Linear Complementarity Systems: A Study in Hybrid Dynamics

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof. dr. M. Rem,
voor een commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen op
dinsdag 30 november 1999 om 16.00 uur

door

Wilhelmus Petrus Maria Hubertina Heemels

geboren te St. Odiliënberg

Dit proefschrift is goedgekeurd door de promotoren:

prof. dr. ir. P.P.J. van den Bosch
en
prof. dr. J.M. Schumacher

Copromotor:

dr. S. Weiland

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Heemels, Wilhelmus P.M.H.

Linear complementarity systems : a study in hybrid dynamics / by
Wilhelmus P.M.H. Heemels. – Eindhoven : Technische Universiteit Eindhoven, 1999.
Proefschrift. – ISBN 90–386–1690–2
NUGI 811
Trefw.: lineaire complementariteitsproblemen / systeemtheorie / nietlineaire
dynamica / dioden.
Subject headings: complementarity / system theory / nonlinear dynamical systems /
variable structure systems / diodes.

Eerste promotor: prof. dr. ir. P.P.J. van den Bosch

Tweede promotor: prof. dr. J.M. Schumacher

Copromotor: dr. S. Weiland

Kerncommissie:

prof. dr. ir. M.L.J. Hautus
dr. A.J. van der Schaft

Promotiecommissie:

prof. dr. ir. W.M.G. van Bokhoven
prof. dr. ir. B.L.R. De Moor
prof. dr. ir. J.F. Groote
prof. dr. H. Nijmeijer
dr. B. Brogliato

The Ph.D. work was financially supported by the Co-operation Centre Tilburg and Eindhoven Universities (SamenwerkingsOrgaan Brabantse Universiteiten).

The Ph.D. work forms a part of the research program of the Dutch Institute of Systems and Control (DISC).

Voorwoord

Het is zaterdagavond kwart over zes als ik besloten heb om het laatste stukje tekst toe te voegen aan hetgeen mijn proefschrift moet gaan vormen. Er waait een stevige wind langs E-hoog en het is zojuist beginnen te regenen. Binnen is het akelig rustig. Een goed moment om eens op de afgelopen vier jaar terug te blikken.

Als eerste denk ik dan aan mijn vriendin Saskia. Volgens mij is zij de aanleiding geweest voor het gezegde "Achter elke sterke man, staat een sterke vrouw." Het eerste deel van dit gezegde moge betwistbaar zijn, maar het tweede deel staat in dit geval buiten kijf. Je bent al die tijd mijn steun en toeverlaat geweest (zeker in de laatste maanden) en al die jaren heb je ontzettend veel plezier aan mijn leven toegevoegd. Misschien zonder het je te realiseren, heb je een enorme bijdrage geleverd aan de totstandkoming van dit proefschrift. Bedankt daarvoor.

Ook mijn ouders en René wil ik bedanken voor alles wat we samen meegemaakt hebben. Jullie hebben mij gesteund in alle keuzes, die ik gemaakt heb, en hebben altijd voor de volledige honderd procent achter mij gestaan. Het feit dat jullie te allen tijde een veilige thuishaven voor mij gevormd hebben, is voor vele zaken, waaronder dit proefschrift, onmisbaar geweest.

Ik ben Hans Schumacher en Siep Weiland dankbaar voor de dagelijkse begeleiding, die zij op zich genomen hebben. Jullie enthousiasme en betrokkenheid heb ik als zeer plezierig en waardevol ervaren. Het was een voorrecht om op zo'n prettige en amicale wijze met jullie te mogen samenwerken.

Paul van den Bosch wil ik bedanken voor de enorme vrijheid en het vertrouwen in de afgelopen periode. De informele en soms filosofische gesprekken over uiteenlopende onderwerpen, heb ik zeer gewaardeerd.

I would like to thank Kanat Çamlıbel for the interesting discussions and the pleasant cooperation that resulted in Chapter 5 and 7 of the thesis.

Verder wil ik Arjan van der Schaft en Malo Hautus bedanken om een gedeelte van hun kostbare tijd op te offeren om de laatste puntjes op de i te zetten in mijn proefschrift.

Belangrijk en stimulerend is ook geweest, dat ik de afgelopen vier jaar met veel plezier naar mijn werk gegaan ben. Al mijn (oud-)collega's van ER wil ik hartelijk bedanken voor de prettige sfeer, die er in de groep heerste. In het bijzonder wil ik mijn kamergenoot Liu Hong en de "AiO's" Yvo Boers, Leon Ariaans, Robert-Jan Gorter, Mario Balenović, Dik de Bruin en Vick van Acht bedanken voor de broodnodige ontspanning op en buiten het werk. De gezellige lunches en tussendoortjes, de voetbalavondjes en de tennislessen maakten het verblijf op vloer vier zeer aangenaam. Ik bedank Barbara Cornelissen, Paul Borghouts en Udo Bartzke naast de sfeer-verhogende bijdragen ook voor de financiële, organisatorische, administratieve, technische en morele ondersteuning, die de drukke werkzaamheden vaak verlichtten.

Als laatste wil ik nog noemen de gezelligheid van zaalvoetbalvereniging Totelos, die ervoor gezorgd heeft dat ik regelmatig op andere gedachten (dan mijn proefschrift)

kwam. Mannen en dames, bedankt. Verder bedank ik Meta voor de "brainstorm-sessie" over mijn kaft.

Ook iedereen die niet met naam (en toenaam) genoemd is, maar toch een bijdrage geleverd heeft aan het ontstaan van dit proefschrift, wil ik bedanken.

Het is inmiddels acht uur geworden. Ik kijk naar buiten en zie dat het droog is. Moe maar voldaan, raap ik mijn spullen bij elkaar en zoek mijn pasje om het gebouw te verlaten. Door de regen is het flink afgekoeld. Ik knoop mijn jas dicht, stap op de fiets en verdwijn langzaam in de duisternis.

Maurice Heemels
Eindhoven, 2 oktober 1999

Abstract

Technological innovation pushes towards the consideration of dynamical systems of a mixed continuous and discrete nature, which are called "hybrid systems." Hybrid systems arise, for instance, from the combination of an analog continuous-time process and a digital time-asynchronous controller. Many consumer products (cars, micro-wave units, washing machines and so on) are controlled by digital embedded software, rendering the overall process a system with mixed dynamics. Also many physical systems display hybrid behavior: the description of multi body dynamics depends crucially on the presence or absence of a contact, models of friction phenomena distinguish between slip and stick phases and electrical circuits contain switching elements like diodes that can be blocking (open circuit) or conducting (short circuit).

From these examples it is obvious that a too general study of hybrid systems will lack decisive power: it will not result in detailed information on individual elements in the studied class. Therefore, one has to consider a subclass of hybrid systems carrying a clear additional structure allowing analysis of its behavior (e.g. well-posedness, simulation methods, stability) and facilitating systematic controller synthesis. However, the chosen subclass must also contain many interesting examples from an application point of view. The class of (linear) complementarity systems satisfies both requirements and is the subject of the thesis. Complementarity systems are described by differential equations, inequalities and logic expressions and form dynamical extensions of the linear complementarity problem (LCP) of mathematical programming.

The study of the complementarity class is motivated by a broad range of physically interesting systems that can be reformulated in terms of the complementarity formalism. Examples include mechanical systems subject to unilateral constraints, Coulomb friction or one-sided springs; electrical networks with diodes; control systems with saturation or deadzones; piecewise linear and variable structure systems; relay systems; hydraulic processes with one-way valves; and sets of equations resulting from optimal control problems with state or control constraints. Moreover, in Chapter 6 it is shown that the class of "projected dynamical systems" also fits into the complementarity framework.

To obtain a well-founded theory, it is essential to define a physically relevant solution concept and answer the classical questions of existence and uniqueness of solutions. Because of the "jump-phenomena" in the system variables and the multimodal behavior, formulating a solution concept for linear complementarity systems (LCS) is non-trivial. The solution trajectories are defined by combining a hybrid point of view and a distributional framework. After the formal introduction of the solution concept, connections are established with the existing literature on mechanical systems and electrical circuits. It is shown that the proposed solution concept is not an artificial one, but that it is in accordance with well-known rules specified for subclasses of complementarity systems.

It is surprising to see that studies of well-posedness in hybrid systems theory are rare. One often simply assumes existence and uniqueness of solutions without giving any verifiable conditions for these properties. In this thesis, we try to fill this gap for linear complementarity systems by deriving necessary and sufficient conditions for well-posedness. Although questions of well-posedness are of interest by themselves, it must be emphasized that they provide basic insights that are important for solving issues of controllability, stability and controller synthesis.

In Chapter 4, existence and uniqueness of "initial solutions" to linear complementarity systems is related to the existence and uniqueness of solutions to a family of static linear complementarity problems (LCPs). This connection is based on the so-called rational complementarity problem, a generalization of the LCP for rational functions, as an intermediate tool. This result allows the exploitation of the extensive literature on LCPs to obtain well-posedness results for linear complementarity systems. The strength of these results is illustrated by applying them to unilaterally constrained mechanical systems, linear relay systems and linear passive electrical circuits with ideal diodes. In Chapter 5 these results are extended to obtain "global existence" and to derive additional properties of electrical circuits with diodes.

The existence of initial solutions does not guarantee the existence of a solution on an interval of nontrivial support (called "local existence") in general due to the possibility of an infinite number of re-initializations at one time instant. In Chapter 3 sufficient conditions for local existence of solutions are derived based on another extension of the LCP, the so-called linear dynamic complementarity problem. The conditions are given in terms of the principal minors of the leading row and column coefficient matrices of the system. Based on these ideas new global existence results are given for linear complementarity systems with low leading row coefficients and bimodal systems (having only two modes).

Besides the solution concept and well-posedness issues, attention is paid to numerical methods for simulation of linear complementarity systems. One category of possible hybrid simulation techniques consists of the so-called "event-driven methods" that consider the simulation interval as a union of disjoint subintervals on which the mode (the set of active constraints) does not change. On a subinterval one must deal with differential and algebraic equations that can be solved by standard integration routines (DAE-simulation). As integration proceeds, one has to monitor certain indicators to determine when the subinterval ends (event-detection). Next, a new mode has to be determined (mode selection) and a possible reset of the continuous state variable must be computed (re-initialization). As the proposed solution concept is closely related to the event-driven method, the mathematical analysis of well-posedness has immediate consequences for this method. In particular, contributions are made to solve the re-initialization and mode selection problems.

As an alternative to event-driven methods, one can use time-stepping techniques that replace the system's equations directly by a "discretized" equivalent. Numerical integration formulas are applied to approximate derivatives and all algebraic conditions are enforced to hold at each time-step. For linear complementarity systems, the

method based on the well-known backward Euler formula results in solving an LCP for every time-step. In Chapter 7 an example is presented, for which the approximating functions do not converge when the step size tends to zero. This indicates that one cannot indiscriminately apply the backward Euler time-stepping method to arbitrary linear complementarity systems. Justification of this particular time-stepping method is thus required. Therefore, we show the consistency of this time-stepping method applied to the class of electrical networks with diodes. Here, "consistency" means the convergence of the approximations to the true solution of the original system in spite of the presence of impulses and switching dynamics, and the fact that the method does not try to trace the event times exactly.

During the achievement of the aforementioned goals and in the overview of applications in Chapter 2, relations between the various subclasses of complementarity systems are revealed. The advantage of finding a common structure of these interesting application fields, is that results obtained in one domain can be transformed or extended to another. Moreover, as a common meeting ground of several mature research areas, complementarity systems have the potential to play a major role in developing systematic methods to overcome analysis and synthesis problems in a wide range of applications. The work in this thesis forms a step in this direction, as it solves various fundamental problems, needed for setting up a general system and control theory for complementarity systems.

Contents

1

Introduction

The objectives of this chapter are to motivate this study, to discuss the connections to the existing literature, to formulate the goals of the thesis and to indicate the difficulties in achieving these goals.

1.1 Hybrid Systems

Technological innovation pushes towards the consideration of systems of a mixed continuous and discrete nature, which are sometimes called "hybrid systems[1]." Hybrid systems arise, for instance, from the combination of an analog continuous-time process and a digital time asynchronous controller. Many consumer products (cars, micro-wave units, washing machines and so on) are controlled by embedded software, rendering the overall process a system with mixed dynamics. Hybrid systems abound in our homes, probably more than we realize.

As an illustrative example of a hybrid system consider the regulation of the temperature in a house. In a simplified description, the heating element is assumed to work either at its maximum power or is completely turned off. In these two modes ("on" and "off") the temperature is governed by different dynamical regimes. The switching between the operating modes is controlled by a logical device (the embedded controller) called the thermostat. The mode is changed from "on" to "off," when (a function depending on) the temperature crosses a certain upper value (determined by the desired temperature). Vice versa, if the temperature drops below a minimum value, the heating is switched "on."

In large industrial processes, hierarchical control methods are being utilized more and more. As an example consider a plant (such as a refinery or a distillation column) in the process industry [126, p. 7]. At the top layer, the whole plant, consisting of several process units is monitored and the best economic operating conditions (quality of final

[1] Term used in this context for the first time by Witsenhausen in 1966 [208].

product, production quantity, use of resources, etc.) are determined for the separate process units. These conditions are passed as targets to multivariable controllers (e.g. a model predictive controller) for the process units. In turn, the MPC controller brings the process unit towards these targets and tries to keep it there. On the lowest level, local single-input single-output controllers (e.g. PID) are implemented for maintaining the level of the fluid in a tank between certain bounds determined by the setpoints of the MPC controller. Embedded controllers take care of the different control layers ranging from the implementation of local digital controllers, to exception handling, safety, alarm detection, switching between operating modes, and starting and stopping procedures (the supervisory layer). The overall system consisting of the physical plant and the embedded controller will form a complex hybrid system. Similar examples include automated highways [73], coordinated submarine systems and air-traffic management [198].

The hybrid nature is not necessarily caused by human intervention in smooth systems. Although many examples originate from adding digital controllers to physical processes, the switching between dynamical regimes is naturally present in a variety of systems. For instance in mechanics, one encounters friction models that make a clear distinction between stick and slip phases. Other examples include models describing the evolution of rigid bodies. In this case the governing equations depend crucially on the fact whether a contact is active or not. The dynamics of a robot arm moving freely in space is completely different from the situation in which it is striking the surface of an object. Backlash in gears and deadzones in cog wheels also result in a multimodal descriptions. It is not difficult to come up with interesting applications in the mechanical area: control of robotic manipulators driving nails or breaking objects [32], vibration control in suspension bridges [98], reduction of rattling in gear boxes of cars, drilling machines [160], simulation of crash-tests, regulating landing maneuvers of aircraft, design of juggling robots [33] and so on.

Examples are not only found in the mechanical domain. Nowadays switches like thyristors and diodes are used in electrical networks for a great variety of applications in both power engineering and signal processing. Examples include switched-capacitor filters, modulators, analog-to-digital converters, switching power converters, duty-ratio control, choppers, etc. In the ideal case, diodes are considered as elements with two (discrete) modes: the blocking mode and the conducting mode. Mode transitions for diodes are governed by state events (sometimes also called "internally induced events"), i.e. certain system variables (current or voltage) changing sign. In duty-ratio control the duration of a switch being open and closed (or the ratio between them in a fixed time interval) is determined by a control system and hence, the transitions are triggered by time events ("externally induced events").

Other sources of multimodal behavior are saturation, hysteresis, sensor and actuator failures. Actuator saturation truncates implemented control values outside the actuator range. Sensors provide reliable and accurate measurements only within a specific region, while outside the region the only available information is whether the measured signal is above the maximum or below the minimum of the sensor range. The mal-

functioning of sensors or actuators have the effect that control signals or measurements are not available and as a consequence, the input-output description changes abruptly. Control design must take switching and impact phenomena into account such that a desirable behavior of the closed loop system is realized.

How dependent our lives are on computer technology is illustrated by the efforts taken to solve the millennium bug. The number of computer-controlled products in our homes will grow even further in the coming years. To support this evolution, new methodologies for the analysis and synthesis of hybrid systems are needed. To guarantee the safety and proper functionality, we have to improve our understanding of the interaction between physical processes, digital controllers and software, as all three parts influence the dynamic behavior of the overall process.

Nowadays, the design of such combined systems is often performed by methods either exclusively tailored for discrete event systems (DES) or time continuous systems. As a result, the models neglect either the continuous or the discrete characteristics of the system. As an example, consider the air-traffic management of an airport. In describing the airport accurately, the model must contain the differential equations determining the trajectories of the aircraft, as well as the human and/or organizational processes realizing the communication and assignments between the aircraft and the traffic control center. An air-traffic controller obtained from a model not incorporating one of these aspects, may fail in practice or will at least show less performance than a controller designed by techniques incorporating both the discrete and the continuous behavior.

Another approach for the combined design consists of separating the analysis and design of the continuous and discrete parts and merging them in the final stage. The synthesis of a digital controller (PID, H_∞, IMC, etc.) for a process in a certain operating point is backed up by the vast literature on systems and control theory. Also the tools for the design of a DES taking care of e.g. mode switching and exception handling are available. However, a combined controller design of the system is currently impossible. The merging of the complete embedded controller with the physical plant is performed in a heuristic and ad hoc manner and requires often years of tuning, prototyping and trouble-shooting, which are extremely expensive and time consuming. The time-to-market and the necessary investments for new products can be decreased considerably, if techniques are available that facilitate combined synthesis of both the discrete and continuous parts.

A practical hybrid control problem encountered in the department of electrical engineering of the Eindhoven University of Technology is concerned with the synchronization of several tools within a mailing system based on low resolution encoders [85]. The company Buhrs-Zaandam B.V. in Zaandam (The Netherlands) builds machines that automatically compose a mailing package consisting of various brochures. The main component enters a conveyor belt and several supplements are added by sheet-feeders. The motions of these devices have to be coordinated. Traditionally, this kind of synchronization was realized by one mechanical axis driving all the tools. To increase flexibility ("plug and play" concept) the feeders are all mounted with a motor and a

controller with inputs the positions of both the conveyor belt and the motor. To keep the overall costs of the system low, the sensors for the sheet-feeder motors are cheap low resolution encoders having only one measurement pulse per revolution of the motor. Hence, the measurements are equidistant in the angular position of the motor, but not in time. The sensor has a state-event character: new pulses are triggered by a system variable crossing a certain threshold. In principle, this asynchronous control problem cannot be solved by standard control design methods, because these require (accurate) measurements to be known after fixed time intervals and the control actions to be updated synchronously in time. Neglecting the asynchronous measurement device and simply applying time synchronous design techniques, leads to unsatisfactory results (especially for low speeds) [85]. This problem has been solved by transforming the *asynchronous* problem for a linear system in the time domain to a *synchronous* problem for a nonlinear system in the (angular) position domain. In the position domain a gain-scheduling approach is applied. The design resulted in a position-synchronous, but time-asynchronous controller that is successfully implemented on the practical set-up. The high performance that was required could not be achieved by standard time-synchronous control techniques. However, the proposed "hybrid" control structure resulted in a cheap and satisfactory solution. This particular hybrid control problem is frequently encountered in industrial environments, since these kinds of sensors are often used (e.g. magnetic/optical disk drives, level sensors for the height of a fluid in a tank, transportation systems where the lateral position is only (exactly) known when a marker has passed [34], and so on). The number of industrial requests for solutions to such practical problems with inherent hybrid aspects will grow in the future.

Fortunately, it is widely recognized by the academic world that mixing different devices and concepts will play an increasingly important role in industry. Starting from their own backgrounds, control engineers [7, 145], computer scientists [162], mathematicians and simulation experts work towards systematic methods to support the development of new products. The increasing interest in this research area has become apparent from a series of workshops on hybrid systems in recent years [2, 5, 6, 76, 99, 133]. For an introduction to the field of hybrid dynamical systems, the reader is referred to [180].

1.1.1 Models for hybrid systems

As models are the ultimate tools for obtaining and dealing with knowledge, not only in engineering, but also in philosophy, sociology and economics, a search has been undertaken for appropriate mathematical models for hybrid systems. A whole range of possible model structures for hybrid systems has already been proposed. An overview of possible modeling techniques has been given for instance in [19, 24]. Mentioned are, among others,

- Timed or hybrid Petri-nets, see e.g. [51];
- Differential automata [195];

- Hybrid automata [28, 130];
- Brockett's model [30];
- Mixed logical dynamic models [15];
- Duration calculus [39]
- Real-time temporal logics [1, 161]
- Timed communicating sequential processes [52, 100]
- Switched bond graphs [189]

We would like to emphasize that this list is by no means exhaustive.

Some of these models start from one domain (DES or differential/difference equations) and include additional elements of the other domain. Hybrid automata, for instance, are derived from finite state machines used in describing DES by replacing the simple clock dynamics inside each discrete state by more involved differential and algebraic equations.

1.1.2 Hybrid automata

To give some impression on what hybrid systems look like, we discuss one interesting hybrid model structure, that complies with our point of view, in some detail.

A widely accepted framework for a hybrid system is a *hybrid automaton* given by the quadruple (Q, Σ, A, G) (notation taken from [28]) where

- Q is a finite set of *modes* (sometimes called *discrete states* or *locations*).
- $\Sigma = \{\Sigma_q\}_{q\in Q}$ is a collection of dynamical systems. For mode q these are given by the ordinary differential equations (ODEs) $\dot{z} = f_q(z)$ or the differential and algebraic equations (DAEs) $f_q(\dot{z}, z) = 0$, where $z(t) \in \mathbb{R}^n$ is a state variable.
- $A = \{A_q\}_{q\in Q}$. $A_q \subset \mathbb{R}^n$ is the *jump set* for mode q consisting of the states from which a mode transition and/or state jump occurs.
- $G = \{G_q\}$ is the set of *jump transition maps* where G_q is a (possibly multi-valued) map from A_q to a subset of $\mathbb{R}^n \times Q$.

A short description of the dynamics is given as follows. Starting in a continuous state $z_0 \in \mathbb{R}^n \setminus A_{q_0}$ in mode q_0, one evolves according to the mode dynamics given by Σ_{q_0} until one reaches — if ever — A_{q_0}, say at the event time τ_1 (the reaching of A_{q_0} is called an *event*). From this set a transition is enabled and *must* be fired instantaneously. The transition is governed by the relation $(z_1, q_1) := G_{q_0}(z(\tau_1^-))$ with $z(\tau_1^-) := \lim_{t\uparrow\tau_1} z(t)$. From this new state z_1 in mode q_1, it is possible that again a transition takes place, i.e. $z_1 \in A_{q_1}$. Otherwise, a continuous phase given by the dynamics Σ_{q_1} will follow.

This framework indicates the behavior of a hybrid system: continuous phases separated by events at which (maybe multiple) discrete actions (re-initialization of the continuous state z and discrete state q) take place.

We would like to stress that it can be nontrivial task to rewrite a physical model description in terms of a hybrid automaton. Especially, the definition of the jump sets and the jump transition maps (re-initialization and switching rules) can be really difficult.

1.1.3 Modeling versus decisive power

The choice of a suitable framework is a trade-off between two conflicting criteria: the modeling power and the decisive power. The modeling power indicates the size of the class of systems allowing a reformulation in terms of the chosen model description. The decisive power is the ability to prove quantitative and qualitative properties of individual systems in the framework. A model structure, which is too broad, (like the hybrid automaton in the previous section) cannot reveal specific properties of a particular element in the model class. The size of a model class is often taken too large for analysis purposes. As indicated by [18], even for the easiest hybrid systems analysis and control problems are often undecidable or require a high computational load. As an example, Tsitsiklis and Blondel [18] consider the elementary hybrid system given by

$$x(k+1) = \begin{cases} A_1 x(k), & \text{when } c^\top x(k) \geq 0, \\ A_2 x(k), & \text{when } c^\top x(k) < 0, \end{cases} \tag{1.1}$$

where A_1, A_2 are matrices and c is a (column) vector of appropriate dimensions. To decide whether this switching system is stable is shown to be NP-hard. Loosely speaking, this means that there is no algorithm that answers the question of stability in polynomial time (as function of the size of A_1, A_2 and c).

The complexity of hybrid systems is also shown by a simple piecewise linear forced Van der Pol oscillator with an ideal diode studied in [103]. The system consists of a capacitor, an inductor, a linear negative resistor, a diode and a sinusoidal voltage source. For the analysis the diode is assumed to be an ideal switch. The system switches between the blocking and conducting mode and the dynamics in the individual modes are linear. For a specific region of the parameter values (which are analytically determined) this system displays chaotic behavior that has been experimentally and numerically verified in [103]. The occurrence of chaos in such a simple system is rather intriguing, but indicates that multimodal systems are extremely complex.

From the previous it is clear that one should not consider a too general class of hybrid systems. But on the other hand, it is also useless to study a model class, which is (almost) empty and does not contain any physically relevant system. To summarize, it is essential to study a class of hybrid systems meeting the following criteria.

- The subclass is small enough: it must carry an additional structure facilitating detailed analysis of its behavior and controller design.

- The subclass is large enough: the class must be nontrivial. It has to contain interesting examples from an application point of view.

It may be clear that several choices of subclasses are possible. In this thesis, we will particularly be interested in so-called *complementarity systems* for reasons that will become clear later.

1.2 Complementarity systems

Inequalities have played an important role in many research fields including mathematical programming and economics (e.g. Leontief economies [114]). It is surprising to see that inequalities have received relatively little attention in systems theory. One reason might be that combining inequalities and differential equations means giving up the smoothness properties that form the basis of much of the theory of dynamical systems. However, in many situations (of which we will see several examples later) it seems reasonable to study dynamics in conjunction with inequalities.

In mathematical programming a key role is played by a special combination of inequalities and equations that is called the *linear complementarity problem* (LCP), which is defined as follows. Given a matrix $M \in \mathbb{R}^{k\times k}$ and a vector $q \in \mathbb{R}^k$, then LCP(q, M) amounts to finding vectors u, $y \in \mathbb{R}^k$ such that

$$y = q + Mu \tag{1.2a}$$

and

$$u_i \geq 0,\ y_i \geq 0,\ \{u_i = 0 \text{ or } y_i = 0\} \text{ for all } i \in \{1, \ldots, k\} \tag{1.2b}$$

or show that no such vectors exist. The operator "or" in (1.2b) must be interpreted in a non-exclusive sense. The conditions (1.2b) are called *complementarity conditions* and can equivalently be written as

$$u \geq 0,\ y \geq 0,\ u^\top y = 0. \tag{1.3}$$

The inequalities must be interpreted componentwise in (1.3). In the literature one often encounters also the more compact notation

$$0 \ \leq \ y \perp u \geq 0, \tag{1.4}$$

where the notation $y \perp u$ expresses the orthogonality between y and u. The LCP has many economic and engineering applications [65] and an extensive literature [47] is available on this problem.

The hybrid systems considered in this thesis can be seen as the dynamical extensions of LCPs and will be called *complementarity systems*. In a mechanical context such combinations of differential equations and complementarity conditions have already been used by Lötstedt [124]. Van der Schaft and Schumacher were one of the first

that formulated the the equations of complementarity systems (or "complementary-slackness systems") in a general setting [177, 179]. In their most general form complementarity systems are described by the differential and algebraic equations

$$0 = F(\dot{z}(t), z(t)) \tag{1.5a}$$

$$y(t) = g(z(t)) \in \mathbb{R}^k \tag{1.5b}$$

$$u(t) = h(z(t)) \in \mathbb{R}^k \tag{1.5c}$$

together with the complementarity conditions

$$0 \leq y(t) \perp u(t) \geq 0 \tag{1.5d}$$

In this formulation $t \in [0, \infty)$ denotes the time variable, $z(t)$ the state and $u(t)$ and $y(t)$ the complementarity variables at time t.

A special complementarity system occurs when (1.5a), (1.5b) and (1.5c) are replaced by an "input/state/output system" of the form

$$\dot{x}(t) = f(x(t), u(t)) \tag{1.6a}$$

$$y(t) = g(x(t), u(t)). \tag{1.6b}$$

These systems are called "semi-explicit" complementarity systems. Moreover, if the input/state/output system is taken to be linear, i.e. $f(x, u) = Ax + Bu$, $g(x, u) = Cx + Du$ for constant matrices A, B, C and D of appropriate dimensions, we obtain a *linear complementarity system* (LCS). Note that an LCS arises also by replacing the static linear relation $y = q + Mu$ in (1.2) by the linear dynamical system

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1.7}$$

$$y(t) = Cx(t) + Du(t). \tag{1.8}$$

In this thesis we will focus mainly on LCS, because we can rely in that case on the broad literature of linear system theory.

The study of complementarity systems can be motivated by a whole range of interesting applications. To give a quick round-up of examples, one might think of

- electrical networks with (ideal) diodes;
- piecewise linear systems;
- mechanical systems subject to unilateral constraints or Coulomb friction;
- switching control systems;
- dynamical systems with saturation, relays or deadzones;
- variable structure systems;
- hydraulic processes with one-way valves;

- sets of equations originating from applying Pontryagin's principle [80, 164] to optimal control problems with state or control constraints.

In Chapter 2 we will give a detailed exposition on the dynamical systems that can be modeled by the complementarity formalism and also on the modeling techniques that have to be used. Moreover, in Chapter 6 it will be shown that also projected dynamical systems [62, 147] allow a complementarity reformulation. Projected dynamical systems are used for studying e.g. economical markets, transportation networks and international trade.

1.3 Common meeting ground of disciplines

The fact that complementarity systems form a common denominator of several mature research areas motivates this study. Revealing the generic structure and developing a general framework for a broad range of applications offers many opportunities. One merit is the possible translation of results from one research field into another. If the differences and the similarities between the subclasses are mapped out clearly, then it becomes transparent which results can be transformed. Specific methodologies and proofs for e.g. constrained mechanical systems could be adapted for fields such as electrical circuit theory or piecewise linear systems. Therefore, it is important to investigate the results in the specific domains and identify how the complementarity is exploited to see whether extension is possible. Of course, one has to realize that not all results are extendable due to additional structure present in a specific domain. For instance, in constrained mechanical systems one has a concept of an energy function (storage function) consisting of kinetic an potential energy, which is not available in the framework of projected dynamical systems. Hence, results obtained by explicit use of an energy concept do not generalize (directly) to projected dynamical systems.

1.3.1 Electrical circuit theory

Modeling and simulation of electrical circuits have attracted much interest in the last decades [13, 20, 40, 43, 44, 75, 108, 121, 136, 172, 201, 203]. Circuit theorists are interested in analysis, verification and automated design of large-scale electronic networks. Modeling techniques in this area are frequently based on piecewise linear representations. For an overview of different canonical representations, the reader is referred to [111]. Piecewise linear representations are convenient, since they allow mixed-level simulations due to the same data structure for all kinds of circuits. Moreover, it allows the preservation of the hierarchy in the network model, which facilitates the replacement of a subcircuit by another subcircuit [119].

Research within circuit theory has concentrated on finding memory efficient canonical representations of networks, static (DC) analysis and development of simulation and synthesis tools for electronic circuits. Dynamical properties are studied by using integration routines (time-stepping methods) to approximate the dynamical system by

a series of static one-step problems, see e.g. [20, 120, 172]. In this way, the simulators could be used for both (DC) and transient analysis. Other numerical methods are more "event-driven" (see also subsection 1.4.3) in the sense that they try to trace the switching times of diodes, thyristors and other discontinuous elements exactly [13,136]. Efficient simulators (e.g. PLANET [110]) and tools for automated network design (e.g. TOP-ICS [119]) have been realized. However, existence and uniqueness of solutions to the *discontinuous dynamical* network models and justifications of the approximations based on time-stepping methods are not considered. One of the goals of this thesis will be to fill these gaps.

1.3.2 Piecewise linear systems

Piecewise linear (PL) systems are studied extensively (also outside the circuit theory community), because they form the simplest extensions of linear systems and moreover, can approximate nonlinear systems with arbitrary accuracy. One of the first studies on dynamical properties of PL (discrete-time) systems are stated in [190]. Sontag considers controllability and stability issues for PL systems and tries to use the obtained tools and methods for controlling other, more general, classes of systems (both discrete and continuous time nonlinear systems) by discrete-time PL systems. Recently, the PL-approach and other switching control schemes in the control society revives, see e.g. [17,27,38,81,104,106,134,149,204] for stability and control, [50] for equivalence of realizations, [118, 202] for observability and controllability and [37, 102] for well-posedness issues. Widely applied switching control techniques such as sliding mode control, gain scheduling and relay feedback [105, 123] can sometimes be formulated in PL description as well. From a more general point of view, the PL systems and switching control architectures can be seen as subsets of the (large) class of *variable structure systems*, which received quite some attention in the literature (see e.g. [68, 200]).

The renewed interest in PL systems in the control community motivates the study of complementarity systems too. As piecewise linear dynamical systems allow a re-formulation in terms of complementarity systems, the results of this thesis contribute to this research field as well.

1.3.3 Constrained mechanical systems

Mechanical systems with impacts and friction phenomena have a long history inspired by the work of well-known pioneers as Newton, Huygens and Poisson. The interest in constrained mechanical systems can be explained by the rich application field: robotics [113]; control of robotic manipulators driving nails, compacting powders or breaking objects (impactors) or transition phase control of a robot arm striking the surface of an object [31]; vibration control in suspension bridges, ships colliding at fenders or rattling gears to reduce wear, damage and noise [98]; simulation of crash-tests; regulating landing maneuvers of spacecraft and so on. For an overview of the available literature

on constrained mechanical systems the interested reader may want to consult [31] for an excellent survey. The study of mechanical systems subject to impacts can be split in different classes [31]. Among them one finds well-posedness studies [11, 124, 139, 144, 158, 181] for various restitution rules (inelastic and elastic) and friction phenomena, numerical schemes and experimental validation [12, 78, 192, 194, 199], analysis [49, 72, 160] and control of constrained mechanical systems [32, 33]. This list is not meant to be encyclopedic and the references serve only as possible entries to the subfields.

Although practical simulation procedures have received a lot of attention, classical questions of existence and uniqueness of solutions have been a little neglected. Recently, the interest for well-posedness issues (which are important for numerical methods as well) has increased. Lötstedt [124] proved *local* existence and uniqueness of *smooth* solutions under rather mild conditions. Of course, for global existence one has to study solution trajectories in a framework allowing impacts. Problems of (global) well-posedness for general nonlinear mechanical systems are extremely complicated, as is demonstrated by the first published existence result of reasonable generality due to [139], which takes a whole book [193, p. 25]. Monteiro Marques' result applies to the single-constrained case and is based on proving convergence of the time-stepping approach of Moreau [140, 144] using techniques from the *sweeping process*.

The problem of existence of solutions for a multi-constrained nonlinear mechanical systems was mentioned as an open problem in [139]. This open problem is partly solved by recent work in [192, 193], which uses a novel time-stepping scheme for rigid body dynamics with inelastic impacts and Coulomb friction based on complementarity problems. The convergence of a subsequence of the approximations has been shown. This results in both a (partial) justification of the applied simulation procedure and a proof of existence of solutions. The question of uniqueness is not posed in this work and the convergence of the whole sequence (instead of a subsequence) has not been shown. However, the ideas and techniques could be used as starting point for obtaining similar results for the class of complementarity systems.

1.3.4 Optimal control problems with state or control constraints

An extensively used methodology for solving optimal control problems is the maximum principle, initiated by Pontryagin et al. [164]. The original maximum principle has been used and extended by many others. Regarding optimal control problems with state constraints a survey can be found in [80]. The maximum principle results in necessary conditions for optimality, although the result is not rigorously established for the general case. Therefore, the statement of the conditions is called an "informal theorem" in [80] and is used mainly as a recipe to find candidates for the optimal control functions. Complementarity appears in these conditions to describe the duality between the state constraints and the corresponding multiplier (see Chapter 2). The resulting equations allow Dirac impulses in the solutions, resulting in discontinuities (jumps) of the adjoint variable (sometimes called co-state). The complementarity

point of view may contribute in obtaining a rigorous proof of this theory. Existence and uniqueness of solutions to the set of necessary conditions could be crucial for proving such a result. However, one has to realize that Pontryagin's principle is a two-point boundary value problem and as a consequence well-posedness requires a different approach. Questions on the smoothness of the adjoint variables, the number of constrained and unconstrained phases (finite or infinite) and the study of the behavior are interesting and mainly open questions. Some first steps in this direction can be found in the appendix of [60]. Dontchev and Kolmanovsky prove that for a linear quadratic regulator problem with a single linear state constraint of index one (meaning the constraint needs to be differentiated once to depend on the control input) the optimal control is piecewise analytic with only a finite number of mode switches between constrained and unconstrained phases. Their line of reasoning may be extendable to linear complementarity systems (without impulsive motions).

In the case of control constraints the applications of the maximum principle results in (depending on the cost functional and control constraint set) differential equations with piecewise linear characteristics. As an example consider a linear quadratic regulator problem with the control constraint set equal to the positive orthant in an Euclidean space. This problem is studied in [96] and [97], where it is shown that the control input is given by a simple (continuous) piecewise linear projection of a linear combination of the state and co-state (adjoint) variable on the positive cone. This projection is similar to the one-sided spring as studied for mechanical systems in [98] and has clear relations to PL systems. We observe that many of the applications of complementarity systems have natural connections to each other. This makes it interesting to study complementarity systems that might reveal the relationships and common structure of these subclasses more clearly.

1.4 Goals

The goals of this thesis are to:

(i) *Formulate a mathematically precise and physically relevant solution concept for the class of linear complementarity systems.*

(ii) *Deduce verifiable conditions that guarantee well-posedness of linear complementarity systems.*

(iii) *Develop numerical methods for the simulation of (linear) complementarity systems and obtain results on convergence of the approximations to assess the validity of the methods.*

(iv) *Show the relations between the physically relevant subclasses of complementarity systems.*

The first two goals are concerned with the fundamental system theoretic basis needed for analysis of linear complementarity systems. It is important to set up a

well-founded theory by defining a clear solution concept and answering the classical questions of existence and uniqueness of solutions (called "well-posedness"). Because of the "jump-phenomena" in the system variables and the multimodal behavior, a solution concept of linear complementarity systems is a non-trivial matter. After proposing such a solution concept, we aim to develop verifiable conditions for well-posedness. Here "verifiable" means algebraic conditions in terms of the parameters (state space parameters in our case) describing the system. Although questions of well-posedness are of interest by themselves, it must be emphasized that provide basic insights that will be important in solving issues of controllability, stability and controller synthesis.

The third objective of the thesis is to investigate numerical methods for simulation of LCS. Simulation is a common tool when analytical solutions or properties of dynamical systems cannot be derived. In some subdisciplines of complementarity systems several numerical methods have already been proposed (e.g. in electrical circuit theory and constrained mechanical systems). This thesis will contribute in particular to the so-called "time-stepping" and "event-driven methods" (see Subsection 1.4.3, for a description of these techniques). Since our solution concept is closely related to the event-driven method, we contribute especially to the re-initialization (determining the new continuous state after a mode change) and mode selection problem (determining the new discrete mode after a mode change). For the time-stepping methods, we will provide a rigorous base in the sense that the convergence of the approximations to a true solution of the original model will be shown (so-called "consistency" of the method).

The final goal involves the search for the relations between the subclasses of complementarity systems, which may result in the transfer of concepts, ideas and theory from one domain into another.

The following subsection will be dedicated to illustrate the importance of each of the four goals just mentioned. As such, these subsections serve as a motivation for the presented work.

1.4.1 Solution concept

A first step in the study of a class of dynamical systems must be the interpretation of the describing equations in terms of their solutions. The solution concept must be general enough to include the behavior observed in the physical process for which the model has been made, and limited enough to discard possible pathological solutions that have no physical meaning at all. In the literature on hybrid systems one often encounters the assumption of non-Zenoness in this context. Non-Zenoness means that only a finite number of events (mode switches and/or re-initializations) are allowed to happen in a finite length time interval. We emphasize that the term non-Zenoness, as used here, will include the requirement that at most finitely many successive jumps (resets or re-initializations) are allowed to take place at one time instant. We will illustrate by some simple examples (described also in [94]) the undesirable consequences of such an assumption.

Example 1.4.1 A physical example which will display nonexistence of solutions under an assumption of non-Zenoness is the three-balls system in which the inelastic impacts are modeled by a succession of simple inelastic impacts (Figure 1.1). Suppose that all

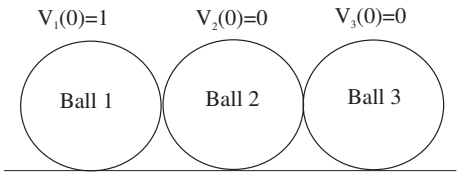

Figure 1.1: Three balls example.

the balls have unit mass and are touching at time 0. The initial velocity $v_1(0)$ of ball 1 is equal to 1 and for balls 2 and 3 equal to $v_2(0) = v_3(0) = 0$. If one assumes that the impact is actually a sequence of simple impacts, first an inelastic collision occurs between ball 1 and 2 resulting in $v_1(0^+) = v_2(0^+) = \frac{1}{2}$, $v_3(0^+) = 0$. Next, ball 2 hits ball 3 resulting in $v_1(0^{++}) = \frac{1}{2}$, $v_2(0^{++}) = v_3(0^{++}) = \frac{1}{4}$ after which ball 1 hits ball 2 again. In this way, a sequence of jumps is generated

$$
\begin{aligned}
v_1 &: \quad 1 \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{3}{8} \quad \tfrac{3}{8} \quad \tfrac{11}{32} \cdots \\
v_2 &: \quad 0 \quad \tfrac{1}{2} \quad \tfrac{1}{4} \quad \tfrac{3}{8} \quad \tfrac{5}{16} \quad \tfrac{11}{32} \cdots \\
v_3 &: \quad 0 \quad 0 \quad \tfrac{1}{4} \quad \tfrac{1}{4} \quad \tfrac{5}{16} \quad \tfrac{5}{16} \cdots
\end{aligned}
$$

which converges to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^\top$ from which a smooth continuation could be defined as a possible solution (the three balls stay touching and attain a velocity $\frac{1}{3}$ after the impact). However, if one assumes non-Zenoness one does not allow a solution containing infinitely many re-initializations at one time instant. In this case there would not exist a solution on a positive length time interval from the initial condition considered above. □

Apart from infinitely many events at one time instant, one has to be careful with accumulation of event times. In many models accumulation of events occurs and has a physical interpretation.

Example 1.4.2 Consider a model of a bouncing ball which is subject to gravitation forces. The model is given by $\ddot{x} = -g$ (x is the height of the ball and g is the gravity constant) and constraint $x \geq 0$. To complete the model we include Newton's restitution rule $\dot{x}(\tau+) = -e\dot{x}(\tau-)$ when $x(\tau-) = 0$ and $\dot{x}(\tau-) < 0$. Here e is the elasticity constant with $0 < e < 1$. Moreover, to complete the model, we include the rule that in case the ball is at rest (i.e. $x(\tau-) = \dot{x}(\tau-) = 0$), the ball stays at rest (meaning that the dynamics change to $\ddot{x} = 0$). The event times $\{\tau_i\}_{i\in\mathbb{N}}$ at which the ball touches the

ground are related through (see [31, p.234])

$$\tau_{i+1} = \tau_i + \frac{2e^i \dot{x}(0)}{g}, i \in \mathbb{N}$$

assuming that $x(0) = 0$ and $\dot{x}(0) > 0$. Hence, $\{\tau_i\}_{i\in\mathbb{N}}$ has a finite limit equal to $\tau^* = \tau_0 + \frac{2\dot{x}(0)}{g-ge} < \infty$. Since the continuous state $(x(t), \dot{x}(t))$ converges to $(0, 0)$ when $t \uparrow \tau^*$ a continuation beyond τ^* can be defined by $(x(t), \dot{x}(t)) = (0, 0), t > \tau^*$. The physical interpretation is that the ball is at rest within a finite time span, but after infinitely many bounces. Hence, the set of event times contains a right-accumulation point[2]. If one does not allow solutions with accumulations of event times, the maximal interval on which a solution can be defined is equal to $[0, \tau^*)$. □

The solution concept that will be used in the thesis will correspond to the *inelastic* impact case for non-smooth mechanical systems (see section 3.8). Consequently, the bouncing ball does not fit in the framework of LCS (at least using the inelastic jump transition rule). However, it indicates that there exist models of physical relevance that require a solution concept including the possibility of right-accumulations of events. An example with right-accumulations of event times that will fit in the solution concept used for linear complementarity systems, is given by the following system adapted from [68].

Example 1.4.3 A time reversed version of a system studied by Filippov [68, p. 116] (also mentioned in [123]) is given by

$$\dot{x}_1 = -\text{sgn}(x_1) + 2\text{sgn}(x_2) \tag{1.9a}$$
$$\dot{x}_2 = -2\text{sgn}(x_1) - \text{sgn}(x_2), \tag{1.9b}$$

where "sgn" denotes the signum-relation given by $\text{sgn}(x) = 1$, if $x > 0$, $\text{sgn}(x) = -1$, if $x < 0$ and $\text{sgn}(x) \in [-1, 1]$ when $x = 0$. Because this system consists of two relay characteristics, it can be modeled as a LCS (see chapter 2). Solutions of this piecewise constant system are spiraling towards the origin, which is an equilibrium. Since $\frac{\mathrm{d}}{\mathrm{d}t}(|x_1(t)| + |x_2(t)|) = -2$ when $x(t) \neq 0$ along trajectories x of the system, solutions reach the origin in finite time (see Figure 1.2 for a trajectory). However, solutions cannot arrive at the origin without going through an infinite number of mode transitions. Since these mode switches occur in a finite time interval, the event times contain a right-accumulation point (i.e. the time that the solution reaches the origin) after which the solution stays at zero.

□

The previous two examples show that *global existence* of solutions (i.e. existence of a solutions defined for all $t \in [0, \infty)$) cannot be achieved with a solution concept

[2]A point $\tau \in \mathcal{E} \subseteq \mathbb{R}$ is a right-accumulation point of $\mathcal{E}$, if there exists $\tau_i \in \mathcal{E}, i \in \mathbb{N}$ with $\tau_i < \tau$ such that $\tau = \lim_{i\to\infty} \tau_i = \tau$. A left-accumulation point is defined by changing "<" into ">."

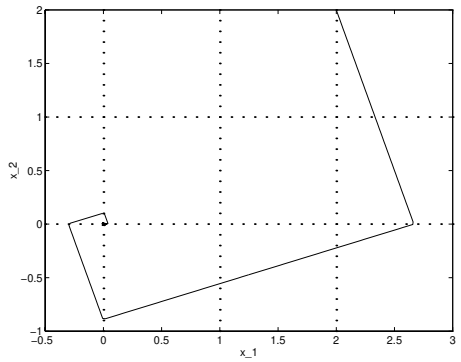

Figure 1.2: Trajectory with initial state $(2, 2)^\top$.

excluding right-accumulations of event times. Admitting left-accumulations of events in the solution concept, may result in nonuniqueness as demonstrated by the next example.

Example 1.4.4 The time-reversed model of (1.9) (which is the original example in [68]) is given by

$$\dot{x}_1 = \text{sgn}(x_1) - 2\text{sgn}(x_2) \tag{1.10a}$$
$$\dot{x}_2 = 2\text{sgn}(x_1) + \text{sgn}(x_2). \tag{1.10b}$$

This system has (infinitely many) solutions corresponding to initial state $x_0 = 0$, if one allows left-accumulations of event times. Hence, uniqueness cannot be inferred. Note that if we only allow right-accumulations of event times, the only solution starting in the origin is the zero solution. Allowing also left-accumulations results in a nondeterministic system, which is undesirable from a point of view of modeling and simulation. In contrast with smooth dynamical systems, time is considered to be asymmetric for hybrid systems, since reversing time is not natural and does not lead to well-posed systems in general. Solutions are therefore considered in a 'forward sense' that accepts right-accumulation and rejects left-accumulations of event times. In some situations we can even exclude the existence of (left-)accumulation points, see Chapter 5 and [94].

Am important observation is that the solutions with left-accumulations of events *do* however satisfy (1.10) in the sense of Carathéodory. A function x is a Carathéodory solution to $\dot{x} = f(x)$ with initial condition $x(0) = x_0$, if the equality[3] $x(t) = x_0 + \int_0^t f(x(\tau))d\tau$ holds for all $t \in [0, \infty)$. Hence, one has to be careful with using 'classical' notions of solutions for hybrid systems. However, if one can prove uniqueness in the sense of Carathéodory, one might be able to show that no left-accumulation

[3]Since the sgn-relation is multi-valued in zero, the equality (=) should be replaced by an inclusion ($\subseteq$) in order to be mathematically precise. However, for the particular example at hand, it makes no difference.

of events occurs. This is the method that will be used for electrical networks with diodes (see Chapter 5). □

A mechanical example displaying similar nonuniqueness due to left-accumulation of event times has been formulated by Bressan as described in [31, p. 58].

The previous examples indicate that there is a clear interaction between the way of modeling, the interpretation of the equations (i.e. the solution concept) and existence and uniqueness of solutions. Care should be taken to formulate a solution concept, since it influences the existence and uniqueness of solutions (and thus the well-posedness of the model). As stated before, the notion of solution must include all relevant trajectories of the original physical system and must not admit pathological solutions representing behavior that is not encountered in practice. Looking at the subclasses of complementarity systems, our solution concept must be able to describe at least

- smooth continuations;
- discontinuities and impulsive motions (as in constrained mechanical systems);
- right-accumulation of event times (as in the bouncing ball and the time-reversed Filippov's example).

1.4.2 Well-posedness

A first criterion for the validity of a mathematical model involves the existence and uniqueness of solutions (given initial conditions). This property is referred to as *well-posedness* and is a fundamental issue for every class of dynamical systems. It is surprising to see that in hybrid systems theory the study of well-posedness is quite rare. One often assumes that solutions exist, that they are unique and have a finite number of events in a finite length time interval. However, verifiable conditions for these properties are not presented in most cases, although it is widely recognized that it is an important issue. Johansson [106] calls well-posedness an important component of a more complete theory for piecewise linear dynamical systems. Imura and Van der Schaft [102] state that there are still few results on the basic problems of uniqueness of solutions to piecewise linear discontinuous systems.

We do not claim that well-posedness is completely neglected. The work [129] studies existence and uniqueness of solutions to hybrid automata. However, as a consequence of the general framework used in [129], their conditions are not verifiable by algebraic relations on the system parameters. For the class of PL systems, [37, 102] study questions of existence and uniqueness of solutions in a Carathéodory sense, thereby not allowing possible discontinuities in state entries and impulsive motions. As a consequence, their results do not apply to constrained mechanical systems or sets of equations resulting from optimal control problems and thus not to the general class of LCS. As mentioned before, in the field of constrained mechanical systems some results are known [124] (local existence and uniqueness of smooth continuations), [139]

(global existence with one constraint and inelastic impacts) and [192] (global existence with inelastic impacts for multiple constraints). Although they rely on the special structure of mechanical systems, some of the ideas can be extended to other classes of complementarity systems like projected dynamical systems as will be discussed in Chapter 6.

A starting point of the well-posedness results in this thesis is the work of Van der Schaft and Schumacher. In [177] necessary and sufficient conditions for *local* well-posedness are given for *bimodal* linear complementarity systems. In this context, bimodal means that there are only two modes. Stated differently, there is only one pair of complementarity variables and thus $u \in \mathbb{R}$, $y \in \mathbb{R}$. In [179] the local existence and uniqueness of *smooth* continuations have been studied for nonlinear semi-explicit complementarity systems (see (1.6) below).

As a last comment in this subsection concerns the fact that for smooth systems the notion of well-posedness often includes the continuous dependence of the solutions on the initial data. In this thesis, we will present an example of linear complementarity systems (in a mechanical context) that displays *discontinuous* dependence on initial states caused by the sensitivity of the solution trajectories to the order in which constraints become active. So, in general this property does not hold for linear complementarity systems. However, in Chapter 7 continuous dependence is proven for a class of linear complementarity systems for which the underlying state space description satisfies a passivity condition.

1.4.3 Numerical methods

Simulation is a common tool (and final escape) when analytical solutions or properties of model equations cannot be derived. However, simulation has to be considered as just executing experiments. The answers obtained are only valid for the experiments carried out. Reliable extrapolation of the results to other operating conditions cannot be guaranteed. As a consequence, simulation is not able to show that a complex (hybrid) model has properties like stability. The reason is that the design of experiments covering all possible operating modes is a tedious and time-consuming activity and is almost always impossible. A simulation can only prove that a system does not have certain properties by executing one particular experiment that contradicts the property.

It is recognized that new techniques are required for approximating the solution trajectories of hybrid systems. Simulators and languages like Omola/Omsim [4], Chi (χ) [14], SHIFT [57], Psi [23], Prosim [187], Modelica [137] and Matlab/Simulink/Stateflow are recently developed or adding hybrid features to their existing simulation environment. Most of the mentioned hybrid simulators can be categorized as event-driven methods according to a classification made by Moreau [140] for numerical techniques used for unilaterally constrained mechanical systems.

Classification of simulation techniques

The paper [140] classifies the literature on simulation techniques for rigid body dynamics with collisions into three categories. We believe that this classification also applies to possible numerical methods for complementarity systems.

• *Event-driven methods*

These methods are based on considering the simulation interval as a union of disjoint subintervals on which the mode (active constraint set) remains unchanged. On each subinterval the inequalities (unilateral constraints) are replaced by a set of equalities (bilateral constraints), that determine the evolution of the system. On each of these subintervals we are dealing with *differential and algebraic equations* (DAE), which can be solved by standard integration routines [29] (*DAE-simulation*). As integration proceeds, one has to monitor certain indicators (remaining inequalities that hold with strict inequality at the interior of the subinterval) to determine when the subinterval ends (*event detection*). At this event time a mode transition occurs, which means that one has to determine what the new mode will be on the next subinterval (*mode selection*). In mechanical terms, one must calculate which contacts persist or release and which contacts are newly formed. This can be a difficult task, since the contacts which release after the event time are not necessarily those for which an unfeasible contact has just been evaluated. An example of this phenomenon due to Delassus is described in [31, p. 117]. An illustration in a complementarity context is Example 3.8.3 below. If the state at the event time is not consistent with the selected mode, a jump is necessary (*re-initialization*). For instance, when two rigid bodies run into each other, a reset of their velocities will be required to prevent violation of the non-interpenetrability constraint. The complete numerical method is based on a repetitive cycle consisting of DAE-simulation, event detection, mode selection and re-initialization. It is possible that multiple mode selections and re-initializations are required before a DAE can be simulated over a subinterval of nontrivial support.

The event-driven methodologies are also used for simulation of switching electrical circuits [13, 136].

• *Smoothing methods*

The idea is to approximately replace the nonsmooth governing relationships by some regularized ones [140]. As an example in a mechanical setting, a non-interpenetrability constraint will be replaced by some stiff repulsion laws and damping actions which are effective as soon as two bodies of the mechanical system come close to each other. It is illustrative in this context to consider Chapter 2 of [31], where it is shown for a simple example that the percussion explained by compliant models (containing stiffness and damping) tend to the hard impact model (with Dirac measures in the reaction force) when the stiffness and damping coefficients tend to infinity.

The dynamics of the resulting approximate system is then governed by differential equations with sufficient smoothness to be handled through standard numerical techniques. Discrete modes do not really exist anymore, so event detection and mode selection are not necessary. Instantaneous jumps are replaced by (finitely) fast motions, so also the problem of re-initialization disappears. A drawback of this method is that an accurate simulation requires the use of very stiff approximate laws. The time-stepping procedures have to resort to very small step-length and possibly also have to enforce numerical stability by introducing artificial terms in the equations [140]. This results in long simulation times and the effect of the artificial modifications may blur the simulation results.

- *Time-stepping methods*[4]

The describing equations are directly replaced by some "discretized" equivalent. Numerical integration routines (see e.g. [71]) are applied to approximate the system equations. In particular, all algebraic relations (like the complementarity conditions) are enforced to hold at each time-step. In this way, one has to solve at each time-step an algebraic problem (sometimes called the "one-step problem") involving information obtained from previous time-steps. For linear complementarity systems, for instance, one has to solve a linear complementarity problem at each time-step. In contrast with event-driven methods, time-stepping methods do not determine the event times accurately, but "overstep" them. The time-stepping methods are used, for instance, in [20, 110, 120, 125, 143, 155, 172, 192, 194]. The work [193, p. 3] states that time-stepping methods (applied to mechanical systems) are based on using integrals of forces over each time-step instead of the instantaneous values of the force functions. The contact laws are not applied moment-by-moment. Instead, they are applied to short-time integrals. In this way there is no clear distinction between finite forces and impulses, which allows the two to be treated on the same level. In terms of event-driven methods, this means that the re-initialization and DAE-simulation are solved by the same technique.

As our solution concept is closely related to the event-driven method, the mathematical analysis of well-posedness has immediate consequences for this method. In particular, contributions are made to solve the re-initialization and mode selection problems. However, the main interest for numerical schemes in this thesis will be on time-stepping methods. The motivation for this is that time-stepping methods are used extensively for switching electrical circuits and unilaterally constrained mechanical systems [20, 110, 120, 125, 143, 155, 172, 192, 194], but the consistency of the method is less clear than for an event-driven method. Our main objective is to give a rigorous base for time-stepping applied to electrical networks with diodes by showing the convergence of (at least a subsequence of) the approximations to the true solution of the original model. As mentioned before, for mechanical systems such proofs can be

[4] Moreau [140] refers to these methods as *contact dynamics*.

found in [192].

A further advantage of the time-stepping approach is that it can be used as a starting point for controlling linear complementarity systems. For smooth systems it is common practice to use sampled data control and to design a controller on the basis of a discretized version of the system. This methodology can be extended to complementarity systems (under certain conditions), because accurate discretized models can be obtained by time-stepping techniques. Since such a discretized model can be rewritten in a discrete-time piecewise linear description (for which stabilization and control problems have already been studied [15, 190]), this opens several possibilities for controller synthesis for complementarity systems.

The use of smoothing as a numerical tool is not investigated in this thesis (and is recommended for further research). The complementarity conditions $u_i \geq 0$, $y_i \geq 0$ and $\{u_i = 0 \text{ or } y_i = 0\}$ could, for instance, be replaced by the piecewise linear function $y_i = \max(0, -\alpha u_i)$, where α will be a parameter approaching infinity. It would be interesting to study whether the solutions of the relaxations (as function of α) converge to the solutions of the original linear complementarity system. Of course, one could also make other, smoother, approximation of the complementarity conditions. As remarked before, a drawback of this approach is that one has to deal with very stiff differential equations whenever one requires accurate approximations of the real solution.

Mode selection

Mode selection refers to the problem of determining the next mode ("active index set" or "discrete state") on the basis of the current continuous state. In the terminology of electrical circuits with diodes, it involves the selection of blocking (current is zero) and conducting (voltage is zero) diodes in the next time frame given the continuous state of the network (voltages across capacitors and currents through inductors). This problem is essential for simulation based on an event-driven method, but has also a clear connection to well-posedness. If from a certain state vector no mode can be chosen, there does not exist a solution starting from this state (deadlock). If multiple modes can be selected, there may be a situation of nonuniqueness of solutions.

A practical example illustrating the need of efficient mode selection methods can be found in [196], where the objective is to verify computer-controlled power converters for the propulsion of a locomotive by a digital real-time ("hardware-in-the-loop") simulation. Instead of testing the obtained control system directly to the locomotive, one connects the inputs and outputs of the designed control system to a real-time simulation of the target process (all loops are closed via the simulator). The demand of such real-time simulation is motivated by the following factors [196]:

- reduction of risk (loss of human life or capital)
- decrease of costs (tests in target system can be extremely expensive)
- lack of availability (designated working environment is not available)

- lack of coverage (not all test states can be reached during regular operation).

Some of these factors play also key roles in e.g. crash-tests, flight simulators or design of a nuclear reactor. In [196] it is indicated that in case idealized models of switches are used, at least the following two separate issues have to be dealt with:

- The causality of the model will change during every mode transition;
- State events will have to be detected and continuous mode equations re-arranged and re-initialized for every mode transition.

It is obvious that the real-time condition asks for efficient mode selectors. The propulsion system of the locomotive contains thyristors that require a frame time (i.e. the calculation time for one simulation step with the entire model after which communication with the control systems takes place) of the simulation of $30\mu s$. The determination of the state of the thyristors within such an extremely fast time frame appeared to be a big problem [196].

So the problem of mode selection is not only important from an academic point of view, but also from an industrial point of view. In [179] the problem of mode selection is treated for complementarity systems in the semi-explicit form, where only smooth continuations are considered. Of course, one has to incorporate the possibility of impulsive continuations and state re-initializations to arrive at a solution for the complete mode selection problem. Therefore, the mode selection problem will be considered in this thesis for linear complementarity systems although some of the applied ideas have further extensions.

1.5 Outline

The purpose of the previous section has been to describe and motivate the separate goals, and to indicate the difficulties in solving the problems. The approaches to these goals are described in individual chapters as indicated by the following outline of the thesis. A nice feature is that the chapters are self-contained as they consist of accepted or submitted papers.

In Chapter 2, the study of complementarity systems is motivated by a whole range of possible applications. We consider unilaterally constrained mechanical systems, piecewise linear systems, electrical circuits with ideal diodes, hydraulic systems with one-way valves, systems of equations originating from applying Pontryagin's maximum principle to optimal control problems with state and/or input constraints, projected dynamical systems, variable structure systems (e.g. relay systems and switching control systems), control systems with saturation or deadzones, etcetera. This chapter consists of the paper [91].

In Chapter 3 a mathematically precise solution concept will be given for linear complementarity systems. To show that this is not an artificial definition without any physical relevance, we will prove that it corresponds to the switching and re-initialization

rules for linear constrained mechanical systems as proposed by Moreau [139] for the inelastic impact case. Moreover, several mode selection methods are proposed and discussed. Using this analysis, we are able to formulate sufficient conditions for *local* well-posedness based on the principal minors of the leading row and column coefficient matrices. Under these conditions, the set of regular states (the continuous states from which smooth continuation is possible without a re-initialization) will exactly be characterized. It will be shown also that after at most one jump of the state variable, smooth continuation is possible. As a final result in this chapter, these results are used to prove *global* well-posedness for bimodal systems and linear complementarity systems of "low index." This chapter is based on the papers [92] and [94].

The well-posedness results in Chapter 3 are based on the mode selection tool called the *linear dynamic complementarity problem* (LDCP). Another equivalent method, called the *rational complementarity problem* (RCP), is studied in Chapter 4. Since the solutions of the RCP have a direct relationship to (initial) solutions of linear complementarity systems, solvability issues of RCPs are essential for well-posedness. In this chapter we will show that the existence and uniqueness of solutions to RCP can be completely characterized by solvability properties of a family of linear complementarity problems, see (1.2). As a consequence, we can rely on the vast literature on the LCP [47] to obtain existence and uniqueness results of solutions to RCP and thus of well-posedness of linear complementarity systems. These results apply (among others) to linear constrained mechanical systems, linear relay systems (using results of [123]) and linear passive complementarity systems (including linear passive electrical circuits and ideal diodes). The results in this chapter have been published in [93].

The article [84] is included as Chapter 5, which discusses the extension of the 'initial well-posedness' results obtained in Chapter 4 to global well-posedness results using an assumption of passivity. Immediate applications of these results are linear passive electrical circuits with ideal diodes. As we will see, for such networks existence of solutions is guaranteed on $[0, \infty)$ and the solutions are unique, even if one allows accumulation of event times. As a byproduct, we obtain that derivatives of Dirac impulses do not occur and Dirac impulses can only occur at the initial time $t = 0$ (at "switch on"). These facts are "common sense truths" in the circuit theory community, but we are not aware of any rigorous proofs of these facts in the literature. Moreover, in proving these results we obtain an exact characterization of the set of regular states.

In Chapter 6 the actual proof is given for the claim of Chapter 2 that projected dynamical systems [62,147] can be cast into the complementarity formalism. Building on the results of [179] and convexity analysis the result is shown. Moreover, we give an alternative proof of global existence of solutions to projected dynamical systems. Usually, this proof has been based on the Skorokhod problem [188]. We believe that the proof presented in this thesis is more direct and that it reveals additional properties of the solution trajectories. The results of Chapter 6 have been reported in [86].

The results on time-stepping methods of [35] are written down in Chapter 7. This chapter starts by presenting an example of a linear complementarity system for which time-stepping based on the backward Euler integration formula fails. This motivates

the need for a rigorous proof of the consistency of the method for certain subclasses of complementarity systems. We will therefore be interested in proving the convergence of the approximations generated by the time-stepping method based on backward Euler as reported in e.g. [20, 120, 172] applied to linear passive electrical circuits with diodes. The same arguments as used in the consistency proof yield also the continuous dependence of the solution trajectories on the initial state. A similar convergence problem is studied for linear complementarity systems of "low index."

Finally, in Chapter 8 the contributions of the thesis are summarized and several open problems, which we believe interesting and relevant for both industry and academia, are recommended for future research.

Kanat Çamlıbel acted as one of my co-authors for the papers on which Chapters 5 and 7 are based, and these results are part of his PhD work too.

2

Applications of complementarity systems

This paper has been presented at the European Control Conference 1999 in Karlsruhe (Germany) [91].

2.1 Introduction

Technological innovation leads to an increasing interest in systems of a mixed continuous/discrete nature (called 'hybrid systems'). Recently, hybrid systems receive a lot of attention both from the control [7] and computer science community [162]. A subclass of hybrid systems consists of complementarity systems as introduced in [177]. In its most general form a complementarity system is governed by the differential and algebraic equations

$$0 = F(\dot{z}(t), z(t)) \tag{2.1a}$$

$$y(t) = g(z(t)) \in \mathbb{R}^k \tag{2.1b}$$

$$u(t) = h(z(t)) \in \mathbb{R}^k \tag{2.1c}$$

together with the complementarity conditions

$$\{y_i(t) = 0 \text{ or } u_i(t) = 0\},\ y_i(t) \geq 0,\ u_i(t) \geq 0 \tag{2.1d}$$

for all $i \in \{1, \ldots, k\}$. The complementarity conditions are similar as those appearing in the linear complementarity problem of mathematical programming [47].

A special complementarity system occurs when (2.1a), (2.1b) and (2.1c) are replaced by an "input-output system" of the form

$$\dot{x}(t) = f(x(t), u(t)) \tag{2.2a}$$

$$y(t) = g(x(t), u(t)). \tag{2.2b}$$

In this case we speak of "semi-explicit" complementarity systems.

If the system is linear, i.e. $f(x,u) = Ax + Bu$, $g(x,u) = Cx + Du$ for constant matrices A, B, C, D, we speak of a linear complementarity system (LCS).

The class of complementarity systems has been investigated in [86, 92, 93, 123, 177, 179]. Several basic issues are studied in these papers: the introduction of a mathematically precise solution concept, existence and uniqueness of solutions, mode selection methods, simulation issues and the study of the particular behavior of these systems. Current and future research will include stability analysis, development of numerical algorithms to approximate solutions and the inclusion of measurement and control variables. The purpose of this paper to show that the analysis of the class of complementarity systems is motivated by a wide range of applications.

2.2 Electrical networks with ideal diodes

Consider a linear electrical network consisting of resistors, inductors, capacitors, gyrators, transformers (RLCGT) and of k ideal diodes. To model this system as a LCS, the network is viewed as the interconnection of an RLCGT network with the diodes. More precisely, the RLCGT components form a multiport network described by a state space representation $\dot{x} = Ax + Bu$, $y = Cx + Du$ [3] with state variable x representing voltages over capacitors and currents through inductors. The input/output variables u and y represent the port variables: the pair (u_i, y_i) denotes the voltage-current variables at the i-th port. Interconnection of the i-th port to an (ideal) diode results in the equations

$$u_i = -V_i,\ y_i = I_i \text{ or } u_i = I_i,\ y_i = -V_i,$$

where V_i and I_i are the voltage across and current through the i-th diode, respectively. Finally, the ideal diode characteristic of the i-th diode is given by (see also fig. 2.1)

$$V_i \leq 0,\ I_i \geq 0,\ \{V_i = 0 \text{ or } I_i = 0\}. \tag{2.3}$$

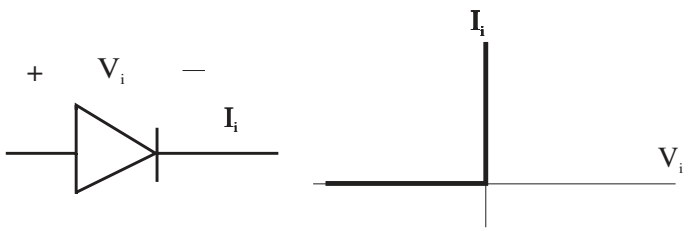

Figure 2.1: The i-th ideal diode characteristic.

2.3 Pipelines with one-way valves

Many chemical and hydraulic processes contain valves that only allow flows in one direction. A lid in the pipe can be opened to one side only, which prevents the fluid or gas from streaming back. The situation is shown in fig. 2.2.

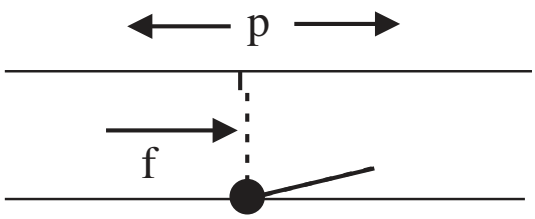

Figure 2.2: A pipeline with a one-way valve.

The flow in the pipe at time t is denoted by $f(t)$ and the pressure over the valve (lid) by $p(t)$. Ideally, only two situations can happen. The lid is either completely closed (dotted situation) or completely open (solid situation). The closed case occurs only if the pressure on the right is larger than the pressure on the left ($p(t) \geq 0$). The flow is then equal to zero ($f(t) = 0$). In the other situation (valve open), the pressure over the valve is zero and the fluid streams in the positive direction ($p(t) = 0$ and $f(t) \geq 0$). Hence, flow and pressure are complementarity variables.

2.4 Constrained mechanical systems

Consider a conservative mechanical system in which q denotes the generalized coordinates and p the generalized momenta. The free motion dynamics can be expressed in terms of the Hamiltonian $H(q, p)$, which has the interpretation of the total energy in the system. The equations are

$$\dot{q} = \frac{\partial H}{\partial p}(q, p) \tag{2.4a}$$

$$\dot{p} = -\frac{\partial H}{\partial q}(q, p). \tag{2.4b}$$

The system is subject to the geometric inequality constraints given by

$$C(q) \geq 0. \tag{2.4c}$$

Friction effects are not modeled here. We refer to subsection 2.5.3 for phenomena like Coulomb friction.

