraint set X which decomposes according to the subsystem decomposition $I_1, \dots, I_k$ then it holds for $x_0 \in X$ and $t \in \mathbb{R}_+$*

$$\varphi_t(x_0) \in X \text{ if and only if } \varphi_t^{I_l}(\Pi_{I_l}(x_0)) \in \Pi_{I_l}(X) \text{ for all } 1 \le l \le k. \tag{4.18}$$

Proof. By Remark 4.18 it holds $z \in X$ if and only if $\Pi_{I_j}(z) \in \Pi_{I_j}(X)$ for all $1 \leq j \leq k$. Using (4.11) we conclude that $\varphi_t(x_0) \in X$ if and only if

$$\varphi_t^{I_j}(\Pi_{I_j}(x_0)) = \Pi_{I_j}(\varphi_t(x_0)) \in \Pi_{I_j}(X) \text{ for all } 1 \leq j \leq k. \qquad \square$$

In Theorem 4.23, we will show that the concepts of the following definition allow a decomposition based on subsystems.

Definition 4.21. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be Lipschitz continuous and constraint set $X \subset \mathbb{R}^n$. We call a set $M \subset X$*

1. *positively invariant for X, if $M \subset X$ and for all $x_0 \in M$ the flow $\varphi_t(x_0)$ is defined for all $t \in \mathbb{R}_+$ and $\varphi_t(x_0) \in X$ for all $t \in \mathbb{R}_+$.*

2. *maximal invariant in X, if M is the largest positively invariant set for X.*

3. *pointwise/uniformly attractive for X, if for all x_0 in the maximal positively invariant set M_+ for X we have as $t \to \infty$ that $\operatorname{dist}(\varphi_t(x_0), M) \to 0$ pointwise/uniformly in x_0.*

4. *pointwise/uniformly repelling for X, if M is pointwise/uniformly attractive for X for the time reversed differential equation.*

5. *stable for X, if for each open neighbourhood U of M there exists an open neighbourhood V of M such that $\varphi_t(V \cap X) \subset U \cap X$ for all $t \in \mathbb{R}_+$.*

6. *asymptotically stable for X, if M is stable and attractive.*

7. *the stable manifold for X of some set N, if M consists of all points x in the maximum positively invariant set M_+ for which it holds $\operatorname{dist}(\varphi_t(x), N) \to 0$ as $t \to \infty$.*

8. *the unstable manifold for X of N, if M is the stable manifold for X of N for the time reversed system.*

9. *global attractor for X, if M is the smallest compact set that uniformly attracts all bounded subsets of the maximal positively invariant set M_+ for X, i.e. for all $R \in \mathbb{R}_+$ we have $\operatorname{dist}(\varphi_t(M_+ \cap B_R(0)), M) \to 0$ as $t \to \infty$.*

10. *is called the region of attraction for X of some target set X_T with time horizon $T \in \mathbb{R}_+$, if*

$$M = \{x \in X : \varphi_t(x) \in X \text{ for all } t \in [0,T] \text{ and } \varphi_T(x) \in X_\mathrm{T}\}. \tag{4.19}$$

11. *is called the reachable set for X of some initial set $X_0 \subset X$ with time horizon $T \in \mathbb{R}_+$, if*

$$M = \{x \in X : \text{ it exists } x_0 \in X_0 \text{ s.t. } \varphi_t(x_0) \in X \text{ for } t \in [0,T], \varphi_T(x_0) = x\}. \tag{4.20}$$

Remark 4.22. *If a dynamical system on $\mathbb{R}^n$ is not equipped with a constraint set, we can define the objects from Definition 4.21 with respect to the constraint set $X = \mathbb{R}^n$. For stable, attractive, and asymptotically stable sets this, then, coincides with the notion from Definition 2.3.*

We recall the "gluing" from (4.12). For index sets $I_1, \dots, I_k \subset [n]$ and $M_l \in \mathbb{R}^{I_l}$ for $l = 1, \dots, k$ we glue together the sets $M_1, \dots, M_k$ to a set $S(M_1, \dots, M_k) \subset \mathbb{R}^n$ by

$$S(M_{I_1}, \dots, M_{I_k}) = \{x \in \mathbb{R}^n : \Pi_{I_l} x \in M_{I_l} \text{ for } l = 1, \dots, k\}$$

Similarly to Proposition 4.15, we follow the decoupling idea illustrated in Figure 4.1. In the following theorem, we show that this decoupling applies to many of the notions from Definition 4.21 for systems with state constraints.

Theorem 4.23. *Let $I_1, \dots, I_k$ induce a subsystem decomposition. Assume that $X \subset \mathbb{R}^n$ decomposes according to $I_1, \dots, I_k$ via $X_1 := \Pi_{I_1}(X), \dots, X_k := \Pi_{I_k}(X)$. Let $M \subset X$ be closed. Then*

1. *If M is positively invariant for X if and only if for $l = 1, \dots, k$ the set $\Pi_{I_l}(M)$ is positively invariant for X_l under the subsystem (I_l, f_{I_l}). Further, if for $l = 1, \dots, k$ the set M_l is positively invariant for X_l under subsystem (I_l, f_{I_l}) then $S(M_1, \dots, M_k)$ is positively invariant for X under the whole system.*

2. *The maximum positively invariant set M_+ for X is given by*

$$M_+ = S(M_1, \dots, M_k) \tag{4.21}$$

   where for $1 \leq l \leq k$ the set M_l is the maximum positively invariant set for X_l under the subsystem (I_l, f_{I_l}).

3. *Let $X_{\mathrm{T}} \subset X$ and $T \in \mathbb{R}_+$. Assume that X_{T} decomposes according to $I_1, \dots, I_k$ via $X_{\mathrm{T},1}, \dots, X_{\mathrm{T},k}$. Then the region of attraction R_T is given by*

$$R_T = S(R_{T,1}, \dots, R_{T,k}) \tag{4.22}$$

   where $1 \leq l \leq k$ the set $R_{T,l}$ is the region of attraction for X_l with target set $X_{\mathrm{T,l}}$ under the subsystem (I_l, f_{I_l}).

4. *Let $X_{\mathrm{T}} \subset X$ and $T \in \mathbb{R}_+$. Assume that X_0 decomposes according to $I_1, \dots, I_k$ via $X_{0,1}, \dots, X_{0,k}$. Then, the reachable set $\mathcal{S}_T$ for X of X_0 is given by*

$$\mathcal{S}_T = S(\mathcal{S}_{T,1}, \dots, \mathcal{S}_{T,k})) \tag{4.23}$$

   where for $l = 1, \dots, k$ the set $\mathcal{S}_{T,l}$ is the reachable set with initial set $X_{\mathrm{T,l}}$ under the subsystem (I_l, f_{I_l}).

5. *For X compact, the global attractor $\mathcal{A}$ for X is given by*

$$\mathcal{A} = S(\mathcal{A}_1, \dots, \mathcal{A}_k) \tag{4.24}$$

   where for $l = 1, \dots, k$ the set $\mathcal{A}_l$ is the global attractor for X_l under the subsystem (I_l, f_{I_l}).

6. *For $l = 1, \dots, k$ let M_l be compact and pointwise/uniform attractive for X_l under the subsystem (I_l, f_{I_l}) then $S(M_1, \dots, M_k)$ is pointwise/uniform attractive for X under the whole system.*

7. *For $l = 1, \dots, k$ let M_l be compact and repelling for X_l for the subsystem (I_l, f_{I_l}) then $S(M_1, \dots, M_k)$ is repelling for X for the whole system.*

8. *If for $l = 1, \dots, k$ the set M_l is closed and asymptotically stable for X_l under the subsystem (I_l, f_{I_l}) for X_l then $S(M_1, \dots, M_k)$ is asymptotically stable under the whole system.*

9. *Let $N \subset X$ be compact and decompose according to $I_1, \dots, I_k$ via $N_1 := \Pi_{I_1}(N), \dots, N_k := \Pi_{I_k}(N)$. Let V be the stable manifold of N for X. For $1 \leq l \leq k$ let M_l be the stable manifold of N_l for X_l under the subsystem (I_l, f_l). Then $\Pi_{I_l}(V) \subset M_l$. Further, for the stable manifold V of N for X it holds*

$$V = S(M_1, \dots, M_k). \tag{4.25}$$

10. *Point 12. holds correspondingly for the unstable manifold.*

Proof. 1. The first part of the statement follows from (4.18) in Lemma 4.20. The second part of the statement is a direct consequence of the first part because $\Pi_{I_l}(S(M_1, \dots, M_k)) \subset M_l$ for all $1 \leq l \leq k$.

2. By 1. the set $S(M_1, \dots, M_k)$ is positively invariant for X. By maximality we have $M_+ \supset S(M_1, \dots, M_k)$. Again by 1., $\Pi_{I_l}(M_+)$ is positively invariant for $\Pi_{I_l}(X) = X_l$ under the subystem (I_l, f_{I_l}) for all $l = 1, \dots, k$ and hence $\Pi_{I_l}(M_+) \subset M_l$ for all $l = 1, \dots, k$. This shows

$$M_+ \subset S(\Pi_{I_1}(M_+), \dots, \Pi_{I_k}(M_+)) \subset S(M_1, \dots, M_k) \subset M_+. \tag{4.26}$$

3. We have to verify that the elements on the right-hand side of (4.22) are exactly the elements in the region of attraction R_T. This follows from Lemma 4.14 and (4.18) in Lemma 4.20 applied to X for $t \in [0, T]$ and X_T for $t = T$.

4. The reachable set $\mathcal{S}_T$ with initial set X_T is the region of attraction for the time-reversed system with target set X_T. The statement follows from 3.

5. By Theorem 2.7, it holds $\mathcal{A} = M_+ \cap M_-$ where M_- denotes the maximum negatively invariant set, i.e. the MPI set for the reversed time direction. By time reversing, the set M_- decouples analog to M_+. For $l = 1, \dots, k$ let M_l^+ respectively M_l^- denote the MPI respectively maximum negatively invariant set for X_l under the subsystem (I_l, f_{I_l}). We get by 2.

$$\begin{aligned} \mathcal{A} &= M_+ \cap M_- = S(M_1^+, \dots, M_k^+) \cap S(M_1^-, \dots, M_k^-) \\ &= \{x \in X : x_{\mathbf{P}(x_i)} \in M_l^+ \cap M_l^- \text{ for } l = 1, \dots, k\} \\ &= \{x \in X : x_{\mathbf{P}(x_i)} \in \mathcal{A}_l \text{ for } l = 1, \dots, k\} \end{aligned}$$

where we used that for the global attractor $\mathcal{A}_l$ for X_l under the subsystem (I_l, f_{I_l}) we have, by Theorem 2.7, $\mathcal{A}_l = M_l^+ \cap M_l^-$.

6. In case of uniform attraction, note first that a compact set can only be uniformly attractive if the maximum positively invariant set M_+ is bounded, hence compact. Hence, by Theorem 2.7, the global attractor exists and a set is uniformly attractive if and only if it contains the global attractor. The statement follows then from 4. For pointwise attraction, note that by 2., $\Pi_{I_l}(M_+)$ is contained in the maximum positively invariant set for X_l under the subsystem (I_l, f_{I_l}). Therefore, we can conclude the statement similarly as in the proof of Proposition 4.15 3.

7. This follows from the statement for attractive sets reversing by the time direction.

8. For $1 \leq l \leq k$, we choose, by Theorem 2.10, a strict Lyapunov function $V_l : M_l^+ \to [0, \infty)$ with $V_l^{-1}(\{0\}) = M_l$ where M_l^+ is the MPI set for X_l under the subsystem (I_l, f_{I_l}). From hereon, we can follow the proof of Proposition 4.15 4. to conclude the statement.

9. For each $l = 1, \ldots, k$, the map Π_{I_l} is a contraction and it follows that $\Pi_{I_l}(V)$ is contained in the stable manifold M_l of N_l for X_l for the subsystem (I_l, f_{I_l}). Now we show (4.25). First, we claim that

$$S(M_1, \ldots, M_k) \subset V. \tag{4.27}$$

First note that, by 1., the set $S(M_1, \ldots, M_k)$ is positively invariant for X. For the claim (4.27), it remains to show that for all $x \in S(M_1, \ldots, M_k)$ it holds $\operatorname{dist}(\varphi_t(x), N) \to 0$. Assume this is not the case. Then there exists $x \in S(M_1, \ldots, M_k)$, $\varepsilon > 0$ and $(t_m)_{m\in\mathbb{N}} \subset \mathbb{R}$ with $t_m \nearrow \infty$ such that

$$\operatorname{dist}(\varphi_{t_m}(x), N) > \varepsilon \text{ for all } m \in \mathbb{N}. \tag{4.28}$$

By assumption, for all $x \in S(M_1, \ldots, M_k)$ it holds $\operatorname{dist}(\varphi^{I_l}_{t_m}(\Pi_{I_l}(x)), N_l) \to 0$ for all $l = 1, \ldots, k$. Since N is compact, so is $N_l = \Pi_{I_l}(N)$ for all $l = 1, \ldots, k$. Hence, $(\Pi_{I_l}(\varphi_{t_m}(\Pi_{I_l}(x))))_{m\in\mathbb{N}} = (\varphi^{I_l}_{t_m}(\Pi_{I_l}(x)))_{m\in\mathbb{N}} \subset \mathbb{R}^{I_l}$ is a bounded sequence. Therefore, $(\varphi_{t_m}(x))_{m\in\mathbb{N}} \subset \mathbb{R}^n$ is bounded and we can find a convergent subsequence $(\varphi_{t_{m_p}}(x))_{p\in\mathbb{N}}$ and $y \in \mathbb{R}^n$ such that $\varphi_{t_{m_p}}(x) \to y$ as $p \to \infty$. For $l = 1, \ldots, k$ we have

$$\Pi_{I_l}(y) = \lim_{p\to\infty} \Pi_{I_l}(\varphi_{t_{m_p}}(x)) = \lim_{p\to\infty} \varphi^{I_l}_{t_{m_p}}(\Pi_{I_l}(x)) \in \Pi_{I_l}(N) = N_l.$$

This states $y \in N$ which contradicts (4.28). Next, we show

$$V \subset S(M_1, \ldots, M_k). \tag{4.29}$$

This follows from the first part. Namely, it holds $\Pi_{I_l}(V) \subset M_l$ for all $l = 1, \ldots, k$. Therefore we get

$$V \subset S(\Pi_{I_1}(V), \ldots, \Pi_{I_k}(V)) \subset S(M_1, \ldots, M_k).$$

10. This follows from reversing the time direction in 10.

□

We end this section with a result about sparse representations of Lyapunov functions.

Proposition 4.24. *Assume $I_1, \dots, I_k$ induces a subsystem decomposition. Assume that there exists a strict Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}_+$ such that for $V^{-1}(\{0\})$ decomposes according to $I_1, \dots, I_k$. Then, for $l = 1, \dots, k$, there exists a Lyapunov function $V_l : \mathbb{R}^{I_l} \to \mathbb{R}_+$ for the system induced by I_l such that*

$$\tilde{V} := \sum_{r=l}^{k} V_l \circ \Pi_{I_k} \tag{4.30}$$

is a strict Lyapunov function for the whole system with $V^{-1}(\{0\}) = \tilde{V}^{-1}(\{0\})$.

Proof. Let $B := V^{-1}(\{0\})$. For $l = 1, \dots, k$, by Corollary 4.9, we can find a strict Lyapunov function V_{I_l} for the subsystem induced by I_l with $V_{I_l}^{-1}(\{0\}) = \Pi_{I_l}(B)$. The function $\tilde{V}$ defined by (4.30) is then a strict Lyapunov function for the whole system and it holds

$$\begin{aligned} \tilde{V}^{-1}(\{0\}) &= \{x \in \mathbb{R}^n : \Pi_{I_l}(x) \in \Pi_{I_l}(B), l = 1, \dots, k\} \\ &= S(\Pi_{I_1}(B), \dots, \Pi_{I_k}(B)) = B = V^{-1}(\{0\}), \end{aligned} \tag{4.31}$$

where we used that B decomposes according to $I_1, \dots, I_k$, i.e. it holds $B = S(\Pi_{I_1}(B), \dots, \Pi_{I_k}(B))$ by Remark 4.18. □

4.4.2 Application to computational methods

Since our decomposition into subsystems provides an exact decomposition of the whole system into lower dimensional problems, it is applicable to a general class of problems and computational methods.

Decomposition methods are beneficial whenever the dimension of the state space has a huge impact on the computational cost. Due to the curse of dimensionality, this is the typical case for many applications.

We propose the general procedure in Algorithm 1. Algorithm 1 describes how we decompose the problem of solving a given (decomposable!) task for the whole system based on subsystems.

For sparse set approximations, we specify Algorithm 1 to the following Algorithm 2. For a given dynamical system (4.1) let M denote one of the decomposable sets from Theorem 4.23 (i.e. an equilibrium point, periodic orbit, the MPI set, the ROA set, the reachable set, the GA, the stable or the unstable manifold). By Theorem 4.23, the gluing step in Algorithm 1 is specified and incorporated in Algorithm 2.

Next, we show that the decoupling procedure preserves certain convergence properties. We consider the following two (pseudo) metrics on subsets on $\mathbb{R}^n$, one is

Algorithm 1 Decoupling procedure for a decomposable task

1: Input: Dynamics $f : \mathbb{R}^n \to \mathbb{R}^n$ (and constraint set X if needed for the task) and a method S for (approximately) solving the task for an arbitrary dynamical system.
2: Subsystem decomposition: Determine sets $I_1, \dots, I_k \subset [n]$ that induce a subsystem decomposition (for which X decomposes accordingly).
3: Solve for the subsystems: Solve the task on each subsystem using the given method S.
4: Glue: Combine the results for the subsystems to a global solution G.
5: **return** G.

Algorithm 2 Decoupling procedure for computing/approximation the set M for the whole system on $\mathbb{R}^n$ induced by $\dot{x} = f(x)$.

1: Input: Dynamics $f : \mathbb{R}^n \to \mathbb{R}^n$ (and constraint set X if needed) and a method S for computing/approximating M for arbitrary systems given their dynamics.
2: Subsystem decomposition: Determine sets $I_1, \dots, I_k \subset [n]$ that induce a subsystem decomposition (for which X decomposes accordingly).
3: For each $l = 1, \dots, k$: Compute/approximate M for each (state constraint) subsystem I_l using S and denote the solution by M_l.
4: Glue the results together via (4.12), i.e.

$$G := \{x \in X : \Pi_{I_l}(x) \in M_l \text{ for } l = 1, \dots, k\}.$$

5: **return** G.

the Hausdorff distance $\mathrm{dist}(\cdot, \cdot)$ and the other is the Lebesgue measures discrepancy, defined by

$$d_\lambda(K_1, K_2) := \lambda(K_1 \Delta K_2) \tag{4.32}$$

where λ is the Lebesgue measure and $K_1 \Delta K_2 = (K_1 \setminus K_2) \cup (K_2 \setminus K_1)$ is the symmetric difference between the sets K_1 and K_2.

Theorem 4.25 ([Schlosser 2020])**.** *Let $I_1, \dots, I_k$ induce a subsystem decomposition. Assume X is compact and decomposes according to $I_1, \dots, I_k$. Let M denote either the set of equilibria, a periodic orbit, the MPI set, the ROA set, the reachable set, the GA, the stable, or the unstable manifold. Let $M = \{x \in \mathbb{R}^n : \Pi_{I_l}(x) \in M_l, 1 \leq l \leq k\}$ be the decomposition from Theorem 4.23. Let $(M_1^{(m)})_{m\in\mathbb{N}}, \dots, (M_k^{(m)})_{m\in\mathbb{N}}$ be a sequence of closed sets with $M_l^{(m)} \subset \mathbb{R}^{I_l}$ for $1 \leq l \leq k$ and*

$$M^{(m)} := S(M_1^{(m)}, \dots, M_k^{(m)}) = \{x \in \mathbb{R}^n : \Pi_{I_l}(x) \in M_l^{(m)}, 1 \leq l \leq k\}.$$

The following hold:

1. *in case of Hausdorff distance (induced by any norm on $\mathbb{R}^n$): If $M_l^{(m)} \supset M_l$*

for all $1 \le l \le k$ *and* $m \in \mathbb{N}$ *and*

$$\operatorname{dist}(M_l^{(m)}, M_l) \to 0 \qquad \text{as } m \to \infty \qquad \text{for all } 1 \le l \le k, \tag{4.33}$$

then $M^{(m)} \supset M$ *and*

$$\operatorname{dist}(M^{(m)}, M) \to 0\,, \quad \text{as } m \to \infty. \tag{4.34}$$

2. *In case of Lebesgue measure: It holds*

$$\begin{aligned} d_\lambda(M, M^{(m)}) &= \lambda(M \Delta M^{(m)}) \\ &\le \sum_{l=1}^{k} \lambda(M_l \Delta M_l^{(m)})) \lambda(\Pi_{[n]\setminus I_l}(X)). \end{aligned} \tag{4.35}$$

In particular, if $M_l^{(m)}$ *converges to* M_l *with respect to* d_λ *for* $1 \le l \le k$*, then* $M^{(m)}$ *converges to* S *with respect to* d_λ *as* $m \to \infty$*.*

Proof. For the first statement note that the inclusion $M_l^{(m)} \supset M_l$ for all $1 \le l \le k, m \in \mathbb{N}$ implies $M^{(m)} \supset M$. To show the claim (4.34), let us assume it does not hold. Then there exists $\varepsilon > 0$ and an unbounded subsequence $(m_r)_{r\in\mathbb{N}}$ such that

$$\operatorname{dist}(M^{(m_r)}, M) > \varepsilon \tag{4.36}$$

and we find points $x^{(m_r)} \in M_l^{(m_r)}$ with $\operatorname{dist}(x^{(m_r)}, M) > \varepsilon$. From boundedness of $M_1, \dots, M_k$ and the assumption (4.33) it follows that there exists $x \in \mathbb{R}^n$ and a subsequence of $(m_r)_{r\in\mathbb{N}}$ which we will still denote by $(m_r)_{r\in\mathbb{N}}$ such that $x^{(m_r)} \to x$ as $r \to \infty$. By assumption (4.33) there exist $y_l^{m_r} \in M_l$ for $l = 1, \dots, k$ with $\|y_l^{m_r} - \Pi_{I_l}(x^{(m_r)})\| \to 0$ as $r \to \infty$. Hence, also $y_l^{m_r} \to \Pi_{I_l}(x)$ as $r \to \infty$ for $l = 1, \dots, k$. Because $M_1, \dots, M_k$ are closed it follows $\Pi_{I_l}(x) \in S_l$ for $l = 1, \dots, k$ and by Theorem 4.23 we get $x \in M$. In particular, we get

$$\varepsilon < \operatorname{dist}(x^{(m_r)}, M) \le \|x^{(m_r)} - x\| \to 0$$

as $m \to \infty$, which is a contradiction. To conclude (4.35) note that

$$M \Delta M^{(m)} \subset \bigcup_{l=1}^{k} \{x \in X : \Pi_{I_l}(x) \in M_l \Delta M_l^{(m)}\}.$$

Applying the Lebesgue measure to this inclusion gives

$$\begin{aligned} \lambda(M \Delta M^{(m)}) &\le \sum_{l=1}^{k} \lambda\left(\{x \in X : \Pi_{I_l}(x) \in M_l \Delta M_l^{(m)}\}\right) \\ &\le \sum_{l=1}^{k} \lambda(M_l \Delta M_l^{(m)}) \lambda(\Pi_{n\setminus I_l}(X)). \end{aligned}$$

□

4.5 A coordinate-free formulation

The concept of subsystems from Section 4.1 is not intrinsic – subsystems as in Definition 4.1 depend on the coordinates the dynamical system is written in. For example, linear dynamics $\dot{x} = Ax$ where A has only non-zero entries but is diagonalizable do not allow any non-trivial subsystem. If a change of coordinates, that diagonalizes A, preserves the constraint set then for the conjugated system any subset $I \subset [n]$ induces a subsystem. Therefore, the formulation of subsystems is not coordinate-free.

For a coordinate-free formulation, we assume that M is a smooth manifold of dimension n and f a vector field on M such that M is positively invariant. The generated flow is denoted by φ_t for $t \in \mathbb{R}$. We generalize subsystems by means of factor systems.

Definition 4.26. *We call $(\mathcal{N}, P)$ a subsystem if $\mathcal{N}$ is a smooth manifold of dimension less than m and $P : M \to \mathcal{N}$ a surjective smooth submersion map such that there exists a flow $\varphi_t^{\mathcal{N}}$ on $\mathcal{N}$ such that*

$$\varphi_t^{\mathcal{N}} \circ P = P \circ \varphi_t. \tag{4.37}$$

The concept of coordinate indices I is not well defined in the coordinate-free setting. Hence, compared to Definition 4.1, we have to specify $\mathcal{N}$ (which plays the role of $\mathbb{R}^I$) and the map P (which plays the role of Π_I). That means a subsystem (I, f_I) from Definition 4.1 corresponds to $(\mathbb{R}^I, \Pi_I)$ corresponding to Definition 4.26 where $M = \mathbb{R}^n$.

Remark 4.27. *If we drop the condition that $\mathcal{N}$ is of lower dimension than M then a subsystem is $(\mathcal{N}, P)$ is a factor system (see for example [Eisner 2015]). We restrict to lower dimensional systems because factor systems can be very large and counter the idea that the subsystem should be of easier nature. One such example is $M = [0, 1]$ with the trivial dynamics $f = 0$. Using space filling curves $P : [0, 1] \to \mathcal{N}$ where $\mathcal{N}$ is a connected compact manifold (with boundary) of arbitrary dimension induces a subsystem.*

Next, we define subsystem decompositions.

Definition 4.28. *We say that subsystems $(\mathcal{N}_j, P_j)$ for j in some index set J form a subsystem decomposition if the map*

$$P := (P_j)_{j \in J} : M \to \prod_{j \in J} \mathcal{N}_j, \qquad P(x) := (P_j(x))_{j \in J} \tag{4.38}$$

is injective.

Remark 4.29. *Let $I_1, \dots, I_k$ induce a subsystem decomposition according to Definition 4.12 then the subsystems $(\mathbb{R}^{I_1}, \Pi_{I_1}), \dots, (\mathbb{R}^{I_k}, \Pi_{I_k})$ form a subsystem decomposition in the sense of Definition 4.28.*

In order to consider state constraints X, we do so as in Definition 4.17. We say that X decomposes according to a subsystem decomposition $(\mathcal{N}_j, \Pi_j)$ for j in some

index set J if

$$X = \{x \in M : P_j(x) \in P_j(X) \text{ for } j \in J\}. \tag{4.39}$$

Remark 4.30. *If $X \subset M = \mathbb{R}^n$ decomposes according to a subsystem decomposition induced by $I_1, \dots, I_k$ then X, in the sense of Definition 4.17, then X decomposes according to the subsystem decomposition $(\mathbb{R}^{I_1}, \Pi_{I_1}), \dots, (\mathbb{R}^{I_k}, \Pi_{I_k})$ in the sense of (4.39).*

Remark 4.31. *Theorem 4.23 can be formulated as a coordinate-free setting under the assumptions that M is positively invariant, $(\mathcal{N}_j, \Pi_j)_{j\in J}$ form a subsystem decomposition for which X decomposes accordingly. The arguments are analog to the ones used in the proof of Theorem 4.23.*

Example 4.32. *We get back to the introductory example of diagonalizable linear dynamics $\dot{x} = Ax$ with matrix $A =\in \mathbb{R}^{n\times n}$. Further, let us assume there exists n linearly independent eigenvectors $u_1, \dots, u_n$ of A, with eigenvalues $\lambda_1, \dots, \lambda_n$. Let $v_1, \dots, v_n$ be the dual basis to $u_1, \dots, u_n$, i.e. for $i, j = 1, \dots, n$ we have $\langle v_i, u_j\rangle = \delta_{ij}$, for the Kronecker delta δ_{ij}. The flow is given by $\varphi_t(x) = \sum_{i=1}^{n} e^{\lambda_i t}\langle x, v_i\rangle u_i$. For the system on $M = \mathbb{R}^n$ and $X = \overline{B_1(0)}$ we find the subsystem decomposition: For $1 \le i \le n$ let $(\mathcal{N}_i, P_i)$ with $\mathcal{N}_i := \mathbb{R}$ and $P_i : M \to \mathcal{N}_i$ with $P_i(x) := \langle x, v_i\rangle$ and corresponding flow $\varphi_t^{(\mathcal{N}_i)}(r) := e^{\lambda_i t} r$.*

4.6 Limitations

We will discuss two limitations, which are both of different nature. One describes the restrictiveness of our notion of subsystems. As a result of constraining to causal independence among the subsystem, many dynamical systems just do not have subsystems. The second one concerns limitations of the decomposition. Theorem 4.23 shows that many important objects from dynamical systems decouple according to their subsystems but this is not true for *all* objects. We will give the example of the weak attractor.

Definition 4.33 (Weak attractor)**.** *Let $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ be a dynamical system. A compact set $\mathcal{A}_{\mathrm{w}} \subset X$ is called a weak attractor if it is the minimal attracting set, i.e., it is the smallest compact set $\mathcal{A}_{\mathrm{W}}$ that is attracting.*

Example 4.34 (The weak attractor does not decompose)**.** *We consider the following (discrete time) dynamical system. Let $\mathcal{N} := \mathbb{N}_0 \cup \{\beta\}$ denote the one-point compactification of $\mathbb{N}_0$ and let $X := \mathcal{N}^3$. The dynamics are given by*

$$\begin{aligned} x_{k+1} &= x_k + 1 \\ y_{k+1} &= \begin{cases} 0, & x_k = 3^l \text{for some } l \in \mathbb{N}_0 \\ y_k = x_k, & \text{else} \end{cases} \\ z_{k+1} &= \begin{cases} 0, & x_k = 2^l \text{for some } l \in \mathbb{N}_0 \\ z_k = x_k, & \text{else.} \end{cases} \end{aligned} \tag{4.40}$$

It decomposes into the subsystem on x, (x, y) and (x, z). The weak attractors for the three subsystems are given by

$$\mathcal{A}_{\mathrm{w}}^{(1)} = \{\beta_x\},\ \mathcal{A}_{\mathrm{w}}^{(1,2)} = \{(\beta_x, 0), (\beta_x, \beta_y\} \text{ and } \mathcal{A}_{\mathrm{w}}^{(1,3)} = \{(\beta_x, 0), (\beta_x, \beta_z)\}. \tag{4.41}$$

Gluing those sets together, by (4.12), would give the set

$$\begin{aligned} S(\mathcal{A}_{\mathrm{w}}^{(1)}, \mathcal{A}_{\mathrm{w}}^{(1,2)}, \mathcal{A}_{\mathrm{w}}^{(1,2)}) &= \{(x, y, z) \in X : x \in A_w^{(1)}, (x, y) \in A_w^{(1,2)}, (x, z) \in A_w^{(1,3)}\} \\ &= \{(\beta_x, 0, 0), (\beta_x, \beta_y, 0), (\beta_x, 0, \beta_z), (\beta_x, \beta_y, \beta_z)\}. \end{aligned}$$

But as we will see now, the element $(\beta_x, 0, 0)$ is not contained in the weak attractor. A trajectory $(x_k, y_k, z_k)_{k\in\mathbb{N}_0}$ for initial value (x_0, y_0, z_0) has $(\beta_x, 0, 0)$ as an accumulation point if and only if the components y_k and z_k vanish infinitely often and simultaneously. But y_k vanishes only when $k = x_k = 3^{l_1}$ and z_k only when $k = x_k = 2^{l_2}$ for $l_1, l_2 \in \mathbb{N}_0$. For $k > 1$ this would mean we have two different prime factorizations of $k \in \mathbb{N}$, hence this cannot happen.

4.7 Detecting subsystems via the sparsity graph

Subsystems are determined by the dependence on the states in each other's dynamics. This dependence between one state and the dynamics of another can be illustrated by a graph, the so-called sparsity graph of the dynamics f. This will allow us to state a simple algorithm for identifying subsystems of a given dynamical system.

Definition 4.35 (Sparsity graph)**.** *Let $f = (f_1, \ldots, f_n) : \mathbb{R}^n \to \mathbb{R}^n$ be a function. The sparsity graph G_f associated to f is defined by:*

1. *The set of nodes is $\{x_1, \ldots, x_n\}$.*
2. *For $i \neq j$ there pair (x_i, x_j) is an edge if f_j depends explicitly on x_i .*

The sparsity graph G_f of a function f can only tell if there is a causal dependence of a state x_i on a state x_j (indirectly through other states) in the dynamical system induced by $\dot{x} = f(x)$. The sparsity graph does not include quantitative information. This is because the sparsity graph forgets about the precise interaction of nodes. In Figure 4.2 we consider a sparsity graph of the function $f : \mathbb{R}^5 \to \mathbb{R}^5$ from Example 4.3.

From the sparsity graph in Figure 4.2 we infer that the underlying vector field f is of the following form

$$f = (f_1(x_1), f_2(x_1, x_2), f_3(x_1, x_3, x_4), f_4(x_1, x_4), f_5(x_1, x_5)).$$

This is consistent with (3.13) but does not tell us more about the explicit form of functions $f_1, \ldots, f_5$. Figure 4.3 provides a graphical illustration of all subsystems for dynamics corresponding to the sparsity graph from Figure 4.2.

Figure 4.3 indicates certain important features related to subsystems. At first, x_1 seems like a source of information – there are only outgoing edges – and is

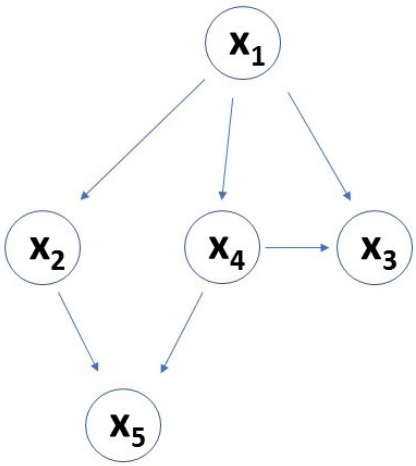

Figure 4.2: The sparsity graph of the function from (3.13)

contained in any non-trivial subsystem. On the other hand, x_3 and x_5 have the character of a sink of information – there is no outgoing edge. Further, we see that any of the subsystems displayed in Figure 4.3 consists of all nodes for which there exists a directed path in G_f to a distinguished node: The blue resp. yellow resp. red resp. green resp. grey subsystem consists of all nodes for which there exists a directed path to x_1 resp. x_2 resp. x_4 resp. x_5 resp. x_3. The related objects will be defined and given a name in the following definition.

Definition 4.36 (Predecessor, leaf, Past)**.** *Let G be a directed graph with nodes $\{x_1, \ldots, x_n\}$.*

1. *A node x_i is a called predecessor of node x_j if $x_i = x_j$ or there is a directed path from x_i to x_j. In that case, x_j is called a successor of x_i.*
2. *A node x_i is called a leaf if it does not have a successor (i.e., all nodes connected to x_i are its predecessors).*
3. *A node x_i is called a root if it does not have a predecessor (i.e., all other nodes connected to x_i are its successors).*
4. *The set of all indices of predecessors of x_i is called the past of x_i and denoted by $P(x_i) \subset [n]$.*

From the Definition 4.36 we conclude the following lemma.

Lemma 4.37. *Let G_f be the sparsity graph for the function $f : \mathbb{R}^n \to \mathbb{R}^n$. The predecessor and successor relation is transitive: If x_k is a predecessor/successor of x_j and x_j is a predecessor/successor of x_i then x_k is a predecessor/successor of x_i. In particular, for a predecessor x_k of x_i we have $P(x_k) \subset P(x_i)$.*

Proof. The first statement follows just by patching paths together, namely the one from x_k to x_j and the one from x_j to x_i. □

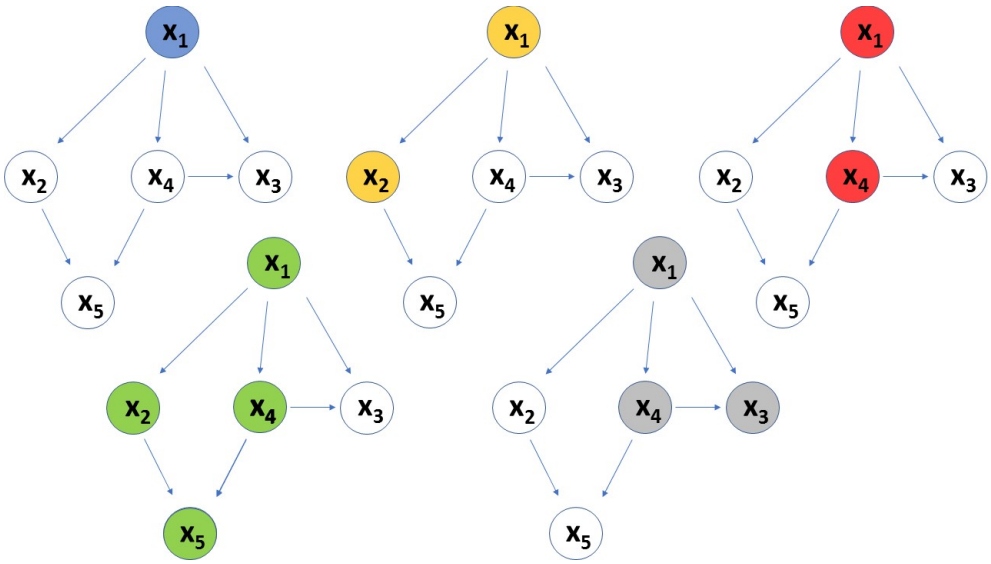

Figure 4.3: Sparsity graph G_f for f from (3.13). Nodes colored with the same color represent a subsystem. The remaining subsystems, which are not presented, are induced by $(\emptyset, f_\emptyset), (\{1,2,4\}, f_{\{1,2,4\}}), (\{1,2,3,4\}, f_{\{1,2,4\}})$ and $(\{1,2,3,4,5\}, f)$. Any of these can be written as a union of the subsystems colored in blue, yellow, red, green, and grey respectively.

In Definition 4.36, we defined the past of a node x_i as all the nodes the dynamics of x_i (indirectly) depend on. This similarity to subsystems is manifested in the following corollary, which states that the pasts of nodes are the building blocks for subsystems.

Corollary 4.38. *For a state x_i, the past $P(x_i)$ induces the smallest subsystem containing x_i. In particular, if I induces a subsystem then I contains the pasts $P(x_i)$ of all its nodes x_i for all $I \in I$.*

Proof. We have already noted, right before Definition 4.35 of the sparsity graph, that a subsystem containing x_i has to contain all states $(x_j)_{j \in P(x_i)}$ corresponding to its past $P(x_i)$. Let us prove that $(P(x_i), f_{P(x_i)})$ indeed induces a subsystem, and hence the minimal one containing x_i. Assume this is not the case. That means that $f_{P(x_i)}$ does not only depend on the variables $x_{P(x_i)}$. Hence, there exists $j \in P(x_i)$ and $k \in [n] \setminus P(x_i)$ such that f_j depends on x_k. That means that x_k is a predecessor of x_j in the sparsity graph of f, in particular, x_k is in the past of x_j and hence has to be contained in the subsystem by the first part of the proof. □

This gives rise to the following characterization of subsystems.

Proposition 4.39 (Characterization of subsystems)**.** *(I, f_I) induces a subsystem if and only if I contains all the pasts of its nodes.*

Proof. By Corollary 4.38, it follows that a subsystem contains all the pasts of its nodes. Now, suppose $I \subset [n]$ is such that all pasts $P(x_i) \subset I$ for all $i \in I$. Let us assume (I, f_I) does not induce a subsystem. That means it exists a component

$i \in I$ such that f_i also depends on a node x_j with $j \notin I$. But f_i depending on x_j means that j is a predecessor of i, in particular, $j \in P(x_i)$ which contradicts our assumption. □

Remark 4.40. *Referring to Remark 4.5, we formulate some of the concepts from Definition 4.36 by topological concepts. Let $\mathcal{T}_f := \{I \subset [n] : I \text{ induces a subsystem}\}$ be the subsystem topology on $[n]$. Then, by Corollary 4.38, the past $P(x_i)$ is the smallest open neighborhood of $i \in [n]$. Furthermore, an index $i \in [n]$ is a leaf respectively root if and only if $\{i\}$ is open respectively closed in that topology.*

4.7.1 The sparsity graph of systems with state constraints

In Section 4.4.1 we had to include the sparsity structure of the constraint set X into the subsystems as well. Therefore, the sparsity graph for state constrained system has to depend on X as well as on the dynamics f.

Definition 4.41 (The sparsity graph $G_{f,\mathcal{J}}$)**.** *Let G_f be the sparsity graph of f and $\mathcal{J}$ a family of index sets $J_1, \dots, J_N \subset \{1, \dots, n\}$ for which X decomposes accordingly. The graph $G_{f,\mathcal{J}}$ has the nodes $x_1, \dots, x_n$ and (x_i, x_l) is an edge of $G_{f,\mathcal{J}}$ if (x_i, x_l) is an edge in G_f or if there exists $1 \leq r \leq N$ such that $i, l \in J_r$.*

Such as the sparsity graph G_f determines subsystems, and so does the sparsity graph $G_{f,\mathcal{J}}$ determine state constrained subsystems. We now investigate how to deduce a subsystem decomposition, for systems with and without state constraints, from the sparsity graph G_f respectively $G_{f,\mathcal{J}}$.

4.8 Constructing a subsystem decomposition

In this Section, we treat the question of finding subsystems of a given dynamical system. That is an important task because a good choice of subsystems is essential for a lower dimensional or more differentiated treatment of the dynamical system.

Our decomposition will be based solely only the sparsity graph G_f respectively $G_{f,\mathcal{J}}$ from the previous section. We observe that their strongly connected components cannot be further decomposed into smaller subsystems. Thus, contracting the strongly connected components allows for a simple way of finding a subsystem decomposition.

4.8.1 Strongly connected components of the sparsity graph

Definition 4.42 (Strongly connected component; [Cormen 2022])**.** *In a directed graph G we call a set of nodes C*

1. *strongly connected, if for each two nodes $x, y \in C$ there exists a path from x to y and a path from y to x.*
2. *strongly connected component, if C is strongly connected and maximal with this property, i.e. C can not be extended to a larger set that is still strongly connected.*

From a sparsity perspective, strongly connected sets C of a sparsity graph describe ensembles of states that are fully influencing each other.

Remark 4.43. *Strongly connected sets can be formulated via the notation from the previous section, namely: A set C is strongly connected if and only if we have $P(x) \supset C$ for all $x \in C$. In particular, any subsystem either contains a strongly connected component or is disjoint from it.*

From a subsystem perspective, Remark 4.43 motivates not to distinguish between the nodes in a strongly connected component. This is why we will introduce the condensation graph in the following subsection.

4.8.2 The condensation graph

In this section, we describe a method of merging multiple nodes in a graph to a single node. This is called node contraction.

Definition 4.44 (Node contraction; [Diestel 2017])**.** *Let $G = (V, E)$ be a (directed) graph with nodes V and edges E. Let $W_1, \dots, W_k \subset V$ be pairwise disjoint sets of nodes and fix elements $\omega_i \in W_i$ for $i = 1, \dots, k$. The graph $G' = (V', E')$ obtained by contracting $W_1, \dots, W_k$ is given by the nodes $V' := \{\omega_1, \dots, \omega_k\} \cup V \setminus \bigcup_{i=1}^{k} W_i$ and (v_1, v_2) for $v_1, v_2 \in V'$ is an edge if one of the following holds:*

1. *$v_1, v_2 \in V$ and $(v_1, v_2) \in E$*
2. *$v_2 \in V$ and $v_1 = \omega_i$ for some $1 \leq i \leq k$ and there exists $w \in W_i$ such that $(w_i, v) \in E$*
3. *$v_1 \in V$ and $v_2 = \omega_i$ for some $1 \leq i \leq k$ and there exists $w \in W_i$ with $(v, w_i) \in E$.*

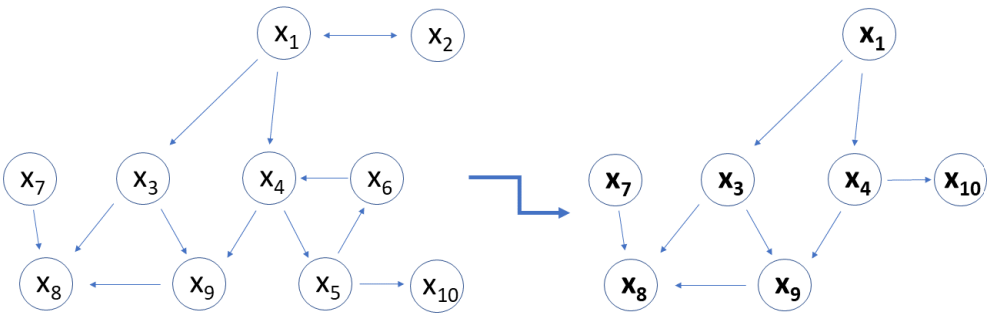

Figure 4.4: Example of a sparsity graph (left) and its condensation graph (right). The strongly connected components are $C_1 = \{x_4, x_5, x_6\}, C_2 = \{x_1, x_2\}$. The graph on the right arises from the sparsity graph (left) by contracting each of the strongly connected components C_1 and C_2.

The objects that we want to contract are the strongly connected components. After contracting them, the graph has a much simpler structure – it is a forest, i.e. acyclic.

Lemma 4.45 ([Tarjan 1972])**.** *Let G be a directed graph and G' be the condensation graph of G obtained from contracting each of the connected components to a single node. Then G' is acyclic.*

The forest structure in the condensation graph is useful because we can directly infer important objects. In our case, these are the leaves. We will use the leaves of the condensation graph of the sparsity graph to construct subsystem decomposition via their past. For that result, we provide the following Lemma.

Lemma 4.46. *Any directed graph without cycles has at least one leaf. Furthermore, for directed graphs without cycles, any node is a predecessor of a leave.*

Proof. Let W be a maximal path in the graph, i.e. a path that cannot be extended in G. Let x be the last node in W. We claim that x is a leaf. If x is not a leaf then there exists an edge (x, y) in G for some node y. By maximality of W we cannot add y to W, which means the edge (x, y) has been used before in W. This means that W has visited x before, i.e. there is a part of W that connects x to itself, i.e. a cycle – contradiction. For the remaining statement let y be an arbitrary node. We can choose a longest path containing this node which has to end in a leaf x, hence y is a predecessor of x. □

4.8.3 Construction of a subsystem decomposition

We use the condensation graph from the previous section to give a simple algorithm for finding a subsystem decomposition.

Proposition 4.47. *Let G_f be the sparsity graph of f and let $v_1, \dots, v_m$ be leaves of its condensation graph, where the connected components $W_1, \dots, W_k \subset \{x_1, \dots, x_n\}$ have been contracted to $\omega_1, \dots, \omega_k$. The pasts $P(v_1), \dots, P(v_m)$ induce a subsystem decomposition (where for $1 \leq i \leq k$ each w_i is re-placed by the nodes W_i it represents).*

Proof. By Corollary 4.38, the pasts $P(v_1), \dots, P(v_m)$ induce subsystems. It remains to show that $P(v_1), \dots, P(v_m)$ induces a subsystem decomposition, i.e. that

$$[n] = \bigcup_{l=1}^{m} \mathbf{P}(v_l). \tag{4.42}$$

Let $1 \leq i \leq n$. If $x_i \in W_l$ for some $1 \leq l \leq k$ then $i \in \mathbf{P}(v_l)$ because of the strong connectivity of W_l, $v_l \in W_l$, and the definition of the condensed sparsity graph. If $x_i \notin W_1 \cup \dots \cup W_k$ then x_i is a node in the condensed sparsity graph. By Lemma 4.46 it follows that x_i is a predecessor for a leaf. Hence also in this case $i \in \bigcup_{l=1}^{m} \mathbf{P}(v_l)$. □

For systems with state constraints, the construction is analog where instead of the sparsity graph G_f we use $G_{f,\mathcal{J}}$ for a family $\mathcal{J}$ of index sets for which X decomposes accordingly. In Algorithm 4.8.3 we state the procedure.

Algorithm 3 Computation of a subsystem decomposition

1: Input: The sparsity graph G_f of f and a family $\mathcal{J}$ of index sets for which X decomposes accordingly.
2: Compute the graph $G_{f,\mathcal{J}}$ from Definition 4.41
3: Compute the strongly connected components $W_1, \dots, W_k$ of $G_{f,\mathcal{J}}$
4: Build the condensed graph $\mathcal{G}_{f,\mathcal{J}}$ of $G_{f,\mathcal{J}}$: Contract each of the strongly connected components W_i to w_i (see Definition 4.44).
5: Find the leaves $v_1, \dots, v_m$ in $\mathcal{G}_{f,\mathcal{J}}$.
6: Define $I_1, \dots, I_m$: Set $I_l := \mathbf{P}(v_l)$ for $l = 1, \dots, m$ (where for $1 \leq i \leq k$ each w_i is re-placed by the nodes W_i it represents).
7: **return** $I_1, \dots, I_m$ inducing a subsystem decomposition.

Remark 4.48. *The complexity of the Algorithm 4.8.3 is at most linear in n, the number of states, and the number of edges m in the graph $G_{f,\mathcal{J}}$. This is because the strongly connected components can be found by a depth first search in $\mathcal{O}(n + m)$ [Cormen 2022] Section 22.5). All the remaining tasks, such as building the sparsity graph, the condensation graph, and finding the leaves, are simple and can also be performed in $\mathcal{O}(n + m)$.*

The largest appearing state space dimension based on the subsystem decomposition obtained from Algorithm 4.8.3 is given by the largest number of predecessors of a node in the graph $G_{f,\mathcal{J}}$. This number is given by

$$\omega := \max_l |\mathbf{P}(v_l)|. \tag{4.43}$$

where $|\mathbf{P}(v_l)|$ is the number of predecessors of the node v_j in the sparsity graph $G_{f,\mathcal{J}}$. We formulate this statement precisely in the following theorem.

Theorem 4.49. *Let $\mathcal{J}$ be a family of index sets for which X decomposes accordingly. Algorithm 4.8.3 gives a subsystem decomposition induced by $I_1, \dots, I_k$ for which X decomposes accordingly such that the largest subsystem contains ω nodes, i.e.*

$$\max_{l=1,\dots,m} |I_l| = \omega,$$

where ω is given by (4.43).

Proof. We will show that

1. $I_1, \dots, I_m$ induce a subsystem decomposition for which X factors accordingly, and

2. the largest number of variables in each of these subsystems is at most ω.

That $I_1, \dots, I_m$ induces a subsystem decomposition can be verified similarly to Proposition 4.47 because the sparsity graph G_f is a subgraph of $G_{f,\mathcal{J}}$. To show (1) it remains to check that X decomposes accordingly. Let us write $J = \{J_1, \dots, J_N\}$ for some $N \in \mathbb{N}$. For $1 \leq r \leq N$ and $i, j \in J_r$, the two nodes x_i and x_j form a circle in the graph $G_{f,\mathcal{J}}$. Therefore, the nodes $(x_j)_{j \in J_r}$ get contracted in the condensation graph of $G_{f,\mathcal{J}}$ for all $r = 1, \dots, N$ and each of the sets I_l can be written as

$$I_l = \bigcup_{l \in Z_k} J_l,$$

where the sets $Z_1, \dots, Z_m$ form a partition of $\{1, \dots, N\}$. We claim that X decomposes according to $I_1, \dots, I_m$ via

$$X = \{x \in \mathbb{R}^n : \Pi_{I_k}(x) \in Y_k, k = 1, \dots, m\} \tag{4.44}$$

where for $k = 1, \dots, m$ the sets Y_k are given by

$$Y_k := \{y \in \mathbb{R}^{I_k} : x \in \mathbb{R}^n, \Pi_{J_l}(x) \in X_l, l \in Z_k\}.$$

We conclude (4.44) via

$$\begin{aligned} X &= \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in X_l, l = 1, \dots, N\} = \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in X_l, l \in \bigcup_{k=1}^{m} Z_k\} \\ &= \{x \in \mathbb{R}^n : \Pi_{I_k}(x) \in Y_k, k = 1, \dots, m\}, \end{aligned}$$

because X decomposes according to $J_1, \dots, J_N$. To verify our second claim 2., recall that each set $I_k \subset [n]$ corresponds to the past of a node v_k in the condensation graph of $G_{f,\mathcal{J}}$. Because the condensation graph contracts only the strongly connected components, I_k is nothing else than the past of the node x_{i_k} with $x_{i_k} = v_k$ in the sparsity graph $G_{f,\mathcal{J}}$. □

Corollary 4.50. *If there are no state constraints, then the Algorithm 4.8.3 is optimal in the following sense: For any other subsystem decomposition, induced by index sets $\tilde{I}_1, \dots, \tilde{I}_l \subset [n]$, it holds*

$$\max_{k=1,\dots,l} |\tilde{I}_k| \geq \omega.$$

Proof. If there are no state constraints we do not need the additional family of index sets $\mathcal{J}$ in Algorithm 4.8.3. Let $I_1, \dots, I_m$ be the output of Algorithm 4.8.3. As in the proof of Theorem 4.49, we see that for each $1 \leq k \leq m$ there exists x_k with $I_k = P(x_k)$. By Corollary 4.38, the set $P(x_k)$ induces the smallest subsystem containing x_k. Now, let $\tilde{I}_1, \dots, \tilde{I}_l$ induce another subsystem decomposition. Then, for each $1 \leq k \leq m$ it holds $x_k \in \tilde{I}_l$ for some l. And thus, by minimality of $P(x_k)$, it holds $I_k = P(x_k) \subset \tilde{I}_l$. It follows with Theorem 4.49

$$\max_{k=1,\dots,l} |\tilde{I}_k| \geq \max_{k=1,\dots,m} |I_k| = \omega.$$ □

As input for Algorithm 4.8.3 we used a family $\mathcal{J}$ of index sets for which X decomposes accordingly. In applications the sparsity of the dynamics often comes with the same sparse structure for X, i.e. X decomposes according to any subsystem decomposition. But this might not always be the case. Finding the best decomposition of X, i.e. a decomposition that minimizes ω from (4.43), is a delicate task. We address a suboptimal procedure based on factoring X into a cartesian product motivated from Remark 4.19. We say that index sets $J_1, \dots, J_N \subset \{1, \dots, n\}$ induce a factorization of X if $J_1, \dots, J_N$ is a partition of $\{1, \dots, n\}$ and

$$X = \{x \in \mathbb{R}^n : \Pi_{J_l}(x) \in P_{J_l}(X) \text{ for } l = 1, \dots, N\}.$$

As noted in Remark 4.19, X decomposes according to the family $\mathcal{J}$ of index sets $J_1, \dots, J_N$ if $J_1, \dots, J_N$ induce a factorization. We propose to use a minimal factorization $\mathcal{J}$ as input to Algorithm 4.8.3 in order to obtain a fine subsystem decomposition. The existence of a minimal factorization is assured by the following lemma from [Schlosser 2020].

Lemma 4.51. *There exists a minimal factorization for X; that is there exist index sets $J_1, \dots, J_N$ that induce a factorization of X, such that for any other factorization induced by $\tilde{J}_1, \dots, \tilde{J}_{\tilde{N}}$ we have for all $l = 1, \dots, \tilde{N}$ that $\tilde{J}_l = \bigcup_{k: J_k \subset \tilde{J}_l} J_k$.*

4.9 Extensions to other systems

In this section, we show that the notion of subsystems and corresponding decompositions of the whole system extend naturally to time dependent systems, differential inclusions, time-delay, stochastic, hybrid, and control systems. The idea is the same: For a system with solution map φ, a set of indices $I \subset [n]$ induces a subsystem if $\Pi_I \circ \varphi(x_0)$ coincides with the solution for the subsystem to the initial value $\Pi_I(x_0)$.

We get started with the case of time dependent systems.

Remark 4.52 (Time dependent systems)**.** *The extension to time-dependent systems $\dot{x}(t) = f(t, x(t)), x(0) = x_0$ is immediate. We treat t as an additional state $x_{n+1}(t) := t$, i.e. $\dot{x}_{n+1} = 1$, $x_{n+1}(0) = 0$ and reduce to the autonomous case. Because in $\dot{x}_{N+1} = 1$ none of the states $x_1, \dots, x_n$ appear, the node x_{n+1} is a root.*

4.9.1 Subsystems for time-delay systems

We consider the following time-delay system

$$\dot{x}(t) = f(x(t), x(t - \tau_1), \dots, x(t - \tau_l)), \; x(-s) = x_0(s) \text{ for } s \in [0, \tau_l] \tag{4.45}$$

for $t \in \mathbb{R}_+$, a continuous initial function $x_0 : [-\tau_l, 0]$, delays $0 < \tau_1 \leq \tau_2 \leq \ldots \leq \tau_l$ and a Lipschitz continuous map

$$f : \mathbb{R}^n \times (\mathbb{R}^n)^l \to \mathbb{R}^n, f(x, y^1, \dots, y^l) \in \mathbb{R}^n. \tag{4.46}$$

Under the above assumptions existence and uniqueness hold for (4.45), see for instance [Fridman 2014].

We follow a familiar structure from Section 4.1, by first defining what a system means.

Definition 4.53. *A set $I \subset [n]$ induces a subsystem for (4.45) if $f_I = \Pi_I \circ f$ only depends on the variables $x_I, y_I^1, \dots, y_I^l$. A family of sets $I_1, \dots, I_k \subset [n]$ induces a subsystem decomposition if $I_1, \dots, I_k$ induce subsystems and $\bigcup_{j=1}^{k} I_j = [n]$.*

For a set I that induces a subsystem, we want to define a time-delay system on $\mathbb{R}^n$ whose solutions correspond to the projection $\Pi_I \circ x(\cdot)$ for solutions $x(\cdot)$ of (4.45). Therefore, we view f_I as a map in the following way

Let (I, f_I) induce a subsystem for (4.45). We view f_I as a map

$$f_I : \mathbb{R}^I \times \mathbb{R}^I \times \cdots \times \mathbb{R}^I \to \mathbb{R}^I, f_I(x_I, y_I^l, \dots, y_I^l) := f(x, y^1, \dots, y^l) \qquad (4.47)$$

for $x_I, y_I^1, \dots, y_I^l \in \mathbb{R}^I$ and $x, y^1, \dots, y^l \in \mathbb{R}^I$ such that $\Pi_I(x) = x_I, \Pi_I(y^1) = y_I^1, \dots, \Pi_I(y^l) = y_I^l$.

Proposition 4.54. *Let $I \subset [n]$ induce a subsystem for the time-delay system (4.45). Then for any solution $x(\cdot)$ of (4.45) the function $w := \Pi_I \circ x(\cdot)$ solves*

$$\dot{w}(t) = f_I(w(t), w(t-\tau_1), \dots, w(t-\tau_l)), w(-s) = \Pi_I(x(s)) \text{ for } s \in [0, \tau_l]. \quad (4.48)$$

Proof. We just check (4.48) using the assumption on f respectively f_I. We have

$$\begin{aligned} \dot{w}(t) &= \Pi_I(\dot{x}(t)) = \Pi_I\left(f(x(t), x(t-\tau_1), \dots, x(t-\tau_l))\right) \\ &= f_I((x_i(t))_{i\in I}, (x_i(t-\tau_1))_{i\in I}, \dots, (x_i(t-\tau_l))_{i\in I}) \\ &= f_I(w(t), w(t-\tau_1), \dots, w(t-\tau_l)). \end{aligned}$$

The initial condition is satisfied by definition of w. □

To obtain a decomposition result as in Theorem 4.23 it is important that solutions of the subsystems give rise to solutions of the whole system.

Proposition 4.55. *Let $I_1, \dots, I_k$ induce a subsystem decomposition. Let $x_0(\cdot) : [-\tau_l, 0] \to \mathbb{R}^n$ be continuous and $x^1(\cdot), \dots, x^k(\cdot)$ be solutions of (4.48) with corresponding initial values $x^j(\cdot) = \Pi_{I_j} \circ x_0(\cdot)$. Then the solution $x(\cdot)$ of (4.45) is the unique function with $x^j(\cdot) = \Pi_{I_j} \circ x(\cdot)$ for all $j = 1, \dots, k$.*

Proof. Proposition (4.54) shows that $x^j(\cdot) = \Pi_{I_j} \circ x(\cdot)$ is satisfied for all $j = 1, \dots, k$. That $I_1, \dots, I_k$ induce a subsystem decomposition implies that a function $z(\cdot)$ with $\Pi_{I_j} \circ z(\cdot) = x^j(\cdot)$ for all $j = 1, \dots, k$, is unique. □

Remark 4.56. *Based on Propositions 4.54 and 4.55 we can apply the techniques from the previous sections, including the decomposition from Section 4.4.1 also to time-delay systems.*

Remark 4.57 (Sparsity graph)**.** *For time-delay systems, the sparsity graph of f from (4.45) has nodes $x_1, \dots, x_n$ and an edge from x_i to x_j if $\dot{x}_j$ depends on at least one of the values $x_i(t), x_i(t-\tau_1), \dots, x_i(t-\tau_l)$, i.e. f_j depends on at least one of the values $x_i(t), x_i(t-\tau_1), \dots, x_i(t-\tau_l)$.*

4.9.2 Subsystems for stochastic differential equations

Consider a stochastic differential equation of the following form

$$dX_t = f(t, X_t)\, dt + \sigma(t, X_t)\, dB_t \tag{4.49}$$

with given initial random variable X_0, $(B_t)_{t\in\mathbb{R}_+}$ a m-dimensional Brownian motion and $f : \mathbb{R}^n \to \mathbb{R}^n$, $\sigma = (\sigma_{i,j})_{\substack{i=1,\dots,n\\ j=1,\dots,m}} : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ being measurable with

$$\|f(x)\|_2 + \sum_{i,j=1}^{n} |\sigma_{i,j}(x)| \le C(1 + \|x\|_2) \text{ for all } x \in \mathbb{R}^n \tag{4.50}$$

for a constant $C \in \mathbb{R}$ and

$$\|f(x) - f(y)\|_2 + \sum_{i,j=1}^{n} |\sigma_{ij}(x) - \sigma_{ij}(y)| \le D\, \|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n \tag{4.51}$$

for some constant D. We refer to [Øksendal 2003] for our notation of stochastic ordinary differential equation, as well as for the existence and uniqueness of a solution $(X_t)_{t\in\mathbb{R}_+}$ of (4.49) under the above conditions, see for instance [Øksendal 2003, Section 5].

Definition 4.58. *A set $I \subset [n]$ induces a subsystem for the stochastic differential equation (4.49) if $f_I = \Pi_I \circ f$ and $\sigma_I := \Pi_I \circ \sigma := (\sigma_{i,j})_{i\in I, 1\le j\le m}$ depend only on the variables x_I. A family of sets $I_1, \dots, I_k \subset [n]$ induces a subsystem decomposition if $I_1, \dots, I_k$ induce subsystems and $\bigcup_{j=1}^{k} I_j = [n]$.*

As usual, for $I \subset [n]$ that induces a subsystem, we treat f_I respectively σ_I as a function on $\mathbb{R}^I$ and we get a corresponding stochastic differential equation (4.52) induced by f_I and σ_I.

Proposition 4.59. *Let I induce a subsystem for (4.49) and $(X_t)_{t\ge 0}$ be the solution of (4.49). For $t \in \mathbb{R}_+$ let $W_t := \Pi_I \circ X_t$. Then the stochastic process $(W_t)_{t\in\mathbb{R}_+}$ is the unique solution of*

$$dY_t = f_I(Y_t)\, dt + \sigma(Y_t)\, dB_t,\ Y_0 = \Pi_I(X_0). \tag{4.52}$$

Proof. That $(X_t)_{t\in\mathbb{R}_+}$ solves (4.49) means that for $0 \le t \le s$ it holds, [Øksendal 2003],

$$X_s - X_t = \int_t^s f(X_r)\, dr + \int_t^s \sigma(X_r)\, dB_r. \tag{4.53}$$

Based on (4.53) we verify the corresponding integral equation for $(W_t)_{t\in\mathbb{R}_+}$, i.e. let $0 \le t \le s$ then

$$\begin{aligned} W_s - W_t &= \Pi_I (X_s - X_t) = \Pi_I \left(\int_t^s f(X_r)\, dr + \int_t^s \sigma(X_r)\, dB_r \right) \\ &= \int_t^s \Pi_I(f(X_r))\, dr + \int_t^s (\Pi_I \circ \sigma)(X_r)\, dB_r \\ &= \int_t^s f_I(r, W_r)\, dr + \int_t^s \sigma_I(r, X_r)\, dB_r. \end{aligned}$$

This proves the claim. □

For a subsystem decomposition, we can recover the solution from the subsystem solution.

Proposition 4.60. *Let $I_1, \dots, I_k \subset [n]$ induce a subsystem decomposition for (4.49) and let $(W_t^1)_{t\in\mathbb{R}_+}, \dots, (W_t^k)_{t\in\mathbb{R}_+}$ be the solutions of the corresponding subsystem equation (4.52). Then $(X_t)_{t\in\mathbb{R}_+}$ is the unique stochastic process with $W_t^j = \Pi_{I_j} \circ X_t$ for all $t \in \mathbb{R}_+$ and all $j = 1, \dots, k$.*

Proof. By Proposition 4.59 it holds $W_t^j = \Pi_{I_j} \circ X_t$ for all $t \in \mathbb{R}_+$ and $j = 1, \dots, k$. Because $I_1, \dots, I_k$ induce a subsystem decomposition, a function Y with $\Pi_{I_j}(Y) = W_t^j$ for all $j = 1, \dots, k$ is uniquely determined for each $t \in \mathbb{R}_+$. □

Remark 4.61 (Sparsity graph)**.** *For stochastic differential equations (4.49) the sparsity graph has nodes $x_1, \dots, x_n$ and there is an edge from x_i to x_j if f_j depends on x_i or σ_{jl} depends on x_i for some $1 \le l \le n$.*

4.9.3 Subsystems for hybrid systems

For hybrid systems, we follow the notation and solution concept from [Goedel 2012]. We consider the equation

$$\begin{cases} \dot{x} \in F(x), & x \in C \\ x_+ \in G(x), & x \in D \end{cases} \tag{4.54}$$

for disjoint continuity and jump sets $C, D \subset \mathbb{R}^n$, set valued maps $F, G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, and initial value $x(0) = x_0 \in \mathbb{R}^n$.

Definition 4.62. *A set $I \subset [n]$ induces a subsystem for (4.54) if $F_I = \Pi_I \circ F$ and $G_I = \Pi_I \circ G$ only depend on the variables x_I.*

We get the following expected proposition.

Proposition 4.63. *Let $I \subset [n]$ induce a subsystem for (4.54) assume that C and D decompose according to $I, [n] \setminus I$ (see Definition 4.17). Let ϕ be a solution of*

(4.54). Then $\Pi_I \circ \phi$ is a solution for the hybrid subsystem

$$\begin{cases} \dot{y} \in F_I(y), & y \in \Pi_I(C) \\ y_+ \in G_I(y), & y \in \Pi_I(D) \end{cases}$$

with initial value $y^j(0) = \Pi_I(x_0)$.

Proof. The statement follows basically from Lemma 4.14 by separating the dynamics in its continuous part on C and its discrete part on D. Note that Lemma 4.14, such as most other results in this chapter, holds true for the discrete time case as well. □

Remark 4.64. *Because the maps F and G are set-valued we do not have uniqueness of solutions, therefore a reconstruction of a global solution from solutions of the subsystem might not be possible. The reason is that in different subsystems, incompatible solutions can be selected.*

Remark 4.65 (Sparsity graph)**.** *For hybrid systems (4.54) we build the sparsity graph based on decompositions of C and D as for state constrained systems in Definition 4.41. Therefore, assume that C decomposes according to $J_1, \dots, J_N \subset [n]$ and D decomposes according to $U_1, \dots, U_M \subset [n]$. The sparsity graph for the hybrid system (4.54) has the nodes $x_1, \dots, x_n$ and there is an edge from x_i to x_j if $\Pi_j \circ F$ depends on x_i or $\Pi_j \circ G$ depends on x_i or $i, j \in J_l$ for some $1 \leq l \leq N$ or $i, j \in U_r$ for some $1 \leq r \leq M$.*

4.9.4 Subsystems for control systems

In this section, we define subsystems for control systems of the form

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = x_0 \in \mathbb{R}^n \tag{4.55}$$

for a continuous vector field $f = (f_1, \dots, f_n) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, which is Lipschitz continuous in x, and controls $u = (u_1, \dots, u_m) : \mathbb{R}_+ \to \mathbb{R}^m$. We call $x(\cdot), u(\cdot)$ a solution of (4.55) if the pair satisfied (4.55).

For control systems, we have to include the control u into the notion of a subsystem.

Definition 4.66. *A set $I \times K \subset [n] \times [m]$ induces a subsystem for (4.55) if f_I depends only on x_I and u_K. A family of set $I_1 \times K_1, \dots, I_N \times K_N \subset [n] \times [m]$ induces a subsystem decomposition if $\bigcup_{l=1}^{n} I_l = [n]$ and $K_l \cap K_r = \emptyset$ for all $l \neq r$.*

If $I \times K$ induces a subsystem for (4.55) then we can treat f_I as a control vector field on $\mathbb{R}^I$ with control inputs $(u_k)_{k \in K}$.

Proposition 4.67. *Let $I \times K$ induce a subsystem for (4.55) and $x(\cdot), u(\cdot)$ be a solution of (4.55). Then $\Pi_I \circ x(\cdot), \Pi_K \circ u(\cdot)$ is a solution for the control subsystem*

$$\dot{y}(t) = f_I(y(t), v_K(t)), \quad y(0) = \Pi_I(x_0). \tag{4.56}$$

Proof. We compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{dt}}\Pi_I(x(t)) &= \Pi_I\left(\dot{x}(t)\right) = \Pi_I(f(x(t),u(t))) = f_I(x_I(t),u_K(t)) \\ &= f_I(\Pi_I(x(t)),\Pi_K(u(t)))\end{aligned}$$

with initial value $\Pi_I(x(0)) = \Pi_I(x_0)$. □

For subsystem decompositions, we can reconstruct the system from its subsystem solutions.

Proposition 4.68. *Let $I_1 \times K_1, \ldots, I_N \times K_N$ be a subsystem decomposition for 4.55 and $(x^1(\cdot),u^1(\cdot)),\ldots,(x^N(\cdot),u^N(\cdot))$ be solutions of the corresponding control subsystem equation (4.56). Then there is a unique function $x(\cdot)$ and a control function $u(\cdot)$ that solve (4.55) and satisfy $\Pi_{I_l}\circ x(\cdot) = x^l(\cdot)$ and $\Pi_{K_l}\circ u(\cdot) = u^l(\cdot)$ for all $1 \le l \le N$.*

Proof. Uniqueness of the function $x(\cdot)$ follows from $\bigcup_{l=1}^{N} I_l = [n]$. To show existence we define the control $u = (u_1,\ldots,u_m)$ by

$$u_i(t) = u_i^l(t) \tag{4.57}$$

for $t \in \mathbb{R}_+$ and l such that $i \in K_l$. If there is no $1 \le l \le N$ with $i \in K_l$ (i.e. f is independent of the control input u_i) we set $u_i(t) = 0$. The condition from Definition 4.66 that the sets $K_1,\ldots,K_N$ are pairwise independent implies that u is well defined. Next we claim that $x(\cdot),u(\cdot)$ solves (4.55). Let $i \in [n]$ and $1 \le l \le N$ with $i \in I_l$ we have

$$\dot{x}_i(t) = f_i(x_I(t),u_K(t)) = (f_{I_l}(x_I(t),u_K(t)))_i = (f(x(t),u(t)))_i = f_i(x(t),u(t))$$

by our choice of $u(\cdot)$ in (4.57) and the fact that I_l induces a subsystem □

Remark 4.69 (Sparsity graph)**.** *The sparsity graph for the control system (4.55) has the nodes $x_1,\ldots,x_n$ and $u_1,\ldots,u_m$, an edge from x_i to x_j if f_j depends on x_i and an bi-directed edge between x_i and u_j if f_i depends on u_j.*

Remark 4.70. *It is possible to merge the notions of subsystems for the different systems from this section. Thus, we can treat stochastic control systems or time-delay hybrid systems, etc.*

Chapter 5

Linear programming problems for global attractors

In this chapter, we present the linear programming problem representations of global attractors from [Schlosser 2021, Schlosser 2022a]. We concentrate on continuous time dynamical systems. Treating discrete time systems is possible in a similar fashion and treated in [Schlosser 2021].

We consider dynamical system induced by differential equations

$$\dot{x} = f(x),\ x(0) = x_0 \in \mathbb{R}^n \tag{5.1}$$

for a Lipschitz continuous vector field f on $\mathbb{R}^n$. We denote the corresponding semi-flow of solutions to (5.1) by φ. The dynamical system on $\mathbb{R}^n$ is equipped with a constraint set X. That is, we are only concerned about those trajectories which stay in X for all positive times. This leads to the following notion of the maximum positively invariant set and global attractor for X.

Definition 5.1 (Maximum positively invariant and global attractor for X)**.** *For a dynamical system $(\mathbb{R}^n, (\varphi_t)_{t\in\mathbb{R}_+})$ and a compact set $X \subset \mathbb{R}^n$ we call the set*

$$M_+ = \{x_0 \in X : \varphi_t(x_0) \in X \text{ for all } t \in \mathbb{R}_+\}$$

the maximum positively invariant (MPI) set for X. The global attractor (GA) for X is the smallest compact set $\mathcal{A} \subset X$ that uniformly attracts M_+, i.e.

$$\lim_{t\to\infty} \operatorname{dist}(\varphi_t(M_+), \mathcal{A}) = 0,$$

where dist *denotes the Hausdorff distance with respect to the euclidean norm in $\mathbb{R}^n$.*

In the following, we present two infinite dimensional linear programming problems that represent the global attractor for X.

5.1 An occupation measure approach

This section is based on the text [Schlosser 2021]. The method strongly builds on the work [Korda 2014] and combines it with a characterization of the global attractor as the maximum invariant (i.e. positively invariant in forward and backward time direction) set.

We follow an established line of reasoning via so-called occupation measures, allowing for a linear but infinite dimensional representation of the dynamical system via measures [Rubio 1975, Vinter 1978, Korda 2014, Henrion 2013, Lasserre 2008].