To obtain a complementarity formulation, we introduce (see also [92,124,160,177, 179]) the Lagrange multiplier u generating the constraint forces needed to satisfy the

unilateral constraints (2.4c). According to the rules of classical mechanics, the system can then be written as

$$\dot{q} = \frac{\partial H}{\partial p}(q, p) \tag{2.5a}$$

$$\dot{p} = -\frac{\partial H}{\partial q}(q, p) + \frac{\partial C^\top}{\partial q}(q)u \tag{2.5b}$$

$$y = C(q) \tag{2.5c}$$

together with the complementarity conditions (2.1d). The conditions (2.1d) express that the Lagrange multiplier u_i is only nonzero, if the corresponding constraint is active ($y_i = 0$). Vice versa, if the constraint is inactive ($y_i > 0$), the corresponding multiplier u_i is necessarily equal to zero.

The control of these systems is a major research topic. Since most control theories are model-based, adequate modeling of dynamical discontinuities and impact phenomena are necessary. Control applications can be found for instance in the field of robotics [31, 49, 113].

2.5 Piecewise linear characteristics

In this section we consider a dynamical system in which certain variables are coupled by means of a static piecewise linear (PL) characteristic. The situation is depicted in fig. 2.3. The variables v, z appear in the dynamics of the system Σ. These variables are related "in closed loop" through a PL relation. As an example one could think of a mechanical system with Coulomb friction or an electrical circuit containing a resistor having a PL behavior (see e.g. [121]).

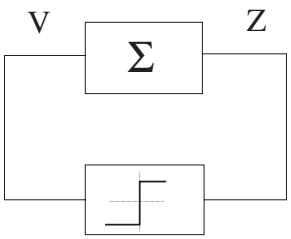

Figure 2.3: System with a PL relation.

2.5.1 A simple max-relation

Let v and z be related through $v = \max(0, z)$. See fig. 2.4. We introduce two auxiliary variables u, y and the algebraic equation $z = u - y$. It is easily verified that adding

the complementarity conditions $u \geq 0$, $y \geq 0$ and $\{y = 0 \text{ or } u = 0\}$, results in $u = v$. Hence, the relation $v = \max(0, z)$ can be replaced by

$$\begin{aligned} z &= u - y && \text{(2.6a)} \\ v &= u && \text{(2.6b)} \\ u \geq 0, \quad y \geq 0, \quad &\{y = 0 \text{ or } u = 0\} && \text{(2.6c)} \end{aligned}$$

resulting in a complementarity system. Hence, any system that can be formulated in terms of 'max' operations (think of 'max-plus systems'), can be cast into a complementarity framework due to the fact that $v = \max(w, z) = w + \max(0, z - w)$.

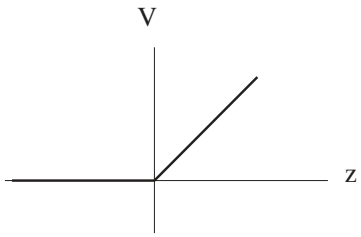

Figure 2.4: A simple max-relation.

Direct applications of this simple relation are one-sided springs. In fig. 2.5 a linear spring is attached to a wall, but not to the cart. Let q denote the position of the cart with respect to the equilibrium of the spring. The spring force $F(q)$ is a nonlinear function of q:

$$F(q) = \begin{cases} -kq, & \text{if } q < 0 \\ 0, & \text{if } q \geq 0 \end{cases} \tag{2.7}$$

with $k > 0$ denoting the spring constant. The interpretation is clear. Only when the spring is pressed ($q < 0$), the spring exerts a nonzero force $-kq$ on the cart. In the other situation where the cart is on the right of the equilibrium ($q \geq 0$), the spring is at rest and the force $F(q)$ is equal to zero. The relation (2.7) can compactly be written as $F(q) = \max(-kq, 0)$. Systems with one-sided springs are studied in e.g. [98].

As a second example consider the following single input control system $\dot{x} = Ax + Bu$ where the control input u is restricted to take nonnegative values only. In [95] one is interested in the existence of a nonnegative state feedback of the form $u = \max(0, Fx)$ where F is a constant row vector resulting in a stable closed loop system $\dot{x} = Ax + B\max(0, Fx)$.

A max-relation also occurs in application of Pontryagin's maximum principle to optimal control problems with control restraint sets being convex polyhedra. The maximum principle yields a two-point boundary problem containing max-relations as shown in [96, 97].

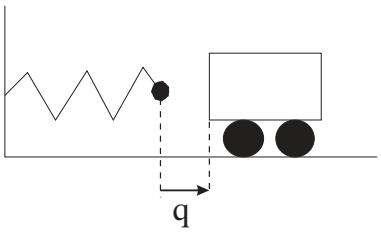

Figure 2.5: One-sided spring.

2.5.2 Piecewise linear (PL) functions

A dynamical system described by an ordinary differential equation and one or more continuous static PL *functions* can be modeled as a complementarity system. To make this plausible, consider the function between v and z as given by fig. 2.6. The function consists of three connected branches with slopes r_i, $i = 1, 2, 3$. The offset at $z = 0$ is equal to g and the slope changes at $z = a_i$, $i = 1, 2$. A description of this function in terms of max-relations is given by (2.8), as is easily verified.

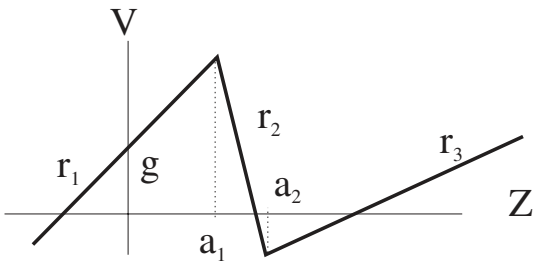

Figure 2.6: An arbitrary PL characteristic.

$$v = g + r_1 z + (r_2 - r_1)\max(z - a_1, 0) + (r_3 - r_2)\max(z - a_2, 0) \tag{2.8}$$

Since the max-relation can be rewritten as a complementarity system, it is obvious that this PL characteristic can be rephrased in terms of a complementarity description.

Applications are for instance saturation and deadzone characteristics (fig. 2.7) which occur in many control systems. Furthermore, devices as bipolar transistors, MOSFET's and p-n junction diodes in electrical network theory are often modeled by PL functions [20, 43, 121].

Finally, it is clear that many continuous nonlinear (static) relation can be suitably approximated by PL functions.

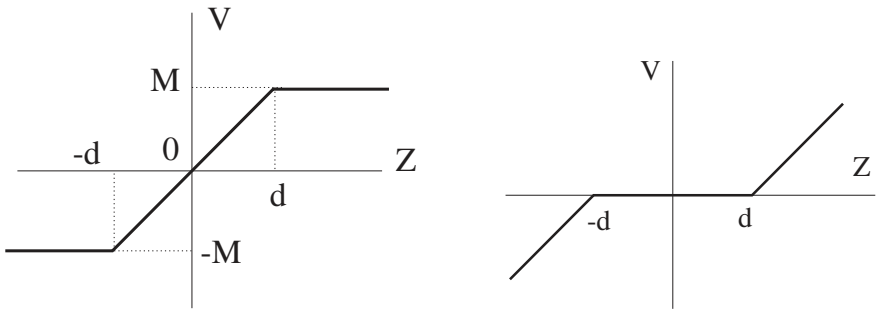

Figure 2.7: Saturation and deadzone characteristic

2.5.3 PL relations

Besides the examples given in the previous subsection, there exist many physically relevant models that are given by PL *relations*, but not by PL *functions*. Examples are mechanical systems with Coulomb friction or relay systems (see fig. 2.8). However, also these systems can be put in a complementarity framework by using an alternative approach. The approach is not given in full detail here, but is sketched by applying it to the example of a Coulomb friction/relay characteristic (see also [112, 123, 160]).

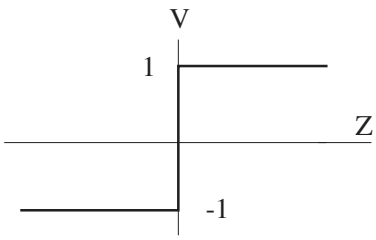

Figure 2.8: Relay or sgn-relation.

The relay characteristic in fig. 2.8 can be described by

$$\begin{aligned} v = 1, \quad & \text{if} \quad z > 0 \\ -1 \le v \le 1, \quad & \text{if} \quad z = 0 \\ v = -1, \quad & \text{if} \quad z < 0, \end{aligned} \tag{2.9}$$

which is sometimes denoted by $v = \mathrm{sgn}(z)$.

Lemma 2.5.1 *The PL relation as given in fig. 2.8 can be described by the equations*

$$u_1 + u_2 = 2 \tag{2.10a}$$
$$y_1 - y_2 = z \tag{2.10b}$$
$$v = \tfrac{1}{2}(u_2 - u_1) \tag{2.10c}$$

together with the complementarity conditions

$$\{u_1 = 0 \text{ or } y_1 = 0\},\ u_1 \geq 0,\ y_1 \geq 0 \tag{2.11}$$
$$\{u_2 = 0 \text{ or } y_2 = 0\},\ u_2 \geq 0,\ y_2 \geq 0. \tag{2.12}$$

□

Proof Due to the complementarity conditions there are $2^2 = 4$ possibilities.

$u_1 = u_2 = 0$: since (2.10a) implies that $2 = 0$, this mode is not feasible.

$u_1 = y_2 = 0$: (2.10a) and (2.10c) give $v = \frac{1}{2}u_2 = 1$. Eq. (2.10b) implies $z = y_1 \geq 0$. This mode corresponds to the right branch in fig. 2.8.

$u_2 = y_1 = 0$: Similar to the previous case, we can derive that this mode corresponds to the left branch.

$y_1 = y_2 = 0$: Eq. (2.10b) implies $z = 0$ and due to (2.10a) and (2.10c) it follows that $-1 \leq v \leq 1$. This corresponds to the middle branch.

Note that in the last mode ($y_1 = y_2 = 0$) the causality between v and z is different then in the other two feasible modes.

The above modeling leads to a complementarity system of the form (2.1), because the algebraic equations (2.10a)-(2.10b) are used. Alternative modeling may lead to a semi-explicit form in case the system Σ (see fig. 2.3) is represented by $\dot{x} = f(x, v)$ and $z = g(x, v)$. Indeed, take

$$u_1 = \tfrac{1}{2}(1 - v) \tag{2.13a}$$
$$y_2 = \tfrac{1}{2}(1 + v) \tag{2.13b}$$
$$z = y_1 - u_2 \tag{2.13c}$$

together with the complementarity conditions on (u_i, y_i). Similarly as in the previous proof, one can check all the four possibilities to verify that the above equations describe the relay characteristic. By suitable substitutions one gets the semi-explicit form

$$\dot{x} = f(x, 1 - 2u_1) \tag{2.14a}$$
$$y_1 = g(x, 1 - 2u_1) + u_2 \tag{2.14b}$$
$$y_2 = 1 - u_1 \tag{2.14c}$$

Other approaches to PL modeling use absolute value functions [44], extended and generalized complementarity problems [37, 201] or state variables [20, 121]. More complicated examples can also be modeled as complementarity systems. Examples can be found in [121], where a "reversed Z-characteristic" has been put in a complementarity system (left picture in fig. 2.9) and in [201], where a model has been derived whose characteristic consists of the edges of a square (right picture).

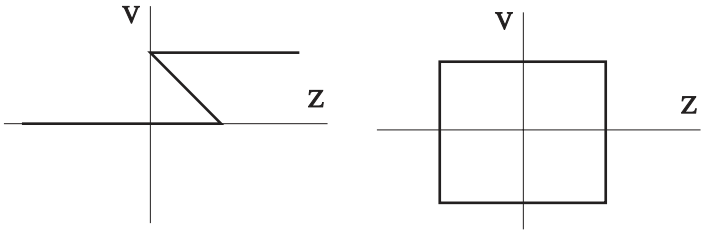

Figure 2.9: Reversed Z-curve and square

Existence and uniqueness of solutions to dynamical systems with PL characteristics are nontrivial. Such well-posedness issues are studied in [37].

2.6 Variable structure systems

2.6.1 Convex definition

Consider a system that switches between two dynamics as a result of inequalities. In fig. 2.10 the state space is separated into two parts by a hypersurface defined by $\phi(x) = 0$. On one side of the surface $C_+ := \{x \in \mathbb{R}^n \mid \phi(x) > 0\}$ the dynamics $\dot{x} = f_+(x)$ holds, on the opposite side $C_- := \{x \in \mathbb{R}^n \mid \phi(x) < 0\}$ the dynamics $\dot{x} = f_-(x)$ is valid.

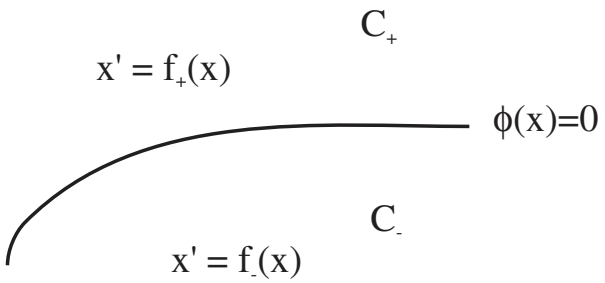

Figure 2.10: Switching dynamics.

A *sliding mode* occurs when in a state x_0, lying on the hypersurface $\phi(x) = 0$, $f_+(x_0)$ points in the direction of C_- and $f_-(x_0)$ points in the direction of C_+ (fig. 2.11).

Hence, from the initial state x_0 it is impossible to go to C_- or C_+, because the dynamics immediately steers you back to the hypersurface satisfying $\phi(x) = 0$. A kind of sliding solution has been formalized by Filippov [68] by the *convex definition* which corresponds to infinitely fast switching. In brief, it states that the sliding mode is given by taking a convex combination of both dynamics $\dot{x} = \lambda f_+(x) + (1-\lambda) f_-(x)$, $0 \leq \lambda \leq 1$ such that x moves along $\phi(x) = 0$.

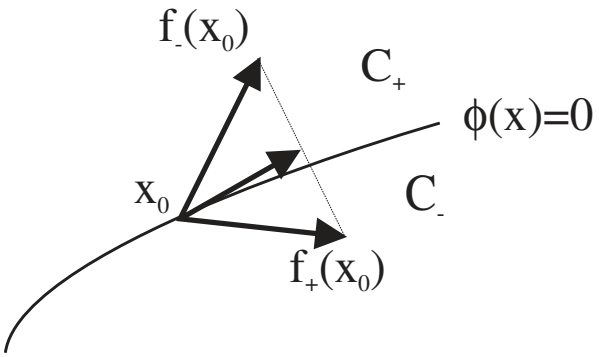

Figure 2.11: Sliding mode.

Proposition 2.6.1 *The variable structure system with solutions according to the convex definition can be modeled by*

$$\dot{x} = \lambda f_+(x) + (1-\lambda) f_-(x) \tag{2.15}$$

and

$$\lambda = 1, \quad \text{if} \quad \phi(x) > 0 \tag{2.16a}$$
$$0 \leq \lambda \leq 1, \quad \text{if} \quad \phi(x) = 0 \tag{2.16b}$$
$$\lambda = 0, \quad \text{if} \quad \phi(x) < 0, \tag{2.16c}$$

i.e. $\lambda = \frac{1}{2} + \frac{1}{2} sgn(\phi(x))$ *with 'sgn' the relation described by (2.9). As seen before, this PL relation allows several complementarity reformulations.* □

Similar techniques as for a single surface, apply to multiple surfaces splitting up the state space.

2.6.2 Equivalent control definition

Another solution concept introduced by Filippov is based on the *equivalent control definition* of sliding modes [68]. This definition is related to "switching control systems." The system given by $\dot{x} = f(x, u)$ with x the state variable is controlled by the

discontinuous feedback (called the "equivalent control")

$$u = \begin{cases} g_+(x), & \xi(x) > 0 \\ g_-(x), & \xi(x) < 0 \end{cases} \tag{2.17}$$

with the function $\xi : \mathbb{R}^n \to \mathbb{R}$ modeling the switching surface. Similar to the previous subsection, a sliding mode occurs when the dynamics $f_+(x) := f(x, g_+(x))$ and $f_-(x) := f(x, g_-(x))$ point outward C_+ and C_-, respectively. The equivalent control definition of a sliding mode picks a convex combination of the control laws instead of a convex combination of $f_+(x)$ and $f_-(x)$ (note that the definitions are equivalent when $f(x, u)$ is affine in u). Formally, the sliding mode is given by the differential and algebraic equations $\dot{x} = f(x, \lambda g_+(x) + (1 - \lambda)g_-(x))$, $\xi(x) = 0$ and valid as long as $\lambda \in [0, 1]$ is satisfied. Obviously, this system can also be modeled as a system $\dot{x} = f(x, \lambda g_+(x) + (1 - \lambda)g_-(x))$ with a characteristic between λ and $\xi(x)$ as in (2.16).

Proposition 2.6.2 *A variable structure system as above with solutions according to the equivalent control definition can be rewritten in terms of a complementarity system.*□

2.7 Optimal control problems with state constraints

An important class of optimal control problems consists of maximizing the criterion $J(x_0, v) := \int_0^T [F(x, v, t)]dt + S(x(T), T)$ by choosing an appropriate control function v subject to the dynamics $\dot{x} = f(x, v, t)$ with initial condition $x(0) = x_0$ and the state constraint $h(x, t) \geq 0$ for all $t \in [0, T]$. Additional requirements like control constraints $g(x, v, t) \geq 0$ and end-point conditions $a(x(T), T) \geq 0$ and $b(x(T), T) = 0$ could be included, but are omitted for brevity.

In the survey [80] Pontryagin's maximum principle [164] is used to obtain necessary conditions for a control input to be optimal.

Introduce the Hamiltonian $H(x, v, \lambda, t) := F(x, v, t) + \lambda^\top f(x, v, t)$. The optimal control v_{opt} satisfies

$$v_{\text{opt}} = \arg\max_v H(x_{\text{opt}}, v, t) \tag{2.18a}$$

$$\dot{x}_{\text{opt}} = \frac{\partial H}{\partial \lambda}(x_{\text{opt}}, v_{\text{opt}}, t) \tag{2.18b}$$

$$\dot{\lambda} = -\frac{\partial H}{\partial x}(x_{\text{opt}}, v_{\text{opt}}, t) - \frac{\partial h}{\partial x}^\top (x_{\text{opt}}, t)u \tag{2.18c}$$

$$y = h(x_{\text{opt}}, t) \tag{2.18d}$$

with complementarity conditions holding between the multiplier u and constraint variables y. The variable λ is called the adjoint or co-state variable. There are additional boundary conditions such that the maximum principle results in a two-point boundary problem. It is possible that jumps occur in the adjoint variable λ. Also for these

jumps additional relations are available. We do not specify all the available conditions, but only illustrate that this kind of optimal control problems fit in the class of complementarity systems.

The formulation in [80] is called an informal theorem, because the result is not rigorously established for the general case. It is presented as a kind of recipe to find possible candidates for the optimal controls.

2.8 Projected dynamical systems

Projected dynamical systems (PDS) have been studied in [62, 147]. These systems are described by differential equations of the form

$$\dot{x}(t) = \Pi_K(x(t), -F(x(t))), \tag{2.19}$$

where F is a vector field, K is a closed convex set, and Π_K is a projection operator that prevents the solution from moving outside the constraint set K. Loosely speaking, a PDS obeys an equation of the form $\dot{x} = -F(x)$ as long as x is contained in the interior of K or $-F(x)$ is "pointing inwards K." When $-F(x)$ is pointing outward and x is at the boundary of K, the operator Π_K projects $-F(x)$ into the direction of K such that the solution stays inside K.

To be precise, the cone of inward normals at $x \in K$ is defined by

$$n(x) = \{\gamma \mid \langle \gamma, x - k \rangle \leq 0 \text{ for all } k \in K\}. \tag{2.20}$$

Given $x \in K$ and $v \in \mathbb{R}^n$, define the projection of the vector v at x with respect to K by

$$\Pi_K(x, v) = v - \langle v, n^*(x) \rangle n^*(x), \tag{2.21a}$$

where

$$n^*(x) \in \arg \max_{n \in n(x),\ \|n\| \leq 1} \langle v, -n \rangle. \tag{2.21b}$$

Definition 2.8.1 The PDS(K, F) is given by

$$\dot{x} = \Pi_K(x, -F(x)). \tag{2.22}$$

□

We consider convex sets K that can be given by finitely many inequalities, i.e. $K = K_h := \{x \in \mathbb{R}^n \mid h(x) \geq 0\}$ with $h : \mathbb{R}^n \to \mathbb{R}^p$ a real-analytic function such that the component functions h_i are convex. ∇h_i denotes the gradient of h_i and is considered to be a row vector. The Jacobian $H(x)$ denotes the matrix in which the i-th row is equal to $\nabla h_i(x)$, i.e. the ij-th element of $H(x)$ is equal to $\frac{\partial h_i}{\partial x_j}(x)$. Moreover, F is assumed to be real-analytic as well. Under suitable assumptions (like a rank condition on the Jacobian $H(x)$ and growth conditions on the vector field $F(x)$, see Chapter 6 for the details) the following result can be proven.

Proposition 2.8.2 *Under suitable assumptions (see Chapter 6) both PDS(K_h, F) and the complementarity system given by*

$$\begin{aligned} \dot{x}(t) &= -F(x(t)) + H^\top(x(t))u(t) && \text{(2.23a)} \\ y(t) &= h(x(t)) && \text{(2.23b)} \\ \{y_i(t) &= 0 \text{ or } u_i(t) = 0\},\ y_i(t) \geq 0,\ u_i(t) \geq 0, && \text{(2.23c)} \end{aligned}$$

have a unique solution defined on $[0, \infty)$ *for any given initial state* x_0*. Moreover, the solutions coincide.* □

PDS are used for studying equilibria of oligopolistic markets, urban transportation networks, traffic systems, international trade, agricultural and energy markets (spatial price equilibria).

2.9 Conclusions

The class of complementarity systems may seem quite restrictive at first sight. The goal of this paper has been to show that this is not the case: a wide variety of interesting discontinuous dynamical systems can be rewritten in a complementarity formalism. Among the applications of complementarity systems are many examples relevant to the systems and control community. We mentioned the switching control systems (variable structure systems), optimal control problems with state and/or control constraints, systems with discontinuous positive feedback and control systems with relays. Furthermore, many challenging questions are still open in the field of control of complementarity systems. These include characterization of stability, controllability, state/output feedback stabilizability and the development of simulation tools. An incentive to continue this line of research is the range of possible applications: control of mechanical systems with Coulomb friction, unilateral constraints and one-sided springs; control of robots; simulations of crash-tests; regulating landing maneuvers of spacecraft; feedback control of dynamical systems with saturating actuators or deadzones; control of traffic systems and economical markets; control, design and verification of switching circuits.

3

Linear Complementarity Systems

This chapter has been published in SIAM Journal on Applied Mathematics [92]. Parts of the chapter have been presented in an abridged form at the IFAC Conference on System Structure and Control in Nantes (France), July 1998 [90] and the Conference on Decision and Control in San Diego (USA), December 1997 [87].

3.1 Introduction

In many technical and economic applications one encounters systems of differential equations and inequalities. For a quick roundup of examples, one may think of the following: motion of rigid bodies subject to unilateral constraints, electrical networks with ideal diodes, optimal control problems with inequality constraints in the states and/or controls, dynamical systems with piecewise linear characteristics like saturation functions, deadzones, relays, Coulomb friction and one-sided springs, projected dynamical systems, dynamic versions of linear and nonlinear programming problems, and dynamic Walrasian economies. It has to be noted that there is considerable inherent complexity in systems of differential equations and inequalities, since nonsmooth trajectories and possibly jumps have to be taken into account. As a result of this, even basic issues such as existence and uniqueness of solutions are difficult to settle. Given the wealth of possible applications however, it is of interest to overcome these difficulties.

In the literature one can find many strands of research dealing with dynamics subject to inequality constraints, some mainly motivated by problems in mechanics, others more closely connected to operations research and economics. The framework of *differential inclusions* (see for instance [9]) gives a general setting for the study of systems in which both differential equations and inequalities play a role. In this chapter, however, we shall be interested in more specific dynamical systems for which uniqueness of solutions holds. Although of course one can get unique solutions from

a differential inclusion by imposing suitable side constraints, we prefer to think of the systems considered in this chapter as systems that switch between modes on the basis of certain inequality constraints, and that behave within each mode as ordinary differential systems rather than as differential inclusions. This "multimodal" way of thinking is natural in a number of applications; in the study of Coulomb friction, one has the transition between stick mode and slip mode, in the study of electrical networks with ideal diodes, there is the transition between the conducting and the blocking mode of each diode, and in the context of dynamic optimization, one has mode transitions when an inactive constraint becomes active or vice versa. A similar point of view may be found in the literature on the so-called "hybrid systems" encompassing both continuous and discrete dynamics, which have recently been a popular subject of study both for computer scientists and for control theorists (see for instance [7, 162]).

Among the studies that have been made of dynamical systems exhibiting some sort of switching behavior, one may mention a number that have been inspired by applications in mechanics [31, 124, 139, 144, 160, 192, 194, 199], in electrical engineering [20, 121], and in operations research [62, 147], as well as general studies such as [68]. The work in this chapter is more general than most of the cited studies in the sense that we do not *a priori* impose conditions on the "index" of the constraints. (The index measures the number of actual constraints following from a given algebraic constraint within the context of a given set of differential equations; the term comes from numerical analysis, see for instance [29].) Our treatment is also general in that we allow an arbitrary finite number of state variables, and an arbitrary finite number of constraints. On the other hand, our work is more restricted, since we consider only linear differential equations; in conjunction with the switching rules, the systems that we study are therefore piecewise linear dynamical systems.

As a consequence of the fact that we are looking at systems of arbitrary index, we have to take into account the possibility of solutions containing *impulses*. The occurrence of such impulses is state-dependent and in this sense our situation is different from the one in [10] where impulses are externally imposed rather than generated by the system itself. One of the main reasons for restricting the development in this chapter to linear dynamics within each mode is the fact that this allows us to treat impulses within a standard distributional framework. Earlier works in the research program that has led to this chapter are [177–179]. The paper [177] uses a solution concept, which is not in accordance with mechanical systems with multiple constraints, while in [179] one considers a nonlinear framework with only *smooth* continuations and no specification of jump or mode switching rules. Finally, in [178] jump and switching rules are given that are only valid for the mechanical case. So, a *complete* specification of the dynamics on a general level is not given so far. Without a complete solution concept, issues of existence and uniqueness of solutions can only be studied partially. The contribution of this chapter is as follows: (i) it gives a complete definition of what is to be understood by a solution of a linear complementarity system; (ii) it gives sufficient conditions for well-posedness of linear complementarity systems, in the sense of existence and uniqueness of solutions; (iii) it presents an effective procedure for generating solutions to linear

complementarity systems. In addition to this, we establish an explicit connection to the literature on mechanical systems that are subject to mode-switching by showing that our formulation agrees with the one of Moreau [144] (see also [31, 139]) for the class of systems covered by both formulations, namely linear mechanical systems.

The chapter is organized as follows. We start with an example, to motivate the ingredients needed for defining a solution concept for complementarity systems. To introduce the notion of solution some mathematical preliminaries as presented in Section 3.3 are required. A definition of the class of linear complementarity systems with its solution concept is given in Section 3.4. The definition relies on a mapping which assigns a "next mode" to each continuous state; several alternative ways of constructing this mapping are discussed in Section 3.5. Sufficient conditions for local existence and uniqueness of solutions follow in Section 3.6. After that, we present a computational example to illustrate the construction of solutions from the definition. In Section 3.8 it will be shown that the proposed solution concept is not an artificial one, but that it complies for linear constrained mechanical systems with the inelastic formulation of Moreau. Finally, conclusions follow in Section 3.9.

In this chapter, the following notational conventions will be in force. $\mathbb{R}$ denotes the real numbers, $\mathbb{R}_+$ the nonnegative real numbers, and $\mathbb{N} := \{0, 1, 2, \dots\}$. For a positive integer l, $\bar{l}$ denotes the set $\{1, 2, \dots, l\}$. If a is a (column) vector with k real components, we write $a \in \mathbb{R}^k$ and denote the ith component by a_i. For two vectors $a, b \in \mathbb{R}^k$, the notation $a \perp b$ means that for all $i \in \bar{k}$ either $a_i = 0$ or $b_i = 0$. Given two vectors $a \in \mathbb{R}^k$ and $b \in \mathbb{R}^l$, then $\mathrm{col}(a, b)$ denotes the vector in $\mathbb{R}^{k+l}$ that arises from stacking a over b. $M \in \mathbb{R}^{m\times n}$ means that M is a real matrix with dimensions $m \times n$. $M^\top$ is the transpose of the matrix M. The kernel of M is denoted by $\mathrm{Ker}\, M$ and the image by $\mathrm{Im}\, M$. Given $M \in \mathbb{R}^{k\times l}$ and two subsets $I \subseteq \bar{k}$ and $J \subseteq \bar{l}$, the (I, J)-submatrix of M is defined as $M_{IJ} := (m_{ij})_{i\in I, j\in J}$. In case $J = \bar{l}$, we also write $M_{I\bullet}$ and if $I = \bar{k}$, we write $M_{\bullet J}$. For a vector a, $a_I := (a_i)_{i\in I}$. The diagonal matrix with diagonal entries $a_1, \dots, a_k$ is denoted by $\mathrm{diag}(a_1, \dots, a_k)$.

The field of rational functions in one indeterminate is denoted by $\mathbb{R}(s)$. Rational vector functions with k components and rational matrices with dimensions $m \times n$ are denoted by $\mathbb{R}^k(s)$ and $\mathbb{R}^{m\times n}(s)$, respectively. For reasons of clarity, we shall systematically use a notation in which vectors over $\mathbb{R}(s)$ are written with an argument s to distinguish between the vector $u \in \mathbb{R}^k$ and the rational vector $u(s) \in \mathbb{R}^k(s)$. A rational matrix is called proper, if for all entries the degree of the numerator is smaller than or equal to the degree of the denominator. A rational matrix is called biproper, if it is square, proper and has a proper inverse. If two rational vectors $u(s), y(s) \in \mathbb{R}^k(s)$ satisfy that for all $i \in \bar{k}$ either $u_i(s) = 0$ or $y_i(s) = 0$, we write $u(s) \perp y(s)$.

The set $C^\infty(\mathbb{R}, \mathbb{R})$ denotes the set of smooth functions, i.e. all functions from $\mathbb{R}$ to $\mathbb{R}$ that are arbitrarily often differentiable. For a smooth function u the i-th derivative is denoted by $u^{(i)}$.

A vector $u \in \mathbb{R}^k$ is called nonnegative, and we write $u \geq 0$, if $u_i \geq 0, i \in \bar{k}$ and positive ($u > 0$), if $u_i > 0, i \in \bar{k}$. If a vector u is not nonnegative, we write $u \ngeq 0$. A sequence of scalars $(u^1, u^2, \dots, u^r)$ is called lexicographically nonnegative,

written as $(u^1, u^2, \ldots, u^r) \succeq 0$, if $(u^1, u^2, \ldots, u^r) = (0, 0, \ldots, 0)$ or $u^j > 0$ where $j := \min\{p \in \bar{r} \mid u^p \neq 0\}$. A sequence of scalars is called lexicographically positive, denoted by $(u^1, u^2, \ldots, u^r) \succ 0$, if $(u^1, u^2, \ldots, u^r) \succeq 0$ and $(u^1, u^2, \ldots, u^r) \neq (0, 0, \ldots, 0)$. For a sequence of vectors $(u^1, u^2, \ldots, u^r)$ with $u^i \in \mathbb{R}^k$, we write $(u^1, u^2, \ldots, u^r) \succeq 0$ when $(u_i^1, u_i^2, \ldots, u_i^r) \succeq 0$ for all $i \in \bar{k}$. Likewise, we write $(u^1, u^2, \ldots, u^r) \succ 0$ when $(u_i^1, u_i^2, \ldots, u_i^r) \succ 0$ for all $i \in \bar{k}$.

For sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \setminus \mathcal{B} := \{x \in \mathcal{A} \mid x \notin \mathcal{B}\}$ and $\mathcal{P}(\mathcal{A})$ denotes the power set of $\mathcal{A}$, i.e. the collection of all subsets of $\mathcal{A}$. For two subspaces V, T of $\mathbb{R}^n$, the notation $V \oplus T = \mathbb{R}^n$ means that V and T form a direct sum decomposition of $\mathbb{R}^n$, i.e. $V + T := \{v + t \mid v \in V,\ t \in T\} = \mathbb{R}^n$ and $V \cap T = \{0\}$.

3.2 Example

Before specifying the class of linear complementarity systems (LCS), we illustrate some of the aspects that play a role in the evolution of such systems by an example of two carts connected by a spring (used also in [177]). The left cart is attached to a wall by a spring. The motion of the left cart is constrained by a completely inelastic stop. The system is depicted in figure 3.1.

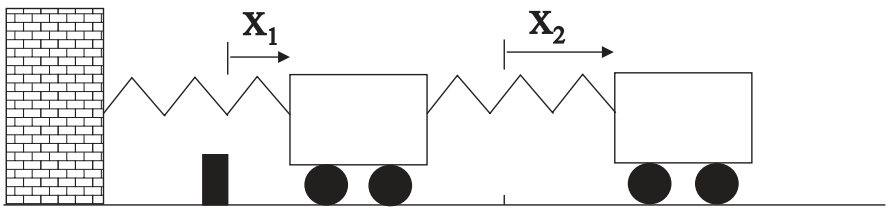

Figure 3.1: Two-carts system.

For simplicity, the masses of the carts and the spring constants are set equal to 1. The stop is placed at the equilibrium position of the left cart. By x_1, x_2 we denote the deviations of the left and right cart, respectively, from their equilibrium positions and x_3, x_4 are the velocities of the left and right cart, respectively. By u, we denote the reaction force exerted by the stop. Furthermore, the variable y is set equal to x_1. Simple mechanical laws lead to the dynamical relations

$$\begin{array}{lcll} \dot{x}_1(t) & = & x_3(t) & \\ \dot{x}_2(t) & = & x_4(t) & \\ \dot{x}_3(t) & = & -2x_1(t) + x_2(t) + u(t) & (3.1) \\ \dot{x}_4(t) & = & x_1(t) - x_2(t) & \\ y(t) & := & x_1(t). & \end{array}$$

To model the stop in this setting, the following reasoning applies. The variable $y(t) = x_1(t)$ should be nonnegative, because it is the position of the left cart with

respect to the stop. The force exerted by the stop can only act in the positive direction implying that $u(t)$ should be nonnegative. If the left cart is not at the stop at time t ($y(t) > 0$), the reaction force vanishes at time t, i.e. $u(t) = 0$. Similarly, if $u(t) > 0$, the cart must necessarily be at the stop, i.e. $y(t) = 0$. This is expressed by the conditions

$$0 \leq y(t) \perp u(t) \geq 0. \tag{3.2}$$

The system can be represented by two modes, depending on whether the stop is active or not. We distinguish between the unconstrained mode ($u(t) = 0$) and the constrained mode ($y(t) = 0$). The dynamics of these modes are given by the following Differential and Algebraic Equations (DAEs)

unconstrained	constrained
$\dot{x}_1(t) = x_3(t)$	$\dot{x}_1(t) = x_3(t)$
$\dot{x}_2(t) = x_4(t)$	$\dot{x}_2(t) = x_4(t)$
$\dot{x}_3(t) = -2x_1(t) + x_2(t)$	$\dot{x}_3(t) = -2x_1(t) + x_2(t) + u(t)$
$\dot{x}_4(t) = x_1(t) + x_2(t)$	$\dot{x}_4(t) = x_1(t) + x_2(t)$
$u(t) = 0$	$y(t) = x_1(t) = 0.$

When the system is represented by either of these modes, the triple (u, x, y) is given by the corresponding dynamics as long as the inequalities in (3.2)

unconstrained	constrained
$y(t) \geq 0$	$u(t) \geq 0$

are satisfied. A mode change is triggered by violation of one of these inequalities. The mode transitions that are possible for the two-carts systems are described below.

- **Unconstrained → Constrained**: The inequality $y(t) \geq 0$ tends to get violated at a time instant $t = \tau$. The left cart hits the stop and stays there. The velocity of the left cart is reduced to zero instantaneously at the time of impact: the kinetic energy of the left cart is totally absorbed by the stop due to a purely inelastic collision. A state for which this happens is, for instance, $x(\tau) = (0, -1, -1, 0)^\top$.
- **Constrained → Unconstrained**: The inequality $u(t) \geq 0$ tends to be violated at $t = \tau$. The right cart is located at or moving to the right of its equilibrium position, so the spring between the carts is stretched and pulls the left cart away from the stop. This happens for example if $x(\tau) = (0, 0, 0, 1)^\top$.
- **Unconstrained → Unconstrained with re-initialization according to constrained mode.** The inequality $y(t) \geq 0$ tends to get violated at $t = \tau$. As an example, consider $x(\tau) = (0, 1, -1, 0)^\top$. At the time of impact, the velocity

of the left cart is reduced to zero just as in the first case. Hence, a state jump (re-initialization) to $(0, 1, 0, 0)^\top$ occurs. The right cart is at the right of its equilibrium position and pulls the left cart away from the stop. Stated differently, from $(0, 1, 0, 0)^\top$ smooth continuation in the unconstrained mode is possible.

This last transition is a special one in the sense that first the constrained mode is active causing the corresponding state jump. After the jump no smooth continuation is possible in the constrained mode resulting in a second mode change back to the unconstrained mode.

From state $x(\tau) = (0, -1, -1, 0)^\top$, we can enter the constrained mode by starting with an instantaneous jump to $x(\tau+) = (0, -1, 0, 0)^\top$. This jump can be modelled as the result of a (Dirac) pulse δ exerted by the stop. In fact, $u = \delta$ results in the state jump $x(\tau+) - x(\tau) = (0, 0, 1, 0)^\top$. This motivates the use of distributional theory as a suitable mathematical framework for describing physical phenomena like collisions with discontinuities in the state vector.

To summarize, the motion of the carts is governed by two systems of Differential and Algebraic Equations (DAEs), called the constrained and the unconstrained mode. A change of mode is triggered by violation of certain inequalities corresponding to the current mode. The time instants at which this occurs, are called "event times." At an event time, the system will switch to a new mode. A mode transition often calls for a state jump or re-initialization. In the example, velocity jumps occur, when the left cart arrives at the stop with negative velocity. In this chapter, the above dynamics will be formalized for the complete class of linear complementarity systems and special attention will be paid to the mode selection problem and well-posedness issues. However, first we recall some facts concerning systems of linear differential and algebraic equations, such as appear in the constrained and unconstrained mode descriptions.

3.3 Mathematical Preliminaries

We consider a linear differential/algebraic system of the form

$$\dot{x}(t) = Kx(t) + Lu(t) \tag{3.3a}$$
$$0 = Mx(t) + Nu(t). \tag{3.3b}$$

The time arguments will often be suppressed for brevity. Throughout this section, $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$. The system parameters K, L, M and N are constant matrices of dimensions $n \times n$, $n \times m$, $r \times n$ and $r \times m$, respectively.

Definition 3.3.1 A state x_0 is said to be *consistent* for (K, L, M, N), if there exist smooth functions u and x such that $x(0) = x_0$ and (3.3) is satisfied. The set of all consistent states for (K, L, M, N) is denoted by $V(K, L, M, N)$ and is called the *consistent subspace*. □

The following sequence of subspaces converges in at most n (dimension of state) steps to $V = V(K, L, M, N)$ (for a proof see [83]):

$$\begin{aligned} V_0 &= \mathbb{R}^n \\ V_{i+1} &= \{x \in \mathbb{R}^n \mid \exists u \in \mathbb{R}^m \text{ such that } Kx + Lu \in V_i,\ Mx + Nu = 0\}. \end{aligned} \tag{3.4}$$

Definition 3.3.2 The quadruple (K, L, M, N) is called *autonomous*, if for every consistent state x_0 the system (3.3) has a unique solution (x, u). □

The system (3.3) is autonomous, if the full-column-rank condition

$$\operatorname{Ker}\begin{bmatrix} L \\ N \end{bmatrix} = \{0\} \tag{3.5}$$

holds together with

$$V(K, L, M, N) \cap T(K, L, M, N) = \{0\} \tag{3.6}$$

where $T(K, L, M, N)$ is the subspace that is obtained as the limit of the sequence

$$\begin{aligned} T_0 &= \{0\} \\ T_{i+1} &= \{x \in \mathbb{R}^n \mid \exists u \in \mathbb{R}^m\ \exists \bar{x} \in T_i \text{ with } x = K\bar{x} + Lu,\ M\bar{x} + Nu = 0\}. \end{aligned} \tag{3.7}$$

This sequence converges in maximally n (dimension of state) steps (proof can be found in [83]). The subspace $T = T(K, L, M, N)$ can be interpreted as the *jump space* associated to (K, L, M, N), i.e. the space along which fast motions will occur that take an inconsistent initial state instantaneously to a point in the consistent subspace V.

To formalize the interpretation of T as a jump space, we introduce the class of impulsive-smooth distributions as studied by Hautus and Silverman [83]. The general form of an impulsive-smooth distribution $\mathtt{u}$ (note the different font used for distributions) is

$$\mathtt{u} = \underbrace{\sum_{i=0}^{l} u^{-i}\delta^{(i)}}_{\mathtt{u}_{imp}} + \mathtt{u}_{reg}, \tag{3.8}$$

where $\delta = \delta^{(0)}$ denotes the delta distribution with support at zero, $\delta^{(r)}$ its r-th distributional derivative, $u^0, u^{-1}, \ldots, u^{-l}$ are coefficients in $\mathbb{R}$ and $\mathtt{u}_{reg}$ is a distribution that can be identified with the restriction to $[0, \infty)$ of some smooth function. The regular part of an impulsive-smooth distribution $\mathtt{u}$ is denoted by $\mathtt{u}_{reg}$ and its impulsive part by $\mathtt{u}_{imp}$. The class of impulsive-smooth distributions will be denoted by C_{imp}. For an element $\mathtt{u}$ of C_{imp} of the form (3.8), we write $\mathtt{u}(0+)$ for the limit value $\lim_{t\downarrow 0} \mathtt{u}_{reg}(t)$.

Having introduced the class C_{imp}, we can replace the system of equations (3.3) by its distributional version

$$\begin{array}{rcl} \dot{\mathtt{x}} & = & K\mathtt{x} + L\mathtt{u} + x_0\delta \\ 0 & = & M\mathtt{x} + N\mathtt{u} \end{array} \tag{3.9}$$

in which the initial condition x_0 appears explicitly, and we can look for a solution of (3.9) in the class of vector-valued impulsive-smooth distributions. In [83] it is shown that under the conditions (3.5) and (3.6) there exists a unique solution $(\mathtt{u}, \mathtt{x}) \in C_{imp}^{m+n}$ to (3.9) for all $x_0 \in V + T$; moreover, the solution is such that $\mathtt{x}(0+)$ is equal to $P_V^T x_0$, the projection of x_0 onto V along the jump space T. In fact, $\mathtt{x}(0+)$ depends only on the impulsive part of $\mathtt{u}$: if $\mathtt{u}_{imp} = \sum_{i=0}^{l} u^{-i}\delta^{(i)}$, then

$$\mathtt{x}(0+) = x_0 + \sum_{i=0}^{l} K^i L u^{-i}. \tag{3.10}$$

Lemma 3.3.3 *Consider the system (3.3) and suppose that the number of inputs (m) equals the number of constraints (r). Then the following statements are equivalent.*

1. (K, L, M, N) *is autonomous.*
2. *The system (3.9) admits a unique impulsive-smooth distribution for each initial condition.*
3. $V(K, L, M, N) \oplus T(K, L, M, N) = \mathbb{R}^n$ *and* $\mathrm{Ker}\begin{bmatrix} L \\ N \end{bmatrix} = \{0\}$.
4. $G(s) := M(sI - K)^{-1}L + N$ *is invertible as a rational matrix.*

□

Proof. The implication $2 \Rightarrow 1$ follows from the definition of an autonomous system. The quadruple (K, L, M, N) is autonomous iff the system $\Sigma : \dot{x} = Kx + Lu$, $y = Mx + Nu$ is left invertible in the sense of [83]. In [83], it is proven that the statements

- the system Σ is left invertible
- $V(K, L, M, N) \cap T(K, L, M, N) = \{0\}$ and $\mathrm{Ker}\begin{bmatrix} L \\ N \end{bmatrix} = \{0\}$
- $G(s)$ is left invertible

are equivalent. Since $G(s)$ is assumed to be square ($m = r$), left invertibility is the same as invertibility. Hence, $1 \Rightarrow 4$. According to [83, Thm. 3.24], invertibility of $G(s)$ implies additionally that $V(K, L, M, N) \oplus T(K, L, M, N) = \mathbb{R}^n$. This proves $4 \Rightarrow 3$. Finally, $3 \Rightarrow 2$ is a consequence of the fact that the assumptions (3.5)-(3.6)

imply that there is a unique solution $(\mathrm{u}, \mathrm{x}) \in C^{m+n}_{imp}$ to (3.9) for all $x_0 \in V + T$, as mentioned earlier. Since $V + T$ is equal to $\mathbb{R}^n$, this implies 2. □

The systems studied in this chapter are described by standard state space equations of linear systems together with complementarity conditions as in the complementarity problems of mathematical programming. Therefore some concepts from complementarity theory will be recalled briefly. The Linear Complementarity Problem (LCP) [47] is defined as follows.

Given a matrix $M \in \mathbb{R}^{k\times k}$ and $q \in \mathbb{R}^k$, find $u, y \in \mathbb{R}^k$ such that

$$y = q + Mu \tag{3.11}$$
$$0 \le y \perp u \ge 0 \tag{3.12}$$

This problem is denoted by LCP(q, M).

Let a matrix M of size $k \times k$ and two subsets I and J of $\bar{k}$ of the same cardinality be given. The (I, J)-*minor* of M is the determinant of the square matrix $M_{IJ} := (m_{ij})_{i\in I, j\in J}$. The (I, I)-minors are also known as the principal minors. M is called a *P-matrix*, if all principal minors are positive. A square matrix M is said to be *positive definite*, if $x^\top Mx > 0$ for all nonzero $x \in \mathbb{R}^n$. Note that a positive definite matrix is not necessarily symmetric according to this definition.

We state the following results.

Theorem 3.3.4 *For given M, the problem LCP(q, M) has a unique solution for all vectors q if and only if M is a P-matrix.* □

Proof. See [47, Thm. 3.3.7]. □

Theorem 3.3.5 *A positive definite matrix is a P-matrix.* □

Proof. [47, Thm. 3.1.6 and Thm. 3.3.7]. □

3.4 Linear Complementarity Systems

In this section, we introduce linear complementarity systems (LCS) and formulate the notion of solution for such systems.

A linear complementarity system is governed by the simultaneous equations

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{3.13a}$$
$$y(t) = Cx(t) + Du(t) \tag{3.13b}$$
$$0 \le y(t) \perp u(t) \ge 0. \tag{3.13c}$$

The notation in (3.13c) is consistent with the notation used in complementarity problems in mathematical programming (see the formulation of the linear complementarity problem in section 3.3). In this section, we will describe how the relations above have to be interpreted to arrive at a notion of solution to such a complementarity system. The functions u, x and y take values in $\mathbb{R}^k$, $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively; A, B, C and D are constant matrices of appropriate dimensions. Note that the dimensions of the variables $y(t)$ and $u(t)$ are the same. Equation (3.13c) states that for every component $i = 1, \ldots, k$ either $u_i(t) = 0$ or $y_i(t) = 0$. The set of indices for which $y_i(t) = 0$, called the *mode* or *active index set*, may change during the time evolution of the system. The system may therefore switch from one 'operation mode' to another. To define the dynamics of (3.13) completely, one has to specify when the mode switches occur, what their effect will be on the state variables, and how a new mode will be selected. We will do this below, extending earlier treatments in [177] (where only systems with a single constraint were considered ($k = 1$), see also Example 3.8.3 for a comparison of the mode selection criteria) and [179], which only treated existence and uniqueness of *smooth* continuations while impulsive motions and re-initialization rules were left out of consideration and only a limited discussion of mode selection criteria could be given. A generalization from smooth to impulsive-smooth continuations is not straightforward. The interpretation of the inequalities for impulsive motions is not obvious. A requirement of such an interpretation will be that it must comply with physical laws for 'real-life' systems included in the class of complementarity systems. In this section, we will formalize a distributional interpretation of the inequalities that agrees with Moreau's re-initialization rules for linear mechanical systems (see Section 3.8).

The system has 2^k modes. Each mode is characterized by the active index set $I \subseteq \bar{k}$, which indicates that $y_i = 0, i \in I$ and $u_i = 0, i \in I^c$ where $I^c := \bar{k} \setminus I = \{i \in \bar{k} \mid i \notin I\}$. For each such mode the laws of motion are given by systems of Differential and Algebraic Equations (DAEs). Specifically, in mode I they are given by

$$\left\{\begin{array}{lcl} \dot{x}(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) + Du(t) \\ y_i(t) & = & 0,\ i \in I \\ u_i(t) & = & 0,\ i \in I^c, \end{array}\right. \tag{3.14}$$

or equivalently,

$$\left\{\begin{array}{lcl} \dot{x}(t) & = & Ax(t) + B_{\bullet I} u_I(t) \\ 0 & = & C_{I\bullet} x(t) + D_{II} u_I(t) \\ y_{I^c}(t) & = & C_{I^c\bullet} x(t) + D_{I^c I} u_I(t) \\ u_{I^c}(t) & = & 0 \\ y_I(t) & = & 0 \end{array}\right. \tag{3.15}$$

The set of consistent states for mode I equals $V(A, B_{\bullet I}, C_{I\bullet}, D_{II})$ and is denoted by V_I. The jump space is given by $T_I := T(A, B_{\bullet I}, C_{I\bullet}, D_{II})$. We call mode I *autonomous*, if the quadruple $(A, B_{\bullet I}, C_{I\bullet}, D_{II})$ is autonomous. A standing assumption in the rest of this chapter will be the following.

Assumption 3.4.1 All modes of the complementarity system (3.13) are autonomous. □

By Lemma 3.3.3 this is equivalent to saying that $G_{II}(s) := C_{I\bullet}(sI - A)^{-1}B_{\bullet I} + D_{II}$ is invertible for each index set $I \subseteq \bar{k}$. Note that the notation $G_{II}(s)$ is consistent in the sense that $G_{II}(s)$ is the (I, I)-submatrix of the rational matrix $G(s) := C(sI - A)^{-1}B + D$. Again by Lemma 3.3.3, assumption 3.4.1 implies that $V_I \oplus T_I = \mathbb{R}^n$ for all $I \subseteq \bar{k}$ and that (3.14) has a unique impulsive-smooth solution for all individual modes given an arbitrary initial state.

3.4.1 Continuous phase

Definition 3.4.2 Given $x_0 \in \mathbb{R}^n$ and $I \subseteq \bar{k}$, we denote the unique distributional solution to (3.14) for mode I and initial state x_0 by $(\mathrm{u}^{x_0,I}, \mathrm{x}^{x_0,I}, \mathrm{y}^{x_0,I}) \in \mathcal{C}_{imp}^{k+n+k}$. □

According to [83, Thm. 3.10], there exists a linear mapping F_I such that (3.14) is satisfied for $x_0 \in V_I$ by taking $u(t) = F_I x(t)$. Substituting this feedback in (3.14) transforms the DAE into an ordinary differential equation (ODE). Hence, the regular part of an impulsive-smooth solution u satisfying (3.14) for a given initial state is a *Bohl function*, i.e. u_{reg} is of the form

$$\mathrm{u}_{reg}(t) = \begin{cases} 0 & (t < 0) \\ Ee^{Gt}v & (t \geq 0) \end{cases} \tag{3.16}$$

for real matrices E, G and a vector v depending on the initial state and the specific mode I.

3.4.2 Re-initialization

If initial states of (3.14) are not consistent, i.e. if $x_0 \notin V_I$, then a re-initialization of the initial state will be necessary as pointed out in Section 3.3. Indeed, if $x_0 \notin V_I$, then the solution to (3.14) will contain a nontrivial impulsive part resulting in an instantaneous jump or re-initialization of the state variable. As discussed in Section 3.3, the re-initialized vector $\mathrm{x}^{x_0,I}(0+)$ is equal to the projection of x_0 onto the consistent subspace V_I along the jump space T_I. That is $\mathrm{x}^{x_0,I}(0+) := P_I x_0$, where P_I is the projection operator $P_{V_I}^{T_I}$.

3.4.3 Event detection

Suppose that the current time, state, and mode are $\tau = 0$, x_0, and I, respectively. Note that due to the time-invariance of the system description (3.13), the assumption $\tau = 0$ is just a normalization. The system (3.13) will be represented by (3.14) for mode I as long as the inequalities in (3.13c)

$$\mathrm{u}_{reg}^{x_0,I}(t) \geq 0 \text{ and } \mathrm{y}_{reg}^{x_0,I}(t) \geq 0 \tag{3.17}$$

are satisfied for $t \geq \tau$. The function $\theta : \mathbb{R}^n \times \mathcal{P}(\bar{k}) \to \mathbb{R}_+$ gives the length of the time interval during which the system evolves in mode I from initial state x_0. Note that we only consider the regular part here. In formal terms, θ is defined as follows.

Definition 3.4.3 The time-to-next-event function $\theta : \mathbb{R}^n \times \mathcal{P}(\bar{k}) \to \mathbb{R}_+$ is defined as

$$\theta(x_0, I) := \inf\{t > 0 \mid \mathrm{u}_{reg}^{x_0,I}(t) \not\geq 0 \text{ or } \mathrm{y}_{reg}^{x_0,I}(t) \not\geq 0\}$$

with the convention $\inf \emptyset = \infty$. □

The next event time after time τ will be $\tau + \theta(x(\tau), I)$ (by time-invariance), when the mode and the state at time τ are equal to I and $x(\tau)$, respectively. Since smooth continuation is not possible in mode I after the event time $\tau + \theta(x(\tau), I)$, a transition to another mode must occur. An important aspect of the solution concept will be how to select the new mode.

To illustrate the definition of θ, consider Example 3.4.4 and 3.4.5 of the two-carts system in the next subsection. In these cases, $\theta((0, -1, 0, 0)^\top, \{1\}) = \frac{\pi}{2}$ and $\theta((0, 1, -1, 0)^\top, \{1\}) = 0$.

3.4.4 Mode selection

The mode selection procedure that we propose is based on the concept of *initial solution*. Loosely speaking, an initial solution with initial state x_0 is a triple $(\mathrm{u}, \mathrm{x}, \mathrm{y}) \in C_{imp}^{k+n+k}$ satisfying (3.14) for some mode I and satisfying (3.17) either on a time interval of positive length or on a time instant at which delta distributions are active. The idea is that an initial solution is a starting trajectory for the "global" solution to (3.13).

Example 3.4.4 Consider the two-carts system with initial state $(0, -1, 0, 0)^\top$. The solution to the constrained mode is $u(t) = \cos t$ and $y(t) = 0$. Hence, it satisfies (3.14) for $I = \{1\}$ on $[0, \infty)$ and (3.17) on $[0, \frac{\pi}{2})$. So, this solution satisfies (3.13) on $[0, \frac{\pi}{2})$. Therefore we admit selection of the constrained mode ($I = \{1\}$) as smooth continuation in this mode is possible. □

Example 3.4.5 From the initial state $x_0 = (0, 1, -1, 0)^\top$ first a state jump occurs to $P_{\{1\}}x_0 = (0, 1, 0, 0)^\top$ governed by the laws of the constrained mode, but no smooth continuation is possible in the constrained mode. Solving the dynamics corresponding to the constrained mode, i.e. (3.14) with $I = \{1\}$, gives $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ with $\mathrm{u} = \delta + \mathrm{u}_{reg}$, where $\mathrm{u}_{reg}(t) = -\cos t$. Although (3.17) is not satisfied on a positive time interval, incorporation of this solution in the definition of initial solutions seems well-motivated on physical grounds. We admit selection of $I = \{1\}$. □

We now make the notion of initial solution more precise. Given an impulsive-smooth distribution $\mathrm{v} \in C_{imp}$, we define the leading coefficient of its impulsive part

by

$$\text{lead}(\mathrm{v}) := \begin{cases} 0, & \text{if } \mathrm{v}_{imp} = 0 \\ v^{-l}, & \text{if } \mathrm{v}_{imp} = \sum_{i=0}^{l} v^{-i}\delta^{(i)} \text{ with } v^{-l} \neq 0. \end{cases} \tag{3.18}$$

Definition 3.4.6 We call a scalar-valued impulsive-smooth distribution $\mathrm{v} \in C_{imp}$ *initially nonnegative*, if

$$\begin{cases} \text{lead}(\mathrm{v}) > 0, & \text{if } \mathrm{v}_{imp} \neq 0 \\ \text{there exists } \varepsilon > 0 \text{ such that for all } t \in [0, \varepsilon)\ \mathrm{v}_{reg}(t) \geq 0, & \text{otherwise.} \end{cases}$$

A vector-valued impulsive-smooth distribution in C_{imp}^k is called initially nonnegative, if each of its components is initially nonnegative. We call an impulsive-smooth distribution u *initially positive*, if u is initially nonnegative and additionally if u_i is regular, then for some $\varepsilon > 0$ $\mathrm{u}_i(t) > 0$, $t \in (0, \varepsilon)$. □

Definition 3.4.7 We call $(\mathrm{u}, \mathrm{x}, \mathrm{y}) \in C_{imp}^{k+n+k}$ an *initial solution* to (3.13) with initial state x_0, if

1. there exists an $I \subseteq \bar{k}$ such that $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ satisfies (3.14) with initial state x_0 in the distributional sense; and
2. u, y are initially nonnegative.

□

Given a state x_0, define the set $\mathcal{I}(x_0)$ by

$$\begin{aligned} \mathcal{I}(x_0) := \{J \subseteq \bar{k} \mid &\text{ there exists an initial solution } (\mathrm{u}, \mathrm{x}, \mathrm{y}) \text{ to (3.13) with} \\ &\text{initial state } x_0 \text{ such that } \mathrm{u}_i = 0,\ i \in J^c \text{ and } \mathrm{y}_i = 0,\ i \in J\}. \end{aligned} \tag{3.19}$$

The set $\mathcal{I}(x_0)$ denotes the set of all possible modes in which an initial solution exists with initial state x_0.

Remark 3.4.8 There may be more than one mode corresponding to a given initial solution $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ to (3.13). With the index set I defined by

$$J := \{i \in \bar{k} \mid \mathrm{u}_i \neq 0\}, \tag{3.20}$$

the complementarity conditions require $\mathrm{y}_i = 0$ for $i \in J$. Hence, $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ is an initial solution in mode J. Consider now the "undetermined index set"

$$K := \{i \in \bar{k} \mid \mathrm{u}_i = 0 \text{ and } \mathrm{y}_i = 0\}.$$

Any mode $J \subseteq I \subseteq J \cup K$ may also be selected and the initial solution $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ satisfies (3.14) for $I = J$ with initial state x_0 as well. As an example consider $x_0 = 0$.

In this case, $(\mathtt{u}, \mathtt{x}, \mathtt{y}) = 0$ is a possible initial solution. J and K as defined above are equal to $\emptyset$ and $\bar{k}$, respectively. Consequently, mode I can be chosen arbitrarily, which means that this initial solution satisfies the mode dynamics for each mode. For a given initial solution, the freedom in the choice of the mode corresponding to this solution is exactly characterized by the undetermined index set. □

Remark 3.4.9 If an initial solution $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ has a nontrivial impulsive part, it can be the case that the corresponding mode is only valid for the time instant 0 itself. This happens when the smooth part $(\mathtt{u}_{reg}, \mathtt{y}_{reg})$ are not initially nonnegative. An example is provided by Example 3.4.5, which explains also the special mode transition as mentioned in Section 3.2. The constrained mode $(\mathcal{I}((0, 1, -1, 0)^\top) = \{\{1\}\})$ is selected only for the re-initialization of the state $(\theta((0, 1, -1, 0)^\top, \{1\}) = 0)$. From the re-initialized state $P_{\{1\}}(0, 1, -1, 0)^\top = (0, 1, 0, 0)^\top$ (see also Section 3.7) a new mode is selected $(\mathcal{I}((0, 1, 0, 0)^\top) = \{\emptyset\})$. In the unconstrained mode a smooth initial solution exists with the re-initialized state $(0, 1, 0, 0)^\top$ as initial state. □

3.4.5 Solution concept

We are now in a position to define a solution concept for (3.13). A point $\tau \in \mathcal{E} \subset \mathbb{R}$ is called a right-accumulation point of $\mathcal{E}$, if there exists a sequence $\{\tau_i\}_{i\in\mathbb{N}}$ such that $\tau_i \in \mathcal{E}$ and $\tau_i < \tau$ for all i and furthermore, $\lim_{i\to\infty} \tau_i = \tau$. A left-accumulation point is defined similarly by interchanging "$<$" by "$>$." A set $\mathcal{E} \subset \mathbb{R}$ is called right-isolated, if it contains no left-accumulation points. We call $\tau \in \mathcal{E}$ isolated, if it is not an accumulation point of $\mathcal{E}$.

Definition 3.4.10 A *solution* to (3.13) on $[0, T_e)$, $T_e > 0$, with initial state x_0, is a quadruple $(\mathcal{E}, x_c, u_c, y_c)$, where $\mathcal{E}$, the set of event times, is a right-isolated closed subset of $[0, T_e)$ with empty interior and

$$\begin{aligned} x_c : &\quad (0, T_e) \setminus \mathcal{E} \quad \to \mathbb{R}^n \\ u_c : &\quad (0, T_e) \setminus \mathcal{E} \quad \to \mathbb{R}^k \\ y_c : &\quad (0, T_e) \setminus \mathcal{E} \quad \to \mathbb{R}^k, \end{aligned}$$

being arbitrarily often differentiable that satisfies the following.

1. $0 \in \mathcal{E}$
2. For $\tau \in \mathcal{E}$, $x_c(\tau+) := \lim_{t\downarrow\tau, t\notin\mathcal{E}} x_c(t) = \lim_{i\to\infty} z_i$, where $\{z_i\}_{i\in\mathbb{N}}$ satisfies

$$\begin{cases} z_{i+1} = P_{I_{i+1}} z_i \\ I_{i+1} \in \mathcal{I}(z_i) \end{cases} \tag{3.21}$$

and

$$z_0 := \begin{cases} x_c(\tau-) := \lim_{t\uparrow\tau, t\notin\mathcal{E}} x_c(t), & \text{if } \tau > 0 \\ x_0, & \text{if } \tau = 0. \end{cases} \tag{3.22}$$

3. For isolated $\tau \in \mathcal{E}$ there exists an $I \in \mathcal{I}(x_c(\tau+))$ such that

$$\tau^* := \min\{t > \tau \mid t \in \mathcal{E}\} = \tau + \theta(x_c(\tau+), I) > 0 \tag{3.23}$$

and $(u_c(t), x_c(t), y_c(t))$ satisfies (3.14) for mode I and for $t \in (\tau, \tau^*)$.

□

$P_{I_{i+1}}$ denotes the projection operator corresponding to mode I_{i+1} as introduced in Subsection 3.4.2. The definition requires that the limits in item 2 and in the first case of (3.22) exist.

The set $\mathcal{E}$ specifies the event times, i.e. the times at which there is a change of mode. Two successive isolated event times (τ and τ^*) are related by 3 in terms of the time-to-next-event function θ (Definition 3.4.3). This requirement is included in the solution concept to exclude redundant event times. The triple (x_c, u_c, y_c) denotes the trajectories in the continuous phases of the complementarity system (as imposed by item 3). Item 2 links the continuous phases at the event times by a series of mode selections and re-initializations. The *multiplicity* $m(\tau)$ of the event time $\tau \in \mathcal{E}$ is defined as the $\min\{i \in \mathbb{N} \mid z_i = x_c(\tau+)\}$, i.e. the number of re-initializations needed before smooth continuation (a continuous phase) is possible. In case $m(\tau) = \infty$, one needs a limiting operation to determine the state just after the event, $x_c(\tau+)$. If $m(\tau)$ is finite, then only a finite number of mode selections and re-initializations (projections) in (3.21) are needed. Item 2 specifies also the initial conditions.

Remark 3.4.11 In the literature of hybrid dynamical systems it is often assumed that only a finite number of events exists in a finite time interval. Solutions with this property are sometimes called *non-Zeno* solutions. The relaxation of our solution concept is twofold. First, we allow that there are infinitely many mode switchings and re-initializations at one time instant. Second, right-accumulation points of event times are included. We incorporate solutions that could be called *right-Zeno* to be consistent with the literature on hybrid systems. As an example of a right-Zeno solution consider the example of a bouncing ball with elastic impacts (with restitution coefficient smaller than one). This system has a right-accumulation point, because the ball is at rest within a finite time span but after infinitely many bounces. Since our solution concept complies with mechanical systems with inelastic impacts (see Section 3.8), the bouncing ball example does not fit in the class of systems that we study, but it indicates that there exist models of physical systems that require right-Zeno solutions. An example of a complementarity system allowing right-Zeno solutions is provided by a time reversed version of a system studied by Filippov [68, p. 116], i.e.

$$\dot{x}_1 = -\text{sgn}(x_1) + 2\text{sgn}(x_2) \tag{3.24a}$$

$$\dot{x}_2 = -2\text{sgn}(x_1) - \text{sgn}(x_2), \tag{3.24b}$$

where "sgn" denotes the signum-function given by $\text{sgn}(x) = 1$, if $x > 0$ and $\text{sgn}(x) = -1$, if $x < 0$. Because this system consists of two relay characteristics, it can be modelled as a linear complementarity system (see Chapter 2). Solutions of this piecewise

constant systems are spiraling towards the origin, which is an equilibrium point. Since $\frac{\mathrm{d}}{\mathrm{d}t}(\mid x_1(t) \mid + \mid x_2(t) \mid) = -2$, solutions reach the origin in finite time. However, solutions cannot arrive at the origin without going through an infinite number of mode transitions; since these mode switches occur in a finite time interval, the event times contain a right-accumulation point (i.e. the time that the solution reaches the origin) after which the solution stays at zero. Left-accumulation points are excluded from Definition 3.4.10 due to the requirement that the event set $\mathcal{E}$ is right-isolated. However, note that the time-reverse of the system (3.24) (which is the original example in [68]) has (infinitely many) left-Zeno solutions corresponding to initial state $x_0 = 0$ in a generalized solution concept that admits left-accumulation points. Such a generalized solution concept results in a nondeterministic system and nonuniqueness of solutions, which is undesirable from a point of view of modelling and simulation. In the solution concept of Definition 3.4.10 the only solution emanating from the origin in Filippov's original example is the zero solution. □

Before we present conditions on the complementarity system to guarantee the existence and uniqueness of solutions, two algebraic mode selection procedures will be introduced.

3.5 Mode selection methods

An essential problem in the definition of the solution concept and in the time simulation of complementarity systems is to find the set of possible continuation modes $\mathcal{I}(x_0)$ for a given state x_0. In fact, this is the construction of a (possibly multi-valued) map from the continuous state space $\mathbb{R}^n$ to the discrete space $\mathcal{P}(\bar{k})$. The determination of $\mathcal{I}(x_0)$ in the previous section is based on finding all initial solutions and the corresponding modes. In this section, we obtain two alternative representations of $\mathcal{I}(x_0)$ that do not require the solution of differential equations.

3.5.1 Rational complementarity problem

As noticed in Section 3.4, the solutions to (3.14) are impulsive-smooth distributions whose regular parts are Bohl functions. Such "Bohl distributions" have rational Laplace transforms. Specifically, the Laplace transform $\hat{\mathrm{u}}(s)$ of $\mathrm{u} = \sum_{i=0}^{l} u^{-i}\delta^{(i)} + \mathrm{u}_{reg}$ with u_{reg} as in (3.16) equals [82]

$$\hat{\mathrm{u}}(s) = \sum_{i=0}^{l} u^{-i}s^i + E(sI - G)^{-1}v \ .$$

Observe that the polynomial part of the Laplace transform corresponds to the impulsive part and the strictly proper part to the regular part of the Bohl distribution.

Lemma 3.5.1 *Let* $\mathrm{v} = \sum_{i=0}^{l} v^{-l}\delta^{(i)} + \mathrm{v}_{reg} \in C_{imp}$ *be a Bohl distribution. The following statements are equivalent.*

1. $\mathtt{v}$ *is initially nonnegative.*
2. *There exists a* $\sigma_0 \in \mathbb{R}$ *such that the Laplace transform* $\hat{\mathtt{v}}(s)$ *satisfies* $\hat{\mathtt{v}}(\sigma) \geq 0$ *for all* $\sigma \in \mathbb{R}, \sigma \geq \sigma_0$.
3. *The sequence given by* $(v^{-l}, v^{-l+1}, \ldots, v^0, \mathtt{v}_{reg}(0), \mathtt{v}^{(1)}_{reg}(0), \mathtt{v}^{(2)}_{reg}(0), \ldots)$ *is lexicographically nonnegative.*

Also the following statements are equivalent.

1. $\mathtt{v}$ *is the zero distribution.*
2. *The Laplace transform* $\hat{\mathtt{v}}(s)$ *is the zero function.*
3. *The sequence given by* $(v^{-l}, v^{-l+1}, \ldots, v^0, \mathtt{v}_{reg}(0), \mathtt{v}^{(1)}_{reg}(0), \mathtt{v}^{(2)}_{reg}(0), \ldots)$ *is the zero sequence.*

□

Proof. Evident. □

Let $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ be an initial solution to (3.13) with initial state x_0. The Laplace transforms of $\mathtt{u}$, $\mathtt{y}$, denoted by $\hat{\mathtt{u}}(s)$, $\hat{\mathtt{y}}(s)$, are rational and satisfy

$$\hat{\mathtt{y}}(s) = C(sI - A)^{-1}x_0 + [C(sI - A)^{-1}B + D]\hat{\mathtt{u}}(s) \text{ and } \hat{\mathtt{y}}(s) \perp \hat{\mathtt{u}}(s) \tag{3.25}$$

for all $i \in \bar{k}$; moreover there exists a $\sigma_0 \in \mathbb{R}$ such that

$$\hat{\mathtt{y}}(\sigma) \geq 0, \ \hat{\mathtt{u}}(\sigma) \geq 0 \tag{3.26}$$

for all $\sigma \in \mathbb{R}$, $\sigma \geq \sigma_0$. The converse is true as well, so the Laplace transforms are rational and satisfy (3.25)-(3.26) iff the corresponding time functions define an initial solution to (3.13).