5.1.1 A linear programming problem for global attractors

In [Korda 2014] a linear programming problem (LP) for approximating the maximum positively invariant set M_+ was presented. Theorem 2.7 shows that the global attractor is characterized as the largest set that is positively invariant forward and backward in time. Hence, combining the LP for the maximum positively invariant set (in forward time) with the same LP in reversed time direction gives the global attractor. To motivate the resulting LP we follow the construction from [Korda 2014]. For an initial measure $\mu_0 \in M(X)$ and a discount factor $\beta > 0$, let μ be the discounted occupation measure. That is, for a measurable set $C \subset X$ the value $\mu(C)$ is defined by

$$\mu(C) := \int_X \int_0^\infty e^{-\beta t} \mathbb{I}_C(\varphi_t(x)) \, dt \, d\mu_0(x). \tag{5.2}$$

Then μ is a well defined measure on X that measures the discounted average time spent in C (one can think of sampling initial conditions at random from the probability distribution given by μ_0). From (5.2) and integration by parts, we get the following relation for all $v \in \mathcal{C}^1(\mathbb{R}^n)$

$$\begin{aligned}
\int_X \nabla v \cdot f \, d\mu &= \int_X \int_0^\infty e^{-\beta t} \nabla v(\varphi_t(x)) f(\varphi_t(x)) \, dt \, d\mu_0(x) \\
&= \int_X \int_0^\infty e^{-\beta t} \frac{\partial}{\partial t} v(\varphi_t(x)) \, dt \, d\mu_0(x) \\
&= -\int_X v \, d\mu_0 + \beta \int_X \int_0^\infty e^{-\beta t} v(\varphi_t(x)) \, d\mu_0(x) = \beta \int_X v \, d\mu - \int_X v \, d\mu_0.
\end{aligned}$$

We call the equation

$$\int_X \nabla v \cdot f \, d\mu = \beta \int_X v \, d\mu - \int_X v \, d\mu_0 \qquad \text{for all } v \in \mathcal{C}^1(\mathbb{R}^n) \tag{5.3}$$

the (continuous-time) Liouville's equation. The Liouville equation can also be seen as a direct application of basic semigroup theory: For $v \in \mathcal{C}^1(\mathbb{R}^n)$, $Av := \nabla v \cdot f$ is the action of the generator of the Koopman semigroup $T_t v := v \circ \varphi_t$. The adjoint semigroup is given by the push forward $P_t \rho(C) := \rho(\varphi_t^{-1}(C))$ for measurable sets $C \subset X$. Hence, we can read the occupation measure μ (5.2) as the Laplace transform $\mu = \int_0^\infty e^{-\beta t} P_t \mu_0 \, dt$ of μ_0. Thus, μ solves the resolvent equation $(\beta \mathrm{Id} - A^*)^{-1} \mu_0 = \mu$ ([Engel 2006] Theorem 1.10), which is nothing else than the Liouville equation (5.3).

Liouville equation and the global attractor We recall first that, by Theorem 2.7, the global attractor is given by the intersection of the maximum positively

invariant in forward and backward time direction. Therefore, we focus on the maximum positively invariant set M_+ first.

In [Korda 2014] it was shown that for any pair (μ, μ_0) satisfying the Liouville equation (5.3), it holds $\text{supp}(\mu_0) \subset M_+$. Thus, to characterize the set M_+, we could maximize the support μ_0 among pairs of measures (μ, μ_0) that satisfy the Liouville equation. This results in the set M_+, see [Korda 2014]. Indeed, we can always choose a measure μ_0 with support $\text{supp}(\mu_0) = M_+$ and μ its corresponding occupation measure (5.2). However, maximization of the support of a measure is computationally challenging. In order to circumvent this obstacle, we follow the same strategy as in [Korda 2014]. Instead of maximizing the support we will maximize the mass $\mu_0(X)$ under the condition that μ_0 is dominated by the Lebesgue measure λ which is equivalent to $\mu_0 + \hat{\mu}_0 = \lambda\big|_X$ for a $\hat{\mu}_0 \in M(X)$. That gives the following linear programming problem

$$\begin{aligned} \lambda(M_+) = \quad & \sup_{\mu_0, \hat{\mu}_0, \mu} && \mu_0(X) \\ & \text{s.t.} && \mu_0, \hat{\mu}_0, \mu \in M(X) \\ & && \int_X \beta v - \nabla v \cdot f \, d\mu = \int_X v \, d\mu_0 \quad \forall v \in \mathcal{C}^1(\mathbb{R}^n) \\ & && \mu_0 + \hat{\mu}_0 \; = \lambda\big|_X \end{aligned} \tag{5.4}$$

That the optimal value of the above LP is indeed the Lebesgue volume of M_+ is shown in [Korda 2014]. To turn to the global attractor we add the invariance conditions for the reversed time direction. That is imposed by adding the Liouville equation (5.3) induced by the vector field $-f$. This yields the following LP

$$\begin{aligned} \sup_{\mu_0, \hat{\mu}_0, \mu_+, \mu_-} \quad & \mu_0(X) \\ \text{s.t.} \quad & \mu_0, \hat{\mu}_0, \mu_+, \mu_- \in M(X) \\ & \int_X \beta v^1 - \nabla v^1 \cdot f \, d\mu_+ = \int_X v^1 \, d\mu_0 \quad \forall v^1 \in \mathcal{C}^1(\mathbb{R}^n) \\ & \int_X \beta v^2 + \nabla v^2 \cdot f \, d\mu_- = \int_X v^2 \, d\mu_0 \quad \forall v^2 \in \mathcal{C}^1(\mathbb{R}^n) \\ & \mu_0 + \hat{\mu}_0 \quad = \lambda\big|_X \end{aligned} \tag{5.5}$$

From the construction of (5.5), we immediately derive the following bound.

Proposition 5.2. *$\lambda(\mathcal{A})$ is a lower bound for (5.5).*

Proof. For a measurable set C let $\mu_0(C) := \lambda(\mathcal{A} \cap C)$ and $\hat{\mu}_0(C) := \lambda((X \setminus \mathcal{A}) \cap C)$. Then $\mu_0, \hat{\mu}_0 \in M(X)$ and $\mu_0 + \hat{\mu}_0 = \lambda\big|_X$. Let μ_+ and μ_- be the corresponding occupation measures with discount factor $\beta > 0$ defined by (5.2) forward and backward in time. So $(\mu_0, \hat{\mu}_0, \mu_+, \mu_-)$ is feasible and has objective value $\mu_0(X) = \lambda(\mathcal{A} \cap X) = \lambda(\mathcal{A}) = \lambda\big|_X(\mathcal{A})$. □

5.1.2 The dual LP

The dual LP of (5.5) is given by

$$\begin{array}{ll} \inf\limits_{w,v^1,v^2} & \int\limits_X w \, d\lambda \\ \text{s.t.} & (w, v_1, v_2) \in \mathcal{C}(\mathbb{R}^n) \times \mathcal{C}^1(\mathbb{R}^n) \times \mathcal{C}^1(\mathbb{R}^n) \\ & -v^1 - v^2 + w \geq \mathbf{1} \\ & w \geq 0 \\ & \beta v^1 - \nabla v^1 \cdot f \geq 0 \\ & \beta v^2 + \nabla v^2 \cdot f \geq 0 \end{array} \tag{5.6}$$

The following important lemma provides geometric insight about the feasible points of the LP (5.6) and is from [Korda 2014, Lemma 3].

Lemma 5.3. *Let (w, v^1, v^2) be feasible for the dual LP (5.6) and M_+ respectively M_- denote the maximum positively respectively negatively invariant set. It holds $v^1 \geq 0$ on M_+, $v^2 \geq 0$ on M_- and hence $w \geq 1$ on $\mathcal{A}$.*

The above lemma indicates what the optimal solution of the dual LP should look like, i.e., that $w = \mathbb{I}_{\mathcal{A}}$. Since $\mathbb{I}_{\mathcal{A}}$ is not continuous on X (if $\mathcal{A}$ is not a connected component of X), which implies that the solution to (5.6) is not attained; however, a minimizing sequence exists and its infimum is equal to the supremum in the primal problem (5.5), i.e., there is no duality gap. This is formalized in the following crucial result.

Theorem 5.4. *For all $\beta > 0$ there is no duality gap and the optimal value of (5.5) and (5.6) is given by $\lambda(\mathcal{A})$. The infimum in the dual program is not attained unless $\mathcal{A} = X$. For a feasible solution (v^1, v^2, w) of the dual problem we have $\mathcal{A} \subset w^{-1}([1, \infty))$.*

This statement shows that the global attractor can be approximated by super-level sets of functions w obtained from the dual problem and as the feasible solutions approach the optimum this approximation gets tight.

Before turning to a more explicit construction, we give a short argument for a proof of Theorem 5.4 by using [Korda 2014]. Because the global attractor is given by the intersection of the sets M_+ and M_-, we can first apply the linear programming problem from [Korda 2014] to find M_+ and in the next step apply the same problem to the dynamical system with reversed time direction to find the part of M_- laying in M_+. Since it was shown in [Korda 2014] that there is no duality gap in each step, there will be no duality gap for the LPs (5.5) and (5.6). The arguments used in [Korda 2014] use infinite-dimensional LP theory while we will give a constructive proof. In the case of regularizing discount factor $\beta > \mathrm{Lip}(f)$ we will construct a sequence of feasible solutions $(v_m^1, v_m^2, w_m)_{m \in \mathbb{N}}$ such that $w_m \to \mathbb{I}_{\mathcal{A}}$ in $\mathrm{L}^1(X, \lambda)$; in particular this is a minimizing sequence by Lemma 5.2.

Before proving Theorem 5.4, let us state the following result [Lee 2003, Theorem 2.29].

Proposition 5.5. *For each closed set $C \subset \mathbb{R}^n$ there exists a bounded function $p \in \mathcal{C}^\infty(\mathbb{R}^n)$ such that $p^{-1}(\{0\}) = C$ and $p(x) \geq 0$ for all $x \in \mathbb{R}^n$.*

Proof. of Theorem 5.4 We start with the easy part, namely that the superlevel sets $w^{-1}([1,\infty))$ give outer approximations of the global attractor $\mathcal{A}$. By Lemma 5.3 we have $w(x) \geq 1$ on $\mathcal{A}$. For the rest we only cover the case $\beta > \mathrm{Lip}(f)$ and additionally assume that the derivative of f is locally Lipschitz continuous as well. We need these technical assumptions in order to guarantee that our construction gives a sufficient regular function (namely $\mathcal{C}^1$). Note that the general case is covered by applying the arguments from [Korda 2014] twice, as mentioned above. The idea is that we will define suitable functions v_m^1, v_m^2 that satisfy the equation $\beta v_m^1 - \nabla v_m^1 f \geq 0$ and $\beta v^2 + \nabla v^2 \cdot f \geq 0$ respectively and build a minimizing sequence based on those functions. Through the functions v^1 and v^2 we want to recognize which points leave X. Without loss of generality we can assume f being globally Lipschitz with globally Lipschitz derivative, because, under our assumptions, one can always modify f outside of X to obtain a globally Lipschitz function on $\mathbb{R}^n$ coinciding with f on X. We denote the corresponding flow also by φ. Note that this flow exists for all times $t \in \mathbb{R}$ and initial values $x_0 \in \mathbb{R}^n$ and we have $x \notin M_+$ if and only if $\varphi_t(x) \notin X$ for some $t \in \mathbb{R}_+$. Let us choose $p \in \mathcal{C}^\infty(\mathbb{R}^n)$ bounded such that $p^{-1}(\{0\}) = X$ and $p > 0$ everywhere else. For $x \in \mathbb{R}^n$ define

$$v^1(x) := -\int_0^\infty e^{-\beta t} p(\varphi_t(x))\, dt. \tag{5.7}$$

Then $v^1(x) < 0$ if and only if there exists a time $t \in \mathbb{R}_+$ for which we have $\varphi_t(x) \notin X$. In particular $v^1 < 0$ on $X \setminus M_+$. We will see that v^1 satisfies $\beta v^1 - \nabla v^1 \cdot f \geq 0$ on X. We have $\|\partial_x \varphi_t(x)\| \leq \overline{M} e^{\mathrm{Lip}(f)t}$ for some $\overline{M} > 0$ and all $t \in \mathbb{R}_+$. Thanks to $\beta > \mathrm{Lip}(f)$ we can interchange integration and differentiation and get for all $x \in \mathbb{R}^n$

$$Dv^1(x) = -\int_0^\infty e^{-\beta t} \partial_x \left(p(\varphi_t(x))\right)\, dt = -\int_0^\infty e^{-\beta t} Dp(\varphi_t(x)) \partial_x \varphi_t(x)\, dt.$$

Further

$$\begin{aligned} \beta v^1(x) &= -\beta \int_0^\infty e^{-\beta t} p(\varphi_t(x))\, dt \overset{p.i.}{=} -p(x) + \int_0^\infty e^{-\beta t} Dp(\varphi_t(x)) f(\overline{\varphi}_t(x))\, dt \\ &= -p(x) - \int_0^\infty e^{-\beta t} Dp(\varphi_t(x)) \partial_x \varphi_t(x) f(x)\, dt = -p(x) + Dv^1(x) f(x), \end{aligned}$$

where we have used in the third line the following relation

$$\partial_{x_0} \varphi_t(x_0) \cdot f(x_0) = \partial_t \varphi_t(x_0) = f(\varphi_t(x_0)).$$

This is the only part where we needed f to have a locally Lipschitz continuous derivative. Since p is vanishing on X we have $\beta v^1 - \nabla v^1 \cdot f = 0$ on X and $v^1(x) < 0$ for $x \notin M_+$. Proceeding similarly backward in time we find $v^2 \in \mathcal{C}^1(\mathbb{R}^n)$ that satisfies $\beta v^2 + \nabla v^2 \cdot f \geq 0$ on X and $v^2(x) < 0$ for $x \notin M_-$. In particular the triple $(v_m^1, v_m^2, w_m) := (m \cdot v^1, m \cdot v^2, \max\{0, 1 + m \cdot v^1 + m \cdot v^2\})$ is feasible and as $m \to \infty$ we have $w_m \searrow \mathbb{I}_{\mathcal{A}}$. It follows from the monotone convergence theorem

that $\int_X w_m \, d\lambda \to \int_X \mathbb{I}_{\mathcal{A}} \, d\lambda = \lambda(\mathcal{A})$. By Proposition 5.2 we know that $\lambda(\mathcal{A})$ is a lower bound for the primal problem while the above shows that $\lambda(\mathcal{A})$ is an upper bound of the dual problem. Weak duality, Theorem 2.21, gives that $\lambda(\mathcal{A})$ is the optimal value for both the primal and dual LP. □

Note that for $\beta > \mathrm{Lip}(f)$ we have constructed a feasible solution (v^1, v^2, w) such that $w^{-1}([1,\infty)) = w^{-1}(\{1\}) = \mathcal{A}$.

5.1.3 Solving the LPs via semidefinite programming

We approach the problem of computationally solving the infinite dimensional LP (5.6) via techniques from polynomial optimization. In order to do so, we assume more algebraic structure on the dynamical system.

Assumption 5.6. *The vector field f is polynomial and X is a compact basic semi-algebraic set, that is, there exist polynomials $p_1, \dots, p_j \in \mathbb{R}[x_1, \dots, x_n]$ such that $X = \{x \in \mathbb{R}^n : p_i(x) \geq 0 \text{ for } i = 1, \dots, j\}$. Further we assume that one of the p_i is given by $p_i(x) = R_X^2 - \|x\|_2^2$ for some large enough $R_X \in \mathbb{R}$.*

We will only state the procedure for the dual LP (5.6) because this will provide guaranteed outer approximations of the global attractor, while this is not the case for the primal problem. In order to solve the infinite dimensional problem we first replace the space of continuous functions with the space of polynomials; this is justified by the Stone-Weierstraß theorem. Then we truncate the degree of the polynomials to get tightenings of the dual problem in form of finite dimensional semidefinite programs (SDPs). Where the SDPs arise from the application of Putinar's Positivstellensatz to reformulate positivity as a sum-of-squares constraint. The corresponding tightenings truncated at degree $k \in \mathbb{N}$ for the problem in continuous time read as

$$\begin{aligned} d_k := \quad & \inf_{v^1, v^2, w, \{q_i\}, \{t_i\}, \{r_i\}, \{s_i\}} \quad \mathbf{w}'\mathbf{l} \\ \text{s.t.} \quad & -v^1 - v^2 + w - 1 = q_0 + \sum_{i=1}^{j} q_i p_i \\ & w(x) = t_0 + \sum_{i=1}^{j} t_i p_i \\ & \beta v^1 - \nabla v^1 \cdot f = r_0 + \sum_{i=1}^{j} r_i p_i \\ & \beta v^2 + \nabla v^2 \cdot f = s_0 + \sum_{i=1}^{j} s_i p_i \end{aligned} \tag{5.8}$$

where $\mathbf{w}'$ is the vector of coefficients of the polynomial w and $\mathbf{l}$ is the vector of the moments of the Lebesgue measure over X (i.e., $\mathbf{l}_\alpha = \int_X x^\alpha \, d\lambda(x)$, $\alpha \in \mathbb{N}^n$, $\sum_i \alpha_i \leq k$), both indexed in the same basis of $\mathbb{R}[x]_k$; hence $\mathbf{w}'\mathbf{l} = \int_X w(x) \, d\lambda(x)$. The decision variables v^1, v^2, w are polynomials in $\mathbb{R}[x]_k$ whereas $q_0, \dots, q_j$, $r_0, \dots, r_j$, $s_0, \dots, s_j$, $t_0, \dots, t_j$ are sum-of-squares of polynomials with degrees such that q_0, t_0, r_0, s_0, $q_i p_i$, $t_i p_i$, $r_i p_i$, $s_i p_i$ are all in $\mathbb{R}[x]_k$ for all $i = 1, \dots, j$. These sum-of-squares optimization

problems translate directly to convex SDPs (see, e.g., [Lasserre 2009, Parrilo 2000]) with high-level modeling software available (e.g., Yalmip [Löfberg 2004], Gloptipoly [Henrion 2009]). That this procedure leads to a convergent hierarchy of optimization problems is stated in the following theorem.

Theorem 5.7. *For all $k \in \mathbb{N}$ we have $d_k \geq d_{k+1}$ and $d_k \to \lambda(\mathcal{A})$ as $k \to \infty$. Further, let (w_k, v_k^1, v_k^2) be optimal for the SDP (5.8) then for*

$$A_k := \{x \in X \mid \min\{v_k^1(x), v_k^2(x)\} \geq 0\}, \tag{5.9}$$

it holds $A_k \supset \mathcal{A}$ and

$$\lim_{k\to\infty} \lambda(A_k \setminus \mathcal{A}) = 0. \tag{5.10}$$

Proof. The inequality $d_k \geq d_{k+1}$ follows immediately since the set of feasible elements is monotonically increasing with k. To prove the convergence of d_k to $\lambda(\mathcal{A})$ as $k \to \infty$, note first that any triple (v^1, v^2, w) that is feasible for the problem (5.8) is feasible for the original dual LPs (5.6). Hence, we have $d_k \geq \lambda(\mathcal{A})$ for all $k \in \mathbb{N}$. Let $\varepsilon > 0$ and (v^1, v^2, w) be feasible for the dual LP (5.6). Then $(v^1 + \varepsilon, v^2 + \varepsilon, w + \varepsilon)$ is strictly feasible and by compactness and the Stone-Weierstraß theorem we can find polynomials $\nu^1, \nu^2, \omega \in \mathbb{R}[x_1, \ldots, x_n]$ such that $\max\{\|v^1 - \nu^1\|_\infty, \|\nabla v^1 - \nabla \nu^1\|_\infty\}, \max\{\|v^2 - \nu^2\|_\infty, \|\nabla v^2 - \nabla \nu^2\|_\infty\} < \frac{\beta}{1+\beta}\varepsilon$ in the continuous time case and $\|w - \omega\|_\infty < \varepsilon$. By the triangle inequality we see that (ν^1, ν^2, ω) is also strictly feasible with objective value

$$\int_X \omega \, d\lambda \leq \int_X w \, d\lambda + \varepsilon\lambda(X).$$

Since $\varepsilon > 0$ was arbitrary we see that the optimal value is unchanged when restricting the decision variable to polynomials. From Putinar's Positivstellensatz [Putinar 1993] it follows now the convergence of d_k to $\lambda(\mathcal{A})$ as $k \to \infty$. For the remaining claim note first that A_k always contains $\mathcal{A}$ by Lemma 5.3. To verify (5.10) we use again Lemma 5.3 and see $w_k \geq 1$ on A_k. From feasibility we have $w \geq 0$ on X and it follows

$$\begin{aligned} \lambda(A_k \setminus \mathcal{A}) &= \lambda(A_k) - \lambda(\mathcal{A}) = \int_{A_k} 1 \, dx - \lambda(\mathcal{A}) \leq \int_{A_k} w_k(x) \, dx - \lambda(\mathcal{A}) \\ &\leq \int_X w(x) \, dx = d_k - \lambda(\mathcal{A}), \end{aligned}$$

and this converges to zero, as $k \to \infty$, by the already proven part of the statement. □

The asymptotic convergence to the global attractor was proven for all parameters $\beta > 0$. But when computing an outer approximation the choice of this parameter has a quantitative effect. Note that the limit case $\beta = 0$, in the primal problem, corresponds to the problem of finding an invariant measure, while large values of β respectively α give high discounting, i.e. the occupation measure takes short time evolution of the system more into account.

Numerical examples Additionally to the Lorenz system shown in Figure 3.11, we present two more numerical examples, one of which has also a strange attractor and the other one has a stable limit cycle. The system with strange attractor has discrete time, and is given by the Hénon map, scaled such that the attractor is inside the unit box,

$$x_{m+1} = \frac{2}{3}(1 + y_m) - 2.1x_m^2, \quad y_{m+1} = 0.45x_m. \tag{5.11}$$

As mentioned, the treatment of discrete time systems is similar and we refer to [Schlosser 2021] for the details. The second example is the Van–der-Pol oscillator

$$\dot{x} = 2y, \quad \dot{y} = -0.8x - 10(x^2 - 0.21)y. \tag{5.12}$$

The numerical approximations of the attractors were generated by simulation of very long trajectories, discarding the initial portions. The SDP problems were solved using MOSEK. The figures Fig. 5.1 and Fig. 5.2 show the outer approximations of the global attractors given by A_k from Theorem 5.7 arising from the tightening SDPs with degree bound $k = 8$ and $k = 10$ for the Hénon map and $k = 12$ for the Van der Pol oscillator.

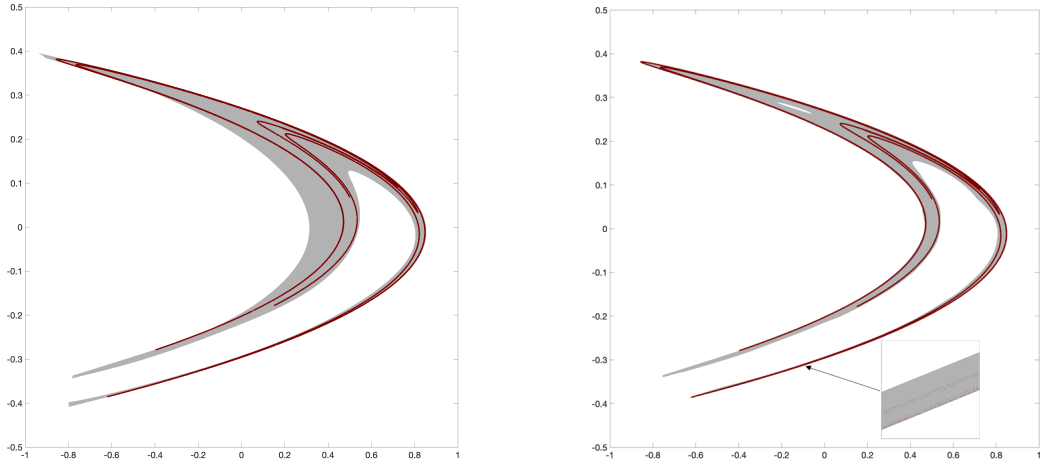

Figure 5.1: Outer approximations for the Hénon attractor. Left: $\beta = -\log(0.05)$ and degree 8 polynomials. Right: The figure shows the intersection of the approximations obtained by degree 6,8 and 10 polynomials.

Although we know that the optimal values $d_k = \int w_k$ of the tightening SDPs decrease monotonically to $\lambda(\mathcal{A})$ it is not guaranteed that the sets A_k are monotonically shrinking towards the global attractor. And it is not to be expected that the sequence of sets A_k shows such a monotone decay. But that on the other hand allows us to get better results by combining lower degree approximations with higher degree ones. In addition to that, the freedom of the choice of the scalar parameter β respectively α allows further refinement. The right pane of Figure 5.1 shows the intersection of the outer approximations obtained by the tightening SDPs up for

degrees $4, 6$ and 8 and a scalar grid of the parameter α. For the Hénon map the outer approximation is given by the grey colored area.

We expect that more complicated topological structures, such as holes, require higher degree polynomials to be identified by our approach. Since we only gave a guaranteed convergence in terms of Lebesgue measure discrepancy, we may not have full control of all topological properties of the outer approximations of the global attractor[1]. The Van der Pol oscillator is an example where the global attractor is given by an asymptotically stable limit cycle. Hence, the solutions to the SDP tightenings have to detect the limit cycle and hence this is connected to the task of finding holes which we have also seen for the Hénon map. Here it is important to choose the set X a bit more carefully. For the left pane in Figure (5.3) we chose $X = \{(x_1, x_2) \in \mathbb{R}^2 : 0.4 \leq \|(x_1, x_2)\|_2 \leq 2\}$ so that the limit cycle is included in X but the initial value $(0, 0)$ corresponding to the trivial solution $x(t) = y(t) = 0$ for all t is not included. The difference is that if $(0, 0)$ is in X, then the limit cycle and its whole interior are the global attractor. This is detected by our approach as shown in the right pane of Figure 5.3. The reason why in that case the attractor is the much larger set is that the interior of the limit cycle is the unstable manifold of the equilibrium point $(0, 0)$, hence contained in the global attractor.

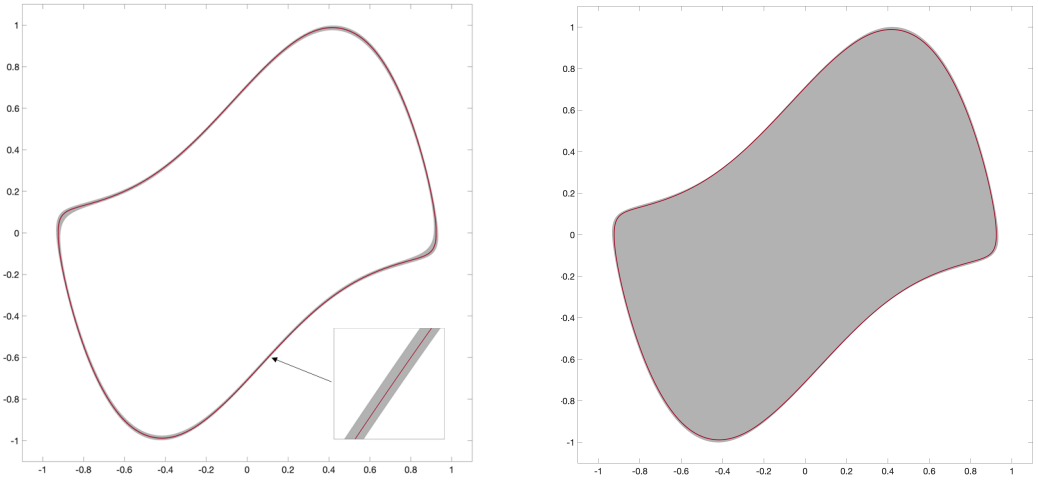

Figure 5.2: Outer approximations of the global attractor for the Van der Pol oscillator. Left: approximation with $X = \{x \mid 0.4 \leq \|x\|_2 \leq 2\}$ (fixed point $(0, 0)$ not included), degree 12 polynomials and $\beta = 0.05$. Right: approximation with $X = \{x \mid \|x\|_2 \leq 2\}$ (fixed point included), degree 12 polynomials and $\beta = 2$.

The figures in Fig. 5.2 show that the global attractor is detected well. But our numerical examples also showed that one has to be careful with numerical issues because, in the case of the Van der Pol oscillator for lower degree polynomials, the

[1] Getting a guaranteed asymptotic control for the topological properties of the attractor would require convergence in the Hausdorff metric. Proving such convergence remains a challenging and so far elusive task for the moment-sum-of-squares approach, here as well as in previous works (e.g., [Korda 2014, Henrion 2013]).

graph of $\min\{v^1, v^2\}$ is very flat around the global attractor which leads to possible round off issues when depicting the superlevel set $X_k = \min\{v^1, v^2\}^{-1}([0, \infty))$. In such situations, in order to obtain provable outer approximation, more careful post-processing of the solutions to the SDP is required, e.g., using the methods of [Henrion 2013, Korda 2014].

5.2 A convex almost Lyapunov function approach

The work that we present in this section combines two existing approaches for approximating global attractors. One is the approach from [Schlosser 2021] from the previous Section 5.1. The other is [Jones 2021a] based on Lyapunov theory. In [Jones 2021a], the authors propose the following optimization problem

$$\begin{aligned} \inf_{J} \quad & \lambda\left(J^{-1}([0,1])\right) \\ \text{s.t.} \quad & J \in \mathcal{C}^1(\mathbb{R}^n) \\ & J(x) \geq 0 \qquad\qquad \text{on } X \\ & \nabla J \cdot f \leq -(J-1) \qquad \text{on } X \\ & \emptyset \neq J^{-1}\left([0,1]\right) \subset \mathring{X}. \end{aligned} \tag{5.13}$$

This approach provides converging supersets $J^{-1}([0,1])$ of the global attractor which have the desirable property of being positively invariant. However, the method from [Jones 2021a] has the disadvantage that the optimization problem (5.13) is not convex and not computationally tractable without the use of heuristics. Incorporating such heuristics came at the price of losing guaranteed convergence [Jones 2021a]. We marry both approaches [Jones 2021a, Schlosser 2021] by combining their techniques and, by doing so, get converging outer approximations of the global attractor consisting of positively invariant sets based on convex optimization via sum-of-squares techniques.

We begin with following the line of reasoning from [Jones 2021a] and specify the type of Lyapunov functions we are interested in.

Definition 5.8. *Let $U \subset \mathbb{R}^n$ be open. A function $V \in \mathcal{C}^1(U)$ is called an exponential Lyapunov function for the dynamical system induced by f if for all $x \in U$*

$$V(x) \geq 0 \text{ and } \nabla V(x) \cdot f(x) \leq -V(x). \tag{5.14}$$

We call V an exponential Lyapunov function for a set $A \subset U$ if V is an exponential Lyapunov function in the above sense and $A := V^{-1}(\{0\})$.

In this section, with regard to Theorem 2.11, we make the following assumption on the basin of attraction of $\mathcal{A}$ (see Definition 2.3), which we denote by $B_f(\mathcal{A})$.

Assumption 5.9. *We assume $B_f(\mathcal{A}) \subset \mathbb{R}^n$ is open, where $\mathcal{A}$ is the GA for X.*

Let M_+ still denote the maximum positively invariant set. We infer directly

from Theorem 2.11 that the global attractor $\mathcal{A}$ can be characterized as follows

$$\begin{array}{lll} \mathcal{A} = \inf\limits_{V} & V^{-1}(\{0\}) & \\ \text{s.t.} & V \in \mathcal{C}^1(\mathbb{R}^n) & \\ & V \geq 0 & \text{on } M_+ \\ & \nabla V \cdot f \leq -V & \text{on } M_+. \end{array} \tag{5.15}$$

Using the ideas from Section 5.1 we can relate the optimization problem (5.15) to the following LP.

$$\begin{array}{lll} p_0^* = \inf & \int\limits_{M_+} w(x)\, dx & \\ \text{s.t.} & (w, V) \in \mathcal{C}(M_+) \times \mathcal{C}^1(\mathbb{R}^n) & \\ & w + V \geq \mathbf{1} & \text{on } M_+ \\ & w \geq 0 & \text{on } M_+ \\ & V \geq 0 & \text{on } M_+ \\ & \nabla V \cdot f \leq -V & \text{on } M_+ \end{array} \tag{5.16}$$

Proposition 5.10. *We have $p_0^* = \lambda(\mathcal{A})$ for p_0^* from (5.16).*

Proof. For any feasible (w, V) the function V is an exponential Lyapunov function for the GA $\mathcal{A}$. By Theorem 2.11 we have $\mathcal{A} \subset V^{-1}(\{0\})$, i.e. $V = 0$ on $\mathcal{A}$. In particular from $w + V \geq 1$ on M_+ it follows $w \geq 1$ on $\mathcal{A}$ and by non-negativity of w, we have $\int\limits_{M_+} w(x)\, dx \geq \lambda(\mathcal{A})$. That means $p_0^* \geq \lambda(\mathcal{A})$. To construct a minimizing sequence for (5.16) let $0 \leq V \in \mathcal{C}^1(B_f(\mathcal{A}))$ be an exponential Lyapunov function for $\mathcal{A}$, i.e. $V^{-1}(\{0\}) = \mathcal{A}$, satisfying $\nabla V \cdot f \leq -V$, according to Theorem 2.11. It holds $M_+ \subset B_f(\mathcal{A})$ and hence for $k \in \mathbb{N}$ the function $w_k := \max\{0, 1 - k \cdot V\}$ is continuous on M_+ with $w_k + k \cdot V \geq 1$, i.e. the pair $(w_k, k \cdot V)$ is feasible for (5.16) for all $k \in \mathbb{N}$. For $x \in \mathcal{A}$ we have $V(x) = 0$, thus $w_k(x) = \max\{0, 1 - k \cdot V(x)\} = \max\{0, 1\} = 1$, and for $x \notin \mathcal{A}$ we have $V(x) > 0$, i.e. $w_k(x) = \{0, 1 - kV(x)\} \searrow 0$ as $k \to 0$. By the monotone convergence theorem it follows $\int\limits_{M_+} w_k(x)\, dx \to \lambda(\mathcal{A})$, hence $p_0^* \leq \lambda(\mathcal{A})$. □

In the next step, we want to get rid of the explicit dependence on the unknown set M_+. Again we do so, as in Section 5.1, by adding the additional decision variable v and the constraint $\beta v - \nabla v \cdot f \geq 0$ on X for a discounting factor $\beta > 0$. The resulting LP has the form

$$\begin{array}{lll} \lambda(\mathcal{A}) = & \inf\limits_{w,V,v} \int\limits_{X} w(x)\, dx & \\ \text{s.t.} & (w, V, v) \in \mathcal{C}(\mathbb{R}^n) \times \mathcal{C}^1(\mathbb{R}^n) \times \mathcal{C}^1(\mathbb{R}^n) & \\ & w + V - v \geq \mathbf{1} & \text{on } X \\ & w \geq 0 & \text{on } X \\ & V + v \geq 0 & \text{on } X \\ & \nabla V \cdot f + V + v \leq 0 & \text{on } X \\ & \beta v - \nabla v \cdot f \geq 0 & \text{on } X \end{array} \tag{5.17}$$

where only known data (i.e. f and X appears).

In the final step, regarding the previous section and how the LP (5.6) was solved, we would like to replace the search spaces $\mathcal{C}(\mathbb{R}^n)$ and $\mathcal{C}^1(\mathbb{R}^n)$ by the space of polynomials $\mathbb{R}[x_1, \dots, x_n]$ and then use the machinery from polynomial optimization as in Section 5.1.3. Unfortunately, there exist polynomial dynamical systems, i.e. where f is polynomial, for which there does not exist a polynomial exponential Lyapunov function [Ahmadi 2018]. But in [Jones 2021a, Goluskin 2018], the authors showed that (slightly) relaxing the notion of exponential Lyapunov function allows for feasible polynomial candidates for such functions.

For $\varepsilon > 0$ we call a function $J \in \mathcal{C}^1(B_f(\mathcal{A}))$ an ε-almost Lyapunov function if

$$J \geq 0 \text{ and } \nabla J \cdot f \leq -J + \varepsilon. \tag{5.18}$$

Almost Lyapunov functions still carry important properties of the attractor, namely the set $J^{-1}([0, \varepsilon])$ contains the attractor, is positively invariant (see Lemma 3.15) and there exist polynomials $p \in \mathbb{R}[x]$ that satisfy (5.18) on X (see the discussion after Lemma 3.15))! The parameter ε can be interpreted as an indicator of how far J is from being an exponential Lyapunov function since 0-almost Lyapunov functions are exactly exponential Lyapunov functions. Because we are still interested in exponential Lyapunov functions we will add a penalty to ε. As it turns out, penalizing ε by the factor $\lambda(X)$ leads to an exact penalty function in the following LP with discounting factor $\beta > 0$

$$\begin{aligned} p^* = \inf_{w,J,\varepsilon,v} \quad & \int_X w(x)\, dx + \varepsilon\lambda(X) & & \\ \text{s.t.} \quad & (w, J, \varepsilon, v) \in \mathcal{C}(X) \times \mathcal{C}^1(\mathbb{R}^n) \times [0, \infty) \times \mathcal{C}^1(\mathbb{R}^n) & & \\ & w + J - v \geq \mathbf{1} & \text{on } X & \\ & w \geq 0 & \text{on } X & \\ & J \geq 0 & \text{on } X & \\ & \nabla J \cdot f + J + v \leq \varepsilon & \text{on } X & \\ & \beta v - \nabla v \cdot f \geq 0 & \text{on } X & \end{aligned} \tag{5.19}$$

The main result in this section is that the LP (5.19) gives the volume of the GA, i.e. $p^* = \lambda(\mathcal{A})$, and induces tight outer approximations of the GA via $J^{-1}([0, \varepsilon])$. This is stated in the following Theorem.