The above observations result in the formulation of the *Rational Complementarity Problem* (terminology introduced in [179]). Note that the formulation of the RCP here is a relaxation of the one in [179], because we allow general rational solutions.

Rational Complementarity Problem. (RCP(x_0)) Let (A, B, C, D) and initial state x_0 be given. Find rational vector functions $y(s)$ and $u(s)$ such that the equalities

$$y(s) = C(sI - A)^{-1}x_0 + [C(sI - A)^{-1}B + D]u(s) \text{ and } y(s) \perp u(s)) \tag{3.27}$$

hold for all $i \in \bar{k}$, and there exists a $\sigma_0 \in \mathbb{R}$ such that for all $\sigma \geq \sigma_0$ we have

$$y(\sigma) \geq 0, \ u(\sigma) \geq 0. \tag{3.28}$$

If $(u(s), y(s))$ is a solution to RCP(x_0), any index set $J \subseteq \bar{k}$ satisfying $u_{J^c}(s) = 0$ and $y_J(s) = 0$ represents a mode J in which an initial solution exists. Hence, it is easily observed that due to the one-to-one relation between initial solutions and solutions to the corresponding RCP the set of possible continuation modes $\mathcal{S}(x_0)$ must be equal to $\mathcal{S}_{\mathrm{RCP}}(x_0)$, where

$$\mathcal{S}_{\mathrm{RCP}}(x_0) = \{I \subseteq \bar{k} \mid \exists (u(s), y(s)) \text{ solution to RCP}(x_0) \text{ such that } u_{I^c}(s) = 0 \text{ and } y_I(s) = 0\}. \quad (3.29)$$

A second algebraic mode selection method can be derived by using the power series expansion of the solutions to RCP(x_0). This is described next.

3.5.2 Linear dynamic complementarity problem

If $(u(s), y(s))$ is a solution to RCP(x_0), then it necessarily has to satisfy $u_{I^c}(s) = 0$ and $y_I(s) = 0$ for some $I \subseteq \bar{k}$. Consequently,

$$\begin{aligned} 0 &= R_{I\bullet}(s)x_0 + G_{II}(s)u_I(s) \\ y_{I^c}(s) &= R_{I^c\bullet}(s)x_0 + G_{I^cI}(s)u_I(s), \end{aligned}$$

where $G(s)$ is the proper transfer function $C(sI - A)^{-1}B + D$, and $R(s)$ is the strictly proper rational matrix $C(sI - A)^{-1}$. Note that $G_{II}(s)$ is invertible by Assumption 3.4.1. This implies that $u_I(s) = -G_{II}^{-1}(s)R_{I\bullet}(s)x_0$ and

$$y_{I^c}(s) = [R_{I^c\bullet}(s) - G_{I^cI}(s)G_{II}^{-1}(s)R_{I\bullet}(s)]x_0.$$

It follows from the representation theory of rational matrix functions (see for instance [116]) that the degree of the polynomial part of $G_{II}^{-1}(s)$ is at most n. Hence, the polynomial parts of the rational functions $u(s)$ and $y(s)$ have degree at most $n - 1$. In terms of time-domain solutions, this means that only derivatives of the Dirac function up to order $n - 1$ can appear in initial solutions. So we can write

$$y(s) = \sum_{i=-n+1}^{\infty} y^i s^{-i} \quad (3.30)$$

and likewise for $u(s)$. To translate the nonnegativity conditions (3.28) to the coefficients of the power series expansion around infinity, we use that $y(s)$ is nonnegative for all sufficiently large real s, if and only if

$$(y^{-n+1}, y^{-n+2}, \ldots) \geq 0 \quad (3.31)$$

and similarly for $u(s)$.

Given the system description (A, B, C, D), the *Markov parameters* of the system are defined by

$$H^i = \begin{cases} D, & \text{if } i = 0 \\ CA^{i-1}B, & \text{if } i = 1, 2, \dots \end{cases} \tag{3.32}$$

Note that

$$G(s) = \sum_{i=0}^{\infty} H^i s^{-i}. \tag{3.33}$$

Using the power series expansions of $y(s)$ and $u(s)$ and (3.33), RCP(x_0) can be reformulated as the *Linear Dynamic Complementarity Problem* (terminology introduced in [179]) by considering the coefficients corresponding to equal powers of s. The formulation here extends the concept of LDCP as introduced in [179], because impulsive motions are included.

Linear Dynamic Complementarity Problem (LDCP$_\kappa(x_0)$) Let a system description (A, B, C, D), an integer $\kappa \geq -n+1$ and an initial state x_0 be given. Let H^i, $i \geq 0$ be given by (3.32). Find sequences $(y^{-n+1}, y^{-n+2}, \dots, y^\kappa)$ and $(u^{-n+1}, u^{-n+2}, \dots, u^\kappa)$ such that the equations

$$y^i = \sum_{j=-n+1}^{i} H^{i-j} u^j, \qquad \text{if } -n+1 \leq i \leq \min(0, \kappa) \tag{3.34a}$$

$$y^i = CA^{i-1}x_0 + \sum_{j=-n+1}^{i} H^{i-j} u^j, \qquad \text{if } 1 \leq i \leq \kappa \tag{3.34b}$$

are satisfied, and for all indices $i \in \bar{k}$ at least one of the following is true:

$$(y_i^{-n+1}, y_i^{-n+2}, \dots, y_i^\kappa) = 0 \quad \text{and} \quad (u_i^{-n+1}, u_i^{-n+2}, \dots, u_i^\kappa) \geq 0 \tag{3.35}$$

$$(y_i^{-n+1}, y_i^{-n+2}, \dots, y_i^\kappa) \geq 0 \quad \text{and} \quad (u_i^{-n+1}, u_i^{-n+2}, \dots, u_i^\kappa) = 0 \tag{3.36}$$

LDCP$_\infty(x_0)$ denotes the problem of finding vector sequences $(u^j)_{j=-n+1}^{\infty}$ and $(y^j)_{j=-n+1}^{\infty}$ that satisfy LDCP$_\kappa(x_0)$ for all $\kappa \geq -n+1$.

If $(u^j)_{j=-n+1}^{\kappa}$ and $(y^j)_{j=-n+1}^{\kappa}$ form a solution to LDCP$_\kappa(x_0)$, then index sets $J \subseteq \bar{k}$ satisfying (3.35), $i \in J$ and (3.36), $i \in J^c$ represent candidate modes for selection.

The complete set of candidates for selection, denoted by $\mathcal{S}^{\kappa}_{\mathrm{LDCP}}(x_0)$, is defined by

$$\mathcal{S}^{\kappa}_{\mathrm{LDCP}}(x_0) := \{J \subseteq \bar{k} \mid \exists (u^j)^{\kappa}_{j=-n+1}, (y^j)^{\kappa}_{j=-n+1} \text{ solution to } \mathrm{LDCP}_{\kappa}(x_0) \text{ such that (3.35) holds for } i \in J \text{ and (3.36) holds for } i \in J^c\}.$$

Theorem 3.5.2 *Let a system (A, B, C, D) be given. The following statements are equivalent when Assumption 3.4.1 holds.*

1. *The equations (3.13) have an initial solution for initial state x_0.*
2. *RCP(x_0) has a solution.*
3. *LDCP$_\infty(x_0)$ has a solution.*

There is a one-to-one correspondence between initial solutions to (3.13), solutions to RCP(x_0), and solutions to LDCP$_\infty(x_0)$. Furthermore, for all $x_0 \in \mathbb{R}^n$,

$$\mathcal{S}(x_0) = \mathcal{S}_{RCP}(x_0) = \mathcal{S}^{\infty}_{LDCP}(x_0).$$

□

Proof. From the derivation of RCP, it follows that *1* and *2* are equivalent. If $(u(s), y(s))$ is a solution to RCP(x_0), then the coefficients of the power series expansion of this solution around infinity form a solution to LDCP$_\infty(x_0)$. Hence, *2* implies *3*.

To see that *3* implies *1*, suppose $(y^{-n+1}, y^{-n+2}, \dots)$, $(u^{-n+1}, u^{-n+2}, \dots)$ is a solution to LDCP$_\infty(x_0)$. Take $I \subseteq \bar{k}$ such that (3.35) holds for $i \in I$ and (3.36) holds for $i \in I^c$. Define $p(0) := x_0 + \sum_{i=0}^{n-1} A^i B u^{-i}$. We first show that $p(0) \in V_I$. To this end, note that $y^i_I = 0$ and $u^i_{I^c} = 0$ for all $i \in \{-n+1, -n+2, \dots\}$. From (3.34b), it follows that $p(0)$ satisfies

$$\begin{aligned}
0 = y^1_I &= C_{I\bullet} p(0) + D_{II} v(0) \\
0 = y^2_I &= C_{I\bullet} A p(0) + D_{II} v(1) + C_{I\bullet} B_{\bullet I} v(0) \\
&\vdots \\
0 = y^{\kappa}_I &= C_{I\bullet} A^{\kappa-1} p(0) + D_{II} v(\kappa-1) + C_{I\bullet} B_{\bullet I} v(\kappa-2) + .. + C_{I\bullet} A^{\kappa-2} B_{\bullet I} v(0) \\
&\vdots \qquad\qquad\qquad\qquad \vdots
\end{aligned} \tag{3.37}$$

where $v(i) = u^{i+1}_I$, $i \geq 0$. Combining algorithm (3.4) and the equations above, it follows that for $l \geq 0$ the states $A^l p(0) + \sum_{i=0}^{l-1} A^i B_{\bullet I} v(l-1-i)$ belong to $V_j(A, B_{\bullet I}, C_{I\bullet}, D_{II})$, , $j \geq 0$. In particular for $l = 0$ this means that $p(0) \in \lim V^j(A, B_{\bullet I}, C_{I\bullet}, D_{II}) = V_I$. This means that there exists a smooth solution $(\mathtt{u}_{reg}, \mathtt{x}_{reg}, \mathtt{y}_{reg})$ to (3.14) for mode I with initial state $x(0) = p(0)$.

By differentiating (3.14) in time and evaluating the resulting equalities at time instant 0 for the solution $(\mathtt{u}_{reg}, \mathtt{x}_{reg}, \mathtt{y}_{reg})$, we observe that $\tilde{v}(i) := \mathtt{u}^{(i)}_{reg,I}(0)$, $i =$

$0, 1, \dots$ satisfies (3.37) as well. To show that this implies that $\tilde{v}^i = v^i$ for all i, observe that due to (3.37) both sequences satisfy the discrete-time analogue of the first two lines of (3.15), i.e.

$$p(i+1) = Ap(i) + B_{\bullet I} v(i); \quad 0 = C_{I\bullet} p(i) + D_{II} v(i), \quad i = 0, 1, 2, \dots \tag{3.38}$$

with initial state $p(0)$. The difference $w(i) := v(i) - \tilde{v}(i)$ satisfies (3.38) with initial state 0. We introduce the formal z-transform

$$w(z) := \sum_{i=0}^{\infty} w^i z^{-i}.$$

Using the z-transform $G_{II}(z)$ of the discrete-time system (see e.g. [117]), we get $0 = G_{II}(z)w(z)$. The invertibility of $G_{II}(z)$ implies that $w(z) = 0$ and hence, $v(i) = \tilde{v}(i)$ for all $i \geq 0$, or equivalently, $u_I^{i+1} = \mathrm{u}_{reg,I}^{(i)}(0)$, $i \geq 0$. This also implies that $y^{i+1} = \mathrm{y}_{reg}^{(i)}(0)$, $i \geq 0$.

We define $\mathrm{u} := \sum_{i=0}^{n-1} u^{-i}\delta^{(i)} + \mathrm{u}_{reg}$, $\mathrm{y} := \sum_{i=1}^{n-1} y^i \delta^{(i-1)} + \mathrm{y}_{reg}$ and let x be the solution to $\dot{\mathrm{x}} = A\mathrm{x} + B\mathrm{u} + x_0\delta$. Obviously, $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ satisfies *1* in Definition 3.4.7. We only have to show that *2* in Definition 3.4.7 is satisfied. Since

$$(y^{-n+1}, y^{-n+2}, \dots) = (y^{-n+1}, \dots, y^0, \mathrm{y}_{reg}^{(0)}(0), \mathrm{y}_{reg}^{(1)}(0), \dots)$$

and

$$(u^{-n+1}, u^{-n+2}, \dots) = (u^{-n+1}, \dots, u^0, \mathrm{u}_{reg}^{(0)}, \mathrm{u}_{reg}^{(1)}, \dots)$$

form a solution to $\mathrm{LDCP}_\infty(x_0)$, (3.35) or (3.36) is satisfied for all $i \in \bar{k}$. According to Lemma 3.5.1, this is equivalent to u and y being initially nonnegative. Consequently $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ is an initial solution with initial state x_0.

The one-to-one correspondence follows easily from the above, because solutions to RCP and initial solutions are related through Laplace transform and its inverse. Solutions to RCP are uniquely transformed to solutions to LDCP by taking the coefficients of a power series expansion around infinity. Moreover, a solution to LDCP is linked to an initial solution by setting the derivatives of an initial solution at zero equal to the LDCP solution as stated above (see also remark 3.5.3). The final statement is a result of the one-to-one correspondence. □

Remark 3.5.3 Note that in the proof of Theorem 3.5.2, a direct link between initial solutions and solutions to $\mathrm{LDCP}_\infty(x_0)$ is given. If $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ is an initial solution with $\mathrm{u} = \sum_{i=0}^{n-1} u^{-i}\delta^{(i)} + \mathrm{u}_{reg}$ and $\mathrm{y} = \sum_{i=0}^{n-1} y^{-i}\delta^{(i)} + \mathrm{y}_{reg}$ for initial state x_0, define $\tilde{u}^i := u^i$, $i = -n+1, \dots, 0$ and $\tilde{u}^{i+1} = \mathrm{u}_{reg}^{(i)}(0)$, $i \geq 0$ and let $\tilde{y}^i$, $i \geq -n+1$ be defined analogously. Then $(\tilde{u}^i)_{i=-n+1}^{\infty}$, $(\tilde{y}^i)_{i=-n+1}^{\infty}$ is a solution to $\mathrm{LDCP}_\infty(x_0)$. We shall use the transformations between $\mathrm{LDCP}_\infty(x_0)$, $\mathrm{RCP}(x_0)$ and initial solutions

frequently. The above proof also yields an alternative way of deriving the LDCP: differentiate the initial solution with incorporation of the impulsive part and evaluate the results at time instant zero. For smooth continuations, this method can also be used in the nonlinear case [124, 179]. □

In the above theorem it is shown that the infinite version of LDCP can be used to select the correct modes. However, under suitable conditions, already the finite version $\mathrm{LDCP}_n(x_0)$ selects the right modes, where n is the dimension of the state variable (see Theorem 3.6.12 below). In [55], it has been shown that $\mathrm{LDCP}_\kappa(x_0)$ for finite κ is a special case of the Generalized Linear Complementarity Problem [201] and the Extended Linear Complementarity Problem [53]. In [201], an algorithm is proposed to find all solutions to GLCP. Such algorithms can be used to efficiently solve the LDCP.

3.6 Well-posedness results

Due to the multimodal and nonlinear behavior of linear complementarity systems, basic questions like existence and uniqueness of solutions given an initial state are nontrivial. It is not difficult to find linear complementarity systems for which no solution exists from certain initial conditions or for which the solution is not unique (see [177]). In this section we will derive conditions guaranteeing well-posedness.

3.6.1 Local well-posedness

Definition 3.6.1 The complementarity system (3.13) is locally well-posed if for each initial state there exists an $\varepsilon > 0$ such that a unique solution on $[0, \varepsilon)$ in the sense of Definition 3.4.10 exists. □

An equivalent way of defining local well-posedness is by requiring that for each state there exists a unique solution on an interval of positive length starting with either a finite number of jumps or an infinite number of jumps with convergence of the event states, followed by smooth continuation on that interval.

Definition 3.6.2 Let (A, B, C, D) be a system with Markov parameters H^i, $i = 0, 1, 2, \ldots$. The leading column indices $\eta_1, \ldots, \eta_k$ of the linear system (A, B, C, D) are defined for $j \in \bar{k}$ as

$$\eta_j := \inf\{i \in \mathbb{N} \mid H^i_{\bullet j} \neq 0\}$$

with the convention $\inf \emptyset = \infty$. The leading row indices $\rho_1, \ldots, \rho_k$ of (A, B, C, D) are defined for $j \in \bar{k}$ as

$$\rho_j := \inf\{i \in \mathbb{N} \mid H^i_{j\bullet} \neq 0\}.$$

□

Since we consider only invertible transfer functions (see Assumption 3.4.1 and Lemma 3.3.3), the leading row and column indices are all finite. Due to the Cayley-Hamilton theorem, we even have $\rho_i \leq n$ and $\eta_i \leq n$. The *leading row coefficient matrix* $\mathcal{M}(A, B, C, D)$ and *leading column coefficient matrix* $\mathcal{N}(A, B, C, D)$ for the system (A, B, C, D) are defined as

$$\mathcal{M}(A, B, C, D) := \begin{pmatrix} H_{1\bullet}^{\rho_1} \\ \vdots \\ H_{k\bullet}^{\rho_k} \end{pmatrix} \text{ and } \mathcal{N}(A, B, C, D) := (H_{\bullet 1}^{\eta_1} \dots H_{\bullet k}^{\eta_k}) \tag{3.39}$$

respectively. We omit the arguments (A, B, C, D), if they are clear from the context.

The main result of this section is stated as follows. Recall that a square matrix is a P-matrix, if all of its principal minors are strictly positive (Section 3.3).

Theorem 3.6.3 *If the leading row coefficient matrix $\mathcal{M}$ and the leading column coefficient matrix $\mathcal{N}$ are both P-matrices, then the linear complementarity system (3.13) is locally well-posed. From each initial condition, at most one state jump occurs before smooth continuation is possible, i.e. the multiplicity of an event time is at most one.* □

Remark 3.6.4 The definition of well-posedness if often taken to include continuous dependence of solutions on initial conditions. Such continuous dependence is not claimed in the above theorem. An example of a linear complementarity system that displays discontinuous dependence on initial conditions will be given in Section 3.8. □

To prove the main result, we first need some auxiliary results.

Lemma 3.6.5 *If the leading row coefficient matrix $\mathcal{M}$ has only nonzero principal minors, then assumption 3.4.1 is satisfied, i.e. all modes are autonomous. The same holds when the leading column coefficient matrix $\mathcal{N}$ has only nonzero principal minors.* □

Proof. Lemma 3.3.3 states that it is sufficient to show that $G_{II}(s)$ is invertible for all $I \subseteq \bar{k}$. For notational convenience, we assume $I = \bar{l}$ for some $l \in \bar{k}$. If $\mathcal{M}$ has only nonzero principal minors, then $\mathcal{M}_{II}$ is invertible. Hence, $G_{II}(s) = \operatorname{diag}(s^{-\rho_1}, \dots, s^{-\rho_l})V(s)$ where $V(s)$ is a biproper matrix, because $V(\infty) = \mathcal{M}_{II}$ is invertible [82, Thm. 4.5]. The reasoning is analogous for the case in which $\mathcal{N}$ has only nonzero minors. □

Definition 3.6.6 A state x_0 of the complementarity system (3.13) is called regular, if there exists a smooth initial solution with initial state x_0. □

A state x_0 is regular if and only if RCP(x_0) has a strictly proper solution. Or equivalently, x_0 is regular if and only if LDCP$_\infty(x_0)$ has a solution with $u^{-n+1} = \ldots = u^0 = 0$.

Under the assumption that the leading row coefficient matrix is a P-matrix, the following result characterizes the regular states. The result is an extension of a similar result in [179] which was derived under the additional assumption of "uniform relative degree" (i.e. $\rho_1 = \rho_2 = \ldots = \rho_k = \rho$). In contrast to [179] we restrict ourselves here to the linear case, but an extension to the nonlinear case is straightforward.

Theorem 3.6.7 *Let a system (A, B, C, D) be given. Suppose that the leading row coefficient matrix $\mathcal{M}$ is a P-matrix. Then $x_0 \in \mathbb{R}^n$ is a regular state of the complementarity system (3.13) if and only if for all $i \in \bar{k}$*

$$(C_{i\bullet}x_0, C_{i\bullet}Ax_0, \ldots, C_{i\bullet}A^{\rho_i-1}x_0) \geq 0. \tag{3.40}$$

Moreover, the smooth continuation is unique. □

Proof. Note that $y_i^{(j)}(0) = C_{i\bullet}A^j x_0$, $j = 0, \ldots, \rho_i - 1$, $i = 1, \ldots, k$, independently of the choice of a smooth input u. Hence, the above condition is necessary to guarantee $y(t) \geq 0$, $t \in [0, \varepsilon)$ for some positive ε.

To prove the converse, we will show that if for all $i \in \bar{k}$ (3.40) holds, the corresponding LDCP$_\infty(x_0)$ has a solution with $u^{-n+1} = \ldots = u^0 = 0$. This is sufficient to show that a smooth initial solution exists. The idea of the proof is to reduce the LDCP$_\infty(x_0)$ to a series of LCPs that can all be solved uniquely. This idea originates in [124].

We will show that LDCP$_\infty(x_0)$ with the additional requirement $y^{-n+1} = \ldots = y^0 = 0$, $u^{-n+1} = \ldots = u^0 = 0$ has a unique solution. From such a solution, it is immediately clear that (3.34a) is satisfied. The remaining equalities can be written as

$$y_i^j = C_{i\bullet}A^{j-1}x_0, \;\; j = 1, 2, \ldots, \rho_i, \;\; i = 1, \ldots, k \tag{3.41}$$

and

$$\begin{pmatrix} y_1^{\rho_1+p} \\ \vdots \\ y_k^{\rho_k+p} \end{pmatrix} = \xi_p(x_0, u^1, \ldots, u^{p-1}) + \mathcal{M}u^p, \tag{3.42}$$

where $\xi_1, \xi_2, \ldots$ are certain linear functions. We denote by $L(l)$, $l \in \mathbb{N}$ the truncated problem of finding u^j, $j = 1, \ldots, l$ and y_i^j, $i \in \bar{k}$, $j = 1, \ldots, \rho_i + l$ satisfying (3.41) and (3.42), $p \in \{1, \ldots, l\}$ together with the requirement that for all indices $i \in \bar{k}$ at least one of the following statements is true:

$$(y_i^1, y_i^2, \ldots, y_i^{\rho_i+l}) = 0 \quad \text{and} \quad (u_i^1, u_i^2, \ldots, u_i^l) \geq 0 \tag{3.43}$$

$$(y_i^1, y_i^2, \ldots, y_i^{\rho_i+l}) \geq 0 \quad \text{and} \quad (u_i^1, u_i^2, \ldots, u_i^l) = 0. \tag{3.44}$$

The problem $L(l)$ is a subproblem of $\mathrm{LDCP}_\infty(x_0)$ and if we find a solution $(y^1, y^2, \dots)$, $(u^1, u^2, \dots)$ satisfying $L(l)$ for all $l \geq 0$, then this solution is a solution to the corresponding $\mathrm{LDCP}_\infty(x_0)$ with $y^{-n+1} = \ldots = y^0 = 0$, $u^{-n+1} = \ldots = u^0 = 0$.

We claim that $L(l)$ has a unique solution for all $l \geq 0$. This is obvious for $l = 0$. We will proceed by induction in the same way as in [124, 179].

We write I_l, J_l, K_l for the active (input) index set, the inactive index set and the undecided index set, respectively, determined by $L(l)$. Formally, for $l \geq 1$, $I_l = \{i \in \bar{k} \mid (u_i^1, \dots, u_i^l) \succ 0\}$, $J_l = \{i \in \bar{k} \mid (y_i^1, \dots, y_i^{\rho_i + l}) \succ 0\}$ and $K_l = \bar{k} \setminus (I_l \cup J_l)$ with y_i^j, $i = 1, \dots, k$, $j = 1, \dots, \rho_i + l$ and u^i, $i = 1, \dots, l$ determined (uniquely) by $L(l)$. For convenience we also define $I_0 := \emptyset$, $J_0 = \{i \in \bar{k} \mid (y_i^1, \dots, y_i^{\rho_i}) \succ 0\}$ and $K_0 = \bar{k} \setminus J_0$.

Note that $L(l-1)$ is a subproblem of $L(l)$, so variables uniquely determined by $L(l-1)$ are automatically uniquely specified for $L(l)$. As a consequence, I_{l-1}, J_{l-1}, K_{l-1} are determined as well. Comparing $L(l)$ with $L(l-1)$, we observe that $L(l)$ has one additional equation: (3.42) for $p = l$. We divide this equation into the three parts given by I_{l-1}, J_{l-1} and K_{l-1}. For notational convenience, we omit all indices depending on l and all superscripts:

$$\begin{pmatrix} y_I \\ y_J \\ y_K \end{pmatrix} = \begin{pmatrix} z_I \\ z_J \\ z_K \end{pmatrix} + \begin{pmatrix} \mathcal{M}_{II} & \mathcal{M}_{IJ} & \mathcal{M}_{IK} \\ \mathcal{M}_{JI} & \mathcal{M}_{JJ} & \mathcal{M}_{JK} \\ \mathcal{M}_{KI} & \mathcal{M}_{KJ} & \mathcal{M}_{KK} \end{pmatrix} \begin{pmatrix} u_I \\ u_J \\ u_K \end{pmatrix} \tag{3.45}$$

with $z = \xi_l(x_0, u^1, \dots, u^{l-1})$. From the definition of I_{l-1}, J_{l-1} and K_{l-1}, we get $y_I = 0$ and $u_J = 0$, because (3.43) or (3.44) should hold. By substituting this result in (3.45), we obtain

$$0 = z_I + \mathcal{M}_{II} u_I + \mathcal{M}_{IK} u_K \tag{3.46}$$

$$y_J = z_J + \mathcal{M}_{JI} u_I + \mathcal{M}_{JK} u_K \tag{3.47}$$

$$y_K = z_K + \mathcal{M}_{KI} u_I + \mathcal{M}_{KK} u_K. \tag{3.48}$$

Since $\mathcal{M}_{II}$ is a principal submatrix of a P-matrix, it is invertible and hence we get from (3.46) that $u_I = -\mathcal{M}_{II}^{-1}(z_I + \mathcal{M}_{IK} u_k)$. Substituting this expression in (3.48) leads to

$$y_K = z_K - \mathcal{M}_{KI}\mathcal{M}_{II}^{-1} z_I + (\mathcal{M}_{KK} - \mathcal{M}_{KI}\mathcal{M}_{II}^{-1}\mathcal{M}_{IK}) u_K \tag{3.49}$$

Due to (3.43) and (3.44) and the definition of K_{l-1}, the complementarity conditions

$$0 \leq u_K \perp y_K \geq 0 \tag{3.50}$$

hold. So, (3.49) and (3.50) constitute an LCP. Since $\mathcal{M}_{KK} - \mathcal{M}_{KI}\mathcal{M}_{II}^{-1}\mathcal{M}_{IK}$ is a Schur complement of a P-matrix, it is itself a P-matrix by Proposition 2.3.5 in [47]. According to Theorem 3.3.4, the corresponding LCP has a unique solution. From u_K we can compute u_I and y_J. Hence, the induction hypothesis has been proven for l. So we find a solution of $\mathrm{LDCP}_\infty(x_0)$ with $u^{-n+1} = \ldots = u^0 = 0$, $y^{-n+1} = \ldots = y^0 = 0$.

and hence a smooth initial solution corresponding to x_0 exists. Since the solution to $\mathrm{LDCP}_\infty(x_0)$ with $u^{-n+1} = \ldots = u^0 = 0$ is unique, the one-to-one correspondence between initial solutions and solutions of $\mathrm{LDCP}_\infty(x_0)$ implies that the corresponding smooth initial solution is unique. □

One can even prove that the initial solution corresponding to a regular initial state is unique and thus smooth. Our next result is concerned with the uniqueness of solutions emanating from a not necessarily regular initial state.

Theorem 3.6.8 *Let a system (A, B, C, D) be given. If the leading column coefficient matrix $\mathcal{N}$ is a P-matrix, then for every state x_0 and every $\kappa \geq 0$, the problem $\mathrm{LDCP}_\kappa(x_0)$ has a solution that is unique except for u_i^j, $i \in \bar{k}$, $j = \kappa - \eta_i + 1, \ldots, \kappa$, which are left undetermined. Furthermore, $u_i^{-n+1} = u_i^{-n+2} = \ldots = u_i^{-\eta_i} = 0$, $i \in \bar{k}$ and $y^{-n+1} = \ldots = y^0 = 0$.* □

Proof. The proof is based on separation of the equalities (3.34) in two parts, (3.34a) and (3.34b), providing the equations for y^i, $i = -n+1, \ldots, 0$ and y^i, $i = 1, \ldots, \kappa$, respectively. For both parts we start an induction that is analogous to the one used in the previous proof: we reduce the LDCP to a series of LCPs which can be solved uniquely. This is done by selecting certain equations from (3.34) for each successive LCP in such a way that only principal submatrices of the leading column coefficient matrix $\mathcal{N}$ appear in these LCPs.

We introduce the index sets $O_j := \{i \in \bar{k} \mid \eta_i = j\}$, $j = 0, 1, \ldots, n$ and $S_j := \bigcup_{i=0}^{j} O_i$, $j = 0, 1, \ldots, n$. So, the η_j-th Markov parameter is the first Markov parameter in which the j-th column is nonzero. O_j is the set of indices i for which the i-th column in the sequence of Markov parameters $(H^0, H^1, \ldots)$ is nonzero for the first time in H^j. S_j is the set of indices i for which the matrix $(H^0_{\bullet i}, H^1_{\bullet i}, \ldots, H^j_{\bullet i})$ is nonzero. As noted before, $\eta_i \leq n$. Hence, $S_n = \bar{k}$. By definition, $H^i_{\bullet S_j^c} = 0$, $i \leq j$ and $S_0 \subseteq S_1 \subseteq S_2 \subseteq \ldots \subseteq S_n$.

After suitable permutation of rows and columns if necessary, there are integers $k_0, \ldots, k_{n+1}$ with $0 = k_0 \leq k_1 \leq k_2 \leq \ldots k_n \leq k_{n+1} = k$ such that $O_j = \{k_j + 1, \ldots, k_{j+1}\}$, $j = 0, 1, \ldots, n$. Then

$$\mathcal{N} = [H^0_{\bullet O_0}\ H^1_{\bullet O_1} \ldots\ H^n_{\bullet O_n}].$$

We claim that for $1 \leq r \leq n$ the problem $\mathrm{LDCP}_{-n+r}(x_0)$ has a solution with

$$u^{-n+1}_{S_{r-1}} = u^{-n+2}_{S_{r-2}} = \ldots = u^{-n+r}_{S_0} = 0 \tag{3.51}$$

$$y^{-n+1} = y^{-n+2} = \ldots = y^{-n+r} = 0. \tag{3.52}$$

The remaining variables $u^{-n+1}_{S^c_{r-1}}, u^{-n+2}_{S^c_{r-2}}, \ldots, u^{-n+r}_{S^c_0}$ are left undetermined. This will be the induction hypothesis.

For $r = 1$, we only have the equation

$$y^{-n+1} = H^0 u^{-n+1} \tag{3.53}$$

with the complementarity conditions $0 \leq y^{-n+1} \perp u^{-n+1} \geq 0$. The complementarity conditions follow from the fact that for each index either (3.35) or (3.36) should hold. Since $H^0_{\bullet S_0^c} = 0$, (3.53) reduces to

$$y^{-n+1} = H^0_{\bullet S_0} u^{-n+1}_{S_0}. \tag{3.54}$$

Since $u^{-n+1}_{S_0^c}$ does not appear in this equation, it is left completely undetermined (except for the condition $u^{-n+1}_{S_0^c} \geq 0$). Considering (3.54) and the complementarity conditions only for y_i^{-n+1}, $i \in S_0$ results in the LCP

$$\begin{aligned} &y^{-n+1}_{S_0} = H^0_{S_0 S_0} u^{-n+1}_{S_0} = \mathcal{N}_{S_0 S_0} u^{-n+1}_{S_0} \\ &0 \leq y^{-n+1}_{S_0} \perp u^{-n+1}_{S_0} \geq 0. \end{aligned}$$

Since $\mathcal{N}_{S_0 S_0}$ is a principal submatrix of $\mathcal{N}$, it is a P-matrix. Theorem 3.3.4 then implies that the above LCP has a unique solution. Obviously, $y^{-n+1}_{S_0} = 0$, $u^{-n+1}_{S_0} = 0$ is the unique solution. From (3.54), $y^{-n+1} = 0$ follows immediately. This proves the induction hypothesis for $r = 1$.

Suppose that the induction hypothesis above holds for $r - 1$, where $2 \leq r \leq n$. Since $\mathrm{LDCP}_{-n+r-1}(x_0)$ is a subproblem of $\mathrm{LDCP}_{-n+r}(x_0)$, we consider only the additional equality in (3.34):

$$\begin{aligned} y^{-n+r} &= H^0 u^{-n+r} + H^1 u^{-n+r-1} + \ldots + H^{r-1} u^{-n+1} \\ &= H^0_{\bullet S_0} u^{-n+r}_{S_0} + H^1_{\bullet S_1} u^{-n+r-1}_{S_1} + \ldots + H^{r-1}_{\bullet S_{r-1}} u^{-n+1}_{S_{r-1}} \\ &= H^0_{\bullet S_0} u^{-n+r}_{S_0} + H^1_{\bullet S_1 \setminus S_0} u^{-n+r-1}_{S_1 \setminus S_0} + \ldots + H^{r-1}_{\bullet S_{r-1} \setminus S_{r-2}} u^{-n+1}_{S_{r-1} \setminus S_{r-2}} \\ &= H^0_{\bullet O_0} u^{-n+r}_{O_0} + H^1_{\bullet O_1} u^{-n+r-1}_{O_1} + \ldots + H^{r-1}_{\bullet O_{r-1}} u^{-n+1}_{O_{r-1}}. \end{aligned} \tag{3.55}$$

The second equality follows from $H^i_{\bullet S_i^c} = 0$, the third one follows from the induction hypothesis (3.51). The last equality is a consequence of $S_j \setminus S_{j-1} = O_j$. Since $u^{-n+1}_{S_{r-1}^c}, u^{-n+2}_{S_{r-2}^c}, \ldots, u^{-n+r}_{S_0^c}$ do not appear in this additional equation, these variables remain undetermined.

Equation (3.55) consists of k scalar equations. Considering only the equalities for y_i^{-n+r}, $i \in S_{r-1}$, we find

$$\begin{aligned} y_{S_{r-1}}^{-n+r} &= \begin{pmatrix} H^0_{S_{r-1}O_0} & H^1_{S_{r-1}O_1} \cdots H^{r-1}_{S_{r-1}O_{r-1}} \end{pmatrix} \begin{pmatrix} u_{O_0}^{-n+r} \\ u_{O_1}^{-n+r-1} \\ \vdots \\ u_{O_{r-1}}^{-n+1} \end{pmatrix} \\ &= \mathcal{N}_{S_{r-1}S_{r-1}} \underbrace{\begin{pmatrix} u_{O_0}^{-n+r} \\ u_{O_1}^{-n+r-1} \\ \vdots \\ u_{O_{r-1}}^{-n+1} \end{pmatrix}}_{=:v_{-r}}. \end{aligned}$$

Since (3.35) or (3.36) should hold for all i, it follows that

$$0 \le y_{S_{r-1}}^{-n+r} \perp v_{-r} \ge 0.$$

This is the LCP we are looking for. Since $\mathcal{N}_{S_{r-1}S_{r-1}}$ (as a submatrix of $\mathcal{N}$) is also a P-matrix, the above LCP has a unique solution (Theorem 3.3.4). Hence, this solution must be $v_{-r} = y_{S_{r-1}}^{-n+r} = 0$. Using this in (3.55) shows that $y^{-n+r} = 0$. In combination with the induction hypothesis for $r-1$, this yields the hypothesis for r. This completes our induction step and hence the proof of our first claim.

To complete the proof, we start a second induction with hypothesis as stated in the formulation of the theorem. Note that this is equivalent to saying: $\mathrm{LDCP}_\kappa(x_0)$ has a unique solution for every state x_0, only $u_{S_0^c}^\kappa, u_{S_1^c}^{\kappa-1}, \ldots, u_{S_{n-1}^c}^{\kappa-n+1}$ are left undetermined. For $\kappa = 0$ this hypothesis is true, for it follows from the previous induction by taking $r = n$. Suppose the hypothesis is true for $\kappa - 1, \kappa \ge 1$. Since $\mathrm{LDCP}_{\kappa-1}(x_0)$ is a subproblem of $\mathrm{LDCP}_\kappa(x_0)$, the variables $u_{S_0}^{\kappa-1}, \ldots, u_{S_{n-1}}^{\kappa-n}, u^{\kappa-n-1}, \ldots, u^{-n+1}$ are already uniquely determined. We set

$$\begin{aligned} I &:= \{i \in \bar{k} \mid (u_i^{-n+1}, u_i^{-n+2}, \ldots, u_i^{\kappa-\eta_i-1}) \succ 0\}, \\ J &:= \{i \in \bar{k} \mid (y_i^{-n+1}, y_i^{-n+2}, \ldots, y_i^{\kappa-1}) \succ 0\} \text{ and} \\ K &:= \bar{k} \setminus (I \cup J). \end{aligned}$$

In comparison with $\mathrm{LDCP}_{\kappa-1}(x_0)$, $\mathrm{LDCP}_\kappa(x_0)$ has the additional equality

$$y^\kappa = \sigma(x_0, u_{S_0}^{\kappa-1}, u_{S_1}^{\kappa-2}, \ldots, u_{S_{n-1}}^{\kappa-n}, u^{\kappa-n-1}, \ldots, u^{-n+1}) + \mathcal{N} \begin{pmatrix} u_{O_0}^\kappa \\ u_{O_1}^{\kappa-1} \\ \vdots \\ u_{O_{n-1}}^{\kappa-n+1} \end{pmatrix}$$

for some function σ. Splitting this equation into three parts according to the index sets I, J, K, we can follow the same reasoning as in the proof of Theorem 3.6.7 to conclude that $y^\kappa, u^\kappa_{O_0}, u^{\kappa-1}_{O_1}, \ldots, u^{\kappa-n+1}_{O_{n-1}}$ are uniquely determined and thus prove the induction hypothesis for κ. □

We are now in a position to prove Theorem 3.6.3.

Proof of Theorem 3.6.3 Lemma 3.6.5 implies that all modes are autonomous. Take an arbitrary initial state x_0. It follows from Theorem 3.6.8 that $\text{LDCP}_\infty(x_0)$ has a unique solution which satisfies $u_i^{-n+1} = u_i^{-n+2} = \ldots = u_i^{-\eta_i} = 0$, $i \in \bar{k}$ and $y^{-n+1} = \ldots = y^0 = 0$. Due to the one-to-one correspondence between initial solutions and solutions to $\text{LDCP}_\infty(x_0)$, an initial solution $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ exists and the solution must be unique as well. In case the initial condition is regular, the initial solution is smooth. In other cases, we have to prove that after the state jump corresponding to $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ smooth continuation is possible. Stated otherwise, we have to show that the re-initialized state $\mathtt{x}(0+)$ is regular. The re-initialization is given by the impulsive part $\mathtt{u}_{imp} = \sum_{i=0}^{n-1} u^{-i}\delta^{(i)}$, where the coefficients u^{-i} follow from $\text{LDCP}_\infty(x_0)$. Since the impulsive part is unique, the re-initialization is unique; it results in $\mathtt{x}(0+) := x_0 + \sum_{i=0}^{n-1} A^i B u^{-i}$ (see (3.10)). The complementarity conditions (3.35) and (3.36) imply that $(y^1, y^2, \ldots, y^n) \succeq 0$. The right hand side of (3.34) contains for $y_i^1, \ldots, y_i^{\rho_i}$, $i \in \bar{k}$ only coefficients corresponding to the impulsive part, i.e. only $u^0, \ldots, u^{-n+1}$. Hence, observe that $(C_{i\bullet}x(0+), \ldots, C_{i\bullet}A^{\rho_i-1}x(0+)) = (y_i^1, \ldots, y_i^{\rho_i}) \succeq 0$, $i \in \bar{k}$. According to Lemma 3.6.7, $\mathtt{x}(0+)$ is a regular state. So after at most one re-initialization, (unique) smooth continuation is guaranteed. □

3.6.2 Global well-posedness

This subsection contains material of the paper [94] and presents two classes of linear complementarity systems that can be proven to be globally well-posed.

Definition 3.6.9 The linear complementarity system (3.13) is globally well-posed, if

1. For each initial state there exists a solution on $[0, \infty)$ in the sense of Definition 3.4.10.
2. If $(\mathcal{E}^j, u_c^j, x_c^j, y_c^j)$, $j = 1, 2$ are two solutions corresponding to the same initial state and both defined on $[0, T_e)$ for arbitrary $T_e > 0$, then

$$(u_c^1, x_c^1, y_c^1)(t) = (u_c^2, x_c^2, y_c^2)(t)$$

for all $t \in [0, T_e)$ with $t \notin \mathcal{E}^1 \cup \mathcal{E}^2$.

We will also use the term "global existence" for the first, and "global uniqueness" for the second statement above. □

Local existence does not imply global existence. A problem arises when the event times contain a right-accumulation point $\tau^* < \infty$ and there is no limit for $x_c(t)$ as $t \uparrow \tau^*$. In fact, this is the only phenomenon that may prevent a local well-posed linear complementarity system from being globally well-posed. Note that local uniqueness of solutions is equivalent to global uniqueness of solutions using the solution concept of Definition 3.4.10 (see also Chapter 4).

Bimodal linear complementarity systems

A LCS is said to be bimodal, if there is only one complementarity pair (u, y) (i.e. $k = 1$). As a consequence, the corresponding LCS has two modes ($I = \emptyset$ and $I = \{1\}$).

Theorem 3.6.10 *Consider a bimodal LCS (3.13) with $C \neq 0$.* [1] *The following statements are equivalent.*

1. *The leading Markov parameter $\mathcal{M} = \mathcal{N}$ is defined (i.e. $\rho_1 = \eta_1 < \infty$) and positive.*

2. *The linear complementarity system (3.13) is locally well-posed.*

3. *The linear complementarity system (3.13) is globally well-posed.*

□

Proof. Thm. 3.6.3 yields $1 \Rightarrow 2$. To prove $2 \Rightarrow 1$, consider the following cases.

1. Suppose the leading Markov parameter $\mathcal{M} = \mathcal{N}$ is defined and negative. According to Lemma 3.6.5 all the modes are autonomous in this situation.

   (a) $D = 0$. [177, Thm. 4.8] claims that the system is not locally well-posed.

   (b) $D < 0$. It can easily be seen from (3.13) that $u = -D^{-1}Cx$ (mode $I = \{1\}$) and $u = 0$ (mode $I = \emptyset$) both generate a smooth initial solution and thus a local solution in the sense of Definition 3.4.10 for an initial state x_0 with $Cx_0 > 0$.

2. In case $\mathcal{M}$ and $\mathcal{N}$ are not defined ($\rho_1 = \eta_1 = \infty$), all Markov parameters are zero. It is clear that y is independent of u in (3.13b). Hence, for any $x_0 \in \mathbb{R}^n$ with $Cx_0 < 0$ there does not exist a solution.

[1] Note that $C = 0$ is a degenerate and uninteresting case, since the complementarity conditions do not involve the state vector x. Any quadruple $(\mathcal{E}, u_c, x_c, y_c)$ with $u(t)$ a solution to LCP(0, D) for all $t \notin \mathcal{E}$ and satisfying (3.13a)-(3.13b) is a solution to (3.13). It can easily be seen that for a scalar D, LCP(0, D) has a unique solution if and only if $D \neq 0$.

As mentioned before, local uniqueness of solutions and global uniqueness of solutions are equivalent (see also Chapter 4). Since global existence implies local existence, we have $3 \Rightarrow 2$. It remains to show that $2 \Rightarrow 3$, i.e. we have to show that local existence implies global existence of solutions.

The mode dynamics are given by $\dot{x} = Ax$ for $I = \emptyset$ and $\dot{x} = A + BF_{\{1\}}$ for $I = \{1\}$ with F_I as in subsection 3.4.1. It can easily be verified that the consistent subspace for $I = \emptyset$ equals $V_\emptyset = \mathbb{R}^n$ and thus the re-initialization operator $P_\emptyset$ as in subsection 3.4.2 is just the identity $\mathcal{1}$. The re-initialization $P_{\{1\}}$ is the projection on $V_{\{1\}}$ along $T_{\{1\}}$.

Let $[0, \tau^*)$ be the maximal interval on which a solution $(\mathcal{E}, u_c, x_c, y_c)$ exists for initial state x_0 and suppose that $\tau^* < \infty$. We drop the supscript c for ease of notation. Time τ^* is a right-accumulation point of events, because otherwise the LCS evolves in either one of the modes on an interval $(\tau^* - \beta, \tau^*)$ for some $\beta > 0$ Then it is clear that $\lim_{t \uparrow \tau^*} x(t)$ exists, because the dynamics within a mode is linear. Consequently, continuation beyond τ^* would be possible due to local existence of solutions.

Without loss of generality we may assume that the initial mode is $\{1\}$. Since τ^* is a right-accumulation of events there are infinitely many cycles consisting of smooth continuation in mode $\{1\}$, smooth continuation in mode $\emptyset$ and then a jump of the state variable according to $P_{\{1\}}$. Consider the state x_b at the beginning of the cycle (after the re-initialization). It is clear that $P_{\{1\}}x_b = x_b \in V_{\{1\}}$. Denote the duration of mode $\{1\}$ by Δ_1 (may be equal to zero) and in mode $\emptyset$ by $\Delta_\emptyset$ and define $x_m = e^{(A+BF_{\{1\}})\Delta_1}x_b$. Note that $x_m \in V_{\{1\}}$ due to invariance of $V_{\{1\}}$ under the dynamics $\dot{x} = (A + BF_{\{1\}})x$. Then we obtain for $x_e := P_{\{1\}}e^{A\Delta_\emptyset}e^{(A+BF_{\{1\}})\Delta_1}x_b$ at the end of the cycle

$$\|x_e - x_b\| \leq \|P_{\{1\}}e^{A\Delta_\emptyset}x_m - \underbrace{x_m}_{=P_{\{1\}}x_m}\| + \|e^{(A+BF_{\{1\}})\Delta_1}x_b - x_b\| \leq$$
$$c_\emptyset\Delta_\emptyset\|P_{\{1\}}\|\|x_m\| + c_{\{1\}}\Delta_1\|x_b\| \leq c(\Delta_\emptyset + \Delta_1)\|x_b\| \leq c\Delta\|x_b\| \quad (3.56)$$

for certain constants $c_\emptyset$, $c_{\{1\}}$ and c, and $\Delta = \Delta_\emptyset + \Delta_1$ the duration of the complete cycle. Consider the sequence of states $\{x_i\}_{i\in\mathbb{N}}$ at the beginning of the cycles and let Δ_i be the duration of the i-th cycle starting in x_i and ending in x_{i+1}. Hence, (3.56) translates into $\|x_{i+1} - x_i\| \leq c\Delta_i\|x_i\|$ and yields $\|x_{i+1}\| \leq (1 + c\Delta_i)\|x_i\|$. Consequently, we have that

$$\|x_{i+1}\| \leq \prod_{j=1}^{i}(1 + c\Delta_j)\|x_0\|.$$

By taking the logarithm of this inequality and using that $\sum_{j=0}^{\infty}\Delta_j = \tau^*$, it can be seen that $\|x_i\| \leq e^{c\tau^*}\|x_0\|$. This implies that $x(t)$ is bounded on $[0, \tau^*)$. For $m > n$ it holds that

$$\|x_m - x_n\| \leq c\sum_{i=n}^{m-1}\Delta_i\|x_i\|.$$

Since $\sum_{i=0}^{\infty} \Delta_i = \tau^*$ and x is bounded on $[0, \tau^*)$ this yields that $\{x_i\}_{i\in\mathbb{N}}$ is a Cauchy sequence and hence has a limit. It is clear that then also $\lim_{t\uparrow\tau^*} x(t)$ must exist. Local existence of solutions implies that a solution can be defined beyond τ^*, which contradicts the definition of τ^*. Hence, $\tau^* = \infty$. □

LCS with low leading row indices

Theorem 3.6.11 *Consider the linear complementarity system (3.13) and let the leading row and column coefficient matrices $\mathcal{M}$ and $\mathcal{N}$ be P-matrices. If the leading row indices ρ_i are contained in $\{0, 1\}$ for all $i \in \bar{k}$, then the linear complementarity system (3.13) is globally well-posed.* □

Proof. According to Thm. 3.6.3 the system (3.13) is locally well-posed. Since local uniqueness is equivalent to global uniqueness of solutions, it remains to show that local existence results in global existence of solutions using the hypothesis in the formulation of the theorem. Define $K := \{i \in \bar{k} \mid \rho_i = 1\}$. The set of regular states $\mathcal{R}$ is equal to $\{x_0 \in \mathbb{R}^n \mid C_{K\bullet}x_0 \geq 0\}$ (Theorem 3.6.7). Since $\mathcal{R}$ is closed, it is invariant under the dynamics. Indeed, if $\mathcal{R}$ is not invariant, there exists an $x_0 \in \mathcal{R}$ such that a local solution $(\mathcal{E}, u_c, x_c, y_c)$ satisfies $x_c(0) = x_0$ and $x_c(t) \notin \mathcal{R}$ for $t \in (0, \varepsilon)$ for some $\varepsilon > 0$. The fact that $x_0 \in \mathcal{R}$ implies the existence of a $0 < \alpha < \varepsilon$ such that (u_c, x_c, y_c) is equal to a smooth initial solution on $[0, \alpha)$. This implies that for initial state $x_c(\tau)$ with $\tau \in (0, \alpha)$ there exists a *smooth* initial solution equal to $t \mapsto (u_c(t + \tau), x_c(t + \tau), y_c(t + \tau))$. Hence, $x_c(\tau) \in \mathcal{R}$ for $\tau \in (0, \alpha)$, which leads to a contradiction.

Suppose that the maximal interval on which a solution (u, x, y) (we omitted the subscript c) with initial state x_0 exists is equal to $[0, \tau^*)$ with $\tau^* < \infty$. Since every event time has at most multiplicity one, we can assume that $x_0 \in \mathcal{R}$ (otherwise take one initial jump). Since $\mathcal{R}$ is invariant under the dynamics of the LCS, it holds that $x(t) \in \mathcal{R}$ for all $t \in [0, \tau^*)$. In a continuous phase there is at most exponential growth, because the solution x is governed in mode I by $\dot{x} = (A + BF_I)x$ with F_I as in subsection 3.4.1. Since in each mode there is at most exponential growth without jumps, it is clear that $x(t)$ is bounded on $[0, \tau^*)$ (say $\|x(t)\| \leq M$ for all $t \in [0, \tau^*)$). Hence, when the solution is given on the interval $(s, t) \subseteq [0, \tau^*)$ by mode I, then

$$\|x(t) - x(s)\| = \|e^{(A+BF_I)(t-s)}x(s) - x(s)\| \leq c_I|t - s|\|x(s)\| \leq c_I M|t - s| \tag{3.57}$$

For arbitrary $(s, t) \subseteq [0, \tau^*)$ with x possibly evolving through several modes we obtain from (3.57) that

$$\|x(t) - x(s)\| \leq M \max_{I\in\mathcal{P}(\bar{k})} c_I \mid t - s \mid .$$

This implies that x is Lipschitz continuous on $[0, \tau^*)$ and thus also uniformly continuous. A standard result in mathematical analysis [169, ex. 4.13] states that $x^* :=$

$\lim_{t\uparrow\tau^*} x(t)$ exists and lies in $\mathcal{R}$ due to closedness of $\mathcal{R}$. Therefore, smooth continuation is possible from x^* beyond τ^*, because of the existence of local solutions. This contradicts the definition of τ^*. Hence, $\tau^* = \infty$. □

3.6.3 Mode selection by a finite LDCP

The next theorem states that in case $\mathcal{N}$ is a P-matrix, it is sufficient to consider $\text{LDCP}_n(x_0)$ (instead of $\text{LDCP}_\infty(x_0)$) for selection of a mode. Hence, only an algebraic problem with a finite number of constraints has to be solved.

Theorem 3.6.12 *Let a system (A, B, C, D) be given. If the leading column coefficient matrix $\mathcal{N}$ is a P-matrix, then from every initial state there exists a unique initial solution to (3.13). This solution evolves in mode I where $I := \{i \in \bar{k} \mid (u_i^{-n+1}, u_i^{-n+2}, \dots, u_i^{n-\eta_i}) \succ 0\}$ where $(u^j)_{j=-n+1}^n$, $(y^j)_{j=-n+1}^n$ constitutes a solution to $\text{LDCP}_n(x_0)$.* □

Proof. Let $(y^{-n+1}, y^{-n+2}, \dots, y^n)$ and $(u^{-n+1}, u^{-n+2}, \dots, u^n)$ be a solution to $\text{LDCP}_n(x_0)$ and let I be defined as in the formulation of the theorem. Define $p(0) := x_0 + \sum_{i=0}^{n-1} A^i B u^{-i}$. Note that this is the state after the jump induced by the impulsive distribution $\sum_{i=0}^{n-1} u^{-i}\delta^{(i)}$ starting from x_0. It follows from the definition of I that $(u_i^{-n+1}, \dots, u_i^{n-\eta_i}) = 0$, $i \in I^c$, and in combination with (3.35), (3.36) the same definition yields $(y_i^{-n+1}, \dots, y_i^n) = 0$, $i \in I$. Using (3.34b), we conclude that $p(0)$ satisfies

$$\begin{aligned}
0 = y_I^1 &= C_{I\bullet}p(0) + D_{II}v(1) \\
0 = y_I^2 &= C_{I\bullet}Ap(0) + D_{II}v(2) + C_{I\bullet}B_{\bullet I}v(1) \\
&\vdots \qquad\qquad\qquad \vdots \\
0 = y_I^n &= C_{I\bullet}A^{n-1}p(0) + D_{II}v(n) + C_{I\bullet}B_{\bullet I}v(n-1) + .. + C_{I\bullet}A^{n-2}B_{\bullet I}v(1)
\end{aligned} \tag{3.58}$$

with $v(i) = u_I^i$. By using (3.4) and the equations above, it can be shown that for all $j = 0, 1, \dots, n$ the vector $p(0) \in V_j(A, B_{\bullet I}, C_{I\bullet}, D_{II})$. Clearly, this implies that $p(0) \in \lim V_j(A, B_{\bullet I}, C_{I\bullet}, D_{II}) = V_n(A, B_{\bullet I}, C_{I\bullet}, D_{II}) = V_I$, for the algorithm converges within n steps (similarly as in the proof of Theorem 3.5.2). Hence, there exists a regular solution $(\mathtt{u}_{reg}, \mathtt{x}_{reg}, \mathtt{y}_{reg})$ to (3.14) in mode I with initial state $p(0)$. We define

$$\begin{aligned}
\tilde{\mathtt{u}} &:= \sum_{i=0}^{n-1} u^{-i}\delta^{(i)} + \mathtt{u}_{reg} \\
\tilde{y} &:= \mathtt{y}_{reg}.
\end{aligned}$$

Furthermore, $\tilde{\mathrm{x}}$ denotes the solution to (3.14) in mode I corresponding to $\tilde{\mathrm{u}}$ and initial state x_0. Note that according to Theorem 3.6.8 $y^{-n+1} = \ldots = y^0 = 0$. Obviously, this is a solution to (3.14) in mode I; so it only remains to show that $\tilde{\mathrm{u}}$, $\tilde{\mathrm{y}}$ are initially nonnegative. We shall do this by proving that $\mathrm{u}_{reg}^{(i)}(0) = u^{i+1}$ for all $i = 0, 1, \ldots, n - \eta_i - 1$ and consequently, $\mathrm{y}_{reg}^{(i)}(0) = y^{i+1}$.

Notice that both $v(i) = u_{reg,I}^{(i-1)}(0)$, $i = 1, \ldots, n$ and $v(i) = u_I^i$, $i = 1, \ldots, n$ satisfy (3.58). We extend the solution of $\mathrm{LDCP}_n(x_0)$ with zeros to get an infinite sequence $(u^{-n+1}, \ldots, u^n, 0, 0, \ldots)$. The difference $w(i) = u_{reg,I}^{(i)}(0) - u_I^{i+1}$, $i \geq 0$ can be taken as an input to the discrete-time system

$$\begin{aligned} q(i+1) &= Aq(i) + B_{\bullet I} w(i), \quad q(0) = 0 \\ \bar{y}(i) &= C_{I\bullet} q(i) + D_{II} w(i) \end{aligned} \tag{3.59}$$

satisfying $\bar{y}(0) = \ldots = \bar{y}(n-1) = 0$. Taking the z-transform of the discrete-time system (3.59) (see e.g. [117]) with input $w(i)$ gives (with some abuse of notation the z-transform of w is denoted by $w(z)$)

$$G_{II}(z) w(z) = \sum_{i=0}^{\infty} \bar{y}(i) z^{-i} = z^{-n} p(z) \tag{3.60}$$

for some proper rational vector function $p(z)$. For notational simplicity, we set $I = \bar{l}$, $l \in \bar{k}$. Since $\mathcal{N}_{II}$ is a P-matrix (and hence invertible), $G_{II}(z)$ can be written as

$$G_{II}(z) = V_2(z) \mathrm{diag}(z^{-\eta_1}, \ldots, z^{-\eta_l}), \tag{3.61}$$

where V_2 is biproper (i.e. proper rational with proper rational inverse), because $V_2(\infty) = \mathcal{N}_{II}$ is invertible (Theorem 4.5 in [82]). Hence, (3.60) yields

$$w(z) = G_{II}^{-1}(z) p(z) = \mathrm{diag}(z^{-\eta_1 - n}, \ldots, z^{-\eta_l - n}) \tilde{p}(z),$$

where $\tilde{p}(z) = V_2^{-1}(z) p(z)$ is proper. The definition of $w(i)$ now implies that

$$\mathrm{u}_{reg,j}^{(i)}(0) = u_j^{i+1},$$

for all $j \in I$ and $i = 0, 1, \ldots, n - \eta_j - 1$.

Since for $j \in I$,

$$(u_j^{-n+1}, \ldots, u_j^0, \mathrm{u}_{reg,j}^{(0)}(0), \ldots, \mathrm{u}_{reg,j}^{(n-\eta_i-1)}(0)) = (u_j^{-n+1}, \ldots, u_j^{n-\eta_i}) \succ 0$$

the distribution $\tilde{\mathrm{u}}_j \in \mathcal{C}_{imp}$ is initially positive for $j \in I$. Note that $\tilde{\mathrm{y}}_I = 0$ by construction of $\tilde{\mathrm{y}}$: $\tilde{\mathrm{y}} = \mathrm{y}_{reg}$ satisfies together with u_{reg} the condition (3.14) for mode I and initial state $p(0)$. Similarly, for $j \in I^c$, $\tilde{\mathrm{u}}_j = 0$. Note that

$$(y^{-n+1}, \ldots, y^0, \mathrm{y}_{reg}^{(0)}, \ldots, \mathrm{y}_{reg}^{(n-1)}) = (y^{-n+1}, \ldots, y^n) \succeq 0,$$

because $\mathrm{u}^{(i)}_{reg,j} = u_j^{i+1}$ for $j \in I$ and $i = 0, 1, \ldots, n - \eta_j - 1$. As a consequence we have that if $(y_j^{-n+1}, \ldots, y_j^n) \succ 0$, then $\tilde{\mathrm{y}}_j \in C_{imp}$ is initially positive. For $j \in I^c$, it may happen that $(y_j^{-n+1}, \ldots, y_j^n) = 0$; however, this implies that $\tilde{\mathrm{y}}_j$ is identically zero. To see this, note that y_{reg,I^c} can be written as the output of the system

$$\begin{aligned} \dot{x} &= (A + BF_I)x \\ \mathrm{y}_{reg,I^c} &= (C_{I^c} + D_{I^c\bullet}F_I)x, \end{aligned}$$

because the input u satisfying (3.14) can be given in feedback form by $u(t) = F_I x(t)$ (see section 3.4). By the Cayley-Hamilton theorem and because the state space dimension of the system is equal to n,

$$(y_j^{-n+1}, \ldots, y_j^0, \mathrm{y}_{reg,j}(0), \mathrm{y}^{(1)}_{reg,j}(0), \ldots, \mathrm{y}^{(n-1)}_{reg,j}(0)) = 0$$

implies

$$(y_j^{-n+1}, y_j^{-n+2}, \ldots, y_j^0, \mathrm{y}_{reg,j}(0), \mathrm{y}^{(1)}_{reg,j}(0), \ldots) = 0.$$

Since $\mathrm{y}_{reg,j}$ is of Bohl type, $\tilde{\mathrm{y}}_j = \mathrm{y}_{reg,j} \in C_{imp}$ is identically zero (Lemma 3.5.1). Hence, $(\tilde{\mathrm{u}}, \tilde{\mathrm{x}}, \tilde{\mathrm{y}})$ is an initial solution to (3.13).

Uniqueness follows from the fact that that $\mathrm{LDCP}_\infty(x_0)$ has a unique solution (Theorem 3.6.8). Indeed, the one-to-one correspondence between initial solutions and solutions to $\mathrm{LDCP}_\infty(x_0)$ implies that there is only one initial solution, which must evolve in the above mode. □

Remark 3.6.13 Since $\mathrm{LDCP}_\infty(x_0)$ has a unique solution, the mode I as defined in the previous theorem (selected by $\mathrm{LDCP}_n(x_0)$) is obviously contained in $\mathcal{I}^\infty_{\mathrm{LDCP}}(x_0) = \mathcal{I}_{\mathrm{RCP}}(x_0)$. Since there is only one corresponding initial solution, it evolves in all the modes contained in $\mathcal{I}^\infty_{\mathrm{LDCP}}(x_0)$. Hence, all selected index sets in $\mathcal{I}^\infty_{\mathrm{LDCP}}(x_0)$ are appropriate. Of course, the additional modes contained in $\mathcal{I}^\infty_{\mathrm{LDCP}}(x_0)$ are characterized by the undetermined index set K as in Remark 3.4.8. □

Remark 3.6.14 Solving $\mathrm{LDCP}_n(x_0)$ can be simplified by using Theorem 3.6.8. This theorem states that the variables y^{-n+1}, y^{-n+2}, …, y^0 and u_i^{-n+1}, u_i^{-n+2}, …, $u_i^{-\eta_i}$, $i \in \bar{k}$ can immediately be set to zero, which reduces the number of equations to be solved. □

In the section below, we illustrate the above theory by means of the two-carts example.

3.7 Algorithm for constructing solutions

In this section, a method will be proposed to construct analytical solutions to linear complementarity systems. The method will be illustrated by applying it to the two-carts example of Section 3.2. We emphasize that it is not the purpose of this chapter to give a *numerical* scheme for the simulation of complementarity systems, although the analytical algorithm may be used as a guideline for the development of such a scheme.

The algorithm is described by the following procedure.

Algorithm 3.7.1 Let x_0 be the initial state and T_e the final time.

0. initialization: Set $z := x_0$, $\mathcal{E} := \{0\}$, and $t' := 0$ as the initial state and time.

1. step one: Select for initial state z a mode $I \in \mathcal{I}(z)$.

2. step two: Consider the following two possibilities:

1. From the state z smooth continuation is possible in mode I, i.e. $z \in V_I$. Go to step four.
2. No smooth continuation is possible in mode I from z, i.e. $z \notin V_I$. Go to step three.

3. step three: Compute the projection P_I of z along T_I onto V_I (Subsection 3.4.2). Set $z := P_I z$. Go to step one.

4. step four: Compute the solution $(\mathrm{u}^{z,I}, \mathrm{x}^{z,I}, \mathrm{y}^{z,I})$ (see Subsection 3.4.1).

5. step five: Determine the next event time $\theta(z, I)$. Define $(u_c(t), x_c(t), y_c(t)) := (\mathrm{u}^{z,I}(t - t'), \mathrm{x}^{z,I}(t - t'), \mathrm{y}^{z,I}(t - t'))$ for $t \in (t', t' + \theta(z, I))$. Set $t' := t' + \theta(z, I)$, $\mathcal{E} := \mathcal{E} \cup \{t'\}$ and $z := x_c(t'-)$. If $t' \geq T_e$ the algorithm terminates. Otherwise, go to step one.

□

The algorithm can be visualized by the flow diagram as given by Figure 3.2.

Remark 3.7.2 Algorithm 3.7.1 produces a solution on $[0, T_e)$ if the following conditions are satisfied.

1. The algorithm does not get into a situation with $t' < T_e$ and $\mathcal{I}(z) = \emptyset$. Such a situation is called "deadlock."
2. All encountered event times have a finite multiplicity. Stated otherwise, the algorithm does not end up in an infinite loop consisting of only re-initializations and mode selections, where a limiting operation is required.
3. The event times do not have a finite accumulation point strictly smaller than T_e.

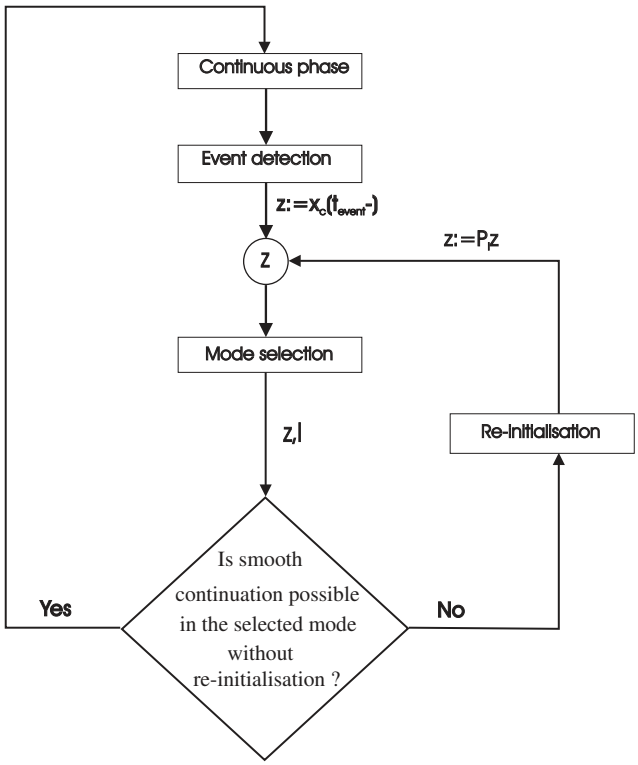

Figure 3.2: Schematic description of complete dynamics

□

Theorem 3.7.3 *Let a system* (A, B, C, D) *be given satisfying the conditions of Theorem 3.6.3. Algorithm 3.7.1 produces a solution on* $[0, T_e)$ *if and only if accumulation of events does not occur on the interval* $[0, T_e]$. □

Proof. By Theorem 3.6.3 the first two conditions mentioned in Remark 3.7.2 are satisfied (deadlock cannot occur and the maximal multiplicity of an event time is one). Therefore the result follows. □

Returning to the two-carts system of Section 3.2, we suppose that the initial state equals

$$x_0 = e^{-A}(0\ -1\ -1\ 0)^\top \approx (0.3202, -0.4335, 0.3716, -1.0915)^\top$$

and $T_e = 3$. Note that for this system the Markov parameters are given by $H^0 = H^1 = 0$ and $H^2 = \mathcal{M} = \mathcal{N} = 1$. Hence, the two-carts system satisfies the sufficient conditions for local well-posedness presented in this chapter. Consequently,

Algorithm 3.7.1 can only fail if the set of event times contains a finite accumulation point $\tau < 3$. According to Algorithm 3.7.1, we start by setting $\mathcal{E} := \{0\}$, $z := x_0$ and $t' := 0$.

Step one This step selects the unconstrained mode ($I = \emptyset \in \mathcal{I}(z)$), because the only initial solution for initial state z is $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ given by $(0, e^{At}z, Ce^{At}z)$. Note that $\mathtt{y}$ is initially nonnegative, because $\mathtt{y}(0+) = x_{01} \approx 0.3202$ is equal to the distance of the cart to the stop which is strictly positive.

Step two This step leads to the decision that smooth continuation in the selected mode is possible, because $z \in V_\emptyset = \mathbb{R}^4$ (every state is consistent for the unconstrained mode).

Step four The unconstrained dynamics is specified by a linear ordinary differential equation; the solution is equal to $\mathtt{u}^{z,I}(t) = 0$, $\mathtt{x}^{z,I}(t) = e^{At}z$, $\mathtt{y}^{z,I}(t) = Ce^{At}z$.

Step five Determining the zero crossing of $\mathtt{y}^{z,I}$ gives $\theta(z, I) := 1$. The corresponding state is equal to $(0, -1, -1, 0)^\top$, which is not regular for the unconstrained mode. Note that $\mathtt{y}^{z,I}_{reg}(1) = 0$, $\dot{\mathtt{y}}^{z,I}_{reg}(1) < 0$, so continuing in the unconstrained mode would violate the inequality constraint $y(t) \geq 0$. Hence, $u_c(t) = 0$, $x_c(t) = e^{A(t-1)}(0\ -1\ -1\ 0)^\top$, $y_c(t) = Ce^{At}(0\ -1\ -1\ 0)^\top$ for $t \in (0, 1)$, $\mathcal{E} = \{0, 1\}$, $t' := 1$ and $z := (0\ -1\ -1\ 0)^\top$. Since $t' < T_e$, we go to step one.

Step one For the purpose of illustrating mode selection by RCP, the dynamical system is transformed to the Laplace domain:

$$(s^4 + 3s^2 + 1)y(s) = \begin{pmatrix} s(s^2+1), & s, & s^2+1, & 1 \end{pmatrix} \begin{pmatrix} x_{10} \\ x_{20} \\ x_{30} \\ x_{40} \end{pmatrix} + (s^2+1)u(s). \tag{3.62}$$

Substituting z for $(x_{10}, x_{20}, x_{30}, x_{40})^\top$ results in

$$(s^4 + 3s^2 + 1)y(s) = -s - s^2 - 1 + (s^2+1)u(s).$$

Since $y(s)$ or $u(s)$ should be zero, there are only two possibilities:

$$\begin{aligned} \text{unconstrained mode: } u(s) = 0; \qquad & y(s) = \frac{-s^2 - s - 1}{s^4 + 3s^2 + 1} \\ \text{constrained mode: } y(s) = 0; \qquad & u(s) = 1 + \frac{s}{s^2+1}. \end{aligned}$$

Since the RCP requires nonnegativity for sufficiently large values of the indeterminate s, the combination $y(s) = 0$, $u(s) = 1 + \frac{s}{s^2+1}$ is the unique solution to RCP(z); so $\mathcal{I}(z) = \mathcal{I}_{\mathrm{RCP}}(z) = \{\{1\}\}$. Hence, the constrained mode must be selected ($I := \{1\}$).