Theorem 5.11. *Let X be compact and $f : \mathbb{R}^n \to \mathbb{R}^n$ be a locally Lipschitz continuous vector field. Let $\mathcal{A}$ be the global attractor for the dynamical system induced by f with constraint set X. Then, for any $\beta > 0$, we have for p^* in (5.19)*

$$p^* = \lambda(\mathcal{A}).$$

Furthermore, for any feasible (w, J, ε, v) we have $J^{-1}([0, \varepsilon])$ is positively invariant and

$$\mathcal{A} \subset K := J^{-1}([0, \varepsilon]) \cap v^{-1}([0, \infty)) \cap X \tag{5.20}$$

with

$$\lambda(K \setminus \mathcal{A}) \leq \int_X w(x)\, dx + \varepsilon\lambda(X) - p^* \tag{5.21}$$

which converges to zero as (w, J, ε, v) *gets optimal for (5.19).*

Proof. The essential observation is that any feasible (w, J, ε, v) satisfies $v \geq 0$ on M_+. This follows from Lemma 5.3) by the last constraint in (5.19). It follows that $\nabla J \cdot f \leq \varepsilon - J$ on M_+, and hence $J \leq \varepsilon$ on $\mathcal{A}$ and $J^{-1}([0, \varepsilon])$ is positively invariant and contains $\mathcal{A}$ by Lemma 3.15, as well as we have $w + J \geq 1$ on M_+, and hence $w \geq 1 - \varepsilon$ on $\mathcal{A}$. That gives

$$\begin{aligned}\int_X w(x)\, dx + \varepsilon\lambda(X) &\geq (1-\varepsilon)\lambda(\mathcal{A}) + \varepsilon\lambda(X) \\ &\geq (1-\varepsilon)\lambda(\mathcal{A}) + \varepsilon\lambda(\mathcal{A}) = \lambda(\mathcal{A}),\end{aligned}$$

i.e. $p^* \geq \lambda(\mathcal{A})$. The remaining inequality $\lambda(\mathcal{A}) \leq p^*$ is the technical part of this proof. We begin by using a construction from [Schlosser 2021] to find a function $v \in \mathcal{C}^1(\mathbb{R}^n)$ with

$$\beta v - \nabla v \cdot f = 0, \quad v = 0 \text{ on } M_+ \text{ and } v < 0 \text{ on } X \setminus M_+. \tag{5.22}$$

We show that for any (w, J) feasible for (5.16) and $\varepsilon > 0$ we can find $k = k(\varepsilon) \in \mathbb{N}$ such that $(\tilde{w}, \tilde{J}, 2\varepsilon, k \cdot v)$ is feasible for (5.19), where $\tilde{w}$ and $\tilde{J}$ are such that the corresponding cost for $(\tilde{w}, \tilde{J}, 2\varepsilon, k \cdot v)$ is close to the cost of (w, J) for the LP (5.16). Since w is only non-negative on M_+ but (5.19) requires to be non-negative on X we choose $\tilde{w}$ with $\tilde{w}(x) := \max\{\hat{w} - r \cdot \operatorname{dist}(x, M_+), 0\}$ for $r > 0$ large enough (where $\hat{w}$ is any continuous extension of w to X, which exists by Tietze's extension theorem) such that

$$\int_X \tilde{w}(x)\, dx \leq \int_{M+} w(x)\, dx + \varepsilon. \tag{5.23}$$

To construct $\tilde{J}$ let $U_1 := J^{-1}([-\varepsilon/2, \infty)) \cap X \supset M_+$ and $U_2 := J^{-1}((-\infty, -\varepsilon]) \cap X$. By [Lee 2013] Theorem 2.29 we can find a non-negative function $\phi \in \mathcal{C}^1(\mathbb{R}^n)$ with $\phi = 0$ on $U_1 \supset M_+$ and $\phi \geq \min_{x \in X} J(x)$ on U_2. Then the function $\tilde{J} := J + \varepsilon + \phi$ is $\mathcal{C}^1$, is non-negative and $\tilde{J}$ (and its derivative) coincides with $J + \varepsilon$ (and its derivative) on M_+. Now we consider the choice of k such that $(\tilde{w}, \tilde{J}, 2\varepsilon, k \cdot v)$ becomes feasible for (5.19), i.e. also the first and fourth constraint in (5.19) are satisfied. Because on M_+ we have $w + J + \varepsilon \geq 1 + \varepsilon > 1$, $\nabla J \cdot f + J \leq 0 < \varepsilon$, $\tilde{w} = w$, $\tilde{J} = J$ and $\nabla \tilde{J} = \nabla J$, there is an open neighbourhood U of M_+ such that

$$\tilde{w} + \tilde{J} > 1 \text{ and } \nabla \tilde{J} \cdot f + \tilde{J} < 2\varepsilon \text{ on } U \tag{5.24}$$

Because v is non-positive and vanishes exactly on $M_+ \subset U_1$ we have $-v \geq \rho$ on $X \setminus U$ for some $\rho > 0$. Let $k \in \mathbb{N}$ with

$$k \geq \rho^{-1} \max_{x \in X \setminus U} \{1 - \tilde{w}(x) - \tilde{J}(x), \nabla \tilde{J}(x) \cdot f(x) + \tilde{J}(x) - 2\varepsilon\}. \tag{5.25}$$

By non-positivity of v and (5.24) we have

$$\tilde{w} + \tilde{J} - k \cdot v > 1 \text{ and } \nabla \tilde{J} \cdot f + \tilde{J} + k \cdot v < 2\varepsilon \text{ on } U$$

For $x \in X \setminus U$ we get by our choice of k, (5.25), that

$$\tilde{w}(x) + \tilde{J}(x) - k \cdot v(x) \geq \tilde{w}(x) + \tilde{J}(x) + k\rho \overset{(5.25)}{\geq} 1$$

and similarly for the constraint $\nabla \tilde{J} \cdot f + \tilde{J} + k \cdot v \leq 2\varepsilon$. Therefore, $(\tilde{w}, \tilde{J}, 2\varepsilon, k \cdot v)$ is feasible for (5.19). Using (5.23) we can bound the corresponding cost $\int_X \tilde{w}(x)\, dx + 2\varepsilon\lambda(X)$ by

$$\int_X \tilde{w}\, dx + 2\varepsilon\lambda(X) \leq \int_{M_+} w(x)\, dx + \varepsilon + 2\varepsilon\lambda(X).$$

Since $\varepsilon > 0$ was arbitrary we conclude $p^* \leq p_0^* = \lambda(\mathcal{A})$. Finally, it remains to show $\mathcal{A} \subset K$, for K given by (5.20), and the estimate (5.21) for any feasible (w, J, ε, v). From Lemma 3.15 it follows $J^{-1}([0, \varepsilon])$ contains the attractor. For $v^{-1}([0, \infty))$ this is true as well because property $\mathcal{A} \subset M_+ \subset v^{-1}([0, \infty))$ (see the first line of the proof). Hence, it follows

$$\mathcal{A} \subset J^{-1}([0, \varepsilon]) \cap v^{-1}([0, \infty)) \cap X = K.$$

Further, by definition of K, we obtain from the first constraint in the LP (5.19)

$$w \geq 1 - J + v \geq 1 - \varepsilon \text{ on } K. \tag{5.26}$$

Non-negativity of w now gives

$$\begin{aligned} \int_X w(x)\, dx + \varepsilon\lambda(X) &\geq \int_K 1 - \varepsilon\, dx + \varepsilon\lambda(X) \\ &= (1 - \varepsilon)\lambda(K) + \varepsilon\lambda(X) \geq \lambda(K). \end{aligned}$$

Subtracting $p^* = \lambda(\mathcal{A})$ on both sides finishes the proof. □

Remark 5.12. *The method in [Jones 2021a] treats **minimal attractors** and not global attractors in the sense of Definition 5.1. But both concepts are closely related and coincide under the additional assumption $\mathcal{A} \subset \mathring{X} \subset X \subset B_f(\mathcal{A})$. Assuming less, namely on that $X \subset B_f(A)$ for the minimal attractor A then removing the decision variable v and the constraint $\beta v - \nabla v \cdot f \geq 0$ in the LP (5.19) leads to an LP representation of the minimal attractor.*

Remark 5.13. *The dual problem of the LP (5.19) acts on the space of Borel measures on X. We did not include the dual problem here for two reasons. First, as for the LP (5.5), it gives less insight into the global attractor; see the discussion before Theorem 3.12. Second, in contrast to the LP (5.5), here the measure formulation does not give more geometric insight into the problem – the geometric interpretation is obtained from the Lyapunov approach in the LP (5.19).*

Solving the LP The LP (5.19) can be solved via the moment-sum-of-squares hierarchy as we did for the LP LP (5.6) as in Section 5.1.3. A careful look at the proof of Theorem 5.11 reveals that we have constructed a minimizing sequence that satisfies the inequality constraints in (5.19) strictly. As in the proof for the convergence of SOS-hierarchy for the LP (5.6), utilizing the Weierstraß approximation theorem and Putinar's Positivstellensatz, shows convergence for the corresponding SOS hierarchy for the LP (5.19). For details, we refer to [Schlosser 2022a].

Numerical examples We illustrate the approach from this section by three numerical examples that have been used in [Schlosser 2021, Jones 2021a]. One is the following globally asymptotically stable system with attractor $\mathcal{A} = \{(0,0)\}$, which does not allow for a polynomial Lyapunov function[Ahmadi 2011]

$$\begin{aligned}\dot{x}(t) &= -2y(t)\left(-x(t)^4 + 2x(t)^2y(t)^2 + y(t)^4\right) - \\ &\quad 2x(t)(x(t)^2 + y(t)^2)\left(x(t)^4 + 2x(t)^2y(t)^2 - y(t)^2\right) \\ \dot{y}(t) &= 2x(t)\left(x(t)^4 + 2x(t)^2y(t)^2 - y(t)^4\right) - \\ &\quad 2y(t)(x(t)^2 + y(t)^2)\left(-x(t)^4 + 2x(t)^2y(t)^2 + y(t)^4\right). \end{aligned} \tag{5.27}$$

The other two examples are the Van der Pol oscillator from (5.12) and the Hénon map from (5.11).

For the Van der Pol oscillator, as in [Jones 2021a] and [Schlosser 2021], the method from this Section works very well and the performance is comparable with the method in[Jones 2021a], see Figure 5.3. In our numerical examples, the approximation from the occupation measures approach from Section 5.1 performs slightly better than the method from this Section.

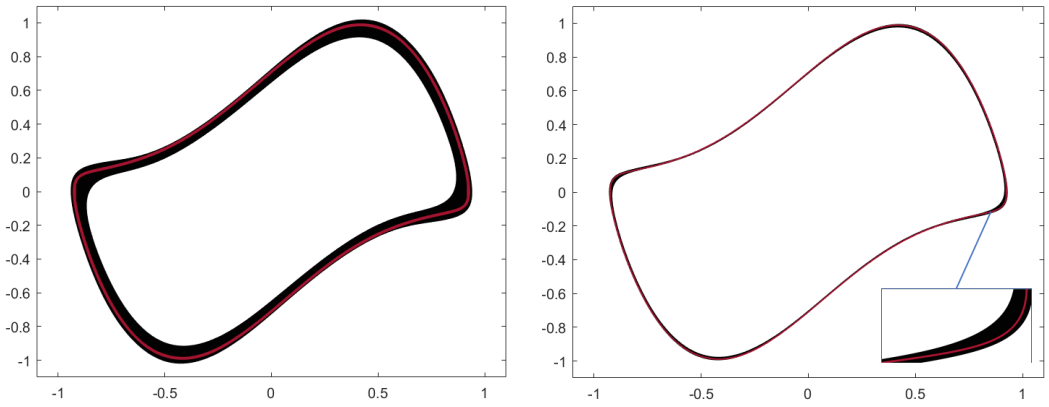

Figure 5.3: Outer approximations (black) of the attractor (red) for the Van der Pol oscillator for $X = \{x : 0.4 \leq \|x\|_2 \leq 2\}$. Left: approximation, degree 12 polynomials, and $\beta = 0.2$. Right: approximation for polynomials up to degree 16 and $\beta = 0.2$.

For the system (5.27) we notice some numerical instabilities in the decision variable ε in the LP 5.19 when solving the corresponding SDPs using Yalmip [Löfberg 2004] and Mosek [ApS 2019] (Figure 5.4 left). Using bisection in $\varepsilon \geq 0$

(for small ε) and solving the corresponding SDPs for fixed ε avoided the mentioned numerical issues.

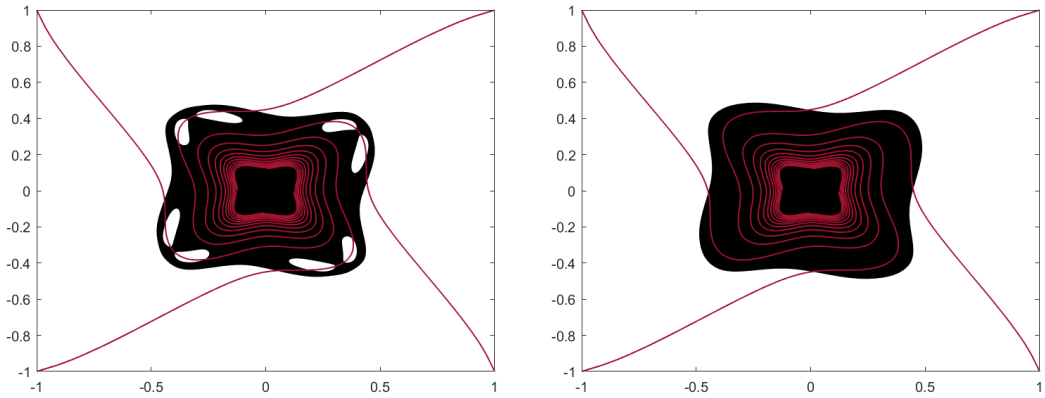

Figure 5.4: Outer approximations (black) of the attractor $\mathcal{A} = \{(0,0)\}$ and trajectories starting from $(1,1),(1,-1),(-1,1),(-1,-1)$ (red) for (5.27) for $X = [-1,1]^2$. Left: approximation by degree 16 polynomials and $\beta = 0.2$, the obtained ε^* in the corresponding SDP is too small and causes incorrect behavior of the set $J^{-1}([0,\varepsilon^*])$, see the white "holes". Right: Outer approximation using bisection on ε and polynomials up to degree 16 with discounting parameter $\beta = 0.2$.

For the Hénon map we utilized freedom of choice in the parameter $\beta > 0$. This discounting parameter $\beta > 0$ can be tuned and several solutions corresponding to different values of β can be intersected to improve the quality of the approximation. Similarly, we can introduce a parameter $\gamma > 0$ to the "almost Lyapunov" constraint by considering $\nabla J \cdot f \leq \varepsilon - \gamma \cdot J$. As for β, small values of γ describe less/slower discounting/decay and should be used when the dynamics towards the attractor is slow. The intersection of solutions for different values of β and γ for the Hénon map is illustrated on the right in Figure 5.5.

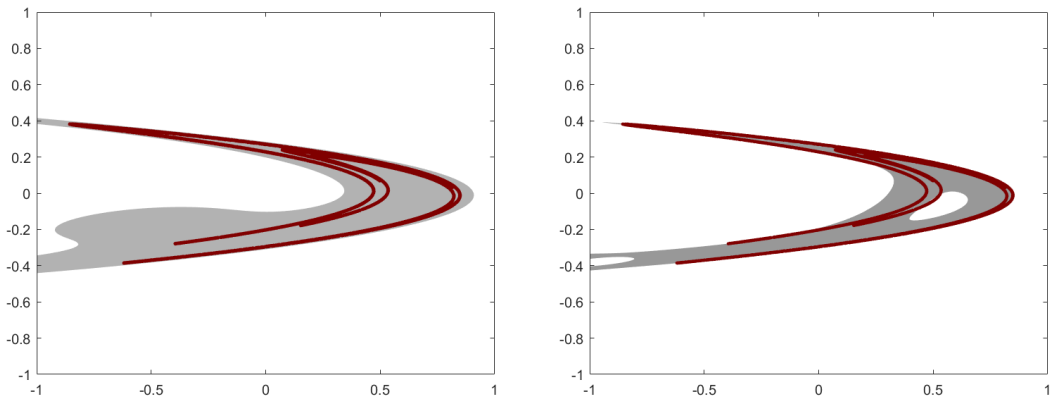

Figure 5.5: Outer approximations (gray) of the attractor (red) for the Hénon map for $X = [-1,1]^2$. Left: approximation, degree 6 polynomials and $\beta = -\log(0.002)$, $\gamma = 0.05$. Right: Intersection of approximation by degree 8 polynomials obtained by different values $\beta = -\log(0.001), -\log(0.002), -\log(0.01)$ and $\gamma = 0.002, 0.05, 0.2$.

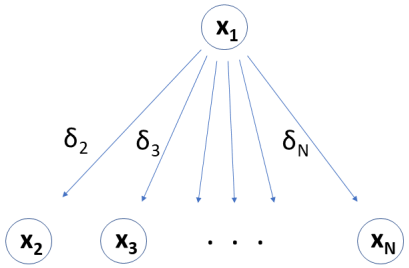

Figure 5.6: Interconnection of Van-Der Pol oscillators in a cherry structure.

5.3 Sparse LPs

We have shown in Theorem 5.7 the approximations that we obtain from the SOS hierarchies (5.8) satisfy the conditions in 2. in Theorem 4.25. Thus, for sparse dynamical systems, we can decouple the computation of the GA into its computation on the subsystems according to Algorithm 2 and get sparse convergent (with respect to Lebesgue measure discrepancy) outer approximations of the global attractor. The same is true for the outer approximations obtained via the LP (5.19) by [Schlosser 2022a, Theorem 3].

We assume that Assumption 5.6 is satisfied and immediately obtain the following result as a corollary from Theorem 4.25.

Corollary 5.14. *Algorithm 2, where in step 2 we use the SDP hierarchy (5.8), produces converging outer approximations of the global attractor $\mathcal{A}$, i.e. $S^{(k)} \supset \mathcal{A}$ for all $k \in \mathbb{N}$ and*

$$d_\lambda(S^{(k)}, \mathcal{A}) = \lambda(S^{(k)} \Delta \mathcal{A}) \to 0 \text{ as } k \to \infty.$$

When using Algorithm 4.8.3 and the minimal factorization $\mathcal{J}$ from Lemma 4.51, for selecting a subsystem decomposition in step 1 in Algorithm 2, the complexity of the corresponding SDPs is determined by the largest number of variables appearing in one of the subsystem, i.e. ω from Theorem 4.49.

Proof. This follows immediately from the convergence result Theorem 5.7 and Theorem 4.25. The complexity statement follows because the largest occurring SDP, i.e. the SDP involving the most variables, is induced by the subsystems containing the most states. Its complexity is determined by the number of states in the subsystem. By Theorem 4.49, that number is given by ω. □

The above Corollary is true in the same way for computations of the ROA set via [Henrion 2013] and the MPI set via [Korda 2014] due to the analog convergence results [Henrion 2013, Theorem 6] and [Korda 2014, Theorem 7].

Numerical examples In [Schlosser 2021], we consider the artificial cherry interconnection of Van der Pol oscillators as in Figure 5.6. We have N states $x^1, \dots, x^N \in$

$\mathbb{R}^2$. For the leaf nodes $x^2, \ldots, x^N$, the dynamics is

$$\begin{aligned}\dot{x}_1^i &= 2x_2^i \\ \dot{x}_2^i &= -0.8x_1^i - 10((x_1^i)^2 - 0.21)x_2^i + \delta_i x_1^1.\end{aligned}$$

For the root note x^1, the dynamics is

$$\begin{aligned}\dot{x}_1^1 &= 2x_2^1 \\ \dot{x}_2^1 &= -0.8x_1^1 - 10((x_1^1)^2 - 0.21)x_2^1.\end{aligned}$$

We illustrate the decoupling procedure by computing outer approximations of the MPI set of this system with respect to the constraint set $[-1.2, 1.2]^{2N}$. We carry out the computation for degree $k = 8$ in the SOS hierarchy and $N = 10$, resulting in a total dimension of the state-space equal to 20. The optimal decoupling in this case is into subsystems (x^1, x^i), $i = 2, \ldots, N$, each of dimension four. Figure 5.7 shows the sections of the MPI set outer approximations when the value at the root node is fixed at $[0.5, -0.1]$. The computation time was 12 seconds.[2] Next we carried out the the computation with $k = 8$ and $N = 26$, resulting in state-space dimension of 52. Figure 5.8 shows the sections of the MPI set outer approximations when the value at the root node is fixed at $[0.5, -0.1]$. The total computation time was 40.3 seconds. It should be mentioned that these problems in dimension 20 or 52 are currently intractable without structure exploitation. Here the sparse structure allowed for decoupling in 9 respectively 25 problems in 4 variables, which were solved in less than a minute in total.

TSSOS: Exploiting term sparsity In [Wang 2021b] we explored the application of TSSOS – term sparsity sum-of-squares – from [Wang 2021a, Wang 2020] for dynamical systems. The sparsity type that we consider in Chapter 4 and that was applied to the previous numerical examples is of correlation type. In contrast to this, term sparsity investigates "algebraic" sparsity in the exponents of the appearing polynomials (that includes the dynamics f as well as the polynomial description of the constraint set X). At each level of the SOS hierarchy, this allows inducing a second hierarchy with the goal of exploiting term sparse structures and reducing computation time, see [Wang 2021a, Wang 2020] for details and applications. In [Wang 2021b] we illustrate by numerical examples that TSSOS can help reduce computation time in some cases. The approach provides a trade-off between computational costs and solution accuracy. However, under certain symmetry conditions for the problem (i.e. symmetries in the dynamics and the constraint set) this approach enjoys convergence (in each level of the SOS hierarchy) and recovers the sign symmetry reduction [Wang 2021b, Theorem 3.11].

[2] All computations were carried out using YALMIP [Löfberg 2004] and MOSEK running on Matlab and 4.2 GHz Intel Core i7, 32 GB 2400MHz DDR4.

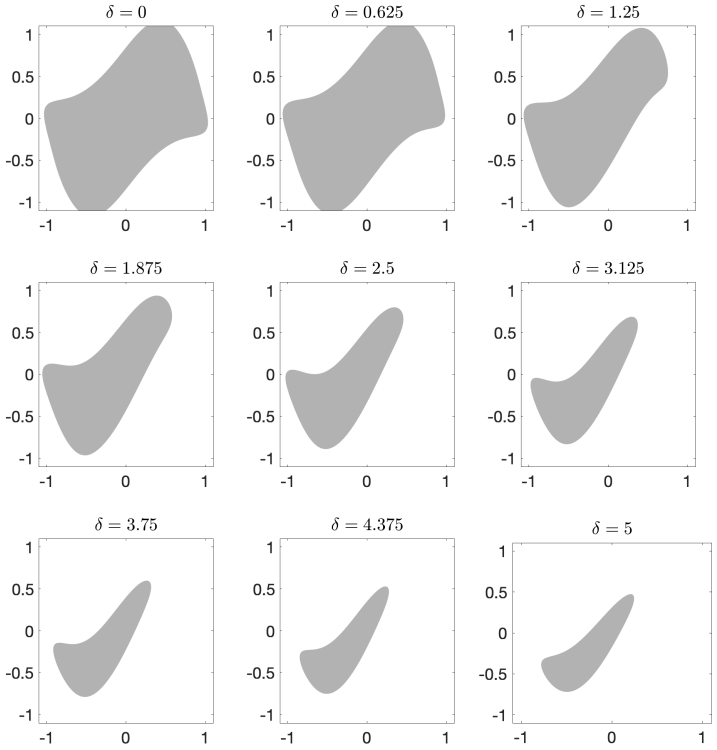

Figure 5.7: Van der Pol oscillators in a cherry structure: The figure shows the outer approximations of the MPI set for $k = 8$ and $N = 10$ for the subsystems given by the cherry–branches.

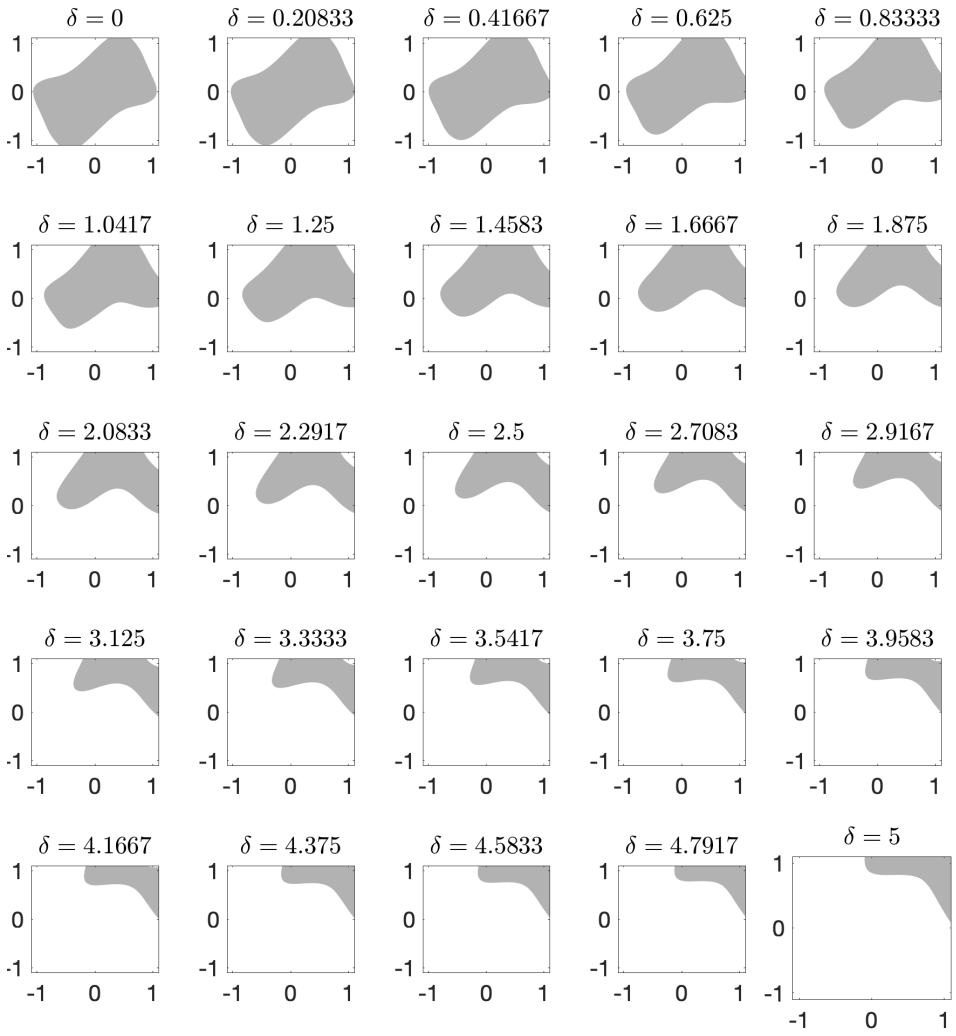

Figure 5.8: Van der Pol oscillators in a cherry structure: The figure shows the outer approximations of the MPI set for $k = 8$ and $N = 26$ for the subsystems given by the cherry–branches

Chapter 6

Koopman semigroup – sparsity structures and domains of reproducing kernel spaces

The Koopman lifting procedure, see Section 2.5, results in a linear operator (semigroup) representation of the dynamical system. The first part of this chapter contains our analysis for Koopman and Perron-Frobenius operators on reproducing kernel Banach spaces from [Ikeda 2022b]. In the second part, we transfer sparsity structures of the dynamical system to "block structures" of these operators, this part is based on [Schlosser 2022b].

6.1 Koopman and Perron-Frobenius analysis on RKBS; discrete time systems

We begin this section with discrete dynamical systems, i.e. we consider a map $f : X \to X$. We will define the Koopman operators and their adjoint operators on reproducing kernel Banach spaces $(\mathcal{B}, \mathcal{B}', \langle \cdot, \cdot \rangle, k)$. Apart from the Koopman lift, RKBSs provide a different way of lifting the dynamics into a linear setting. The main idea is that we can define an operator K_f transporting the dynamics f into the RKBS. This is done via the feature map $x \mapsto k(x, \cdot)$ by

$$K_f k(x, \cdot) := k(f(x), \cdot). \tag{6.1}$$

We illustrate this in Figure 6.1.

There is a beautiful connection between the operator K_f and the Koopman operator on the RKBS. Namely, it turns out, that the Koopman operator is the adjoint of the operator K_f. In the following we exploit this relation between those two operators.

6.1.1 Definitions of the Koopman and Perron-Frobenius operator

To guarantee that the map K_f from (6.1) is well defined, it is useful to assume that $k(x_1, \cdot), \ldots, k(x_n, \cdot)$ are linearly independent for any $n \in \mathbb{N}$ and any choice of pairwise distinct point $x_1, \ldots, x_n \in X$.

Assumption 6.1. *We assume that the set $\{k(x, \cdot) : x \in X\} \subset \mathcal{B}'$ is linearly independent.*

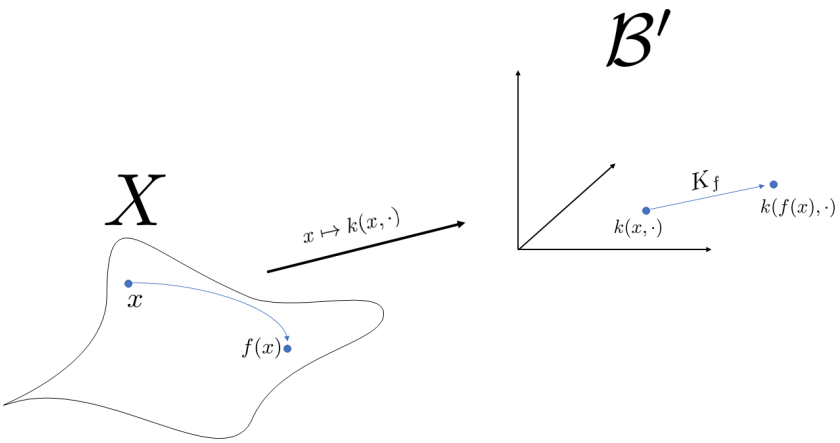

Figure 6.1: Illustration of the operator K_f.

Remark 6.2. *For RKHS the $\{k(x,\cdot) : x \in X\}$ is linearly independent if and only if k is a strictly positive kernel, that is, the kernel k satisfies for all $n \in \mathbb{N}$, $(a_1,\dots,a_n) \in \mathbb{C}^n \setminus \{0\}$, $(x_1,\dots,x_n) \in X^n$*

$$\sum_{i,j=1}^{n} a_i \overline{a}_j k(x_i,x_j) > 0.$$

Definition 6.3 (Koopman and Perron-Frobenius operator)**.** *Let $(\mathcal{B},\mathcal{B}',\langle\cdot,\cdot\rangle,k)$ be an RKBS with kernel such that Assumption 6.1 is satisfied. Let $f : X \to X$ be given dynamics. The Koopman operator $T_f : \mathcal{B} \supset D(T_f) \to \mathcal{B}$ is defined by*

$$T_f g := g \circ f \quad \text{for } g \in D(T_f) := \{h \in \mathcal{B} : h \circ f \in \mathcal{B}\}. \tag{6.2}$$

The Perron-Frobenius operator $K_f : \operatorname{Span}\{k(x,\cdot) : x \in X\} \to \operatorname{Span}\{k(x,\cdot) : x \in X\} \subset \mathcal{B}'$ is defined by

$$K_f k(x,\cdot) := k(f(x),\cdot) \text{ for } x \in X \tag{6.3}$$

and extended linearly to $\operatorname{Span}\{k(x,\cdot) : x \in X\}$.

Remark 6.4. *It is shown in [Cowen 1995, Theorem 1.4] that if a bounded operator K leaves the set $\{k(x,\cdot) : x \in X\}$ invariant then K is a Perron-Frobenius operator.*

Assumption 6.1 guarantees that extending (6.3) linearly to $\operatorname{Span}\{k(x,\cdot) : x \in X\}$ is well defined. To define the adjoint operator of the Perron-Frobenius operator K_f, we first need that K_f is densely defined. Therefore, we recall the notion of density from [Lin 2022] – a set $W \subset \mathcal{B}$ respectively $W' \subset \mathcal{B}'$ is called dense with respect to $\langle\cdot,\cdot\rangle$ if

$$\langle w, g\rangle = 0 \text{ for all } w \in W \text{ implies } g = 0 \tag{6.4}$$

and analog for W'

$$\langle v, w' \rangle = 0 \text{ for all } w' \in W' \text{ implies } v = 0. \tag{6.5}$$

In the case of $W = \mathcal{B}$ and $W' = \mathcal{B}'$, the conditions (6.4) and (6.5) state that the bilinear form $\langle \cdot, \cdot \rangle$ is non-degenerate. Condition (6.4) is a reformulation of the map ϕ from (2.83) being injective and (6.5) states that we can embed $\mathcal{B}$ into $(\mathcal{B}')^*$, the dual space of $\mathcal{B}'$. Hence, the conditions (6.4) and (6.5) describe foremost an algebraic property of the bilinear form $\langle \cdot, \cdot \rangle$ and therefore should not be mistaken with the notion of density with respect to the topologies on $\mathcal{B}$ and $\mathcal{B}'$.

Remark 6.5. *The set* $\operatorname{Span}\{k(x, \cdot) : x \in X\}$ *is dense in* $\mathcal{B}'$ *with respect to* $\langle \cdot, \cdot \rangle$ *because for any* $g \in \mathcal{B}$ *with* $0 = \langle g, h \rangle$ *for all* $h \in \mathcal{B}'$*, we have in particular* $g(x) = \langle g, k(x, \cdot) \rangle = 0$*, i.e.* g *is the zero function. For reflexive RKBS, in particular RKHS, we get also that* $\mathcal{B}$ *is dense in* $\mathcal{B}$*.*

The following result states that the Perron-Frobenius operator is adjoint (with respect to $\langle \cdot, \cdot \rangle$) to the Koopman operator and extends the result from the RKHS setting in [Rosenfeld 2020]. Note that we use the notation A' for the adjoint with respect to a bilinear form $\langle \cdot, \cdot \rangle$ and A^* for the classical adjoint operator.

Lemma 6.6. *Let* $(\mathcal{B}, \mathcal{B}', \langle \cdot, \cdot \rangle, k)$ *be an RKBS with kernel satisfying Assumption 6.1. Then* K_f *is densely defined with respect to* $\langle \cdot, \cdot \rangle$ *and we have* $T_f = K_f'$*.*

Proof. Since we assume that the set $\{k(x, \cdot) : x \in X\}$ is linearly independent the Perron-Frobenius operator is well defined. By Remark 6.5 $\operatorname{Span}\{k(x, \cdot) : x \in X\}$ is dense in $\mathcal{B}'$ with respect to $\langle \cdot, \cdot \rangle$ and by Lemma 2.77 the adjoint of K_f exists and is unique. To check that T_f is the adjoint of K_f let $g \in D(K_f')$ then for all $x \in X$ we have

$$\begin{aligned} K_f' g(x) &= \langle K_f' g, k(x, \cdot) \rangle = \langle g, K_f k(x, \cdot) \rangle = \langle g, k(f(x), \cdot) \rangle \\ &= g(f(x)) = T_f g(x). \end{aligned}$$

This shows that T_f is at least an extension of K_f'. For $g \in D(T_f)$, i.e. $g \in \mathcal{B}$ such that $g \circ f \in \mathcal{B}$ we have

$$\begin{aligned} \langle g, K_f k(x, \cdot) \rangle &= \langle g, k(f(x), \cdot) \rangle = g(f(x)) = (g \circ f)(x) \\ &= \langle g \circ f, k(x, \cdot) \rangle = \langle T_f g, k(x, \cdot) \rangle. \end{aligned}$$

Hence we have $K_f' = T_f$. □

6.1.2 Basic properties

In Theorem 6.7 we present a collection of fundamental properties of the Koopman and Perron-Frobenius operator on RKBS. Before stating these properties, we want to put them into context with existing results for Koopman and Perron-Frobenius operators on RKHS. The first three statements in Theorem 6.7 transfer from classical arguments for composition operators; in particular it shows that the information

about the dynamical system is incorporated in the Koopman operator (statement 3. in Theorem 6.7). Statement 4. is an extension from existing results for the RKHS setting and can be found in [Rosenfeld 2019, Paulsen 2016]. Statement 6. is a transfer of a classical result for adjoint operators to the RKBS setting and statement 8. relates to kernel-mean embeddings from [Klus 2020].