Step two Since the solution to RCP(z) is not strictly proper, the answer to the question in the decision block in Figure 3.2 is negative, so we have to re-initialize.

Step three Using (3.4) and (3.7), we can compute the consistent states and the jump space:

$$T_{\{1\}} = \operatorname{Im}\begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix};\ V_{\{1\}} = \operatorname{Ker}\begin{pmatrix} 1\,0\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix} = \operatorname{Im}\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

To re-initialize we have to project z onto $V_{\{1\}}$ along $T_{\{1\}}$, which results in

$$z := P_{\{1\}}z = P_{V_{\{1\}}}^{T_{\{1\}}}z = (0, -1, 0, 0)^\top.$$

Step one We have to solve RCP(z):

$$(s^4 + 3s^2 + 1)y(s) = -s + (s^2 + 1)u(s)$$

together with the complementarity conditions. The only solution is $y(s) = 0$, $u(s) = \frac{s}{s^2+1}$ resulting in $I := \{1\}$.

Step two Since the solution to RCP(z) is strictly proper, smooth continuation in the selected mode is possible. The physical interpretation is clear: the left cart hits the stop. Instantaneously, the velocity is put to zero and the right cart keeps the left cart pushed against the stop.

Step four The dynamics of the constrained mode is given by a set of DAEs. However, these can easily be translated into an ODE (note that there must exist a linear mapping $F_{\{1\}}$ such that $u(t) = F_I x(t)$ satisfies the mode dynamics; see Subsection 3.4.1). The input u must be chosen in such a way, that it keeps y identically zero. Since $y = x_1$, $\dot{y} = x_3$, $\ddot{y} = 2x_1 + x_2 + u$, u should equal $-2x_1 - x_2$. (Note that $F_I = (-2\ -1\ 0\ 0)$ is a possible choice, but is not the only choice. $F_I = (\alpha\ -1\ \beta\ 0)$ is an alternative for every α and β, because $x_1 = x_3 = 0$ for consistent states.) Hence, the dynamics in the constrained mode is given by $x_1 = x_3 = 0$, $\ddot{x}_2 = -x_2$, $u = -x_2$. Solving this set of equations for initial state z gives $\mathrm{u}^{z,I}(t) = \cos t$, $\mathrm{x}_1^{z,I}(t) = 0$, $\mathrm{x}_2^{z,I}(t) = -\cos t$ and $\mathrm{y}^{z,I}(t) = 0$. Note that we could also have concluded this by taking the inverse Laplace transform of the solution $(u(s), y(s))$ to the RCP in the last mode selection.

Step five An event is detected at $\theta(z, I) = \inf\{t \geq 0 \mid \cos(t) < 0\} = \frac{\pi}{2}$. The piece of $(u_c(t), x_c(t), y_c(t))$ on $(1, 1 + \frac{\pi}{2})$ is given by the initial solution above as described in Algorithm 3.7.1. $\mathcal{E} := \{0, 1, 1 + \frac{\pi}{2}\}$, $t' := 1 + \frac{\pi}{2}$ and $z := (0, 0, 0, 1)^\top$. Since $t' < 3 = T_e$, we proceed with step one.

Step one This time LDCP will be demonstrated as a mode selection tool. Since the conditions of Theorem 3.6.12 are satisfied, a finite version of the LDCP can be used

for mode selection: $\mathrm{LDCP}_4(z)$ reads

$$\begin{aligned} y^{-3} &= 0 \\ y^{-2} &= 0 \\ y^{-1} &= u^{-3} \\ y^{0} &= u^{-2} \\ y^{1} &= u^{-1} - 2u^{-3} \\ y^{2} &= u^{0} - 2u^{-2} + u^{-3} \\ y^{3} &= u^{1} - 2u^{-1} + u^{-2} + 3u^{-3} \\ y^{4} &= 1 + u^{2} - 2u^{0} + u^{-1} + 3u^{-2} - 3u^{-3}, \end{aligned}$$

together with complementarity conditions (3.35) and (3.36). Setting $y^i = 0$, $i \in \{-3, \ldots, 4\}$ leads to $(u^{-3}, \ldots, u^1, u^2) = (0, \ldots, 0, -1) \prec 0$. Hence, (3.35) does not hold. It is obvious that setting $u^i = 0$, $i \in \{-3, \ldots, 4\}$ leads to $(y^{-3}, \ldots, y^3, y^4) = (0, \ldots, 0, 1) \succeq 0$ so that (3.36) holds. Hence, $\mathcal{S}^4_{\mathrm{LDCP}}(z) = \{\emptyset\}$ and the unconstrained mode must be selected ($I := \emptyset$).

Step two Since the impulsive part of u is zero, i.e. $u^{-3} = u^{-2} = u^{-1} = u^0 = 0$, smooth continuation is possible. This can also be observed from the fact that $(0, 0, 0, 1)^\top$ is a consistent state for the unconstrained mode. In terms of the physical system: the right cart is on the right of its equilibrium and pulls the left cart away from the stop.

Step four and five Determining a new piece of $(u_c(t), x_c(t), y_c(t))$ leads to $u_c(t) = 0$, $x_c(t) = e^{A(t-1-\frac{\pi}{2})}(0, 0, 0, 1)^\top$ and $y_c(t) = Ce^{A(t-1-\frac{\pi}{2})}(0, 0, 0, 1)^\top$ in the same way as before. The next event time $1 + \frac{\pi}{2} + \theta(z, I)$ is strictly larger than $T_e = 3$ so that the algorithm halts with a complete solution on $[0, 3)$.

The computed trajectory is plotted in figure 3.3. Note the complementarity between u and x_1 and the discontinuity in the derivative of x_1 at time $t = 1$.

To show that the particular mode transition mentioned in Section 3.2 can be handled properly by the proposed algorithm, we take the initial state $z_0 = x_0 = (0, 1, -1, 0)^\top$ (labeling of z_0 as in (3.21)). Substituting this initial condition in (3.62) results in

$$(s^4 + 3s^2 + 1)y(s) = s - s^2 - 1 + (s^2 + 1)u(s).$$

Solving $\mathrm{RCP}(z_0)$ (step one) leads to $y(s) = 0$ and $u(s) = 1 - \frac{s}{s^2+1}$ and so $\mathcal{S}_{\mathrm{RCP}}(z_0) = \{\{1\}\}$. We select the constrained mode ($I_1 = \{1\}$). Smooth continuation is not possible in the selected mode (step two), because the solution to RCP is not strictly proper. Re-initialization (step three) leads to $z_1 := P_{\{1\}}z_0 = (0, 1, 0, 0)^\top$. $\mathrm{RCP}(z_1)$ has to be considered (step one):

$$(s^4 + 3s^2 + 1)y(s) = s + (s^2 + 1)u(s).$$

Notice that setting $y(s)$ equal to zero results in $u(s) = -\frac{s}{s^2+1}$, the strictly proper part of the solution of $\mathrm{RCP}(x_0)$. This is not a valid choice. The only solution is $u(s) = 0$

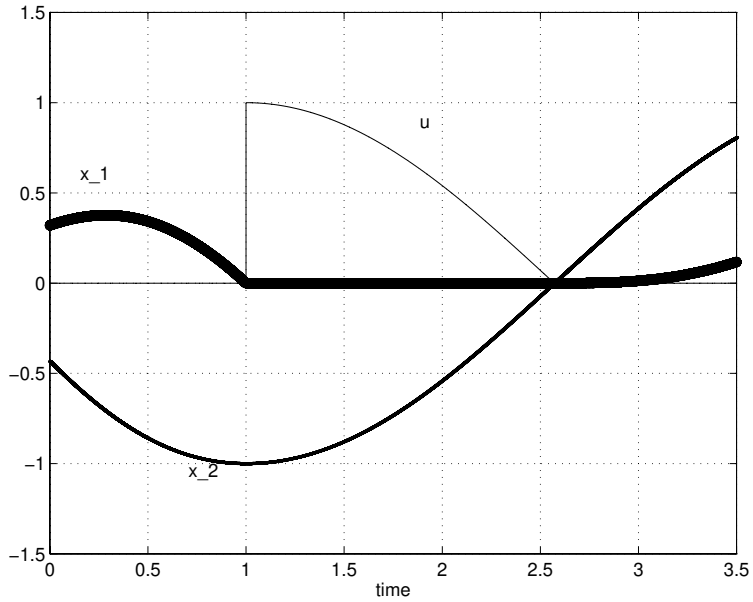

Figure 3.3: Solution trajectory of two-carts system.

and $y(s) = \frac{s}{(s^4+3s^2+1)}$, which corresponds to the unconstrained mode, i.e. $I_2 = \emptyset$. Since the solution of RCP(z_1) is strictly proper, smooth continuation is possible in the unconstrained mode (step two) and we can go to step four and five to compute the smooth continuation.

3.8 Mechanical Systems

In this section, it will be shown that the proposed mode selection rule coincides with the one of Moreau [139, 144] when these rules are applied to the class of systems that are covered by both frameworks, to wit, linear mechanical systems.

We will focus on linear mechanical systems whose dynamics in free motion is given by the differential equations

$$M\ddot{q}(t) + D\dot{q}(t) + Kq(t) = 0 \tag{3.63}$$

where q denotes the vector of generalized coordinates. Furthermore, M denotes the generalized mass matrix, which is assumed to be positive definite, D denotes the damping matrix and K is the elasticity matrix. The system is subject to unilateral constraints given by

$$Eq(t) \geq 0, \tag{3.64}$$

where E has full row rank. Furthermore, we assume that impacts are purely inelastic.

To obtain a complementarity formulation, we introduce the constraint forces $E^\top u$ with u the corresponding Lagrange multiplier needed to satisfy the unilateral constraints. Moreover, define the state vector x as $\mathrm{col}(q, \dot{q})$. According to the rules of classical mechanics, the system can then be written as follows (with omission of all time arguments)

$$\dot{x} = \underbrace{\begin{pmatrix} 0 & I \\ -M^{-1}K & -M^{-1}D \end{pmatrix}}_{A} x + \underbrace{\begin{pmatrix} 0 \\ M^{-1}E^\top \end{pmatrix}}_{B} u \tag{3.65a}$$

$$y = \underbrace{(E \;\; 0)}_{C} x \tag{3.65b}$$

$$0 \le y \perp u \ge 0 \tag{3.65c}$$

for all i. This systems satisfies $\rho_i = \eta_i = 2, i \in \bar{k}$; note that $\mathcal{M} = \mathcal{N} = EM^{-1}E^\top$ is positive definite and hence a P-matrix (Theorem 3.3.5). Hence, the system is locally well-posed (Theorem 3.6.3).

We consider only initial states $x_0 = \mathrm{col}(q_0, \dot{q}_0)$ with $Eq_0 \ge 0$. We call these points feasible. In the two-carts system, this means that we do not consider initial states for which the left cart starts on the left of the stop. In [139,144] no jumps occur in q itself, but jumps occur in the velocities $\dot{q}$. These jumps are governed in the inelastic impact case by the following minimization problem, where $J := \{i \in \bar{k} \mid E_{i\bullet}q_0 = 0\}$.

Minimization Problem 3.8.1 Let an initial state $x_0 = \mathrm{col}(q_0, \dot{q}_0)$ be given. The new state after re-initialization, denoted by $x(0+) = \mathrm{col}(q(0+), \dot{q}(0+))$, is determined by

$$\begin{aligned} q(0+) &= q_0 \\ \dot{q}(0+) &= \arg \min_{\{w \mid E_{J\bullet}w \ge 0\}} \tfrac{1}{2}(w - \dot{q}_0)^\top M (w - \dot{q}_0). \end{aligned}$$

□

The notation "arg min" denotes the set of vectors in the constrained set that minimize the criterion over the constrained set. Note that the minimization problem has a unique solution. The problem reflects a kind of "principle of economy": among the kinematically admissible right velocities, the one is chosen that is nearest in the kinetic metric [139, p. 75]. Observe that if we prove that jumps in our formulation correspond to the above minimization problem, then it follows that the feasible set $\{x \in \mathbb{R}^n \mid Cx \ge 0\}$ is invariant under the dynamics as introduced in Section 3.4, since the smooth dynamics do not take the solution outside this set.

The Kuhn-Tucker conditions [115] for the minimization problem give necessary conditions for optimality. The vector $\dot{q}(0+)$ is the minimizing argument only if there exists a Lagrange multiplier λ such that

$$M(\dot{q}(0+) - \dot{q}_0) - E_{J\bullet}^\top \lambda = 0 \tag{3.66}$$

$$0 \leq \lambda \perp E_{J\bullet}\dot{q}(0+) \geq 0. \tag{3.67}$$

The equality (3.66) is equivalent to

$$\dot{q}(0+) = \dot{q}_0 + M^{-1}E_{J\bullet}^{\top}\lambda \tag{3.68}$$

and therefore $\dot{y}(0+) = E\dot{q}(0+)$ and λ satisfy the following LCP with $\dot{y}_0 := E\dot{q}_0$:

$$\dot{y}_J(0+) = \dot{y}_0 + E_{J\bullet}M^{-1}E_{J\bullet}^{\top}\lambda \tag{3.69}$$

$$0 \leq \dot{y}_J(0+) \perp \lambda \geq 0. \tag{3.70}$$

According to Theorem 3.3.4, this LCP has a unique solution, because $E_{J\bullet}M^{-1}E_{J\bullet}^{\top}$ is a P-matrix. Since the minimization problem 3.8.1 is convex, the Kuhn-Tucker conditions are even sufficient for optimality. Hence, the LCP (3.69)-(3.70) is equivalent to the minimization problem for determining the jumps. This observation was also made in [178]. Notice that once this LCP is solved, the required jumps are known, because $\dot{q}(0+)$ then follows from (3.68).

We will prove now that $\mathrm{LDCP}_n(x_0)$ (and hence $\mathrm{LDCP}_\infty(x_0)$ and $\mathrm{RCP}(x_0)$) are equivalent to the optimization problem in the sense that both methods produce the same state jumps and select the same mode.

Theorem 3.8.2 *For linear mechanical systems of the form (3.65) with M positive definite and E of full row rank, the re-initialization by means of $LDCP_n(x_0)$ (or $LDCP_\infty(x_0)$ or $RCP(x_0)$) agrees with Moreau's rule for the inelastic impact case [139], [144] for feasible initial states. Linear mechanical complementarity systems are locally well-posed.* □

Proof. Since the row coefficient matrix and the column coefficient matrix are P-matrices, local well-posedness follows from Theorem 3.6.3. Furthermore, Theorem 3.6.8 states that $u^{-2} = u^{-3} = \ldots = u^{-n} = 0$. Because we start from a feasible state x_0, it follows that also $u^{-1} = 0$. Indeed, the first relevant LCP in the $\mathrm{LDCP}_n(x_0)$ (as in the proof of Theorem 3.6.7) is given by

$$y^1 = Cx_0 + CABu^{-1}$$

with the corresponding complementarity conditions. Since this LCP has a unique solution, the solution must satisfy $u^{-1} = 0$, because $Cx_0 \geq 0$. Hence, $y^{-n+1} = y^{-n+2} = \ldots = y^0 = 0$ and $y^1 = Cx_0$. The next relevant equality in (3.34) is

$$y^2 = CAx_0 + CABu^0. \tag{3.71}$$

We define J again as $\{i \in \bar{k} \mid C_i x_0 = 0\}$. Since one of the expressions (3.35) or (3.36) has to be satisfied for $i \in J$, the conditions

$$y_i^2 \geq 0,\ u_i^0 \geq 0,\ y_i^2 \perp u_i^0,\ i \in J$$

have to hold. Because $y_i^1 > 0$ for elements $i \in J^c$, $0 = u_i^0 = u_i^1 = \ldots = u_i^n$ must hold to satisfy (3.36). Considering only $i \in J$, we can write down the LCP following from (3.71) and the above complementarity conditions:

$$y_J^2 = C_{J\bullet}Ax_0 + C_{J\bullet}AB_{\bullet J}u_J^0 \tag{3.72}$$

$$0 \leq y_J^2 \perp u_J^0 \geq 0. \tag{3.73}$$

This LCP is identical to the LCP (3.69) and (3.70). This shows that the re-initialization by means of $\mathrm{LDCP}_n(x_0)$ leads to the same result as minimization problem 3.8.1. □

From this proof, we see that for feasible initial states only proper rational solutions to RCP occur, i.e. jumps only take place along Im B.

Example 3.8.3 To illustrate the equivalence of Moreau's rule and the complementarity rule, consider the two-carts system of Section 3.2 extended with a hook. See figure 3.4.

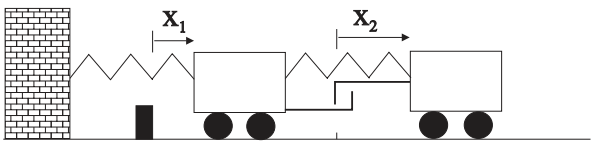

Figure 3.4: Two-carts system with hook.

The complementarity description is given by

$$\begin{aligned}
\dot{x}_1(t) &= x_3(t) \\
\dot{x}_2(t) &= x_4(t) \\
\dot{x}_3(t) &= -2x_1(t) + x_2(t) + u_1(t) + u_2(t) \\
\dot{x}_4(t) &= x_1(t) - x_2(t) - u_2(t) \\
y_1(t) &:= x_1(t) \\
y_2(t) &:= x_1(t) - x_2(t)
\end{aligned}$$

where u_1, u_2 denote the reaction forces exerted by the stop and hook, respectively. These equations are completed by the complementarity conditions (3.13c). Taking

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix};\ D = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix};\ K = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix};\ E = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix} \tag{3.74}$$

leads to a description as in the beginning of this section.

Using the minimization problem to determine the re-initialization and mode selection in case of an initial state $(x_{10}, x_{20}, x_{30}, x_{40})^\top$ with $x_{10} = x_{20} = 0$ results in

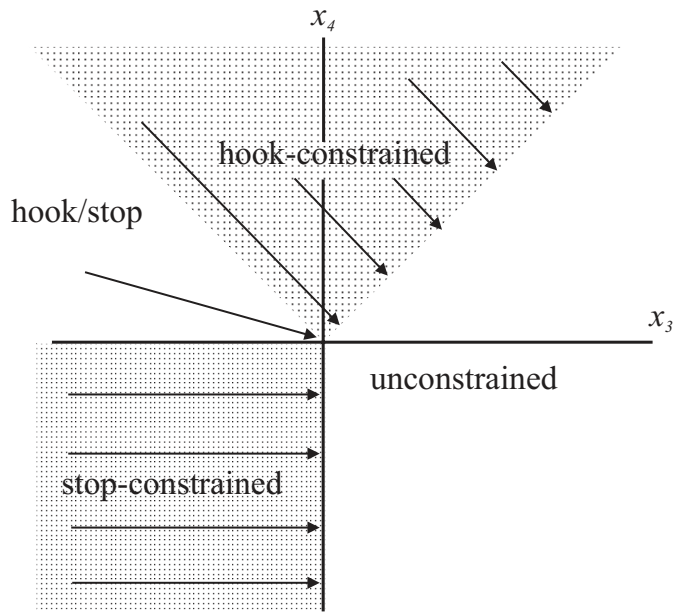

Figure 3.5: Re-initialization scheme

the alternatives shown in figure 3.5. Note that the minimization problem consists of finding the minimal distance to the feasible set (area indicated by "unconstrained"). The arrows denote the re-initialization directions.

To illustrate that RCP(x_0) gives the same results, the equations corresponding to (3.25) are given below:

$$\begin{aligned}(s^4+3s^2+1)y_1(s) &= (s^2+1)x_{30}+x_{40}+(s^2+1)u_1(s)+s^2u_2(s)\\ (s^4+3s^2+1)y_2(s) &= s^2x_{30}-(s^2+1)x_{40}+s^2u_1(s)+(2s^2+1)u_2(s).\end{aligned}$$

To determine the continuous states x_0 for which the stop-constrained mode ($I = \{1\}$) is selected, $y_1(s) \equiv 0$ and $u_2(s) \equiv 0$ are inserted in the equations above. Next we solve for $u_1(s)$ and $y_2(s)$, which leads to

$$\begin{aligned}u_1(s) &= -x_{30}-\frac{1}{s^2+1}x_{40}\\ y_2(s) &= \frac{1}{s^4+3s^2+1}[-s^2-1-\frac{-s^2}{s^2+1}]x_{40}.\end{aligned}$$

Entering the stop-constrained mode is only allowed if for sufficiently large values of the indeterminate s the above two expressions are nonnegative (see (3.26)). This requires $x_{30} \leq 0$ and $x_{40} \leq 0$. This indeed corresponds to the indicated area for the stop-constrained mode in figure 3.5. Note that the polynomial parts of u_1 and u_2

equal $-x_{30}$ and 0, respectively. Hence, $\mathtt{u}_{imp} = (-x_{30},\ 0)^\top \delta$ for the corresponding initial solution $(\mathtt{u}, \mathtt{x}, \mathtt{y})$. According to (3.10), the state jump equals $B(-x_{30},\ 0)^\top = (0,\ 0,\ -x_{30},\ 0)^\top$. This agrees with the direction of the arrows in figure 3.5. Similarly, the other modes and re-initialization directions can be verified.

This example shows also that the mode selection procedure that was suggested in [177] does not always agree with Moreau's impact rule for the inelastic case. This fact has already been mentioned in [178] without giving an explicit example. It is proposed there that if I is the current mode and violation of (3.17) occurs at time τ in state $x(\tau)$, the new mode is given by

$$J := (I \setminus \Gamma_2) \cup \Gamma_1,$$

where

$$\begin{aligned} \Gamma_1 &:= \{i \in I^c \mid \mathtt{y}_{reg,i}^{x(\tau),I} < 0,\ t \in (\tau, \tau+\varepsilon) \text{ for some } \varepsilon > 0\} \\ \Gamma_2 &:= \{i \in I \mid \mathtt{u}_{reg,i}^{x(\tau),I} < 0,\ t \in (\tau, \tau+\varepsilon) \text{ for some } \varepsilon > 0\}. \end{aligned}$$

In words, this means that constraints that are active or inactive according to mode I will become inactive or active, respectively, if their corresponding inequalities would be violated by continuation of the solution in mode I. In the example, this means that if we are in the unconstrained mode ($I = \emptyset$) and we arrive in $x(\tau) = (0, 0, -1, 2)^\top$, the selected mode should be $J = \{1, 2\}$, the hook/stop constrained mode. This does not agree with the minimization problem illustrated in figure 3.5, which indicates the hook-constrained mode. A physical argument against the choice in [177] in the indicated situation, might be that removing the stop does not lead to violation of $y_1(t) \geq 0$.

The above example also illustrates the fact that the solutions of linear complementarity systems do not always depend continuously on the initial state. The discontinuous dependence is caused by the sensitivity of solutions to the order in which constraints become active. Consider the initial states $x_0(\varepsilon) = (\varepsilon, \varepsilon, -2, 1)^\top,\ \varepsilon \geq 0$. For $\varepsilon = 0$ the solution is a jump to $(0, 0, 0, 0)^\top$, after which the system stays in its equilibrium position. For $\varepsilon > 0$, first the hook becomes active, resulting in a jump to $(\varepsilon, \varepsilon, -\frac{1}{2}, -\frac{1}{2})^\top$. This is followed by a regular continuation in the hook-constrained mode until the left cart hits the stop. The state just before the impact is $(0, 0, -\frac{1}{2} + g(\varepsilon), -\frac{1}{2} + g(\varepsilon))^\top$ for some continuous function $g(\varepsilon)$ with $g(0) = 0$. Re-initialization yields the new state $(0, 0, 0, -\frac{1}{2} + g(\varepsilon))^\top$, which converges to $(0, 0, 0, -\frac{1}{2})^\top$ if $\varepsilon \downarrow 0$. Obviously, the system has a discontinuity in $(0, 0, -2, 1)^\top$. One may also note that the sequence of initial states $x_0(\varepsilon) = (0, -\varepsilon, -2, 1)$, $\varepsilon \geq 0$ leads after two re-initializations for $\varepsilon \downarrow 0$ to the limit state $(0, 0, \frac{1}{2}, \frac{1}{2})$. This alternative limit corresponds to a situation in which first the stop-constrained and then the hook-constrained mode is active. □

3.9 Conclusions

The main purpose of this chapter has been to define a new class of dynamical systems called "linear complementarity systems". The definition builds on ideas from

linear system theory and from mathematical programming, and is motivated in part by systems of differential equations and algebraic inequalities that have been studied in mechanics and in electrical network theory. Applications are envisaged for instance in the modelling of power converters and other electrical networks that depend on controlled switching, in linear-quadratic control problems subject to linear inequality constraints, and in the study of piecewise linear systems.

A linear complementarity system can be viewed as a dynamical system that switches between several operating modes, and behaves as a linear system within each mode. The state spaces corresponding to different modes are in general not all of the same dimension, although they are naturally embedded in one encompassing space; in relation to this, state trajectories may exhibit discontinuities when a mode switch takes place. To give a precise definition of what is to be understood by a solution of a complementarity system, one has to be precise about the conditions under which a transition from one given mode to another given mode can take place, and one has to specify the associated jumps of the state variable. For mode selection, we have used ideas from mathematical programming, in particular from the theory of the linear complementarity problem [47]; for the determination of jumps we have relied on linear system theory, more specifically the geometric theory of linear systems [83].

When a class of dynamical systems is introduced, a first concern should be to give conditions for existence and uniqueness of solutions. We have given such conditions in terms of leading row and column coefficient matrices. Several methods for mode selection have been discussed, and a method for generating solutions has been presented. Also, we have shown that our notion of solution agrees with the one proposed by Moreau [139] for the class of systems that both solution concepts apply to.

In spite of the length of this chapter, it is clear that many issues remain to be investigated. The method that we have shown for constructing solutions only allows us to establish existence of solutions on intervals that do not contain accumulation points of the set of event times. To overcome this problem it seems necessary to work with sequences of approximating solutions, which may be generated for instance by time-stepping methods; compare the work by Stewart and Trinkle [192, 194]. A related issue is to provide conditions under which numerical solution methods for piecewise linear systems (see for instance [121]) can be shown to be consistent. The rational complementarity problem that has been discussed only briefly here is expected to play a crucial role in such investigations; see Chapter 4 for a more extensive treatment of the RCP.

Of course, all of the well-known topics of interest in dynamical systems theory can also be addressed in the context of complementarity systems: conditions for stability, existence of limit cycles, occurrence of chaos, and so on. Control of mechanical systems with unilateral constraints is discussed by Brogliato [31]. Perhaps the main challenge is to effectuate the interaction between the various fields of research that find a common meeting ground in complementarity systems.

4

The Rational Complementarity Problem

This chapter has been published in Linear Algebra and its Applications [93]. Parts of the chapter have been presented in an abridged form at the American Control Conference in Philadelphia (USA), June 1998 [88] and the Conference on Decision and Control in Tampa (USA), December 1998 [89].

4.1 Introduction

The classical *linear complementarity problem* (LCP) can be formulated as follows. Given a real k-dimensional vector q and a real $k \times k$ matrix M, find k-dimensional vectors y and u such that $y = q + Mu$ and for all indices i we have $y_i \geq 0$, $u_i \geq 0$, and at least one of y_i and u_i is zero. The LCP and various ramifications and generalizations of it play an important role in many optimization and equilibrium problems, and for this reason the LCP has been studied extensively in mathematical programming; see [47] for a comprehensive treatment. The rational complementarity problem (RCP), which is the main subject of this chapter, is a variation of the LCP in which the field of real numbers is replaced by the field $\mathbb{R}(s)$ of rational functions with real coefficients. To formulate a complementarity problem over $\mathbb{R}(s)$, we equip the field of rational functions with a suitable order to be defined below.

The RCP is motivated by its relations to a class of discontinuous dynamical systems, called linear complementarity systems (LCS) as studied in [87, 92, 177, 179]. Linear complementarity systems are specified by linear differential equations and inequalities similar to those appearing in the linear complementarity problem. Typical examples of such systems include mechanical systems subject to unilateral constraints, electrical networks with diodes, systems subject to relays and saturation characteristics, optimization problems with state constraints and systems with Coulomb friction. The dynamics of the complementarity class consists of continuous-time phases separated by state-events resulting in re-initializations of the continuous state of the system. In fact, in each continuous-time phase (called 'modes') the system is governed by its own

characteristic dynamic laws. The RCP plays a crucial role for LCS as it couples the continuous state to a corresponding mode. Systems in which continuous dynamics and switching rules are connected are called 'hybrid dynamical systems.' Hybrid systems have recently drawn much attention, see e.g. [7, 162]. In this field of research existence and uniqueness of solutions are often assumed, and sufficient conditions are rarely given. In previous papers [87,92,177,179] well-posedness results for LCS were obtained based on the so-called *linear dynamic complementarity problem*, a version of the complementarity problem based on taking derivatives of the LCS. The RCP has only been mentioned without exploiting its possibilities. In establishing a relationship between RCP and LCS, conditions for existence and uniqueness of solutions to LCS are derived in this chapter. These conditions are more general than the ones in [87, 92, 177, 179].

There is a connection between the RCP and a parameterized form of the LCP; this relation is explored in detail in this chapter. There are also relations between the RCP and certain generalizations of the LCP. Specifically, we discuss the *order complementarity problem* (OCP) that was defined in [22] as well as a version of the LCP defined over a general totally ordered field. We illustrate that certain results can be derived on an abstract level; however for the main part of the chapter we opt for a concrete treatment heading directly towards establishing the connection between RCP and a parameterized LCP. It is this connection (plus the body of knowledge already available for LCP) which enables us to establish well-posedness results for LCS. As specific applications we discuss linear mechanical systems with unilateral inelastic constraints, passive linear electrical networks with ideal diodes (and more generally linear dissipative systems with complementarity conditions), and linear systems with relays (based on LCP-results in [123]). The earlier well-posedness results in [87, 92, 177, 179] do not cover these special subclasses of complementarity systems.

The outline of the chapter is as follows. In the next two sections, we introduce some notational conventions and several complementarity problems: LCP, RCP, OCP and an 'abstract linear complementarity problem.' In section 4.4 necessary and sufficient conditions guaranteeing existence and uniqueness of solutions to RCP will be presented in terms of LCPs. In section 4.5 LCS will be introduced together with its solution concept. The connection between solutions to RCP and initial solutions to LCS will be stated. In the next section three physically relevant subclasses of complementarity systems are considered for which well-posedness results are obtained.

4.2 Notation

In this chapter, the following notational conventions will be in force. $\mathbb{N}$ denotes the natural numbers $\{0, 1, 2, \dots\}$, $\mathbb{R}$ the real numbers, $\mathbb{R}_+$ the nonnegative real numbers and $\mathbb{C}$ the complex numbers. For a positive integer l, $\bar{l}$ denotes the set $\{1, 2, \dots, l\}$. If a is a (column) vector with k components, we denote its i-th component by a_i. Given two vectors $a \in \mathbb{R}^k$ and $b \in \mathbb{R}^l$, then $\mathrm{col}(a, b)$ denotes the vector in $\mathbb{R}^{k+l}$

that arises from stacking a over b. The support of a vector $a \in \mathbb{R}^k$ is defined as $\operatorname{supp} a := \{i \in \bar{k} \mid a_i \neq 0\}$. $M^\top$ is the transpose of the matrix $M \in \mathbb{C}^{m\times n}$ and M^* denotes the complex conjugate transpose. A matrix $M \in \mathbb{C}^{m\times m}$ is called positive semi-definite if $2 \operatorname{Re} x^* M x = x^*(M + M^*)x \geq 0$ for all $x \in \mathbb{C}^m$. This is denoted by $M \geq 0$. In case strict inequality holds for all nonzero vectors x, we call the matrix positive definite and write $M > 0$. By $\mathcal{I}$ we denote the identity matrix of any dimension.

Given $M \in \mathbb{R}^{k\times l}$ and two subsets $I \subseteq \bar{k}$ and $J \subseteq \bar{l}$, the (I, J)-submatrix of M is defined as $M_{IJ} := (M_{ij})_{i\in I, j\in J}$. In case $J = \bar{l}$, we also write $M_{I\bullet}$ and if $I = \bar{k}$, we write $M_{\bullet J}$. The (I, I)-submatrices are sometimes called the principal submatrices. For a vector a, $a_I := (a_i)_{i\in I}$. A matrix $M \in \mathbb{R}^{k\times l}$ generates a convex cone, denoted by $\operatorname{pos} M$, obtained by taking nonnegative linear combinations of the columns of M. Formally,

$$\operatorname{pos} M := \{q \in \mathbb{R}^k \mid q = Mv \text{ for some } v \in \mathbb{R}^l_+\}.$$

By $\mathbb{R}(s)$ we denote the field of real rational functions in one variable. For reasons of clarity and cohesion, we shall systematically use a notation in which vectors over $\mathbb{R}(s)$ are written with an argument s and (vectors of) time functions appear with an argument t. Vectors over $\mathbb{R}$ are written without argument; distributions are also written without an argument, but in a different font. If $p(s) = 0$ for all s, we write (to avoid misunderstandings) $p(s) \equiv 0$. If $p(s)$ is not the zero polynomial, we write $p(s) \not\equiv 0$. $M(s) \in \mathbb{R}^{k\times l}(s)$ means that $M(s)$ is a $k \times l$ matrix with entries in $\mathbb{R}(s)$. Furthermore, the kernel of a rational matrix $M(s) \in \mathbb{R}^{k\times l}(s)$ over $\mathbb{R}(s)$ is denoted by $\ker_{\mathbb{R}(s)} M(s)$. The dimension of a linear subspace L of $\mathbb{R}^k(s)$ over $\mathbb{R}(s)$ is denoted by $\dim_{\mathbb{R}(s)} L$. A rational matrix is called (strictly) proper, if for all entries the degree of the numerator is smaller than or equal to (strictly smaller than) the degree of the denominator.

A vector $u \in \mathbb{R}^k$ is called nonnegative, and we write $u \geq 0$, if $u_i \geq 0$ for all $i \in \bar{k}$ and positive ($u > 0$), if $u_i > 0$ for all $i \in \bar{k}$. If two vectors $u, y \in \mathbb{R}^k$ satisfy that for all i at least one of u_i and y_i is zero, we write $u \perp y$. Similarly, we write $u(s) \perp y(s)$ for two rational vectors $u(s), y(s) \in \mathbb{R}^k(s)$, if for all i at least one of $u_i(s) \equiv 0$ and $y_i(s) \equiv 0$ is satisfied.

The set of arbitrarily often differentiable functions from $\mathbb{R}$ to $\mathbb{R}^m$ is denoted by $C^\infty(\mathbb{R}; \mathbb{R}^m)$.

4.3 Complementarity Problems

In this section, we introduce several instances of the complementarity problem. One of the fundamental results in the literature on complementarity problems will be examined for all versions of the complementarity problem considered here.

The linear complementarity problem (LCP) [47] is defined as follows.

Definition 4.3.1 (Linear complementarity problem) Given a matrix $M \in \mathbb{R}^{k\times k}$ and

a vector $q \in \mathbb{R}^k$. LCP(q, M) amounts to finding u, $y \in \mathbb{R}^k$ such that

$$\begin{aligned} y &= q + Mu && (4.1)\\ y &\geq 0,\ u \geq 0 && (4.2)\\ & y \perp u && (4.3) \end{aligned}$$

□

Recall that (4.3) implies that for all $i \in \bar{k}$ $y_i = 0$ or $u_i = 0$. Furthermore, it is evident that (4.2)-(4.3) can be replaced by $u \wedge y = 0$, where $\wedge$ denotes the componentwise minimum of two vectors.

LCP(q, M) is called *solvable*, if there exist u, $y \in \mathbb{R}^k$ satisfying (4.1), (4.2) and (4.3). LCP(q, M) is called *feasible*, if there exist u, $y \in \mathbb{R}^k$ that satisfy (4.1) and (4.2).

In [47], a wealth of theoretical and algorithmical results have been gathered concerning this fundamental problem in mathematical programming. We recall some notations and concepts from [47].

If we rewrite (4.1) as

$$q = -Mu + \mathbb{1}y = (-M\ \mathbb{1}) \begin{pmatrix} u \\ y \end{pmatrix}, \tag{4.4}$$

we see that we have to express q as an element of the cone pos $(-M\ \mathbb{1})$. However, this has to be done in a special way. In general, when $q = Az$ with $z_i \neq 0$, we say that the representation uses the column $A_{\bullet i}$ of A. The condition $y \perp u$ requires that in expressing q as an element of the cone pos $(-M\ \mathbb{1})$ not both $-M_{\bullet i}$ and $\mathbb{1}_{\bullet i}$ may be used.

Definition 4.3.2 Given $M \in \mathbb{R}^{k\times k}$, $J \subseteq \bar{k}$, $K \subseteq \bar{k}$, $J \cap K = \emptyset$ we define the matrix $C_M(J, K) \in \mathbb{R}^{k\times \text{card}(J\cup K)}$ as [1]

$$C_M(J, K) := (-M_{\bullet J}\ \mathbb{1}_{\bullet K}). \tag{4.5}$$

We define the *complementarity matrix* $C_M(J) \in \mathbb{R}^{k\times k}$ (relative to M) by

$$C_M(J) := C_M(J, J^c)$$

with $J^c := \bar{k} \setminus J := \{i \in \bar{k} \mid i \notin J\}$. The associated cone pos $C_M(J)$ is called a *complementarity cone* (relative to M). □

If $M \in \mathbb{R}^{k\times k}$, there are 2^k complementarity cones. From the discussion above Definition 4.3.2, it follows that if for some $q \in \mathbb{R}^k$ a solution to LCP(q, M) exists, then q has to be an element of a complementarity cone pos $C_M(J)$ for some $J \subseteq \bar{k}$.

[1] "card" denotes the cardinality of a set. For a finite set the cardinality is equal to the number of elements in the set.

Hence, the collection of vectors q for which a solution to LCP(q, M) exists is exactly the union of all complementarity cones of M, i.e.

$$\text{LCP}(q, M) \text{ has a solution iff } q \in \bigcup_{J \subseteq \bar{k}} \text{pos}\, C_M(J). \tag{4.6}$$

Hence, the existence of solutions to LCP(q, M) for all $q \in \mathbb{R}^k$ is equivalent to the union in (4.6) being equal to $\mathbb{R}^k$.

If we assume that all complementarity matrices of M are invertible, a necessary and sufficient condition for existence and uniqueness of solutions to LCP(q, M) for all q is that the 2^k complementarity cones of M form a 'partition' of the space $\mathbb{R}^k$. We call such a set of 2^k cones a partition of the vector space $\mathbb{R}^k$, if the union of the cones is the whole vector space and the intersection of any pair of cones is a lower dimensional cone (called 'face' or 'edge') [171].

For index sets I, $J \subseteq \bar{k}$ with the same number of elements the (I, J)-*minor* of M is the determinant of the square matrix $M_{IJ} := (M_{ij})_{i \in I, j \in J}$. The (I, I)-minors are also known as the principal minors. M is called a *P-matrix*, if all principal minors are strictly positive.

The following result is classical.

Theorem 4.3.3 *For given $M \in \mathbb{R}^{k \times k}$, the problem LCP($q$, M) has a unique solution for all vectors $q \in \mathbb{R}^k$ if and only if M is a P-matrix.* □

Proof. See [47, 171]. □

In this chapter we shall be motivated to consider a problem in which the role of the real numbers in the LCP is taken over by the field $\mathbb{R}(s)$ of rational functions with real coefficients. To formulate the "rational complementarity problem" it is convenient to first introduce a total ordering on $\mathbb{R}(s)$. One can define many orderings on $\mathbb{R}(s)$, but we shall be particularly interested in the following one.

Definition 4.3.4 A rational function $f(s) \in \mathbb{R}(s)$ will be said to be *nonnegative* if

$$\exists \sigma_0 \in \mathbb{R} \quad \forall \sigma \in \mathbb{R} \quad \{\, \sigma > \sigma_0 \;\Rightarrow\; f(\sigma) \geq 0 \,\}.$$

If this condition holds we write $f(s) \succeq 0$. □

In other words, a rational function $f(s)$ is nonnegative if and only if $f(\sigma)$ is nonnegative for all sufficiently large real σ. It is easily verified that the binary relation $\succeq$ so defined is indeed a total ordering on $\mathbb{R}(s)$. Indeed, a nonzero rational function must be either eventually positive or eventually negative, since a rational function can have only finitely many poles and zeros. The ordering defined above can also be described as the one induced by the lexicographic ordering of the coefficients of the Laurent series around infinity. On the rational vectors $\mathbb{R}^k(s)$ a partial ordering induced by the ordering in Definition 4.3.4 can be introduced as follows. We write for $f(s) \in \mathbb{R}^k(s)$

that $f(s) \succeq 0$ if and only if $f_i(s) \succeq 0$ for $i = 1, \ldots, k$. After these preparations, the RCP can now be stated as follows.

Definition 4.3.5 (Rational complementarity problem) Let a rational vector $q(s) \in \mathbb{R}^k(s)$ and a rational matrix $M(s) \in \mathbb{R}^{k\times k}(s)$ be given. The *rational complementarity problem* with *data* $q(s)$ and $M(s)$, denoted by RCP($q(s)$, $M(s)$), is the problem of finding rational k-vectors $u(s) \in \mathbb{R}^k(s)$ and $y(s) \in \mathbb{R}^k(s)$ such that

$$y(s) = q(s) + M(s)u(s) \text{ and } 0 \preceq u(s) \perp y(s) \succeq 0. \tag{4.7}$$

Any pair of rational vectors satisfying the above conditions is said to be a *solution* of RCP($q(s)$, $M(s)$). □

Writing out the RCP explicitly in terms of the ordering yields: find *rational* vector functions $u(s)$ and $y(s)$ such that

$$y(s) = q(s) + M(s)u(s) \text{ and } y^\top(s)u(s) = 0 \tag{4.8}$$

hold for all $s \in \mathbb{R}$ and there exists a $\sigma_0 \in \mathbb{R}$ such that for all $\sigma \geq \sigma_0$ we have

$$y(\sigma) \geq 0,\ u(\sigma) \geq 0. \tag{4.9}$$

The latter formulation of the RCP($q(s)$, $M(s)$) is used in [179].

Clearly, RCP is strictly analogous to LCP and one may expect that results like Theorem 4.3.3 will *mutatis mutandis* be valid for RCP. We shall prove below that this is indeed the case, but we shall also establish a relation between RCP and a parameterized version of LCP. Since a large body of results on LCP is available, it will prove to be convenient to have such a relation. First let us discuss how RCP fits into various possible generalizations of LCP.

Firstly, we note that $\mathbb{R}(s)$ can be looked at as an (infinite-dimensional) vector space over $\mathbb{R}$, and hence the same holds for $\mathbb{R}^k(s)$. Obviously the partial order $\succeq$ is compatible with the vector space structure of $\mathbb{R}^k(s)$ as a vector space over $\mathbb{R}$; moreover, for each two elements $f(s)$ and $g(s)$ there is a maximum $f(s) \vee g(s)$ and a minimum $f(s) \wedge g(s)$ (coinciding with the componentwise maximum and minimum), so that $\mathbb{R}^k(s)$ is actually a (real) *vector lattice* [159]. Therefore, RCP can be looked at as a special case of the *order complementarity problem* which is defined in [22]. This fact was pointed out to us by Kanat Çamlibel.

Definition 4.3.6 (Order complementarity problem) Let X be a vector lattice. Let a vector $q \in X$ and a linear mapping $M : X \to X$ be given. The *order complementarity problem* with *data* given by q and M (denoted by OCP(q, M)) is the problem of finding vectors u and y in X such that

$$y = q + Mu \text{ and } u \wedge y = 0. \tag{4.10}$$

Any pair of vectors (u, y) satisfying the above conditions is said to be a *solution* to OCP(q, M). □

To formulate a statement analogous to Theorem 4.3.3 for OCP, first the notion of a mapping of type (P) has to be introduced. In the definition below (taken from [22, Def. 2.10.b]) the notations $x^+ := x \vee 0$ and $x^- := -(x \wedge 0)$ are used for the positive and the negative parts of x.

Definition 4.3.7 Let X be a vector lattice. A linear mapping $M : X \to X$ is said to be of type (P) if the conditions

$$(Mx)^+ \wedge x^+ = 0 \text{ and } (Mx)^- \wedge x^- = 0 \tag{4.11}$$

hold only for $x = 0$. □

The definition could be summarized as: M is a mapping of type (P) if it does not reverse the sign of any nonzero vector. The result for OCP that is most closely to Theorem 4.3.3 is now the following [22, Thm. 2.14].

Theorem 4.3.8 *Let X be a vector lattice. A linear mapping $M : X \to X$ is of type (P) if and only if for each $q \in X$ the problem OCP(q, M) has at most one solution.* □

A real matrix is of type (P) if and only if it is a P-matrix (cf. [66], [47, Thm. 3.4.4]). In the general context of OCP, however, the type-(P) property is not strong enough to guarantee existence of solutions, as is shown by an example in [22].

Of course, it would be possible to consider a generalized OCP with vector lattices over $\mathbb{R}(s)$ rather than over $\mathbb{R}$. However, in this way we would not make use of the fact that in the rational complementarity problem we are dealing with a space that is finite-dimensional as a vector space over $\mathbb{R}(s)$. So, rather than looking at RCP as a special case of an OCP formulated over $\mathbb{R}(s)$, we will look at it as a special case of an abstract version of the standard LCP. This abstract version can be formulated as follows.

Definition 4.3.9 (Abstract linear complementarity problem) Consider a totally ordered field $(\mathrm{F}, \geq)$. Let q be a vector in F^k and let M be a matrix over F of size $k \times k$. The *linear complementarity problem over* F with *data* given by q and M ($\mathrm{LCP}_\mathrm{F}(q, M)$) is the problem of finding vectors u and y in F^k such that

$$y = q + Mu \text{ and } u \wedge y = 0. \tag{4.12}$$

Any pair of vectors (u, y) satisfying the above condition is said to be a *solution* to $\mathrm{LCP}_\mathrm{F}(q, M)$. □

Obviously, RCP is the same as $\mathrm{LCP}_{\mathbb{R}(s)}$, while $\mathrm{LCP}_\mathbb{R}$ is the standard LCP. So if we can prove that Theorem 4.3.3 and related results can be generalized to LCP_F, then we get immediate corollaries for the rational complementarity problem. Unfortunately it appears that the proofs of Theorem 4.3.3 that are available in the literature (for instance [47,171]) do not readily extend to the abstract case because of their dependence

on geometric intuition and/or topological properties of the real line. Below we shall present a proof of the abstract analogue of Theorem 4.3.3 on the basis of an indirect argument using a result from mathematical logic known as "Tarski's principle". Further on in the chapter we shall however use a different approach, using more concrete reasoning to obtain results that are formulated only for RCP; this will suffice for the intended applications to certain dynamical systems.

First we establish that in the context of an arbitrary totally ordered field, a matrix is a P-matrix if and only if it is of type (P) in the sense of Def. 4.3.7. The standard proof of this fact (see [47,66]) makes use of eigenvalues in a way that does not extend to general ordered fields.

Lemma 4.3.10 *Let $(\mathrm{F}, \geq)$ be a totally ordered field. The following properties are equivalent for matrices $M \in \mathrm{F}^{k\times k}$.*

(i) *All principal minors of M are positive.*

(ii) *If $x \in \mathrm{F}^k$ satisfies $(Mx)_i x_i \leq 0$ for all $i \in \{1, \ldots, k\}$, then $x = 0$.*

□

Proof. The proof of the implication from (i) to (ii) as given in [66] is directly applicable to the case in which the real line is replaced by an arbitrary totally ordered field, so we only need to prove the implication in the reverse direction. The proof will be given by induction with respect to the size of the principal submatrices of M. So suppose that (ii) holds, and consider first the minors corresponding to principal submatrices of M of size 1, i. e. the diagonal elements of M. Let e_p denote the p-th unit vector. Since obviously $(Me_p)_i(e_p)_i = 0$ for $i \neq p$, condition (ii) implies $M_{pp} = (Me_p)_p(e_p)_p > 0$. Assume now that all minors of principal submatrices of sizes up to $j-1$ are positive, and suppose that there is a principal submatrix M_{II} of size j such that $\det M_{II}$ is nonpositive. Take $p \in I$ and define $\tilde{I} := I \setminus \{p\}$. Let N be the matrix defined by

$$N = \lambda e_p e_p^\top, \quad \lambda = -\frac{\det M_{II}}{\det M_{\tilde{I}\tilde{I}}}. \tag{4.13}$$

Note that by our assumptions $\lambda \geq 0$. Since $(M+N)_{II}$ is obtained from M_{II} by adding λ times the p-th unit vector with $\mathrm{card}(I)$ components to the p-th column of M_{II}, and since the determinant of a matrix is linear as a function of each of its columns, we have

$$\det(M+N)_{II} \;=\; \det M_{II} + \lambda \det M_{\tilde{I}\tilde{I}} \;=\; 0.$$

Therefore, there exists a nonzero vector x_I such that $(M+N)_{II}x_I = 0$. Let x be the vector defined by $x_i = (x_I)_i$ for $i \in I$ and $x_i = 0$ for $i \notin I$. Write $y = Mx$, and note that $y_I = M_{II}x_I = -N_{II}x_I$. Consequently, for $i \notin I$ we have $y_i x_i = 0$ because $x_i = 0$, for $i \in \tilde{I}$ the relation $y_i x_i = 0$ holds because $y_i = 0$, and finally

$y_p x_p = -\lambda x_p^2 \leq 0$. Therefore condition (ii) is violated and we have reached a contradiction. □

To get the analogue of Theorem 4.3.3 for the abstract version of LCP we shall appeal to some ideas in mathematical logic, in particular a result known as *Tarski's principle*. We briefly review the most pertinent facts; see [167] for a complete treatment. A totally ordered field $(\mathrm{F}, \geq)$ is said to be *real closed* if its ordering $\geq$ is unique and there is no proper algebraic extension field of F that has an ordering extending $\geq$. It can be shown that a totally ordered field is real closed if and only if $\mathrm{F}(\sqrt{-1})$ is algebraically closed. For example, $\mathbb{R}$ is real closed but $\mathbb{R}(s)$ is not. It follows from Zorn's lemma that every totally ordered field admits an algebraic order extension that is real closed; by a theorem of Artin and Schreier [8], the real closure is unique up to isomorphism. An *elementary property* of a totally ordered field is one that can be stated in first-order logic (allowing quantification over individual elements but not over sets) using the algebraic operations and the order relation. Tarski's principle [167, Cor. 5.3] asserts that real closed fields are indistinguishable from $\mathbb{R}$ on the basis of elementary properties; so any elementary property that can be shown to hold in $\mathbb{R}$ is true in every real closed field.

Theorem 4.3.11 *Let* $(\mathrm{F}, \geq)$ *be a totally ordered field. The following statements are equivalent for matrices* M *in* $\mathrm{F}^{k\times k}$.

(i) *For all* $q \in \mathrm{F}^k$, *the problem* $LCP_{\mathrm{F}}(q, M)$ *has a unique solution.*

(ii) *All principal minors of* M *are positive.*

□

Proof. We have already shown in the foregoing lemma that (ii) is equivalent to the statement that M is of type (P). The implication from (i) to (ii) then follows as in [16, p. 274] (see also [22, Thm. 2.14]), since the argument given there, which proceeds from the assumption that M is of type (P), is valid over an arbitrary totally ordered field. It remains to prove the reverse implication. For this, note that the property expressed in the theorem is (for each given k) an elementary property. Since the statement is true for $\mathbb{R}$ by Theorem 4.3.3, it follows from Tarski's principle that the statement is also true for the real algebraic closure $\bar{\mathrm{F}}$ of F. In particular, if all principal minors of M are positive, then there exists for each given $q \in \mathrm{F}^k$ a unique pair of vectors y and u in $\bar{\mathrm{F}}^k$ such that $y = q + Mu$ and $y \wedge u = 0$. Let $I \subset \bar{k}$ be the set of indices i for which $y_i = 0$, and let $\tilde{M}$ be the matrix of size $k \times k$ whose j-th column equals the j-th column of $-M$ if $j \in I$, and is equal to the j-th unit vector if $j \notin I$. Note that $\tilde{M}$ is invertible, since its determinant is (up to a sign) a principal minor of M. Define $v = \tilde{M}^{-1}q \in \mathrm{F}^k$. Because $u_I = 0$ and $y_{I^c} = 0$ we must have $v_I = y_I$ and $v_{I^c} = u_{I^c}$, and in particular it follows that both y and u must actually belong to F^k. So we have constructed a solution to $\mathrm{LCP}_{\mathrm{F}}(q, M)$. Since the solution is unique over $\bar{\mathrm{F}}$, it is certainly also unique over F. □

In particular it follows that the rational complementarity problem $\mathrm{RCP}(q(s), M(s))$

has a unique solution for all $q(s)$ if and only if all principal minors of $M(s)$ are positive in the ordering that we defined on $\mathbb{R}(s)$. A corollary that is specific to RCP is the following.

Corollary 4.3.12 *For a rational matrix $M(s) \in \mathbb{R}^{k\times k}(s)$, the problem RCP($q(s)$, $M(s)$) has a unique solution for all $q(s) \in \mathbb{R}^k(s)$ if and only if there exists a $\sigma_0 \in \mathbb{R}$ such that for all $\sigma \geq \sigma_0$ the problem LCP(q, $M(\sigma)$) is uniquely solvable for all $q \in \mathbb{R}^k$.* □

Proof. According to Theorem 4.3.11, the first statement is true if and only if

$$\forall I \subset \bar{k} \quad \exists \sigma_0 \in \mathbb{R} \quad \forall \sigma \in \mathbb{R} \quad \{\sigma \geq \sigma_0 \Rightarrow \det M_{II}(\sigma) > 0\} \tag{4.14}$$

whereas the second statement can be reformulated as (Theorem 4.3.3)

$$\exists \sigma_0 \in \mathbb{R} \quad \forall \sigma \in \mathbb{R} \quad \forall I \subset \bar{k} \quad \{\sigma \geq \sigma_0 \Rightarrow \det M_{II}(\sigma) > 0\}. \tag{4.15}$$

Since the first quantification in (4.14) is over a finite set, the two statements are equivalent.
□

Note that the corollary is actually equivalent to Theorem 4.3.11 as applied to RCP. The connection between RCP and LCP as given in the corollary will be of crucial importance below to show well-posedness results for certain dynamical systems. Actually, we shall need some refinements of the corollary. Not in all cases does an "abstract" approach lead directly to a statement relating RCP and a parameterized LCP. Interchanging quantifiers is involved and this is not always as easy as in the proof above. Below we shall follow a "concrete" approach, in which we aim directly for connections between results connected to RCP and corresponding results connected to a parameterized LCP.

4.4 Relation between RCP and LCP

Let $q(s) \in \mathbb{R}^k(s)$ and $M(s) \in \mathbb{R}^{k\times k}$ be given. For any particular $\sigma \in \mathbb{R}$ the data of RCP (4.8)-(4.9) defines a standard LCP($q(\sigma)$, $M(\sigma)$). So, a connection between the RCP and the corresponding parameterized set of LCPs must exist, especially considering Corollary 4.3.12.

The first refinement of Corollary 4.3.12 is concerned with the question of existence of solutions to RCP independently of uniqueness. Note that the theorem below applies to RCP($q(s)$, $M(s)$) for a specific $q(s)$ and does not state a result for all possible $q(s) \in \mathbb{R}^k(s)$ as in Corollary 4.3.12. Therefore, the result below is much stronger. The proof is given in a direct way and not via the abstract route that was indicated in the previous section.

Theorem 4.4.1 *Let $q(s) \in \mathbb{R}^k(s)$ and $M(s) \in \mathbb{R}^{k\times k}(s)$ be given. RCP($q(s)$, $M(s)$) has a solution if and only if there exists a $\sigma_0 \in \mathbb{R}$ such that LCP($q(\sigma)$, $M(\sigma)$) has a solution for all $\sigma \geq \sigma_0$.* □

We would like to stress that the solvability of RCP($q(s)$, $M(s)$) is not completely characterized by the solvability of LCP($q(\infty)$, $M(\infty)$) where $q(\infty)$ and $M(\infty)$ denote the limits of $q(\sigma)$ and $M(\sigma)$ for $|\ \sigma\ | \longrightarrow \infty$, if they exist[2].

Example 4.4.2 Take

$$q(s) = (-1\ -\frac{1}{s}\ \ 1)^\top \text{ and } M(s) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Then RCP($q(s)$, $M(s)$) has no solutions, while LCP($q(\infty)$, $M(\infty)$) has uncountably many.

Conversely, RCP($q(s)$, $M(s)$) with

$$q(s) = (-1\ \ -1)^\top \text{ and } M(s) = \begin{pmatrix} 1+\frac{1}{s} & -1 \\ -1 & 1+\frac{1}{s} \end{pmatrix}$$

admits a solution (note that $M(\sigma)$ is a P-matrix for all nonnegative real σ), although LCP($q(\infty)$, $M(\infty)$) is unsolvable. □

□

Before we prove Theorem 4.4.1, we introduce some auxiliary concepts and results. Consider the equation

$$w = Mz, \quad z \geq 0 \tag{4.16}$$

for given vector $w \in \mathbb{R}^k$ and matrix $M \in \mathbb{R}^{k\times l}$. The solution set, defined as $S := \{z \geq 0 \mid w = Mz\}$, is a convex polyhedron (i.e. the intersection of finitely many closed halfspaces).

Definition 4.4.3 A solution z to (4.16) is said to be *basic* if $M_{\bullet \operatorname{supp} z}$ has full column rank. □

Remark 4.4.4 By convention, the matrix with no columns has full column rank. In this way, $z = 0$ is a basic solution to (4.16) with $w = 0$. □

Lemma 4.4.5 *If a solution to (4.16) exists, then there exists a basic solution as well.* □

[2]If the limits do not exist or are zero, one could perform some scaling on the equations of the RCP. Solvability of RCP($q(s)$, $M(s)$) is equivalent to solvability of RCP($D_1(s)q(s)$, $D_1(s)M(s)D_2(s)$) for diagonal rational matrices $D_i(s)$ where the diagonal elements are equal to some (negative, zero or positive) power of s.

Proof. See Theorem 2.6.12 in [47]. □

Definition 4.4.6 Let $q \in \mathbb{R}^k$ and $M \in \mathbb{R}^{k\times k}$ be given. A solution (u, y) to LCP(q, M) is basic, if col(u, y) is a basic solution to $q = (-M\ I)z$, $z \geq 0$. □

Lemma 4.4.7 *Let $q \in \mathbb{R}^k$ and $M \in \mathbb{R}^{k\times k}$ be given. If a solution to* LCP(q, M) *exists, then there exists a basic solution as well.* □

Proof. Let (u, y) be a solution to LCP(q, M). Consider the problem $q = (-M\ I)_{\bullet J} z$, $z \geq 0$ with $J = \text{supp}\,(\text{col}(u, y))$. Since this problem has a solution, Lemma 4.4.5 yields that it has a basic solution as well. Since this basic solution uses a subset of the columns used by col(u, y), it is clear that the complementarity conditions still hold for the basic solution. □

The last lemma before we can prove Theorem 4.4.1 is the following. We omit the proof which can be based on the Smith-McMillan form of rational matrices [132, Thm.2.3].

Lemma 4.4.8 *If $G(s)$ is a rational matrix, then the set of $\lambda \in \mathbb{C}$ for which $G(\lambda)$ has dependent columns coincides with the zero set of some polynomial.* □

Proof of Theorem 4.4.1 We divide the pairs (J, K) with $J, K \subseteq \bar{k}$ and $J \cap K = \emptyset$ in two sets $\mathcal{L}_{ind}$ and $\mathcal{L}_{dep}$ depending on the fact whether the columns of $C_{M(s)}(J, K)$ are independent over $\mathbb{R}(s)$ or not. By Lemma 4.4.8, there exist polynomials $p_{J,K}(s)$ satisfying for all $\lambda \in \mathbb{C}$, $p_{J,K}(\lambda) = 0$ if and only if $C_{M(\lambda)}(J, K)$ has dependent columns. Then $\mathcal{L}_{ind}$ and $\mathcal{L}_{dep}$ are given by

$$\begin{aligned} \mathcal{L}_{ind} &:= \{(J, K) \mid J, K \subseteq \bar{k},\ J \cap K = \emptyset,\ p_{J,K}(s) \not\equiv 0\} \\ \mathcal{L}_{dep} &:= \{(J, K) \mid J, K \subseteq \bar{k},\ J \cap K = \emptyset,\ p_{J,K}(s) \equiv 0\}. \end{aligned}$$

We take $\sigma_1 \geq \sigma_0$ (σ_0 as in the formulation of Theorem 4.4.1) such that σ_1 is larger than all real zeros of all the polynomials $p_{J,K}(s)$ that are not identically zero. As a consequence, if there exists a $\sigma \geq \sigma_1$ such that the real matrix $C_{M(\sigma)}(J, K)$ has (in)dependent columns, then the real matrix $C_{M(\sigma)}(J, K)$ has (in)dependent columns for all $\sigma \geq \sigma_1$.

Note that for $(J, K) \in \mathcal{L}_{ind}$, we have $q(s) \in C_{M(s)}(J, K)$ (for all s) if and only if the columns of the matrix $(q(s)\ C_{M(s)}(J, K))$ are dependent over $\mathbb{R}(s)$. Hence, we can apply Lemma 4.4.8 to get polynomials $r_{J,K}(s)$ satisfying for $(J, K) \in \mathcal{L}_{ind}$ and for $\sigma \in \mathbb{R}, \sigma > \sigma_1$, $r_{J,K}(\sigma) = 0$ if and only if $q(\sigma) \in C_{M(\sigma)}(J, K)$. Since the $r_{J,K}(s)$ are polynomials, we can find a real $\sigma_2 \geq \sigma_1$ (by taking it larger than all real zeros of all nonzero polynomials $r_{J,K}(s)$) with the property that if for some $(J, K) \in \mathcal{L}_{ind}$ there holds $q(\sigma) \in C_{M(\sigma)}(J, K)$ for certain real $\sigma \geq \sigma_2$, then $q(\sigma) \in C_{M(\sigma)}(J, K)$ for all $\sigma \in \mathbb{R}$. All pairs $(J, K) \in \mathcal{L}_{ind}$ for which $r_{J,K}(s) \equiv 0$ are denoted by $\mathcal{L}_{ind}^{con}$.

Finally, take $\sigma_3 \geq \sigma_2$ such that all components of the solutions of

$$q(s) = C_{M(s)}(J, K) \begin{pmatrix} u_J(s) \\ y_K(s) \end{pmatrix} \tag{4.17}$$

for $(J, K) \in \mathcal{L}_{ind}^{con}$ do not change sign anymore for $s \geq \sigma_3$. Since $C_{M(s)}(J, K)$ has independent columns over $\mathbb{R}(s)$ for $(J, K) \in \mathcal{L}_{ind}^{con}$, this solution is unique and rational. Hence, $\sigma_3 \geq \sigma_2$ has to be taken larger than all real zeros and poles of all nonzero entries of all the solutions to (4.17) corresponding to $(J, K) \in \mathcal{L}_{ind}^{con}$.

Take $\sigma \geq \sigma_3$. Since $\sigma \geq \sigma_3 \geq \sigma_0$, we have by the hypothesis of Theorem 4.4.1 that LCP$(q(\sigma), M(\sigma))$ has a solution (u, y) (by Lemma 4.4.7 we may assume that it is basic), that results in writing

$$q(\sigma) = C_{M(\sigma)}(I, I^c) \begin{pmatrix} u_I \\ y_{I^c} \end{pmatrix} \tag{4.18}$$

for some $I \subseteq \bar{k}$ and $\operatorname{col}(u_I, y_{I^c}) \geq 0$. The columns corresponding to indices that are not contained in $\operatorname{supp} \operatorname{col}(u_I, y_{I^c})$ are omitted resulting in

$$q(\sigma) = C_{M(\sigma)}(J, K) \operatorname{col}(u_J, y_K) \tag{4.19}$$

with $K \subseteq I^c$, $J \subseteq I$. Moreover, $C_{M(\sigma)}(J, K)$ has full column rank, because the solution (u, y) is basic. Hence, $(J, K) \in \mathcal{L}_{ind}$. By definition of σ_2, the fact that (4.19) is true for σ, and $\sigma \geq \sigma_2$, it follows that $(J, K) \in \mathcal{L}_{ind}^{con}$. This means that (4.17) has a solution $\operatorname{col}(u_J(s), y_K(s))$ for (J, K). Since $\operatorname{col}(u_J(\sigma), y_K(\sigma))$ satisfies (4.19) and $C_{M(\sigma)}(J, K)$ has full column rank, it is clear that $\operatorname{col}(u_J(\sigma), y_K(\sigma)) = \operatorname{col}(u_J, y_K) \geq 0$. Since $\operatorname{col}(u_J(s), y_K(s))$ does not change sign for $s \geq \sigma_3$, it is clear that $\operatorname{col}(u_J(s), y_K(s)) \geq 0$ for all $s \geq \sigma_3$. By introducing $u_{I \setminus J}(s) = 0$ and $y_{I^c \setminus K}(s) = 0$, $(u(s), y(s))$ is a solution to RCP$(q(s), M(s))$.

The other way around is easy. If $(u(s), y(s))$ is a solution to RCP$(q(s), M(s))$ satisfying $y(\sigma) \geq 0$, $u(\sigma) \geq 0$ for all $\sigma \geq \sigma_0$, then $(u(\sigma), y(\sigma))$ is a solution to LCP$(q(\sigma), M(\sigma))$ for all $\sigma \geq \sigma_0$. □

Next, the question of uniqueness of solutions to RCP$(q(s), M(s))$ is considered. We shall actually prove the following fairly general version.

Theorem 4.4.9 *Let $E \in \mathbb{R}^{l \times k}$, $q(s) \in \mathbb{R}^k(s)$ and $M(s) \in \mathbb{R}^{k \times k}(s)$ be given. The following statements are equivalent.*

1. *Any pair of solutions $(u^i(s), y^i(s))$, $i = 1, 2$ to RCP$(q(s), M(s))$ satisfies $Eu^1(s) = Eu^2(s)$ for all s.*
2. *There exists a real number σ_0 such that for all $\sigma \geq \sigma_0$ any pair of solutions (u^i, y^i), $i = 1, 2$ to LCP$(q(\sigma), M(\sigma))$ satisfies $Eu^1 = Eu^2$.*

□

From this it follows easily that uniqueness of solutions to LCP($q(\sigma)$, $M(\sigma)$) for all sufficiently large σ is equivalent to the uniqueness of the solution to RCP($q(s)$, $M(s)$).

Corollary 4.4.10 *Let $q(s) \in \mathbb{R}^k(s)$ and $M(s) \in \mathbb{R}^{k\times k}(s)$ be given.*
RCP($q(s)$, $M(s)$) has at most one solution if and only if there exists a real number σ_0 such that for all $\sigma \geq \sigma_0$ LCP($q(\sigma)$, $M(\sigma)$) has at most one solution. □

Proof. Take $E = \mathcal{I}$ in Theorem 4.4.9 and note that $u(s)$ determines $y(s)$ uniquely in the RCP and that u determines y uniquely in the LCP. □

Note that Corollary 4.4.10 is stronger than Corollary 4.3.12, because it treats uniqueness independently of existence of solutions and moreover, it states a uniqueness result for *separate* rational k-vectors instead of for all rational k-vectors.

Also uniqueness of solutions to RCP($q(s)$, $M(s)$) does not follow from uniqueness properties of solutions to LCP($q(\infty)$, $M(\infty)$) (provided the limits exist).

Example 4.4.11 Take

$$q(s) = (-1 \;\; -1)^\top \text{ and } M(s) = \begin{pmatrix} 1 + \frac{1}{s} & 1 \\ 1 & 1 \end{pmatrix}.$$

LCP($q(\infty)$, $M(\infty)$) has multiple solutions, while RCP($q(s)$, $M(s)$) has only one solution, because $M(\sigma)$ is a P-matrix for all $\sigma > 0$ (see Theorem 4.3.3 and Corollary 4.4.10).
□

□

The remainder of this section is devoted to the proof of Theorem 4.4.9, for which some preliminary results are needed.

Definition 4.4.12 Let C be a convex set. Then $z \in C$ is called an *extreme point* of C, if for all $z^1, z^2 \in C$ and for all $\lambda \in [0, 1]$

$$z = \lambda z^1 + (1 - \lambda)z^2,\; z^1 \neq z^2 \implies \lambda \in \{0, 1\}.$$

□

Lemma 4.4.13 *A solution to (4.16) is basic if and only if it is an extreme point of the solution set S.* □

Proof. See Theorem 2.6.13 in [47]. □

The following Lemma is known as Goldman's resolution theorem (Theorem 1 in [74], Theorem 2.6.23 in [47]). The vector in $\mathbb{R}^k$ with all components equal to 1 is denoted by e.

Lemma 4.4.14 *The solution set S of (4.16) has a finite number of extreme points, say $\{p^1, \ldots, p^r\}$. Define P as the* convex hull *of the extreme points of S (i.e. $P := \{\sum_{i=1}^r \alpha_i p_i \mid \alpha_i \geq 0,\ \sum_{i=1}^r \alpha_i = 1\}$) and define the cone $C := \{x \geq 0 \mid Mx = 0\}$. Then it holds that*

$$S = P + C.$$

Furthermore, if $Y := \{z \geq 0 \mid Mz = 0,\ e^\top z = 1\} \neq \emptyset$, then Y has a finite number of extreme points, say $\{y^1, \ldots, y^l\}$ and C equals pos $(y^1, \ldots, y^l)$. *$Y = \emptyset$ if and only if $C = \{0\}$.* □

Lemma 4.4.15 *Let E be a matrix in $\mathbb{R}^{l\times k}$. Suppose that (4.16) has (at least) two solutions z^i, $i = 1, 2$ with $Ez^1 \neq Ez^2$, but that any pair of basic solutions z^i_{bas}, $i = 1, 2$ satisfies $Ez^1_{bas} = Ez^2_{bas}$. Then there exists an index set I such that* ker $M_{\bullet I}$ *is nontrivial, no vectors in* ker $M_{\bullet I}$ *have components of opposite sign and this kernel is spanned by a vector $v \geq 0$ with $Ev \neq 0$ (in particular,* dim ker $M_{\bullet I} = 1$*).* □

Proof. According to Lemma 4.4.14 the solution set S of (4.16) can be written as $P + C$ with P and C as in Lemma 4.4.14. Since $Ep^1 = \ldots = Ep^r$ and $Ez^1 \neq Ez^2$, it is obvious that one of the extreme points of Y, as defined in Lemma 4.4.14, must be outside the kernel of E, say y^1. Take $I := \text{supp}\, y^1$. Note that $0 \neq y^1 \in \ker M_{\bullet I}$ and that $Ey^1 \neq 0$. Since y^1 is an extreme point of Y (or equivalently, y^1 is a basic solution to $Mz = 0$, $e^\top z = 1$, $z \geq 0$), $\ker M_{\bullet I} \cap \ker e_I^\top = \{0\}$ implying that $\dim \ker M_{\bullet I} \leq 1$. Hence, $\ker M_{\bullet I}$ is spanned by y^1 which has no components of opposite sign, because it is contained in Y. □

Remark 4.4.16 If no vectors in a nontrivial subspace V have components of opposite sign, then its dimension must be equal to one. Indeed, take two nonzero vectors $z^1 \geq 0$ and $z^2 \geq 0$ contained in V. Consider $z^1 - \alpha z^2$. When α increases from zero, all components must change from nonnegative to nonpositive at the same time, i.e. we must have $z^1 = \alpha z^2$ for some α. □

Lemma 4.4.17 *Let E be a matrix in $\mathbb{R}^{l\times k}$. Suppose that LCP(q, M) has (at least) two solutions (u^i, y^i), $i = 1, 2$ with $Eu^1 \neq Eu^2$, but that any pair of basic solutions (u^i_{bas}, y^i_{bas}), $i = 1, 2$ satisfies $Eu^1_{bas} = Eu^2_{bas}$. Then there exist a particular basic solution (u_{bas}, y_{bas}) and disjoint index sets J, K such that*

- supp $u_{bas} \subseteq J$, supp $y_{bas} \subseteq K$;
- *no vectors in* ker $C_M(J, K)$ *have components of opposite sign; and*
- *there is a vector* col$(z, w) \geq 0$ *with $w_{K^c} = 0$ and $z_{J^c} = 0$ such that* col(z_J, w_K) *spans* ker$C_M(J, K)$ *and $Ez \neq 0$.*

□

Proof. The set of all solutions of LCP(q, M) can be written as the union of the solution sets of $q = (-M\ \mathbb{1})\text{col}(u, y)$, $u_{J^c} = 0$, $y_J = 0$, $u \geq 0$ and $y \geq 0$ for all index sets $J \subseteq \bar{k}$. Consider an index set J whose corresponding system of equalities and inequalities allows at least two solutions $\text{col}(u^1, y^1)$, $\text{col}(u^2, y^2)$ with $Eu^1 \neq Eu^2$ and proceed as in the proof of Lemma 4.4.15. Note that such index sets must exist, because otherwise the hypothesis, that multiple solutions $(u^i, y^i), i = 1, 2$ to LCP(q, M) satisfy $Eu^1 \neq Eu^2$, is contradicted. □

Proof of Theorem 4.4.9 Suppose multiple solutions $(u^i(s), y^i(s))$, $i = 1, 2$ to RCP$(q(s), M(s))$ exist satisfying $Eu^1(s) \not\equiv Eu^2(s)$. Then $(u^i(\sigma), y^i(\sigma)), i = 1, 2$ form different solutions to LCP$(q(\sigma), M(\sigma))$ with $Eu^1(\sigma) \neq Eu^2(\sigma)$ for all $\sigma \in \mathbb{R}$ sufficiently large.

To prove the converse, we consider the collection of (J, K)-pairs with $J \cap K = \emptyset$ satisfying

$$\dim_{\mathbb{R}(s)} \ker_{\mathbb{R}(s)} C_{M(s)}(J, K) = 1.$$

We denote this set by $\mathcal{L}_1$. Let $\eta^{J,K}(s)$ be a polynomial vector spanning $\ker C_{M(s)}(J, K)$ for $(J, K) \in \mathcal{L}_1$. We define $\sigma_4 \in \mathbb{R}_+$ such that the components of $\eta^{J,K}(\sigma)$ for $(J, K) \in \mathcal{L}_1$ do not change sign anymore for $\sigma \in \mathbb{R}, \sigma \geq \sigma_4$.

Take $\sigma_5 \in \mathbb{R}_+$ such that for all (J, K)-pairs with $J, K \subseteq \bar{k}$ and $J \cap K = \emptyset$ the following is true:

$$\dim \ker C_{M(\sigma)}(J, K) = \dim_{\mathbb{R}(s)} \ker_{\mathbb{R}(s)} C_{M(s)}(J, K) \text{ for all } \sigma \geq \sigma_5.$$

We define $\sigma_6 := \max_{i \in \bar{5}} \sigma_i$ with σ_1, σ_2 and σ_3 as defined in the proof of Theorem 4.4.1. We claim that if there exists a real number $\sigma > \sigma_6$ with the property that LCP$(q(\sigma), M(\sigma))$ has multiple solutions (u^i, y^i), $i = 1, 2$ with $Eu^1 \neq Eu^2$, then there exist also multiple solutions $(u^i(s), y^i(s)), i = 1, 2$ to RCP$(q(s), M(s))$ with the property $Eu^1(\sigma) \neq Eu^2(\sigma)$.

Lemma 4.4.7 claims the existence of a basic solution to LCP$(q(\sigma), M(\sigma))$. If there exist two (or more) basic solutions (u^i_{bas}, y^i_{bas}), $i = 1, 2$ with $Eu^1_{bas} \neq Eu^2_{bas}$ the construction of the proof of Theorem 4.4.7 can be used to find two different solutions to RCP$(q(s), M(s))$. Note that the constructed solutions differ at $s = \sigma$.

If any pair of basic solutions (u^i_{bas}, y^i_{bas}), $i = 1, 2$ satisfies $Eu^1_{bas} = Eu^2_{bas}$, then Lemma 4.4.17 guarantees the existence of disjoint index sets J, K and a basic solution (u_{bas}, y_{bas}) with $\text{supp}\, u_{bas} \subseteq J$, $\text{supp}\, y_{bas} \subseteq K$ such that $\ker C_{M(\sigma)}(J, K)$ is nontrivial and no vectors in $\ker C_{M(\sigma)}(J, K)$ have components of opposite sign. Remark 4.4.16 states that $\dim \ker C_{M(\sigma)}(J, K) = 1$. The definition of σ_5 implies that $\dim_{\mathbb{R}(s)} \ker_{\mathbb{R}(s)} C_{M(s)}(J, K) = 1$ and the definition of σ_4 implies that the corresponding null vector $\eta^{J,K}(s)$, as defined above, does not change sign anymore beyond σ_4. Since $\eta^{J,K}(\sigma)$ spans $\ker C_{M(\sigma)}(J, K)$, it has no components of opposite sign. Without loss of generality we may assume that all components are nonnegative resulting in $\eta^{J,K}(s)$ having only nonnegative components for $s \geq \sigma_4$. The vector polynomial $\eta^{J,K}(s)$ can

be split in its J-part and K-part as $\mathrm{col}(\tilde{z}(s), \tilde{w}(s))$. We define $\mathrm{col}(z(s), w(s))$ by setting $z_J(s) := \tilde{z}(s)$, $z_{J^c}(s) = 0$, $w_K(s) = \tilde{w}(s)$ and $w_{K^c}(s) = 0$. Moreover, according to Lemma 4.4.17 we have $Ez(\sigma) = E_{\bullet J}\eta_J^{J,K}(\sigma) \neq 0$.

The construction as in the proof of Theorem 4.4.7 can be applied to the basic solution (u_{bas}, y_{bas}) of LCP$(q(\sigma), M(\sigma))$ to find a solution $(u(s), y(s))$ to RCP$(q(s), M(s))$ with $y_i(s) = 0$ if $i \notin \operatorname{supp} y_{bas}$ and $u_i(s) = 0$ if $i \notin \operatorname{supp} u_{bas}$. Looking at the support of $\mathrm{col}(z(s), w(s))$, it is observed that we can add a nonnegative multiple of $(z(s), w(s))$ to the solution $(u(s), y(s))$ without destroying the complementarity conditions. Furthermore, since $\mathrm{col}(z(s), w(s))$ has only nonnegative components for $s \geq \sigma_4$ the inequality conditions (4.9) remain valid for $(u^\alpha(s), y^\alpha(s)) := (u(s), y(s)) + \alpha(z(s), w(s))$, $\alpha \geq 0$. Hence, in this way we constructed an infinite number of solutions to the RCP$(q(s), M(s))$. Note that $Ez(\sigma) \neq 0$ implies that the constructed RCP-solutions satisfy $Eu^{\alpha_1}(\sigma) \neq Eu^{\alpha_2}(\sigma)$ if $\alpha_1 \neq \alpha_2$. □

The importance of the previously presented theorems is that the existence and uniqueness of solutions to RCP is related to existence and uniqueness of solutions to LCPs. A wealth of existence and uniqueness results concerning solutions to LCPs is already available in the literature (see [47]). These results can be applied to prove existence and uniqueness results for RCPs as is demonstrated by three classes of RCPs having a relation to dynamical systems. The relationship between RCP and a class of dynamical systems with discontinuous dynamics and impulsive motions is treated in the next section.

4.5 Relation between RCP and linear complementarity systems

In this section the relation of the RCP to *linear complementarity systems* will be discussed.

4.5.1 Linear complementarity systems

A linear complementarity system (LCS) is governed by the simultaneous equations

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{4.20a}$$

$$y(t) = Cx(t) + Du(t) \tag{4.20b}$$

$$0 \leq y(t) \perp u(t) \geq 0 \tag{4.20c}$$

The functions $u(t)$, $x(t)$, $y(t)$ take values in $\mathbb{R}^k$, $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively; A, B, C and D are constant matrices of appropriate dimensions. Equation (4.20c) implies that for all t and for every component $i = 1, \ldots, k$ at least one of $u_i(t) = 0$ and $y_i(t) = 0$ must be satisfied. This results in a multimodal system with 2^k modes, where each mode is characterized by a subset I of $\bar{k}$, indicating that $y_i(t) = 0$ if $i \in I$ and $u_i(t) = 0$ if

$i \in I^c$ with $I^c = \bar{k} \setminus I$. For each such mode the laws of motion are given by Differential and Algebraic Equations (DAEs). Specifically, in mode I they are given by

$$\begin{aligned} \dot{x} &= Ax + Bu && \text{(4.21a)} \\ y &= Cx + Du && \text{(4.21b)} \\ y_i &= 0,\ i \in I && \text{(4.21c)} \\ u_i &= 0,\ i \in I^c. && \text{(4.21d)} \end{aligned}$$

The mode will vary during the time evolution of the system. The system evolves in a certain mode as long as the inequality conditions in (4.20c) are satisfied. At the event of a mode transition, the system may display jumps (re-initialization) of the state variable. In the next subsection these phenomena will be formalized, which will result in a mathematically exact solution concept.

4.5.2 Solution concept of LCS

The solution concept of linear complementarity systems is based on a distributional framework as in [83]. This distributional framework is needed, because we have to be able to consider "impulsive motions." To make this plausible, consider a mechanical systems subject to some unilateral constraint, e.g. a particle moving around in a space which contains a wall. If the particle hits the wall with a nonzero velocity, a jump (a very fast motion) occurs in the velocity that can be modeled as the result of a Dirac pulse appearing in the reaction force exerted by the wall. Since such mechanical systems can be modeled as LCS, the previous motivates the choice for a distributional set-up as in [83] from which we recall some concepts below.

The set of distributions defined on $\mathbb{R}$ with support on $[0, \infty)$ is denoted by $\mathcal{D}'_+$ (see e.g. [183]). Particular examples of elements of $\mathcal{D}'_+$ are the delta distribution (or "Dirac pulse") and its derivatives. We denote the delta distribution by δ and its r-th derivative by $\delta^{(r)}$. Linear combinations of these particular distributions will be called *impulsive distributions*, that is, a distribution $\mathtt{u} \in \mathcal{D}'_+$ is an impulsive distribution, if it can be written as $\mathtt{u} = \sum_{i=0}^{l} u^{-i}\delta^{(i)}$ for scalars u^{-i}, $i = 0, \ldots, l$. A special subclass of $\mathcal{D}'_+$ is the set of regular distributions in $\mathcal{D}'_+$. These are distributions that are smooth on $[0, \infty)$. Formally, a distribution $\mathtt{u} \in \mathcal{D}'_+$ is *smooth* on $[0, \infty)$, if a function $v(t) \in C^\infty(\mathbb{R}; \mathbb{R})$ exists such that

$$\mathtt{u}(t) = \begin{cases} 0 & (t < 0) \\ v(t) & (t \geq 0). \end{cases}$$

Note that we use a different font for distributions to distinguish between the distribution $\mathtt{u}$, vectors $u \in \mathbb{R}^k$, (time-)functions $u(t)$ and rational functions $u(s)$.

Definition 4.5.1 [83] An *impulsive-smooth distribution* is a distribution $\mathtt{u} \in \mathcal{D}'_+$ of the form $\mathtt{u} = \mathtt{u}_{imp} + \mathtt{u}_{reg}$, where $\mathtt{u}_{imp}$ is impulsive and $\mathtt{u}_{reg}$ is smooth on $[0, \infty)$.

The class of these distributions is denoted by C_{imp}. If the regular part of an impulsive-smooth distribution is of the form

$$\mathrm{u}_{reg}(t) = \begin{cases} 0 & (t < 0) \\ Fe^{Gt}H & (t \geq 0) \end{cases} \tag{4.22}$$

for constant real matrices F, G and vector H of appropriate dimensions, we call the distribution of *Bohl type* or a *Bohl distribution*. □

Given an impulsive-smooth distribution $\mathrm{u} = \mathrm{u}_{imp} + \mathrm{u}_{reg} \in C_{imp}$, we define the leading coefficient of its impulsive part by

$$\mathrm{lead}(\mathrm{u}) := \begin{cases} 0, & \text{if } \mathrm{u}_{imp} = 0 \\ u^{-l} & \text{if } \mathrm{u}_{imp} = \sum_{i=0}^{l} u^{-i}\delta^{(i)} \text{ with } u^{-l} \neq 0. \end{cases} \tag{4.23}$$

Definition 4.5.2 We call a scalar-valued impulsive-smooth distribution $\mathrm{v} \in C_{imp}$ *initially nonnegative*, if

$$\begin{cases} \mathrm{lead}(\mathrm{v}) > 0, & \text{in case } \mathrm{v}_{imp} \neq 0 \\ \mathrm{v}_{reg}(t) \geq 0, \text{ for all } t \in [0, \varepsilon) \text{ for certain } \varepsilon > 0, & \text{otherwise.} \end{cases}$$

A scalar-valued impulsive-smooth distribution v is called *initially positive*, if v is initially nonnegative and additionally, if the impulsive part v_{imp} is equal to zero, it is required that $\mathrm{v}_{reg}(t) > 0$, for all $t \in (0, \varepsilon)$ for some $\varepsilon > 0$ (note that the interval is open from the left). An impulsive-smooth distribution in C_{imp}^k is called initially nonnegative (positive), if each of its components is initially nonnegative (positive). □

The initial nonnegativity or positiveness of a Bohl distribution can completely be characterized by its Laplace transform. This is not the case for general impulsive-smooth distributions. The simple proof of the following lemma is omitted.

Lemma 4.5.3 *1. Suppose that the Laplace transform of $\mathrm{u} \in C_{imp}^k$, denoted by $\hat{\mathrm{u}}(s)$, exists*[3]*. If u is initially positive, then there exists a $\sigma_0 \in \mathbb{R}$ such that the Laplace transform satisfies $\hat{\mathrm{u}}(\sigma) > 0$ for all real $\sigma \geq \sigma_0$. For a Bohl distribution the reverse statement holds as well.*

2. Suppose that $\mathrm{u} \in C_{imp}^k$ is of Bohl type and denote its Laplace transform by $\hat{\mathrm{u}}(s)$. There exists a $\sigma_0 \in \mathbb{R}$ such that the Laplace transform $\hat{\mathrm{u}}(s)$ satisfies $\hat{\mathrm{u}}(\sigma) \geq 0$ for all $\sigma \geq \sigma_0$ if and only if u is initially nonnegative.

3. Suppose $u(t)$ is a piecewise continuous function with $u(t) = 0, t < 0$ such that the Laplace transform, denoted by $\hat{u}(s)$, exists. Furthermore, assume the existence of a constant $\epsilon > 0$ such that $u(t) \geq 0$ for all $t \in [0, \epsilon]$ and $u(t) > 0$ for all $t \in (t_b, t_f) \subset [0, \delta]$ with $t_b < t_f$. Then there exists a $\sigma_0 \in \mathbb{R}$ such that $\hat{u}(\sigma) > 0$ for all $\sigma \geq \sigma_0$.

□

[3] We say that the Laplace transform exists, if the Laplace transform can be defined on a nontrivial half space of the complex plane.

To show that the reverse of statement 1 and statement 2 is not true for general impulsive-smooth functions, we consider the following counterexamples.

Example 4.5.4 We define for $\tau \in \mathbb{R}$ the functions $f_\tau(t) \in C^\infty(\mathbb{R}; \mathbb{R})$ as

$$f_\tau(t) = \begin{cases} 0, & t \leq \tau \\ e^{-\frac{1}{t-\tau}}, & t > \tau. \end{cases} \tag{4.24}$$

It can be verified that this defines indeed a class of C^∞-functions with derivatives equal to zero in $t = \tau$. A counterexample for the reverse of statement 1 is $f_1(t)$. The function $-f_1(t)$ shows also that statement 2 cannot be generalized to C_{imp}. □

□

Next, we define the concept of a distributional solution to a system of the form $\dot{x} = Kx + Lu$, $y = Mx + Nu$ with K, L, M and N constant matrices of appropriate dimensions. Let an initial condition x_0 (at time instant 0) be given. We replace the system by its distributional equivalent [83]:

$$\dot{\mathtt{x}} = K\mathtt{x} + L\mathtt{u} + x_0\delta \tag{4.25a}$$
$$\mathtt{y} = M\mathtt{x} + N\mathtt{u}, \tag{4.25b}$$

where $\dot{\mathtt{x}}$ denotes the distributional derivative of $\mathtt{x}$.

Definition 4.5.5 [83] A triple $(\mathtt{u}, \mathtt{x}, \mathtt{y}) \in \mathcal{D}_+^{\prime(m+n+r)}$ is a (distributional) solution to $\dot{x} = Kx + Lu$, $y = Mx + Nu$ with initial condition $x(0) = x_0$, if $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ satisfies (4.25) as an equality of distributions. □

In [83], it is shown that for equations of the form (4.25) there is for every $\mathtt{u} \in C_{imp}^m$ a unique pair $(\mathtt{x}, \mathtt{y}) \in D_+^{\prime(n+r)}$ such that $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ is a solution to (4.25) for given x_0; moreover $(\mathtt{x}, \mathtt{y}) \in C_{imp}^{n+r}$. Hence, given an initial state x_0, $\mathtt{u}$ can be seen as an input, because it uniquely determines $(\mathtt{x}, \mathtt{y})$. An important observation is that a nontrivial impulsive part of $\mathtt{u}$ may result in a re-initialization (also called "jump" or "impulsive motion") of the state. If $\mathtt{u}_{imp} = \sum_{i=0}^{l} u^{-i}\delta^{(i)}$ for vectors $u^{-i} \in \mathbb{R}^m$, then a jump will take place according to

$$\mathtt{x}_{reg}(0+) := \lim_{t \downarrow 0} \mathtt{x}_{reg}(t) = x_0 + \sum_{i=0}^{l} A^i B u^{-i}. \tag{4.26}$$

Next we will consider equations of the form (4.25) with the additional requirement that $\mathtt{y} = 0$.

Definition 4.5.6 A state x_0 is said to be *consistent* for (K, L, M, N), if there exists a regular input u such that

$$\begin{array}{rcl} \dot{\mathrm{x}} & = & K\mathrm{x} + L\mathrm{u} + x_0\delta \\ 0 & = & M\mathrm{x} + N\mathrm{u} \end{array} \tag{4.27}$$

is satisfied. $V(K, L, M, N)$ denotes the set of all consistent states for the system (K, L, M, N) and is called the *consistent subspace*.

□

The next lemma specifies a particular form of the regular inputs satisfying (4.27).

Lemma 4.5.7 *Consider (4.27) with K, L, M, N constant matrices of appropriate dimensions and write $V = V(K, L, M, N)$. There exists a matrix F of appropriate dimensions such that $(K + LF)V \subseteq V$ and $(M + NF)V = \{0\}$.* □

Proof. See Theorem 3.10 in [83]. □

The previous lemma shows that $V = V(K, L, M, N)$ can be made *invariant* by applying a feedback law $u(t) = Fx(t)$. By this we mean, that if $x_0 \in V$, then the solution of the closed-loop dynamics (i.e. after applying the feedback law) $\dot{x}(t) = Kx(t) + Lu(t) = (K + LF)x(t)$ with $x(0) = x_0$ satisfies $x(t) \in V$ for all $t \in \mathbb{R}_+$. This is a consequence of $(K + LF)V \subseteq V$. Furthermore, since $(M + NF)V = \{0\}$, it even holds that $Mx(t) + Nu(t) = (M + NF)x(t) = 0$. Note that the corresponding open-loop control function $u(t) = Fx(t) = Fe^{(A+BF)t}x_0$ is a Bohl function.

After these preliminaries we can define an initial solution to (4.20) given an initial state (see also Chapter 3).

Definition 4.5.8 We call $(\mathrm{u}, \mathrm{x}, \mathrm{y}) \in C_{imp}^{k+n+k}$ an *initial solution* to (4.20) with initial state x_0, if there exists an $I \subseteq \bar{k}$ such that

1. $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ is a solution to (4.21a)-(4.21b) with initial state x_0 in the distributional sense;
2. u and y satisfy (4.21c)-(4.21d) as equalities of distributions; and
3. u, y are initially nonnegative.

□

Obviously, an initial solution only satisfies the equations (4.20) "temporarily." In case an initial solution has a nontrivial impulsive part, only the re-initialization as given in (4.26) forms a piece of the global solution. If the initial solution $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ is smooth, the restriction $(\mathrm{u}, \mathrm{x}, \mathrm{y})\,|_{[0,\varepsilon)}$ satisfies the equations (4.20) on the interval $[0, \varepsilon)$, where ε is given by

$$\varepsilon := \inf\{t > 0 \mid \mathrm{u}_{reg,i}(t) < 0 \text{ or } \mathrm{y}_{reg,i}(t) < 0 \text{ for some } i \in \bar{k}\} \tag{4.28}$$

Only if $\varepsilon = \infty$ $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ forms a global solution to the LCS (4.20). If $\varepsilon < \infty$, the global solution is continued with a part of a different initial solution corresponding to initial state $\mathtt{x}_{reg}(\varepsilon)$. Such a definition of a (global) solution to (4.20) based on concatenation of initial solutions is formalized below. Given a state x_0, we define $\mathcal{I}(x_0)$ by

$$\mathcal{I}(x_0) := \{I \subseteq \bar{k} \mid \text{there exists an initial solution } (\mathtt{u}, \mathtt{x}, \mathtt{y}) \text{ to (4.20) that satisfies (4.21) for mode } I\}. \quad (4.29)$$

The set $\mathcal{I}(x_0)$ denotes the set of possible modes that can be selected from x_0. In Chapter 3 it has been shown that several other mode selection methods yield the same set of continuation modes (under some mild assumptions). One of them is the RCP.

Definition 4.5.9 A *solution* to (4.20) on $[0, T_e)$, $T_e > 0$ with initial state x_0 consists of a 6-tuple $(\mathcal{D}, \tau, x_e, u_c(t), x_c(t), y_c(t))$ where $\mathcal{D}$ is either $\{0, \ldots, N\}$ for some $N \geq 0$ or $\mathbb{N}$,

$$\begin{aligned}
\tau : \quad & \mathcal{D} && \to [0, T_e) \\
x_e : \quad & \mathcal{D} && \to \mathbb{R}^n \\
u_c(t) : \quad & (0, T_e) \setminus \tau(\mathcal{D}) && \to \mathbb{R}^k \\
x_c(t) : \quad & (0, T_e) \setminus \tau(\mathcal{D}) && \to \mathbb{R}^n \\
y_c(t) : \quad & (0, T_e) \setminus \tau(\mathcal{D}) && \to \mathbb{R}^k,
\end{aligned}$$

that satisfies the following.

1. There exists a function $I : \mathcal{D} \to 2^{\bar{k}} := \{J \mid J \subseteq \bar{k}\}$ with $I(i) \in \mathcal{I}(x_e(i))$.

2. On an interval $(a, b) \subseteq [0, T_e)$ with $a = \tau(i) < b$ for certain $i \in \mathcal{D}$ and $(a, b) \cap \tau(\mathcal{D}) = \emptyset$, $(u_c(t), x_c(t), y_c(t))$ is smooth and is equal to a smooth initial solution $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ in mode $I(i)$ with initial state $x_e(i)$ (i.e. $(u_c(t), x_c(t), y_c(t)) = (\mathtt{u}(t-a), \mathtt{x}(t-a), \mathtt{y}(t-a))$ for all $t \in (a, b)$). Furthermore, $u_c(t) \geq 0$ and $y_c(t) \geq 0$ hold for all $t \in (a, b)$.

3. (a) $\tau(0) = 0$

   (b) If $\mathcal{D} = \mathbb{N}$ then $\sup_{i \in \mathcal{D}} \tau(i) = T_e$.

4. $x_e(0) = x_0$.

5. If $\tau(i+1) > \tau(i)$, then $x_e(i+1) = \lim_{t \uparrow \tau(i+1)} x_c(t)$. If $\tau(i+1) = \tau(i)$, then there must exist an initial solution $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ in mode $I(i)$ with initial state $x_e(i)$ such that $x_e(i+1) = \lim_{t \downarrow 0} \mathtt{x}_{reg}(t)$ for all i with $i \in \mathcal{D}$, $i+1 \in \mathcal{D}$.

□

The interpretation of these notions and requirements will briefly be given. The function τ specifies the event times: the times at which the active mode changes. The set $I(i)$ denotes the active mode between $\tau(i)$ and $\tau(i+1)$. The triple $(x_c(t), u_c(t), y_c(t))$ denotes the trajectories in the continuous phases of the complementarity system (as imposed by item 2.) and $x_e(i)$ denotes the event state at time $\tau(i)$. Items $3(a)$ and 4 specify the initial conditions. Item $3(b)$ requires that the 6-tuple defines a solution on $[0, T_e)$ in case that T_e is an accumulation point of event times. The relation between two successive event states is described in 5. in case of smooth continuation and in case of re-initialization. In this definition there is some redundancy allowed in the number of events (size of $\mathcal{D}$) and the event times. Given a solution $(\mathcal{D}, \tau, x_e, u_c(t), x_c(t), y_c(t))$, one could add — without violating the requirements — between any two event times $\tau(i)$ and $\tau(i+1)$ with $\tau(i) < \tau(i+1)$ an additional event time $\tilde{\tau}$ by introducing $x_e(\tilde{\tau}) = x_c(\tilde{\tau})$. Similarly, one could also add a void re-initialization, when a regular initial solution exists from a certain state.

In Chapter 3 a more general solution concept is given. The extensions are twofold. The solution as in Definition 4.5.9 allows only finitely many re-initializations at one time instant, while the solution concept in Chapter 3 may have infinitely many re-initializations as long as the event states converge. However, sufficient conditions are known that guarantee that at most one re-initialization is required before smooth continuation is possible, see Chapter 3. These conditions are formulated in terms of *leading column and row coefficient matrices* being P-matrices. The second extension is concerned with possibly continuing a solution after an accumulation point of events (i.e. the existence of a $\tau^* < \infty$ such that $\lim_{i\to\infty} \tau(i) = \tau^*$). Using the solution concept above the largest interval on which a solution can be defined is $[0, \tau^*)$. However, in Chapter 3 the solution concept includes continuation from an accumulation point, if the state trajectory $x_c(t)$ has a left limit at τ^*.

In Chapter 3 a method has been proposed to construct analytical solutions to a LCS (4.20). This method can be used as a first set-up for simulation tools. The method can briefly be summarized as follows. Starting from an initial state x_0 one constructs an initial solution (see also next subsection for the relation to RCP). If the initial solution is smooth, there exists an interval $[0, \varepsilon)$ with $\varepsilon > 0$ as in (4.28) such that all the equations in (4.20) are satisfied. To determine ε one has to detect when the inequalities $u(t) \geq 0$ and $y(t) \geq 0$ are violated. In this way a smooth piece $(u_c(t), x_c(t), y_c(t))$ is constructed on $[0, \varepsilon)$. From $x_c(\varepsilon)$ one must find a new initial solution.

If the initial solution corresponding to x_0 has a nontrivial impulsive part, the re-initialized state according to (4.26) must be computed. Next one determines a new initial solution with the re-initialized state as new initial condition and one considers the two possibilities (impulsive or smooth initial solution) again. This cycle is repeated till a solution is constructed on the desired interval $[0, T_e)$.

Currently numerical simulation techniques based on time-stepping methods as in [121] (electrical circuits) and [192] (mechanical systems with impacts and friction) are under study.

4.5.3 Relation between existence and uniqueness of solutions to RCP and LCS

A special form of RCP($q(s)$, $M(s)$) arises when

$$q(s) := C(s\mathbb{1} - A)^{-1}x_0 \text{ and } M(s) := C(s\mathbb{1} - A)^{-1}B + D \tag{4.30}$$

for $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$, $D \in \mathbb{R}^{k\times k}$ and $x_0 \in \mathbb{R}^n$. We denote this case of RCP by RCP(x_0) assuming that A, B, C, D are clear from the context.

We generalize a result presented in Chapter 3. In Chapter 3 the following theorem was proven under an additional constraint on the separate mode dynamics (4.21) implying that all initial solutions are automatically Bohl distributions. The theorem below expresses that solvability of the RCP is related to existence of *initial* solution. Note that this is a local result, since it does not claim existence of a global solution as in Definition 4.5.9.

Theorem 4.5.10 *The following statements are equivalent.*

1. *The equations* (4.20) *have an initial solution for initial state* x_0.
2. *The equations* (4.20) *have an initial solution for initial state* x_0 *of Bohl type.*
3. *RCP*(x_0) *has a solution.*

Furthermore, there is a one-to-one correspondence between initial solutions to (4.20) of Bohl type and solutions to RCP(x_0). *More specifically,* $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ *is an initial solution to (4.20) of Bohl type if and only if its Laplace transform* $(\hat{\mathtt{u}}(s), \hat{\mathtt{x}}(s), \hat{\mathtt{y}}(s))$ *is such that* $(\hat{\mathtt{u}}(s), \hat{\mathtt{y}}(s))$ *is a solution to RCP*(x_0) *and*

$$\hat{\mathtt{x}}(s) = (s\mathbb{1} - A)^{-1}x_0 + (s\mathbb{1} - A)^{-1}B\hat{\mathtt{u}}(s). \tag{4.31}$$

The initial Bohl solution is smooth if and only if the corresponding solution to RCP(x_0) *is strictly proper.* □

The equivalence between 2 and 3 is proven in Theorem 3.5.2 together with the one-to-one correspondence between initial solutions of Bohl type with initial state x_0 and solutions to RCP(x_0) as described above. Evidently, statement 2 implies statement 1. The converse implication is far from trivial and will be a consequence of the proof of Theorem 4.5.14.

Of course, one may wonder whether a similar statement as in Theorem 4.5.10 can be made about uniqueness. The next example shows that this is not the case.

Example 4.5.11 Consider the complementarity system (4.20) with

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix};\ B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix};\ C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix};\ D = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The corresponding RCP(x_0) with $x_0 = (0, 0)^\top$ has a unique solution $u(s) = y(s) = (0, 0)^\top$ for all s. However, we can construct uncountably many different initial solutions (note that these cannot be Bohl due to the one-to-one correspondence between initial solutions of Bohl type and solutions to the RCP). For all $\tau > 0$ the functions $u_1(t) = f_\tau(t)$, $u_2(t) = -f_\tau(t)$ and $y_1(t) = y_2(t) = 0$ constitute an initial solution to (4.20) with initial state $(0, 0)^\top$, where $f_\tau(t)$ are the functions introduced in Example 4.5.4.
□

□

This example demonstrates that multiple initial solutions may exist in certain situations, although there is only one Bohl initial solution (or equivalently, only one solution to the corresponding RCP). However, we can introduce an equivalence relation on the space of impulsive-smooth distributions such that all initial solutions belong to the same equivalence class, in case there is only one initial solution of Bohl type.

We introduce the following notation. Consider the distributions $\mathtt{g} = \mathtt{g}_{imp} + \mathtt{g}_{reg} \in D_+'^k$, $\mathtt{h} = \mathtt{h}_{imp} + \mathtt{h}_{reg} \in D_+'^k$ with $\mathtt{g}_{imp}$, $\mathtt{h}_{imp}$ impulsive and $\mathtt{g}_{reg}$, $\mathtt{h}_{reg}$ piecewise continuous. These distributions could be called *impulsive-piecewise continuous*. For an $\varepsilon > 0$ we write

$$\mathtt{g} \mid_{(0,\varepsilon)} = \mathtt{h} \mid_{(0,\varepsilon)} \text{ if } \mathtt{g}_{reg} \mid_{[0,\varepsilon)} = \mathtt{h}_{reg} \mid_{[0,\varepsilon)} .$$

Similarly, we write

$$\mathtt{g} \mid_{[0,\varepsilon)} = \mathtt{h} \mid_{[0,\varepsilon)} \text{ if } \mathtt{g}_{reg} \mid_{(0,\varepsilon)} = \mathtt{h}_{reg} \mid_{(0,\varepsilon)} \text{ and } \mathtt{g}_{imp} = \mathtt{h}_{imp}.$$

Definition 4.5.12 Let $\mathtt{g}$, $\mathtt{h}$ be two C_{imp}^k-functions. We shall say that $\mathtt{g}$ is *equivalent* to $\mathtt{h}$, $\mathtt{g} \sim \mathtt{h}$, if and only if there exists an $\varepsilon > 0$ such that $\mathtt{g} \mid_{[0,\varepsilon)} = \mathtt{h} \mid_{[0,\varepsilon)}$. This is an equivalence relation and the equivalence classes are called *germs*. We say that two initial solutions $(\mathtt{u}^1, \mathtt{x}^1, \mathtt{y}^1)$, $(\mathtt{u}^2, \mathtt{x}^2, \mathtt{y}^2)$ *are in the same germ* or are *unique up to germ equivalence* if $\mathrm{col}(\mathtt{u}^1, \mathtt{x}^1, \mathtt{y}^1) \sim \mathrm{col}(\mathtt{u}^2, \mathtt{x}^2, \mathtt{y}^2)$. □

This definition extends an equivalence relation on C^∞-functions and the corresponding equivalence classes (also called germs) as used in differential geometry, see e.g. [21]. The following lemma states that the Bohl distributions can be embedded in the space of germs.

Lemma 4.5.13 *Each germ contains at most one Bohl distribution.* □

Proof. Bohl functions are real-analytic. Hence, $\mathtt{g} \mid_{[0,\varepsilon)} = \mathtt{h} \mid_{[0,\varepsilon)}$ implies $\mathtt{g} = \mathtt{h}$ for two Bohl distributions $\mathtt{g}$, $\mathtt{h}$. □

The set of Bohl distributions can be embedded (using the above lemma) in the set of germs in C_{imp}. However, not all germs contain a Bohl distribution as can be seen from the equivalence class containing $f_0(t)$ (defined in Example 4.5.4).

The uniqueness result that we are after is formulated as follows. The proof is given later in this section.

Theorem 4.5.14 *Let $E \in \mathbb{R}^{l\times k}$ be given. The following statements are equivalent.*

1. *The relation $E\mathrm{u}^1 \sim E\mathrm{u}^2$ holds for any pair of initial solutions $(\mathrm{u}^j, \mathrm{x}^j, \mathrm{y}^j)$, $j = 1, 2$ to (4.20) with initial state x_0.*
2. *The relation $Eu^1(s) \equiv Eu^2(s)$ holds for any pair of solutions $(u^j(s), y^j(s))$, $j = 1, 2$ to* RCP(x_0).

□

Remark 4.5.15 Consider a linear complementarity system (4.20) with parameters (A, B, C, D). Suppose that $\ker E \subseteq \ker B$. Then it is evident, that statement 1 in Theorem 4.5.14 implies that for any pair of initial solutions $(\mathrm{u}^j, \mathrm{x}^j, \mathrm{y}^j)$, $j = 1, 2$ to (4.20) with initial state x_0, also $\mathrm{x}^1 \sim \mathrm{x}^2$ is true. If in addition, $\ker E \subseteq \ker D$, then also $\mathrm{y}^1 \sim \mathrm{y}^2$ holds. □

An immediate corollary is the following (take E equal to the identity matrix).

Theorem 4.5.16 *All initial solutions to (4.20) with initial state x_0 are unique up to germ equivalence if and only if RCP(x_0) has a unique solution.* □

Remark 4.5.17 Returning to example 4.5.11, it is obvious that all the indicated initial solutions are contained in one germ with as a representative the initial solution of Bohl type (as stated in Theorem 4.5.16). □

One may wonder if each germ of initial solutions contains a Bohl initial solution. The above theorem implies that this is true (due to the one-to-one correspondence between Bohl initial solutions and solutions to RCP), when there is only one Bohl initial solution. However, the following counterexample shows that the collection of germs of initial solutions can not be identified by the Bohl initial solutions in general.

Example 4.5.18 Consider the complementarity system (4.20) with

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix};\ B = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix};\ C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix};\ D = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

For initial state $x_0 = (0, 0)^\top$ the function $u_1(t) = u_2(t) = f_0(t)$ (see Example 4.5.4), $y_1(t) = y_2(t) = 0$ is an initial solution. However, this function is not equivalent to a Bohl distribution as noted before. □

To prove Theorem 4.5.14 one technical result is needed. It is possible that the Laplace transform of an initial solution does not exist. The next lemma shows that the initial solution can be modified for large time-values such that the Laplace transform exists and satisfies the conditions of RCP except the rationality.

Lemma 4.5.19 *If there exists an initial solution* $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ *to (4.20) with initial state* x_0*, then there exists an impulsive-piecewise continuous distribution* $(\tilde{\mathtt{u}}, \tilde{\mathtt{x}}, \tilde{\mathtt{y}})$ *and an* $\epsilon > 0$ *such that*

1. $(\tilde{\mathtt{u}}, \tilde{\mathtt{x}}, \tilde{\mathtt{y}})$ *is Laplace transformable with Laplace transform* $(\hat{\mathtt{u}}(s), \hat{\mathtt{x}}(s), \hat{\mathtt{y}}(s))$*;*
2. $(\tilde{\mathtt{u}}, \tilde{\mathtt{x}}, \tilde{\mathtt{y}}) \mid_{[0,\epsilon)} = (\mathtt{u}, \mathtt{x}, \mathtt{y}) \mid_{[0,\epsilon)}$*;*
3. *The relations (with* $q(s)$ *and* $M(s)$ *as in (4.30))*

$$\hat{\mathtt{y}}(s) = q(s) + M(s)\hat{\mathtt{u}}(s) \text{ and } \hat{\mathtt{y}}(s) \perp \hat{\mathtt{u}}(s) \tag{4.32}$$

hold for all $s \in \mathbb{C}$ *and there exists a* $\sigma_0 \in \mathbb{R}$ *such that for all* $\sigma \geq \sigma_0$ *we have* $\hat{\mathtt{y}}(\sigma) \geq 0,\ \hat{\mathtt{u}}(\sigma) \geq 0$*.*

□

Proof. Let $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ be an initial solution to (4.20). For i such that $\mathtt{u}_{imp,i} = 0$ define $\tau_i^{\mathtt{u}} = \inf\{t > 0 \mid \mathtt{u}_{reg,i}(t) < 0\}$ and define $\tau_i^{\mathtt{y}}$ similarly if $\mathtt{y}_{imp,i} = 0$. Note that the defined $\tau_i^{\mathtt{u}}$ and $\tau_i^{\mathtt{y}}$ are strictly positive due to initial nonnegativity of $\mathtt{u}$ and $\mathtt{y}$. Take $\epsilon > 0$ such that ϵ is smaller than all defined $\tau_i^{\mathtt{u}}$ and $\tau_i^{\mathtt{y}}$.

We introduce the index sets J, K by

$$J := \{i \in \bar{k} \mid \mathtt{u}_i \mid_{[0,\epsilon]} = 0\} \quad K := \{i \in \bar{k} \mid \mathtt{y}_i \mid_{[0,\epsilon]} = 0\}.$$

We define $V := V(A, B_{\bullet J^c}, C_{K\bullet}, D_{K,J^c})$ (see Definition 4.5.6). It is clear that $\mathtt{x}_{reg}(t) \in V$ for $t \in (0, \epsilon)$ and hence $\mathtt{x}_{reg}(\epsilon) = \lim_{t \uparrow \epsilon} \mathtt{x}_{reg}(t) \in V$. We now take a feedback law F as in 2 of Lemma 4.5.7 making the subspace V invariant under the closed-loop dynamics $\dot{\xi} = (A + B_{\bullet J^c}F)\xi$ (note the discussion after Lemma 4.5.7).

We introduce a new distribution $\tilde{\mathtt{u}}$ by $\tilde{\mathtt{u}} = \mathtt{u}_{imp} + \tilde{\mathtt{u}}_{reg}$ (note that the impulsive part is unchanged) with

$$\tilde{\mathtt{u}}_{reg,j}(t) = \begin{cases} \mathtt{u}_{reg,j}(t), & t \in [0, \epsilon] \\ 0, & t > \epsilon \text{ and } j \in J \\ F_{j\bullet}\xi, & t > \epsilon \text{ and } j \in J^c, \end{cases}$$

where $\xi(t)$ is the solution to $\dot{\xi}(t) = (A + B_{\bullet J^c}F)\xi(t)$ with initial condition $\xi(\epsilon) = x(\epsilon)$. Note that $\xi(t)$ is a Bohl function.

The existence of the Laplace transforms denoted by $(\hat{\mathtt{u}}(s), \hat{\mathtt{x}}(s), \hat{\mathtt{y}}(s))$ is easily established, because $\tilde{\mathtt{u}}$ is at most exponentially increasing. Furthermore, the second statement in the formulation of the lemma follows by construction.

Taking $\tilde{\mathtt{y}}$ as the corresponding output of (4.20a)-(4.20b) with initial state x_0, it is obvious that the first part of (4.32) is satisfied for all s. That the second part of (4.32) holds for all s follows from the construction which is such that $\tilde{\mathtt{u}}_J = 0$ and $\tilde{\mathtt{y}}_K = 0$. Note that the union of the index sets J and K is equal to $\bar{k}$ because of the complementarity satisfied by the initial solution $(\mathtt{u}, \mathtt{x}, \mathtt{y})$. It is clear that for all i with

$\tilde{\mathrm{u}}_{imp,i} \neq 0$ the Laplace transform satisfies $\hat{\mathrm{u}}_i(\sigma) > 0$ for sufficiently large $\sigma \in \mathbb{R}$. Indeed, the impulsive part $\tilde{\mathrm{u}}_{imp,i}$ is equal to $\mathrm{u}_{imp,i}$, which has a positive leading coefficient. In case $\tilde{\mathrm{u}}_{imp,i} = 0$ the definitions of ϵ and J imply that for all $i \in J^c$ $\mathrm{u}_{reg,i}(t) \geq 0$ for all $t \in [0, \epsilon]$ and there exists a nonempty interval $(t_b, t_f) \subseteq [0, \epsilon]$ such that $\mathrm{u}_{reg,i}(t) > 0$ for $t \in (t_b, t_f)$. Applying statement 3 of Lemma 4.5.3 yields $\hat{\mathrm{u}}(\sigma) \geq 0$ for all σ sufficiently large. Similar remarks can be made for $\hat{\mathrm{y}}(s)$. □

Proof of Theorem 4.5.14 If RCP(x_0) has multiple solutions $(u^j(s), y^j(s))$, $j = 1, 2$ with $Eu^1(s) \not\equiv Eu^2(s)$, the inverse Laplace transforms result in initial solutions $(\mathrm{u}^j, \mathrm{x}^j, \mathrm{y}^j)$, $j = 1, 2$ of Bohl type such that $E\mathrm{u}^1$ and $E\mathrm{u}^2$ are different. According to Lemma 4.5.13 this implies that $E\mathrm{u}^1$ and $E\mathrm{u}^2$ are contained in different germs.

Suppose there exist initial solutions $(\mathrm{u}^1, \mathrm{x}^1, \mathrm{y}^1)$, $(\mathrm{u}^2, \mathrm{x}^2, \mathrm{y}^2)$ with $E\mathrm{u}^1$ and $E\mathrm{u}^2$ in different germs. According to the previous lemma there exist an $\epsilon > 0$ and impulsive-piecewise continuous distributions $(\tilde{\mathrm{u}}^j, \tilde{\mathrm{x}}^j, \tilde{\mathrm{y}}^j)$, $j = 1, 2$ satisfying the conditions $1 - 3$ of Lemma 4.5.19 with respect to $(\mathrm{u}^j, \mathrm{x}^j, \mathrm{y}^j)$.

Two cases can be distinguished: either $E\mathrm{u}^1_{imp} \neq E\mathrm{u}^2_{imp}$ or $E\mathrm{u}^1_{imp} = E\mathrm{u}^2_{imp}$ and $E\mathrm{u}^1_{reg}(t) \neq E\mathrm{u}^2_{reg}(t)$ for some $t \in (0, \epsilon)$. In the latter case the continuity of both functions implies that $E\mathrm{u}^1_{reg}(t) \neq E\mathrm{u}^2_{reg}(t)$ for all $t \in (t_b, t_f) \subseteq [0, \epsilon]$ for certain $t_b \neq t_f$. Hence, the same holds for the related impulsive-piecewise continuous distributions $\tilde{\mathrm{u}}^1$ and $\tilde{\mathrm{u}}^2$. It is clear that the Laplace transforms of these impulsive-piecewise continuous distributions, denoted by $(\hat{\mathrm{u}}^j(s), \hat{\mathrm{x}}^j(s), \hat{\mathrm{y}}^j(s))$, $j = 1, 2$ are not rational in general and thus $(\hat{\mathrm{u}}^j(s), \hat{\mathrm{y}}^j(s))$ do not form solutions to RCP(x_0). However, since $(\hat{\mathrm{u}}^j(s), \hat{\mathrm{y}}^j(s))$, $j = 1, 2$ satisfy (4.8) for all s and (4.9) for all $\sigma \geq \sigma_0$, $(\hat{\mathrm{u}}^j(\sigma), \hat{\mathrm{y}}^j(\sigma))$, $j = 1, 2$ satisfy LCP($q(\sigma), M(\sigma)$) with $q(s)$ and $M(s)$ as in (4.30).

We intend to invoke Theorem 4.4.9 to find multiple solutions $(u^j(s), y^j(s))$, $j = 1, 2$ to RCP(x_0). Suppose that the conditions of this theorem are not satisfied, i.e. assume that there exists an $\sigma_0 \in \mathbb{R}$ such that for all $\sigma \geq \sigma_0$

$$E\hat{\mathrm{u}}^1(\sigma) - E\hat{\mathrm{u}}^2(\sigma) = 0. \tag{4.33}$$

We reconsider the two cases above. In the first case we have $E\tilde{\mathrm{u}}^1_{imp} \neq E\tilde{\mathrm{u}}^2_{imp}$. It is clear that this contradicts (4.33). Similarly, in the second case (i.e. $E\tilde{\mathrm{u}}^1_{imp} = E\tilde{\mathrm{u}}^2_{imp}$) (4.33) becomes

$$\int_0^\infty [E\tilde{\mathrm{u}}^1_{reg}(t) - E\tilde{\mathrm{u}}^2_{reg}(t)]e^{-\sigma t}dt = 0$$

for all $\sigma \geq \sigma_0$. Since in the second case the regular parts differ on the interval (t_b, t_f), the above equation cannot hold for all $\sigma \geq \sigma_0$. Hence, the conditions of Theorem 4.4.9 are satisfied and multiple solutions $(u^j(s), y^j(s))$, $j = 1, 2$ to RCP(x_0) with $Eu^1(s) \not\equiv Eu^2(s)$ do exist. □

Remark 4.5.20 The proof of Theorem 4.5.10 can easily be derived from the proof above. Similarly, we can construct a solution to LCP($q(\sigma), M(\sigma)$) for all sufficiently

large σ by taking the Laplace transform of the corresponding impulsive-piecewise continuous distribution satisfying the conditions of Lemma 4.5.19. Instead of invoking Theorem 4.4.9, one has to use Theorem 4.4.1 to prove the relation between existence of initial solutions and the existence of solutions to the corresponding RCP. □

The following corollary shows how the equivalence relation for initial solutions can be used to establish 'global' uniqueness of the global solution. The proof is based on the fact that only the 'nonnegative part' of the initial solution returns in the global solution.

Theorem 4.5.21 *Let $T_e > 0$ and $E \in \mathbb{R}^{l\times k}$ be such that* $\ker E \subseteq \ker B$ *with B as in (4.20). Suppose that $Eu^1(s) \equiv Eu^2(s)$ for all initial states x_0 and any pair of solutions $(u^j(s), y^j(s))$, $j = 1, 2$ to RCP(x_0). Then each pair of global solutions $(\mathcal{D}^j, \tau^j, x_e^j, u_c^j(t), x_c^j(t), y_c^j(t))$, $j = 1, 2$ on $[0, T_e)$ to (4.20) for equal initial state satisfies $Eu_c^1(t) = Eu_c^2(t)$ and $x_c^1(t) = x_c^2(t)$ for all $t \in [0, T_e)$ with $t \notin \tau^1(\mathcal{D}^1) \cup \tau^2(\mathcal{D}^2)$. If in addition,* $\ker E \subseteq \ker D$ *with D as in (4.20), then $y_c^1(t) = y_c^2(t)$ for all $t \in [0, T_e)$ with $t \notin \tau^1(\mathcal{D}^1) \cup \tau^2(\mathcal{D}^2)$.* □

The relevance of the assumption $\ker E \subseteq \ker B$ is mentioned in Remark 4.5.15 and will also become clear in the proof. Situations in which $\ker B$ is nontrivial occur for instance in the mechanical systems treated in the next section.

Proof. The proof is based on the following observations. According to the hypothesis of the theorem, Theorem 4.5.14 and Remark 4.5.15, we must have that any pair of initial solutions $(\mathrm{u}^j, \mathrm{x}^j, \mathrm{y}^j)$, $j = 1, 2$ with the same initial state, satisfies $E\mathrm{u}^1 \sim E\mathrm{u}^2$ and $\mathrm{x}^1 \sim \mathrm{x}^2$. This will be called the *similarity property* in the proof. Secondly, note that for a global solution as in Definition 4.5.9, $(u_c(t + \tilde{t}), x_c(t + \tilde{t}), y_c(t + \tilde{t}))$ for some $\tilde{t} \notin \tau(\mathcal{D})$ is equal to a smooth initial solution with initial state $x_c(\tilde{t})$ on a closed interval of positive length with left end-point zero.

Define

$$t^* := \inf\{t \in [0, T_e) \setminus (\tau^1(\mathcal{D}^1) \cup \tau^2(\mathcal{D}^2)) \mid Eu_c^1(t) \neq Eu_c^2(t) \text{ or } x_c^1(t) \neq x_c^2(t)\}$$

with the convention $\inf \emptyset = \infty$. In case $t^* = \infty$, we are finished, because then the claim of the theorem is true. Hence, suppose $t^* < \infty$. Without loss of generality we may assume that no void re-initializations occur meaning that $\tau(i) = \tau(i + 1)$ and $x_e(i) = x_e(i + 1)$. It is clear that in these cases $\tau(i + 1)$ can be removed from the set of event times without essentially changing the global solution.

We can distinguish three cases.

1. $t^* \in \tau^1(\mathcal{D}^1) \cap \tau^2(\mathcal{D}^2)$. Let $j^i_{\min}$ and $j^i_{\max}$ be the minimal and maximal integer j in $\mathcal{D}^i$, respectively, such that $\tau^i(j) = t^*$ for $i = 1, 2$. In case $t^* = 0$, it is clear that $x_e^1(j^1_{\min}) = x_e^2(j^2_{\min})$. If $t^* > 0$, Definition 4.5.9 (item 5.) and the definition of t^* imply that $x_e^1(j^1_{\min}) = \lim_{t\uparrow t^*} x_c^1(t) = \lim_{t\uparrow t^*} x_c^2(t) = x_e^2(j^2_{\min})$. The definition of re-initializations (item 5.) and the similarity property yield by induction that $x_e^1(j^1_{\min}+r) = x_e^2(j^2_{\min}+r)$ for all $0 \leq r \leq \min(j^1_{\max} - j^1_{\min}, j^2_{\max} -$

$j^2_{\min}$). Since no void re-initializations occur, the similarity property implies that $j^1_{\max} - j^1_{\min} = j^2_{\max} - j^2_{\min}$. Hence, for both global solutions we have that $\tau^i(j^i_{\max}+1) > \tau^i(j^i_{\max}) = t^*$ with the same initial state $x^1_e(j^1_{\max}) = x^2_e(j^2_{\max})$. Recall the way that $(u^i_c(t), x^i_c(t), y^i_c(t))$ is defined on $(\tau^i(j^i_{\max}), \tau^i(j^i_{\max}+1))$ as a piece of an initial solution (see item 2. of Definition 4.5.9). According to the similarity property, it is then clear that

$$Eu^1_c(t) = Eu^2_c(t) \text{ and } x^1_c(t) = x^2_c(t)$$

for all $t \in [t^*, t^*+\varepsilon)$ for some $\varepsilon > 0$. This contradicts the definition of t^*.

2. $t^* \in \tau^1(\mathcal{D}^1) \setminus \tau^2(\mathcal{D}^2)$ (or $t^* \in \tau^2(\mathcal{D}^2) \setminus \tau^1(\mathcal{D}^1)$). Note that $t^* > 0$, because 0 is always an event time. Let j be the smallest integer in $\mathcal{D}^1$ such that $\tau^1(j) = t^*$. According to Definition 4.5.9, $x^1_e(j) = \lim_{t\uparrow\tau^1(j)} x^1_c(t) = \lim_{t\uparrow t^*} x^2_c(t) = x^2_c(t^*)$. Since $t^* \notin \tau^2(\mathcal{D}^2)$, $(u^2_c(t+t^*), x^2_c(t+t^*), y^2_c(t+t^*))$ is equal to a smooth initial solution with initial state $x^1_e(j)$ on a closed interval of positive length with left end-point equal to zero. The similarity property implies that the state of any other initial solution from $x^1_e(j)$ must be equivalent to the state of this smooth one. This implies that $\tau^1(j+1) > \tau^1(j)$, because otherwise a void re-initialization would take place. Due to (again) the similarity property,

$$Eu^1_c(t) = Eu^2_c(t) \text{ and } x^1_c(t) = x^2_c(t)$$

for all $t \in [t^*, t^*+\varepsilon)$ for some $\varepsilon > 0$. This contradicts the definition of t^*.

3. $t^* \notin \tau^1(\mathcal{D}^1) \cup \tau^2(\mathcal{D}^2)$. Note that $t^* > 0$. Both $(u^j_c(t+t^*), x^j_c(t+t^*), y^j_c(t+t^*))$, $j = 1, 2$ are equal to smooth initial solutions with the same initial state $x^1_c(t^*) = x^2_c(t^*)$ on a closed interval with positive length and left end-point zero. The similarity property guarantees

$$Eu^1_c(t) = Eu^2_c(t) \text{ and } x^1_c(t) = x^2_c(t)$$

for all $t \in [t^*, t^*+\varepsilon)$ for some $\varepsilon > 0$. This contradicts the definition of t^*.

Hence, $t^* = \infty$ and thus the proof is complete.

The case in which additionally $\ker E \subseteq \ker D$ holds can be proven analogously. The similarity property includes then also $\mathrm{y}^1 \sim \mathrm{y}^2$ as in Remark 4.5.15. □

Particular choices of E lead to uniqueness of the complete global solution or the state trajectory of the global solution.

Definition 4.5.22 We say that (4.20) has the *unique flow part property*, if every pair of global solutions $(\mathcal{D}^j, \tau^j, x^j_e, u^j_c(t), x^j_c(t), y^j_c(t))$, $j = 1, 2$ to (4.20) on an arbitrary time-interval $[0, T_e)$ with arbitrary initial state x_0 satisfies $u^1_c(t) = u^2_c(t)$, $x^1_c(t) = x^2_c(t)$ and $y^1_c(t) = y^2_c(t)$ for all $t \in [0, T_e)$ with $t \notin \tau^1(\mathcal{D}^1) \cup \tau^2(\mathcal{D}^2)$.

We say that (4.20) has the *unique state part property*, if any pair of global solutions $(\mathcal{D}^j, \tau^j, x_e^j, u_c^j(t), x_c^j(t), y_c^j(t))$, $j = 1, 2$ to (4.20) on an arbitrary interval $[0, T_e)$ with arbitrary initial state x_0 satisfies $x_c^1(t) = x_c^2(t)$ for all $t \in [0, T_e)$ with $t \notin \tau^1(\mathcal{D}^1) \cup \tau^2(\mathcal{D}^2)$. □

Corollary 4.5.23 *Consider a linear complementarity system given by the quadruple* (A, B, C, D).

- *Suppose that* $Bu^1(s) \equiv Bu^2(s)$ *is true for any pair of solutions* $(u^j(s), y^j(s))$, $j = 1, 2$ *to RCP*(x_0) *for all initial states* x_0. *Then the LCS (4.20) has the unique state part property.*
- *Suppose that* $u^1(s) \equiv u^2(s)$ *is true for any pair of solutions* $(u^j(s), y^j(s))$, $j = 1, 2$ *to RCP*(x_0) *for all initial states* x_0. *Then the LCS (4.20) has the unique flow part property.*

□

4.6 Well-posedness results

By combining the results of the sections 4.4 and 4.5, existence and uniqueness of initial solutions can be related to solvability properties of parameterized sets of LCPs. This will now be exploited to obtain well-posedness results for linear mechanical systems subject to unilateral constraints, linear relay systems and electrical networks containing ideal diodes. Establishing (unique) solvability of the LCPs can be a nontrivial task in certain situations, as we will see.

4.6.1 Well-posedness results of linear mechanical systems

We consider linear mechanical systems given by

$$M\ddot{q} + D\dot{q} + Kq = 0, \tag{4.34}$$

where q denotes the vector of generalized coordinates. Moreover, M denotes the generalized mass matrix (or inertia matrix), which is assumed to be positive definite, D denotes the damping matrix and K the stiffness matrix. Suppose now that the system is subject to frictionless unilateral constraints given by

$$Fq \geq 0 \tag{4.35}$$

with F some matrix of appropriate dimensions. Furthermore, we assume that impacts are purely inelastic. Then (4.34) is replaced by

$$M\ddot{q} + D\dot{q} + Kq = F^\top u \tag{4.36}$$

together with complementarity conditions on u and Fq. $F^\top u$ are the constraint forces and u are the multipliers corresponding to the unilateral constraints. This formulation can be cast into a linear complementarity system by introducing the state vector $x = \mathrm{col}(q, \dot{q})$ resulting in

$$\dot{x} = \underbrace{\begin{pmatrix} 0 & I \\ -M^{-1}K & -M^{-1}D \end{pmatrix}}_{A} x + \underbrace{\begin{pmatrix} 0 \\ M^{-1}F^\top \end{pmatrix}}_{B} u \tag{4.37a}$$

$$y = \underbrace{(F \;\; 0)}_{C} x \tag{4.37b}$$

together with the complementarity conditions (4.20c) on the reaction force u and the displacement y. Note that the B-matrix has full column rank if and only if F has full row rank; hence, if the unilateral constraints are dependent, $\ker B$ is nontrivial. This is for instance the case if an equality constraint is described by two inequalities in (4.35). Note that such a dependence was taken into account in Theorem 4.5.21.

Of course, the linear setting chosen here is quite restrictive in comparison with recent advances in the field of nonlinear mechanical systems with inequality constraints [72, 131, 192]. In fact, results as in Theorem 4.6.6 below were proven already in [124, 142] for nonlinear mechanical systems by differentiation of the relevant system's variables. The purpose of this section is merely an illustration of the general theory developed in this chapter. We will show that Theorem 4.6.6 can be obtained quite easily by using the RCP.

RCP(x_0) for a linear mechanical system as above is equal to RCP($q(s)$, $M(s)$) with

$$M(s) := C(s\mathbb{1} - A)^{-1}B = F(s^2M + sD + K)^{-1}F^\top \tag{4.38a}$$

$$q(s) := C(s\mathbb{1} - A)^{-1}x_0 = F(s^2M + sD + K)^{-1}[(sM + D)q_0 + M\dot{q}_0] \tag{4.38b}$$

with $\mathrm{col}(q_0, \dot{q}_0) = x_0$. To prove solvability of the corresponding LCP($q(\sigma)$, $M(\sigma)$) for sufficiently large $\sigma \in \mathbb{R}$, we use the following lemma from [46].

Lemma 4.6.1 [46] *If $G = NPN^\top$ for some positive definite (not necessarily symmetric) matrix P and some matrix N and $c \in \mathrm{Im}\, G$, then the problem*

$$y = c + Gu,\; 0 \le y \perp u \ge 0$$

has solutions. If (u^1, y^1) and (u^2, y^2) are two solutions, then $y^1 = y^2$ and $Gu_1 = Gu_2$. □

We also need the following.

Lemma 4.6.2 *Let $P \in \mathbb{R}^{k\times k}$ and $N \in \mathbb{R}^{l\times k}$ be matrices with P positive definite (but not necessarily symmetric). Then the following holds:*

$$\begin{aligned} \ker NPN^\top &= \ker N^\top \\ \operatorname{Im} NPN^\top &= \operatorname{Im} N = \operatorname{Im} NP. \end{aligned}$$

□

Proof. If $NPN^\top v = 0$, then $v^\top NPN^\top v = 0$ is true implying that $N^\top v = 0$. This proves the first identity above, because the converse is trivial. The second statement follows by duality. □

Remark 4.6.3 Note that all matrices $G = NPN^\top$ for some matrix N and some positive definite matrix P are nonnegative definite (but not necessarily symmetric). However, the converse statement that all nonnegative matrices can be written in the above form, is not true. A counterexample is provided by $G = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$. Indeed, if $G = NPN^\top$, then Lemma 4.6.2 implies that $\ker N^\top = \ker G = \{0\}$. However, for $v = \operatorname{col}(1, 1)$ we have $v^\top Gv = 0$ and hence (see proof of Lemma 4.6.2) $N^\top v$ is equal to 0, which contradicts the triviality of the kernel of $N^\top$. □

□

Theorem 4.6.4 *RCP($q(s)$, $M(s)$) with $q(s)$ and $M(s)$ as in (4.38) (or equivalently RCP(x_0) with the matrices A, B, C as in (4.37)) has for each x_0 a solution.* □

Proof. Obviously, we have for sufficiently large σ that $(\sigma^2 M + \sigma D + K)^{-1}$ is positive definite, because M is positive definite. According to Lemma 4.6.1 and Theorem 4.4.1, left to prove is that for sufficiently large σ, $q(\sigma)$ as in (4.38b) belongs to $\operatorname{Im} M(\sigma)$. However, this is immediate from Lemma 4.6.2, because

$$q(\sigma) \in \operatorname{Im} F(\sigma^2 M + \sigma D + K)^{-1} = \operatorname{Im} F(\sigma^2 M + \sigma D + K)^{-1} F^\top = \operatorname{Im} M(\sigma) \tag{4.39}$$

for sufficiently large σ. □

Theorem 4.6.5 *Consider a linear mechanical system of the form (4.37) with initial state x_0. The corresponding RCP(x_0) may have multiple solutions, say $(u^1(s), y^1(s))$ and $(u^2(s), y^2(s))$. However, these solutions satisfy $Bu^1(s) \equiv Bu^2(s)$.* □

Proof. Take σ_0 such that $R(\sigma) := (\sigma^2 M + \sigma D + K)^{-1}$ is positive definite for all $\sigma \geq \sigma_0$. Suppose that there exist two solutions (u^i, y^i), $i = 1, 2$ to LCP($q(\sigma)$, $M(\sigma)$) for some $\sigma \geq \sigma_0$. According to Lemma 4.6.1, we have

$$M(\sigma)u^1 = M(\sigma)u^2(s). \tag{4.40}$$

Lemma 4.6.2 states that $\ker M(\sigma) = \ker F R(\sigma) F^\top = \ker F^\top$ holds for all $\sigma \geq \sigma_0$. Hence, $F^\top(u^1 - u^2) = 0$. The form of the matrix B as in (4.37) now implies that $Bu^1 = Bu^2$. Invoking Theorem 4.4.9 completes the proof. □

For linear mechanical systems the following well-posedness result follows from Theorem 4.6.4, Theorem 4.6.5, Theorem 4.5.10 and Corollary 4.5.23.

Theorem 4.6.6 *Consider a constrained mechanical system given by (4.37) and (4.20c). For each initial state x_0 there exists an initial solution. Furthermore, the constrained mechanical system has the unique state part property (as defined in Definition 4.5.22).* □

For the case of *independent* unilateral constraints (i.e. F has full row rank), it has already been proven in Chapter3, that after at most one nonsmooth initial solution, a smooth initial solution occurs, i.e. for each initial state there exists an $\varepsilon > 0$ such that a solution in the sense of Definition 4.5.9 exists on $[0, \varepsilon)$ with $\tau(1) > \tau(0)$ or $\tau(2) > \tau(1) = \tau(0)$. It is also shown that the initial solutions with possible jumps agree with the jump rules as proposed by Moreau in the case of inelastic collisions [139,144].

4.6.2 Well-posedness of linear relay systems

In this subsection, we consider a system given by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{4.41a}$$
$$y(t) = Cx(t) + Du(t) \tag{4.41b}$$

with $u(t) \in \mathbb{R}^k$, $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^k$ and A, B, C, D are matrices of appropriate dimensions. Each pair (u_i, y_i) is connected by an ideal relay (or Coulomb friction characteristic) with a relation as given in figure 4.1 (note the minus sign in front of u_i). The vectors d_1 and $d_2 \in \mathbb{R}^k$ in this figure are constant vectors with

$$d_1 \geq 0,\ d_2 \geq 0,\ d_1 + d_2 > 0. \tag{4.42}$$

Several approaches are known that cast the relay/Coulomb friction characteristic into a complementarity description by introducing several auxiliary variables, see e.g. [112, 123, 160]. In [123] a corresponding rational complementarity problem RCP($q(s)$, $M(s)$) has been formulated with

$$M(s) = \begin{pmatrix} G^{-1}(s) & -G^{-1}(s) \\ -G^{-1}(s) & G^{-1}(s) \end{pmatrix} \tag{4.43a}$$
$$q(s) = \begin{pmatrix} -G^{-1}(s)T(s)x_0 + \frac{1}{s}d_1 \\ G^{-1}(s)T(s)x_0 + \frac{1}{s}d_2 \end{pmatrix}, \tag{4.43b}$$

where x_0 is the initial condition of (4.41) and

$$T(s) := C(s\mathbb{1} - A)^{-1}$$
$$G(s) := C(s\mathbb{1} - A)^{-1}B + D.$$

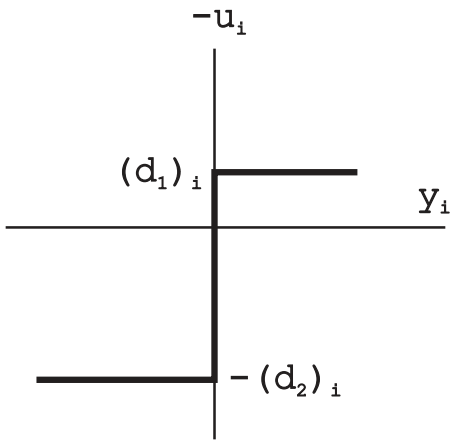

Figure 4.1: The i-th relay characteristic.

We assume that $G(s)$ is invertible as a rational matrix. Similarly as for a standard LCS, the RCP($q(s)$, $M(s)$) has a solution if and only if the system (4.41) with initial condition x_0 has an initial solution. All initial solutions corresponding to the same initial state are unique up to germ equivalence if and only if this RCP admits at most one solution.

We consider an LCP(q, M) with M and q of the following structure.

$$M = \begin{pmatrix} G^{-1} & -G^{-1} \\ -G^{-1} & G^{-1} \end{pmatrix} \tag{4.44a}$$

$$q = \begin{pmatrix} -G^{-1}v + d_1 \\ G^{-1}v + d_2 \end{pmatrix}, \tag{4.44b}$$

where G is an invertible matrix, v is some vector and d_1 and d_2 are vectors satisfying (4.42).

The assumptions in the following theorem do not require M to be a P-matrix. According to Theorem 4.3.3 this implies that LCP(q, M) does not have a unique solution for all arbitrary vectors q. In [123] the special structure of q and M in (4.44) is exploited to prove the following result.

Theorem 4.6.7 [123] *If G is a P-matrix, then the LCP(q, M) with q and M as in (4.44) has a unique solution for each x_0 and each d_1, d_2 satisfying (4.42).* □

As a corollary of the theorems 4.4.1 and 4.4.9, we get the following statement.

Lemma 4.6.8 *If there exists a $\sigma_0 \in \mathbb{R}$ such that $G(\sigma)$ is a P-matrix for all $\sigma \geq \sigma_0$, then RCP($q(s)$, $M(s)$) with $q(s)$ and $M(s)$ as in (4.43) has a unique solution for all x_0.* □

As a consequence of Theorem 4.5.10, Theorem 4.5.16, and Corollary 4.5.23, we get the main result of this subsection.

Theorem 4.6.9 *Consider the linear relay system given by (4.41) and k ideal relay characteristics. If $G(\sigma) := C(\sigma \mathbb{1} - A)^{-1}B + D$ of (4.41) is a P-matrix for all $\sigma \geq \sigma_0$ for some $\sigma_0 \in \mathbb{R}$, then for all x_0 there exist initial solutions of the relay system (4.41) with initial state x_0, and all these initial solutions are unique up to germ equivalence. Furthermore, the linear relay system has the unique flow part property (as defined in Definition 4.5.22).* □

In [123], it has been shown that all initial solutions are regular distributions and hence the state trajectory $x_c(t)$ of the global solution as in Definition 4.5.9 is continuous in the sense that $\lim_{t\uparrow\tau(i)} x_c(t) = \lim_{t\downarrow\tau(i)} x_c(t)$. Between event times $x_c(t)$ is even smooth.