Theorem 6.7. *Let $f, \tilde{f} : X \to X$ be two maps and $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS on X with kernel k satisfying Assumption 6.1. Then*

1. $K_f K_{\tilde{f}} = K_{f\circ\tilde{f}}$
2. *if f is a bijection then $K_f^{-1} = K_{f^{-1}}$*
3. *If $\mathcal{B}$ is dense in $\mathcal{B}$ with respect to $\langle\cdot,\cdot\rangle$ then*

$$f = \hat{f} \text{ if and only if } K_f = K_{\hat{f}}.$$

4. *T_f is closed (with respect to the weak as well as norm topology). In particular, T_f is bounded if and only if $D(T_f) = \mathcal{B}$.*
5. *Assume X is compact and $\mathcal{B}$ has the universal property (see Definition 2.57), $f : X \to X$ is continuous and one of the following holds*

   (a) *The map ϕ from (2.83) is an isomorphism*

   (b) *$x \mapsto k(x,\cdot) \in \mathcal{B}'$ is continuous*

   Then, if X contains infinitely many elements, the operator K_f is not closed with respect to $\langle\cdot,\cdot\rangle$.
6. *If T_f is densely defined then K_f is closable. If the map ϕ from (2.83) is an isomorphism and $\mathcal{B}$ is reflexive then the converse is true as well, i.e. if K_f is closable then T_f is densely defined.*
7. *Assume the map ϕ from (2.83) is an isomorphism. If $D(T_f) = \mathcal{B}$ then K_f can be extended to a bounded operator on $\mathcal{B}'$. If, in addition, $\mathcal{B}$ is reflexive then the converse is true as well.*
8. *Under the assumptions of point 5., the operator K_f can be extended to*

$$D := \left\{ \int_X k(x,\cdot)\, d\mu(x) : \mu \in M(X) \right\} \tag{6.6}$$

   by

$$\bar{K}_f \left(\int_X k(x,\cdot)\, d\mu(x) \right) := \int_X k(f(x),\cdot)\, d\mu(x) \text{ for } \mu \in M(X). \tag{6.7}$$

   where $M(X)$ denotes the set of Borel measures on X.

Proof. We have for all $x \in X$

$$\begin{aligned} K_f K_{\tilde{f}} k(x,\cdot) &= K_f k(\tilde{f}(x),\cdot) = k(f(\tilde{f}(x)),\cdot) = k((f\circ\tilde{f})(x),\cdot) \\ &= K_{f\circ\tilde{f}} k(x,\cdot). \end{aligned}$$

Hence $K_f K_{\tilde{f}} = K_{f\circ\tilde{f}}$ on $\mathrm{Span}\{k(x,\cdot) : x \in X\}$ and if follows the first statement. In particular it follows $K_f^{-1} = K_{f^{-1}}$ if f is invertible. For the third statement for $f = \hat{f}$ it is obvious that also $K_f = K_{\hat{f}}$. Assume now $K_f = K_{\hat{f}}$. Then for $x \in X$ and all $h \in \mathcal{B}$

$$0 \quad = \quad \langle h, (K_f - K_{\hat{f}})k(x,\cdot)\rangle = \langle h, k(f(x),\cdot) - k(\hat{f}(x),\cdot)\rangle.$$

Hence, since we assumed $\mathcal{B}$ to be dense in $\mathcal{B}$ with respect to $\langle\cdot,\cdot\rangle$, it follows $k(f(x),\cdot) = k(\hat{f}(x),\cdot)$. From Assumption 6.1, it follows $f(x) = \hat{f}(x)$. The fourth statement follows from $T_f = K_f'$ (by Lemma 6.6), Lemma 2.77 and the closed graph theorem. We will show the fifth statement at last once we have proven 8. For 6., if T_f is densely defined then $B := T_f'$ is a closed extension of K_f. If $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle)$ is reflexive then the second statement in 6. follows directly from [Ikeda 2022b, Proposition B.8]. For 7., assume that $D(T_f) = \mathcal{B}$ then by 4. we have that T_f is bounded. The idea is to use the adjoint of T_f together with the isomorphism ϕ to define a natural candidate for an extension of K_f. We define the bounded operator $T := \phi^{-1} T_f^* \phi : \mathcal{B}' \to \mathcal{B}'$, where $T_f^* : \mathcal{B}^* \to \mathcal{B}^*$ denotes the (classical) adjoint of T_f. We claim that T extends K_f. To check this let $x \in X$ and $g \in \mathcal{B}$, then by definition of ϕ and Lemma 6.6

$$\begin{aligned} \langle g, Tk(x,\cdot)\rangle &= \langle g, \phi^{-1} T_f^* \phi k(x,\cdot)\rangle = (T_f^* \phi k(x,\cdot))(g) \\ &= (\phi k(x,\cdot))(T_f g) = \langle T_f g, k(x,\cdot)\rangle \\ &= \langle g, K_f k(x,\cdot)\rangle. \end{aligned}$$

From which it follows $Tk(x,\cdot) = K_f k(x,\cdot)$ because ϕ is injective (or in other words, $\mathcal{B}$ is dense in $\mathcal{B}$ with respect to $\langle\cdot,\cdot\rangle$). For the second statement of 7. we assume that K_f has a bounded extension $K : \mathcal{B}' \to \mathcal{B}'$ and $\mathcal{B}$ reflexive and want to show that T_f is bounded. The idea is very similar but the adjoint of K is an operator on $\mathcal{B}'^* \cong \mathcal{B}^{**}$ and in order to find an operator on $\mathcal{B}$ we use that $\mathcal{B}$ is reflexive. That $\mathcal{B}$ is reflexive means that the map

$$J : \mathcal{B} \to \mathcal{B}^{**}, J(b)(b^*) := b^*(b) \tag{6.8}$$

is an isomorphism. We define the candidate operator

$$U := J^{-1}(\phi^*)^{-1} K^* \phi^* J : \mathcal{B} \to \mathcal{B}. \tag{6.9}$$

The operator U from (6.9) is bounded and we claim that $U = T_f$. To check this let

$g \in \mathcal{B}$ and $x \in X$. Then playing with the definition of ϕ, ϕ^* and J gives

$$\begin{aligned} Ug(x) &= \langle Ug, k(x,\cdot)\rangle = \phi(k(x,\cdot))(Ug) \\ &= \phi(k(x,\cdot))\left(J^{-1}(\phi^*)^{-1}K^*\phi^*Jg\right) \\ &= \left((\phi^*)^{-1}K^*\phi^*Jg\right)(\phi(k(x,\cdot)) \\ &= (K^*\phi^*Jg)\,(\phi^{-1}\phi(k(x,\cdot)) = (K^*\phi^*Jg)k(x,\cdot) \\ &= (\phi^*Jg)(Kk(x,\cdot)) = (\phi^*Jg)k(f(x),\cdot) \\ &= Jg(\phi k(f(x),\cdot)) = \phi(k(f(x),\cdot))(g) \\ &= \langle g, k(f(x),\cdot) = g(f(x)) = T_f g(x). \end{aligned}$$

To show statement 8., we separate the two cases of assumptions (a) and (b) from 5. In the case of (b) note first that the (Bochner) integrals in (6.6) and (6.7) exist due to the continuity assumptions on k and f. By choosing μ to be a dirac delta δ_y for some $y \in X$ we get

$$\bar{K}_f k(y,\cdot) = \bar{K}_f\left(\int_X k(x,\cdot)\, d\delta_y(x)\right) = \int_X k(f(x),\cdot)\, d\delta_y(x) = k(f(y),\cdot).$$

That shows that $\bar{K}_f$ extends K_f. It remains to show that (6.7) is well defined. That means whenever there are two measures $\mu, \nu \in M(X)$ with

$$\int_X k(x,\cdot)\, d\mu(x) = \int_X k(x,\cdot)\, d\nu(x) \tag{6.10}$$

then also $\int_X k(f(x),\cdot)\, d\mu(x) = \int_X k(f(x),\cdot)\, d\nu(x)$. This follows trivially once we have shown that the representation of (6.10) is unique, i.e. (6.10) implies $\mu = \nu$. From (6.10) we get for all $g \in \mathcal{B}$ by continuity of the bilinear form

$$\begin{aligned} \int_X g(x)\, d\mu(x) &= \int_X \langle g, k(x,\cdot)\rangle\, \mu(x) = \left\langle g, \int_X k(x,\cdot)\, d\mu(x)\right\rangle \\ &= \left\langle g, \int_X k(x,\cdot)\, d\nu(x)\right\rangle = \int_X g(x)\, d\nu(x). \end{aligned}$$

The universal property together with the Riesz-Markov representation theorem implies now that $\mu = \nu$. Now assume (a) from 5 instead of (b). The isomorphism ϕ is defined by $\phi : \mathcal{B}^* \to \mathcal{B}'$ with $b^*(b) = \langle b, \phi(b^*)\rangle$ for all $b^* \in \mathcal{B}^*$. To show that we can extend K_f by (6.7), we first show that the argument in (6.7) as well as the proposed image have representations based on the embedding $i : \mathcal{B} \to \mathcal{C}(X)$ (more precisely its adjoint $i^* : M(X) \to \mathcal{B}^*$), the isomorphism ϕ, and $b \in \mathcal{B}$, and the Perron-Frobenius operator P_f on $M(X)$ from (6.46). Note that, here the term

$\int_X k(x,\cdot)\, d\mu(x)$ is understood in the weak sense, that is, for each $g \in \mathcal{B}$ we have

$$\langle g, \int_X k(x,\cdot)\, d\mu(x)\rangle := \int_X \langle g, k(x,\cdot)\rangle\, d\mu(x) = \int_X g(x)\, d\mu(x). \tag{6.11}$$

Next, we claim that $\int_X k(x,\cdot)\, d\mu(x)$, is nothing else than $\phi(i^*\mu)$ for all $\mu \in M(X)$. This can be seen as follows: For any $g \in \mathcal{B}$ we have

$$\begin{aligned}\langle g, \phi(i^*\mu)\rangle &= (i^*\mu)(g) = \int_X i(g)(x)\, d\mu(x) = \int_X g(x)\, d\mu(x)\\ &\overset{(6.11)}{=} \langle g, \int_X k(x,\cdot)\, d\mu(x)\rangle.\end{aligned}$$

Similarly for the right-hand side of (6.7). Namely, for any $g \in \mathcal{B}$

$$\begin{aligned}\left\langle g, \int_X k(f(x),\cdot)\, d\mu(x)\right\rangle &= \int_X g(f(x))\, d\mu(x) = \int_X g\, dP_f\mu\\ &= (i^*(P_f\mu))(g) = \langle g, \phi(i^*(P_f\mu))\rangle.\end{aligned}$$

where P_f denotes the Perron-Frobenius operator on $M(X)$ from (2.51). That means (6.7) states that we want to extend K_f to the range of $\phi \circ i^*$, i.e. D, by setting

$$\bar{K}_f(\phi(i^*\mu)) := \phi(i^* P_f \mu)). \tag{6.12}$$

First let us check that this is well defined. The universal property implies i^* is injective and hence $\phi \circ i^*$ is injective, too – hence (6.12) is well defined. Finally, to see that $\bar{K}_f$ is indeed an extension of K_f we have show that $\bar{K}_f k(x,\cdot) = k(f(x),\cdot)$ for all $x \in X$. As in the previous case we use that $\phi(i^*\delta_x) = k(x,\cdot)$ for any $x \in X$, from which it follows

$$\begin{aligned}\bar{K}_f k(x,\cdot) &= \bar{K}_f(\phi(i^*\delta_x)) = \phi(i^*(P_f\delta_x)) = \phi(i^*\delta_{f(x)})\\ &= k(f(x),\cdot).\end{aligned}$$

This shows 8. under assumption (b) from 5. Last, it remains to show 5. The property that is important in this proof is that weak* convergence of measures μ_n to $\mu \in M(X)$, denoted by $\mu_n \overset{*}{\rightharpoonup} \mu$, implies

$$\left\langle g, \int_X k(x,\cdot)\, d\mu_n(x)\right\rangle \to \langle g, \int_X k(x,\cdot)\, d\mu(x)\rangle \tag{6.13}$$

for all $g \in \mathcal{B}$. This follows directly from the weak* convergence of μ_n, namely

$$\left\langle g, \int_X k(x,\cdot)\, d\mu_n(x)\right\rangle = \int_X g\, d\mu_n \to \int_X g\, d\mu = \langle g, \int_X k(x,\cdot)\, d\mu(x)\rangle.$$

We use the extension $\bar{K}_f$ of K_f from 8. and show $\bar{K}_f = K_f$ if K_f was closed with respect to $\langle \cdot, \cdot \rangle$. But this will lead to a contradiction because we will see that the domain of $\bar{K}_f$ is strictly greater than the domain of K_f. Let $\mu \in M(X)$. We may assume that μ represents a non-negative measure – otherwise, apply the Hahn-Jordan decomposition to μ. By scaling we may assume that μ is a probability measure. Then for $n \in \mathbb{N}$ there exist $x_1^{(n)}, \dots, x_{k_n}^{(n)} \in X$ and $\lambda_1^{(n)}, \dots, \lambda_{k_n}^{(n)} \geq 0$ with $\sum_{i=1}^{k_n} \lambda_i^{(n)} = 1$ such that

$$\mu_n := \sum_{i=1}^{k_n} \lambda_i^{(n)} \delta_{x_i^{(n)}} \overset{*}{\rightharpoonup} \mu \quad \text{as } n \to \infty. \tag{6.14}$$

By continuity of the Perron-Frobenius operator P_f on $M(X)$ from (6.46) we then also have

$$\sum_{i=1}^{k_n} \lambda_i^{(n)} \delta_{f(x_i^{(n)})} = P_f \mu_n \overset{*}{\rightharpoonup} P_f \mu \tag{6.15}$$

For for any $g \in \mathcal{B}$ we get from (6.14)

$$\langle g, \sum_{i=1}^{k_n} \lambda_i^{(n)} k(x_i^{(n)}, \cdot) \rangle = \left\langle g, \int_X k(x, \cdot)\, d\mu_n(x) \right\rangle \to \left\langle g, \int_X k(x, \cdot)\, d\mu(x) \right\rangle$$

and from (6.15)

$$\begin{aligned} \langle g, K_f \sum_{i=1}^{k_n} \lambda_i^{(n)} k(x_i^{(n)}, \cdot) \rangle &= \langle g, \sum_{i=1}^{k_n} \lambda_i^{(n)} k(f(x_i^{(n)}), \cdot) \rangle = \left\langle g, \int_X k(f(x), \cdot)\, d\mu_n(x) \right\rangle \\ &= \left\langle g, \int_X k(x, \cdot)\, dP_f \mu_n(x) \right\rangle \to \left\langle g, \int_X k(x, \cdot)\, dP_f \mu(x) \right\rangle \end{aligned}$$

Because we assumed that K_f was closed with respect to $\langle \cdot, \cdot \rangle$ it follows in particular that $\int_X k(x, \cdot)\, d\mu(x) \in D(K_f) = \operatorname{Span}\{k(x, \cdot) : x \in X\}$. That means we can find $m \in \mathbb{N}$, $y_1, \dots, y_m \in X$, $a_1, \dots, a_m \in \mathbb{R}$ with

$$\int_X k(x, \cdot)\, d\mu(x) = \sum_{i=1}^{m} a_i k(y_i, \cdot) \text{ in } \mathcal{B}', \tag{6.16}$$

which means for all $g \in \mathcal{B}$ we have

$$\int_X g\, d\mu = \sum_{i=1}^{m} a_i g(y_i) = \int_X g\, d \left(\sum_{i=1}^{m} a_i \delta_{y_i} \right).$$

From the universal property it follows $\mu = \sum_{i=1}^{m} a_i \delta_{y_i}$, i.e. μ is atomic. Since μ was arbitrary that means all Borel measures $\mu \in M(X)$ are atomic – but this is not

true when X contains infinitely many points, see [Ikeda 2022b, Lemma C.1]. □

Remark 6.8 (Invariant kernels)**.** *An easy (but restrictive, see [Das 2019]) setting that guarantees boundedness of the operator T_f on an RKHS $\mathcal{H}$ with kernel k is invariance of k, i.e. for all $x, y \in X$ it holds*

$$k(f(x), f(y)) = k(x, y). \tag{6.17}$$

In this case T_f and K_f are isometries, due to

$$\begin{aligned}\left\| K_f \sum_{i=1}^{n} a_i k(x_i, \cdot) \right\|^2 &= \left\| \sum_{i=1}^{n} a_i k(f(x_i), \cdot) \right\|^2 = \sum_{i,j=1}^{n} a_i \overline{a}_j k(f(x_i), f(x_j)) \\ &= \sum_{i,j=1}^{n} a_i \overline{a}_j k(x_i, x_j) = \|k(x, \cdot)\|^2\end{aligned}$$

for all $n \in \mathbb{N}$ and $a_1, \ldots, a_n \in \mathbb{C}$. More generally, by the same arguments, the Perron-Frobenius operator is bounded with $\|K_t\| \leq M$ if and only if we have

$$\sum_{i,j=1}^{n} a_i \overline{a}_j k(f(x_i), f(x_j)) \leq M \sum_{i,j=1}^{n} a_i \overline{a}_j k(x_i, x_j). \tag{6.18}$$

In contrast to (6.17) the condition (6.18) is typically not easily verified.

One possibility of defining an RKBS, such that the Koopman operators are bounded, uses conjugacy and follows the classical concept for dynamical systems that sometimes (local) charts give better insight into the dynamics. In the following proposition, we use the notation from Lemma 2.72 in the preliminary section.

Proposition 6.9. *Let $f : X \to X$ and $(\mathcal{B}, \mathcal{B}', \langle \cdot, \cdot \rangle, k)$ be an RKBS on X with kernel. Let $g : Y \to Y$ such that there exists a bijective function $\phi : Y \to X$ with $\phi \circ g = f \circ \phi$. Let $(\mathcal{B}_\phi, \mathcal{B}'_\phi, \langle \cdot, \cdot \rangle_\phi, k_\phi)$ be the corresponding pullback RKBS with kernel from Lemma 2.72. Then for the Perron-Frobenius operators K_f on $\mathcal{B}'$ and K_g on $\mathcal{B}'_\phi$ it holds*

$$T_\phi K_f = K_g T_\phi. \tag{6.19}$$

In particular if K_f is bounded on $\mathcal{B}'$ then so is K_g with $\|\overline{K}_g\| = \|\overline{K}_f\|$.

Proof. By Lemma 2.72 we have that T_ϕ is an isometric isomorphism. Hence, it remains to show (6.19). For any $x \in X$ we have

$$\begin{aligned}T_\phi K_f k(x, \cdot) &= T_\phi k(f(x), \cdot) = k(f(x), \phi(\cdot)) = k(\phi(g(\phi^{-1}(x)), \phi(\cdot)) \\ &= k_\phi(g(\phi^{-1}(x)), \cdot) = K_g k_\phi(\phi^{-1}(x), \cdot) = K_g k(x, \phi(\cdot)) \\ &= K_g T_f k(x, \cdot)\end{aligned}$$

□

6.1.3 Examples

In this section, we present several examples from the literature. Example 6.10 recovers the case of the Koopman operator acting $\mathcal{C}(X)$ from an RKBS perspective. Other examples treat holomorphic dynamics, as Example 6.11, point out limitations of the approach, as in [Ishikawa 2021] or Examples 6.13 and 6.12, or Sobolev spaces in Example 6.14.

Those examples, particularly the limiting ones, demonstrate that not any RKBS fits the dynamical system at hand, and properties of the dynamical system, such as linearity or regularity, have to be considered for the choice of the kernel.

We start with the classical example of the Koopman operator on $\mathcal{C}(X)$. We view $\mathcal{C}(X)$ as an RKBS as in Example 2.69.

Example 6.10 ($\mathcal{C}(X)$ as an RKBS.)**.** *As in Example 2.69, for compact X, we view $\mathcal{C}(X)$ equipped with the supremum norm $\|\cdot\|_\infty$ as an RKBS with a kernel $k : X \times X \to \mathbb{R}$ continuous such that $\operatorname{Span}\{k(\cdot, x) : x \in X\}$ is a dense subset of $\mathcal{C}(X)$. For the Koopman operator T_f for a continuous discrete time dynamics $f : X \to X$, it holds $D(T_f) = \mathcal{B} = \mathcal{C}(X)$ because $g \circ f$ is continuous whenever g is. And hence K_f can be extended to a bounded operator on $\mathcal{B}'$ by Theorem 6.7. For the examples $k(x, y) = 1 - |x - y|$, $k(x, y) = e^{xy}$ and $k(x, y) = (1 + y)^x$ for $X = [0, 1]$ we get*

$$K_f : \begin{cases} 1 - |x - \cdot| \mapsto 1 - |f(x) - \cdot| \\ e^{x\cdot} \mapsto e^{f(x)\cdot} \\ (1 + \cdot)^x \mapsto (1 + \cdot)^{f(x)}. \end{cases} \tag{6.20}$$

In the next example, we consider Hardy spaces $H^p(D)$ from Example 2.61 and Remark 2.73, where $p \geq 1$ and D is the unit disc $D \subset \mathbb{C}$.

Example 6.11. *The Hardy space $H^p(D)$ consists of all analytic functions on D for which the following norm is finite*

$$\|g\|_{H^p} := \sup_{0 \leq r < 1} \left(\int_0^{2\pi} |f(re^{i\theta})|^p \, d\theta \right)^{\frac{1}{p}}. \tag{6.21}$$

The kernel is given by the Szegö kernel $k(z, w) := \frac{1}{1 - z\overline{w}}$ and turns $B := H^p(D)$ into an RKBS where we take the dual-pairing $\langle \cdot, \cdot \rangle$ of $H^p(D)$ and its topological dual space (with respect to the norm $\|\cdot\|_{H^p}$) and we set $\mathcal{B}' := \overline{\operatorname{Span}\{k(z, \cdot) : z \in D\}}$ where the closure is taken in the topological dual space of $H^p(D)$. By [Cowen 1995, Theorem 3.6] a holomorphic automorphism $f : D \to D$ has a bounded Koopman operator on $H^p(D)$ with $\|T_f\|^p = \frac{1 + |f(0)|}{1 - |f(0)|}$. For dynamics given by a Möbiustransform $f(z) := \lambda \frac{z - a}{1 - z\overline{a}}$ for $a, \lambda \in \mathbb{C}$ with $|\lambda| = 1$ and $|a| < 1$ we define the map $\phi(z) := \frac{z - \gamma}{1 - z\overline{\gamma}}$ with the unique fixed point $\gamma \in D$ of f. It can be shown that the pull-back kernel k_ϕ is an invariant kernel for f and thus T_f is an isometry (see Remark 6.8) on the pullback RKBS (see Lemma 2.72) and K_f and can be extended to an isometry on the same space as well. In [Russo 2022] the authors go further and consider weighted composition operators on the Hardy space and show stronger boundedness results for this setting.

The following example shows that even a small perturbation can cause the Koopman operator of the perturbed system to have a trivial domain on the same RKBS for which the unperturbed system induces a bounded Koopman operator.

Example 6.12. *Consider the dynamical system on* $[0,1]$ *given by the linear map* $f : [0,1] \to [0,1]$ *with* $f(x) := qx$ *for some* $0 < q < 1$*. Because* f *is linear we are tempted to use the RKBS* $\mathcal{H} = \mathbb{R}$ *(interpreted as an RKBS on a set* $X = \{x\}$ *containing only one single element* x*) with the euclidean inner product. Let us consider the following perturbed system given by*

$$g(x) := \sin(q \arcsin(x)).$$

Differentiating f *and* g *shows that, around the origin,* g *behaves similarly to* f *and since the origin is attracting we could guess that both systems are closely related - which is true but we need to use the right concept of relatedness. First, we note that* $\mathbb{R}$ *does not work as an RKHS for* g *anymore because* g *is not linear and* T_g *will have trivial domain* $D(T_g) = \{0\}$*. This is connected to the problem that the Perron-Frobenius is not well defined because* $\operatorname{Span}\{k(x,\cdot) : x \in X\}$ *is not linearly dependent (due to the finite dimension of* $\mathbb{R}$*). The maps* f *and* g *are conjugated by the function* $\psi : [0,1] \to [0,1]$ *given by* $\psi(x) = \frac{1}{\pi}\arcsin(x))$ *because we have* $\psi \circ g = f \circ \psi$*. Hence, we can apply Lemma 6.9 and define a kernel* k_ψ *by* $k_\psi(x,y) = \pi^{-2}\arcsin(x)\arcsin(y)$*. Thus,* T_g *is well defined and bounded on the corresponding RKHS* $\mathcal{H}_\psi$*.*

In the above example, we have seen a case where the domain of the Koopman operator is trivial. Cases in which the domain of the Koopman operator can be large, even the whole space, but where the Koopman can never be compact are RKHS induced by positive definite maps $u : \mathbb{R}^d \to \mathbb{C}^d$ with kernel $k(x,y) := u(x-y)$ from Example 2.62. In [Ikeda 2022a] it is shown that under certain conditions on the spectral density w of u (see Example 2.62 for the role of w), no composition operator is compact on $\mathcal{H}$.

The class of RKHS induced by positive definite functions includes the class of shift-invariant kernels. This class covers the important and popular example of the Gaussian kernel RKHS.

Example 6.13 (Shift invariant kernels)**.** *A kernel* k *on* $\mathbb{R}^n$ *(or any group) is called shift invariant if for all* $x, y, a \in \mathbb{R}^n$ *we have* $k(x+a, y+a) = k(x,y)$*. Kernels of the form*

$$k(x,y) = h(\|x-y\|) \tag{6.22}$$

for some positive definite function h*, are typical examples for shift-invariant kernels. A function* h *is called positive definite if the corresponding kernel (6.22) is positive definite. For example the Gauss kernel with parameter* $\sigma > 0$ *given by*

$$k(x,y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\|x-y\|^2}{2\sigma^2}} \tag{6.23}$$

is positive definite and the corresponding RKHS is dense in the space of continuous functions on $\mathbb{R}^n$ *that vanish at infinity, see [Sriperumbudur 2011]. In [Ishikawa 2021]*

it was shown that for the Gaussian kernel RKHS, the only dynamics that induce bounded Koopman and Perron-Frobenius operators are affine ones. Nevertheless, Koopman analysis has been successfully applied to forecasting in [Kawahara 2016, Alexander 2020] and system identification in [Rosenfeld 2019] in this setting.

In Examples 2.63 and 2.74, we interpreted Sobolev spaces of high enough regularity as RKBS. We state a simple condition for which diffeomorphisms induce bounded Koopman and Perron-Frobenius operators on these spaces.

Example 6.14 (Sobolev space)**.** *For $\Omega \subset \mathbb{R}^n$ open and bounded with C^1 boundary. For $k \in \mathbb{N}$ and $p \in [2, \infty)$ we denote by $W^{k,p}(\Omega)$ the Sobolev space from Example 2.74. We assume $k > \frac{n}{p}$ so that $\mathcal{B} := \mathrm{W}^{k,p}(\Omega)$ turns into an RKBS with the universal property, see Example 2.63. For simplicity we treat the case $n = k = 1$, $p > 1$ and $\Omega = (0,1)$. In this case, the kernel can be given explicitly, see 2.77, and we can investigate the Perron-Frobenius operator directly K_f. Nevertheless, it is easier to verify that the Koopman operator is bounded. By Theorem 6.7 the boundedness of the Koopman operator is equivalent to $D(T_f) = \mathcal{B}$. We show that for $\Omega = (0,1) \subset \mathbb{R}$ and diffeomorphic $f : [0,1] \to [0,1]$ such that f' and $\frac{1}{f'}$ are bounded we indeed have $D(T_f) = \mathcal{B}$. For $g \in \mathcal{B} \cap \mathcal{C}^1(U)$ we get*

$$\|T_f g\|_{\mathrm{L}^p}^p = \int_0^1 g(f(x))^p \, dx = \int_{f(0)}^{f(1)} g(y)^p \frac{1}{f'(f^{-1}(y))} \, dy \leq \|\frac{1}{f'}\|_\infty \|g\|_{\mathrm{L}^p}^p$$

and from $(T_f g)' = (g \circ f)' = g' \circ f \cdot f'$ we get

$$\begin{aligned} \|(T_f g)'\|_{\mathrm{L}^p}^p &= \int_0^1 g'(f(x))^p f'(x)^p \, dx = \int_{f(0)}^{f(1)} g'(y)^p f'(f^{-1}(y))^{p-1} \, dy \\ &\leq \|f'\|_\infty^{p-1} \|g'\|_{\mathrm{L}^p}^p. \end{aligned}$$

This shows that $T_f|_{\mathcal{H} \cap \mathcal{C}^1(U)} : \mathcal{H} \cap \mathcal{C}^1(U) \to \mathcal{H}$ is a bounded operator. Since $\mathcal{H} \cap \mathcal{C}^1(U)$ is dense in $\mathcal{H}$ we can uniquely extend $T_f|_{\mathcal{H} \cap \mathcal{C}^1(U)}$ to a bounded operator T on $\mathcal{H}$. Using that T_f is closed (see Theorem 6.7 4.) it follows that $T = T_f$, i.e. that T_f is bounded and it follows from Theorem 6.7 7. that $K_f = T_f^*$ is bounded, too. We refer to [Menovschikov 2021] for detailed investigations of composition operators on Sobolev spaces.

6.2 The Koopman and Perron-Frobenius semigroup on RKBS for continuous time systems

In this section, we define the Koopman and Perron-Frobenius semigroups for continuous time dynamical systems $(X, (\varphi_t)_{t \in \mathbb{R}_+})$ – similar to the discrete case. In contrast to the discrete time case, we want to investigate the infinitesimal generator in addition to the semigroup, see Definition 2.48 for the definition of the generator. We relate the generator to the vector field f inducing the dynamical system. Finally,

we give a geometric condition under which the Koopman and Perron-Frobenius semigroup are strongly continuous and consist of bounded operators. In the last part of this chapter, we mention how symmetry and sparsity of the dynamical system are transferred to the Koopman and Perron-Frobenius semigroup based on the treatment of those semigroups on $\mathcal{C}(X)$ in [Salova 2019, Schlosser 2022b]

6.2.1 Koopman and Perron-Frobenius one-parameter semigroup on RKBS and their generators

We begin with the definition of the Koopman and Perron-Frobenius semigroup.

Proposition 6.15. *Let $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS with kernel k and $(T_t)_{t\in\mathbb{R}_+}$ be the Koopman operator for a dynamical system with semiflow φ. Let $\{k(x,\cdot) : x \in X\}$ be linearly independent. Then we define the Perron-Frobenius semigroup of linear operators $(K_t)_{t\in\mathbb{R}_+}$ with $K_t : \operatorname{Span}\{k(x,\cdot) : x \in X\} \to \operatorname{Span}\{k(x,\cdot) : x \in X\}$ for $t \in \mathbb{R}_+$ by linearly extending*

$$K_t k(x,\cdot) = k(\varphi_t(x),\cdot). \tag{6.24}$$

Further $K_t' = T_t$ for $t \in \mathbb{R}_+$.

Proof. For each t the operator T_t coincides with the composition operator T_{φ_t} from Section 6.1. So the result follows from Definition 6.3 and Lemma 2.78. □

Remark 6.16. *As mentioned in the proof of Proposition 6.15 for each $t \in \mathbb{R}_+$ the operator T_t coincides with T_{φ_t} from Section 6.1. Similarly for $K_t = K_{\varphi_t}$. In particular, Theorem 6.7 holds for T_t and K_t for each $t \in \mathbb{R}_+$.*

We should not expect continuity of the map $t \mapsto T_t$ respectively $t \mapsto K_t$ with respect to the operator norms. This is illustrated by our guiding examples of the Koopman semigroup on $\mathcal{C}(X)$ from Example 2.43 and on $\mathrm{L}^2(X,\mu)$ from Example 2.42. As long as the spaces $\mathcal{C}(X)$ and $\mathrm{L}^2(X,\mu)$ are not finite dimensional or the system is not trivial, then $t \mapsto T_t$ is not continuous on $\mathcal{C}(X)$ respectively $\mathrm{L}^2(X,\mu)$ with respect to the operator norm. Further, norm continuity of the semigroup would imply that the generator is bounded, see [Engel 2006, Theorem 2.12] and we have already seen in (2.69) that the generator for the Koopman semigroup is a differential operator and thus unbounded in many situations. We, therefore, begin with addressing strong continuity of the semigroup in Remark 6.17, see Definition 2.44 the definition of strong continuity.

Remark 6.17. *If the map $x \mapsto k(x,\cdot) \in \mathcal{B}'$ is continuous then the Perron-Frobenius semigroup $(K_t)_{t\in\mathbb{R}_+}$ is strongly continuous on $\operatorname{Span}\{k(x,\cdot) : x \in X\}$. This follows because for any $n \in \mathbb{N}$, $a_1,\dots,a_n \in \mathbb{R}$ respectively $\mathbb{C}$ and $x_1,\dots,x_n$ we have*

$$\|(K_t - \mathrm{Id})\left(\sum_{i=1}^n a_i k(x_i,\cdot)\right)\| \quad\leq\quad \sum_{i=1}^n |a_i| \|k(\varphi_t(x_i),\cdot) - k(x_i,\cdot)\|$$

which converges to 0 as $t \to 0$, due to continuity of φ and $x \mapsto k(x,\cdot)$.

For differentiability properties of the orbits of the semigroups we turn to their generators. The generator A of the Koopman semigroup and C of the Perron-Frobenius semigroup are defined by

$$Ah := \lim_{t\to 0} \frac{1}{t}\left(T_t h - h\right) \quad , \quad Cg := \lim_{t\to 0} \frac{1}{t}\left(K_t g - g\right) \tag{6.25}$$

whenever the limit exists.

Remark 6.18. *To treat the Perron-Frobenius semigroup from a semigroup perspective we would like to apply generator theorems, such as the Hille-Yosida theorem [Engel 2006, Theorem 3.5]. Unfortunately, those generator theorems typically require knowledge of the closedness and domain of the generator. In the case of the operator C, the closedness is less accessible. On the contrary closedness of the generator of the Koopman semigroup is known for RKHS, see Remark 6.19, but we lack a priori information about the domain of the generator.*

For the Koopman semigroup, we get, as in (2.69), that A is given by the directional derivative in direction of the vector field f. If we assume that $X \subset \mathbb{R}^n$ is open and that the dynamical system $(X, (\varphi_t)_{t\in\mathbb{R}_+})$ induced by a differential equation $\dot{x} = f(x)$ then, it holds for all $x \in X$ and each $h \in \mathcal{C}^1(X)$ for which Ah is defined

$$\begin{aligned} Ah(x) &= \langle Ah, k(x,\cdot)\rangle = \langle \lim_{t\to 0} \frac{U_t h - h}{t}, k(x,\cdot)\rangle = \lim_{t\to 0} \frac{1}{t}\langle U_t h - h, k(x,\cdot)\rangle \\ &= \lim_{t\to 0} \frac{h(\varphi_t(x)) - h(x)}{t} = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} h(\varphi_t(x)) = Dh(x)\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \varphi_t(x) \\ &= Dh(x) f(x). \end{aligned} \tag{6.26}$$

Remark 6.19. *In the case that Y is an RKHS consisting of continuously differentiable functions the generator A from (6.26) is a closed operator, see [Rosenfeld 2019, Theorem 4.2].*

The asymmetry between the Koopman and Perron-Frobenius semigroup carries over to their generators: The description for the generator C is less explicit on a general element $g \in \mathcal{B}'$, on the other hand, we can certify certain elements belonging to its domain. In [Rosenfeld 2019], for RKHS, it was shwon that certain path-integrals belong to the domain of the infinitesimal generator C of $(K_t)_{t\in\mathbb{R}_+}$. We define those path integrals also for RKBS now.

Definition 6.20. *Let $T > 0$ and $t \mapsto k(\varphi_t(x), \cdot) \in \mathcal{B}'$ be continuous on $[0, T]$. For $x \in X$ and $I_{T,x} \in \mathcal{B}^*$ be defined by*

$$I_{T,x} g := \int_0^T g(\varphi_t(x))\, dt. \tag{6.27}$$

We can identify $I_{T,x}$ with the element in $\mathcal{B}'$ given by

$$b'_{T,x} := \int_0^T k(\varphi_t(x), \cdot)\, dt. \tag{6.28}$$

Remark 6.21. *The continuity assumption in Definition 6.20 is used to guarantee that the (Riemann) integral (6.28) exists. Because φ is continuous by assumption, the continuity of $t \mapsto k(\varphi_t(x), \cdot)$ holds if the feature map $x \mapsto k(x, \cdot)$ continuous. In order to weaken regularity assumptions on the feature map Bochner's theorem on Bochner integrals can be evoked – in our case of RKBS, this would typically require more regularity of the space $\mathcal{B}$ and $\mathcal{B}'$, such as reflexivity for example.*

Assumption 6.22. *We assume that the feature map $x \mapsto k(x, \cdot) \in \mathcal{B}'$ is continuous.*

Now we would like to extend K_t to $I_{T,x}$, as we did in Theorem 6.7 8., by

$$K_t I_{T,x} = \int_0^T k(\varphi_{t+s}(x), \cdot)\, ds. \tag{6.29}$$

In order to guarantee that (6.29) is well defined we will use the universal property.

Assumption 6.23. *We assume that X is compact and $\mathcal{B}$ satisfies the universal property from Definition 2.66, i.e. that $\mathcal{B}$ is dense in $\mathcal{C}(X)$.*

Lemma 6.24. *Under Assumptions 6.1, 6.22 and 6.23, the term $K_t I_{T,x}$ from (6.29) is well defined.*

Proof. We want to argue as in Theorem 6.7 8. Therefore, it suffices to note that $I_{T,x} = \int_X k(y, \cdot)\, d\mu(y)$ for the measure μ given the action $\int_X g\, d\mu = \int_0^T g(\varphi_t(x))\, dt$ for $g \in \mathcal{C}(X)$. The result follows from Theorem 6.7 8. □

Finally, we can show that $I_{T,x}$ is contained in the domain of the generator C of the Perron-Frobenius semigroup following the arguments from the RKHS setting in [Rosenfeld 2020].