4.6.3 Well-posedness of dissipative systems with complementarity conditions

Let us consider a linear complementarity system (4.20), in which the dynamical system given by (4.20a)-(4.20b) is *dissipative* in the following sense.

Definition 4.6.10 [206] The system (A, B, C, D) given by (4.20a)-(4.20b) with supply rate $u^\top y$ is said to be *dissipative*, if there exists a nonnegative function $S : \mathbb{R}^n \to \mathbb{R}_+$ such that for all $t_0 \leq t_1$, and all locally square integrable functions $(u(t), x(t), y(t))$ from $\mathbb{R}$ to $\mathbb{R}^{k+n+k}$ satisfying (4.20a)-(4.20b) the inequality

$$S(x(t_0)) + \int_{t_0}^{t_1} u^\top(t)y(t)dt \geq S(x(t_1))$$

holds. A function S satisfying the conditions above is called a *storage function*. □

The above inequality is called the *dissipation inequality*. We shall also use the assumption of minimality of the system description, which is standard in the literature on dissipative dynamical systems, see e.g. [206]. The triple (A, B, C) in (4.20a)-(4.20b) is called *minimal*, if it is controllable and observable. In algebraic terms this means that

$$\operatorname{rank}(B\ AB\ \ldots\ A^{n-1}B) = n \text{ and } \operatorname{rank}(C^\top\ C^\top A^\top\ \ldots\ C^\top(A^\top)^{n-1}) = n. \quad (4.45)$$

We state the following results from [206].

Theorem 4.6.11 [206] *Consider the system (A, B, C, D) as in (4.20a)-(4.20b) and assume that (A, B, C) is minimal. Then (A, B, C, D) is dissipative with respect to the supply rate $u^\top y$ if and only if the transfer matrix $M(s) := C(s\mathbb{1} - A)^{-1}B + D$ is* positive real, *i.e. the poles of the entries of $M(s)$ have nonpositive real parts and $M(s) + M^*(s) \geq 0$ for all s with $\operatorname{Re} s > 0$.* □

Theorem 4.6.12 [206] *Consider the system (A, B, C, D) as in (4.20a)-(4.20b) and assume that (A, B, C) is minimal. The system is dissipative with respect to the supply rate $u^\top y$ if and only if there exists a symmetric positive definite matrix K such that $S(x) = x^\top K x$ defines a storage function.* □

Now we are in a position to prove the main result of this subsection.

Theorem 4.6.13 *If the linear complementarity system given by (4.20) is such that (A, B, C, D) is dissipative with respect to the supply rate $u^\top y$ and the triple (A, B, C) is minimal, then the corresponding RCP(x_0) has for each x_0 a solution. RCP(x_0) may have multiple solutions. However, we have $Bu^1(s) \equiv Bu^2(s)$ for all pairs of solutions $(u^j(s), y^j(s))$, $j = 1, 2$ to RCP(x_0).* □

Proof. Since $M(s)$ is positive real, $M(\sigma)$ is positive semi-definite for each nonnegative real σ. According to [47, Thm. 3.1.2] this implies that if the LCP($C(\sigma \mathbb{1} - A)^{-1}x_0$, $M(\sigma)$) is feasible (see Section 4.3) for a definition), then it is solvable. So, if we can show that for all $\sigma > 0$ LCP($C(\sigma \mathbb{1} - A)^{-1}x_0$, $M(\sigma)$) is feasible, then we proved according to Theorem 4.4.1 that RCP(x_0) has a solution.

Suppose that there exists a $\sigma > 0$ such that LCP($C(\sigma \mathbb{1} - A)^{-1}x_0$, $M(\sigma)$) is not feasible. This means that the set of inequalities $y = C(\sigma \mathbb{1} - A)^{-1}x_0 + M(\sigma)u \geq 0$, $u \geq 0$ does not have a solution $y \in \mathbb{R}^k$, $u \in \mathbb{R}^k$. Rewriting this in the standard form used in Farkas' lemma [135] yields that

$$(-M(\sigma) \quad \mathbb{1}) \begin{pmatrix} u \\ y \end{pmatrix} = C(\sigma \mathbb{1} - A)^{-1}x_0, \quad \begin{pmatrix} u \\ y \end{pmatrix} \geq 0$$

does not have a solution. Then, Farkas' lemma [135] implies that there exists a vector u_0 such that

$$\begin{aligned} 0 &\leq u_0; && (4.46)\\ 0 &\geq u_0^\top M(\sigma); && (4.47)\\ 0 &> u_0^\top C(\sigma \mathbb{1} - A)^{-1}x_0. && (4.48) \end{aligned}$$

Observe that the following trajectories

$$\begin{aligned} u(t) &= u_0 e^{\sigma t} && (4.49)\\ x(t) &= (\sigma \mathbb{1} - A^\top)^{-1}C^\top u_0 e^{\sigma t} && (4.50)\\ y(t) &= M^\top(\sigma) u_0 e^{\sigma t} && (4.51) \end{aligned}$$

form a solution of

$$\begin{aligned} \dot{x}(t) &= A^\top x(t) + C^\top u(t)\\ y(t) &= B^\top x(t) + D^\top u(t). \end{aligned}$$

Note that the system with parameters $(A^\top, C^\top, B^\top, D^\top)$ results in the transfer matrix $M^\top(s)$. Furthermore, note that $(A^\top, C^\top, B^\top)$ is minimal, because (A, B, C)

is minimal and that $M^\top(s)$ is positive real, because $M(s)$ is positive real. Hence, the system $(A^\top, C^\top, B^\top, D^\top)$ is dissipative according to Theorem 4.6.11.

Substituting (4.49)-(4.51) in the dissipation inequality of $(A^\top, C^\top, B^\top, D^\top)$, we get for $t_0 < t_1$

$$S(x(t_0)) + \int_{t_0}^{t_1} u_0^\top M^\top(\sigma) u_0 e^{2\sigma t} dt \geq S(x(t_1)), \tag{4.52}$$

where we take $S(x) = x^\top K x$ as a storage function for $(A^\top, C^\top, B^\top, D^\top)$ with K symmetric and positive definite as in Theorem 4.6.12. Note that $u_0^\top M^\top(\sigma) u_0 = 0$ due to the fact that $M^\top(\sigma)$ is positive semi-definite and (4.46)-(4.47). Hence, the integral in (4.52) is zero resulting in $0 \leq S(x(t_1)) \leq S(x(t_0))$. Since $\lim_{t_0 \to -\infty} x(t_0) = 0$ (see (4.50) and recall that $\sigma > 0$), we get $x^\top(t_1) K x(t_1) = S(x(t_1)) = 0$ for all $t_1 \in \mathbb{R}$. But this means that $x(t_1) = 0$ for all $t_1 \in \mathbb{R}$, because K is positive definite. Since $(\sigma \mathbb{1} - A^\top)$ is invertible for every $\sigma > 0$, (4.50) implies $C^\top u_0 = 0$ which contradicts (4.48). This proves the existence part of the theorem.

To prove the uniqueness part, we use similar reasoning as for the existence part. Suppose LCP$(C(\sigma \mathbb{1} - A)^{-1} x_0, M(\sigma))$ has for some $\sigma > 0$ multiple solutions (u^1, y^1) and (u^2, y^2). According to [47, Thm.3.1.7], then we must have that $[M^\top(\sigma) + M(\sigma)](u^1 - u^2) = 0$. Observing that $u(t) = e^{\sigma t}(u^1 - u^2)$, $x(t) = (\sigma \mathbb{1} - A)^{-1} B(u^1 - u^2) e^{\sigma t}$, $y(t) = M(\sigma)(u^1 - u^2) e^{\sigma t}$ are trajectories of the system (A, B, C, D), we can conclude analogously as above by using the dissipation inequality for (A, B, C, D) that $B(u^1 - u^2) = 0$. According to Theorem 4.4.9 this implies that any pair of solutions to RCP(x_0) $(u^j(s), y^j(s))$, $j = 1, 2$ satisfies $Bu^1(s) \equiv Bu^2(s)$. □

The main theorem of this subsection is now a consequence of Theorem 4.5.10 and Theorem 4.5.23.

Theorem 4.6.14 *A linear complementarity system (4.20) with (A, B, C, D) dissipative with respect to the supply rate $u^\top y$ and (A, B, C) minimal, has for each initial state x_0 an initial solution. Moreover the corresponding LCS has the unique state part property (as defined in Definition 4.5.22).* □

An example of a linear complementarity system with (A, B, C, D) dissipative with respect to the supply rate $u^\top y$ is a linear electrical network consisting of resistors, capacitors, inductors, gyrators, transformers and k ideal diodes. To model such a network as a complementarity system, we first extract the diodes and replace them by ports with two terminals. Associated with these two terminals are two variables: the current entering one terminal and leaving the other and the voltage across these terminals. The resulting multiport network can be described by a state space representation (A, B, C, D) [3] with input/output (u/y) variables representing the port variables. For the i-th port, we have that either u_i is the current entering the port and y_i the voltage across the port or vice versa. To include the ideal diodes in the electrical network, we

add the ideal diode characteristics to the port variables. These are (with a sign change with respect to the usual conventions in circuit theory)

$$0 \leq y(t) \perp u(t) \geq 0. \tag{4.53}$$

Together with the (A, B, C, D)-system this constitutes an example of the systems considered in this subsection.

4.7 Conclusions

The main results in this chapter can be split in two categories. The first category deals with the existence and uniqueness of solutions to the RCP. Both existence and uniqueness are completely characterized in terms of properties of corresponding parameterized LCPs for large parameter values. The proofs rely on convexity theory and properties of rational functions. Since a wealth of theoretical and numerical results is known for LCPs, this provides many methods to answer solvability issues of RCPs.

The second part of the chapter has shown the relation of the RCP to a class of hybrid dynamical systems: the linear complementarity class. A relation has been established between the existence of initial solutions to a linear complementarity system and the existence of solutions to the RCP. It appears that a similar relation for uniqueness is less trivial, because an example shows that it is possible that multiple initial solutions exist for a fixed initial state, although there is only one solution to the corresponding RCP. This has led to the introduction of an equivalence relation among the initial solutions. In terms of this equivalence relation, a uniqueness relation between solutions of RCP and initial solutions has been stated. The results on initial solutions have been translated to the global solution of a complementarity system.

The obtained results have been exploited to prove existence and uniqueness results of physical processes like mechanical systems subject to unilateral constraints, dissipative systems with complementarity conditions like electrical networks with diodes, and systems with relays and/or Coulomb friction. The set of examples presented here gives a flavor of the systems that can be modeled as complementarity systems and indicates the relevance of the complementarity class and the results presented here.

The proofs of the well-posedness results that we have obtained are constructive in nature, in the sense that they present specific algorithms which determine the status ("active" or "inactive") of all complementarity conditions given an initial condition. In other words, these algorithms solve the "mode selection problem". Algorithms of this type are important in the *simulation* of hybrid systems. In this chapter we have not considered the numerical issues related to mode selection problems; this is an important subject for further research.

5

Linear passive complementarity systems: well-posedness

This chapter is based on the paper [84], which has been submitted for publication. A preliminary version [36] of this chapter has been presented at the Conference on Decision and Control 1999 in Phoenix (USA). Kanat Çamlıbel acted as one of my co-authors in these papers, and these papers are also part of his PhD-work.

5.1 Introduction

Nowadays switches like thyristors and diodes are used in electrical networks for a great variety of applications in both power engineering and signal processing. For the simulation of the transient behavior of such networks the switches are often modeled ideally [13,136,154,197,203]. It is well-known that ideal modeling causes the network model to be of a mixed discrete and continuous nature. In particular, the circuit evolves through multiple topologies (modes) depending on the (discrete) states of the diodes. The mode transitions are triggered by inequalities and may result in discontinuities and Dirac impulses in the network's variables, see e.g. [56, 136, 146, 154, 175, 197, 203]. Several numerical methods have been proposed to deal with these phenomena and simulation of circuits with nonsmooth characteristics is well established by now [13, 20,41,42,63,120,121,136,172].

However, little attention has been paid to the question if and in what sense the computed time functions converge to the true solution of the network model. The simulation methods can be distinguished in two categories depending on whether or not the software attempts to find the exact times at which events take place. The convergence of "event tracking" methods might be inferred from a combination of standard results on the convergence of numerical algorithms for root finding and for simulation of smooth dynamical systems. For "time stepping" methods, which are a

popular alternative to event tracking methods in a number of applications [20,120,172], the issue of convergence is less clear. It is the objective of this chapter and of chapter 7 to provide a rigorous basis for the use of time stepping methods in the simulation of internally switched electrical circuits.

Before we are able to prove consistency of a numerical routine, we have to establish what is meant by a transient "true solution" of a dynamical network with ideal diodes. In this chapter we provide a mathematical framework that allows the precise formulation of a solution concept for these continuous/discrete networks. The framework will be borrowed from the theory of linear complementarity systems [92, 93, 123, 177, 179]. These systems can be seen as dynamical extensions of the *linear complementarity problem* [47], which (together with a number of variants) has been used extensively in the study of piecewise-linear electrical networks [20, 53, 63, 111, 120, 121, 191, 201].

The definition of true solutions is coupled to the question of existence and uniqueness of solutions of the network model (called well-posedness). Much effort has been invested in considering existence and uniqueness of solutions to *static* (DC) models of electrical networks [40, 45, 69, 70, 75, 150, 152, 165, 166, 173, 174]. However, studies of the dynamic equivalent are rare. The only papers known to the authors dealing with existence and uniqueness of (dynamic) RLC-networks are [58, 153]. Since an ideal diode cannot be reformulated as a current or voltage-controlled resistor, the obtained results in [58, 153] do not cover the networks considered here.

The main purposes of the chapter are the following.
(i) Define a mathematically precise solution concept for linear passive networks with diodes.
(ii) Prove (global) existence and uniqueness of solutions.
(iii) Establish regularity properties of the solutions. In particular, it will be rigorously proven that derivatives of Dirac impulses do not occur (even for inconsistent initial states) and Dirac impulses occur only at the initial time. Moreover, it will turn out that the set of switching times is a right-isolated set, meaning that for all time instants there exists a positive length time interval in which the diodes do not change their state. Chapter 7 will use these results to prove consistency of a transient simulation technique based on time-stepping.

The outline of the chapter is as follows. In Section 5.2 linear passive networks with diodes will be reformulated as linear complementarity systems. In Section 5.3, we describe the evolution of the network model within a given mode (i.e. with a fixed state of the diodes). Next, an extension of the linear complementarity problem will be introduced, which will play an important role in the proof of well-posedness. In Section 5.5 the solution concept is introduced. Finally, the proof of global well-posedness is presented.

The following notations will be in force. $\mathbb{N}$ denotes the set of natural numbers $\{0, 1, 2, \dots\}$, $\mathbb{R}$ the real numbers, $\mathbb{R}_+$ the nonnegative real numbers (including zero) and $\mathbb{C}$ the complex numbers. For a positive integer l, $\bar{l}$ denotes the set $\{1, 2, \dots, l\}$. If a is a (column) vector with k components, we denote its i-th component by a_i. $M^\top$ is the transpose of the matrix $M \in \mathbb{C}^{m\times n}$ and M^* denotes the complex conjugate transpose.

A matrix $M \in \mathbb{C}^{m\times m}$ is called nonnegative definite if $\operatorname{Re} x^*Mx = \frac{1}{2}x^*(M+M^*)x \geq 0$ for all $x \in \mathbb{C}^m$. This is denoted by $M \geq 0$. In case strict inequality holds for all nonzero vectors x, we call the matrix positive definite and write $M > 0$. By $\mathit{1}$ we denote the identity matrix of any dimension. Given $M \in \mathbb{R}^{k\times l}$ and two subsets $I \subseteq \bar{k}$ and $J \subseteq \bar{l}$, the (I, J)-submatrix of M is defined as $M_{IJ} := (M_{ij})_{i\in I, j\in J}$. In case $J = \bar{l}$, we also write $M_{I\bullet}$. If $I = \bar{k}$, the notation $M_{\bullet J}$ is sometimes used.

By $\mathbb{R}(s)$ we denote the field of real rational functions in one variable. $M(s) \in \mathbb{R}^{k\times l}(s)$ means that $M(s)$ is a $k \times l$ matrix with entries in $\mathbb{R}(s)$. A rational vector or matrix is called (strictly) proper, if for all entries the degree of the numerator is smaller than or equal to (strictly smaller than) the degree of the denominator.

A vector $u \in \mathbb{R}^k$ is called nonnegative, and we write $u \geq 0$, if $u_i \geq 0$ for all $i \in \bar{k}$ and positive ($u > 0$), if $u_i > 0$ for all $i \in \bar{k}$. If two vectors $u, y \in \mathbb{R}^k$ are orthogonal, i.e. $u^\top y = 0$, we write $u \perp y$. We write $u(s) \perp y(s)$ for two rational vectors $u(s), y(s) \in \mathbb{R}^k(s)$, if for all i at least one of $u_i(s) \equiv 0$ and $y_i(s) \equiv 0$ is satisfied.

The set of arbitrarily often differentiable functions from $\mathbb{R}$ to $\mathbb{R}^m$ is denoted by $C^\infty(\mathbb{R}; \mathbb{R}^m)$. $\mathcal{L}_2^k(t_0, t_1)$ denotes the set of all measurable functions v from (t_0, t_1) to $\mathbb{R}^k$ for which the integral $\int_{t_0}^{t_1} \|v(\tau)\|^2 d\tau$ is finite.

5.2 Linear passive networks with ideal diodes

Linear electrical networks consisting of (linear) resistors, inductors, capacitors, gyrators, transformers (RLCGT) and ideal diodes can be described in a complementarity formulation as mentioned in e.g. [20, 121]. Indeed, the RLCGT-network can be described by the state space model

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{5.1a}$$

$$y(t) = Cx(t) + Du(t) \tag{5.1b}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ matrices of appropriate dimensions [3]. The state variable $x(t)$ usually represents the voltages across capacitors and currents through inductors at time t. The pair (u_i, y_i) denotes the voltage-current variables at the connections to the diodes, i.e.

$$u_i = -V_i,\ y_i = I_i \text{ or } u_i = I_i,\ y_i = -V_i,$$

where V_i and I_i are the voltage across and current through the i-th diode, respectively. The ideal diode characteristics are given by Figure 5.1 and described by the relations

$$V_i \leq 0,\ I_i \geq 0,\ \{V_i = 0 \text{ or } I_i = 0\}. \tag{5.2}$$

The top branch of the characteristic corresponds to the conducting mode and the left branch to the blocking mode.

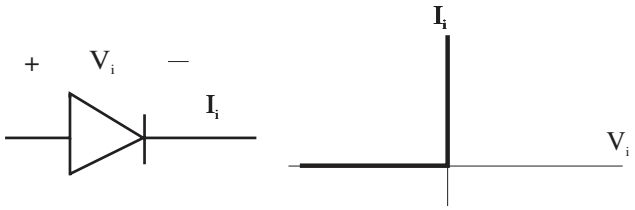

Figure 5.1: The ideal diode characteristic.

By suitable substitutions the following system description is obtained.

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) && \text{(5.3a)} \\ y(t) &= Cx(t) + Du(t) && \text{(5.3b)} \\ 0 \leq y(t) &\perp u(t) \geq 0. && \text{(5.3c)} \end{aligned}$$

In this formulation $t \in \mathbb{R}_+$ denotes the time variable, $x(t)$ the state, and $u(t)$ and $y(t)$ the complementarity variables at time t. The system (5.3) is called a *linear complementarity system* as introduced in [177] and further studied in [92,93,123,179]. We use the notation LCS(A, B, C, D) to indicate the system given by (5.3). Note that (5.3c) is equivalent to

$$y_i(t) \geq 0,\ u_i(t) \geq 0,\ \{y_i(t) = 0 \text{ or } u_i(t) = 0\}$$

for all $i \in \bar{k}$.

Since (5.3a)-(5.3b) is a model for the RLCGT-multiport network consisting of resistors, capacitors, inductors, gyrators and transformers, the quadruple (A, B, C, D) has special properties (see [3]). To be precise, the square system (A, B, C, D) is passive (or in the terms of [206], *dissipative* with respect to the supply rate $u^\top y$) as defined below.

Definition 5.2.1 [206] A system (A, B, C, D) given by (5.1) is called *passive*, or *dissipative* with respect to the supply rate $u^\top y$, if there exists a nonnegative function $V : \mathbb{R}^n \to \mathbb{R}_+$, called a *storage function*, such that for all $t_0 \leq t_1$ and all time functions $(u, x, y) \in \mathcal{L}_2^{k+n+k}(t_0, t_1)$ satisfying (5.1) the following inequality holds:

$$V(x(t_0)) + \int_{t_0}^{t_1} u^\top(t) y(t) dt \geq V(x(t_1)).$$

□

The above inequality is called the *dissipation inequality*. The storage function represents a notion of "stored energy" in the network.

A standing technical assumption throughout the remainder of the chapter will be the following.

Assumption 5.2.2 B has full column rank and (A, B, C) is a minimal representation, i.e. the matrices $[B\ AB\ \dots\ A^{n-1}B]$ and $[C^\top\ A^\top C^\top \dots (A^\top)^{n-1}C^\top]$ have full rank. □

The minimality of the system description (A, B, C) is standard in the literature on dissipative dynamical systems, see e.g. [206]. The following lemma gives several equivalent characterizations for passivity.

Proposition 5.2.3 [206] *Consider a system (A, B, C, D) with (A, B, C) a minimal representation. The following statements are equivalent.*

- *(A, B, C, D) is passive.*
- *The transfer matrix $G(s) := C(s\mathbb{1} - A)^{-1}B + D$ is* positive real, *i.e. $x^*[G(\lambda) + G^*(\lambda)]x \geq 0$ for all complex vectors x and all $\lambda \in \mathbb{C}$ with $\mathrm{Re}\,\lambda > 0$ and λ no eigenvalue of A.*
- *The matrix inequalities*

$$\begin{pmatrix} -A^\top K - KA & -KB + C^\top \\ -B^\top K + C & D + D^\top \end{pmatrix} \geq 0 \tag{5.4a}$$

and

$$K = K^\top \geq 0 \tag{5.4b}$$

have a solution K.

Moreover, in case (A, B, C, D) is passive, all solutions to the linear matrix inequalities (5.4) are positive definite and K is a solution to (5.4a) if and only if $V(x) = \frac{1}{2}x^\top Kx$ defines a storage function of the system (A, B, C, D). □

5.3 Dynamics in a given mode

Equation (5.3c) implies that for all t and for every $i = 1, \dots, k$ $u_i(t) = 0$ or $y_i(t) = 0$ must be satisfied (diode is conducting or blocking). This results in a multimodal system with 2^k modes, where each mode is characterized by a subset I of $\bar{k}$, indicating that $y_i(t) = 0$ if $i \in I$ and $u_i(t) = 0$ if $i \in I^c$ with $I^c = \bar{k} \setminus I$. For each such mode (also called "topology," "configuration," or "discrete state") the laws of motion are given by differential and algebraic equations (DAEs). Specifically, in mode I they are given by

$$\dot{x} = Ax + Bu \tag{5.5a}$$
$$y = Cx + Du \tag{5.5b}$$
$$y_i = 0,\ i \in I \tag{5.5c}$$
$$u_i = 0,\ i \in I^c. \tag{5.5d}$$

Hence, a mode complies with the circuit, where the states of the diodes are fixed and replaced by short and open connections.

The mode will vary during the time evolution of the system (diodes go from conducting to blocking or vice versa). The system evolves in a certain mode as long as the inequality conditions in (5.3c) are satisfied. At the event of a mode transition, the system may display jumps of the state variable x. Jumping phenomena are well-known in the theory of unilaterally constrained mechanical systems [31], where at impacts the velocities of the colliding bodies change instantaneously. These discontinuous and impulsive motions are also observed in electrical networks (see e.g. [56, 136, 146, 154, 175, 197, 203]) and consequently, a distributional framework will be needed to obtain a mathematically precise solution concept. We restrict ourselves to the Dirac distribution denoted by δ and its derivatives, where $\delta^{(i)}$ denotes the i-th (distributional) derivative of δ.

Definition 5.3.1 [83] An *impulsive-smooth distribution* is a distribution u of the form $\mathrm{u} = \mathrm{u}_{imp} + \mathrm{u}_{reg}$, where

- u_{imp} is a linear combination of δ and its derivatives, i.e.

$$\mathrm{u}_{imp} = \sum_{i=0}^{l} u^{-i}\delta^{(i)}$$

  for vectors $u^{-i} \in \mathbb{R}^k$, $i = 0, \ldots, l$ and

- u_{reg} is an arbitrarily often differentiable function from $[0, \infty)$ to $\mathbb{R}^k$ such that $\mathrm{u}_{reg}^{(m)}(0+) = \lim_{t\downarrow 0} \frac{d^m \mathrm{u}_{reg}}{dt^m}(t)$ is defined and finite for all $m = 0, 1, 2, \ldots$.

The class of these distributions is denoted by C_{imp}^k. For a distribution $\mathrm{u} \in_{imp}^k$, u_{imp} is called the impulsive part and u_{reg} is called the smooth part. In case $\mathrm{u}_{imp} = 0$ we call u a *regular* or *smooth* distribution. If the Laplace transform of an impulsive-smooth distribution is rational, we call the distribution of *Bohl type* or a *Bohl distribution*. For a smooth Bohl distribution, we will use the term *Bohl function*. □

Lemma 5.3.2 *Consider the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ such that Assumption 5.2.2 is satisfied and (A, B, C, D) represents a passive system. Then the following holds.*

1. *For all $I \subseteq \bar{k}$ and for all initial states x_0, there exists a unique solution $(\mathrm{u}, \mathrm{x}, \mathrm{y}) \in C_{imp}^{k+n+k}$ satisfying (the dynamics for mode I given by)*

$$\begin{aligned} \dot{\mathrm{x}} &= A\mathrm{x} + B\mathrm{u} + x_0\delta && (5.6a)\\ \mathrm{y} &= C\mathrm{x} + D\mathrm{u} && (5.6b)\\ \mathrm{y}_i &= 0,\ i \in I && (5.6c)\\ \mathrm{u}_i &= 0,\ i \in I^c && (5.6d) \end{aligned}$$

   in the distributional sense. We denote this solution by $(\mathrm{u}^{x_0,I}, \mathrm{x}^{x_0,I}, \mathrm{y}^{x_0,I})$.

2. *For all modes I there exist matrices F^I and K^I such that the smooth parts $(u, x, y) := (\mathrm{u}_{reg}^{x_0,I}, \mathrm{x}_{reg}^{x_0,I}, \mathrm{y}_{reg}^{x_0,I})$ of $(\mathrm{u}^{x_0,I}, \mathrm{x}^{x_0,I}, \mathrm{y}^{x_0,I})$ for arbitrary initial state x_0 are Bohl functions and satisfy*

$$\dot{x} = F^I x \tag{5.7}$$
$$u = K^I x \tag{5.8}$$
$$y = Cx + Du. \tag{5.9}$$

The matrices F^I and K^I only depend on the mode I and not on the particular x_0 at hand.

□

Proof.
1. The existence and uniqueness of a solution for (5.6) for all initial states x_0 is equivalent to the transfer matrix $G_{II} := C_{I\bullet}(s\mathbb{1} - A)^{-1}B_{\bullet I} + D_{II}$ being invertible as a rational matrix [83, Prop. 3.23, Thm. 3.24, Thm. 3.26] (see also Lemma 3.3.3 in the thesis). This can also be seen from (5.14)-(5.15) below. Hence, suppose that $\det G_{II}(s) \equiv 0$. Then there exists a rational vector $v(s) \not\equiv 0$ such that $G_{II}(s)v(s) \equiv 0$. Take $\sigma > 0$ such that $v(\sigma) \neq 0$ and $\sigma\mathbb{1} - A$ is invertible. Define $\bar{u}$ as

$$\bar{u}_i := \begin{cases} 0 & \text{if } i \notin I \\ v_i(\sigma) & \text{if } i \in I \end{cases}$$

The triple

$$u(t) = \bar{u}e^{\sigma t} \tag{5.10}$$
$$x(t) = (\sigma\mathbb{1} - A)^{-1}B\bar{u}e^{\sigma t} \tag{5.11}$$
$$y(t) = G(\sigma)\bar{u}e^{\sigma t} \tag{5.12}$$

satisfies the system equations (5.1), where $G(s) = C(s\mathbb{1} - A)^{-1}B + D$. Since (A, B, C, D) is passive, there exists a $K > 0$ such that the dissipation inequality

$$x^\top(t_0)Kx(t_0) + \int_{t_0}^{t_1} u^\top(t)y(t)dt \geq x^\top(t_1)Kx(t_1) \tag{5.13}$$

holds for all t_0 and t_1 with $t_1 \geq t_0$. It can be verified that $u^\top(t)y(t) = e^{2\sigma t}\bar{u}^\top G(\sigma)\bar{u} = e^{2\sigma t}v(\sigma)^\top G_{II}(\sigma)v(\sigma) = 0$ for all t. By letting t_0 tend to $-\infty$, (5.13) results in

$$0 \geq x^\top(t_1)Kx(t_1)$$

for all t_1. Due to $K > 0$, this implies that $x(t_1) = 0$ for all t_1. From (5.11) it follows that $B\bar{u} = 0$. Since B is of full column rank, $\bar{u} = 0$ and hence also $v(\sigma) = 0$. We reached a contradiction and hence proved the first statement.

2. This statement follows from [83, Thm. 3.10]. □

Remark 5.3.3 From the proof of Lemma 5.3.2 it can be inferred that there exists a $\sigma_0 \in \mathbb{R}$ such that for all $\sigma \geq \sigma_0$ the principal minors of $G(\sigma)$ are positive, i.e. $\det G_{II}(\sigma) > 0$ for all $I \subseteq \bar{k}$. In terms of [47, Def. 3.3.1] $G(\sigma)$ is a P-matrix for all sufficiently large σ. This is most easily seen from the positive realness of $G(s)$, which implies that $G(\sigma)$ is nonnegative definite for all $\sigma > 0$. Since a nonnegative definite matrix has only nonnegative principal minors [47, p. 153] and $\det G_{II}(s) \not\equiv 0$ (as shown in the proof of Lemma 5.3.2), the statement follows. □

The solutions $(\mathrm{u}^{x_0,I}, \mathrm{x}^{x_0,I}, \mathrm{y}^{x_0,I})$ have *rational* Laplace transforms, denoted by $(\hat{\mathrm{u}}^{x_0,I}(s), \hat{\mathrm{x}}^{x_0,I}(s), \hat{\mathrm{y}}^{x_0,I}(s))$, which satisfy

$$
\begin{aligned}
s\hat{\mathrm{x}}^{x_0,I}(s) &= A\hat{\mathrm{x}}^{x_0,I}(s) + B\hat{\mathrm{u}}^{x_0,I}(s) + x_0 && \text{(5.14a)}\\
\hat{\mathrm{y}}^{x_0,I}(s) &= C\hat{\mathrm{x}}^{x_0,I}(s) + D\hat{\mathrm{u}}^{x_0,I}(s) && \text{(5.14b)}\\
\hat{\mathrm{y}}_I^{x_0,I}(s) &= 0 && \text{(5.14c)}\\
\hat{\mathrm{u}}_{I^c}^{x_0,I}(s) &= 0. && \text{(5.14d)}
\end{aligned}
$$

We introduce $G(s) = C(s\mathbb{1} - A)^{-1}B + D$ and $R(s) = C(s\mathbb{1} - A)^{-1}$. Since $G_{II}(s)$ is invertible as a rational matrix (see proof of Lemma 5.3.2), the equations (5.14) can be explicitly solved. This yields that the Laplace transforms $(\hat{\mathrm{u}}^{x_0,I}(s), \hat{\mathrm{x}}^{x_0,I}(s), \hat{\mathrm{y}}^{x_0,I}(s))$ are given by

$$
\begin{aligned}
\hat{\mathrm{u}}_I^{x_0,I}(s) &= -G_{II}^{-1}(s)R_{I\bullet}(s)x_0 && \text{(5.15a)}\\
\hat{\mathrm{u}}_{I^c}^{x_0,I}(s) &= 0 && \text{(5.15b)}\\
\hat{\mathrm{x}}^{x_0,I}(s) &= (s\mathbb{1} - A)^{-1}Bx_0 + (s\mathbb{1} - A)^{-1}B\hat{\mathrm{u}}^{x_0,I}(s) && \text{(5.15c)}\\
\hat{\mathrm{y}}_{I^c}^{x_0,I}(s) &= [R_{I^c\bullet}(s) - G_{I^cI}(s)G_{II}^{-1}(s)R_{I\bullet}(s)]x_0 && \text{(5.15d)}\\
\hat{\mathrm{y}}_I^{x_0,I}(s) &= 0. && \text{(5.15e)}
\end{aligned}
$$

Hence, the solutions of the mode dynamics (5.6) are one-to-one related (by the Laplace transform and its inverse) to solutions satisfying (5.14). On the basis of this relation, we can prove that only Dirac impulses (and not its derivatives) show up in passive electrical networks with diodes.

Theorem 5.3.4 *Consider matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ such that Assumption 5.2.2 is satisfied and (A, B, C, D) represents a passive system. Then for each $x_0 \in \mathbb{R}^n$ and $I \subseteq \bar{k}$ the Laplace transform $\hat{\mathrm{u}}^{x_0,I}(s)$ is proper.* □

Proof. Denote $\hat{\mathrm{u}}^{x_0,I}(s)$ by $u(s)$ for brevity. The triple

$$
\begin{aligned}
\bar{u}(t) &= u(\sigma)e^{\sigma t} && \text{(5.16)}\\
\bar{x}(t) &= (\sigma\mathbb{1} - A)^{-1}Bu(\sigma)e^{\sigma t} && \text{(5.17)}\\
\bar{y}(t) &= G(\sigma)u(\sigma)e^{\sigma t} && \text{(5.18)}
\end{aligned}
$$

satisfies (5.1) for all $\sigma \in \mathbb{R}$ such that $\sigma \mathbb{1} - A$ is nonsingular. It follows from passivity that there exists a $K > 0$ such that for all t_1 and t_0 with $t_1 \geq t_0$

$$\bar{x}^\top(t_1)K\bar{x}(t_1) - \bar{x}^\top(t_0)K\bar{x}(t_0) \leq \int_{t_0}^{t_1} \bar{u}^\top(t)\bar{y}(t)dt. \tag{5.19}$$

By substituting (5.16)-(5.18) into the dissipation inequality (5.19), one obtains

$$u^\top(\sigma)B^\top(\sigma\mathbb{1} - A)^{-\top}K(\sigma\mathbb{1} - A)^{-1}Bu(\sigma) \leq \frac{1}{2\sigma}u^\top(\sigma)G(\sigma)u(\sigma). \tag{5.20}$$

Since $K > 0$, B has full column rank, and $(\sigma\mathbb{1} - A)^{-1}$ is strictly proper, there exists an $\alpha > 0$ such that

$$\frac{\alpha}{\sigma^2}\|u(\sigma)\|^2 \leq u^\top(\sigma)B^\top(\sigma\mathbb{1} - A)^{-\top}K(\sigma\mathbb{1} - A)^{-1}Bu(\sigma) \tag{5.21}$$

for all sufficiently large σ. We know from (5.14) that $u\top(s)y(s) = 0$, where $y(s) := \hat{\mathrm{y}}^{x_0,I} = C(s\mathbb{1} - A)^{-1}x_0 + G(s)u(s)$. Hence, the right-hand side of (5.20) satisfies

$$\begin{aligned}\frac{1}{2\sigma}u^\top(\sigma)G(\sigma)u(\sigma) &= -\frac{1}{2\sigma}u^\top(\sigma)C(\sigma\mathbb{1} - A)^{-1}x_0 \\ &\leq \frac{1}{2\sigma}\|C(\sigma\mathbb{1} - A)^{-1}x_0\|\|u(\sigma)\| \\ &\leq \frac{\beta}{2\sigma^2}\|u(\sigma)\|\|x_0\| \end{aligned} \tag{5.22}$$

The last inequality follows from the existence of a $\beta > 0$ such that $\|C(\sigma\mathbb{1}-A)^{-1}\| \leq \frac{\beta}{\sigma}$ for all sufficiently large σ. Thus, (5.20), (5.21) and (5.22) yield

$$\|u(\sigma)\| \leq \frac{\beta}{2\alpha}\|x_0\| \tag{5.23}$$

for all sufficiently large σ. Hence, $u(s)$ must be proper. □

The fact that solutions of linear passive networks with ideal diodes do not contain derivatives of Dirac impulses is widely believed true, but the authors are not aware of any previous rigorous proof. The framework proposed here makes it possible to prove this intuition.

To summarize the discussion so far, it has been shown that instead of considering impulsive-smooth distributions as the solution space within a mode, we can restrict ourselves to Bohl distributions with impulsive part containing only Dirac impulses and not its derivatives (i.e. Bohl distributions with *proper* Laplace transforms).

Consider a solution to (5.6) for mode I and initial state x_0. An important observation is that a nontrivial impulsive part of $\mathrm{u}^{x_0,I}$ will result in a re-initialization (jump) of the state. If $\mathrm{u}_{imp} = u^0\delta$, then a jump will take place according to

$$\mathrm{x}_{reg}(0+) := \lim_{t\downarrow 0} \mathrm{x}_{reg}(t) = x_0 + Bu^0. \tag{5.24}$$

The proof can be found in [83].

5.4 Rational complementarity problem

In the previous section the dynamics within a mode (i.e. with a fixed state of the diodes) has been considered, while the inequality conditions have been neglected. However, a solution $(\mathrm{u}^{x_0,I}, \mathrm{x}^{x_0,I}, \mathrm{y}^{x_0,I})$ within a mode (5.6) will only be valid on an "initial" interval due to a change of mode (diode going from conducting to blocking or vice versa) triggered by the inequality constraints. Therefore, we would like to express some kind of "local nonnegativity." We call a (smooth) Bohl function v *initially nonnegative* if there exists an $\varepsilon > 0$ such that $v(t) \geq 0$ for all $t \in [0, \varepsilon)$. Note that a Bohl function v is initially nonnegative if and only if there exists a $\sigma_0 \in \mathbb{R}$ such that its Laplace transform $\hat{v}(\sigma) \geq 0$ for all $\sigma \geq \sigma_0$. Hence, there is a connection between small time values for time functions and large values for the indeterminate s in the Laplace transform. This fact is closely related to the well-known initial value theorem (see e.g. [59]). The definition of initial nonnegativity for Bohl distributions will be based on this observation (see also Chapters 3 and 4).

Definition 5.4.1 We call a Bohl distribution v *initially nonnegative*, if its Laplace transform $\hat{\mathrm{v}}(s)$ satisfies $\hat{v}(\sigma) \geq 0$ for all sufficiently large real σ. □

Remark 5.4.2 To relate the definition to the time domain, note that a scalar-valued Bohl distribution v without derivatives of the Dirac impulse (i.e. $\mathrm{v}_{imp} = v^0\delta$ for some $v^0 \in \mathbb{R}$) is initially nonnegative if and only if

1. $v^0 > 0$, or
2. $v^0 = 0$ and there exists an $\varepsilon > 0$ such that $v_{reg}(t) \geq 0$ for all $t \in [0, \varepsilon)$.

□

Definition 5.4.3 We call a Bohl distribution $(\mathrm{u}, \mathrm{x}, \mathrm{y}) \in \mathcal{C}_{imp}^{k+n+k}$ an *initial solution* to (5.3) with initial state x_0, if there exists an $I \subseteq \bar{k}$ such that

1. $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ satisfies (5.6) for mode I and initial state x_0 in the distributional sense (i.e. $(\mathrm{u}, \mathrm{x}, \mathrm{y}) = (\mathrm{u}^{x_0,I}, \mathrm{x}^{x_0,I}, \mathrm{y}^{x_0,I})$) and
2. u, y are initially nonnegative.

□

Example 5.4.4 Consider the system $\dot{x}(t) = u(t)$, $y(t) = x(t)$ together with (5.3c). This represents a system consisting of a capacitor connected to a diode. The current in the network is equal to u and the voltage across the capacitor is equal to $y = x$. For initial state $x(0) = x_0 = 1$, $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ with $\mathrm{u} = 0$ (no current) and $\mathrm{y}(t) = \mathrm{x}(t) = 1$ for all $t \in \mathbb{R}$ is an initial solution. This corresponds to the case that the diode is always blocking and there is no (nonzero) current in the network. To demonstrate that the

distributional framework is needed, consider initial state $x_0 = -1$ for which $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ with $\mathrm{u} = \delta$, $\mathrm{x}(t) = \mathrm{y}(t) = 0$, $t > 0$ is the unique initial solution. This corresponds to an instantaneous discharge of the capacitor at time instant 0. Note that a state jump occurs at time 0 from -1 to $P_{\{1\}}(-1) = 0$. □

We emphasize that an initial solution only satisfies the equations (5.3) in the following temporary sense. In case an initial solution has a nontrivial impulsive part, only the re-initialization as given in (5.24) forms a piece of the "global solution." If the initial solution $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ is smooth, the largest interval on which $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ satisfies the equations (5.3) is equal to $[0, \varepsilon)$, where ε is given by

$$\varepsilon := \inf\{t > 0 \mid \mathrm{u}_{reg,i}(t) < 0 \text{ or } \mathrm{y}_{reg,i}(t) < 0 \text{ for some } i \in \bar{k}\}. \tag{5.25}$$

Example 5.4.5 Consider the network depicted in Figure 5.2 with $R_1 = 2\,\Omega$, $R_2 = 1\,\Omega$, $L = 1\,H$ and $C = 1\,F$. We introduce the variables x_1 as the voltage across the capacitor, x_2 the current through the inductor, $-u$ the voltage across the diode and y the current through the diode. The system is governed by the equations

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 - 2x_2 + u \\ y &= x_2 + u \end{aligned}$$

together with the complementarity conditions (5.3c). For initial condition $x_1(0) = -1$,

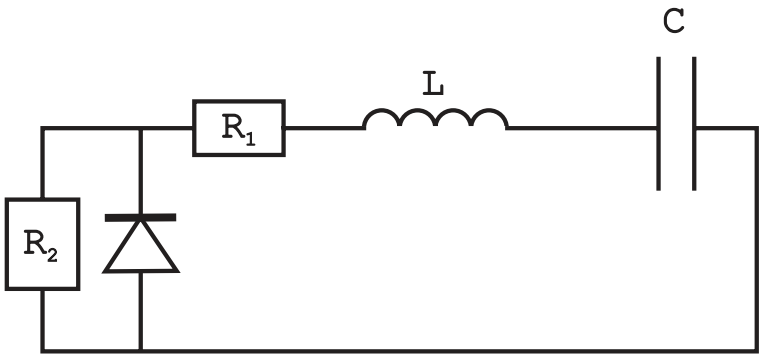

Figure 5.2: A simple network.

$x_2(0) = 2$, it can be verified that the unique initial solution (in the conducting mode) is given by $\mathrm{u} = 0$, $\mathrm{x}_1(t) = (t-1)e^{-t}$, $\mathrm{y}(t) = \mathrm{x}_2(t) = (2-t)e^{-t}$, $t > 0$. This initial solution forms a part of the (global) solution on the interval $[0, \varepsilon) = [0, 2)$. The time $t = 2$ is determined by the violation of the inequality constraint $y(t) \geq 0$ corresponding to the current through the diode becoming negative. This causes the diode to go from conducting to blocking. To determine the next part of the global solution, we have

to find a continuation from initial state $x(2) = (e^{-2}, 0)^\top$, i.e. determining an initial solution with initial state $(e^{-2}, 0)^\top$ (after a suitable shift of the time axis). □

As a solution within a mode exists and is unique given an initial state, it still might be possible that there is more than one initial solution. Since there are 2^k (k the number of diodes) modes, the maximum number of initial solutions is equal to 2^k. The other extreme is that there is no initial solution at all, i.e. no solution within a mode satisfies the initial nonnegativity conditions. We will start our investigation of well-posedness for linear passive complementarity systems by studying existence and uniqueness of initial solutions. An important tool in existence and uniqueness of initial solutions is the *rational complementarity problem* (RCP).

Definition 5.4.6 (Rational complementarity problem) Let the vector $x_0 \in \mathbb{R}^n$ and matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ be given. The *rational complementarity problem* RCP(x_0, A, B, C, D) is the problem of finding rational k-vectors $u(s) \in \mathbb{R}^k(s)$ and $y(s) \in \mathbb{R}^k(s)$ such that

1. for all $s \in \mathbb{C}$

$$y(s) = C(s\mathbb{1} - A)^{-1}x_0 + [C(s\mathbb{1} - A)^{-1}B + D]u(s) \text{ and } u(s) \perp y(s), \tag{5.26}$$

and

2. there exists a $\sigma_0 \in \mathbb{R}$ satisfying for all $\sigma > \sigma_0$

$$y(\sigma) \geq 0 \quad \text{and } u(\sigma) \geq 0. \tag{5.27}$$

Any pair of rational vectors $(u(s), y(s))$ satisfying the above conditions is said to be a *solution* to RCP(x_0, A, B, C, D). If A, B, C and D are clear from the context, we also write RCP(x_0) for brevity. □

From the definition of initial nonnegativity and (5.14), the following important relation is clear from Chapter 3.

Theorem 5.4.7 *Consider the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ and assume that all modes of LCS(A, B, C, D) are autonomous. Then the following statements hold.*

- *All initial solutions are of Bohl type.*
- *There is a one-to-one correspondence between initial solutions to (5.3) and solutions to RCP(x_0). More specifically, $(\mathrm{u}, \mathrm{x}, \mathrm{y})$ is an initial solution to (5.3) if and only if its Laplace transform $(\hat{\mathrm{u}}(s), \hat{\mathrm{x}}(s), \hat{\mathrm{y}}(s))$ is such that $(\hat{\mathrm{u}}(s), \hat{\mathrm{y}}(s))$ is a solution to RCP(x_0) and*

$$\hat{\mathrm{x}}(s) = (s\mathbb{1} - A)^{-1}x_0 + (s\mathbb{1} - A)^{-1}B\hat{\mathrm{u}}(s). \tag{5.28}$$

- *The following statements are equivalent.*
  1. *There exists a unique initial solution to LCS*(A, B, C, D) *for initial state* x_0.
  2. *RCP*(x_0) *has a unique solution.*
- *The initial solution is smooth if and only if the corresponding solution to RCP*(x_0) *is strictly proper. Similarly, the initial solution has an impulsive part containing only Dirac distributions (and not its derivatives) if and only if the corresponding solution to RCP*(x_0) *is proper.*

□

As a consequence, studying existence and uniqueness of initial solutions is equivalent to studying existence and uniqueness of solutions to RCPs. In Chapter 4 necessary and sufficient conditions for existence and uniqueness of solutions to RCPs have been presented in terms of families of *linear complementarity problems* (cf. Definition 5.4.10 below). Based on this relation and the literature on linear complementarity problems the following result has been proven in Chapter 4.

Theorem 5.4.8 *Consider matrices* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ *and* $D \in \mathbb{R}^{k\times k}$ *such that Assumption 5.2.2 is satisfied and* (A, B, C, D) *represents a passive system. Then RCP*(x_0) *has a unique solution for all* x_0. □

Theorem 5.4.7 yields now the following corollary.

Theorem 5.4.9 *Consider matrices* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ *and* $D \in \mathbb{R}^{k\times k}$ *such that Assumption 5.2.2 is satisfied and* (A, B, C, D) *represents a passive system. From each initial state* x_0 *there exists exactly one initial solution to LCS*(A, B, C, D). □

According to Theorem 5.4.7 there exists a one-to-one relation between initial solutions and solutions to RCP. Properties of the solutions to RCP (e.g. strict properness) translate directly to properties of initial solutions. In the next theorem we will therefore study the solutions to RCPs. We need the following concepts to formulate the theorem.

Definition 5.4.10 Let a real vector $q \in \mathbb{R}^k$ and a real matrix $M \in \mathbb{R}^{k\times k}$ be given. LCP(q, M) is the problem of finding a real vector $z \in \mathbb{R}^k$ such that $0 \leq z \perp (q + Mz) \geq 0$ or show that no such z exists. □

For an extensive survey on LCPs, we refer to [47]. The set of all solutions z to LCP(q, M) will be denoted by SOL(q, M).

Remark 5.4.11 If $(u(s), y(s))$ is a solution to RCP(x_0, A, B, C, D), then $u(\sigma)$ is a solution to LCP$(C(\sigma \mathbb{1} - A)^{-1}x_0, G(\sigma))$ for all sufficiently large (real) σ, where $G(s) = C(s\mathbb{1} - A)^{-1}B + D$. □

Remark 5.4.12 We shall employ the following standard observation on LCP-solutions several times. If $z_i \in \mathrm{SOL}(q_i, M_i)$ with $i \in \{1, 2\}$ then

$$\begin{aligned}(z_1 - z_2)^\top((q_1 + M_1 z_1) - (q_2 + M_2 z_2)) \\ = -z_1^\top(q_2 + M_2 z_2) - z_2^\top(q_1 + M_1 z_1) \leq 0\end{aligned}$$

□

Finally, a *dual cone* is defined as follows [47].

Definition 5.4.13 Let $\mathcal{Q}$ be a nonempty set in $\mathbb{R}^k$. The *dual cone* of $\mathcal{Q}$, denoted by $\mathcal{Q}^*$, is defined as the set

$$\mathcal{Q}^* = \{w \in \mathbb{R}^k \mid w^\top v \geq 0 \text{ for all } v \in \mathcal{Q}\}.$$

□

Theorem 5.4.14 *Consider matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ such that Assumption 5.2.2 is satisfied and (A, B, C, D) represents a passive system. Denote the solution set of LCP$(0, D)$ by $\mathcal{Q} := SOL(0, D)$. Furthermore, let $(u_{x_0}(s), y_{x_0}(s))$ be the (unique) solution to RCP(x_0). The following assertions hold:*

1. *For all $x_0 \in \mathbb{R}^n$, $C(x_0 + Bu^0) \in \mathcal{Q}^*$ where $u^0 = \lim_{s\to\infty} u_{x_0}(s)$.*
2. *$u_{x_0}(s)$ is strictly proper if and only if $Cx_0 \in \mathcal{Q}^*$.*
3. *$\lim_{s\to\infty} u_{x_0}(s) \in \mathcal{Q}$.*

□

Proof.
1: In view of Remark 5.4.11 and Remark 5.4.12, we have for each $v \in \mathcal{Q} := \mathrm{SOL}(0, D)$ that

$$(u_{x_0}(\sigma) - v)^\top(C(\sigma \mathbb{1} - A)^{-1}x_0 + G(\sigma)u_{x_0}(\sigma) - Dv) \leq 0$$

for all sufficiently large σ. Since $D \geq 0$ ((5.4a) yields $D + D^\top \geq 0$) and $G(\sigma) = C(\sigma \mathbb{1} - A)^{-1}B + D$, we obtain

$$(u_{x_0}(\sigma) - v)^\top(C(\sigma \mathbb{1} - A)^{-1}x_0 + C(\sigma \mathbb{1} - A)^{-1}Bu_{x_0}(\sigma)) \leq 0 \tag{5.29}$$

for all sufficiently large σ. Multiplying this relation by σ and letting σ tend to infinity give

$$(u^0 - v)^\top(Cx_0 + CBu^0) \leq 0$$

Since $\mathcal{Q}$ is a cone, we have for all $\lambda \geq 0$ and all $v \in \mathcal{Q}$

$$(u^0 - \lambda v)^\top(Cx_0 + CBu^0) \leq 0$$

and hence,

$$\lambda v^\top (Cx_0 + CBu^0) \geq u^{0\top}(Cx_0 + CBu^0).$$

It follows that $v^\top(Cx_0 + CBu^0) \geq 0$ for all $v \in \mathcal{Q}$ and thus $C(x_0 + Bu^0) \in \mathcal{Q}^*$.

2: "only if": Suppose that the solution $(u_{x_0}(s), y_{x_0}(s))$ to RCP(x_0) is such that $u_{x_0}(s)$ is strictly proper. According to statement *1*, $Cx_0 \in \mathcal{Q}^*$, because $u^0 = 0$.

"if": Suppose that $Cx_0 \in \mathcal{Q}^*$. We know that $u_{x_0}(s)$ is proper. Take the power series expansion of $u_{x_0}(s)$ around infinity as

$$u_{x_0}(s) = u^0 + u^1 s^{-1} + u^2 s^{-2} + \cdots \tag{5.30}$$

By substituting (5.30) into

$$u_{x_0}^\top(s) y_{x_0}(s) = u_{x_0}^\top(s)(C(s\mathbb{1} - A)^{-1}x_0 + G(s)u_{x_0}(s)) = 0,$$

we obtain by considering the coefficients corresponding to s^0 and s^{-1}

$$u^{0\top} D u^0 = 0 \tag{5.31}$$

$$u^{0\top} C x_0 + u^{0\top} D u^1 + u^{1\top} D u^0 + u^{0\top} C B u^0 = 0 \tag{5.32}$$

Since $(u_{x_0}(s), y_{x_0}(s))$ is the solution to RCP(x_0), $u^0 \geq 0$ and $Du^0 \geq 0$. Together with (5.31), this gives $u^0 = \lim_{s\to\infty} u_{x_0}(s) \in \mathcal{Q}$ (this proves statement *3*). The relation (5.31) also implies

$$(D + D^\top)u^0 = 0 \tag{5.33}$$

According to Theorem 5.2.3, passivity of the system implies the existence of a symmetric $K > 0$ such that

$$\begin{bmatrix} A^\top K + KA & KB - C^\top \\ B^\top K - C & -(D + D^\top) \end{bmatrix} \leq 0 \tag{5.34}$$

Premultiplying (5.34) by $(\gamma z^\top \; u^{0\top})$ and postmultiplying by $(\gamma z^\top \; u^{0\top})^\top$ for arbitrary $z \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$, yields (use (5.33))

$$\gamma^2 z^\top (A^\top K + KA) z + 2\gamma z^\top (KB - C^\top) u^0 \leq 0$$

Considering this expression as an inequality for a quadratic form in γ, yields that $z^\top(KB - C^\top)u^0 \leq 0$. Since z is arbitrary, we obtain

$$(KB - C^\top)u^0 = 0 \tag{5.35}$$

Now, (5.32) and (5.33) give

$$u^{0\top} C x_0 + u^{0\top} C B u^0 = 0 \tag{5.36}$$

On the other hand, from (5.35), we obtain $u^{0\top}CBu^0 = u^{0\top}B^\top KBu^0$. Since $u^0 \in \mathcal{Q}$ and $Cx_0 \in \mathcal{Q}^*$, (5.36) gives

$$0 \geq -u^{0\top}Cx_0 = u^{0\top}CBu^0 = u^{0\top}B^\top KBu^0 \geq 0$$

Finally, positive definiteness of K and the full column rank of B imply $u^0 = 0$, i.e. $u_{x_0}(s)$ is strictly proper.

3: This has already been shown in the proof of statement *2*. □

Theorem 5.4.14 has several immediate consequences.

Definition 5.4.15 A state x_0 is called *regular* for LCS(A, B, C, D), if the corresponding initial solution is smooth. The collection of regular states is denoted by $\mathcal{R}$. □

Since strictly proper Laplace transforms correspond to smooth Bohl distributions (i.e. Bohl functions), statement *2* in Theorem 5.4.14 gives a characterization of the regular states: $x_0 \in \mathcal{R}$ if and only if $Cx_0 \in \mathcal{Q}^*$ with $\mathcal{Q} = \text{SOL}(0, D)$. As we shall see, this characterization plays a key role in the proof of global existence of solutions as the set of such initial states will be proven to be invariant under the dynamics.

According to [47, Cor. 3.8.10 and Thm 3.1.7 (c)] and because $D \geq 0$ one has $Cx_0 \in \mathcal{Q}^*$ if and only if LCP(Cx_0, D) is solvable. Hence, a test for deciding the regularity of an initial state consist of determining whether or not a certain LCP has a solution. In [13] it is stated that a well-designed circuit does not contain Dirac impulses. As a consequence, the characterization of $\mathcal{R}$ forms a verification of the synthesis of the network containing diodes.

To give an idea about the structure of the cone $\mathcal{Q}^*$ and $\mathcal{R}$, a few examples are in order.

Example 5.4.16 Consider the following situations.
(a) If $D = 0$, then $\mathcal{Q} = \mathbb{R}^k_+$ and $\mathcal{Q}^* = \mathbb{R}^k_+$. Hence, $\mathcal{R} = \{x_0 \in \mathbb{R}^n \mid Cx_0 \geq 0\}$.
(b) If $D = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, then $\mathcal{Q} = \{\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \mid u_1 \geq 0 \text{ and } u_2 = 0\}$. Consequently, $\mathcal{Q}^* = \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \mid y_1 \geq 0\}$ and thus $\mathcal{R} = \{x_0 \in \mathbb{R}^n \mid C_{1\bullet}x_0 \geq 0\}$.
(c) If D is positive definite, it follows that $\mathcal{Q} = \{0\}$, which implies that $\mathcal{Q}^* = \mathbb{R}^k$ and thus $\mathcal{R} = \mathbb{R}^n$. □

A direct implication of the statements *1* and *2* in Theorem 5.4.14 is that, if smooth continuation is not possible for x_0, it is possible after one re-initialization. Indeed, by (5.24) the state after re-initialization is equal to $x_0 + Bu^0$, if the impulsive part of the (unique) initial solution is equal to $u^0\delta$. According to the fact that the Laplace transform of an initial solution is a solution to the corresponding RCP (which is automatically proper), it follows that $\lim_{s\to\infty} u_{x_0}(s) = u^0$ is indeed the coefficient determining the

impulsive part. Since $C(x_0 + Bu^0) \in \mathcal{Q}^*$, it follows from statement *2* that $x_0 + Bu^0$ is a regular state. Hence, from $x_0 + Bu^0$ there exists a smooth initial solution. To summarize this discussion, we formulate a local existence result.

Theorem 5.4.17 *Consider matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ such that Assumption 5.2.2 is satisfied and (A, B, C, D) represents a passive system. For all initial states x_0, there exists a unique local solution. To be specific, for all x_0 there exists a unique Bohl distribution (u, x, y) defined on $[0, \varepsilon)$ for some $\varepsilon > 0$ such that*

1. *There exists an initial solution $(\bar{\mathrm{u}}, \bar{\mathrm{x}}, \bar{\mathrm{y}})$ such that*

$$(u_{imp}, x_{imp}, y_{imp}) = (\bar{\mathrm{u}}_{imp}, \bar{\mathrm{x}}_{imp}, \bar{\mathrm{y}}_{imp})$$

   with $\mathrm{u}_{imp} = u^0\delta$ for some $u^0 \in \mathbb{R}^k$,

2. *$x(0+) = x_0 + Bu^0$, and*

3. *for all $t \in (0, \varepsilon)$*

$$\begin{aligned} x(t) &= x(0+) + \int_0^t [Ax(\tau) + Bu(\tau)]d\tau \\ y(t) &= Cx(t) + Du(t) \\ 0 &\leq u(t) \perp y(t) \geq 0. \end{aligned}$$

□

5.5 Solution concept and global well-posedness

In the Chapters 3 and 4 a (global) solution concept has been introduced that is based on concatenation of initial solutions. In principle, this allows impulses at any mode transition time (necessary for e.g. unilaterally constrained mechanical systems). In the context of linear passive electrical networks with diodes, such a general notion of solution will not be needed. In fact, the solution concept as formulated in Theorem 5.4.17 will be extended such that mode changes are possible. This will be achieved by dropping the Bohl requirement and allowing $\mathcal{L}_2$ functions as regular parts. The function space $\mathcal{L}_\delta(0, T)$ consists of the distributions of the form $\mathrm{u} = \mathrm{u}_{imp} + \mathrm{u}_{reg}$, where $\mathrm{u}_{imp} = u^0\delta$ with $u_0 \in \mathbb{R}$ and $\mathrm{u}_{reg} \in \mathcal{L}_2(0, T)$.

Definition 5.5.1 Consider matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ such that Assumption 5.2.2 is satisfied and (A, B, C, D) represents a passive system. Let a time horizon $T > 0$ and initial state x_0 be given. $(u, x, y) \in \mathcal{L}_\delta^{k+n+k}(0, T)$ is called a solution to LCS(A, B, C, D) on $[0, T]$, if

1. There exists an initial solution $(\bar{u}, \bar{x}, \bar{y})$ such that

$$(u_{imp}, x_{imp}, y_{imp}) = (\bar{u}_{imp}, \bar{x}_{imp}, \bar{y}_{imp})$$

with $u_{imp} = u^0\delta$ for some $u^0 \in \mathbb{R}^k$,

2. $x(0+) = x_0 + Bu^0$, and

3. for almost all $t \in (0, T)$

$$\begin{aligned} x(t) &= x(0+) + \int_0^t [Ax(\tau) + Bu(\tau)]d\tau \\ y(t) &= Cx(t) + Du(t) \\ 0 &\leq u(t) \perp y(t) \geq 0. \end{aligned}$$

□

We have already proven local well-posedness (Theorem 5.4.17). The question arises whether global well-posedness is also guaranteed.

5.5.1 Global existence

We now come to the main existence result of this chapter.

Theorem 5.5.2 *Consider matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$ such that Assumption 5.2.2 is satisfied and (A, B, C, D) represents a passive system. Then, for all initial states x_0 and all $T > 0$ the system LCS(A, B, C, D) has a solution on $[0, T]$ in the sense of Definition 5.5.1.* □

Proof. The construction of a solution will be based on concatenation of initial solutions. Theorem 5.4.17 implies that a solution (u, x, y) exists on $[0, \tau_1)$ (take τ_1 as large as possible, i.e. equal to ε as in (5.25)) from initial state x_0. Note that $x(0+) \in \mathcal{R}$ and that (u, x, y) is part of a smooth initial solution with initial state $x(0+)$. Since $t \to (u, x, y)(t + \rho)$ forms a smooth initial solution for any $\rho \in (0, \varepsilon)$, we have that $x(\rho) \in \mathcal{R}$ for all $\rho \in (0, \varepsilon)$. Since (u, x, y) is a Bohl function, the limit $\lim_{t\uparrow\varepsilon} x(t) = x(\varepsilon)$ exists. The closedness of $\mathcal{R}$ (follows from statement *2* in Theorem 5.4.14) implies that $x(\varepsilon) \in \mathcal{R}$. Due to local existence of solutions and $x(\varepsilon) \in \mathcal{R}$, there exists a smooth continuation (a smooth initial solution) from $x(\varepsilon)$ that defines a solution on $[0, \tau_2)$ with $\tau_2 > \tau_1$. This construction can be repeated as long as the limit $\lim_{t\uparrow\tau} x(t)$ exists, where $[0, \tau)$ is the time-interval on which a solution has been generated so far. The reason that a global solution (on $[0, T]$) does not exist might be that the intervals of continuation $[\tau_i, \tau_{i+1})$ are getting smaller and smaller such that $\lim_{i\to\infty} \tau_i = \tau^* < T$ and $\lim_{t\uparrow\tau^*} x(t)$ does not exist. To complete the proof we will show the existence of the latter limit in any circumstances.

Suppose the maximal interval on which a solution (u, x, y) can be defined is $[0, \tau^*)$, $\tau^* < T$. According to Lemma 5.3.2 there is at most exponential growth ($\dot{x} = F^I x$) between mode changes. Since x is continuous on $(0, \tau^*)$, this implies that x is bounded (say $\|x(t)\| \leq M$ for all $t \in [0, \tau^*)$) On an interval $(s, t) \subseteq [0, \tau^*)$ where (u, x, y) is governed by the dynamics $\dot{x} = F^I x$ of mode I, the following estimate holds

$$\|x(t) - x(s)\| = \|e^{F^I(t-s)}x(s) - x(s)\| \leq c_I \mid t - s \mid \|x(s)\| \leq c_I M \mid t - s \mid \tag{5.37}$$

Note that the matrix function $t \to \frac{e^{tF^I} - \mathbb{1}}{t}$ is bounded (by c_I) on $[0, \tau^*)$. Hence, for $(s, t) \subseteq [0, \tau^*)$ with x possibly evolving through several modes we get from (5.37) that

$$\|x(t) - x(s)\| \leq M \max_{I \subseteq \bar{k}} c_I \mid t - s \mid .$$

This implies that x is Lipschitz continuous on $[0, \tau^*)$ and thus also uniformly continuous. A standard result in mathematical analysis [169, ex. 4.13] states that $x^* := \lim_{t \uparrow \tau^*} x(t)$ exists. From the construction above it can be derived that $x(t) \in \mathcal{R}$ for all $t \in [0, \tau^*)$ and hence, $x^* \in \mathcal{R}$, which implies that smooth continuation is possible (local existence) from x^* beyond τ^*. This contradicts the definition of τ^*. Hence, existence of a solution on $[0, T]$ is guaranteed. □

5.5.2 Uniqueness

It can easily be seen that the solutions obtained by the construction in Theorem 5.5.2 must be unique, because the initial solutions are unique (see Chapter 4). However, it might be possible that a different construction yields other solutions. The following theorem states that this is not the case.

Theorem 5.5.3 *Consider matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times k}$, $C \in \mathbb{R}^{k \times n}$ and $D \in \mathbb{R}^{k \times k}$ such that Assumption 5.2.2 is satisfied and (A, B, C, D) represents a passive system. Then for all initial states x_0 and all final times $T > 0$ there exists at most one solution $(u, x, y) \in \mathcal{L}_\delta^{k+n+k}[0, T]$ to LCS(A, B, C, D) in the sense of Definition 5.5.1.* □

Proof. Suppose that two solutions (u, x, y) and (u', x', y') exist in the sense of Definition 5.5.1. According to Corollary 5.4.9 there exists exactly one initial solution from the initial state x_0. This implies that the impulsive parts of (u, x, y) and (u', x', y') must be the same and moreover, that the re-initialization from x_0 must be unique such that $x(0+) = x'(0+)$. Clearly, $(u - u', x - x', y - y')$ satisfies (5.1) from initial state 0. The dissipation inequality yields

$$\int_0^t [u(\tau) - u'(\tau)]^\top [y(\tau) - y'(\tau)] d\tau \geq [x(t) - x'(t)]^\top K [x(t) - x'(t)]$$

for all $t \in (0, \infty)$. From the fact that u, u', y, y' are nonnegative almost everywhere and the complementarity of (u, y) and (u', y'), we obtain

$$\int_0^t [u(\tau) - u'(\tau)]^\top [y(\tau) - y'(\tau)] d\tau \leq 0.$$

Hence,

$$[x(t) - x'(t)]^\top K [x(t) - x'(t)] \leq 0$$

for all $t \in (0, \infty)$. Since $K > 0$, we obtain $x(t) = x'(t)$ for all t. Since B is of full column rank, this gives $u = u'$ and $y = y'$ almost everywhere. □

Since the global solution is unique, the solution must be equal to the one constructed in the proof of Theorem 5.5.2. This characterizes the nature of solutions to linear passive complementarity systems. Between mode changes the trajectories are of Bohl type and thus real-analytic. Moreover, the set $\mathcal{E}$ of mode transition times is right-isolated, meaning that for all $\tau \in \mathcal{E}$ there exists an $\alpha > 0$ such that $(\tau, \tau + \alpha) \cap \mathcal{E}$ is empty.

Remark 5.5.4 Note that the uniqueness of solutions to LCS(A, B, C, D) would not be lost, if jumps are allowed for time instants $\tau > 0$ satisfying item *1* and *2* of Definition 5.5.1. The reason is the invariance of the regular states $\mathcal{R}$, that implies that $x(t) \in \mathcal{R}$ for all $t \in (0, T]$. □

Remark 5.5.5 Since the set of mode transition times $\mathcal{E}$ is right-isolated, there do no exist left-accumulation points[1] of mode transition times. However, we cannot exclude the existence of right-accumulation points in general on the basis of this chapter. Using a result in [94] it can be proven that for a linear passive network with one diode satisfying assumption 5.2.2 and $D = 0$ also right-accumulations do not occur. □

5.6 Conclusions

Linear passive electrical circuits with ideal diodes have been studied in the context of linear complementarity systems, with the aim of establishing a rigorous base for the analysis of numerical methods for the transient simulation of switched electrical networks. Chapter 7 will deal with the question whether the solutions approximated by a time-stepping method [20,120,172] converge as a function of time to the true solution of the network model. To answer such a question, one needs of course a definition of what should be understood by the transient true solution. This question has been dealt with in this chapter and formal proofs were given for the existence and uniqueness of

[1] A point τ is called a left-accumulation point of $\mathcal{E} \subseteq \mathbb{R}$, if there exists a sequence $\{\tau_i\}_{i \in \mathbb{N}}$ such that $\tau_i > \tau$ and $lim_{i \to \infty} \tau_i = \tau$. A right-accumulation point is defined by changing ">" into "<".

solutions. Moreover, several regularity properties of the solutions have been proven of which Chapter 7 will benefit. In particular, it has been shown that derivatives of Dirac impulses do not occur, Dirac impulses happen only at the initial time instant and the set of regular states has been exactly characterized.

Networks with internally triggered switches have discrete as well as continuous characteristics. From this point of view, the chapter proposes a systematic modeling framework and a precise notion of solution for a class of networks of such a mixed nature. Systems consisting of continuous dynamics (differential equations) and switching logic are sometimes called "hybrid systems" and receive currently much attention from both control theorists [7, 145] and computer scientists [162]. Hybrid systems are encountered in various research programs ranging from switching controllers, unilaterally constrained mechanical systems, piecewise linear systems, switched electrical networks to hydraulic systems with valves. Since the underlying problems for these systems are essentially the same, all these research programs may benefit from a general theory as is currently being developed for complementarity systems.

6

Projected dynamical systems in a complementarity formalism

This chapter is based on the report [86], which has been submitted for publication in Operations Research Letters.

6.1 Introduction

In this chapter, we connect two classes of discontinuous dynamical systems. One is the class of *projected dynamical systems* introduced by Dupuis and Nagurney [62] and further developed in [147]. These systems are described by differential equations of the form

$$\dot{x}(t) = \Pi_K(x(t), -F(x(t))), \tag{6.1}$$

where F is a vector field, K is a closed convex set, and Π_K is a projection operator that prevents the solution from moving outside the constraint set K (cf. section 6.2 below for a precise definition). These systems are used for studying the behavior of oligopolistic markets, urban transportation networks, traffic networks, international trade, agricultural and energy markets (spatial price equilibria). Their stationary points can be characterized by means of variational inequalities; one may therefore say that projected dynamical systems provide a dynamic extension of variational inequalities.

We shall compare projected dynamical systems with *complementarity systems*, which may be considered as dynamical extensions of complementarity problems(cf. section 6.3). Applications of complementarity systems include (see Chapter 2) electrical networks with diodes, mechanical systems subject to unilateral constraints or Coulomb friction, control systems with relays, saturation characteristics or deadzones, variable structure systems, dynamical systems with static piecewise linear relations, hydraulic systems with one-way valves and optimal control problems with state or control

constraints. Complementarity systems are nonsmooth dynamical systems; they switch between several dynamical regimes and may show impulsive motions resulting in discontinuities of some system variables. Since complementarity systems are subject to both continuous dynamics and discrete switching, one may also consider them as a subclass of *hybrid dynamical systems* [7, 162]. Because of the nonsmoothness of trajectories, the formulation of a solution concept for complementarity systems is nontrivial (see [92, 177, 179]). Questions of (local) existence and uniqueness of solutions have been studied under various assumptions in [37, 92, 93, 123, 177, 179].

It is well known that variational inequalities and complementarity problems are closely related; see for instance [79]. It is therefore reasonable to expect that projected dynamical systems and complementarity systems are also related. In this chapter we show that there is indeed a natural relationship. Specializing to the stationary points, we obtain as a corollary the classical result which states that, under mild conditions, variational inequalities may be rewritten as mixed nonlinear complementarity problems [79, Prop. 2.2]. Moreover, we obtain a proof of existence and uniqueness of solutions of projected dynamical systems that is independent of the original proof by Dupuis and Nagurney [62] and in particular does not use the Skorokhod problem [188]. Complementarity systems have already been used extensively in the engineering literature (see for instance [124, 160, 179]) and the establishment of a relation between the domains of projected dynamical systems and of complementarity systems makes it possible to compare and transfer analytic and computational techniques between the two.

The following notational conventions and terminology will be used. If k is a positive integer, $\bar{k}$ denotes the set $\{1, \ldots, k\}$. For an index set $I \subseteq \bar{k}$, we denote its complement with respect to $\bar{k}$ by $I^c := \{i \in \bar{k} \mid i \notin I\}$. The cardinality of a set I will be denoted by $|I|$. A vector $u \in \mathbb{R}^k$ is said to be nonnegative (nonpositive) if $u_i \geq 0$ ($u_i \leq 0$) for all $i \in \bar{k}$, and in this case we write $u \geq 0$ ($u \leq 0$). Given a matrix $M \in \mathbb{R}^{k \times l}$ and subsets $I \subseteq \bar{k}$ and $J \subseteq \bar{l}$, we denote the submatrix $(M_{ij})_{i \in I, j \in J}$ by M_{IJ}. In case $I = \bar{k}$ we write $M_{\bullet J}$ rather than $M_{\bar{k}J}$, and similarly if $J = \bar{l}$ we use $M_{I\bullet}$. The transpose of a matrix M is denoted by $M^\top$. In the Euclidean space $\mathbb{R}^k$ the standard inner product is denoted by $\langle \cdot, \cdot \rangle$ and for $u, v \in \mathbb{R}^k$ we write $u \perp v$ if $\langle u, v \rangle = u^\top v = 0$. We denote the restriction of a function $f : [0, T] \to \mathbb{R}$ to an interval $(a, b) \subseteq [0, T]$ by $f|_{(a,b)}$. A function $f : \mathbb{R}^n \to \mathbb{R}^p$ will be said to be real-analytic and convex if its component functions $f_i : \mathbb{R}^n \to \mathbb{R}$ are real-analytic and convex.

6.2 Projected dynamical systems

In this section we recall the definition of projected dynamical systems (PDS) [62, 147]. The defining ingredients are a closed convex set K, which usually corresponds to the constraint set of a particular application, and a vector field F whose domain contains K. The projected dynamics is described by the equation $\dot{x}(t) = -F(x(t))$ on the interior of K, but on the boundary a modification is applied to prevent the solution

from leaving the constraint set.

To be more precise, let a closed and convex set $K \subseteq \mathbb{R}^n$ be given. The cone of inward normals at $x \in K$ is defined by

$$n(x) = \{\gamma \mid \langle \gamma, x - k \rangle \leq 0 \text{ for all } k \in K\}. \tag{6.2}$$

Note that $n(x) = \{0\}$, when x is contained in the interior of K. Given $x \in K$ and $v \in \mathbb{R}^n$, define the projection of the vector v at x with respect to K by

$$\Pi_K(x, v) = v - \langle v, n^*(x) \rangle n^*(x), \tag{6.3a}$$

where

$$n^*(x) \in \arg \max_{n \in n(x),\ \|n\| \leq 1} \langle v, -n \rangle. \tag{6.3b}$$

Note that $\Pi_K(x, v)$ is well-defined even though $n^*(x)$ may not be uniquely specified by (6.3b). The *projected dynamical system* PDS(F, K) corresponding to a closed convex set K and a vector field F on K is defined by

$$\dot{x}(t) = \Pi_K(x(t), -F(x(t))). \tag{6.4}$$

The ordinary differential equation (6.4) has a discontinuous right hand side and is therefore not covered by the standard theory of differential equations. The following notion of solution is proposed in [147].

Definition 6.2.1 [147] An absolutely continuous function $x : [0, T] \to K$ is a *solution* to PDS(F, K) on $[0, T]$ with initial state $x_0 \in K$ if $x(0) = x_0$ and (6.4) holds almost everywhere in $[0, T]$.

The definition (6.3) of the projection operator Π_K is convenient for the development below. An alternative definition is the following one. For $x \in K$ and $v \in \mathbb{R}^n$ define

$$\Pi_K(x, v) = \lim_{\delta \to 0} \frac{P_K(x + \delta v) - x}{\delta}, \tag{6.5}$$

where P_K is the projection operator that assigns to each vector x in $\mathbb{R}^n$ the vector in K that is closest to x in the Euclidean norm $\|\cdot\|$ (i. e. $P_K x = \arg\min_{k \in K} \|x - k\|$). It has been proven in [61] that the formulations in (6.3) and (6.5) are equivalent when K is convex and compact with nonempty interior. In [62] the same result is stated under the assumption that K is a convex polyhedron (i. e. an intersection of finitely many closed half-spaces). □

6.3 Complementarity systems

A complementarity system may be specified (in "semi-explicit affine form", see [177]) by functions $f : \mathbb{R}^n \to \mathbb{R}^n$, $g_i : \mathbb{R}^n \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^p$. The defining equations

for the complementarity system corresponding to f, g_i and h are

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{p} g_i(x(t))u_i(t) \tag{6.6a}$$

$$y(t) = h(x(t)) \tag{6.6b}$$

$$0 \leq y(t) \perp u(t) \geq 0 \tag{6.6c}$$

The relation (6.6c) implies that for all i at least one of the equalities $u_i(t) = 0$ and $y_i(t) = 0$ must be satisfied. Hence, for all times t there exists an index set J such that $u_i(t) = 0$, $i \notin J$ and $y_i(t) = 0$, $i \in J$. In the engineering literature this index set is sometimes called the *active index set*, *mode* or *discrete state* of the system at time t. The mode may change during the time evolution of the system. The times at which this happens are called *event times*.

In general a complementarity system may not have a continuous solution, even when the defining functions f, g and h are smooth, and so one needs to introduce larger function spaces to define solutions (cf. [92,93,177,179]). Although the solution concept below is not the most general one, it suffices for the purpose of the chapter. We need the notion of *right-isolated* sets. A subset $\mathcal{E}$ of $\mathbb{R}$ is said to be right-isolated if for each $t \in \mathcal{E}$ there exists an $\varepsilon > 0$ such that $(t, t+\varepsilon) \cap \mathcal{E} = \emptyset$.

Definition 6.3.1 A continuous function $x : [0, T] \to \mathbb{R}^n$ is called a *solution* to (6.6) with initial state x_0 on the interval $[0, T]$, if $x(0) = x_0$ and there exist a right-isolated set $\mathcal{E} \subset [0, T]$ and two functions $u : [0, T] \to \mathbb{R}^p$, $y : [0, T] \to \mathbb{R}^p$ such that for any interval $(a, b) \subseteq [0, T]$ with $(a, b) \cap \mathcal{E} = \emptyset$ the following conditions hold:

1. the restriction $(u, x, y)|_{(a,b)}$ is real-analytic and satisfies (6.6a–6.6b) for all $t \in (a, b)$;

2. there exists an index set $J \subseteq \bar{p}$ such that $u_{J^c}(t) = 0$, $y_J(t) = 0$, $u_J(t) \geq 0$ and $y_{J^c}(t) \geq 0$ for all $t \in (a, b)$.

□

This definition allows solutions that exhibit accumulations of event times ("Zeno solutions"). Since $\mathcal{E}$ is right-isolated, such accumulations only take place forward in time. Note that a similar restriction is not present in Def. 6.2.1.

By considering several types of dynamics in (6.6a–6.6b), one may define several classes of complementarity systems such as *linear* complementarity systems [92,177] and *Hamiltonian* complementarity systems [177]. For the purpose of this chapter we shall be particularly interested in *gradient-type complementarity systems*; these systems are related to the *gradient systems* that have been studied in [176]. To specify a gradient-type complementarity system, take functions $F : \mathbb{R}^n \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^p$. Let the gradients of the component functions $h_i(x)$ of $h(x)$ be denoted by $\nabla h_i(x)$ (taken to be row-vectors) and let $H(x)$ denote the matrix whose i-th row is equal to

$\nabla h_i(x)$ (i.e. the Jacobian matrix of h at x). The gradient-type complementarity system GTCS(F, h) is given by the equations (6.6):

$$\dot{x}(t) = -F(x(t)) + \sum_{i=1}^{p} [\nabla h_i(x(t))]^\top u_i(t) \tag{6.7a}$$

$$y(t) = h(x(t)) \tag{6.7b}$$

$$0 \leq y(t) \perp u(t) \geq 0 \tag{6.7c}$$

which is a special case of (6.6). Equation (6.7a) can compactly be written in terms of the Jacobian H of h as

$$\dot{x}(t) = -F(x(t)) + [H(x(t))]^\top u(t). \tag{6.8}$$

The above definition makes implicit use of the standard inner product of $\mathbb{R}^n$, but it would also be possible to use a coordinate-free treatment as in [176]. There is a closer analogy with the gradient systems studied by Van der Schaft when in (6.7) the function F is defined as the gradient of some potential function. In that case (6.7) is referred to as a *gradient complementarity system*.

6.4 Projected dynamical systems as complementarity systems

In this section we consider projected dynamical systems specified by a vector field F and a convex set K, and we provide conditions under which these systems can be rewritten as gradient-type complementarity systems. It will be assumed that the convex set K can be represented by means of finitely many inequalities.

Assumption 6.4.1 The set K allows a representation in the form

$$K = \{x \in \mathbb{R}^n \mid h(x) \geq 0\} \tag{6.9}$$

where $h : \mathbb{R}^n \to \mathbb{R}^p$ is real-analytic and convex. □

If h represents K as in (6.9), we define for $x \in K$ the *active index set* $I(x)$ as

$$I(x) := \{i \in \bar{p} \mid h_i(x) = 0\}. \tag{6.10}$$

To prevent technical complications that would obscure the main line of reasoning, we shall use the following constraint qualification in conjunction with Assumption 6.4.1.

Assumption 6.4.2 For h as in (6.9) and H the Jacobian of h, the matrix $H_{I(x)\bullet}(x)$ has full row rank for all $x \in K$. □

Concerning the vector field F, we shall use the following assumptions.

Assumption 6.4.3 The vector field F is real-analytic. □

Assumption 6.4.4 There exists a constant $B \in \mathbb{R}$ such that F satisfies the linear growth condition

$$\|F(x)\| \leq B(1 + \|x\|) \text{ for all } x \in K. \tag{6.11}$$

□

Assumption 6.4.5 There exists a constant $C \in \mathbb{R}$ such that

$$\langle -F(x) + F(y), x - y \rangle \leq C\|x - y\|^2 \quad \text{for all } x, y \in K. \tag{6.12}$$

□

Remark 6.4.6 Assumption 6.4.1 implies that K is convex and closed. A characterization of K as in (6.9) is possible in all applications of projected dynamical systems mentioned in [147]. In [62] it is even assumed that K is a convex polyhedron. Assumptions 6.4.4 and 6.4.5 are used in [147] to prove existence and uniqueness of solutions to the projected dynamical system specified by F and K. □

The following theorem is the main result of this chapter. The theorem will be proved in the next section.

Theorem 6.4.7 *Let a set $K \subseteq \mathbb{R}^n$, a vector field $F : \mathbb{R}^n \to \mathbb{R}^n$ and a function $h : \mathbb{R}^n \to \mathbb{R}^p$ be given such that Assumptions 6.4.1–6.4.5 are satisfied. For all initial states $x_0 \in K$, both the projected dynamical system PDS(F, K) and the gradient-type complementarity system GTCS(F, h) have a unique solution defined on* $[0, \infty)$. *Moreover, these solutions coincide.* □

Remark 6.4.8 It will follow from the proof given below that without Assumption 6.4.4 the theorem still holds, except that the solutions are not guaranteed to exist on $[0, \infty)$. To be specific, suppose that $[0, T_1)$ is the maximal interval on which a solution can be defined for PDS(F, K). Similarly, let $[0, T_2)$ be the maximal interval for which GTCS(F, h) admits a solution. Then $T := T_1 = T_2 > 0$, both solutions are unique on $[0, T)$, and the solutions are equal to each other. □

Remark 6.4.9 The constraint qualification Assumption 6.4.2 is introduced here for simplicity. In the literature on complementarity systems, weaker assumptions have been used. Specifically, Lötstedt [124] uses the condition that the Jacobian matrix $H(x)$ should have locally *constant* row rank to prove the existence and uniqueness of solutions to equations representing unilaterally constrained mechanical systems. □

Remark 6.4.10 Theorem 6.4.7 provides some additional information about the solutions to PDS(F, K). Under the assumptions of the theorem, solutions to projected dynamical system are real-analytic on the open intervals belonging to a set of the form $[0, \infty) \setminus \mathcal{E}$. Moreover, the exceptional set (the set of event times) $\mathcal{E}$ is a right-isolated set. □

Remark 6.4.11 It follows in particular that, under the conditions of Theorem 6.4.7, the stationary points of the projected dynamical system PDS(F, K) coincide with those of the gradient-type complementarity system GTCS(F, h). When K is a convex polyhedron, the stationary points $\bar{x}$ of PDS(F, K) are given by the variational inequality [147, Lemma 1]

$$\langle F(\bar{x}), x - \bar{x} \rangle \ \geq \ 0 \quad \forall x \in K. \tag{6.13}$$

The stationary points $\bar{x}$ of GTCS(F, h) are given by the mixed nonlinear complementarity problem

$$0 \ = \ -F(\bar{x}) + \textstyle\sum_{i=1}^{p} (\nabla h_i(\bar{x}))^\top u_i \tag{6.14a}$$

$$y \ = \ h(\bar{x}) \tag{6.14b}$$

$$0 \ \leq \ y \perp \ u \ \geq 0. \tag{6.14c}$$

In this way we recover the well-known result (see for instance [79, Prop. 2.2]) that, under a suitable constraint qualification, variational inequalities may be rewritten as mixed nonlinear complementarity problems. □

6.5 Proof of the main result

We start with a characterization of the projection Π_K in terms of a minimization problem. The proof will be given below on the basis of a duality argument.

Theorem 6.5.1 *Let $K \subset \mathbb{R}^n$ be of the form (6.9) for a real-analytic and convex function $h : \mathbb{R}^n \to \mathbb{R}^p$. For all $x \in K$ and $v \in \mathbb{R}^n$, we have*

$$\Pi_K(x, v) = \arg \min_{w \in W(x)} \|w - v\| \tag{6.15}$$

where $W(x)$ is the "cone of admissible velocities" given by

$$W(x) \ = \ \{w \in \mathbb{R}^n \mid \nabla h_i(x) w \geq 0 \text{ for all } i \in I(x)\}. \tag{6.16}$$

□

The duality result that we use to prove Theorem 6.5.1 is stated in Prop. 6.5.2 below. The notation C^0 is used to denote the polar cone (see e. g. [168, p. 121]) of a set $C \subseteq \mathbb{R}^n$:

$$C^0 = \{x \in \mathbb{R}^n \mid \langle x, y \rangle \leq 0 \text{ for all } y \in C\}. \tag{6.17}$$

Proposition 6.5.2 *Let $W \subseteq \mathbb{R}^n$ be a closed convex cone with nonempty interior and let $v \in \mathbb{R}^n$ be given. Define w^* by*

$$w^* = \arg\min_{w \in W} \|w - v\| \tag{6.18}$$

and let z^* be such that

$$z^* \in \arg\max_{z \in W^0,\, \|z\| \leq 1} \langle v, z \rangle. \tag{6.19}$$

Then

$$w^* = v - \langle v, z^* \rangle z^*. \tag{6.20}$$

□

Proof. We apply the Fenchel duality theorem [127, p. 201] to the convex function $f(w) := \|w - v\|$ defined on $C := \mathbb{R}^n$ and the concave function $g(w) := 0$ defined on $D := W$. One easily computes (cf. for instance [168, Section 12]) that the conjugate sets of C and D are $C^* = \{z \in \mathbb{R}^n \mid \|z\| \leq 1\}$ and $D^* = -W^0$, and that the conjugate functions of f and g are given by $f^*(z) = \langle v, z \rangle$ for $z \in C^*$ and $g^*(z) = 0$ for $z \in D^*$. From the Fenchel duality theorem, we therefore have

$$\min_{w \in W} \|w - v\| = \max_{z \in W^0,\, \|z\| \leq 1} \langle v, z \rangle. \tag{6.21}$$

Now, suppose first that $\min_{w \in W} \|w - v\| > 0$; then $\|z^*\| = 1$. In this case, there exists a real number $\alpha \geq 0$ such that $w^* - v = -\alpha z^*$ [127, p. 136]. We have $-\alpha = -\|\alpha z^*\| = -\|w^* - v\| = -\langle v, z^* \rangle$ by (6.21); this proves (6.20). Next, suppose that $\min_{w \in W} \|w - v\| = 0$. Then $v \in W$ and hence $w^* = v$. We have $\max_{z \in W^0, \|z\| \leq 1} \langle v, z \rangle = 0$, so that $\langle v, z^* \rangle = 0$ and consequently equation (6.20) is also correct in this case. □

Remark 6.5.3 The proof implies that $\langle w^*, z^* \rangle = 0$. Together with the conditions $w^* \in W$, $z^* \in W^0$, and $v = w^* + \langle v, z^* \rangle z^*$, this shows that $\langle v, z^* \rangle z^*$ is actually the projection $P_{W^0} v$ of v onto the cone W^0 [141, p. 238]. □

Proof of Theorem 6.5.1 Fix an arbitrary $x \in K$. From [168, Cor. 23.7.1, 23.8.1] it follows that the cone of inward normals of K at x, denoted by $n(x)$, and the cone of inward normals of $W(x)$ at 0, denoted by $n_{W(x)}(0)$ satisfy

$$n(x) = n_{W(x)}(0) = \{\gamma \in \mathbb{R}^n \mid \gamma = \sum_{i \in I(x)} [\nabla h_i(x)]^\top \lambda_i \text{ for certain } \lambda_i \geq 0\}. \tag{6.22}$$

By definition of the cone of inward normals and the polar cone (see (6.2) and (6.17)), $n_{W(x)}(0)$ is equal to $-W(x)^0$. Hence, $n(x) = -W(x)^0$. The claim now follows immediately by applying Prop. 6.5.2 to $W = W(x)$ and using that $W^0 = -n(x)$. □

Next we establish a connection to a linear complementarity problem (LCP). See [47] for an extensive treatment of LCPs.

Theorem 6.5.4 *Let a subset K of $\mathbb{R}^n$ be of the form (6.9) for a real-analytic and convex function $h : \mathbb{R}^n \to \mathbb{R}^p$. Fix $x \in K$. Let H be the Jacobian matrix of h at x, and let $I := I(x) = \{i \mid h_i(x) = 0\}$ be the active index set. Then we have*

$$\Pi_K(x, v) \;=\; v + H_{I\bullet}^{\top} u \tag{6.23a}$$

where the vector $u \in \mathbb{R}^{|I|}$ solves the LCP

$$0 \;\le\; u \perp H_{I\bullet} v + H_{I\bullet}[H_{I\bullet}]^{\top} u \;\ge\; 0. \tag{6.23b}$$

□

Proof. By Theorem 6.5.1, the vector $\Pi_K(x, v)$ is the projection of v onto the cone $W(x)$ defined in (6.16). In terms of the notation introduced in the statement of the theorem, we have

$$W(x) \;=\; \{w \in \mathbb{R}^n \mid H_{I\bullet} w \ge 0\}. \tag{6.24}$$

The fact that the projection onto this cone can be found from (6.23) is well-known; one may for instance use the Kuhn-Tucker conditions. An alternative approach is to use the result by Moreau [141] which states that in order to compute the projection of a vector v in a Hilbert space on a closed cone W, it is enough to find w and w^0 such that $v = w + w^0$, $w \in W$, $w^0 \in W^0$, and $w \perp w^0$; the projection $P_W v$ is then given by w. In our case $W(x)$ is given by (6.24) so that the polar cone $W^0(x)$ can be written as

$$W^0(x) \;=\; \{w^0 \in \mathbb{R}^n \mid w^0 = -[H_{I\bullet}]^{\top} u \text{ for some } u \ge 0\}. \tag{6.25}$$

Therefore the three conditions of the LCP (6.23b) are exactly the conditions that ensure, by Moreau's theorem, that $\Pi_K(x, v)$ is given by (6.23). Note in particular that the condition $[H_{I\bullet}]^{\top} u \perp v + H_{I\bullet}^{\top} u$ is equivalent to $u \perp H_{I\bullet} v + H_{I\bullet}[H_{I\bullet}]^{\top} u$. □

The discussion so far may be summarized as follows.

Corollary 6.5.5 *A function $x : [0, T] \to \mathbb{R}^n$ is a solution to the projected dynamical system (6.4) if and only if there exists a locally integrable function $u : [0, T] \to \mathbb{R}^p$ such that, with $I(x)$ the active index set as in (6.10) and $H(x)$ the Jacobian matrix of*

h at $x \in K$, one has for almost all $t \in [0, T]$:

$$\dot{x}(t) \;=\; -F(x(t)) + [H_{I(x(t))\bullet}(x(t))]^{\top} u_{I(x(t))}(t) \tag{6.26a}$$

$$u_{I(x(t))^c}(t) \;=\; 0 \tag{6.26b}$$

$$0 \;\leq\; u_{I(x(t))}(t) \;\perp\; -H_{I(x(t))\bullet}(x(t))F(x(t)) + H_{I(x(t))\bullet}(x(t))[H_{I(x(t))\bullet}(x(t))]^{\top} u_{I(x(t))}(t) \;\geq\; 0. \tag{6.26c}$$

□

In the proof of the main theorem we shall use the following result, which can easily be derived from Theorem 3.2 in [179]. The quoted theorem gives a local existence and uniqueness result for complementarity systems of the form (6.6).

Theorem 6.5.6 *Let real-analytic functions $F : \mathbb{R}^n \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ be given. Take $x_0 \in \mathbb{R}^n$ such that $h(x_0) \geq 0$. If Assumption 6.4.2 is satisfied, then there exists an $\varepsilon > 0$ such that GTCS(F, h) has a solution x on $[0, \varepsilon)$ with initial condition x_0. Moreover, this solution is unique.* □

Proof. Define $I = I(x_0)$ as in (6.10) and apply Theorem 3.2 in [179] to the system GTCS(F, h_I), *i.e.* $\dot{x}(t) = -F(x(t)) + [H_{I\bullet}(x(t))]^{\top} u_I(t)$ and $0 \leq h_I(x(t)) \perp u_I(t) \geq 0$ with $I = I(x_0)$. Since $h_i(x_0) > 0$ for $i \notin I(x_0)$, it is clear that continuous solutions to GTCS(F, h_I) with initial state x_0 are solutions to GTCS(F, h) for sufficiently small t, and vice versa.

Note that $H_{I\bullet}(x_0)[H_{I\bullet}(x_0)]^{\top}$ is positive definite due to Assumption 6.4.2 and hence, is also a P-matrix (*i.e.* has only positive principal minors) [47, Thm. 3.1.6 and Thm. 3.3.7]. Consequently, Theorem 3.2 in [179] applies to GTCS(F, h_I) and the result follows. □

Now we are in a position to prove the main result of this chapter.

Proof of Theorem 6.4.7 Take $x_0 \in K$. According to Theorem 6.5.6 there exists a real-analytic triple (u, x, y) that satisfies (6.7) on $[0, \varepsilon)$. In particular, there exists an index set $J \subseteq \bar{p}$ such that $y_J(t) = 0$ and $u_{J^c}(t) = 0$ for all $t \in [0, \varepsilon)$.

We now want to show that the trajectory x that has been defined in this way on $[0, \varepsilon)$ is also a solution to PDS(F, K) on $[0, \varepsilon)$. It is immediately clear that (6.26a) is satisfied because it is just another way of writing (6.8). For $x \in K$, define $I(x)$ as in (6.10). From the fact that $y_J(t) = 0$ on $[0, \varepsilon)$ it follows that $J \subseteq I(x(t))$ for $t \in [0, \varepsilon)$. Therefore $I(x(t))^c \subseteq J^c$ and so $u_{I(x(t))^c}(t) = 0$ for $t \in [0, \varepsilon)$. Hence, (6.26b) is satisfied. It remains to show that $u_{I(x(t))}(t)$ satisfies the LCP (6.26c) on $[0, \varepsilon)$. It is clear from (6.7c) that the inequality $u_{I(x(t))}(t) \geq 0$ is satisfied on $[0, \varepsilon)$. For $t \in [0, \varepsilon)$, we have

$$0 \;=\; \dot{y}_J(t) = -H_{J\bullet}(x(t))F(x(t)) + H_{J\bullet}(x(t))[H_{J\bullet}(x(t))]^{\top} u_J(t). \tag{6.27}$$

Dropping all arguments now to lighten the notation, we have from $u_{J^c} = 0$ and $J \subseteq I$ that

$$(H_{I\bullet}[H_{I\bullet}]^\top u_I)_J \;=\; H_{J\bullet}[H_{J\bullet}]^\top u_J. \tag{6.28}$$

Since obviously $(H_{I\bullet}F)_J = H_{J\bullet}F$, it follows from (6.27) and from $u_{J^c} = 0$ that the orthogonality condition in (6.26c) holds. The final inequality in (6.26c) follows by expressing $\dot{y}_i(t)$ similarly to (6.27), and noting that $\dot{y}_i(t) \geq 0$ whenever $y_i(t) = 0$ (i. e. whenever $i \in I(x(t))$), because otherwise the inequality $y_i(t) \geq 0$ on $[0, \varepsilon)$ would be violated.

If the limit $\lim_{t\uparrow\varepsilon} x(t) =: x(\varepsilon)$ exists, the existence of a solution to (6.7) starting from $x(\varepsilon)$ on $[\varepsilon, \varepsilon+\varepsilon_1)$ for some $\varepsilon_1 > 0$ follows from Theorem 6.5.6. Hence, we have a solution (x, u, y) to (6.7) on $[0, \varepsilon + \varepsilon_1)$ in the sense of Definition 6.3.1. In the same way as above, it can be shown that x is a solution of PDS(F, K) on $[0, \varepsilon + \varepsilon_1)$.

We now have to show that actually a solution to GTCS(F, h) can be constructed on all of $[0, \infty)$. In principle it might happen that the above construction only leads to a solution on some interval $[0, T)$ with $T < \infty$. To proceed by contradiction, assume that we are in such a situation. The following estimates hold for $0 \leq t \leq T$:

$$\begin{aligned}
\|x(t)\| &\leq \|x_0\| + \int_0^t \|\Pi_K(x(\tau), -F(x(\tau)))\| d\tau \\
&\leq \|x_0\| + \int_0^t \|F(x(\tau))\| d\tau \\
&\leq \|x_0\| + BT + B\int_0^t \|x(\tau)\| d\tau.
\end{aligned}$$

The second step follows easily from the definition of Π_K (see [147, Eq. (2.19)]) and the third inequality is a consequence of (6.11). Using Gronwall's lemma we see from this that $x(\cdot)$ is bounded on $[0, T)$; say $\|x(t)\| \leq M$ for $t \in [0, T)$ for some constant $M > 0$. It follows in particular that no "finite escape time" can occur. Moreover, it follows that the solution x is Lipschitz continuous and hence uniformly continuous on $[0, T)$. Indeed, for $0 \leq t < s < T$ we have

$$\begin{aligned}
\|x(t) - x(s)\| &\leq \int_t^s \|\Pi_K(x(\tau), -F(x(\tau)))\| d\tau \\
&\leq \int_t^s \|F(x(\tau))\| d\tau \\
&\leq B\int_t^s (1 + \|x(\tau)\|) d\tau \\
&\leq B(1 + M)(s - t).
\end{aligned}$$

By a standard result in analysis (see for instance [169, Exc. 4.13]) this implies that the limit $x(T) := \lim_{t\uparrow T} x(t)$ exists. Since by continuity arguments $h(x(T)) \geq 0$, continuation is possible beyond T according to Theorem 6.5.6, and we have reached

a contradiction. Therefore, it follows that there is a unique solution of the gradient-type complementarity system GTCS(F, h) on $[0, \infty)$ which is also a solution of the projected dynamical system PDS(F, K). The uniqueness of solutions to PDS(F, K) follows from Assumption 6.4.5 as in [147, p. 33]. □

Remark 6.5.7 The existence of solutions to PDS(F, K) on $[0, \infty)$ is shown in [147] by a method based on the Skorokhod Problem [188]. The proof above provides an alternative argument. In fact the proof shows that Assumptions 6.4.1–6.4.3 are sufficient for *local* existence of solutions to PDS(F, K). With the additional Assumption 6.4.4, one can prove existence on $[0, \infty)$. The argument to prove uniqueness uses Assumption 6.4.5 and is essentially due to Filippov [67]. □

6.6 Conclusions

We have shown that, under mild conditions, projected dynamical systems can be rewritten as gradient-type complementarity systems. This result may be looked at as a dynamic version of the well known fact that, under suitable conditions, variational inequalities may be rewritten as mixed nonlinear complementarity problems. The class of gradient-type complementarity systems is a subclass of the class of complementarity systems which has received a considerable amount of attention in the engineering and applied physics literature. The establishment of a connection between the domains of projected dynamical systems and complementarity systems facilitates the transfer of techniques from one domain to the other. As an interesting bonus, we have obtained a new, and in the authors' opinion more direct, proof for the existence of solutions to projected dynamical systems.

7

Consistency of a time-stepping method

This chapter is mainly based on the paper [35], which is submitted for publication. Kanat Çamlıbel acted as one of my co-authors in this paper, and this chapter is also part of his PhD-work. In the appendix of this chapter, we added a treatise on the use of time-stepping methods for linear complementarity system with low leading row coefficients. This appendix does not appear in the paper [35], but is closely related to the material of [35].

7.1 Introduction

This chapter continues the work presented in Chapter 5 in the direction of transient simulation of electrical networks with ideal diodes. In particular, we will be interested in the time-stepping method that is based on the well-known backward Euler integration routine [71], which has already been applied for the numerical approximation of electrical networks [20, 120, 121] and unilaterally constrained mechanical systems [125,140,155,192,194]. The advantages of the method are that it is straightforward to implement and many algorithms (e.g. Lemke's algorithm [47], Katzenelson's algorithm [109] and others [121]) are available to solve the one-step problems consisting of linear complementarity problems (LCPs).

In [120] the use of a time-stepping method based on backward Euler (or higher order linear multistep integration methods [71] like the trapezoidal rule) has been proposed also for the class of general linear complementarity systems, i.e. linear time-invariant dynamical systems coupled with ideal diode characteristics (complementarity conditions). By an example (cf. Example 7.3.3 below), it will be shown that the method is not suited for any arbitrary linear complementarity system. This example indicates, that although the method has proven itself in practice, one should not indiscriminately apply it to general dynamical systems with mixed continuous and discrete dynamics.

A justification of the numerical scheme in the sense of showing convergence of the approximating time functions to a true solution of the dynamical system seems required considering the example mentioned above. The importance of such a rigorous validation is also stressed by considering the problems that might occur due to changing configurations of the network, the possibility of Dirac impulses and the discontinuities of the system's variables.

Convergence problems of time-stepping methods for mechanical systems subject to unilateral constraints or friction have been studied by Stewart [192, 193]. He shows that for a broad class of nonlinear constrained mechanical systems there always exists some sequence of approximating time functions that converge to a true solution of the mechanical model. However, the convergence of the complete sequence has not been shown in [192, 193]. The conditions used in [192, 193] (oriented towards mechanical systems) do not cover electrical networks containing ideal diodes, which will be the subject of this chapter. Specifically, we will show that for the class of linear electrical passive circuits with ideal diodes, the 'backward Euler time-stepping method' is consistent. Consistency indicates that for any arbitrary (so not only a special sequence) sequence of time steps, which tends to zero, the corresponding approximations converge to the true transient solution of the network model. Using the same arguments, we will also show that the real transient solutions depend continuously on the initial states. Of course, this is a covenient property for simulation, since small numerical erros will not have a large influence on the outcome of the algorithm.

Although the results are written down for networks containing ideal diodes (internally controlled switches) only, externally controlled switches can easily be included without destroying the convergence proof. The results presented here form a justification of the 'backward Euler time-stepping scheme' in the field of switched electrical networks.

The outline of the chapter is as follows. In section 7.2 the preliminaries on linear complementarity systems and passivity are stated. The time-stepping method that will be studied is introduced in section 7.3. Moreover, a fairly general result on consistency of the numerical method is formulated for linear complementarity systems. In the next section, this result is applied to linear passive complementarity systems, i.e. passive linear systems coupled to ideal diode characteristics. The continuous dependence of solution trajectories on the initial states is mentioned in section 7.4 as well. The conclusions follow in section 7.5. The proofs of the results can be found in section 7.6, after which an additional consistency result is given in the appendix of this chapter. Specifically, for a class of LCS with leading row coefficients equal to zero or one, the existence is proven of a (special) sequence of time steps for which the corresponding approximations converge to a true solution of the model. The proof is based on the fairly general result presented in section 7.3.

Throughout the chapter, $\mathbb{R}$ ($\mathbb{R}^n$) denotes the set of (n-tuples of) real numbers and $\mathbb{R}_+$ the set of nonnegative real numbers, i.e. $\mathbb{R}_+ = [0, \infty)$. For any $x, y \in \mathbb{R}^n$, $x \perp y$ means that $x^\top y = 0$. Inequalities for vectors are always meant to hold componentwise. The Euclidean and maximum norm of a vector $x \in \mathbb{R}^n$ will be denoted by $\|x\| :=$

$\sqrt{\sum_{i=1}^{n} x_i^2}$ and $\|x\|_\infty := \max_{i\in\bar{n}} |x_i|$, respectively. For a positive integer n, $\overline{n}$ denotes the set $\{1, 2, \ldots, n\}$. For a real number $r \in \mathbb{R}$, we use the notation $\lceil r \rceil$ to denote the smallest integer larger than or equal to r. The set of real matrices with n rows and m columns is denoted by $\mathbb{R}^{n\mathrm{x}m}$. For any $A \in \mathbb{R}^{n\mathrm{x}m}$, $J \subseteq \overline{n}$, and $K \subseteq \overline{m}$, A_{JK} denotes the submatrix obtained by taking the rows corresponding to the elements of J and columns corresponding to the elements of K. If $J = \overline{n}$ ($K = \overline{m}$), we also write $A_{\bullet K}$ ($A_{J\bullet}$). For any $A \in \mathbb{R}^{n\mathrm{x}m}$ $\|A\| := \sup_{\|x\|=1} \|Ax\|$ denotes the matrix norm induced by the Euclidean vector norm. A square matrix $A \in \mathbb{R}^{n\mathrm{x}n}$ is said to be nonnegative (positive) definite if $x^\top Ax \geq 0$ ($x^\top Ax > 0$) for all $0 \neq x \in \mathbb{R}^n$. We write $\sigma(A)$ for the set of eigenvalues of A and $\rho(A) := \max_{\lambda\in\sigma(A)} |\lambda|$ for the spectral radius of A. By the symmetric part of A, we mean the matrix $\frac{1}{2}(A + A^\top)$. The identity matrix is denoted by I. The set of n-tuples of square integrable functions on (t_0, t_1) is denoted by $\mathcal{L}_2^n(t_0, t_1)$. The notation $\langle x, y\rangle$ denotes the inner product of $x, y \in \mathcal{L}_2^n(t_0, t_1)$, i.e. $\langle x, y\rangle = \int_{t_0}^{t_1} x^\top(t)y(t)dt$. The norm on $\mathcal{L}_2^n(t_0, t_1)$ is defined by $\|x\| = \langle x, x\rangle^{1/2}$. Moreover, $x|_\Omega$ denotes the restriction of x to the interval Ω. We say that the sequence $\{x_k\} \subset \mathcal{L}_2^n(t_0, t_1)$ *converges* (*weakly converges*) to x if $\lim_{k\to\infty} \|x_k - x\| = 0$ ($\lim_{k\to\infty} \langle x_k - x, y\rangle = 0$ for all $y \in \mathcal{L}_2^n(t_0, t_1)$). The matrix triple (A, B, C) with $A \in \mathbb{R}^{n\mathrm{x}n}$, $B \in \mathbb{R}^{n\mathrm{x}m}$ and $C \in \mathbb{R}^{m\mathrm{x}n}$ is said to be *minimal*, when $\operatorname{rank}\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = n$ and $\operatorname{rank}\begin{bmatrix} C^\top & C^\top A^\top & \cdots & C^\top (A^\top)^{n-1} \end{bmatrix} = n$.

7.2 Preliminaries

We begin by briefly recalling the *linear complementarity problem* (LCP) of mathematical programming. For an extensive survey on the problem, the reader is referred to [47].

Problem 7.2.1 (LCP(q, M)) Given $q \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n\mathrm{x}n}$, find $z \in \mathbb{R}^n$ such that

$$z \geq 0 \tag{7.1a}$$

$$q + Mz \geq 0 \tag{7.1b}$$

$$z^\top(q + Mz) = 0 \tag{7.1c}$$

□

We say that z *solves* LCP(q, M) if z satisfies (7.1). The set of all solutions of LCP(q, M) will be denoted by SOL(q, M). Sometimes we also say that (z, w) is a solution to LCP(q, M), when z satisfies (7.1) and $w = q + Mz$. Note that the so-called complementarity conditions (7.1) are similar to the ideal diode characteristic $v \leq 0$, $i \geq 0$, and $iv = 0$. Not surprisingly, the linear complementarity problem plays a major role in the analysis of the networks with ideal diodes. Indeed, as discussed in chapter 5, linear networks with ideal diodes can be modeled as linear complementarity systems,

which are dynamical versions of the linear complementarity problem, of the form

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{7.2a}$$
$$y(t) = Cx(t) + Du(t) \tag{7.2b}$$
$$0 \leq u(t) \perp y(t) \geq 0, \tag{7.2c}$$

where $u(t) \in \mathbb{R}^m, x(t) \in \mathbb{R}^n, y(t) \in \mathbb{R}^m$ and A, B, C, and D are matrices of appropriate dimensions. We denote (7.2) by LCS(A, B, C, D) and associate to (A, B, C, D) the transfer matrix $G(s) = C(sI - A)^{-1}B + D$.

Before precisely defining the solution concept of LCS(A, B, C, D), we need to mention several spaces of functions and distributions, which play a crucial role in the sequel. The space $\mathcal{B}$ denotes the space of Bohl functions, i.e. functions having rational Laplace transforms. The space $\mathcal{B}_\delta$ consists of the distributions of the form $u = u_{imp} + u_{reg}$, where $u_{imp} = u_0\delta$ is called the *impulsive part* with $u_0 \in \mathbb{R}$ and $u_{reg} \in \mathcal{B}$ is called the *regular part*. A distribution $u \in \mathcal{B}_\delta^n$ is said to be *initially nonnegative*, if its Laplace transform $\hat{u}(s)$ satisfies $\hat{u}(\sigma) \geq 0$ for all sufficiently large $\sigma \in \mathbb{R}$. In a similar fashion, the space $\mathcal{L}_\delta(0, \tau)$ consists of the distributions of the form $u = u_{imp} + u_{reg}$ where $u_{imp} = u_0\delta$ is called the *impulsive part* with $u_0 \in \mathbb{R}$ and $u_{reg} \in \mathcal{L}_2(0, \tau)$ is called the *regular part*. We say that the sequence of distributions $\{u_0^k\delta + u_{reg}^k\} \subset \mathcal{L}_\delta(0, \tau)$ converges (weakly) to $u_0\delta + u_{reg}$, if $\{u_0^k\}$ converges to u_0 and $\{u_{reg}\}$ converges (weakly) to u_{reg} in $\mathcal{L}_2$ sense.

Next, we remind the notion of *initial solution* which has a considerable importance in the analysis of linear complementarity systems.

Definition 7.2.2 $(u, x, y) \in \mathcal{B}_\delta^{m+n+m}$ is an *initial solution*[1] of LCS(A, B, C, D) with initial state x_0, if there exists an index set $J \subseteq \overline{m}$ such that

$$\dot{x} = Ax + Bu + x_0\delta$$
$$y = Cx + Du$$
$$u_i = 0 \text{ if } i \in J$$
$$y_i = 0 \text{ if } i \notin J$$

hold in the distributional sense, and u and y are initially nonnegative. □

It can be shown that there is a one-to-one relation between the initial solutions to LCS(A, B, C, D) with initial state x_0 and the *proper* solutions of the so-called *rational complementarity problem*.

Problem 7.2.3 (RCP(x_0, A, B, C, D)) Given $x_0 \in \mathbb{R}^n$ and (A, B, C, D) with $A \in \mathbb{R}^{nxn}$, $B \in \mathbb{R}^{nxm}$, $C \in \mathbb{R}^{mxn}$ and $D \in \mathbb{R}^{mxm}$, find $\hat{u}(s) \in \mathbb{R}^m(s)$ and $\hat{y}(s) \in \mathbb{R}^m(s)$

[1] Note that the definition of initial solutions as formulated here is more restrictive than the one used in chapter 5 as it only allows the Dirac distribution in its impulsive part and not its derivatives. However, this notion of initial solution suffices for the purposes of the chapter. Note that in chapter 5 it is proven, that the notion as stated here is not restrictive for passive LCS satisfying a full rank and a minimality condition.

such that

$$\hat{y}(s) = C(sI - A)^{-1}x_0 + [C(sI - A)^{-1}B + D]\hat{u}(s)$$
$$\hat{u}(s) \perp \hat{y}(s)$$

for all $s \in \mathbb{C}$ and $\hat{u}(\sigma) \geq 0$ and $\hat{y}(\sigma) \geq 0$ for all sufficiently large $\sigma \in \mathbb{R}$. □

The following proposition states the above mentioned one-to-one relation, which is given by the Laplace transform and its inverse. This connection indicates the relevance of the rational complementarity problem to the study of LCS.

Proposition 7.2.4 *(u, x, y) is an initial solution of LCS(A, B, C, D) with initial state x_0 if and only if its Laplace transform $(\hat{u}(s), \hat{x}(s), \hat{y}(s))$ is such that $(\hat{u}(s), \hat{y}(s))$ is a proper solution of RCP(x_0, A, B, C, D) and $\hat{x}(s) = (sI-A)^{-1}x_0+(sI-A)^{-1}B\hat{u}(s)$.* □

Now, we can give a precise definition of what is meant by a (global) solution of LCS(A, B, C, D).

Definition 7.2.5 We call the triple $(u, x, y) \in \mathcal{L}_\delta^{m+n+m}(0, \tau)$ a *(global) solution* to LCS(A, B, C, D) on $[0, \tau]$ with initial state x_0, if

1. There exists an initial solution $(\bar{u}, \bar{x}, \bar{y})$ such that

$$(u_{imp}, x_{imp}, y_{imp}) = (\bar{u}_{imp}, \bar{x}_{imp}, \bar{y}_{imp})$$

2. The equations

$$\dot{x} = Ax + Bu + x_0\delta$$
$$y = Cx + Du$$

hold in the distributional sense.

3. For almost all $t \in [0, \tau]$, $0 \leq u_{reg}(t) \perp y_{reg}(t) \geq 0$.

□

Notice that the above definition is just a restatement of the one given in chapter 5 in terms of distributions.

The first item in the definition 7.2.5 imposes a relation between the impulsive part and the rest of the solution. In the following example, we illustrate the necessity of such a connection.

Example 7.2.6 Consider the simple circuit depicted in the figure 7.1. By denoting the voltage across the capacitor and the diode by v_c and v_d, respectively and the current through the diode by i_d, one can obtain circuit equations as □

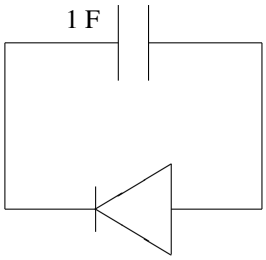

Figure 7.1:

$$\dot{v}_c = -i_d$$
$$v_d = v_c$$
$$0 \geq v_d \perp i_d \geq 0.$$

It can be rewritten in the form of a linear complementarity system as

$$\dot{x} = u \tag{7.3a}$$
$$y = x \tag{7.3b}$$
$$0 \leq u \perp y \geq 0, \tag{7.3c}$$

where $u = i_d$, $x = -v_c$, and $y = -v_d$. For the initial state $x_0 = -1$, the triple $(u, x, y) = (a\delta, a-1, a-1)$ with $a \geq 1$ satisfies the last two items of definition 7.2.5. However, from a physical point of view $(a\delta, a-1, a-1)$ is only a solution for initial state $x_0 = -1$ in case $a = 1$, since this is the only situation complying with the circuit under study (an instantaneous and complete discharge of the capacitor). Note that $(u, x, y) = (\delta, 0, 0)$ is indeed the unique initial solution.

In the sequel, we confine ourselves to linear passive complementarity systems. To be reasonably self-contained, we shall quickly review the notion of passivity and its characterizations in terms of the state representation and the transfer matrix of the system.

Definition 7.2.7 [206] The system (A, B, C, D) given by (7.2a)-(7.2b) is said to be *passive (dissipative with respect to the supply rate $u^\top y$)* if there exists a function $V : \mathbb{R}^n \to \mathbb{R}_+$, called *storage function*, such that

$$V(x(t_0)) + \int_{t_0}^{t_1} u^\top(t) y(t) dt \geq V(x(t_1))$$

holds for all t_0 and t_1 with $t_1 \geq t_0$, and all $(u, x, y) \in \mathcal{L}_2^{m+n+m}(t_0, t_1)$ satisfying (7.2a)-(7.2b). □

We state a well-known theorem on passive systems which is sometimes called the positive real lemma.

Lemma 7.2.8 [206] *Assume that* (A, B, C) *is minimal. Then the following statements are equivalent:*

1. (A, B, C, D) *is passive.*
2. *The matrix inequalities*

$$K = K^\top \geq 0 \text{ and } \begin{bmatrix} A^\top K + KA & KB - C^\top \\ B^\top K - C & -(D + D^\top) \end{bmatrix} \leq 0$$

   have a solution.

3. $G(s)$ *is positive real, i.e.,* $G(\lambda) + G^\top(\bar{\lambda}) \geq 0$ *for all* $\lambda \in \mathbb{C}$ *with* $\lambda \notin \sigma(A)$ *and* $\text{Re } \lambda > 0$.

Moreover, if (A, B, C, D) *is passive all solutions* K *of the linear matrix inequalities in item 2 are positive definite.* □

Throughout the chapter, we will frequently use the following assumption.

Assumption 7.2.9 (A, B, C) is a minimal representation and B is of full column rank. □

The proof of the following theorem can be found in chapter 5 and deals with the existence and uniqueness of solutions to linear passive complementarity systems.

Theorem 7.2.10 *Suppose that* (A, B, C, D) *is such that assumption 7.2.9 holds and* (A, B, C, D) *is passive. Let* $\tau > 0$ *be given. For each* x_0*, there exists a unique solution* $(u, x, y) \in \mathcal{L}_\delta^{m+n+m}(0, \tau)$ *of LCS*(A, B, C, D) *on* $[0, \tau]$ *with initial state* x_0. □

7.3 The backward Euler time-stepping method

For the numerical approximation of the solutions of switched electrical networks the following time-stepping scheme has been used frequently [20, 120, 121]. For LCS the method consists of discretizing the system description by applying the well known backward Euler integration routine and imposing the complementarity conditions at every time step. This comes down to the computation of u_{k+1}^h, y_{k+1}^h, and x_{k+1}^h given x_k^h through the linear complementarity problem given by

$$\frac{\mathrm{x}_{k+1}^h - \mathrm{x}_k^h}{h} = A\mathrm{x}_{k+1}^h + B\mathrm{u}_{k+1}^h \tag{7.4a}$$

$$\mathrm{y}_{k+1}^h = C\mathrm{x}_{k+1}^h + D\mathrm{u}_{k+1}^h \tag{7.4b}$$

$$0 \leq \mathrm{y}_{k+1}^h \perp \mathrm{u}_{k+1}^h \geq 0 \tag{7.4c}$$

Here $\bullet_k^h$ denotes the value at the kth step of the corresponding variable for the fixed step size $h > 0$. Based on this scheme, one can construct approximations of the transient response of a LCS by applying the algorithm below.

Algorithm 7.3.1 $(\{\mathrm{u}_k^h\}, \{\mathrm{x}_k^h\}, \{\mathrm{y}_k^h\})$ =Approx. $(A, B, C, D, \tau, h, x_0)$

1. $N_h = \lceil \frac{\tau}{h} \rceil$
2. $\mathrm{x}_{-1}^h := x_0$
3. $k := -1$
4. solve the *one-step problem*

$$\mathrm{y}_{k+1}^h = C(I - hA)^{-1}\mathrm{x}_k^h + [D + hC(I - hA)^{-1}B]\mathrm{u}_{k+1}^h$$
$$0 \le \mathrm{u}_{k+1}^h \perp \mathrm{y}_{k+1}^h \ge 0$$

   for the variables u_{k+1}^h and y_{k+1}^h
5. $\mathrm{x}_{k+1}^h := (I - hA)^{-1}\mathrm{x}_k^h + h(I - hA)^{-1}B\mathrm{u}_{k+1}^h$
6. $k := k + 1$
7. if $k < N_h$ goto 4
8. stop.

□

The one-step problem is given by a linear complementarity problem in step 4. In general, the LCP may have multiple solutions or may have no solution at all. We shall proceed by assuming unique solvability of the problem. The assumption is introduced here for reasons of generality. Later on we will prove that the assumption is implied by passivity.

Assumption 7.3.2 For all sufficiently small $h > 0$, LCP$(hC(I - hA)^{-1}\bar{x}, G(h^{-1}))$ has a unique solution for all $\bar{x}$, where $G(h^{-1})$ is given by $hC(I - hA)^{-1}B + D$. □

This assumption implies that for all sufficiently small $h > 0$, algorithm 7.3.1 generates an output, which is unique. Hence, for a given initial state x_0 and step size $h > 0$ (sufficiently small), we can define the approximations (u^h, x^h, y^h) given by

$$u_{imp}^h = h\mathrm{u}_0^h\delta \tag{7.5a}$$
$$x_{imp}^h = h\mathrm{x}_0^h\delta \tag{7.5b}$$
$$y_{imp}^h = h\mathrm{y}_0^h\delta \tag{7.5c}$$

$$\left.\begin{array}{l} u_{reg}^h(t) = \mathrm{u}_l^h \\ x_{reg}^h(t) = \mathrm{x}_l^h \\ y_{reg}^h(t) = \mathrm{y}_l^h \end{array}\right\} \quad \text{whenever } (l-1)h \le t < lh, \tag{7.5d}$$

where u_k^h, x_k^h and y_k^h, $k = 0, 1, \ldots, N_h$ have been obtained from algorithm 7.3.1. The most important goal of the chapter is to prove that for a passive system these approximations converge in a suitable sense to the actual solution of the system. This will be called *consistency* of the numerical method. In the following example, we illustrate algorithm 7.3.1 is not always consistent even if assumption 7.3.2 holds.

Example 7.3.3 Consider the linear complementarity system (consisting of a triple integrator with complementarity conditions)

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= u \\ y &= x_1 \\ 0 \leq u \perp y &\geq 0 \end{aligned}$$

with the initial state $x_0 = \begin{pmatrix} 0 & -1 & 0 \end{pmatrix}^\top$. It can be easily calculated that $(I - hA)^{-1} = \begin{pmatrix} 1 & h & h^2 \\ 0 & 1 & h \\ 0 & 0 & 1 \end{pmatrix}$ and $G(h^{-1}) = h^3$. By solving the one-step problem for $k = -1$ ($x_{-1}^h = (0 \ -1 \ 0)^\top$)

$$\begin{aligned} \mathrm{y}_0^h &= -h + h^3 \mathrm{u}_0^h \\ 0 \leq \mathrm{u}_0^h &\perp \mathrm{y}_0^h \geq 0, \end{aligned}$$

we get $(\mathrm{u}_0^h, \mathrm{y}_0^h) = (h^{-2}, 0)$. Hence, $\mathrm{x}_0^h = \begin{pmatrix} 0 & 0 & h^{-1} \end{pmatrix}^\top$. For $k = 0$, the one-step problem

$$\begin{aligned} \mathrm{y}_1^h &= h + h^3 \mathrm{u}_1^h \\ 0 \leq \mathrm{u}_1^h &\perp \mathrm{y}_1^h \geq 0, \end{aligned}$$

yields $(\mathrm{u}_1^h, \mathrm{y}_1^h) = (0, h)$ and $\mathrm{x}_1^h = \begin{pmatrix} \begin{pmatrix} h & 1 & h^{-1} \end{pmatrix} \end{pmatrix}^\top$. By repeating the calculations, it can be verified that algorithm 7.3.1 yields $(\mathrm{u}_k^h, \mathrm{y}_k^h) = (0, \frac{k(k+1)}{2} h)$ for $k \neq 0$. It is clear from (7.5d) that

$$\|y_{reg}^h\| \geq \Big(\int_{(N_h-2)h}^{(N_h-1)h} \|\mathrm{y}_{(N_h-1)}^h\|^2 \, dt\Big)^{1/2} = \frac{(N_h - 1)N_h}{2} h^{3/2} = O(h^{-1/2})$$

whenever $N_h \geq 2$. Therefore, y_{reg}^h is far from being convergent and even not bounded as h converges to zero. □

This example indicates that one should be cautious in applying a time-stepping method to LCS. A verification of a numerical scheme in the sense of showing consistency is consequently needed. The following theorem states conditions that imply consistency.

Theorem 7.3.4 *Consider LCS(A, B, C, D) such that assumption 7.3.2 holds. Let $\tau > 0$ and $x_0 \in \mathbb{R}^n$ be given. Also let (u^h, x^h, y^h) be given by (7.5) via algorithm 7.3.1. Suppose that there exists $\alpha > 0$ such that for all sufficiently small h*

$$\|h\mathrm{u}_0^h\| \leq \alpha \text{ and } \|u_{reg}^h\| \leq \alpha.$$

Then, we have the following statements:

1. *There exists a unique initial solution of LCS(A, B, C, D) with initial state x_0 in the sense of definition 7.2.2.*
2. *The triple $\{(u_{imp}^h, x_{imp}^h, y_{imp}^h)\}$ converges to $(u_{imp}, 0, y_{imp})$, when h tends to zero. Moreover, (u_{imp}, y_{imp}) is of the form $(u_0\delta, y_0\delta)$ with $u_0, y_0 \in \mathbb{R}^m$ such that $(u_{imp}, 0, y_{imp})$ is equal to the impulsive part of the unique initial solution corresponding to initial state x_0.*
3. *Let $\{h_k\}$ converge to zero. Suppose that D is nonnegative definite. Then,*
   (a) *There exists a subsequence $\{h_{k_l}\} \subseteq \{h_k\}$ such that $(\{u^{h_{k_l}}\}, \{y^{h_{k_l}}\})$ converges weakly to some (u, y) and $\{x^{h_{k_l}}\}$ converges to some x.*
   (b) *(u, x, y) is a solution of LCS(A, B, C, D) on $[0, \tau]$ with initial state x_0.*
   (c) *If the solution (u, x, y) is unique for initial state x_0 in the sense of definition 7.2.5, then the complete sequence $(\{u^{h_k}\}, \{y^{h_k}\})$ converges weakly to (u, y) and $\{x^{h_k}\}$ converges to x.*

□

Proof. See section 7.6. □

7.4 Main results for passive LCS

We now show that the conditions of theorem 7.3.4 are satisfisfied in the case of passive linear complementarity systems so that the following result holds.

Theorem 7.4.1 *Consider the LCS(A, B, C, D) such that assumption 7.2.9 holds and (A, B, C, D) is passive. Let $\tau > 0$ and $x_0 \in \mathbb{R}^n$ be given. Let (u, x, y) be the solution of LCS(A, B, C, D) on $[0, \tau]$ with the initial state x_0. Also let (u^h, x^h, y^h) be given by (7.5) via algorithm 7.3.1. Then, $(\{u^h\}, \{y^h\})$ converges weakly to (u, y) and $\{x^h\}$ converges to x as the step size h tends zero.* □

Proof. See section 7.6. □

The above theorem assumes exact computations. In implementing the backward Euler time-stepping method numerical errors will of course be introduced. To give

some justification that also in the case of (small) numerical errors the method is still suitable, we study the issue of the dependence of the solution trajectories on the initial conditions. For general LCS such a property does not hold (see e.g. example 3.8.3 in chapter 3). However, in the special case of linear passive complementarity systems, the continuous dependence holds. To formulate this in a mathematically precise way, we have to introduce some nomenclature. Let $\mathcal{H}$ be a Hilbert space. We say that $T : \mathbb{R}^n \to \mathcal{H}$ is *continuous* (*weakly continuous*), if continuity is considered with respect to the strong (weak) topology on $\mathcal{H}$. In other words, T is continuous (weakly continuous), if for all convergent (weakly convergent) sequences $\{x_k\}$, $\{Tx_k\}$ converges (weakly converges) to $T(\lim_{k\to\infty} x_k)$.

Theorem 7.4.2 *Consider the LCS*(A, B, C, D) *such that assumption 7.2.9 holds and* (A, B, C, D) *is passive. Let* $\tau > 0$ *be given. Define the operators* $x_0 \mapsto (u, y)$ *and* $x_0 \mapsto x$*, where* (u, x, y) *is the solution of LCS*(A, B, C, D) *on* $[0, \tau]$ *with the initial state* x_0*. The operators* $x_0 \mapsto (u, y)$ *and* $x_0 \mapsto x$ *are weakly continuous and continuous, respectively.* □

Proof. See section 7.6. □

7.5 Conclusions

In this chapter, we studied the consistency of a time-stepping method based on the backward Euler integration routine. The method has already proven itself in practice for the transient simulation of piecewise linear electrical circuits and constrained mechanical systems. However, one cannot indiscriminately apply this method for general classes of discontinuous systems as shown by an example in this chapter. The main result of the chapter is therefore concerned with presenting a rigorous proof of the consistency of the backward Euler time-stepping method for a class of linear complementarity systems, to wit linear passive electrical networks with ideal diodes. In spite of the mixed continuous and discrete behaviour of the circuit, the possibility of Dirac impulses occurring at the initial time, and the fact that the time-stepping method does not try to locate the event times exactly, we have shown the convergence of the approximations to the actual transient solution of the network model. Using almost the same arguments, we have also proven the continuous dependence of the true transient solutions on the initial state. For simulation of linear passive networks with ideal diodes, this has the important consequence that numerical errors do not have a large influence on the outcomes of the approximation method. These results provide a justification for the use of time-stepping methods.

Of course, it would be interesting to generalize these results to other systems of a mixed continuous and discrete nature. In particular, we are currently studying the consistency of the backward Euler method for dynamical systems with relay switches and other subclasses of linear complementarity systems. For many system where the

backward Euler time-stepping scheme does not generate proper output (like the triple integrator), it is useful to consider extensions of the time-stepping algorithm that are consistent.

7.6 Proofs

7.6.1 Preliminaries

For ease of reference, we recall some standard results on weakly convergent sequences.

Lemma 7.6.1 *[209] The following statements hold in every Hilbert space $\mathcal{H}$.*

1. *Every bounded sequence has a weakly convergent subsequence.*
2. *If all weakly convergent subsequences of a bounded sequence have the same weak limit, then the sequence itself converges weakly to this limit.*
3. *Assume that $\{v_k\} \subset \mathcal{H}$ converges weakly to v and $\{w_k\} \subset \mathcal{H}$ converges to w. Then*
   (a) *There exists $\alpha > 0$ such that $\|v_k\| \leq \alpha$ for all k and $\|v\| \leq \alpha$.*
   (b) *$\{Sv_k\}$ converges weakly to Sv whenever $S : \mathcal{H} \to \mathcal{H}$ is a continuous linear operator.*
   (c) *$\{\langle v_k, w_k\rangle\}$ converges to $\langle v, w\rangle$.*

□

In the following lemma, we state some results for the matrix inverse $(I - hA)^{-1}$.

Lemma 7.6.2 *Let $A \in \mathbb{R}^{nxn}$. The following statements hold:*

1. *$\|(I - hA)^{-1}\| \leq \dfrac{1}{1 - \lambda h}$ for all h with $\lambda h < 1$ where λ is the largest eigenvalue of $\frac{1}{2}(A + A^\top)$.*
2. *There exists an $\alpha > 0$ such that $\|(I - hA)^{-1}\| \leq \alpha$ for all sufficiently small h.*
3. *If $\{r_k h_k\}$ converges to t then $\{(I - h_k A)^{-r_k}\}$ converges to e^{At} uniformly in t on any bounded interval.*

□

Proof. *1:* By the Wazewski inequality (see e.g. [207, theorem 8.1]), $\|e^{At}\| \leq e^{\lambda t}$ for all t where λ is the largest eigenvalue of $\frac{1}{2}(A + A^\top)$. Theorem 1.5.3 in [156] gives now the desired inequality.

2: It can easily be verified by using item 1 that

$$\|(I - hA)^{-1}\| \leq \frac{1}{1-\alpha}$$

whenever $\lambda h \leq \alpha < 1$.

3: This follows from [156, theorem 3.5.3]. □

7.6.2 Proof of theorem 7.3.4 item 1 and 2

For proving theorem 7.3.4, we start by considering the items *1* and *2*, which are concerned with the existence and uniqueness of initial solutions and the convergence of the impulsive parts of the approximations to the impulsive part of this initial solution. Note that the latter is needed to show that the limit of the approximations exists and satisfies definition 7.2.5 item *1*.

We shall use the following proposition which establishes the relation between the solutions of the one-step problem and the solutions of the rational complementarity problem.

Proposition 7.6.3 *Consider matrices $A \in \mathbb{R}^{n\text{x}n}$, $B \in \mathbb{R}^{n\text{x}m}$, $C \in \mathbb{R}^{m\text{x}n}$ and $D \in \mathbb{R}^{m\text{x}m}$ such that assumption 7.3.2 holds. We have the following statements for all $x_0 \in \mathbb{R}^n$.*

1. *RCP(x_0, A, B, C, D) has a unique solution.*
2. *For all sufficiently small h,*

$$\hat{u}(h^{-1}) = h\mathrm{u}_0^h$$
$$\hat{x}(h^{-1}) = h\mathrm{x}_0^h$$
$$\hat{y}(h^{-1}) = h\mathrm{y}_0^h$$

where $(\hat{u}(s), \hat{y}(s))$ is the solution of RCP(x_0, A, B, C, D) and $\hat{x}(s) = (sI - A)^{-1}x_0 + (sI - A)^{-1}B\hat{u}(s)$.

□

Proof.
1: Observe the basic fact that if LCP(q, M) is solvable, then LCP(αq, M) is also solvable for any $\alpha \geq 0$. As a consequence, assumption 7.3.2 implies together with the identity $h(I - hA)^{-1} = (h^{-1}I - A)^{-1}$ that for all sufficiently small h, LCP($C(h^{-1}I - A)^{-1}x_0$, $G(h^{-1})$) has a unique solution. From theorem 4.4.1 and corollary 4.4.10 in chapter 4, we can conclude that RCP(x_0, A, B, C, D) has a unique solution.

2: Let $(\hat{u}(s), \hat{y}(s))$ be the solution of RCP(x_0, A, B, C, D). It can be easily seen that $\hat{u}(h^{-1})$ solves LCP($C(h^{-1}I - A)^{-1}x_0$, $G(h^{-1})$) for all sufficiently small h. Note

that if z is a solution of LCP(q, M), then αz is a solution of LCP(αq, M) for $\alpha \geq 0$. Therefore, $h^{-1}\hat{u}(h^{-1})$ solves LCP($C(I - hA)^{-1}x_0$, $G(h^{-1})$) for all sufficiently small h due to the identity $h^{-1}(h^{-1}I - A)^{-1} = (I - hA)^{-1}$. Stated differently, for all sufficiently small h

$$\hat{u}(h^{-1}) = h\mathrm{u}_0^h \tag{7.6a}$$

$$\hat{x}(h^{-1}) = h\mathrm{x}_0^h \tag{7.6b}$$

$$\hat{y}(h^{-1}) = h\mathrm{y}_0^h, \tag{7.6c}$$

where $\hat{x}(s) = (sI - A)^{-1}x_0 + (sI - A)^{-1}B\hat{u}(s)$. □

Proof of theorem 7.3.4 items 1 and 2

1: From propositon 7.6.3 item 1, it is known that RCP(x_0, A, B, C, D) is uniquely solvable. Let $(\hat{u}(s), \hat{y}(s))$ denote this unique solution and $\hat{x}(s) = (sI - A)^{-1}x_0 + (sI - A)^{-1}B\hat{u}(s)$. Since $\|h\mathrm{u}_0^h\|$ is bounded for sufficiently small h by the hypothesis of the theorem, $\hat{u}(s)$ is proper due to proposition 7.6.3 item 2. It follows that $\hat{x}(s)$ is strictly proper and $\hat{y}(s)$ is proper. Clearly, propositon 7.2.4 implies that the inverse Laplace transform of $(\hat{u}(s), \hat{x}(s), \hat{y}(s))$ is the unique initial solution of LCS(A, B, C, D) with initial state x_0.

2: Let $(\hat{u}(s), \hat{x}(s), \hat{y}(s))$ be the Laplace transform of the unique initial solution of LCS(A, B, C, D) with initial state x_0. Proposition 7.2.4 implies that $\hat{u}(s)$ and $\hat{y}(s)$ are proper and $\hat{x}(s)$ is strictly proper. Then, the impulsive part of the initial solution $(u_{imp}, x_{imp}, y_{imp})$ is of the form $(u_0\delta, 0, y_0\delta)$ where $u_0 = \lim_{s\to\infty}\hat{u}(s)$ and $y_0 = \lim_{s\to\infty}\hat{y}(s)$. It is clear from (7.5a)-(7.5c) and proposition 7.6.3 item 2 that $(u_{imp}^h, x_{imp}^h, y_{imp}^h)$ converges to $(u_{imp}, 0, y_{imp})$ as h tends zero. □

7.6.3 Order complementarity problem

In this subsection, an infinite dimensional version of the LCP will be considered. This so-called *order complementarity problem* (OCP) has strong relations to (the regular parts of) the solutions of LCS on one hand. On the other, it is possible to embed the discretizations obtained from the backward Euler time-stepping method in the OCP.

To be specific, we briefly recall OCP for the function space $\mathcal{L}_2(0, \tau)$. More details on the OCP can be found in [22].

Problem 7.6.4 $(OCP(q, T))$ Given $q \in \mathcal{L}_2^m(0, \tau)$ and $T : \mathcal{L}_2^m(0, \tau) \to \mathcal{L}_2^m(0, \tau)$, find $z \in \mathcal{L}_2^m(0, \tau)$ such that

$$z(t) \geq 0 \tag{7.7a}$$

$$q(t) + (Tz)(t) \geq 0 \tag{7.7b}$$

for almost all $t \in [0, \tau]$ and

$$\langle z, q + Tz \rangle = 0. \tag{7.7c}$$

□

If z satisfies (7.7), we say that z solves $OCP(q, T)$. In this case, we sometimes also state that (z, w) solves $OCP(q, T)$, where $w = q + Tz$.

Note that the conditions given in item 3 of definition 7.2.5 may be equivalently written as

$$u_{reg}(t) \geq 0 \tag{7.8a}$$

$$y_{reg}(t) \geq 0 \tag{7.8b}$$

for almost all $t \in [0, \tau]$ and

$$\langle u_{reg}, y_{reg} \rangle = 0. \tag{7.8c}$$

Hence, by associating the operator $T_{(A,B,C,D)}$ defined by

$$(T_{(A,B,C,D)}u)(t) = Du(t) + \int_0^t Ce^{A(t-s)}Bu(s)ds$$

to LCS(A, B, C, D), the solutions of LCS(A, B, C, D) can be identified with the solutions of certain OCPs in the following manner.

Proposition 7.6.5 *The following statements hold.*

1. *If $(u, x, y) \in \mathcal{L}^n_\delta(0, \tau)$ is a solution of LCS(A, B, C, D) on $[0, \tau]$ with initial state x_0, then u_{reg} is the solution of $OCP(Ce^{A\cdot}x_0^+|_{[0,\tau]}, T_{(A,B,C,D)})$, where $x_0^+ = x_0 + Bu_0$ and $u_{imp} = u_0\delta$.*
2. *If $u \in \mathcal{L}^m_2((0, \tau))$ is a solution of $OCP(Ce^{A\cdot}x_0|_{[0,\tau]}, T_{(A,B,C,D)})$, then (u, x, y) is a solution of LCS(A, B, C, D) on $[0, \tau]$ with initial state x_0 where*

$$x = e^{A\cdot}x_0|_{[0,\tau]} + T_{(A,B,I,0)}u$$
$$y = Cx + Du.$$

□

7.6.4 The time-stepping method in an OCP formulation

The approximations of (7.5) by the backward Euler time-stepping scheme can also be formulated as the solutions of certain OCPs. To formalize this, we introduce the operators $\tilde{C}_h : \mathbb{R}^{nN_h} \to \mathbb{R}^{mN_h}$, $\tilde{D}_h : \mathbb{R}^{mN_h} \to \mathbb{R}^{mN_h}$, $R_h : \mathcal{L}^m_2(0, \tau) \to \mathbb{R}^{mN_h}$,

$Q_h : \mathbb{R}^{mN_h} \to \mathbb{R}^{nN_h}$, and $P_h^j : \mathbb{R}^{jN_h} \to \mathcal{L}_2^j(0,\tau)$ for given $\tau > 0$ and h with $N_h = \lceil \tau/h \rceil$.

$$\tilde{C}_h := \begin{bmatrix} C & 0 & \cdots & 0 \\ 0 & C & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & C \end{bmatrix} \qquad \tilde{D}_h := \begin{bmatrix} D & 0 & \cdots & 0 \\ 0 & D & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & D \end{bmatrix}$$

$$R_h u := \frac{1}{h} \begin{bmatrix} \int_0^h u(s)ds \\ \int_h^{2h} u(s)ds \\ \vdots \\ \int_{(N_h-1)h}^{\tau} u(s)ds \end{bmatrix}$$

$$Q_h := h \begin{bmatrix} (I-hA)^{-1}B & 0 & \cdots & 0 \\ (I-hA)^{-2}B & (I-hA)^{-1}B & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ (I-h_A)^{-N_h}B & (I-hA)^{-N_h+1}B & \cdots & (I-hA)^{-1}B \end{bmatrix}$$

$$(P_h^j \mathrm{w})(t) := \mathrm{w}_{\overline{lj}\setminus\overline{(l-1)j}} \text{ if } t \in [(l-1)h, lh) \text{ for } l = 1, 2, \ldots, N_h.$$

For ease of reference, we summarize some of the properties of these operators, which will be needed in the the sequel. Without loss of generality , we can assume that $N_h h = \tau$.