Proposition 6.25. *Under Assumptions 6.1–7, for $T > 0$ and $x \in X$, we have $I_{T,x} \in D(C)$, where C denotes the generator of the Perron-Frobenius semigroup. In particular, C is densely defined.*

Proof. By Lemma 6.24 the element $I_{T,x}$ is contained in the domain of K_t for all $t \in \mathbb{R}_+$. For the generator of the Perron-Frobenius semigroup, we get as $t \to 0$

$$\begin{aligned}
\frac{1}{t}\left(K_t I_{T,x} - I_{T,x}\right) &= \frac{1}{t}\left(\int_0^T k(\varphi_{t+s}(x), \cdot) - k(\varphi_s(x), \cdot)\, ds\right) \\
&= \frac{1}{t}\left(\int_T^{T+t} k(\varphi_t(x), \cdot)\, ds - \int_0^t k(\varphi_t(x), \cdot)\, ds\right) \\
&\to k(\varphi_T(x), \cdot) - k(x, \cdot),
\end{aligned}$$

i.e. $I_{T,x} \in D(C)$ and $CI_{T,x} = k(\varphi_T(x), \cdot) - k(x, \cdot)$. To check that C is densely defined let $x \in X$. Since $D(C)$ is a linear subspace we have $\frac{1}{T} I_{T,x} \in D(C)$ for all $T > 0$ and by definition of $I_{T,x}$ we get $k(x, \cdot) = \lim_{T \to 0} \frac{1}{T} I_{T,x} \in \overline{D(C)}$. Hence, C is densely defined. □

Under additional smoothness properties of the kernel, we can even verify that the points $k(x, \cdot)$ belong to $D(C)$. To make this precise we restrict to RKHS $\mathcal{H}$ and fix the following preliminaries. Let $X \subset \mathbb{R}^n$ be open and let the kernel k be $\mathcal{C}^1$. Let $\partial_{x_i} k$ denote the derivative of k with respect to the first variable in direction of the i-th standard basis vector e_i. By [Saitoh 2016, Theorem 2.6] we have $\partial_{x_i} k(x, \cdot) \in \mathcal{H}$ for all $i = 1, \dots, n$ and fixed x, and further

$$\partial_{x_i} k(x, \cdot) = \lim_{h \to 0} \frac{1}{h} (k(x + he_i, \cdot) - k(x, \cdot)) \tag{6.30}$$

converges in $\mathcal{H}$ – in other words the feature map $x \mapsto k(x, \cdot) \in \mathcal{H}$ is $\mathcal{C}^1$. In particular, we get that for fixed x the map $t \mapsto K_t k(x, \cdot) = k(\varphi_t(x), \cdot)$ is continuously differentiable if φ is the flow is differentiable in t. If φ is induced by a differential equation $\dot{x} = f(x)$ for a locally Lipschitz continuous vector field $f = (f_1, \dots, f_n)$ we get $k(x, \cdot) \in D(C)$ and

$$\begin{aligned} Ck(x, \cdot) &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} K_t k(x, \cdot) = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} k(\varphi_t(x), \cdot) \\ &= \sum_{l=1}^{n} \partial_{x_l} k(x, \cdot) f_l(x) \in \operatorname{Span}\{k(x, \cdot) : x \in X\} \subset \mathcal{H}. \end{aligned} \tag{6.31}$$

Remark 6.26. *For the generator A of the Koopman semigroup it is not clear whether $k(x, \cdot)$ is in the domain of A. If so, A acts on $k(x, \cdot)$ by*

$$Ak(x, \cdot) = \sum_{i=1}^{n} \partial_{y_i} k(x, \cdot) f_i(\cdot),$$

where ∂_{y_i} denotes the derivative of k with respect to the second variable in the direction of e_i. Hence, it is not a priori clear whether $Ak(x, \cdot)$ is an element of $\mathcal{H}$.

In Proposition 6.27 we give a geometric condition on the generator C that assures boundedness of the Koopman and Perron-Frobenius semigroup.

Proposition 6.27. *Let $X \subset \mathbb{R}^n$ be open and $\mathcal{H}$ be an RKHS on X with kernel $k \in \mathcal{C}^1(X \times X)$ $(X, (\varphi_t)_{t \in \mathbb{R}_+})$ be a dynamical system induced by a locally Lipschitz vector field $f : X \to \mathbb{R}^n$. Assume Assumption 6.1. For $\omega > 0$ the following are equivalent*

1. *The Koopman semigroup is a strongly continuous semigroup with $\|U_t\| \leq e^{\omega t}$ for all $t \in \mathbb{R}_+$*
2. *The Perron-Frobenius semigroup can be extended to a strongly continuous semigroup $(\bar{K}_t)_{t \in \mathbb{R}_+}$ of bounded operators on $\mathcal{H}$ with $\|\bar{K}_t\| \leq e^{\omega t}$ for all $t \in \mathbb{R}_+$*

3. *For all* $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{R}$ *respectively* $\mathbb{C}$ *and* $x_1, \dots, x_n \in X$ *we have*

$$\mathrm{Re} \sum_{i,j,l} a_i \overline{a}_j f_l(x_i) \partial_{x_l} k(x_i, x_j) \leq \omega \sum_{i,j=1}^{n} a_i \overline{a}_j k(x_i, x_j). \tag{6.32}$$

Proof. Since $T_t = K_t^*$ for all $t \in \mathbb{R}_+$, i.e. the Koopman semigroup is the adjoint semigroup of the Perron-Frobenius semigroup, the strong continuity of one semigroup implies the strong continuity of the other, [Engel 2006, p. 9], because $\mathcal{H}$ is reflexive. In the rest of the proof, we show that 2. and 3. are equivalent. The essential observation is that for $g \in \mathrm{Span}\{k(x, \cdot) : x \in X\}$ we have by (6.31) for all $t \in \mathbb{R}_+$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}} \|K_t g\|^2 &= \frac{\mathrm{d}}{\mathrm{dt}} \langle K_t g, K_t g \rangle = \langle \frac{\mathrm{d}}{\mathrm{dt}} K_t g, K_t g \rangle + \langle K_t g, \frac{\mathrm{d}}{\mathrm{dt}} K_t g \rangle \\ &= \langle C K_t g, K_t \rangle + \langle K_t g, C K_t g \rangle = 2\mathrm{Re} \langle C K_t g, K_t \rangle. \end{aligned} \tag{6.33}$$

Representing g as $g := \sum_{i=1}^{n} a_i k(x_i, \cdot)$ and evaluating in $t = 0$ gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}} \|K_t g\|^2 \big|_{t=0} &= 2\mathrm{Re} \langle \sum_{i=1}^{n} a_i \sum_{l=1}^{n} \partial_{x_l} k(x_i, \cdot) f_l(x_i), \sum_{j=1}^{n} a_j k(x_j, \cdot) \rangle \\ &= 2\mathrm{Re} \sum_{i,j,l} a_i \overline{a}_j f_l(x_i) \partial_{x_l} k(x_i, x_j). \end{aligned} \tag{6.34}$$

Equation (6.34) is the central object connecting 2. and 3. We begin by showing that 3 implies 2. To do so, we first show that $\|K_t g\| \leq e^{\omega t} \|g\|$ for all $g \in \mathrm{Span}\{k(x, \cdot) : x \in X\}$. Condition (6.32) implies for $g = \sum_{i=1}^{n} a_i k(x_i, \cdot)$ by (6.34)

$$\frac{\mathrm{d}}{\mathrm{dt}} \|K_t g\|^2 \big|_{t=0} \leq 2\omega \sum_{i,j=1}^{n} a_i \overline{a}_j k(x_i, x_j) = 2\omega \|g\|^2. \tag{6.35}$$

Because $g \in \mathrm{Span}\{k(x, \cdot) : x \in X\}$ was arbitrary in (6.35), $\mathrm{Span}\{k(x, \cdot) : x \in X\}$ is K_t invariant for all $t \in \mathbb{R}_+$ and $(K_t)_{t \in \mathbb{R}_+}$ is a semigroup we get for the map $u : \mathbb{R}_+ \to \mathbb{R}_+, u(t) := \|K_t g\|^2$ that

$$\dot{u}(t) = \frac{\mathrm{d}}{\mathrm{dt}} \|K_t g\|^2 = \frac{\mathrm{d}}{\mathrm{ds}} \|K_s K_t g\|^2 \big|_{s=0} \leq 2\omega \|K_t g\|^2 = 2\omega u(t).$$

By Gronwall's lemma it follows $u(t) \leq e^{2\omega t} u(0) = e^{2\omega t} \|g\|^2$, i.e. $\|K_t g\| = \sqrt{u(t)} \leq e^{\omega t} \|g\|$ for all $g \in \mathrm{Span}\{k(x, \cdot) : x \in X\}$. That shows $\|K_t\| \leq e^{\omega t}$ on $\mathrm{Span}\{k(x, \cdot) : x \in X\}$. Further, K_t is strongly continuous on $\mathrm{Span}\{k(x, \cdot) : x \in X\}$ by Remark 6.17. Because $\mathrm{Span}\{k(x, \cdot) : x \in X\}$ is dense in $\mathcal{H}$, it follows then that $(K_t)_{t \in \mathbb{R}_+}$ can be extended to a strongly continuous semigroup on $\mathcal{H}$, see [Engel 2006, Proposition 1.3], with the desired growth bound. For the remaining implication, 2. implies 3., we argue similarly. From $\|K_t\| \leq e^{\omega t}$ we get

$$\|K_t g\|^2 \leq e^{2\omega t} \|g\|^2 \tag{6.36}$$

for all $g \in \operatorname{Span}\{k(x,\cdot) : x \in X\}$. Evaluating (6.36) in $t = 0$ we see that both sides coincide. For the derivative with respect to t in $t = 0$ this implies

$$\frac{\mathrm{d}}{\mathrm{dt}}\|K_t g\|^2\big|_{t=0} \leq \frac{\mathrm{d}}{\mathrm{dt}} e^{2\omega t}\|g\|^2\big|_{t=0} = 2\omega\|g\|^2. \tag{6.37}$$

Choosing $g = \sum_{i=1}^{n} a_i k(x_i,\cdot)$ the inequality (6.37) coincides with (6.32) by (6.34). □

Remark 6.28. *In Proposition 6.27, we made strong use of the explicit computation of*

$$\|\sum_{i=1}^{n} a_i k(x_i,\cdot)\|^2 = \sum_{i=1}^{n} a_i \overline{a}_j k(x_i, x_j). \tag{6.38}$$

Such an explicit expression of (6.38) is not available in RKBS in general. Hence, a similar result on RKBS $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ would require further explicit knowledge of expressing the norm in $\mathcal{B}'$ by k.

The condition (6.32) is a geometric condition that connects the dynamics f with the kernel k. For $\omega = 0$ and $n = 1$ and real RKHS $\mathcal{H}$ it resembles a Lyapunov condition and at first only states that $k(\varphi_t(x), x)$ is decreasing in time. Due to the symmetry of k, it follows for all $x \in X$,

$$\frac{\mathrm{d}}{\mathrm{dt}}\Big|_{t=0} k(\varphi_t(x), \varphi_t(x)) = \nabla_x k(x,x) f(x) + \nabla_x k(x,x) f(x) \leq 0$$

which means that $V(x) := k(x,x) = \|k(x,\cdot)\|^2$ is a Lyapunov function. The condition (6.32) for $\omega = 0$ extends this concept to the full RKHS because it states that $\hat{V}(g) := \frac{1}{2}\|g\|^2$ is decaying in time since for all $t \in \mathbb{R}_+$, we have $\frac{\mathrm{d}}{\mathrm{dt}}\hat{V}(K_t g) = \operatorname{Re}\langle C K_t g, K_t g\rangle \leq 0$, that is just a reformulation of the Perron-Frobenius semigroup being contractive on $\mathcal{H}$.

6.2.2 Symmetry and sparsity patterns

In this section, we describe how the symmetry concept from [Salova 2019] for Koopman operators carries over to RKBS. Similarly for the concept of factor systems as in [Eisner 2015, Schlosser 2022b].

Definition 6.29 (Symmetry)**.** *A map $\psi : X \to X$ is called a symmetry for the discrete time dynamical system induced by $f : X \to X$ if $\psi \circ f = f \circ \psi$.*

Remark 6.30. *Typically, the map ψ is assumed to be invertible. In that case, we have $\psi^{-1} \circ f \circ \psi = f$. For continuous time systems $(X, (\varphi_t)_{t\in\mathbb{R}_+})$, symmetry means that $\psi \circ \varphi_t = \varphi_t \circ \psi$ for all $t \in \mathbb{R}_+$. In other words, ψ maps solutions of the dynamical system to, again, solutions of the dynamical system. If the continuous time dynamical system is induced by the differential equation $\dot{x} = f(x)$ then an invertible smooth map $\psi : X \to X$ is a symmetry if $f = (D\psi^{-1} \circ \psi) \cdot (f \circ \psi)$.*

The next proposition states that symmetries induce a commutation relation between the Koopman and Perron-Frobenius operators and their corresponding operators induced by the symmetry map.

Proposition 6.31. *Let $f : X \to X$ be the (discrete) dynamics, $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS on X with kernel k and ψ a symmetry for f. Let T_ψ and K_ψ be the Koopman and Perron-Frobenius operator with respect to ψ on the RKBS. Then, the relation*

$$T_f T_\psi = T_\psi T_f \tag{6.39}$$

holds on the set

$$\{g \in \mathcal{B} : g \in D(T_f) \cap D(T_\psi), T_f g \in D(T_\psi), T_\psi g \in D(T_f)\}$$

and

$$K_\psi K_f = K_f K_\psi \quad \text{on } \operatorname{Span}\{k(x,\cdot) : x \in X\}. \tag{6.40}$$

Proof. This follows directly from the definition of symmetry. We only show it for the Perron-Frobenius operator. For $x \in X$ we have

$$\begin{aligned} K_\psi K_f k(x,\cdot) &= K_\psi k(f(x),\cdot) = k(\psi(f(x)),\cdot) = k(f(\psi(x)),\cdot) = K_f k(\psi(x),\cdot) \\ &= K_f K_\psi k(x,\cdot). \end{aligned}$$

□

The commutation relation (6.39) in Proposition 6.31 is particularly useful when the domains of T_ψ and T_f are known. The easiest case is when both T_ψ and T_f induce bounded operators, i.e. when $D(T_\Psi) = D(T_f) = \mathcal{B}$.

For sparsity, we follow the notion of factor systems, see [Eisner 2015, p. 15] and its application to Koopman operators in [Schlosser 2022b]

Definition 6.32 (Factor system)**.** *Let $f : X \to X$ be a discrete dynamical system on X and $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS. We call a triple (Y, Π, F) a factor system if Y is a set, $\Pi : X \to Y$ and $F : Y \to Y$ such that*

$$\Pi \circ f = F \circ \Pi. \tag{6.41}$$

By similar arguments to the symmetry case, we get the following proposition.

Proposition 6.33. *Let $f : X \to X$ be a discrete dynamical system, (Y, Π, F) be a factor system, $(\mathcal{B}, \mathcal{B}', \langle\cdot,\cdot\rangle, k)$ be an RKBS on X and $(\mathcal{B}_Y, \mathcal{B}_Y', \langle\cdot,\cdot\rangle_Y, k_Y)$ be an RKBS for Y. Let $K_\Pi : \operatorname{Span}\{k(x,\cdot) : x \in X\} \to \operatorname{Span}\{k_Y(y,\cdot) : y \in Y\}$ defined by linear extension of $K_\Pi k(x,\cdot) := k_Y(\Pi(x),\cdot)$. Then*

$$K_\Pi K_f = K_F K_\Pi. \tag{6.42}$$

For the Koopman operators T_f and T_f corresponding to f and F and $T_\Pi : D(T_\Pi) \to \mathcal{B}$ defined by $T_\Pi g := g \circ \Pi$ on $D(T_\Pi) := \{g \in \mathcal{B}_Y : g \circ \Pi \in \mathcal{B}\}$ we have $T_f T_\Pi = T_\Pi T_f$ on

$$\{g \in \mathcal{B} : g \in D(T_\Pi) \cap D(T_f), D(T_\Pi) \in D(T_f), T_f g \in D(T_\Pi)\}. \tag{6.43}$$

Proof. The proof is similar to the proof of Proposition 6.31. For all $x \in X$ we have

$$K_\Pi K_f k(x,\cdot) = K_\Pi k(f(x),\cdot) = k(\Pi(f(x)),\cdot) = k(F(\Pi(x)),\cdot) = K_F K_\Pi k(x,\cdot)$$

and we get (6.42). Similarly for the Koopman operator we have for all g in the set given in (6.43)

$$T_f U_\Pi g = T_f (g \circ \Pi) = g \circ \Pi \circ f = g \circ F \circ \Pi = T_\Pi T_f g. \qquad \square$$

The commutation and intertwining relations in Propositions 6.31 and 6.33, even though being similar, should be interpreted differently. For symmetries the commutation relation (6.40) implies that the operators share eigenspaces - which can be exploited for dynamic mode decomposition as in [Salova 2019, Prashant 2006]. Sparsity on the other hand intends to reduce the dynamical system to another (preferably lower dimensional) one and to exploit the structure of the system on Y computationally, as we did in [Schlosser 2022b].

6.3 Sparse structures for the Koopman and Perron-Frobenius operator

In this section we return to sparsity from Chapter 4 and investigate how it translates to spectral objects of the Koopman semigroup on $\mathcal{C}(X)$. Because we focus on subsystems, we return to dynamical systems on $\mathbb{R}^n$, i.e. we assume that

$$(\mathbb{R}^n, (\varphi_t)_{t\in\mathbb{R}_+}) \tag{6.44}$$

is a dynamical system and we refer to Section 3.1 for the notion of subsystems.

Remark 6.34. *We focus on continuous time systems but the treatment of discrete time dynamical systems is the same – we only have to replace the continuous time objects with their discrete time analogs.*

In this section, we restrict the dynamics to a compact set $X \subset \mathbb{R}^n$. And carry the following assumption in this section.

Assumption 6.35. *X is positively invariant.*

We focus on the Koopman semigroup on $\mathcal{C}(X)$, the space of continuous functions on X, respectively the Perron-Frobenius semigroup on $\mathcal{M}(X)$, the space of Borel measures on X. We recall the Definitions 2.41 and 2.51.

Definition 6.36. *The Koopman semigroup $(T_t)_{t\in\mathbb{R}_+}$ is the family of operators $T_t : \mathcal{C}(X) \to \mathcal{C}(X)$ for $t \in \mathbb{R}_+$ which are given by*

$$T_t g := g \circ \varphi_t. \tag{6.45}$$

The Perron-Frobenius semigroup $(P_t)_{t\in\mathbb{R}_+}$ is the adjoint semigroup of the Koopman semigroup and consists of the operators $P_t : \mathcal{M}(X) \to \mathcal{M}(X)$ for $t \in \mathbb{R}_+$ given by

$$P_t = (\varphi_t)_\#. \tag{6.46}$$

We introduced the Koopman and Perron-Frobenius semigroups in the Preliminary section in Section 2.5. For more of the many inspiring results and applications

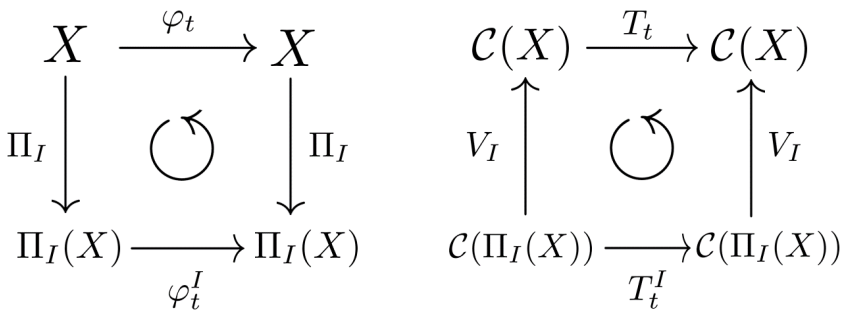

Figure 6.2: Illustration of intertwining between T_t and T_t^I via V_I from (6.48).

of Koopman theory there are many texts to mention, among these are [Budisic 2012, Eisner 2015, Koopman 1931, Küster 2021, Mezić 1999, Kühner 2021, Korda 2018a].

6.3.1 Sparse properties of the Koopman operator induced by subsystems

In the following, we will be most interested in eigenfunctions and eigenmeasures, i.e. eigenvectors of the Koopman and Perron-Frobenius operators respectively.

Definition 6.37. *We say $g \in \mathcal{C}(X)$ respectively $\mu \in \mathcal{M}(X)$ is an eigenfunction respectively eigenmeasure with eigenvalue $\lambda \in \mathbb{C}$ of the Koopman respectively Perron-Frobenius operator if $g \neq 0$ respectively $\mu \neq 0$ and for all $t \in [0, \infty)$ we have $T_t g = e^{\lambda t} g$ and $P_t \mu = e^{\lambda t} \mu$ respectively. For the Perron-Frobenius operator, an eigenmeasure with eigenvalue $\lambda = 0$ is called an invariant measure.*

Eigenfunctions respectively eigenmeasures are the simplest elements corresponding to the spectrum of the two semigroups. They allow to generalize the concept of "diagonalizing" the dynamics or give insight into ergodic properties of the dynamical system, see for example [Budisic 2012, Eisner 2015].

If an index set $I \subset \{1, \dots, n\}$ induces a subsystem then the Koopman semigroup T_t^I for a subsystem acts on $\mathcal{C}(\Pi_I(X))$. The corresponding Perron-Frobenius operator for the subsystem is denoted by P_t^I. In Proposition 6.38 we state that the Koopman respectively Perron-Frobenius operator for the whole systems are intertwined with the corresponding operators for the subsystem. We say an operator $V : W \to Z$ intertwines an operator $S : Z \to Z$ with $T : W \to W$ if

$$SV = VT. \tag{6.47}$$

We will see that operators $V_I : \mathcal{C}(\Pi_I(X)) \to \mathcal{C}(X)$ and $V_I^* : M(X) \to \mathcal{M}(\Pi_I(X))$ given by

$$V_I g := g \circ \Pi_I \text{ and } V_I^* \mu := (\Pi_I)_{\#} \mu \tag{6.48}$$

intertwine T_t and T_t^I respectively P_t^I and P_t for all $t \in \mathbb{R}_+$. That the operator V_I from (6.48) intertwines T_t and T_t^I is a consequence of the functorial nature of the Koopman operators and is illustrated in Figure 6.2.

Proposition 6.38 ([Eisner 2015, p. 208, 233])**.** *Let X be positively invariant and I induce a subsystem. Then*

1. *V_I is injective and intertwines T_t and T_t^I and V_I^* intertwines P_t^I and P_t for all $t \in \mathbb{R}_+$. If X is compact then V_I^* is the dual of V_I and surjective.*
2. *If $g \in \mathcal{C}(\Pi_I(X))$ is an eigenfunction with eigenvalue λ for the Koopman operator T_t^I for the corresponding subsystem then $\hat{g} \in \mathcal{C}(X)$ defined by $\hat{g} := g \circ \Pi_I$ is an eigenfunction with eigenvalue λ of the Koopman operator T_t for the whole system.*
3. *If $\mu \in \mathcal{M}(X)$ is an eigenmeasure with eigenvalue λ of the Perron-Frobenius operator P_t, so is the push forward measure of μ by Π_I, i.e. $\mu_I := (\Pi_I)_\# \mu$, an eigenmeasure with eigenvalue λ for the Perron-Frobenius operator for the subsystem P_t^I.*

The statements 2. and 3. in the following Proposition 6.38 is a direct consequence of the intertwining property 1.

The converse question – constructing eigenfunctions for the subsystem from eigenfunctions for the whole system and analog constructing eigenmeasures for the whole system from eigenmeasures for the subsystem – is less straightforward. We treat that problem in Sections 6.3.2 and 6.3.3 in Theorems 6.39 and 6.44.

6.3.2 Construction of eigenmeasures from eigenmeasures of subsystems

Now, we present that for given invariant measures for subsystems satisfying necessary compatibility conditions we can construct (or glue together) those measures to obtain an invariant measure for the whole system.

Theorem 6.39. *Let $(I_1, f_{I_1}), \ldots, (I_N, f_{I_N})$ induce a subsystem decomposition and assume that X is compact and factors according to $I_1, \ldots, I_N$. For $k = 1, \ldots, N$ let $\mu_k \in \mathcal{M}(\Pi_{I_k}(X))$ be an invariant probability measure for the subsystem induced by I_k. Then there exists an invariant probability measure $\mu \in \mathcal{M}(X)$ such that*

$$(\Pi_{I_k})_\# \mu = \mu_k \text{ for all } k = 1, \ldots, N \tag{6.49}$$

if and only if for all $k, l \in \{1, \ldots, N\}$

$$(\Pi_{I_k \cap I_l})_\# \mu_k = (\Pi_{I_k \cap I_l})_\# \mu_l. \tag{6.50}$$

Proof. Necessity of (6.50) follows from $\Pi_{I_k} \circ \Pi_{I_l} = \Pi_{I_l} \circ \Pi_{I_k} = \Pi_{I_k \cap I_l}$ because for any $k, l \in \{1, \ldots, N\}$ we get

$$\begin{aligned} (\Pi_{I_k \cap I_l})_\# \mu_k &= (\Pi_{I_l} \circ \Pi_{I_k})_\# \mu_k = (\Pi_{I_l})_\# (\Pi_{I_k})_\# \mu_k = (\Pi_{I_l})_\# \mu_k \\ &= (\Pi_{I_l})_\# (\Pi_{I_k})_\# \mu = (\Pi_{I_l} \circ \Pi_{I_k})_\# \mu. \end{aligned}$$

Replacing the roles of k and l and using that Π_{I_k} and Π_{I_l} commute we see that $(\Pi_{I_k \cap I_l})_\# \mu_k = (\Pi_{I_l} \circ \Pi_{I_k})_\# \mu = (\Pi_{I_k \cap I_l})_\# \mu_l$, i.e. (6.49). For the sufficiency part we

consider the set

$$K := \{\mu \in \mathcal{M}(X)_+ : \mu(X) = 1, (\Pi_{I_k})_\# \mu = \mu_k, k = 1, \dots, N\} \tag{6.51}$$

and will show that it is non-empty, convex, compact (with respect to the weak-* topology), and P_t-invariant for all $t \in \mathbb{R}_+$. The result then follows from the Markov-Kakutani theorem [Eisner 2015, Theorem 10.1]. To show that K is non-empty we recall that by [Ambrosio 2013, Lemma 2.1] it is possible to glue two probability measures with coinciding common marginals together "along the marginal". That means for probability measures $\mu \in \mathcal{M}(X \times Y)$ and $\nu \in \mathcal{M}(X \times Z)$ with $(\Pi_X)_\# \mu = (\Pi_X)_\# \nu$ there exists a probability measure $\gamma \in \mathcal{M}(X \times Y \times Z)$ with $(\Pi_{X\times Y})_\# \gamma = \mu$ and $(\Pi_{X\times Z})_\# \gamma = \nu$. The compatibility condition (6.50) guarantees that the common marginals of the measures μ_k coincide and we can apply [Ambrosio 2013, Lemma 2.1] (inductively) to glue together the measures μ_k to a probability measure $\gamma \in M(\mathbb{R}^n)$. Note that the measure γ is not invariant already. A careful look at the proof of [Ambrosio 2013, Lemma 2.1] reveals that the condition that X decomposes according to $I_1, \dots, I_N$ assures that the support of such a glued measure γ is contained in X. Hence, the set K is non-empty. To check convexity and weak-* closedness note that for each $1 \le k \le n$ the operators $(\Pi_{I_k})_\# : \mathcal{M}(X) \to \mathcal{M}(\Pi_{I_k}(X))$ are linear, bounded and continuous with respect to the weak-* topology. It follows that K is convex and weak-* closed. Hence, the set $W = \bigcap_{k=1}^{N} (V^*_{I_k})^{-1}(\{\mu_k\})$ is weak-* closed and convex as the intersection of weak-* closed convex sets. The constraint $\mu(X) = 1$ implies that K is a (closed, convex) subset of the set of probability measures – hence it is compact with respect to the weak-* topology. To check that K is P_t invariant for all $t \in \mathbb{R}_+$ let $\mu \in K$, $t \in \mathbb{R}_+$ and $1 \le k \le N$. Then $P_t\mu(X) = \mu(\varphi_t^{-1}(X)) = \mu(X) = 1$ and

$$\begin{aligned}(\Pi_{I_k})_\# P_t\mu &= (\Pi_{I_k})_\#(\varphi_t)_\#\mu = (\Pi_{I_k} \circ \varphi_t)_\#\mu = (\varphi_t^{I_k} \circ \Pi_{I_k})_\#\mu \\ &= (\varphi_t^{I_k})_\#(\Pi_{I_k})_\#\mu \overset{(6.51)}{=} P_t^I \mu_k = \mu_k\end{aligned}$$

where the last equality holds because μ is an invariant measure for the subsystem, i.e. $P_t^I \mu_k = \mu_k$. That shows invariance of K with respect to P_t for all $t \in \mathbb{R}_+$. Further, for all $t, s \in \mathbb{R}_+$ the operators P_t and P_s commute, due to

$$P_t P_s = P_{t+s} = P_s P_t. \tag{6.52}$$

Because the operators P_t are bounded for all $t \in \mathbb{R}_+$ they are also continuous with respect to the weak-* topology and we can apply the Markov-Kakutani theorem [Eisner 2015, Theorem 10.1] to the family of operators $(P_t)_{t\in\mathbb{R}_+}$ and the set K from (6.51). This gives a measure $\mu \in K$ that satisfies $P_t\mu = \mu$ for all $t \in \mathbb{R}_+$, i.e. μ is an invariant probability measure with the given marginals $(\Pi_{I_k})_\#\mu = \mu_k$ for $k = 1, \dots, N$. □

6.3.3 Eigenfunctions of the Koopman operator based on eigenfunctions of subsystems

For two topological spaces X and Z we have a canonical way of projecting a measure in $\mathcal{M}(X \times Z)$ to measures in $\mathcal{M}(X)$ and $\mathcal{M}(Z)$. For functions $g \in \mathcal{C}(X \times Z)$ it is not so clear how to project g onto $\mathcal{C}(X)$ and $\mathcal{C}(Z)$. The evaluation map $g(\cdot, \cdot) \mapsto g(x_0, \cdot)$ for some $x_0 \in X$ does not send eigenfunctions to eigenfunctions in general. But we will see that the so-called principal eigenvalues have a certain decomposition property. The decomposition of principal eigenfunctions will be based on their uniqueness; we use such uniqueness results from [Kvalheim 2021].

Definition 6.40. *For systems with globally exponentially attractive fixed point x^* we call an eigenfunction $g \in \mathcal{C}^1(X)$ principal eigenfunction for the Koopman operator if $Dg(x^*) \neq 0$.*

The set of eigenfunctions of the Koopman operator can be large. The method from [Korda 2020b] provides a possibility of constructing arbitrarily many eigenfunctions. Therefore, principal eigenfunctions are motivated by the important attempt to single out some very characteristic eigenfunctions. The underlying idea is to find a "basis" of eigenfunctions, in the sense that all other eigenfunctions can be constructed by products and sums of the functions in the "basis". If g and h are eigenfunctions with eigenvalues $\lambda, \theta \neq 0$ then for any $r > 0$ also g^r and $g \cdot h$ are eigenfunction with eigenvalue $r\lambda$ and $\lambda + \theta$. However, if g and h are differentiable and non-constant and $r > 1$ then

$$Dg^r(x^*) = rg^{r-1}(x^*)Dg(x^*) = 0$$

as well as

$$D(g \cdot h)(x^*) = h(x^*)\nabla g(x^*) + g(x^*)\nabla h(x^*) = 0$$

because $g(x^*) = h(x^*) = 0$ for non constant g and h. Thus the condition $\nabla g(x^*) \neq 0$ restricts to eigenfunctions that are not obtained by powers or products of other eigenfunctions.

The vectors $\nabla g(x^*)$ for a principal eigenfunction g are limited to certain values [Kvalheim 2021]. To specify the values $\nabla g(x^*)$ can take we assume from now on that the dynamical system (6.44) is induced by a differential equation

$$\dot{x} = f(x)\ , x(0) = x_0 \in X \tag{6.53}$$

for a $\mathcal{C}^1$ function $f : \mathbb{R}^n \to \mathbb{R}^n$.

For a principal eigenfunction g with eigenvalue λ we have $g(\varphi_t(x)) = e^{\lambda t}g(x)$ and differentiating this relation at $t = 0$ with respect to g gives

$$\nabla g(x) \cdot f(x) = \lambda g(x).$$

Differentiating this relation with respect to x and evaluating in x^* gives

$$\nabla g(x*)^T Df(x^*) = \nabla g(x*)^T Df(x^*) + (D^2 g(x^*))^T f(x^*) = \nabla(\nabla g \cdot f)(x^*) = \lambda \nabla g(x^*)$$

where we have used that $f(x^*) = 0$ because x^* is an equilibrium point. We see that λ is an eigenvalue of $Df(x^*)^T$ and $g(x^*)$ a corresponding eigenvector.

Definition 6.41. *We call a function $g \in \mathcal{C}^1(X)$ a principal eigenfunction of the Koopman operator tangential to a vector $0 \neq v \in \mathbb{R}^n$ if $\nabla g(x^*) = v$.*

The following lemma addresses the eigenvectors of $Df_I(x^*)$ for subsystems induced by I.

Lemma 6.42. *Let I induce a subsystem for f. Then for any $x \in X$*

$$Df(x)^T \cdot D\Pi_I^T = D\Pi_I^T \cdot Df_I(\Pi_I(x))^T \text{ and } D\Pi_I \cdot Df(x) = Df_I(\Pi_I(x))D\Pi_I(x) \tag{6.54}$$

where the derivative $D\Pi_I$ of Π_I is the (constant) matrix with rows consisting of the standard basis vectors $(e_i)_{i\in I}$ and $\cdot$ denotes the matrix product. In particular, if $w = (w_i)_{i\in I}$ is an eigenvector of $Df_I(\Pi_I(x))^T$ with eigenvalue λ then so is $\overline{w}$ for $Df(x)^T$, for

$$\overline{w} := D\Pi_I^T w = \begin{cases} w_k, & k \in I \\ 0, & else \end{cases} \tag{6.55}$$

and if v is an eigenvector of $Df(x)$ with eigenvalue λ and $\Pi_I(v) \neq 0$ then $\Pi_I(v)$ is an eigenvector of $Df_I(\Pi_I(x))$ with eigenvalue λ.

Proof. From the subsystem equation (4.5) it follows from linearity of Π_I

$$D\Pi_I \cdot Df = D(\Pi_I \circ f) = D(f_I \circ \Pi_I) = (Df_I \circ \Pi_I) \cdot D\Pi_I. \tag{6.56}$$

We obtain (6.54) by taking the transpose in (6.56). If w is an eigenvector of $Df_I(x)^T$ with eigenvalue λ then we get

$$Df(x)^T\overline{w} = Df(x)^T \cdot D\Pi_I^T w = \Pi_I^T \cdot Df_I(x)^T w = \lambda\Pi_I^T w = \lambda\overline{w}.$$

If v is an eigenvector of $Df(x)$ with eigenvalue λ we have by (6.54)

$$Df_I(\Pi_I(x))\Pi_I(v) = Df_I(\Pi_I(x)) \cdot D\Pi_I v = D\Pi_I \cdot Df(x)v = \lambda D\Pi_I(v) = \lambda\Pi_I(v).$$

The condition $\Pi_I(v) \neq 0$ guarantees that $\Pi_I(v)$ is an eigenvector of $Df_I(\Pi_I(x)^*)^T$. □

For the uniqueness of eigenfunctions, the concept of resonance is crucial. It is motivated by the property that for eigenfunctions $g_1, \ldots, g_k$ with eigenvalues $\lambda_1, \ldots, \lambda_k$ and $r_1, \ldots, r_k \in \mathbb{N}$ we have $\prod_{i=1}^{k} g_i^{r_i}$ is an eigenfunction with eigenvalue $\sum_{i=1}^{k} r_i\lambda_i$.

Definition 6.43 (Resonance condition; [Mezić 2017] p. 12)**.** *We say a matrix $A \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ (with algebraic multiplicity) is resonant if*

there exists some $i \in \{1, \dots, n\}$ *and* $m_1, \dots, m_n \in \mathbb{N}$ *with* $\sum_{j=1}^{n} m_j \geq 2$ *such that*

$$\lambda_i = \sum_{j \neq i}^{n} m_j \lambda_j.$$

We say A *is resonant of order* k *if* $(m_1, \dots, m_n)$ *can be chosen such that* $\sum_{j=1}^{n} m_j \leq k$.

Non-resonance (of order k), i.e. not being resonant (of order k), and regularity is what is needed to guarantee the existence and uniqueness of principal eigenfunctions [De la Llave 1999, Kvalheim 2021]. Uniqueness allows us to verify that the principal eigenfunction uniquely corresponds to a principal eigenfunction for a subsystem and vice versa.

Theorem 6.44. *Assume there exists a globally exponentially stable fixed point* x^* *and let* I *induce a subsystem. Assume* $Df(x^*)^T$ *is diagonalizable and, for simplicity, that all eigenvalues* $\lambda_1, \dots, \lambda_n$ *have algebraic multiplicity one with corresponding eigenvectors* $v_1, \dots, v_n$. *Assume that the eigenvalues are non-resonant of order* k *with* $k > \max_{i,j} \frac{\mathrm{Re}(\lambda_i)}{\mathrm{Re}(\lambda_j)}$. *Assume* f *is* k*-times continuously differentiable. Then there exist* n *uniquely determined principal eigenfunctions of the whole system tangential to* $v_1, \dots, v_n$ *and exactly* $|I|$ *pairwise distinct principal eigenfunctions for the subsystem, each tangential to one of the vectors* $\Pi_I(v_j)$ *for some* j, *and they induce principal eigenfunctions for the whole system.*

Proof. By [Kvalheim 2021, Proposition 6], the assumptions guarantee the existence and uniqueness of n principal eigenfunctions $g_1, \dots, g_n$ tangential to $v_1, \dots, v_n$. Next, we show that the assumption on non-resonance of $Df(x^*)$ carries over to $Df_I(\Pi_I(x^*))$ and we can use [Kvalheim 2021, Proposition 6] for the subsystem as well. By Lemma 6.42 we get that the spectrum of $Df_I(\Pi_I(x^*))$ is contained in the spectrum of $Df(x^*)$. Further, it also follows from Lemma 6.42 that the geometric multiplicity of each eigenvalue λ for $Df_I(\Pi_I(x^*))$ is at most the geometric multiplicity of λ for $Df(x^*)$, i.e. at most 1. Non-resonance (of order k) of $Df_I(\Pi_I(x^*))$ follows now from non-resonance (of order k) of $Df(x^*)$. For each basis of eigenvectors $w_1, \dots, w_{|I|}$ of $Df_I(\Pi_I(x^*))^T$, we use [Kvalheim 2021, Proposition 6], to guarantee the existence and uniqueness of $|I|$ many principal eigenfunctions $h_1, \dots, h_{|I|}$ tangential $w_1, \dots, w_{|I|}$. It remains to show that each of the $w_1, \dots, w_{|I|}$ can be chosen to be of the form $w_j = \Pi_I(v_{i(j)})$ for $1 \leq j \leq |I|$ and some $1 \leq i(j) \leq n$. Lemma 6.42 states that for $j = 1, \dots, |I|$, the vectors $D\Pi_I^T w_j$ are eigenvectors of $Df(x^*)^T$. From the assumption that all eigenvalues λ of $Df(x^*)$ have algebraic (hence also geometric) multiplicity one, it follows that there exist unique $i(j) \in \{1, \dots, n\}$ and $0 \neq r_j \in \mathbb{R}$ such that $r_j D\Pi_I^T w_j = v_{i(j)}$. That means $r_j h_j$ is a principal eigenfunction tangential to $r_j w_j = \Pi_I(D\Pi_I^T(r_j w_j)) = \Pi_I(v_{i(j)})$. And Proposition 6.38 implies that $\tilde{g}_j := r_j h_j \circ \Pi_I$ is a principal eigenfunction (because $\nabla \tilde{g}_j(x^*) = v_{i(j)} \neq 0$) of the whole system. □

The following corollary addresses the question of whether we can find all the

principle eigenfunctions by searching for them in subsystems. The answer is positive.

Corollary 6.45. *let $(I_1, f_{I_1}), \dots, (I_N, f_{I_N})$ induce a subsystem decomposition. Under the assumptions from Theorem 6.44 any principal eigenfunction for the whole system is already a principal eigenfunction for one of the subsystems.*

Proof. Let g be a principal eigenfunction of the whole system with $w := \nabla g(x^*)$. Then w is an eigenvector of $Df(x^*)^T$ with eigenvalue λ. Hence, λ is also an eigenvalue of $Df(x^*)$. Let v be its corresponding eigenvector. From $\bigcup_{k=1}^{N} I_k = \{1, \dots, n\}$ it follows that for at least one $k \in \{1, \dots, N\}$ we have $\Pi_{I_k}(v) \neq 0$. From Lemma 6.42 we get that λ is an eigenvalue of $Df_{I_k}(\Pi_{I_k}(x^*))$ (with eigenvector $\Pi_{I_k}(v)$). Hence, λ is also an eigenvalue of $Df_{I_k}(\Pi_{I_k}(x^*))^T$. As in the proof of Theorem 6.44, we see that there exists an eigenfunction h with eigenvalue λ for the subsystem induced by I_k such that $\tilde{g} := h \circ \Pi_{I_k}$ is an eigenfunction (with eigenvalue λ) of the whole system. Because we assumed that the eigenvalues are simple, by scaling, we get $\nabla \tilde{g}(x^*) = \nabla g(x^*)$. The uniqueness of principal eigenfunctions implies $\tilde{g} = g$. That shows that g is induced by a principal eigenfunction from a subsystem, namely h. □

Finding eigenfunctions for the subsystems is not answered by Theorem 6.44 and remains a general task (as for finding invariant measures). A partial answer to that question is given in [Korda 2020b] or by the use of Laplace averages for which Proposition 6 and Remark 14 from [Kvalheim 2021] provide a condition under which the Laplace averages exist.

Remark 6.46. *A coordinate-free formulation of the results in this Section and the previous Section 6.3.2 is only partially possible. The coordinate-free notion of subsystems from Section 4.5 allows for generalizing Theorem 6.44 and Corollary 6.45 because we did not really need the fact that the map Π_I is a projection. The situation is different in Theorem 6.39, where we used in [Ambrosio 2013, Lemma 2.1] which is based on a cartesian decomposition of the space.*

Remark 6.47. *We consider smooth factor systems instead of factor systems – where no smoothness is assumed – in order to rule out pathologies for Π as for instance space-filling curves. Since the dimension of the image of a smooth map can not exceed the dimension of the domain (by Sard's theorem, for instance) we see that for smooth factor systems the dimension of Y is necessarily at most the dimension of X. Further smoothness is needed in order to formulate an analog version of Theorem 6.44 where regularity played an essential role.*

6.3.4 Computational applications to dynamic mode decomposition and invariant measures

In the spirit of Section 4.4.2, we show that a priori knowledge of subsystems can be used to reduce computational complexity dynamic mode decomposition (DMD) and of computation of invariant measures.

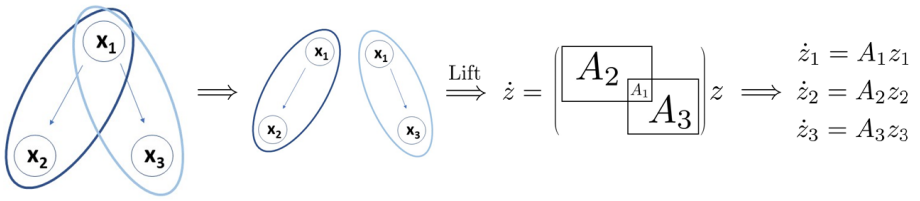

Figure 6.3: Illustration of sparse EDMD: 1. Identification of subsystems, 2. The subsystems induce a block structure for the Koopman operator (lift), 3. Exploitation of the block structure via decoupling of the subsystems.

6.3.4.1 Sparse dynamic mode decomposition

In this section, we describe how extended dynamic mode decomposition (EDMD) ([Schmid 2010]) can benefit from exploiting subsystems, i.e. in this context utilizing knowledge of subsystems. Theorems 6.39 and 6.44 indicate that this allows to still capture (some) important spectral properties of the dynamical system.

The idea is simple: Instead of applying the EDMD to the whole system we use EDMD separately for each of the separate subsystems. This is illustrated by Figure 6.3.

Depending on the choice of the dictionary, i.e. functions $\mathbf{\Psi} = (\psi_1, \dots, \psi_l)$ on which we apply EDMD, this comes with the following advantage

1. In case the number l of dictionary functions is fixed: On a lower dimensional space the same number of dictionary functions allows a better resolution of the geometry of the space. For example, if we choose a dictionary with 1000 radial basis functions (as used to obtain Figures 6.4, 6.5 and 6.6) for the whole state space $\mathbb{R}^6$ as well as for $\mathbb{R}^2$ and respectively $\mathbb{R}^4$ we get a better geometric description of the spaces $\mathbb{R}^2$ and $\mathbb{R}^4$ compared to $\mathbb{R}^6$, allowing a finer EDMD approximation; see Figures 6.4, 6.5 and 6.6.

2. In case the number l of dictionary functions depends on the dimension of the space: Typically l is larger for larger dimensions. This, for instance, is the case when the dictionary consists of (trigonometric) polynomials up to a certain degree. The dimension of the space of polynomials of degree up to d in dimension n is given by $\binom{n+d}{n}$ and hence grows combinatorial in the space dimension n. This relates to the curse of dimensionality and underlines the beneficial impact of lowering the dimension of the space.

This leads to a sparse EDMD stated in Algorithm 6.3.4.1.

Remark 6.48. *For a sparse EDMD based purely on data, without any a priori knowledge of sparsity of the system, we propose to first infer the sparsity graph based on [Granger 1969, Peters 2022] and to use this graph subsequently within Algorithm 6.3.4.1.*

We will illustrate the algorithm at the example of the coupled Duffing equations

Algorithm 4 Sparse dynamic mode decomposition

1: Input: snapshots $Y = [y_1, \dots, y_m], Y^+ = [y_1^+, \dots, y_m^+] \in \mathbb{R}^{n\times m}$ where $y_k^+ = \phi(y_k)$ for $k = 1, \dots, m$ for the underlying (unknown) dynamics ϕ.
2: Determine subsystems $I_1, \dots, I_N$
3: Split the snapshots into snapshots of the subsystems:
For $j = 1, \dots, N$ and $k = 1, \dots, m$ split y_k into $y_k(j) := \Pi_{I_j}(y_k)$ and split y_k^+ into $y_k^+(j) := \Pi_{I_j}(y_k^+)$.
4: Choose dictionaries:
For each $j = 1, \dots, N$ choose a dictionary: $\mathbf{\Psi}^{(j)} = (\psi_1^{(j)}, \dots, \psi_{l_j}^{(j)})^T$.
5: Compute the lifted states: For each $j = 1, \dots, N$ compute

$$Y_{\text{lift}}(j) := [\Psi^{(j)}(y_1(j)), \dots, \Psi^{(j)}(y_m(j))],$$
$$Y_{\text{lift}}^+(j) := [\Psi^{(j)}(y_1^+(j)), \dots, \Psi^{(j)}(y_m^+(j))]$$

6: Compute approximation matrices for the Koopman operators for the subsystems:
For $j = 1, \dots, N$ compute $\mathbf{K}^{(j)}$ for the Koopman operator on the subsystem induced by I_j based on the corresponding snapshots and dictionaries, i.e.

$$\mathbf{K}^{(j)} \in \arg\min_{A\in\mathbb{R}^{l_j\times l_j}} \left\| Y_{\text{lift}}^+(j) - AY_{\text{lift}}(j) \right\|^2 \tag{6.57}$$

7: **return** $\mathbf{K}^{(1)}, \dots, \mathbf{K}^{(N)}$.

(6.58) and (6.59)

$$\begin{aligned} \dot{x}_1^1 &= \frac{1}{2}x_2^1 \\ \dot{x}_2^1 &= -\frac{1}{2}\delta x_2^1 - 2x_1^1(\beta+\alpha)(x_1^1)^2 \end{aligned} \tag{6.58}$$

for $\delta = 0.5$, $\beta = -1$ and $\alpha = 1$ and for x^2, x^3 the dynamics is for $i = 2, 3$ for $\gamma_1 = 1$ and $\gamma_2 = 2$

$$\begin{aligned} \dot{x}_1^i &= \frac{1}{2}x_2^i \\ \dot{x}_2^i &= -\delta x_2^i - 2x_1^i(\beta+\alpha)(x_1^i)^2 + \frac{1}{2}\gamma_i x_1^1. \end{aligned} \tag{6.59}$$

For a simpler notation, we denote the pairs $(x_1^i, x_2^i) \in \mathbb{R}^2$ by $x^i \in \mathbb{R}^2$. Then we use the subsystems induced by $I_1 = \{1\}$, $I_2 = \{1, 2\}$ and $I_3 = \{1, 3\}$ representing the subsystems on the sates $x^1 = (x_1^1, x_2^1)$ and $(x^1, x^i) = (x_1^1, x_2^1, x_1^i, x_2^i)$ for $i = 2, 3$.

For our numerical example we chose $\mathbf{\Psi}$, $\mathbf{\Psi}^{(1)}$, $\mathbf{\Psi}^{(2)}$ and $\mathbf{\Psi}^{(3)}$ to be a dictionary of $l = 1000$ (350 respectively) thin-plate radial basis functions of the form $\phi_i(x) = \|x - c_i\|^2 \log(\|x - c_i\|)$ with uniformly at random sampled centers $c_i \in [-1, 1]^6$, $c_i \in [-1, 1]^4$ and $c_i \in [-1, 1]^2$ for the corresponding subsystems.

The approximations of the Koopman operator for the subsystems can be used for state estimation or computation of (principal) eigenfunctions. This is based on (4.6), Proposition 6.38, and Theorem 6.44. We illustrate state estimation in Figures 6.4, 6.5 and 6.6 based on classical EDMD on the whole system and the sparse EDMD from Algorithm 6.3.4.1 for the coupled Duffing equations (6.58) and (6.59). We use 500 sample trajectories sampled by step size 0.25 for 25 time steps. Figures 6.4, 6.5 and 6.6 display a comparison of the state estimation via EDMD and sparse EDMD. For the initial value $x_0 = (-0.3, -0.3, 0.7, 0.5, 0.3, 0.2)$, we compare estimations of the whole state $x_0(t) = (x_0^1(t), x_0^2(t), x_0^3(t))$ obtained by classical EDMD on the whole system with estimations of the states based on the sparse EDMD. For the sparse approach, we estimate the state $x_0^1(t)$, using only $\mathbf{K}^{(1)}$, the approximation of the Koopman operator on the subsystem induced by $I_1 = \{1\}$, the dictionary $\mathbf{\Psi}^{(1)}$ and snapshots $x^{(1)}(k) = \Pi_{I_1}(x(k)) = x^1(k)$, the state $x_0^2(t)$, using $\mathbf{K}^{(2)}$ obtained from the dictionary $\mathbf{\Psi}^{(2)}$ and snapshots $x^{(2)}(k) = \Pi_{I_2}(x(k)) = (x^1, x^2)$ and similarly for $x_0^3(t)$, using based on the subsystem induced by I_3, for 25 time steps.

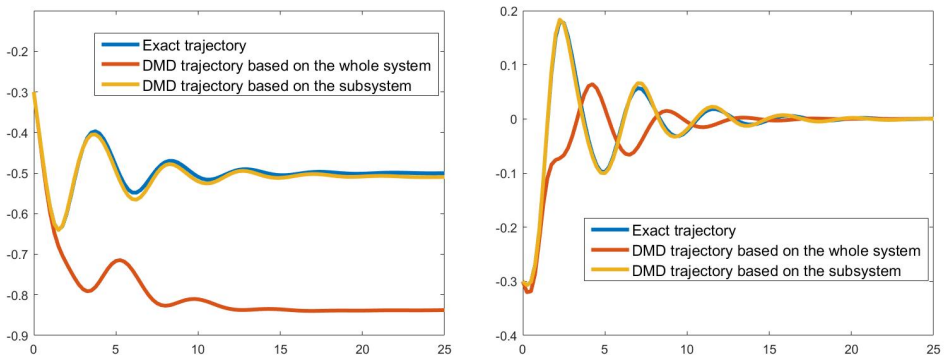

Figure 6.4: DMD approximations based on the whole system and on the subsystem induced by $I = \{1, 2\}$ for (6.58) with initial value $x_0 = (-0.3, -0.3, 0.7, 0.5, 0.3, 0.2)$, trained on the same data. Left: DMD approximation for x_1^1, right: DMD approximation for x_2^1. For the DMD for the whole system, 1000 randomly generated radial basis functions were used while 350 radial basis functions were sufficient for the subsystem.

Eigenfunction computation Related to the sparse EDMD are principal eigenfunctions. Without further restrictions, the system (6.58), (6.59) does not satisfy the assumptions of Theorem 6.44, because the equilibrium point $x^* = (0, 0)$ is not exponentially attracting. Therefore, we want to present a decomposition of the principal eigenfunctions by an example that has the same sparse structure as the systems (6.58), (6.59) but where the principal eigenfunctions can be calculated explicitly.

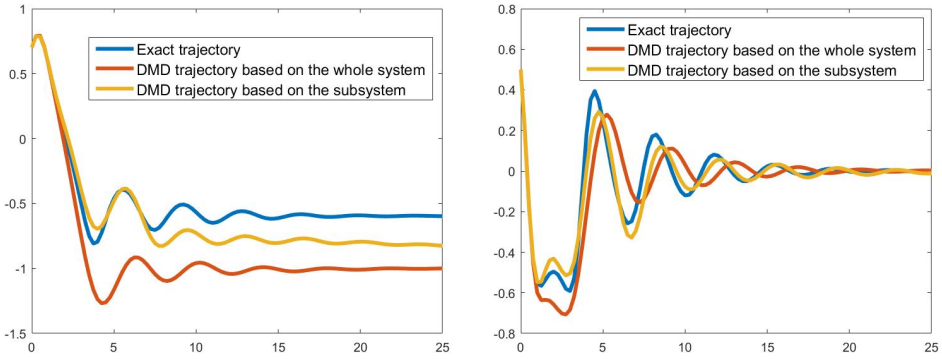

Figure 6.5: DMD approximations based on the whole system and on the subsystem induced by $I = \{1, 2, 3, 4\}$ for (6.58) with initial value $x_0 = (-0.3, -0.3, 0.7, 0.5, 0.3, 0.2)$, trained on the same data with randomly generated 1000 radial basis functions. Left: DMD approximation for x_1^2, right: DMD approximation for x_2^2.

We consider the following system given by

$$\begin{aligned} \dot{x} &= -2x \\ \dot{y} &= -4y - x^2 \\ \dot{z} &= -z + 5x^3 \end{aligned}$$

It has the same sparsity structure as (6.4), (6.59), that is the subsystems act on x, (x, y) and (x, z). The origin is exponentially globally asymptotically stable and hence the assumptions of Theorem 6.44 are satisfied. The subsystem on (x, y) can be found in [Perko 2013] and [Lan 2013] and we find the principal eigenfunctions

$$\begin{aligned} \phi_1(x, y, z) &= x \\ \phi_2(x, y, z) &= y + \frac{1}{4}\ln(x^2)x^2 \\ \phi_3(x, y, z) &= z + x^3. \end{aligned} \tag{6.60}$$

To check that the functions ϕ_1, ϕ_2, ϕ_3 are principal eigenfunctions we note first that they are $\mathcal{C}^1$ and satisfy $\nabla\phi_i(0, 0, 0) \neq 0$ for $i = 1, 2, 3$. That they are eigenfunctions can be checked using the explicit solution to the subsystem on (x, y), see [Lan 2013], or by verifying the generator condition (6.53). In accordance to Theorem 6.44, we see that the function ϕ_1 only depends on x, i.e. arises from the principal eigenfunction of the system $\dot{x} = -2x$, namely from the identity function on $\mathbb{R}$. Similarly, the principal eigenfunctions ϕ_2 and ϕ_3 depend only on (x, y) and (x, z) respectively, and hence arise from principal eigenfunctions for their corresponding subsystems – as it has to be the case, according to Theorem 6.44.

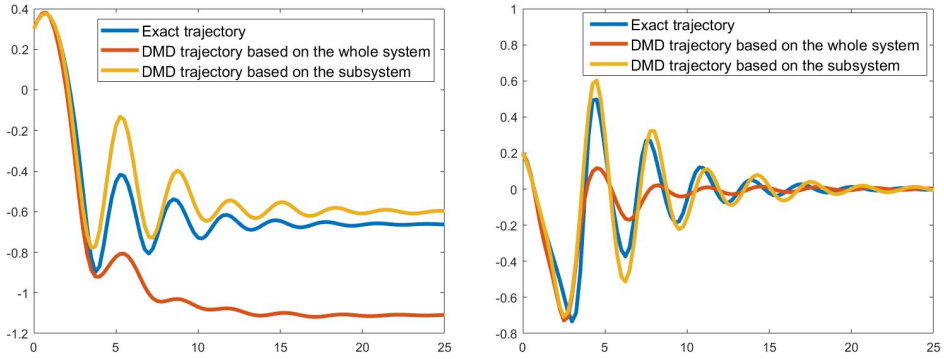

Figure 6.6: DMD approximations based on the whole system and on the subsystem induced by $I = \{1, 2, 5, 6\}$ for (6.58) with initial value $x_0 = (-0.3, -0.3, 0.7, 0.5, 0.3, 0.2)$, trained on the same data with randomly generated 1000 radial basis functions. Left: DMD approximation for x_1^3, right: DMD approximation for x_2^3.

6.3.4.2 Sparse computation of invariant measures

In this section, we propose a sparse computation of invariant measures for the approach from [Magron 2019a] and [Korda 2021] for polynomial dynamical systems based on convex optimization. Before stating the approach we want to shortly present the underlying idea. For this purpose, it is easier to consider discrete time dynamical systems, i.e. $x_{k+1} = f(x_k)$ for some continuous function $f : \mathbb{R}^n \to \mathbb{R}^n$. A measure μ on X is invariant if and only if we have

$$\langle g, \mu \rangle = \int_X g \, d\mu = \int_X g \circ f \, d\mu = \langle g \circ f, \mu \rangle \tag{6.61}$$

for all g in (a dense subset of) $\mathcal{C}(X)$. To follow [Korda 2021] we formulate the following linear optimization problem for extremal invariant probability measures

$$\begin{aligned} p^* \quad := \quad & \min_{\mu} \quad \langle G, \mu \rangle = \int_X G \, d\mu \\ & \text{s.t.} \quad \mu \in \mathcal{M}(X)_+ \\ & \qquad \mu(X) = 1 \\ & \qquad \langle g, \mu \rangle = \langle g \circ f, \mu \rangle \text{ for all } g \in \mathcal{C}(X) \end{aligned} \tag{6.62}$$

Where $G \in \mathcal{C}(X)$ represents a cost and can be used to identify specific invariant measures that we are interested in. Assume $I_1, \ldots, I_N$ induce a subsystem decomposition and the cost G is adapted to the sparse structure – that means G can be written as

$$G = \sum_{k=1}^{N} G_k \qquad \text{for } G_k \in \mathcal{C}(\Pi_{I_k}(X)) \text{ for } k = 1, \ldots, N. \tag{6.63}$$

We formulate corresponding sparse linear programming problems (6.64) and (6.65)

$$\begin{aligned} p_1^* \quad := \quad & \min_{\mu} \quad \int_X G \, d\mu \\ & \text{s.t.} \quad \mu \in \mathcal{M}(X)_+ \\ & \qquad \mu(X) = 1 \\ & \qquad \int_X h \circ f_{I_k} \, d\mu = \int_X h \, d\mu \quad \text{for all } h \in \mathcal{C}(\Pi_{I_k}(X)) \text{ for } k = 1, \dots, N \end{aligned} \tag{6.64}$$

and

$$\begin{aligned} p_2^* \quad := \quad & \min_{\mu_1, \dots, \mu_N} \quad \sum_{k=1}^N \langle G_k, \mu_k \rangle \\ & \text{s.t.} \quad \mu_k \in \mathcal{M}(\Pi_{I_k}(X))_+ \qquad k = 1, \dots, N \\ & \qquad \langle \mathbf{1}, \mu_k \rangle = 1 \qquad k = 1, \dots, N \\ & \qquad \langle h \circ f_{I_k}, \mu_k \rangle = \langle h, \mu_k \rangle \quad \forall h \in \mathcal{C}(\Pi_{I_k}(X)) \text{ for } k = 1, \dots, N \\ & \qquad \langle h, \mu_k \rangle = \langle h, \mu_l \rangle \qquad \forall h \in \mathcal{C}(\Pi_{I_k \cap I_l}(X)) \text{ for } k, l = 1, \dots, N. \end{aligned} \tag{6.65}$$

The linear programming problem (6.64) is clearly a relaxation of (6.62) since invariance is only required for the marginals corresponding to the subsystems. The linear programming problem (6.65) is a reduction to a vector of measures on lower dimensional spaces, and is what we recommend for practical computations, while (6.64) allows more degrees of freedom.

Proposition 6.49. *Let $(I_1, f_{I_1}), \dots, (I_N, f_{I_N})$ induce a subsystem decomposition and for which X decomposes accordingly. Let $G \in \mathcal{C}(X)$ such that G can be written as in (6.63). Then for (6.62) we have $p^* = p_1^* = p_2^*$ from (6.64) and (6.65). Further, there exists an invariant measure μ for the whole system with $p^* = \langle G, \mu \rangle$ such that μ is optimal for (6.64) and $\mu_k := (\Pi_{I_k})_\# \mu$ for $k = 1, \dots, N$ form an optimal (feasible) point for (6.65).*

Proof. First note that the feasible sets for (6.62), (6.64) and (6.65) are weak-* compact and the cost terms are weak-* continuous and therefore minimizers exist. We have already noted that (6.64) is a relaxation of (6.62), i.e. $p_1^* \leq p^*$. We also see that each feasible point μ for (6.64) induces a feasible point $(\mu_1, \dots, \mu_N)$ for (6.65) by $\mu_k = (\Pi_{I_k})_\# \mu$. It follows

$$p_2^* \leq p_1^* \leq p^* \tag{6.66}$$

and it remains to show $p_2^* \geq p^*$. Therefore, let $(\mu_1, \dots, \mu_N)$ be feasible for (6.65). Theorem 6.39 implies that we can find an invariant measure μ for the whole system with corresponding marginals $(\mu_1, \dots, \mu_N)$. Since G is sparse we get

$$p^* \quad \leq \quad \langle G, \mu \rangle = \sum_{k=1}^N \langle G_k \circ \Pi_{I_k}, \mu \rangle = \sum_{k=1}^N \langle G_k, (\Pi_{I_k})_\# \mu \rangle = \sum_{k=1}^N \langle G_k, \mu_k \rangle.$$

Since $(\mu_1, \dots, \mu_N)$ was an arbitrary feasible point for (6.65) we get $p^* \leq p_2^*$. In fact, we have shown that an optimal point $(\mu_1, \dots, \mu_N)$ for (6.65) is induced by an optimal point μ for (6.62). The same measure is also optimal for (6.62). □

Remark 6.50 (Numerical solution of the measure LPs)**.** *In [Schlosser 2022b], we show that the sparse LP (6.65) can be solved via the moment-approach from [Korda 2021, Magron 2019a] leading to a convergent hierarchy of semidefinite programs. That approach is dual to the sum-of-squares hierarchy we have discussed in Section 5.1.3. For details on this duality we refer to [Lasserre 2009, Korda 2021, Magron 2019a]. In [Schlosser 2022b], we also mention that for particle methods the reconstruction of a global measure from (extremal) measures for the subsystem is obtained naturally, see [Schlosser 2022b, Lemma 15.]*

Chapter 7

Perspectives

In Section 4.6, we state limitations of our approach towards sparsity from a theoretical and practical perspective. This raises the question if there is a more general notion of subsystems that allows for decomposing dynamical systems that are less sparse and which extends the list of objects that can be decomposed. A natural way of addressing this task is by trying to make use of information on the dynamical system that remained invisible to our approach – such as the quantitative relation between the states and their dynamics. It is clear, that incorporating quantitative elements of the dynamics into a practical approach is delicate and requires different tools from analysis. There has been done promising work in that direction [Anderson 2012a, Mischaikow 2002, Elkin 2008, Al Maruf 2018]. Therefore, it is intriguing to me to investigate how our approach can be merged with such techniques. In [Wang 2021b] we examined one path in that direction and many more interesting ways are yet to be explored.

Another quantitative extension, for our methods on computing the global attractor from Chapter 5, concerns effectiveness and convergence rates of the proposed numerical method. Work in this direction could build up on [Korda 2017, Korda 2018b] and the recent improvements on the effectiveness of moment-sum-of-squares hierarchy [Baldi 2022]. In this line of reasoning, the LP (5.19) seems more appealing than the LP (5.6). The reason is that Theorem 2.11 induces smooth candidate solutions for (5.19) and convergence analysis can be paired with approximation rates for polynomial approximation of smooth maps.

A qualitative improvement of our results on approximation of the global attractor could aim towards topological properties of the global attractor and its approximations. An open question, that remains challenging for all set approximations via sum-of-squares methods, is the convergence of the approximations with respect to the Hausdorff distance. In this thesis, we only consider the much weaker convergence with respect to Lebesgue measure discrepancy zero. In that situation, many topological properties of the approximated object are not necessarily maintained for its approximations. In contrast to this, convergence with respect to the Hausdorff distance preserves many such properties and is therefore desirable. A possibility to address topological properties in a different way is through inner approximations of the global attractor. Inner approximations of the region of attraction and the maximum positively invariant via sum-of-squares methods have been treated in [Korda 2013, Jones 2021b, Oustry 2019]. However, inner approximations of global attractors by basic semialgebraic sets can be very limited – for instance when the global attractor is the orbit of a non-algebraic curve or when the attractor is a “strange attractor”.

Interesting and challenging tasks are given by extending our computation of the

global attractor to different classes of systems such as partial differential equations and control systems. A promising path for an extension in this direction could follow [Korda 2022, Henrion 2020], where occupation measures are utilized providing approaches toward partial differential equations and applications in control systems. In the case of control systems, the computation of attractors could be used to glimpse at controller design.

In view of the active field of data science, it is intriguing to formulate a data-driven variant of our methods from Chapter 5. A related path was explored in [Korda 2020a] for computing control-invariant sets from data. Pillars of such an approach include sampling theory and data-analysis-based optimization. For the latter, there have been recent beautiful results in [Rudi 2020] that connect closely to sum-of-squares methods.

The field of data science is already well intertwined with Koopman and Perron-Frobenius theory. Currently, the most popular method in data analysis inspired by Koopman theory is dynamic mode decomposition. In our work on Koopman theory on reproducing kernel Banach spaces, we did not investigate the dynamic mode decomposition in reproducing kernel spaces [Kawahara 2016]. But this direction of research is versatile and many open and practically important problems remain. Among these are, first of all, choosing a "good" reproducing kernel space, as well as profound spectral analysis including spectral properties of the dynamic mode decomposition.

When we considered classical extended dynamic mode decomposition for sparse dynamical systems in Section 6.3.4.1, we assumed to know the sparse structure of the dynamical system. Practically, this situation is restrictive because dynamic mode decomposition is often used as a purely data-driven approach without further knowledge of the system. In Section 3.3.2, we discussed existing attempts for exploiting evidence of sparse data. A possible approach, that connects closer to our work on sparse structures of the Koopman operator, is to infer sparsity from the data first, and then apply the sparse dynamic mode decomposition from Section 6.3.4.1. Such a perspective could build on inferring a sparsity graph from data, which has been investigated, for instance, by [Harnack 2017, Paluš 2018, Granger 1969, Peters 2022].

Finally, the numerical examples we performed are scientific ones. I would be happy to test or see if – or hopefully *that* – the methods developed in this work can be beneficial for certain practically relevant problems.

Les grandes personnes sont
bien étranges.

Le petit prince
Antoine de Saint-Exupéry

Les grandes personnes sont
décidément bien bizzares.

Le petit prince
Antoine de Saint-Exupéry

Notation

$\mathbb{Z}$	The integers
$\mathbb{N}$	The positive integers
$\mathbb{N}_0$	The non-negative integers
$[n]$	The set $\{1, \ldots, n\}$ for $n \in \mathbb{N}$
$X \cong Y$	The spaces X and Y are isomorphic
Id	Identity map $\mathrm{Id}(x) := x$
$g \otimes h$	Product of functions g and h defined by $(g \otimes h)(x, y) := (g(x), h(y))$
$\mathbb{R}$	The real numbers
$\mathbb{R}_+$	The non-negative real numbers
$\mathbb{C}$	The complex numbers
$\overline{z}$	The complex conjugate of a complex number z
f^n	The n-fold composition of f with itself, $f^n = \underbrace{f \circ \cdots \circ f}_{n \text{ times}}$
$\mathbb{R}[x]$	The ring of real polynomials
$\deg(p)$	The degree of a polynomial p
x^α	For a multi-index $\alpha \in \mathbb{N}$ denotes the monomial $x^\alpha := (x_1^{\alpha_1}, \ldots, x_n^{\alpha_n})$
$\lvert\alpha\rvert$	The degree $\lvert\alpha\rvert := \alpha_1 + \ldots + \alpha_n$ for a multi-index $\alpha \in \mathbb{N}_0^n$
$\mathbb{R}[x]_d$	The ring of polynomials of degree at most d
Σ	The set of sums-of-squares polynomials, see (2.31)
$\mathcal{Q}(p_1, \ldots, p_m)$	The quadratic module generated by $p_1, \ldots, p_m$, see (2.33)
$\mathrm{Pre}(p_1, \ldots, p_m)$	The preordering generated by $p_1, \ldots, p_m$, see (2.39)
$\mathcal{K}(p_1, \ldots, p_m)$	The set $\{x \in \mathbb{R}^n : p_1(x) \geq 0, \ldots, p_m(x) \geq 0\}$
$\mathbf{1}$	The constant one function, i.e. $\mathbf{1}(x) := 1$ for all x

$\frac{\mathrm{d}}{\mathrm{dt}}$	Time derivative
$\dot{x}$	Time derivative of a curve $x : (a, b) \to \mathbb{R}^n$
$\mathcal{C}(X)$	Space of continuous functions on X
$\mathcal{C}(X)_+$	Space of non-negative continuous functions on X
$\|\cdot\|$	Norm
$\|\cdot\|_2$	Euclidean norm on $\mathbb{R}^n$
$B_r(x)$	Ball of radius r centered at x
$B_r(M)$	Set of all points with distance less than r to M
$\mathbb{S}$	unit sphere in $\mathbb{C}$
$\|\cdot\|_\infty$	Supremum norm (of a function or a vector)
$\|\cdot\|_{\infty,X}$	Supremum norm (of a function) on a set X
$\mathcal{C}^k(X)$	Space of k-times continuously differentiable functions on X
$C^\infty(X)$	space of smooth functions on X
∇	Gradient operator
Dg	Derivative of a function g
$\nabla g \cdot f$	Pointwise euclidean inner product of ∇g and a vector field f
$M(X)$	Space of signed Borel measures on X
$M(X)_+$	Space of (non-negative) Borel measures on X
$\langle g, \mu \rangle$	the integral $\int_X g \, d\mu$ for $g \in \mathcal{C}(X)$ and $\mu \in M(X)$
$h_\#$	Pushforward of a map h
rank	Rank of a matrix
$A \succeq 0$	The matrix A is positive semidefinite
$\overline{X}$	The closure of X (with respect to a given topology)
$\mathring{X}$	The interior of X (with respect to a given topology)
$\mathrm{dist}(A, B)$	one-sided Hausdorff distance between sets A and B (in $\mathbb{R}^n$) given by $\sup_{x \in A} \inf_{y \in B} \|x - y\|_2$
$\|A\|$	Cardinality of a set A

$\mathcal{H}$	(reproducing kernel) Hilbert space
$\langle \cdot, \cdot \rangle$	Inner product, dual pairing or bilinear form
$\mathcal{B}$	(reproducing kernel) Banach space
$\mathcal{B}'$	adjoint space for a reproducing kernel Banach space
$(\mathcal{B}, \mathcal{B}', \langle \cdot, \cdot \rangle, k)$	Reproducing kernel Banach space with kernel (see Definition 2.67)
V^*	Dual space of a normed vector space V
T^*	Adjoint operator of an operator T between two normed vector spaces
$\mathbb{R}^I$	$\prod_{i \in I} \mathbb{R}$ for an index set I
Π_I	Projection from $\mathbb{R}^n$ onto $\mathbb{R}^I$, given by $\Pi_I(x_1, \ldots, x_n) = (x_i)_{i \in I}$
f_I	$\Pi_I \circ f$, see Definition 4.1
(I, f_I)	Subsystem, see Definition 4.1

Bibliography

[Abanin 2017] A. V. Abanin and T. I. Abanina. *Composition operators on Hilbert spaces of entire function.* Russ Math., vol. 61, pages 1–4, 2017. (Cited on pages 86 and 89.)

[Ahmadi 2011] A. A. Ahmadi, M. Krstic and P. A. Parrilo. *A globally asymptotically stable polynomial vector field with no polynomial Lyapunov function.* In 2011 50th IEEE Conference on Decision and Control and European Control Conference, pages 7579–7580. IEEE, 2011. (Cited on page 145.)

[Ahmadi 2018] A. A. Ahmadi and B. El Khadir. *A globally asymptotically stable polynomial vector field with rational coefficients and no local polynomial Lyapunov function.* Systems & Control Letters, vol. 121, pages 50–53, 2018. (Cited on pages 75 and 142.)

[Al Maruf 2018] A. Al Maruf, S. Kundu, E. Yeung and M. Anghel. *Decomposition of Nonlinear Dynamical Networks Via Comparison Systems.* 2018 European Control Conference (ECC), pages 190–196, 2018. (Cited on pages 70, 71 and 185.)

[Alexander 2020] R. Alexander and D. Giannakis. *Operator-theoretic framework for forecasting nonlinear time series with kernel analog techniques.* Physica D: Nonlinear Phenomena, vol. 409, page 132520, 2020. (Cited on pages 85 and 162.)

[Ambrosio 2013] L. Ambrosio and N. Gigli. A user's guide to optimal transport, volume 2062. Springer, Berlin, Heidelberg, 2013. (Cited on pages 173 and 177.)

[Anderson 2010] J. Anderson and A. Papachristodoulou. *Dynamical System Decomposition for Efficient, Sparse Analysis.* 49th IEEE Conference on Decision and Control, Dec. 2010. (Cited on page 71.)

[Anderson 2011a] J. Anderson, Y.-C. Chang and A. Papachristodoulou. *Model decomposition and reduction tools for large-scale networks in systems biology.* Automatica, vol. 47, pages 1165–1174, 06 2011. (Cited on pages 70 and 71.)

[Anderson 2011b] J. Anderson, A. Teixeira, H. Sandberg and A. Papachristodoulou. *Dynamical system decomposition using dissipation inequalities.* Proceedings of the IEEE Conference on Decision and Control, pages 211–216, 12 2011. (Cited on pages 70 and 71.)

[Anderson 2012a] J. Anderson. *Dynamical System Decomposition and Analysis Using Convex Optimization.* PhD thesis, 2012. (Cited on pages 70, 71 and 185.)

[Anderson 2012b] J. Anderson and A. Papachristodoulou. *A Decomposition Technique for Nonlinear Dynamical System Analysis.* IEEE Transactions on Automatic Control, vol. 57, no. 6, pages 1516–1521, 2012. (Cited on page 71.)

[Anosov 1988] D. V. Anosov and V. I. Arnold. Dynamical systems i: ordinary differential equations and smooth dynamical systems. Springer, 1988. (Cited on page 9.)

[ApS 2019] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.* 2019. (Cited on page 145.)

[Arendt 1986] W. Arendt, A. Grabosch, G. Greiner, U. Moustakas, R. Nagel, U. Schlotterbeck, U. Groh, H. Lotz and F. Neubrander. One-parameter semigroups of positive operators, volume 1184. Springer, 1986. (Cited on page 40.)

[Artin 1927] E. Artin. *Über die zerlegung definiter funktionen in quadrate.* In Abhandlungen aus dem mathematischen Seminar der Universität Hamburg, volume 5, pages 100–115. Springer, 1927. (Cited on pages 22 and 24.)

[Baddoo 2021] P.J. Baddoo, B. Herrmann, B.J. McKeon, J.N. Kutz and S.L. Brunton. *Physics-informed dynamic mode decomposition (piDMD)*, 2021. (Cited on pages 93 and 94.)

[Balakrishnan 2021] S. Balakrishnan, A. Hasnain, R. Egbert and E. Yeung. *The Effect of Sensor Fusion on Data-Driven Learning of Koopman Operators*, 2021. (Cited on page 94.)

[Baldi 2022] L. Baldi and B. Mourrain. *On the effective Putinar's Positivstellensatz and moment approximation.* Math. Program., 2022. (Cited on pages 35 and 185.)

[Barvinok 2002] A. Barvinok. A course in convexity, volume 54. American Mathematical Soc., 2002. (Cited on pages 17 and 18.)

[Bhatia 2006] N. P. Bhatia and G. P. Szegö. Dynamical systems: stability theory and applications, volume 35. Springer, 2006. (Cited on pages 9 and 14.)

[Blekherman 2006] G. Blekherman. *There are significantly more nonegative polynomials than sums of squares.* Israel Journal of Mathematics, vol. 153, no. 1, pages 355–380, 2006. (Cited on page 22.)

[Blekherman 2012] G. Blekherman, P. Parrilo and R. R. Thomas. Semidefinite optimization and convex algebraic geometry. SIAM, 2012. (Cited on page 32.)

[Boyd 1997] S. Boyd and L. Vandenberghe. *Semidefinite programming relaxations of non-convex problems in control and combinatorial optimization.* In Communications, Computation, Control, and Signal Processing, pages 279–287. Springer, 1997. (Cited on pages 18, 21 and 32.)

[Boyd 2004] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge university press, 2004. (Cited on page 21.)

[Brézis 2011] H. Brézis. Functional analysis, sobolev spaces and partial differentialzequations. Springer-Verlag New York, 2011. (Cited on pages 48 and 52.)

[Budisic 2012] M. Budisic, R. Mohr and I. Mezić. *Applied Koopmanism.* Chaos, vol. 22, page 047510, 2012. (Cited on pages 37, 85, 93 and 171.)

[Bullo 2019] F. Bullo. Lectures on network systems, volume 1. Kindle Direct Publishing Santa Barbara, CA, 2019. (Cited on page 58.)

[Carswell 2003] B. Carswell, B. D. MacCluer and A. Schuster. *Composition operators on the Fock space.* Acta Sci. Math., vol. 69, pages 871–887, 2003. (Cited on page 89.)

[Chacon 2007] G. A. Chacon, G. R. Chacon and J. Gimenez. *Composition operators on spaces of entire functions.* Proc. Amer. Math. Soc., vol. 135, pages 2205–2218, 2007. (Cited on page 89.)

[Chakravorty 2001] M. Chakravorty and D. Das. *Voltage stability analysis of radial distribution networks.* International Journal of Electrical Power & Energy Systems, vol. 23, no. 2, pages 129–135, 2001. (Cited on pages 57 and 68.)

[Chen 2018] M. Chen, S. L. Herbert, M. S. Vashishtha, S. Bansal and C. J. Tomlin. *Decomposition of reachable sets and tubes for a class of nonlinear systems.* IEEE Transactions on Automatic Control, vol. 63, no. 11, pages 3675–3688, 2018. (Cited on pages 59, 60, 63 and 95.)

[Çömez 2021] D. Çömez. *Modern Ergodic Theory: From a Physics Hypothesis to a Mathematical Theory with Transformative Interdisciplinary Impact.* In Handbook of the Mathematics of the Arts and Sciences, pages 1969–1992. Springer, 2021. (Cited on page 37.)

[Cormen 2022] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein. Introduction to algorithms. MIT press, 2022. (Cited on pages 119 and 122.)

[Cowen 1995] C. Cowen and B. Maccluer. Composition operators on spaces of analytic functions. CRC press, 1995. (Cited on pages 86, 89, 152 and 160.)

[Das 2018] S. Das and D. Giannakis. *Koopman spectra in Reproducing kernel Hilbert spaces.* Applied and Computational Harmonic Analysis, vol. 49, no. 2, pages 573–607, 2018. (Cited on pages 85 and 86.)

[Das 2019] S. Das and D. Giannakis. *Delay-Coordinate Maps and the Spectra of Koopman Operators.* Journal of Statistical Physics, vol. 175, no. 4, pages 1107—-1145, 2019. (Cited on page 159.)

[Das 2021] S. Das, D. Giannakis and J. Slawinska. *Reproducing kernel Hilbert space compactification of unitary evolution groups.* Applied and Computational Harmonic Analysis, vol. 54, pages 75–136, 2021. (Cited on page 86.)

[Dashkovskiy 2011] S. Dashkovskiy, H. Ito and F. Wirth. *On a Small Gain Theorem for ISS Networks in Dissipative Lyapunov Form.* European Journal of Control, vol. 17, 07 2011. (Cited on page 71.)

[De la Llave 1999] R. De la Llave and R. Obaya. *Regularity Of The Composition Operator In Spaces Of Hölder Functions*. Discrete and Continuous Dynamical Systems, vol. 5, page 2998, 1999. (Cited on page 176.)

[Dellnitz 2001] M. Dellnitz, G. Froyland and O. Junge. *The algorithms behind GAIO—Set oriented numerical methods for dynamical systems*. In Ergodic theory, analysis, and efficient simulation of dynamical systems, pages 145–174. Springer, 2001. (Cited on page 73.)

[Dellnitz 2002] M. Dellnitz and O. Junge. *Set Oriented Numerical Methods for Dynamical Systems. Handbook of Dynamical Systems II: Towards Applications*, 2002. (Cited on page 73.)

[Dette 2021] H. Dette and A. A. Zhigljavsky. *Reproducing kernel Hilbert spaces, polynomials, and the classical moment problem*. SIAM/ASA Journal on Uncertainty Quantification, vol. 9, no. 4, pages 1589–1614, 2021. (Cited on page 47.)

[Diestel 2017] R. Diestel. Graph theory. Springer Publishing Company, Incorporated, 5th édition, 2017. (Cited on page 120.)

[Doan 2017] M.-L. Doan, L.-H. Khoi and T. Le. *Composition operators on Hilbert spaces of entire functions of several variables*. Integral Equ. Oper. Theory, vol. 88, no. 3, pages 301–330, 2017. (Cited on page 89.)

[Eisner 2015] T. Eisner, B. Farkas, M. Haase and R. Nagel. Operator theoretic aspects of ergodic theory. Berlin, Heidelberg (Springer), 2015. (Cited on pages 37, 38, 40, 85, 92, 97, 98, 99, 114, 168, 169, 171, 172 and 173.)

[Elkin 2008] V. I. Elkin. *Subsystems on Nonlinear Control Dynamical Systems*. Doklady Mathematics, vol. 78, no. 2, pages 804—-806, 2008. (Cited on pages 71 and 185.)

[Elkin 2012] V. I. Elkin. Reduction of nonlinear control systems: A differential geometric approach. Kluwer Academic Publishers, 2012. (Cited on page 71.)

[Elstrodt 1996] J. Elstrodt. Maß-und integrationstheorie, volume 7. Springer, 1996. (Cited on page 16.)

[Engel 2006] K.-J. Engel and R. Nagel. A short course on operator semigroups. Springer Science & Business Media, 2006. (Cited on pages 5, 39, 40, 41, 132, 163, 164 and 167.)

[Fantuzzi 2020] G. Fantuzzi and D. Goluskin. *Bounding extreme events in nonlinear dynamics using convex optimization*. SIAM Journal on Applied Dynamical Systems, vol. 19, no. 3, pages 1823–1864, 2020. (Cited on page 82.)

[Fridman 2014] E. Fridman. Introduction to time-delay systems. Birkhäuser, 2014. (Cited on page 124.)

[Froyland 2021] G. Froyland, D. Giannakis, B. R. Lintner, M. Pike and J. Slawinska. *Spectral analysis of climate dynamics with operator-theoretic approaches*. Nature communications, vol. 12, no. 1, pages 1–21, 2021. (Cited on pages 37 and 86.)

[Giesl 2015] P. Giesl and S. Hafstein. *Review on computational methods for Lyapunov functions*. Discrete & Continuous Dynamical Systems-B, vol. 20, no. 8, page 2291, 2015. (Cited on page 73.)

[Goedel 2012] R. Goedel, R. G. Sanfelice and A. Teel. *Hybrid dynamical systems: modeling stability, and robustness*, 2012. (Cited on page 127.)

[Goluskin 2018] D. Goluskin. *Bounding extreme values on attractors using sum-of-squares optimization, with application to the Lorenz attractor*. arXiv preprint arXiv:1807.09814, 2018. (Cited on pages 73, 82 and 142.)

[Goluskin 2019] D. Goluskin and G. Fantuzzi. *Bounds on mean energy in the Kuramoto–Sivashinsky equation computed using semidefinite programming*. Nonlinearity, vol. 32, no. 5, page 1705, 2019. (Cited on page 73.)

[Goluskin 2020] D. Goluskin. *Bounding extrema over global attractors using polynomial optimisation*. Nonlinearity, vol. 33, no. 9, page 4878, 2020. (Cited on pages 73 and 82.)

[Granger 1969] C. Granger. *Investigating causal relations by econometric models and cross-spectral methods*. Econometrica: journal of the Econometric Society, pages 424–438, 1969. (Cited on pages 57, 178 and 186.)

[Harnack 2017] D. Harnack, E. Laminski, M. Schünemann and K. R. Pawelzik. *Topological causality in dynamical systems*. Physical review letters, vol. 119, no. 9, page 098301, 2017. (Cited on page 186.)

[Henrion 2009] D. Henrion, J.-B. Lasserre and J. Löfberg. *GloptiPoly 3: moments, optimization and semidefinite programming*. Optimization Methods & Software, vol. 24, no. 4-5, pages 761–779, 2009. (Cited on page 137.)

[Henrion 2013] D. Henrion and M. Korda. *Convex computation of the region of attraction of polynomial control systems*. IEEE Transactions on Automatic Control, vol. 59, no. 2, pages 297–312, 2013. (Cited on pages 73, 78, 82, 131, 139, 140 and 147.)

[Henrion 2020] D. Henrion, M. Korda and J.-B. Lasserre. Moment-sos hierarchy, the: Lectures in probability, statistics, computational geometry, control and nonlinear pdes, volume 4. World Scientific, 2020. (Cited on page 186.)

[Hilbert 1888] D. Hilbert. *Über die darstellung definiter formen als summe von formenquadraten*. Mathematische Annalen, vol. 32, no. 3, pages 342–350, 1888. (Cited on pages 22 and 25.)

[Ikeda 2022a] M. Ikeda, I. Ishikawa and Y. Sawano. *Composition operators on reproducing kernel Hilbert spaces with analytic positive definite functions.* Journal of Mathematical Analysis and Applications, vol. 511, no. 1, page 126048, 2022. (Cited on pages 86, 89 and 161.)

[Ikeda 2022b] M. Ikeda, I. Ishikawa and C. Schlosser. *Koopman and Perron–Frobenius operators on reproducing kernel Banach spaces.* Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 32, page 123143, 2022. (Cited on pages 55, 84, 89, 151, 155 and 159.)

[Ishikawa 2018] I. Ishikawa, K. Fujii, M. Ikeda, Y. Hashimoto and Y. Kawahara. *Metric on Nonlinear Dynamical Systems with Perron–Frobenius Operators.* Adv. Neural Inf. Process. Syst., vol. 31, pages 911—-919, 2018. (Cited on page 85.)

[Ishikawa 2021] I. Ishikawa. *Bounded weighted composition operators on functional quasi-Banach spaces and stability of dynamical systems*, 2021. (Cited on pages 89, 160 and 161.)

[Jones 2021a] M. Jones and M. M. Peet. *A converse sum of squares Lyapunov function for outer approximation of minimal attractor sets of nonlinear systems*, 2021. (Cited on pages 72, 73, 75, 76, 77, 80, 81, 140, 142, 144 and 145.)

[Jones 2021b] M. Jones and M. M. Peet. *Converse Lyapunov functions and converging inner approximations to maximal regions of attraction of nonlinear systems.* arXiv preprint arXiv:2103.12825, 2021. (Cited on page 185.)

[Jovanović 2014] M. R. Jovanović, P. J. Schmid and J. W. Nichols. *Sparsity-promoting dynamic mode decomposition.* Physics of Fluids, vol. 26, no. 2, page 024103, 2014. (Cited on page 94.)

[Katznelson 2004] Y. Katznelson. An introduction to harmonic analysis. Cambridge University Press, 2004. (Cited on page 48.)

[Kawahara 2016] Y. Kawahara. *Dynamic Mode Decomposition with Reproducing Kernels for Koopman Spectral Analysis.* Adv. Neural Inf. Process. Syst., vol. 29, 2016. (Cited on pages 85, 87, 162 and 186.)

[Klus 2020] S. Klus, I. Schuster and K. Muandet. *Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces.* Journal of Nonlinear Science, vol. 30, no. 1, pages 283–315, 2020. (Cited on pages 85 and 154.)

[Koopman 1931] B. O. Koopman. *Hamiltonian systems and transformation in Hilbert space.* Proceedings of the National Academy of Sciences, vol. 17, no. 5, pages 315–318, 1931. (Cited on pages 37 and 171.)

[Korda 2013] Milan Korda, Didier Henrion and Colin N Jones. *Inner approximations of the region of attraction for polynomial dynamical systems.* IFAC Proceedings Volumes, vol. 46, no. 23, pages 534–539, 2013. (Cited on page 185.)

[Korda 2014] M. Korda, D. Henrion and C. Jones. *Convex computation of the maximum controlled invariant set for polynomial control systems.* SIAM Journal on Control and Optimization, vol. 52, no. 5, pages 2944–2969, 2014. (Cited on pages 73, 74, 78, 82, 131, 132, 133, 134, 135, 139, 140 and 147.)

[Korda 2017] Milan Korda, Didier Henrion and Colin N Jones. *Convergence rates of moment-sum-of-squares hierarchies for optimal control problems.* Systems & Control Letters, vol. 100, pages 1–5, 2017. (Cited on page 185.)

[Korda 2018a] M. Korda and I. I. Mezić. *Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control.* Automatica, vol. 93, pages 149–160, 2018. (Cited on pages 85 and 171.)

[Korda 2018b] Milan Korda and Didier Henrion. *Convergence rates of moment-sum-of-squares hierarchies for volume approximation of semialgebraic sets.* Optimization Letters, vol. 12, pages 435–442, 2018. (Cited on page 185.)

[Korda 2020a] M. Korda. *Computing controlled invariant sets from data using convex optimization.* SIAM Journal on Control and Optimization, vol. 58, no. 5, pages 2871–2899, 2020. (Cited on page 186.)

[Korda 2020b] M. Korda and I. Mezić. *Optimal Construction of Koopman Eigenfunctions for Prediction and Control.* IEEE Transactions on Automatic Control, vol. 65, pages 5114–5129, 2020. (Cited on pages 174 and 177.)

[Korda 2021] M. Korda, D. Henrion and I. Mezić. *Convex Computation of Extremal Invariant Measures of Nonlinear Dynamical Systems and Markov Processes.* Journal of Nonlinear Science, vol. 31, no. 1, 2 2021. (Cited on pages 182 and 184.)

[Korda 2022] M. Korda and R. Rios-Zertuche. *The gap between a variational problem and its occupation measure relaxation.* arXiv preprint arXiv:2205.14132, 2022. (Cited on pages 73 and 186.)

[Küster 2015] K. Küster. *The Koopman Linearization of Dynamical Systems.* 2015. (Cited on page 85.)

[Küster 2021] K. Küster. *Decompositions of dynamical systems induced by the Koopman operator.* Analysis Mathematica, vol. 47, no. 1, pages 149–173, 2021. (Cited on pages 13 and 171.)

[Kutz 2016] J. N. Kutz, S. L. Brunton, B. W. Brunton and J. Proctor. Dynamic mode decomposition: Data-driven modeling of complex systems. SIAM, 2016. (Cited on page 94.)

[Kvalheim 2021] M. D. Kvalheim and S. Revzen. *Existence and uniqueness of global Koopman eigenfunctions for stable fixed points and periodic orbits.* Physica D, vol. 425, page 132959, 2021. (Cited on pages 85, 93, 174, 176 and 177.)

[Kwee 2018] A. T. Kwee, M.-F. Chiang, P. K. Prasetyo and E.-P. Lim. *Traffic-cascade: Mining and visualizing lifecycles of traffic congestion events using public bus trajectories.* In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pages 1955–1958, 2018. (Cited on page 58.)

[Kühner 2021] V. Kühner. *What can Koopmanism do for attractors in dynamical systems?* The Journal of Analysis, vol. 29, pages 449—-471, 2021. (Cited on pages 12 and 171.)

[Lan 2013] Y. Lan and I. Mezić. *Linearization in the large of nonlinear systems and Koopman operator spectrum.* Physica D: Nonlinear Phenomena, vol. 242, no. 1, pages 42–53, 2013. (Cited on page 181.)

[Lasserre 2001] J.-B. Lasserre. *Global optimization with polynomials and the problem of moments.* SIAM Journal on optimization, vol. 11, no. 3, pages 796–817, 2001. (Cited on pages 20, 21, 31 and 36.)

[Lasserre 2008] J.-B. Lasserre, D. Henrion, C. Prieur and E. Trélat. *Nonlinear optimal control via occupation measures and LMI relaxations.* SIAM Journal on Control and Optimization, vol. 47, pages 1643–1666, 2008. (Cited on pages 73 and 131.)

[Lasserre 2009] J.-B. Lasserre. Moments, positive polynomials and their applications, volume 1. World Scientific, 2009. (Cited on pages 20, 29, 31, 34, 78, 137 and 184.)

[Lasserre 2015] J.-B. Lasserre. An introduction to polynomial and semi-algebraic optimization, volume 52. Cambridge University Press, 2015. (Cited on pages 31, 34, 78 and 82.)

[Lee 2003] J. M. Lee. *Smooth Maps.* In Introduction to Smooth Manifolds, pages 30–59. Springer, 2003. (Cited on pages 11 and 134.)

[Lee 2013] J. M. Lee. *Smooth manifolds.* Springer, 2013. (Cited on page 143.)

[Li 2022] M. Li, X. Zhou, Y. Wang, L. Jia and M. An. *Modelling cascade dynamics of passenger flow congestion in urban rail transit network induced by train delay.* Alexandria Engineering Journal, vol. 61, no. 11, pages 8797–8807, 2022. (Cited on page 58.)

[Lin 2022] R. Lin, H. Zhang and J. Zhang. *On reproducing kernel Banach spaces: Generic definitions and unified framework of constructions.* Acta Mathematica Sinica, English Series, vol. 38, no. 8, pages 1459–1483, 2022. (Cited on pages 48, 49, 50, 86 and 152.)

[Liu 2019] D. Liu, S. Guo, P. Liu, L. Xiong, H. Zou, J. Tian, Y. Zeng, Y. Shen and J. Zhang. *Optimisation of water-energy nexus based on its diagram in cascade reservoir system.* Journal of Hydrology, vol. 569, pages 347–358, 2019. (Cited on page 68.)

[Liu 2020] Shenyu Liu, Daniel Liberzon and Vadim Zharnitsky. *Almost Lyapunov functions for nonlinear systems*. Automatica, vol. 113, page 108758, 2020. (Cited on page 76.)

[Löfberg 2004] J. Löfberg. *YALMIP: A toolbox for modeling and optimization in MATLAB*. In 2004 IEEE international conference on robotics and automation (IEEE Cat. No. 04CH37508), pages 284–289. IEEE, 2004. (Cited on pages 137, 145 and 148.)

[Magron 2019a] V. Magron, M. Forets and D. Henrion. *Semidefinite approximations of invariant measures for polynomial system*. American Institute of Mathematical sciences, vol. 24, no. 12, pages 6745–6770, 2019. (Cited on pages 182 and 184.)

[Magron 2019b] V. Magron, P.-L. Garoche, D. Henrion and X. Thirioux. *Semidefinite approximations of reachable sets for discrete-time polynomial systems*. SIAM Journal on Control and Optimization, vol. 57, no. 4, pages 2799–2820, 2019. (Cited on page 73.)

[Marshall 2008] M. Marshall. Positive polynomials and sums of squares. American Mathematical Soc., 2008. (Cited on pages 16, 20, 22, 24, 25 and 28.)

[Marx 2018] S. Marx, T. Weisser, D. Henrion and J.-B. Lasserre. *A moment approach for entropy solutions to nonlinear hyperbolic PDEs*. arXiv preprint arXiv:1807.02306, 2018. (Cited on page 73.)

[Meiss 2007] J. D. Meiss. Differential dynamical systems. SIAM, 2007. (Cited on pages 9 and 11.)

[Menovschikov 2021] A. Menovschikov and A. Ukhlov. *Composition operators on Sobolev spaces, Q-mappings and weighted Sobolev inequalities*. arXiv preprint arXiv:2110.09261, 2021. (Cited on page 162.)

[Mezić 1999] I. Mezić and S. Wiggins. *A method for visualization of invariant sets of dynamical systems based on the ergodic partition*. Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 9, no. 1, pages 213–218, 1999. (Cited on pages 13, 37, 85 and 171.)

[Mezić 2005] I. Mezić. *Spectral properties of dynamical systems, model reduction and decompositions*. Nonlinear Dynamics, vol. 41, no. 1, pages 309–325, 2005. (Cited on pages 13 and 85.)

[Mezić 2017] I. Mezić. *Koopman Operator Spectrum and Data Analysis*, 2017. (Cited on page 175.)

[Miller 2021] J. Miller, D. Henrion, M. Sznaier and M. Korda. *Peak Estimation for Uncertain and Switched Systems*. In 2021 60th IEEE Conference on Decision and Control (CDC), pages 3222–3228. IEEE, 2021. (Cited on page 73.)

[Mischaikow 2002] K. Mischaikow. *Topological techniques for efficient rigorous computation in dynamics.* Acta Numerica, vol. 11, pages 435–477, 2002. (Cited on pages 71 and 185.)

[Mohr 2020a] R. Mohr and I. Mezić. *Koopman Spectrum and Stability of Cascaded Dynamical Systems.* In The Koopman Operator in Systems and Control, pages 99–129. Springer, 2020. (Cited on page 68.)

[Mohr 2020b] R. Mohr and I. Mezić. *Koopman Spectrum and Stability of Cascaded Dynamical Systems.* In The Koopman Operator in Systems and Control, pages 99–129. Springer, 2020. (Cited on page 91.)

[Motzkin 1967] T. S. Motzkin. *The arithmetic-geometric inequality.* Inequalities (Proc. Sympos. Wright-Patterson Air Force Base, Ohio, 1965), pages 205–224, 1967. (Cited on page 22.)

[Øksendal 2003] B. Øksendal. Stochastic differential equations. Springer-Verlag Berlin Heidelberg, 2003. (Cited on page 126.)

[Oustry 2019] A. Oustry, M. Tacchi and D. Henrion. *Inner approximations of the maximal positively invariant set for polynomial dynamical systems.* IEEE Control Systems Letters, vol. 3, no. 3, pages 733–738, 2019. (Cited on page 185.)

[Paluš 2018] M. Paluš, A. Krakovská, J. Jakubík and M. Chvosteková. *Causality, dynamical systems and the arrow of time.* Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 28, no. 7, page 075307, 2018. (Cited on page 186.)

[Pan 2021] S. Pan, N. Arnold-Medabalimi and K. Duraisamy. *Sparsity-promoting algorithms for the discovery of informative Koopman-invariant subspaces.* Journal of Fluid Mechanics, vol. 917, page A18, 2021. (Cited on page 94.)

[Parrilo 2000] P. Parrilo. Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. California Institute of Technology, 2000. (Cited on pages 15, 73 and 137.)

[Paulsen 2016] V. I. Paulsen and M. Raghupathi. An introduction to the theory of reproducing kernel hilbert spaces, volume 152. Cambridge university press, 2016. (Cited on pages 43, 45, 46, 47 and 154.)

[Perko 2013] L. Perko. Differential equations and dynamical systems, volume 7. Springer Science & Business Media, 2013. (Cited on pages 9, 11 and 181.)

[Peters 2022] J. Peters, S. Bauer and N. Pfister. *Causal models for dynamical systems.* In Probabilistic and Causal Inference: The Works of Judea Pearl, pages 671–690. ACM, 2022. (Cited on pages 57, 178 and 186.)

[Prajna 2004] S. Prajna, P. Parrilo and A. Rantzer. *Nonlinear control synthesis by convex optimization.* IEEE Transactions on Automatic Control, vol. 49, no. 2, pages 310–314, 2004. (Cited on page 73.)

[Prashant 2006] G. Prashant, M. Hessel-von Molo, M. and M. Dellnitz. *Symmetry of attractors and the Perron-Frobenius operator*. Journal of Difference Equations and Applications, vol. 12, no. 11, pages 1147–1178, 2006. (Cited on page 170.)

[Prestel 2013] A. Prestel and C. Delzell. Positive polynomials: from hilbert's 17th problem to real algebra. Springer Science & Business Media, 2013. (Cited on page 25.)

[Putinar 1993] M. Putinar. *Positive polynomials on compact semi-algebraic sets*. Indiana University Mathematics Journal, vol. 42, no. 3, pages 969–984, 1993. (Cited on pages 29, 35 and 137.)

[Rantzer 2001] A. Rantzer. *A dual to Lyapunov's stability theorem*. Systems & Control Letters, vol. 42, no. 3, pages 161–168, 2001. (Cited on page 73.)

[Regot 2011] S. Regot, J. Macia, N. Conde, K. Furukawa, J. Kjellén, T. Peeters, S. Hohmann, E. De Nadal, F. Posas and R. Solé. *Distributed biological computation with multicellular engineered networks*. Nature, vol. 469, no. 7329, pages 207–211, 2011. (Cited on page 58.)

[Riener 2013] C. Riener, T. Theobald, L. J. Andrén and J.-B. Lasserre. *Exploiting symmetries in SDP-relaxations for polynomial optimization*. Mathematics of Operations Research, vol. 38, no. 1, pages 122–141, 2013. (Cited on page 82.)

[Robinson 2003] J. C. Robinson. *Infinite-Dimensional Dynamical Systems: An introduction to dissipative parabolic PDEs and the theory of global attractors. Cambridge texts in applied mathematics*. Appl. Mech. Rev., vol. 56, no. 4, pages B54–B55, 2003. (Cited on pages 11, 12 and 13.)

[Rosenfeld 2019] J. A. Rosenfeld, B. Russo, R. Kamalapurkar and T. T. Johnson. *The Occupation Kernel Method for Nonlinear System Identification*. 58th Conference on Decision and Control, pages 6455–6460, 2019. (Cited on pages 86, 154, 162 and 164.)

[Rosenfeld 2020] J. Rosenfeld, R. Kamalapurkar, L. Forest Gruss and T.T. Johnson. *Dynamic Mode Decomposition for Continuous Time Systems with the Liouville Operator*, 2020. (Cited on pages 153 and 165.)

[Rosenfeld 2022] J. A. Rosenfeld, R. Kamalapurkar, L. Gruss and T. T. Johnson. *Dynamic mode decomposition for continuous time systems with the Liouville operator*. Journal of Nonlinear Science, vol. 32, no. 1, pages 1–30, 2022. (Cited on pages 47 and 85.)

[Rubio 1975] J. E. Rubio. *Generalized curves and extremal points*. SIAM Journal on Control, vol. 13, no. 1, pages 28–47, 1975. (Cited on pages 73 and 131.)

[Rudi 2020] A. Rudi, U. Marteau-Ferey and F. Bach. *Finding Global Minima via Kernel Approximations*, 2020. (Cited on pages 48, 85 and 186.)

[Rudin 1991] W. Rudin. *Functional analysis, mcgrawhill*. Inc, New York, vol. 45, page 46, 1991. (Cited on pages 15, 16, 27 and 53.)

[Rudin 2006] W. Rudin. *Real and Complex Analysis 3rd edn (eds Devine, PR) ch. 3, 49–50*, 2006. (Cited on page 15.)

[Russo 2022] B. P. Russo and J. A. Rosenfeld. *Liouville operators over the Hardy space*. Journal of Mathematical Analysis and Applications, vol. 508, no. 2, page 125854, 2022. (Cited on page 160.)

[Saitoh 2016] S. Saitoh and Y. Sawano. Theory of reproducing kernels and applications. Springer, 2016. (Cited on pages 43, 44, 45, 47, 85, 89 and 166.)

[Salova 2019] A. Salova, J. Emenheiser, A. Rupe, J. P. Crutchfield and R. M. D'Souza. *Koopman operator and its approximations for systems with symmetries*. Chaos, vol. 29, page 093128, 2019. (Cited on pages 89, 91, 163, 168 and 170.)

[Schaefer 1974] H. H. Schaefer. *Banach lattices*. In Banach Lattices and Positive Operators, pages 46–153. Springer, 1974. (Cited on pages 15, 84 and 85.)

[Scheffold 1971] E. Scheffold. *Das Spektrum von Verbandsoperatoren in Banachverbänden*. Mathematische Zeitschrift, vol. 123, no. 2, pages 177–190, 1971. (Cited on page 85.)

[Schlosser 2020] Corbinian Schlosser and Milan Korda. *Sparse moment-sum-of-squares relaxations for nonlinear dynamical systems with guaranteed convergence*. arXiv preprint arXiv:2012.05572, 2020. (Cited on pages 55, 56, 72, 82, 95, 102, 112 and 124.)

[Schlosser 2021] C. Schlosser and M. Korda. *Converging outer approximations to global attractors using semidefinite programming*. Automatica, vol. 134, page 109900, 2021. (Cited on pages 55, 72, 73, 74, 80, 81, 82, 83, 131, 138, 140, 143, 145 and 147.)

[Schlosser 2022a] C. Schlosser. *Converging approximations of attractors via almost Lyapunov functions and semidefinite programming*. IEEE Control Systems Letters, vol. 6, pages 2912–2917, 2022. (Cited on pages 55, 72, 75, 76, 77, 80, 81, 131, 145 and 147.)

[Schlosser 2022b] Corbinian Schlosser and Milan Korda. *Sparsity Structures for Koopman and Perron–Frobenius Operators*. SIAM Journal on Applied Dynamical Systems, vol. 21, no. 3, pages 2187–2214, 2022. (Cited on pages 55, 84, 92, 151, 163, 168, 169, 170 and 184.)

[Schmid 2010] P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. Cambridge University Press, 2010. (Cited on page 178.)

[Schmüdgen 1991] K. Schmüdgen. *The K-moment problem for compact semi-algebraic sets*. Math. Ann., vol. 289, no. 1, pages 203—-206, 1991. (Cited on pages 25 and 26.)

[Schrijver 2003] A. Schrijver. Combinatorial optimization: polyhedra and efficiency, volume 24. Springer, 2003. (Cited on pages 6 and 7.)

[Sinai 1989] Y. G. Sinai. Dynamical systems ii: Ergodic theory with applications to dynamical systems and statistical mechanics. Springer, 1989. (Cited on pages 37, 38, 97, 98 and 99.)

[Sriperumbudur 2011] B. K. Sriperumbudur, K. Fukumizu and G. Lanckriet. *Universality, Characteristic Kernels and RKHS Embedding of Measures.* Journal of Machine Learning Research, vol. 12, no. 7, 2011. (Cited on page 161.)

[Strogatz 2001] S. H. Strogatz. *Exploring complex networks.* nature, vol. 410, no. 6825, pages 268–276, 2001. (Cited on page 58.)

[Tacchi 2020a] M. Tacchi, C. Cardozo, D. Henrion and J.-B. Lasserre. *Approximating regions of attraction of a sparse polynomial differential system.* IFAC-PapersOnLine, vol. 53, no. 2, pages 3266–3271, 2020. (Cited on pages 69 and 71.)

[Tacchi 2020b] M. Tacchi, J.-B. Lasserre and D. Henrion. *Stokes, Gibbs and volume computation of semi-algebraic sets.* arXiv preprint arXiv:2009.12139, 2020. (Cited on page 75.)

[Tamsir 2011] A. Tamsir, J. Tabor and C. Voigt. *Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'.* Nature, vol. 469, no. 7329, pages 212–215, 2011. (Cited on page 58.)

[Tao 2008] T. Tao and T. Ziegler. *The primes contain arbitrarily long polynomial progressions.* Acta Mathematica, vol. 201, no. 2, pages 213–305, 2008. (Cited on page 37.)

[Tarjan 1972] R. Tarjan. *Depth-first search and linear graph algorithms.* SIAM journal on computing, vol. 1, no. 2, pages 146–160, 1972. (Cited on page 121.)

[Teel 2000] A. R. Teel and L. Praly. *A smooth Lyapunov function from a class-estimate involving two positive semidefinite functions.* ESAIM: Control, Optimisation and Calculus of Variations, vol. 5, pages 313–367, 2000. (Cited on page 14.)

[Teschl 2012] G. Teschl. Ordinary differential equations and dynamical systems, volume 140. American Mathematical Soc., 2012. (Cited on page 9.)

[Vinter 1978] R. B. Vinter and R. M. Lewis. *The equivalence of strong and weak formulations for certain problems in optimal control.* SIAM Journal on Control and Optimization, vol. 16, no. 4, pages 546–570, 1978. (Cited on pages 73 and 131.)

[Wang 2020] J. Wang, V. Magron, J.-B. Lasserre and N. H. A. Mai. *CS-TSSOS: Correlative and term sparsity for large-scale polynomial optimization.* arXiv preprint arXiv:2005.02828, 2020. (Cited on page 148.)

[Wang 2021a] J. Wang, V. Magron and J.-B. Lasserre. *TSSOS: A Moment-SOS hierarchy that exploits term sparsity*. SIAM Journal on Optimization, vol. 31, no. 1, pages 30–58, 2021. (Cited on pages 82 and 148.)

[Wang 2021b] J. Wang, C. Schlosser, M. Korda and V. Magron. *Exploiting Term Sparsity in Moment-SOS hierarchy for Dynamical Systems.*, 2021. (Cited on pages 55, 72, 82, 148 and 185.)

[Williams 2015] M. O. Williams, C. W. Rowley and I. G. Kevrekidis. *A kernel-based method for data-driven koopman spectral analysis*. Journal of Computational Dynamics, vol. 2, no. 2, pages 247–265, 2015. (Cited on pages 85 and 86.)

[Wörmann 1998] T. Wörmann. Strikt positive polynome in der semialgebraischen geometrie. na, 1998. (Cited on page 28.)

[Young 2017] J. Young, T. Hatakeyama and K. Kaneko. *Dynamics robustness of cascading systems*. PLoS computational biology, vol. 13, no. 3, page e1005434, 2017. (Cited on page 68.)

[Zagabe 2023] C. M. Zagabe and A. Mauroy. *Uniform global stability of switched nonlinear systems in the Koopman operator framework*. arXiv preprint arXiv:2301.05529, 2023. (Cited on page 86.)

[Zhu 2014] Z. Zhu, H. G. Golay and D. A. Barbie. *Targeting pathways downstream of KRAS in lung adenocarcinoma*. Pharmacogenomics, vol. 15, no. 11, pages 1507–1518, 2014. (Cited on page 68.)