Proposition 7.6.6 *Let* $\mathrm{v}, \mathrm{w} \in \mathbb{R}^{mN_h}$ *and* $\mathrm{x} \in \mathbb{R}^{nN_h}$. *The following statements hold.*

1. $R_h P_h^m \mathrm{v} = \mathrm{v}$.
2. $\mathrm{v} \geq 0$ *if and only if* $P_h^m \mathrm{v}(t) \geq 0$ *for (almost) all* $t \in [0, \tau]$.
3. $\langle P_h^m \mathrm{v}, P_h^m \mathrm{w} \rangle = h\, \mathrm{v}^\top \mathrm{w}$.
4. $D P_h^m \mathrm{v} = P_h^m \tilde{D}_h \mathrm{v}$.
5. $C P_h^n \mathrm{x} = P_h^m \tilde{C}_h \mathrm{x}$.

□

Proof. Evident from the definitions of P_h^j, R_h, $\tilde{C}_h$ and $\tilde{D}_h$. □

It can easily be seen that $(\tilde{\mathrm{u}}_h, \tilde{\mathrm{y}}_h)$ solves $\mathrm{LCP}(\tilde{C}_h \tilde{\mathrm{q}}_h, \tilde{D}_h + \tilde{C}_h Q_h)$, where

$$\tilde{\mathrm{u}}_h = \begin{bmatrix} \mathrm{u}_1^h \\ \mathrm{u}_2^h \\ \vdots \\ \mathrm{u}_{N_h}^h \end{bmatrix}, \tilde{\mathrm{y}}_h = \begin{bmatrix} \mathrm{y}_1^h \\ \mathrm{y}_2^h \\ \vdots \\ \mathrm{y}_{N_h}^h \end{bmatrix}, \text{ and } \tilde{\mathrm{q}}_h = \begin{bmatrix} (I-hA)^{-1}\mathrm{x}_0^h \\ (I-hA)^{-2}\mathrm{x}_0^h \\ \vdots \\ (I-hA)^{-N_h}\mathrm{x}_0^h \end{bmatrix}.$$

Indeed, LCP($\tilde{C}_h\tilde{\mathrm{q}}_h$, $\tilde{D}_h + \tilde{C}_h Q_h$) is pieced together from N_h one-step problems of algorithm 7.3.1 step 4. The following lemma will complete the puzzle by formulating the approximations as solutions of OCPs and showing convergence properties of solutions to a family of OCPs.

Lemma 7.6.7 *Let $T'_h = P_h^n Q_h R_h$ and $q'_h = P_h^n \tilde{\mathrm{q}}_h$. The following statements hold.*

1. *For all sufficiently small h, (u^h_{reg}, y^h_{reg}) as given by (7.5) solves $OCP(Cq'_h, D + CT'_h)$.*
2. *$\{q'_h(\cdot)\}$ converges to $e^{A\cdot}(x_0 + Bu_0)$ with u_0 as in item 2 of theorem 7.3.4 as h tends zero.*
3. *$\{T'_h u^h_{reg} - T_{(A,B,I,0)} u^h_{reg}\}$ converges to 0 as h tends zero.*

□

Proof.

1: Since $(\tilde{\mathrm{u}}_h, \tilde{\mathrm{y}}_h)$ solves LCP($\tilde{C}_h\tilde{\mathrm{q}}_h$, $\tilde{D}_h + \tilde{C}_h Q_h$), we have

$$\tilde{\mathrm{u}}_h \geq 0 \tag{7.10a}$$

$$\tilde{\mathrm{y}}_h = \tilde{C}_h\tilde{\mathrm{q}}_h + (\tilde{D}_h + \tilde{C}_h Q_h)\tilde{\mathrm{u}}_h \geq 0 \tag{7.10b}$$

$$\tilde{\mathrm{u}}_h^\top \tilde{\mathrm{y}}_h = 0. \tag{7.10c}$$

Note that $u^h_{reg} = P_h^m \tilde{\mathrm{u}}_h$ and $y^h_{reg} = P_h^m \tilde{\mathrm{y}}_h$ due to (7.5) and the definition of P_h^m. Hence, (7.10a) and (7.10b) together with proposition 7.6.6 item 2 implies that

$$u^h_{reg}(t) \geq 0 \text{ and } y^h_{reg}(t) \geq 0 \text{ for (almost) all } t \in [0, \tau], \tag{7.11}$$

while proposition 7.6.6 items 1, 4 and 5 yield

$$\begin{aligned} y^{h_k}_{reg} &= P_k^m \tilde{\mathrm{y}}_k \\ &= P_k^m \tilde{C}_k \tilde{\mathrm{q}}_k + P_k^m \tilde{D}_k \tilde{\mathrm{u}}_k + P_k^m \tilde{C}_k Q_k R_k P_k^m \tilde{\mathrm{u}}_k \\ &= Cq'_k + (D + CT'_k) u^{h_k}_{reg}. \end{aligned} \tag{7.12}$$

Moreover, we have that

$$\begin{aligned} \langle u^h_{reg}, y^h_{reg} \rangle &= \langle P_h^m \tilde{\mathrm{u}}_h, P_h^m \tilde{\mathrm{y}}_h \rangle \\ &= h\, \tilde{\mathrm{u}}_h^\top \tilde{\mathrm{y}}_h \\ &= 0 \end{aligned} \tag{7.13}$$

from proposition 7.6.6 item 3, and (7.10c). Clearly, (7.11), (7.12) and (7.13) imply that (u^h_{reg}, y^h_{reg}) solves $OCP(Cq'_h, D + CT'_h)$.

2: Note that from algorithm 7.3.1 step 5 we have

$$\mathrm{x}_0^h := (I - hA)^{-1}x_0 + h(I - hA)^{-1}B\mathrm{u}_0^h. \tag{7.14}$$

Let $\hat{u}(s)$ be the solution of RCP(x_0, A, B, C, D) and $u_0 = \lim_{s\to\infty}\hat{u}(s)$. As shown in the proof of theorem 7.3.4 item 2, $h\mathrm{u}_0^h$ converges to u_0 as h tends zero. Then, (7.14) implies that

$$\{\mathrm{x}_0^h\} \text{ converges to } x_0 + Bu_0 \tag{7.15}$$

as h tends zero. Note that

$$q_h'(t) = (I - hA)^{-\lceil t/h\rceil}\mathrm{x}_0^h.$$

Hence, from the triangle inequality we get

$$\begin{aligned}\|q_h'(\cdot) - e^{A\cdot}(x_0 + Bu_0)\| &\\ &\le \|(I - hA)^{-\lceil \cdot/h\rceil}\mathrm{x}_0^h - e^{A\cdot}\mathrm{x}_0^h\| + \|e^{A\cdot}\mathrm{x}_0^h - e^{A\cdot}(x_0 + Bu_0)\| \\ &\le \Big(\int_0^\tau \|(I - hA)^{-\lceil t/h\rceil} - e^{At}\|^2\,dt\Big)^{1/2}\|\mathrm{x}_0^h\| + \\ &\qquad + \Big(\int_0^\tau \|e^{At}\|^2\,dt\Big)^{1/2}\|\mathrm{x}_0^h - (x_0 + Bu_0)\|.\end{aligned}$$

Since $\{\lceil t/h\rceil h\}$ converges to t as h tends zero, lemma 7.6.2 item 3 and (7.15) reveal that the right hand side converges to zero.

3: Note that

$$\begin{aligned}(T_h'u_{reg}^h)(t) &= \sum_{p=1}^{l} h(I - hA)^{-(l-p+1)}B\mathrm{u}_p^h \\ &= \sum_{p=1}^{l}\int_{(p-1)h}^{ph}(I - hA)^{-(l-p+1)}B\mathrm{u}_p^h ds\end{aligned}$$

and also that

$$(T_{(A,B,I,0)}u_{reg}^h)(t) = \sum_{p=1}^{l-1}\int_{(p-1)h}^{ph} e^{A(t-s)}B\mathrm{u}_p^h ds + \int_{(l-1)h}^{t} e^{A(t-s)}B\mathrm{u}_l^h ds$$

with $l = \lceil t/h\rceil$. By exploiting the triangle inequality, we get

$$\begin{aligned}\|(T_h'u_{reg}^h)(t) - (T_{(A,B,I,0)}u_{reg}^h)(t)\| \le &\\ \sum_{p=1}^{\lceil t/h\rceil}\int_{(p-1)h}^{ph} &\|(I - hA)^{-(\lceil t/h\rceil - \lceil s/h\rceil + 1)} - e^{A(t-s)}\|\,\|B\mathrm{u}_p^h\|ds\end{aligned} \tag{7.16}$$

since $(p-1)h < s \leq ph$ gives $p = \lceil s/h \rceil$. Clearly, $\{(\lceil t/h \rceil - \lceil s/h \rceil + 1)h\}$ converges to $t - s$ as h tends zero. We already know from the hypothesis that $\|\mathrm{u}_p^h\|$ is bounded for $p \neq 0$. Therefore, from lemma 7.6.2 item 3 we can conclude that the right hand side converges to zero uniformly in t on any bounded interval. It follows that $\{T_h' u_{reg}^h - T_{(A,B,I,0)} u_{reg}^h\}$ converges to zero in $\mathcal{L}_2(0,\tau)$ as h tends zero. □

7.6.5 Convergence of solutions to order complementarity problems

From the previous subsection, it is obvious that the convergence problem for the time-stepping method can be reduced to convergence of the solutions of a sequence of OCPs. The following theorem provides a general framework in which we shall prove the convergence of the regular parts of the approximations obtained by the backward Euler time-stepping method. Before stating the theorem, we need to define the concept of *compact operators*.

Definition 7.6.8 Let $\mathcal{H}$ be a Hilbert space. $T : \mathcal{H} \to \mathcal{H}$ is said to be a *compact operator*, if for any weakly convergent sequence $\{u_k\} \subset \mathcal{H}$, $\{Tu_k\}$ is a (strongly) convergent sequence. □

Theorem 7.6.9 *Let $T : \mathcal{L}_2^m(0,\tau) \to \mathcal{L}_2^m(0,\tau)$ be a compact operator and let $S : \mathcal{L}_2^m(0,\tau) \to \mathcal{L}_2^m(0,\tau)$ be a linear continuous nonnegative definite (i.e. $\langle v, Sv\rangle \geq 0$ for all $v \in \mathcal{L}_2^m(0,\tau)$) operator. Suppose that there exist sequences $\{q_k\}$ and $\{T_k\}$ such that $\{q_k\}$ converges to q and $OCP(q_k, S+T_k)$ is solvable for all k. Let z_k be a solution of $OCP(q_k, S+T_k)$. If $\{z_k\}$ converges weakly to z and $\{T_k z_k - T z_k\}$ converges to zero then z solves $OCP(q, S+T)$.* □

Proof. In order to prove the theorem, one should show that z, which is the weak limit of $\{z_k\}$, satisfies

$$z(t) \geq 0 \tag{7.17a}$$

$$q(t) + ((S+T)z)(t) \geq 0 \tag{7.17b}$$

for almost all $t \in [0,\tau]$ and

$$\langle z, q + (S+T)z\rangle = 0. \tag{7.17c}$$

Since z_k solves $OCP(q_k, S+T_k)$, we have

$$z_k(t) \geq 0 \tag{7.18a}$$

$$q_k(t) + ((S+T_k)z_k)(t) \geq 0 \tag{7.18b}$$

for almost all $t \in [0, \tau]$ and

$$\langle z_k, q_k + (S + T_k)z_k \rangle = 0 \tag{7.18c}$$

for all k. Now, (7.17a) follows from (7.18a) and the weak closedness of the set $\{v \mid v(t) \geq 0$ for almost all $t \in [0, \tau]\}$ (see [170, theorem 3.12]. Lemma 7.6.1 item 3b and definition 7.6.8 imply that

$$\{Sz_k\} \text{ converges weakly to } Sz \tag{7.19a}$$

and

$$\{Tz_k\} \text{ converges to } Tz. \tag{7.19b}$$

As a consequence of (7.19b), we have

$$\{T_k z_k\} \text{ converges to } Tz \tag{7.19c}$$

since $\{T_k z_k - Tz_k\}$ converges to zero by assumption. The equations (7.19a), (7.19c) and the convergence of $\{q_k\}$ imply that $\{q_k + (S + T_k)z_k\}$ converges weakly to $q + (S + T)z$. Hence, (7.17b) follows from (7.18b) and the weak closedness of $\{v \mid v(t) \geq 0$ for almost all $t \in [0, \tau]\}$. Now, it remains to show that (7.17c) holds. Equation (7.18c) gives

$$\langle z_k, Sz_k \rangle = -\langle z_k, q_k + T_k z_k \rangle.$$

The convergence of $\{q_k\}$ and the weak convergence of $\{z_k\}$, together with (7.19c) and lemma 7.6.1 item 3c, imply that

$$\lim_{k\to\infty} \langle z_k, Sz_k \rangle = \lim_{k\to\infty} -\langle z_k, q_k + T_k z_k \rangle = -\langle z, q + Tz \rangle.$$

We also have from (7.17a) and (7.17b) that

$$\langle z, q + (S + T)z \rangle \geq 0.$$

Thus,

$$\langle z, Sz \rangle \geq -\langle z, q + Tz \rangle = \lim_{k\to\infty} \langle z_k, Sz_k \rangle. \tag{7.20}$$

The nonnegative definiteness of S implies

$$\langle z_k - z, S(z_k - z) \rangle \geq 0. \tag{7.21}$$

Since $\lim_{k\to\infty} \langle z, Sz_k \rangle = \lim_{k\to\infty} \langle z_k, Sz \rangle = \langle z, Sz \rangle$ due to the fact that $\{z_k\}$ converges weakly to z and lemma 7.6.1 items 3b and 3c, we get

$$\lim_{k\to\infty} \langle z_k, Sz_k \rangle \geq \langle z, Sz \rangle \tag{7.22}$$

by letting k tend to infinity in (7.21). Together with (7.20), this yields

$$\lim_{k\to\infty} \langle z_k, Sz_k\rangle = \langle z, Sz\rangle. \tag{7.23}$$

Combining (7.23), (7.19c), the convergence of $\{q_k\}$ to q and lemma 7.6.1 item 3c results in

$$\lim_{k\to\infty} \langle z_k, q_k + (S+T_k)z_k\rangle = \langle z, q + (S+T)z\rangle. \tag{7.24}$$

Finally, (7.17c) follows from (7.24) and (7.18c). □

7.6.6 Completing the proof of theorem 7.3.4

The proofs of item 1 and 2 in theorem 7.3.4 have already been shown. The remaining items will be proven in this subsection.

Proofs for items 3a, 3b and 3c of theorem 7.3.4 *3a:* The convergence of impulsive parts has already been shown in the proof of item 2. Hence, we must show that the claim about the regular parts holds. By the hypothesis of the theorem, we know that $\|u^h_{reg}\|$ is bounded for sufficiently small h. According to lemma 7.6.1 item 1, the existence of a weakly convergent subsequence of $\{u^{h_k}_{reg}\}$, say $\{u^{h_{k_l}}_{reg}\}$, is clear. Let u_{reg} denote the weak limit of this subsequence, and also let q'_{h_k} and T'_{h_k} be defined as in lemma 7.6.7. Since $T_{(A,B,I,0)}$ is a compact operator (see e.g. [170, exercise 4.15]), it follows from definition 7.6.8 that $\{T_{(A,B,I,0)}u^{h_{k_l}}_{reg}\}$ converges (strongly) to $T_{(A,B,I,0)}u_{reg}$. Then, lemma 7.6.7 item 3 implies that

$$\{T'_{h_{k_l}}u^{h_{k_l}}_{reg}\} \text{ converges to } T_{(A,B,I,0)}u_{reg}. \tag{7.25}$$

Note that

$$x^{h_{k_l}}_{reg} = q'_{h_{k_l}} + T'_{h_{k_l}}u^{h_{k_l}}_{reg} \tag{7.26a}$$

and

$$y^{h_{k_l}}_{reg} = Cq'_{h_{k_l}} + (D + CT'_{h_{k_l}})u^{h_{k_l}}_{reg}. \tag{7.26b}$$

It is clear from lemma 7.6.7 item 2, (7.26a) and (7.25) that $\{x^{h_{k_l}}_{reg}\}$ converges to $x_{reg} := e^{A\cdot}(x_0 + Bu_0)|_{[0,\tau]} + T_{(A,B,I,0)}u_{reg}$. Since $\{Du^{h_{k_l}}_{reg}\}$ converges weakly to Du_{reg} due to lemma 7.6.1 item 3b, it follows from lemma 7.6.7 item 2, (7.26b) and (7.25) that $\{y^{h_{k_l}}_{reg}\}$ converges weakly to $y_{reg} := Ce^{A\cdot}(x_0 + Bu_0)|_{[0,\tau]} + T_{(A,B,C,D)}u_{reg}$.

3b: Item 2 of Theorem 7.3.4 (see also the proof) states the convergence of the triple $(u^{h_k}_{imp}, x^{h_k}_{imp}, y^{h_k}_{imp})$ to

$$(u_{imp}, 0, y_{imp}) = (u_0\delta, 0, y_0\delta) = (\bar{u}_{imp}, \bar{x}_{imp}, \bar{y}_{imp}), \tag{7.27}$$

where $(\bar{u}, \bar{x}, \bar{y}) \in \mathcal{B}_\delta^{m+n+m}$ is the unique initial solution for initial state x_0. Hence, we also have that $y_{imp} = Du_{imp}$ due to $x_{imp} = 0$. Let us define in the framework of theorem 7.6.9

- $T = T_{(A,B,C,0)}$,
- $S = D$,
- $q_l = Cq'_{h_{k_l}}$, and
- $T_l = CT'_{h_{k_l}}$.

It can be checked that

- T is compact ([170, exercise 4.15]),
- S is nonnegative definite (by the hypothesis $D \geq 0$),
- $\{q_l\}$ converges to $Ce^{A\cdot}(x_0 + Bu_0)|_{[0,\tau]}$ (from lemma 7.6.7 item 2)
- $OCP(q_l, S + T_l)$ is solvable (from lemma 7.6.7 item 1), and
- $\{T_l u_{reg}^{h_{k_l}} - T u_{reg}^{h_{k_l}}\}$ converges to zero (from lemma 7.6.7 item 3).

Then, theorem 7.6.9 implies that u_{reg} solves $OCP(Ce^{A\cdot}(x_0 + Bu_0)|_{[0,\tau]}, T_{(A,B,C,D)})$. Due to proposition 7.6.5 item 2, $(u_{reg}, x_{reg}, y_{reg})$ is a solution of LCS(A, B, C, D) on $[0, \tau]$ with the initial state $x_0 + Bu_0$ (with u_0 as in (7.27)), where

$$x_{reg} = e^{A\cdot}(x_0 + Bu_0)|_{[0,\tau]} + T_{(A,B,I,0)}u_{reg}$$
$$y_{reg} = Cx_{reg} + Du_{reg}.$$

Equivalently,

$$\dot{x}_{reg} = Ax_{reg} + Bu_{reg} + (x_0 + Bu_0)\delta \tag{7.28a}$$
$$y_{reg} = Cx_{reg} + Du_{reg} \tag{7.28b}$$

holds in the distributional sense and

$$0 \leq u_{reg}(t) \perp y_{reg}(t) \geq 0 \tag{7.28c}$$

for almost all $t \in [0, \tau]$. Since $u_{imp} = u_0\delta$, $y_{imp} = Du_{imp}$ and $x_{imp} = 0$, (7.28a) and (7.28b) yield

$$\dot{x} = Ax + Bu + x_0\delta \tag{7.29a}$$
$$y = Cx + Du \tag{7.29b}$$

Clearly, (7.27), (7.29) and (7.28c) imply that (u, x, y) is a solution of LCS(A, B, C, D) on $[0, \tau]$ with initial state x_0.

4: We have already proven that the complete sequence of the impulsive parts $(u_{imp}^{h_k}, x_{imp}^{h_k}, y_{imp}^{h_k})$ converges. Note that the sequence of regular parts $(u_{reg}^{h_k}, x_{reg}^{h_k}, y_{reg}^{h_k})$ is bounded by assumption. Moreover, following the proof of item 3 above, it is clear that every converging subsequence $(u_{reg}^{h_{k_l}}, x_{reg}^{h_{k_l}}, y_{reg}^{h_{k_l}})$ converges to a solution of the LCS(A, B, C, D) with initial state $x_0 + Bu_0$. Since this solution is unique, every converging subsequence of the bounded sequence of regular parts has the same limit. Applying theorem 7.6.1 item 2 completes the proof. □

7.6.7 Some results on LCPs

We will present in this subsection some results on LCPs, that will be needed to prove the main result (theorem 7.4.1) for linear passive complementarity systems.

Proposition 7.6.10 *Let $M \in \mathbb{R}^{n\text{x}n}$ be a positive definite matrix and z_i the unique solution of LCP(q_i, M) for $i = 1, 2$. Then,*

$$\|z_1 - z_2\| \leq \frac{n^{3/2}}{\mu(M)}\|q_1 - q_2\|$$

where $\mu(M)$ denotes the smallest eigenvalue of the symmetric part of M, i.e. $\frac{1}{2}(M + M^\top)$. □

Proof. By From Lemma 7.3.10 and proposition 5.10.10 in [47], we have

$$\|z_1 - z_2\|_\infty \leq \frac{n}{\mu(M)}\|q_1 - q_2\|_\infty. \tag{7.30}$$

Since $\|z\| \leq n^{1/2}\|z\|_\infty$ and $\|z\|_\infty \leq \|z\|$ for all $z \in \mathbb{R}^n$, (7.30) yields

$$\|z_1 - z_2\| \leq \frac{n^{3/2}}{\mu(M)}\|q_1 - q_2\|.$$

□

Using the passivity of (A, B, C, D), we can compute a lower bound on $\mu(G(h^{-1}))$ with $G(s) := C(sI - A)^{-1}B + D$, that will be useful for the application of proposition 7.6.10.

Lemma 7.6.11 *Consider the matrices $A \in \mathbb{R}^{n\text{x}n}$, $B \in \mathbb{R}^{n\text{x}m}$, $C \in \mathbb{R}^{m\text{x}n}$ and $D \in \mathbb{R}^{m\text{x}m}$ such that assumption 7.2.9 holds and (A, B, C, D) is passive. Let $\mu(N)$ denote the smallest eigenvalue of the symmetric part of a matrix N and define $G(s) := C(sI - A)^{-1}B + D$. The following statements hold.*

1. $D \geq 0$.
2. $u \neq 0$ *and* $u^\top Du = 0$ *implies that* $u^\top CBu > 0$.
3. *There exists* $\alpha > 0$ *such that* $\mu(D + hCB) \geq \alpha h$ *for all sufficiently small* h.
4. *There exists* $\beta > 0$ *such that* $\mu(G(h^{-1})) \geq \beta h$ *for all sufficiently small* h.

□

Proof. *1:* This is clear from lemma 7.2.8 item 2.

2: Assume that $u \neq 0$ and $u^\top Du = 0$. We claim that $(KB - C^\top)u = 0$, where K is a solution of the linear matrix inequalities in lemma 7.2.8 item 2. Suppose it is not true, i.e. $(KB - C^\top)u \neq 0$. Then, there exists $x \in \mathbb{R}^n$ such that $x^\top(KB - C^\top)u > 0$. Hence, for sufficiently small $\lambda > 0$

$$\begin{bmatrix} \lambda x \\ u \end{bmatrix}^\top \begin{bmatrix} A^\top K + KA & KB - C^\top \\ B^\top K - C & -(D + D^\top) \end{bmatrix} \begin{bmatrix} \lambda x \\ u \end{bmatrix} = \\ = \lambda^2 x^\top (A^\top K + KA)x + 2\lambda x^\top (KB - C^\top)u > 0. \quad (7.31)$$

Obviously, (7.31) contradicts lemma 7.2.8 item 2. Hence, $(KB - C^\top)u = 0$ and thus $u^\top CBu = u^\top B^\top KBu > 0$, because K is positive definite and B has full column rank.

3: Note that $a_1 + a_2 h \geq b_1 + b_2 h$ for all sufficiently small $h > 0$ if and only if $(a_1 > b_1)$ or $(a_1 = b_1$ and $a_2 \geq b_2)$. Since

$$(u^\top Du > 0) \text{ or } (u^\top Du = 0 \text{ and } u^\top CBu \geq \min_{\substack{v^\top Dv = 0 \\ \|v\| = 1}} v^\top CBv)$$

holds for all u with $\|u\| = 1$ due to items 1 and 2. From this we obtain that for all u with $\|u\| = 1$

$$u^\top Du + h\, u^\top CBu \geq h \min_{\substack{v^\top Dv = 0 \\ \|v\| = 1}} v^\top CBv \quad \text{for all sufficiently small } h > 0.$$

This yields by using [128, property 5.2.2.1(Rayleigh-Ritz theorem)]

$$\mu(D + hCB) = \min_{\|u\| = 1} u^\top (D + hCB)u \\ \geq h \min_{\substack{v^\top Dv = 0 \\ \|v\| = 1}} v^\top CBv$$

for all sufficiently small $h > 0$. Since

$$\min_{\substack{v^\top Dv = 0 \\ \|v\| = 1}} v^\top CBv > 0$$

according to item 2, the proof of this part is complete.

4: It is known from matrix theory (see e.g. [128, property 9.13.4.9]) that

$$\mu(N_1 + N_2) \geq \mu(N_1) + \mu(N_2)$$

for all square matrices N_1 and N_2. Hence, we get

$$\begin{aligned} \mu(G(h^{-1})) &\geq \mu(D + hCB) + h^2\mu(CA(I - hA)^{-1}B) \\ &\geq \beta h \qquad \text{(from item 3)} \end{aligned}$$

for some $\beta > 0$ and all sufficiently small h. □

The following auxiliary lemma will be needed in the sequel.

Lemma 7.6.12 *Let $\mathcal{P} = \{x \in \mathbb{R}^n \mid Ax \geq b\}$ be a given nonempty polyhedron with $A \in \mathbb{R}^{n\text{x}m}$ and $b \in \mathbb{R}^m$ and let x^* be equal to* $\arg\min_{x\in\mathcal{P}} \|x\|$. *There exists an index set $J \subseteq \overline{n}$ such that $x^* = \arg\min_{A_{J\bullet}x=b_J} \|x\|$.* □

Proof. Consider the convex quadratic optimization problem

$$\min_{Ax\geq b} \tfrac{1}{2}x^\top x.$$

The well-known Kuhn-Tucker conditions are necessary and sufficient for this problem because of its convexity (see for instance [47, section 1.2]), i.e. x^* is the solution of the optimization problem above if and only if there exists a $u \in \mathbb{R}^m$ such that

$$\begin{gathered} x^* = A^\top u \\ Ax^* \geq b \\ u \geq 0 \\ u^\top(Ax^* - b) = 0. \end{gathered}$$

Take such a vector u, define $J = \{j \mid u_j > 0\}$ and $v = u_J$. Then, x^* satisfies

$$x^* = (A_{J\bullet})^\top v \tag{7.32a}$$

$$A_{J\bullet}x^* = b_J. \tag{7.32b}$$

Note that (7.32) are necessary and sufficient (Kuhn-Tucker) conditions for the convex quadratic minimization problem

$$\min_{A_{J\bullet}x=b_J} \tfrac{1}{2}x^\top x.$$

□

To formulate the next lemma, we need to define the concept of a dual cone.

Definition 7.6.13 For any nonempty set $\mathcal{Q} \subset \mathbb{R}^m$, the set

$$\{w \in \mathbb{R}^m \mid w^\top v \geq 0 \text{ for all } v \in \mathcal{Q}\}$$

is called the *dual cone* of $\mathcal{Q}$ and denoted by $\mathcal{Q}^*$. □

Lemma 7.6.14 *Let $M \in \mathbb{R}^{n\times n}$ be nonnegative definite and $\mathcal{Q} = \mathrm{SOL}(0, M)$. We have the following statements.*

1. *LCP(q, M) is solvable if and only if $q \in \mathcal{Q}^*$,*
2. *For each $q \in \mathcal{Q}^*$, there exists a unique least-norm solution $z^* \in \mathrm{SOL}(q, M)$ such that $\|z^*\| \leq \|z\|$ for all $z \in \mathrm{SOL}(q, M)$,*
3. *There exists $\alpha > 0$ such that for all $q \in \mathcal{Q}^*$*

$$\|z^*(q)\| \leq \alpha \|q\|,$$

where $z^*(q)$ denotes the least-norm solution (see item 2) of LCP(q, M).

□

Proof. *1-2:* These statements follow from [47, cor. 3.8.10 and thm. 3.1.7(c)], respectively, because $\mathrm{SOL}(q, M)$ is a nonempty polyhedron when $q \in \mathcal{Q}^*$.

3: Define

$$\alpha(A) = \begin{cases} 0 & \text{if } A = 0 \\ \max\limits_{\substack{y \in \mathrm{im}\, A \\ \|y\|=1}} \min\limits_{Ax=y} \|x\| & \text{if } A \neq 0 \end{cases}$$

It is well-known that $\arg\min_{Ax=y} \|x\| = A^\dagger y$ for all $y \in \mathrm{im}\, A$, where $A^\dagger$ denotes the pseudoinverse of A (see [127, p.163]). Clearly, $y \mapsto \|A^\dagger y\|$ is a continuous function on $\mathrm{im}\, A$, because the pseudoinverse is linear and bounded and thus continuous [127, p.165]. Then, this mapping achieves its minimum on the set $\{y \mid y \in \mathrm{im}\, A \text{ and } \|y\| = 1\}$, which is compact. Hence, the quantity $\alpha(A)$ is well-defined for all A. Define

$$\alpha := \sqrt{2} \max_{J \subseteq \overline{n}} \max_{K \subseteq \overline{3n}} \alpha\left(\begin{bmatrix} I \\ -I_{J^c \bullet} \\ M \\ -M_{J\bullet} \end{bmatrix}_{K\bullet} \right).$$

For any $q \in \mathcal{Q}^*$, we know from the items 1 and 2 that $\mathrm{LCP}(q, M)$ is solvable and that there exists a unique least-norm solution $z^*(q)$. Let $J := \{j \mid z^*_j(q) > 0\}$. Clearly, $\mathcal{P} := \{v \mid v_J \geq 0, v_{J^c} = 0, q_J + M_{JJ} v_J = 0, \text{ and } q_{J^c} + M_{J^c J} v_J \geq 0\} \subseteq \mathrm{SOL}(q, M)$ and $z^*(q) \in \mathcal{P}$. Note that $\mathcal{P}$ is a polyhedron, since it is of the form $\{v \mid Av \geq b\}$ with

$$A = \begin{bmatrix} I \\ -I_{J^c \bullet} \\ M \\ -M_{J\bullet} \end{bmatrix} \text{ and } b = \begin{bmatrix} 0 \\ 0 \\ -q \\ q_J \end{bmatrix}.$$

Moreover, it is obvious that $z^*(q) = \arg\min_{Av \geq b} \|v\|$. According to lemma 7.6.12 there exists an index set $K \subseteq \overline{3n}$ such that $z^*(q) = \arg\min_{A_{K\bullet}v = b_K} \|v\|$. Thus, we have $\|z^*(q)\| \leq \alpha(A_{K\bullet})\|b_K\|$. Note that $\|b_K\|^2 \leq \|b\|^2 \leq \|q\|^2 + \|q_J\|^2 \leq 2\|q\|^2$ and $\sqrt{2}\alpha(A_{K\bullet}) \leq \alpha$. Consequently,

$$\|z^*(q)\| \leq \alpha \, \|q\|.$$

□

7.6.8 Proof of theorem 7.4.1

After these preliminary results on LCPs, the proof of the main result on linear passive complementarity systems is in order. The proof will be based on showing that the requirements of theorem 7.3.4 are fulfilled this class of linear complementarity systems.

Lemma 7.6.15 *Consider LCS*(A, B, C, D) *such that assumption 7.2.9 holds and the quadruple* (A, B, C, D) *is passive. Then LCP*$(hC(I - hA)^{-1}\bar{x}, G(h^{-1}))$ *has a unique solution for each* $\bar{x} \in \mathbb{R}^n$ *and all sufficiently small* h *(independent of* $\bar{x}$*).* □

Proof. Lemma 7.6.11 item 4 together with [47, theorem 3.1.6] implies unique solvability for each $\bar{x}$ and all sufficiently small h.
□

Lemma 7.6.16 *Consider LCS*(A, B, C, D) *such that assumption 7.2.9 holds and the quadruple* (A, B, C, D) *is passive. Let* $\tau > 0$ *and* $\mathcal{Q} = \mathrm{SOL}(0, D)$*, i.e.*

$$\mathcal{Q} = \{z \in \mathbb{R}^m \mid z \geq 0, \, Dz \geq 0 \text{ and } z^\top Dz = 0\},$$

be given. Also let $(\{\mathrm{u}_k^h\}, \{\mathrm{x}_k^h\}, \{\mathrm{y}_k^h\})$ *be produced by algorithm 7.3.1. The following statements hold for all sufficiently small* h*.*

1. $C\mathrm{x}_k^h \in \mathcal{Q}^*$ *for all* $k \neq -1$*.*
2. *There exists* $\alpha > 0$ *independent of* x_0 *such that* $\|\mathrm{u}_k^h\| \leq \alpha\|x_0\|$ *for all* $k \neq 0$*.*

□

Proof.

1: It is evident from (7.4b) and (7.4c) that u_k^h solves LCP$(C\mathrm{x}_k^h, D)$ when $k \neq -1$. Since D is nonnegative definite (lemma 7.6.11 item 1), we have that $C\mathrm{x}_k^h \in \mathcal{Q}^*$ due to [47, cor. 3.8.10].

2: All inequalities involving h are meant to hold for all sufficiently small h, and $\alpha_1, \alpha_2, \ldots, \alpha_5$ are suitably chosen positive constants in this proof. Note that LCP$(C\mathrm{x}_k^h, D)$ is solvable for all $k \neq -1$ due to item 1 and [47, corollary 3.8.10]. Let

u^* be the least-norm solution of LCP($C\mathrm{x}_k^h$, D). Clearly, u^* solves also LCP($C\mathrm{x}_k^h - hC(I-hA)^{-1}Bu^*$, $G(h^{-1})$). According to proposition 7.6.10, we have

$$\|\mathrm{u}_{k+1}^h - u^*\| \leq \frac{m^{3/2}}{\mu(G(h^{-1}))}\|C(I-hA)^{-1}\mathrm{x}_k^h - C\mathrm{x}_k^h + hC(I-hA)^{-1}Bu^*\|,$$

since u_{k+1}^h solves LCP($C(I-hA)^{-1}\mathrm{x}_k^h$, $G(h^{-1})$) and $G(h^{-1}) > 0$ for all sufficiently small h. By using the triangle inequality and lemma 7.6.11 item 4, we obtain

$$\|\mathrm{u}_{k+1}^h - u^*\| \leq \frac{\alpha_1}{h}\|C[(I-hA)^{-1} - I]\mathrm{x}_k^h\| + \alpha_1\|C(I-hA)^{-1}Bu^*\|.$$

Note that $(I-hA)^{-1} - I = hA(I-hA)^{-1}$. It can be easily verified that lemma 7.6.2 item 2 and lemma 7.6.14 item 3 result in

$$\|\mathrm{u}_{k+1}^h - u^*\| \leq \alpha_2\|\mathrm{x}_k^h\|. \tag{7.33}$$

Consequently, we get

$$\|\mathrm{u}_{k+1}^h\| \leq \|u^*\| + \|\mathrm{u}_{k+1}^h - u^*\| \leq \alpha_3\|\mathrm{x}_k^h\| \tag{7.34}$$

by applying the triangle inequality and employing lemma 7.6.14 item 3 and (7.33). It follows that

$$\begin{aligned}\|\mathrm{x}_{k+1}^h\| &\leq \|\mathrm{x}_k^h\| + \|\mathrm{x}_{k+1}^h - \mathrm{x}_k^h\| \\ &\leq \|\mathrm{x}_k^h\| + \|[(I-hA)^{-1} - I]\mathrm{x}_k^h + h(I-hA)^{-1}B\mathrm{u}_{k+1}^h\| \quad \text{(from (7.4a))} \\ &\leq (1+\alpha_4 h)\|\mathrm{x}_k^h\|. \quad \text{(from lemma 7.6.2 item 2)}\end{aligned} \tag{7.35}$$

Since $\lim_{h\to 0}(1+\alpha_4 h)^{N_h} = e^{\alpha_4\tau}$ (lemma 7.6.2 item 3), (7.35) implies now that

$$\|\mathrm{x}_k^h\| \leq \alpha_5\|\mathrm{x}_{-1}^h\| = \alpha_5\|x_0\| \tag{7.36}$$

for some $\alpha_5 > 0$. Finally, (7.34) and (7.36) establish the desired inequality. □

After all these preliminaries, we can prove the theorem 7.4.1.

Proof of theorem 7.4.1 According to lemma 7.6.15, assumption 7.3.2 holds. Then, proposition 7.6.3 item 1 implies that RCP(x_0, A, B, C, D) has a unique solution, say $(\hat{u}(s), \hat{y}(s))$. It is known from theorem 5.3.4 that $\hat{u}(s)$ is proper. Therefore, boundedness of $\|h\mathrm{u}_0^h\|$ for all sufficiently small h follows from proposition 7.6.3 item 2. On the other hand, D is nonnegative definite due to item 1 of lemma 7.6.11 and

$$\|u_{reg}^h\| = (\int_0^\tau \|u_{reg}^h(t)\|\,dt)^{1/2} \leq \alpha\tau^{1/2}\|x_0\| \tag{7.37}$$

due to (7.5) and lemma 7.6.16 item 2. Finally, it is known from theorem 7.2.10 that (u, x, y) is the unique solution on $[0, \tau]$ with initial state x_0. As a consequence of

theorem 7.3.4 item 33c, for any sequence $\{h_k\}$, which converges zero, $\{(u^{h_k}, y^{h_k})\}$ converges weakly to (u, y) and $\{x^{h_k}\}$ converges to x. In other words, $\{(u^h, y^h)\}$ converges weakly to (u, y) and $\{x^h\}$ converges to x as h tends zero. □

7.6.9 Proof of theorem 7.4.2

In this subsection, the continuous dependence of the solution trajectories on the initial states will be proven as formulated in theorem 7.4.2.

Proof of theorem 7.4.2 Let the sequence $\{\bar{x}_k\} \subset \mathbb{R}^n$ converge to $\bar{x} \in \mathbb{R}^n$. Denote the solution of LCS(A, B, C, D) on $[0, \tau]$ with the initial states $\bar{x}_k$ and $\bar{x}$ by (u^k, x^k, y^k) and (u, x, y), respectively. Then, it should be shown that

1. $\{(u^k_{imp}, x^k_{imp}, y^k_{imp})\}$ converges $(u_{imp}, x_{imp}, y_{imp})$,
2. $\{(u^k_{reg}, y^k_{reg})\}$ converges weakly to (u_{reg}, y_{reg}) and $\{x^k_{reg}\}$ converges (strongly) to x_{reg}.

1: Let $(u^k_{imp}, x^k_{imp}, y^k_{imp}) = (u^k_0\delta, x^k_0\delta, y^k_0\delta)$. Also let $u^k_0(h)$ and $u_0(h)$ be the solutions of the one-step problems LCP$(C(I-hA)^{-1}\bar{x}_k, hC(I-hA)^{-1}B+D)$ and LCP$(C(I-hA)^{-1}\bar{x}, hC(I-hA)^{-1}B+D)$, respectively. From proposition 7.6.10 and lemma 7.6.11 item 4, we get

$$\|u^k_0(h) - u_0(h)\| \leq \frac{\alpha}{h}\|C(I-hA)^{-1}\|\,\|\bar{x}_k - \bar{x}\|$$

for sufficiently small h. By multiplying the inequality above by h and using lemma 7.6.2 item 2, we obtain

$$\|hu^k_0(h) - hu_0(h)\| \leq \alpha'\|\bar{x}_k - \bar{x}\| \tag{7.38}$$

for sufficiently small h. On the other hand, it is already known from the proof of theorem 7.3.4 item 2 that $\lim_{h\to 0} hu^k_0(h) = u^k_0$ and $\lim_{h\to 0} hu_0(h) = u_0$. Thus, (7.38) yields

$$\|u^k_0 - u_0\| \leq \alpha'\|\bar{x}_k - \bar{x}\|. \tag{7.39}$$

Clearly, $\{u^k_0\}$ converges to u_0 and thus $\{u^k_{imp}\}$ converges to u_{imp}. Since $x^k_{imp} = 0$ and $y^k_{imp} = Du^k_{imp}$, it follows that $\{(u^k_{imp}, x^k_{imp}, y^k_{imp})\}$ converges to $(u_{imp}, x_{imp}, y_{imp})$.

2: Observe that $(u^k_{reg}, x^k_{reg}, y^k_{reg})$ and $(u_{reg}, x_{reg}, y_{reg})$ are the unique solutions of LCS(A, B, C, D) on $[0, \tau]$ with the initial states $\bar{x}_k + Bu^k_0$ and $\bar{x} + Bu_0$, respectively. Moreover, $\{\bar{x}_k + Bu^k_0\}$ converges to $\bar{x} + Bu_0$ as shown in the proof of item 1 above. Lemma 7.6.16 item 2 together with (7.37) implies that for some $\beta > 0$ independent of $\bar{x}_k + Bu^k_0$, $\|u^k_{reg}\| \leq \beta\|\bar{x}_k + Bu^k_0\|$ for all k. This implies that the sequence $\{u^k_{reg}\}$

is bounded, since the sequence $\{\bar{x}_k + Bu_0^k\}$ is convergent. Hence, there exists at least one weakly convergent subsequence of $\{u_{reg}^k\}$ according to lemma 7.6.1 item 3a. Take any such subsequence of $\{u_{reg}^k\}$, say $\{u_{reg}^{k_l}\}$. Define

- $T = T_{(A,B,C,0)}$,
- $S = D$,
- $q_l = Ce^{A\cdot}(\bar{x}_k + Bu_0^k)$, and
- $T_l = T$.

It can be checked that

- T is compact ([170, exercise 4.15]),
- S is nonnegative definite (by lemmma 7.6.11 item 1),
- $\{q_l\}$ converges to $Ce^{A\cdot}(\bar{x}+Bu_0)|_{[0,\tau]}$ (this follows from $\|q_l - Ce^{A\cdot}(\bar{x}+Bu_0)\| \leq \|Ce^{A\cdot}\|\,\|\bar{x}_k - \bar{x}\|$)
- $OCP(q_l, S + T_l)$ is solvable (from proposition 7.6.5 item 1), and
- $\{T_l u_{reg}^{k_l} - T u_{reg}^{k_l}\} = 0$.

As a consequence, the sequence $\{u_{reg}^{k_l}\}$ converges weakly to the solution u_{reg} of $OCP(Ce^{A\cdot}(\bar{x} + Bu_0)|_{[0,\tau]}, T_{(A,B,C,D)})$ according to theorem 7.6.9. Since u_{reg} is unique due to proposition 7.6.5 item 2 and theorem 7.2.10, the reasoning above states that any weakly converging subsequence of $\{u_{reg}^k\}$ has the same limit. Lemma 7.6.1 item 2 implies now that the whole sequence $\{u_{reg}^k\}$ converges weakly to u_{reg}. Note that proposition 7.6.5 item 2 and uniqueness of the solutions of LCS(A, B, C, D) yield that

$$x_{reg}^k = e^{A\cdot}(\bar{x}_k + Bu_0^k)|_{[0,\tau]} + T_{(A,B,I,0)}u_{reg}^k \tag{7.40a}$$

$$y_{reg}^k = Cx_{reg}^k + Du_{reg}^k \tag{7.40b}$$

and

$$x_{reg} = e^{A\cdot}(\bar{x} + Bu_0)|_{[0,\tau]} + T_{(A,B,I,0)}u_{reg} \tag{7.40c}$$

$$y_{reg} = Cx_{reg} + Du_{reg} \tag{7.40d}$$

Then, convergence of $\{x_{reg}^k\}$ to x_{reg} and weak convergence of $\{y_{reg}^k\}$ to y_{reg} follow from (7.40), convergence of $\{\bar{x}_k + Bu_0^k\}$ to $\bar{x} + Bu_0$ and compactness of $T_{(A,B,I,0)}$. □

7.7 Appendix: LCS with low leading row coefficients

In this appendix, we study the consistency of the backward Euler time-stepping scheme for linear complementarity systems given by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{7.41a}$$
$$y(t) = Cx(t) + Du(t) \tag{7.41b}$$
$$0 \leq u(t) \perp y(t) \geq 0 \tag{7.41c}$$

satisfying certain additional conditions. To formulate these conditions, we recall some results from Chapter 3.

Definition 7.7.1 Let (A, B, C, D) be a system with Markov parameters H^i, $i = 0, 1, 2, \ldots$ defined by $H^0 = D$ and $H^i = CA^{i-1}B$ for $i = 1, 2, \ldots$. The leading row coefficients $\rho_1, \ldots, \rho_m$ of (A, B, C, D) are defined for $j \in \bar{m}$ as

$$\rho_j := \inf\{i \in \mathbb{N} \mid H^i_{j\bullet} \neq 0\}$$

with the convention $\inf \emptyset = \infty$. In case the leading row coefficients are all finite, we define the *leading row coefficient matrix* $\mathcal{M}(A, B, C, D)$ as

$$\mathcal{M}(A, B, C, D) := \begin{pmatrix} H^{\rho_1}_{1\bullet} \\ \vdots \\ H^{\rho_m}_{k\bullet} \end{pmatrix}. \tag{7.42}$$

We omit the arguments (A, B, C, D), if they are clear from the context. □

The convergence results in this appendix will be obtained under the following assumption.

Assumption 7.7.2 The leading row coefficients of (A, B, C, D) satisfy $\rho_j \in \{0, 1\}$ for all $j \in \bar{m}$, the leading row coefficient matrix $\mathcal{M}$ is a P-matrix, and D is nonnegative definite. □

We would like to use Theorem 7.3.4 to prove the following result.

Theorem 7.7.3 *Consider the linear complementarity system (7.41) and assume that Assumption 7.7.2 holds. Let $\tau > 0$ and $x_0 \in \mathbb{R}^n$ be given. There exists a sequence $\{h_k\}$ of time steps such that the associated approximations $(u^{h_k}, x^{h_k}, y^{h_k})$ (see (7.5)) generated by the backward Euler time-stepping scheme satisfy the following.*

- *$(u^{h_k}_{reg}, y^{h_k}_{reg})$ converges weakly in $\mathcal{L}^{m+m}_2(0, \tau)$ to (u_{reg}, y_{reg}).*
- *$x^{h_k}_{reg}$ converges in $\mathcal{L}^n_2(0, \tau)$-sense to x_{reg}.*

- *The sequence* $\{(u_{imp}^{h_k}, x_{imp}^{h_k}, y_{imp}^{h_k})\}$ *converges to* $(u_{imp}, 0, y_{imp})$*, where the impulsiv part* (u_{imp}, y_{imp}) *is equal to* $(u_0\delta, y_0\delta)$ *for some* $u_0, y_0 \in \mathbb{R}^m$*.*

- *The triple* (u, x, y) *is a solution of LCS*(A, B, C, D) *on* $(0, \tau)$ *with the initial state* x_0 *in the sense of Definition 7.2.5.*

Moreover, if the solution (u, x, y) *is unique in the sense of Definition 7.2.5, then the above three statements hold for any arbitrary sequence of time steps going to zero.* □

Uniqueness of the solutions in the sense of Definition 7.2.5 can for instance be proven for projected dynamical systems for which the underlying dynamics is linear and the constraint set is a convex polyhedron (under the full rank condition of Chapter 6) by using the argument as in [147, p.33] and exploiting the full rank condition.

7.7.1 Preliminaries

The following results from [47] will be used in the sequel.

Theorem 7.7.4 *Let* $M \in \mathbb{R}^{k\times k}$ *be a P-matrix. For any two vectors* q *and* q' *in* $\mathbb{R}^k$*,*

$$\|z - z'\|_\infty \leq c(M)^{-1}\|q - q'\|_\infty,$$

where z *and* z' *denote the unique solutions to the LCPs* (q, M) *and* (q', M)*, respectively. The constant* $c(M)$ *is defined as*

$$c(M) := \min_{\|z\|_\infty=1} \{\max_{i\in\bar{m}} z_i(Mz)_i\}.$$

□

Proof. See [47, Thm. 7.3.10(a)]. □

A lower bound on $c(M)$ is provided by excercise 5.11.19 in [47].

Theorem 7.7.5 *Let* $M \in \mathbb{R}^{m\times m}$ *be a P-matrix. Define* $\delta(M) := \min\{\sigma(M_{II}) \mid I \subseteq \bar{m}\}$*, where* $\sigma(M_{II})$ *denotes the smallest of the real eigenvalues (if any exists) of* M_{II}*. Moreover,* $\xi(M) := \max_{i\neq j} \|M_{ij}\|$*. Then, the following inequality is true:*

$$c(M) \geq \frac{\delta}{(1 + \frac{\xi(M)}{\delta(M)})^{2(m-1)}}.$$

□

Proof. Excercise 5.11.19 in [47]. □

7.7.2 Proof of the main result

As seen before in this chapter, the resulting LCP that must be solved every time step is given by

$$y_{i+1}^h = C(\mathcal{I} - Ah)^{-1}x_i^h + [C(\frac{1}{h}\mathcal{I} - A)^{-1}B + D]u_{i+1}^h \tag{7.43a}$$

$$0 \leq y_{i+1}^h \perp u_{i+1}^h \geq 0. \tag{7.43b}$$

After computing the solution to this LCP, the state on the next time step can be calculated from

$$x_{i+1}^h = (\mathcal{I} - Ah)^{-1}x_i^h + (\frac{1}{h}\mathcal{I} - A)^{-1}Bu_{i+1}^h \tag{7.44}$$

and a new LCP can be solved again ($i := i + 1$). The step size is taken constant. To approximate a solution trajectory on the interval $(0, \tau)$ for initial state x_0, we set $x_{-1}^h := x_0$ and follow the procedure as described above (see algorithm 7.3.1).

We start by proving that the one-step problem is solvable and the solutions are uniformly bounded.

Theorem 7.7.6 *Suppose that (A, B, C, D) satisfies Assumption 7.7.2. Then the following statements hold.*

1. *LCP$(q, C(\frac{1}{h}\mathcal{I} - A)^{-1}B + D))$ has a unique solution for all $q \in \mathbb{R}^m$ and all sufficiently small $h > 0$.*
2. *For all $i \geq 1$ it holds that $C_{K\bullet}x_i^h \geq 0$ with $K := \{j \in \bar{m} \mid \rho_j = 1\}$.*
3. *Let a fixed $\tau > 0$ be given, then there exists an $\alpha > 0$ such that $\|u_i^h\| \leq \alpha\|x_0\|$ for all $i = 1, \ldots, \lceil \frac{\tau}{h} \rceil$ and all sufficiently small h.*
4. *Let x_0 be given. For all sufficiently small h, it holds that $\|hu_0^h\| \leq \alpha$ for some α.*

□

Remark 7.7.7 Note that for $i = 0$ $C_{K\bullet}x_i^h \geq 0$ does not necessarily hold. As a consequence (see the proof of the theorem), the bound given in statement 3 does not hold for $i = 0$. According to Chapter 3, the condition $C_{K\bullet}x_0 \geq 0$ is equivalent to x_0 being a *regular state*, i.e. a state from which no re-initialization is required before smooth continuation is possible. □

Proof. Statement *1* will be proven during the proof of statement *3*. Statement *2* follows from the observation that $C_{K\bullet}x_{i+1}^h = (y_{i+1,}^h)_K \geq 0$, because $D_{K\bullet} = 0$.

Without loss of generality we may assume that $K := \{i \in \bar{m} \mid \rho_i = 1\} = \{l+1, \ldots, m\}$ for some $l \in \bar{m}$ (otherwise re-arrange the complementarity pairs). Note that this implies that

$$\mathcal{M} = \begin{pmatrix} D_{K^c\bullet} \\ (CB)_{K\bullet} \end{pmatrix}.$$

Moreover, $G(h^{-1})$ can be factorized as

$$G(h^{-1}) = \begin{pmatrix} \mathbb{1} & 0 \\ 0 & h\mathbb{1} \end{pmatrix} \begin{pmatrix} D_{K^c\bullet} + (CB)_{K^c\bullet}h + \dots \\ (CB)_{K\bullet} + (CAB)_{K\bullet}h + \dots \end{pmatrix} =: \Lambda(h)V(h^{-1}) \quad (7.45)$$

with $\lim_{h\downarrow 0} V(h^{-1}) = \mathcal{M}$.

By premultiplying (7.43a) by $\Lambda^{-1}(h)$ and observing that $y^h_{i+1} \geq 0$ if and only if $\Lambda^{-1}(h)y^h_{i+1} \geq 0$, it follows that u^h_{i+1} is a solution to (7.43) if and only if it is a solution to

$$\tilde{y}^h_{i+1} = \Lambda^{-1}(h)C(\mathbb{1} - Ah)^{-1}x^h_i + V(h^{-1})u^h_{i+1} \quad (7.46a)$$

$$0 \leq \tilde{y}^h_{i+1} \perp u^h_{i+1} \geq 0, \quad (7.46b)$$

where $\tilde{y}^h_{i+1}$ is defined as $\Lambda^{-1}(h)y^h_{i+1}$. Note that for sufficiently small h, $V(h^{-1})$ is a P-matrix and consequently, this one-step problem has a unique solution (thereby proving statement *1*). To arrive at the boundedness of u^h_i independently of h, we will utilize Thm. 7.7.4 and 7.7.5 for $V(h^{-1})$. Since $\lim_{h\downarrow 0} V(h^{-1}) = \mathcal{M}$, it follows that ξ as in Theorem 7.7.5 satisfies $\xi(V(h^{-1})) \leq \alpha_0$ for all h sufficiently small. Moreover, $\lim_{h\downarrow 0} V(h^{-1}) = \mathcal{M}$ implies that the quantity $\delta(V(h^{-1}))$ as in Theorem 7.7.5 satisfies $\lim_{h\downarrow 0} \delta(V(h^{-1})) = \delta(\mathcal{M})$, because the eigenvalues of a (sub)matrix depend continuously on the entries of the matrix. Hence, there exists an $0 < \varepsilon < \delta(\mathcal{M})$ such that for sufficiently small h, it holds that

$$0 < \alpha_1 := \delta(\mathcal{M}) - \varepsilon \leq \delta(V(h^{-1})) \leq \delta(\mathcal{M}) + \varepsilon. \quad (7.47)$$

Theorem 7.7.5 implies now that

$$c(V(h^{-1}))^{-1} \leq \frac{\left(1+\frac{\xi(V(h^{-1}))}{\delta(V(h^{-1}))}\right)^{2(m-1)}}{\delta(V(h^{-1}))} \leq \frac{\left(1+\frac{\alpha_0}{\alpha_1}\right)^{2(m-1)}}{\alpha_1} =: \alpha_2.$$

This upperbound for $c(V(h^{-1}))^{-1}$ and Theorem 7.7.4 yield now that for any vector q in $\mathbb{R}^k$,

$$\|u^h_{i+1} - z\|_\infty \leq \alpha_2 \|\Lambda^{-1}(h)C(\mathbb{1} - Ah)^{-1}x^h_i - q\|_\infty, \quad (7.48)$$

where z denotes the unique solution of the LCP(q, $V(h^{-1})$). The idea is to find a q such that $z = 0$ and a suitable bound on u^h_{i+1} is obtained. Note that

$$\Lambda^{-1}(h)C(\mathbb{1} - Ah)^{-1}x^h_i = \begin{pmatrix} C_{K^c\bullet}(\mathbb{1} - Ah)^{-1}x^h_i \\ h^{-1}C_{K\bullet}(\mathbb{1} - Ah)^{-1}x^h_i \end{pmatrix}. \quad (7.49)$$

If we take (see statement 2 of the theorem)

$$q = \begin{pmatrix} 0 \\ h^{-1}C_{K\bullet}x^h_i \end{pmatrix} \geq 0,$$

it is obvious that the unique solution to LCP(q, $V(h^{-1})$) is equal to $z = 0$. According to (7.49), there exists an $\alpha_3 > 0$ such that

$$\|\Lambda^{-1}(h)C(\mathbb{1} - Ah)^{-1}x_i^h - q\|_\infty = \| \begin{pmatrix} C_{K^c\bullet}(\mathbb{1} - Ah)^{-1} \\ C_{K\bullet}A + C_{K\bullet}A^2\frac{h}{2} + .. \end{pmatrix} x_i^h \|_\infty \leq \alpha_3 \|x_i^h\|_\infty$$

for all sufficiently small h. Equation (7.48) implies now that for sufficiently small h

$$\|u_{i+1}^h\|_\infty \leq \underbrace{\alpha_2\alpha_3}_{=:\alpha_4} \|x_i^h\|_\infty. \tag{7.50}$$

To complete the proof, we have to bound $\|x_i^h\|$ for all $i = 1, 2, \ldots, \lceil \frac{\tau}{h} \rceil$. By applying the triangle inequality and using (7.50), we obtain

$$\begin{aligned} &\|x_{i+1}^h\|_\infty \leq \|x_i^h\|_\infty + \|x_{i+1}^h - x_i^h\|_\infty \leq \\ &\|x_i^h\|_\infty + \|[(\mathbb{1} - hA)^{-1} - \mathbb{1}]x_i^h + h(\mathbb{1} - hA)^{-1}Bu_{i+1}^h\|_\infty \leq (1 + \alpha_5 h)\|x_i^h\|_\infty \end{aligned} \tag{7.51}$$

for some $\alpha_5 > 0$ and all sufficiently small h. Since $\lim_{h\to 0}(1+\alpha_5 h)^{\lceil \frac{\tau}{h} \rceil} = e^{\alpha_5 \tau}$, (7.51) implies that

$$\|x_i^h\| \leq \alpha_6 \|x_0\| \tag{7.52}$$

for some $\alpha_6 > 0$ and all sufficiently small h. The proof is now completed by combining (7.50) and (7.52).

Statement *4* is proven as follows. Denote the unique solution (observe that $G(\sigma) := C(\sigma\mathbb{1} - A)^{-1}B + D$ is a P-matrix for sufficiently large $\sigma \in \mathbb{R}$) of the rational complementarity problem RCP($(C(s\mathbb{1} - A)^{-1}x_0, C(s\mathbb{1} - A)^{-1}B + D)$) by $(u(s), y(s))$. It can be seen from the results in the beginning of Section 3.5.2 that $u(s)$ must be proper. Indeed, since $(u(s), y(s))$ is a solution to an RCP, there exists an index set $I \subseteq \bar{k}$ such that $u_{I^c}(s) \equiv 0$ and $u_I(s) = -G_{II}^{-1}(s)C_{I\bullet}(s\mathbb{1} - A)^{-1}x_0$. By replacing h by s^{-1} in the decomposition of (7.45) and using the diagonality of $\Lambda(s)$, we obtain that $G_{II}(s) = \Lambda_{II}(s^{-1})V_{II}(s)$ with $\lim_{s\to\infty} V_{II}(s) = \mathcal{M}_{II}$. Invertibility of $\mathcal{M}_{II}$ follows from the fact that its determinant is equal to a principal minor of the P-matrix $\mathcal{M}$. This means that the rational matrix $V_{II}(s)$ has a proper inverse. Since the inverse of $\Lambda(s^{-1})$ has at most a polynomial part of degree one, the expression for $u_I(s)$ and the strict properness of $C_{I\bullet}(s\mathbb{1} - A)^{-1}x_0$ yield that $u(s)$ is proper.

Since the unique solution (u_0^h, y_0^h) to LCP($h^{-1}C(h^{-1}\mathbb{1} - A)^{-1}x_0, C(h^{-1}\mathbb{1} - A)^{-1}B + D$) coincides with $(h^{-1}u(h^{-1}), h^{-1}y(h^{-1}))$ for sufficiently small h and $u(s)$ is proper, the result follows. □

The proof of the main result 7.7.3 in this appendix follows now by combining theorem 7.3.4 and the theorem we have just proven.

8

Concluding remarks

8.1 Summary of contributions

The contributions in this thesis are oriented towards the fundamental issues for a class of discontinuous dynamical systems. Questions related to the solution concept, well-posedness and reliable numerical schemes are, of course, of independent interest. However, such a rigorous foundation is also indispensable for the analysis of the dynamical behavior (stability, controllability, observability, etc.) and controller synthesis. The results presented here can be looked upon as the basis needed for developing systematic controller design methodologies.

8.1.1 Linear complementarity systems

Specific forms of complementarity systems have been used for a long time in particular applications such as mechanical systems with inelastic unilateral constraints and electrical networks with ideal diodes; see for instance the work of Lötstedt [124]. The idea of coupling complementarity conditions to a *general* input/output dynamical system was first put forward by Van der Schaft and Schumacher [177]. It was noted in this paper that to formulate complete dynamics for complementarity systems one needs to specify mode selection rules and jump rules, and on both topics proposals were formulated. The mode selection rule proposed in [177] is fairly simple (it is not based on the solution of a complementarity problem), and in the case of several pairs of complementary variables it leads to results that are not always satisfactory. In particular, the mode selection rule is not consistent with physical laws for mechanical systems with impacts. An alternative mode selection rule was proposed in [179] for nonlinear dynamics in the case of smooth continuations only. However, the rule was not complete as it did not solve the mode selection problem when impulsive motions are required. Under the assumption of linear dynamics, the rule proposed in [179] was extended to general (not necessarily smooth) continuations in Chapter 3 of this thesis.

In this way, a new class of dynamical systems called *linear complementarity systems* has been introduced based on the equations

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) && \text{(8.1a)} \\ y(t) &= Cx(t) + Du(t) && \text{(8.1b)} \\ 0 \leq y(t) &\perp u(t) \geq 0. && \text{(8.1c)} \end{aligned}$$

The definition of this class of dynamical systems and the study of its properties constitute the main contributions of this thesis.

(Linear) complementarity systems have interesting connections to existing research areas. Firstly, linear complementarity systems can be seen as dynamical extensions of the linear complementarity problem and form as such a bridge between linear system theory and mathematical programming. Secondly, the combination of inequalities and differential equations causes the system description to be of hybrid nature as it contains both continuous and discrete dynamics. As a consequence, (linear) complementarity systems form a subclass of hybrid dynamical systems. Thirdly, the equations (8.1a)-(8.1b) form a standard state space description in input/state/output form. Hence, adding the complementarity relations means 'closing the loop' of a linear control system by a discontinuous feedback. From this point of view, linear complementarity systems are related to control theory. Finally, it has been shown that unilaterally constrained mechanical systems, projected dynamical systems, optimal control problems with inequality constraints, electrical networks with diodes and piecewise linear systems allow a description in terms of the complementarity formalism. Although the complementarity systems seem of a rather specific form at first sight, it turns out that it is a nontrivial class of dynamical systems with many interesting fields of application.

The links to these existing research fields motivate the study of (linear) complementarity systems in the sense that the obtained results have a broad range of applications and may also yield ideas that are extendable to more general classes of hybrid dynamical systems.

8.1.2 Solution concept

The specification of the dynamics of a linear complementarity system in this thesis is based on the notion of an *initial solution*, which itself uses the impulsive-smooth distributional theory that has been developed in [83]. To support the contention that the solution concept that we obtain in this way is physically relevant, we show that the concept agrees with Moreau's formulation in case of unilaterally constrained linear mechanical systems with inelastic impacts. Moreover, the results on electrical networks with ideal diodes (Chapter 5) indicate that the solution concept complies with applications in circuit theory as well.

8.1.3 Well-posedness

The largest part of this thesis is concerned with the problem of existence and uniqueness of solutions. Three phenomena have been described that could obstruct the global existence (i.e. on the interval $[0, \infty)$) of solutions. A first problem could be that no initial solution exists from a given initial state ("deadlock"). This means that neither a smooth continuation nor a re-initialization is possible. In case deadlock can be excluded and uniqueness of initial solutions is true (for arbitrary initial state), the system might be called *initially* well-posed. An initially well-posed system does not necessarily have solutions (starting from a given initial state) on an interval of the form $[0, \varepsilon)$ for some $\varepsilon > 0$. The reason is the possible occurrence of a sequence of only re-initializations taking place at one time instant without convergence of the corresponding event states. Indeed, in this case it is not possible to define a smooth continuation after the (infinitely many) jumps. A *locally* well-posed system does not display such behavior by definition and guarantees consequently the existence and uniqueness of solutions on a time interval with positive length. Finally, a finite (right-)accumulation point of event times may prevent a locally well-posed system from being *globally* well-posed, when the left limit of the state variable does not exist at the accumulation point. In this case the solution cannot be defined beyond the accumulation point. The investigation of all these phenomena in the context of linear complementarity systems has resulted in the contributions summarized below.

Lötstedt [124] showed existence and uniqueness of *smooth* continuations for a class of mechanical complementarity systems. Necessary and sufficient conditions for local well-posedness of linear complementarity systems with one pair of complementary variables ("bimodal systems") were provided by Van der Schaft and Schumacher in [177]. The same authors gave in [179] a sufficient condition for existence and uniqueness of *smooth* continuations in real-analytic nonlinear complementarity systems. The results in this thesis pertain to linear complementarity systems so that all initial conditions may be considered, including the ones that give rise to impulsive solutions. A sufficient condition for local well-posedness of systems in this class has been given. The condition is more general than the one obtained in [179] since there is no assumption of "uniform relative degree". Instead, it is required that the leading row and column coefficient matrices have positive principal minors. From the proof of this result, we obtained two interesting byproducts. Firstly, the set of regular states (states from which smooth continuation is possible without re-initialization) has explicitly been characterized. Secondly, it has been shown that after, at most, one re-initialization, smooth continuation is possible. In terms of multiplicities, this means that every event time has, at most, multiplicity one. These results immediately apply to linear mechanical systems subject to independent unilateral constraints. In addition, we proved for bimodal systems and linear complementarity systems whose leading row coefficients are either zero or one that global existence of solutions is guaranteed under the 'local well-posedness conditions.'

The conditions for local well-posedness are sufficient and the question arises,

whether they are necessary as well. We studied this problem for bimodal systems and Theorem 3.6.10 states the *equivalence* between "local well-posedness," "global well-posedness" and the leading Markov parameter being positive (in this case equivalent to the condition of the leading row and column coefficient matrices having only positive principal minors). The investigation of well-posedness has been continued by the study of the *rational complementarity problem* (RCP), which resulted in necessary and sufficient conditions for initial well-posedness. To be specific, we proved that the existence and uniqueness of an initial solution (or equivalently, the existence and uniqueness of an allowed re-initialization or a smooth continuation) is equivalent to unique solvability of a family of linear complementarity problems (LCPs). The strength of this connection is that dynamical properties of a linear complementarity system are related to solvability characteristics of static problems for which an extensive literature is already available. This strength has been demonstrated by showing initial well-posedness for linear mechanical systems with (possibly dependent) inequality constraints, linear relay systems and electrical networks with ideal diodes.

In Chapter 5, the results on existence and uniqueness of initial solutions have been generalized to obtain global existence of solutions and much stronger statements on uniqueness for linear passive electrical networks with ideal diodes. The passivity of the underlying state space description has turned out to be an elegant assumption, which results in detailed information on the nature of the solutions. In addition to proving global existence of solutions, we have shown that derivatives of Dirac impulses do not occur in the solution trajectories, that Dirac impulses and discontinuities in the state variable occur only at the initial time $t = 0$, and that the set of event times is right-isolated. The interpretation of the latter result is that for all time instants there exists a positive length time interval in which the diodes do not change from conducting to blocking or vice versa. Note that this excludes the existence of left-accumulation points in the set of event times. Furthermore, we have explicitly characterized the set of regular states in terms of the dual cone of the solution set to the homogeneous LCP associated with the 'feedthrough term' D.

From a more general point of view, the results on existence and uniqueness of solutions, presented in the thesis, contribute to fill a gap present in hybrid systems theory in which studies of well-posedness are rare.

8.1.4 Time-stepping methods

In addition to the contributions to the event driven methods by the results on mode selection (RCP, LCP and LDCP) and re-initialization, the main emphasis – from a numerical point of view – has been to provide a rigorous mathematical basis for time-stepping. In particular, we concentrated on the time-stepping method based on the well-known backward Euler integration formula. In practice, this method has already proven to be useful for the transient simulation of piecewise linear electrical circuits [20, 120, 121] and constrained mechanical systems [125, 140, 155, 192, 194]. The advantages of the method for linear complementarity systems are that it is straightforward to im-

plement, and that many algorithms (provided by Lemke [47], Katzenelson [109] and others [121]) are available to explicitly solve the one-step problems consisting of linear complementarity problems. However, as illustrated by an example in Chapter 7, one should not indiscriminately apply time-stepping methods to approximate solutions of arbitrary linear complementarity systems. Also the hybrid nature of the dynamics and the fact that the event times are not traced exactly cause the consistency of the method to be uncertain. As a consequence, it is essential to identify classes of linear complementarity systems for which consistency of the numerical scheme can be shown. In his work on the existence of solutions to nonlinear mechanical complementarity systems, Stewart [192, 193] has shown the convergence of time-stepping approximations for a suitably chosen sequence of time steps. A similar result has been proven in this thesis for linear complementarity systems with leading row coefficient matrices being either zero or one under the condition that the leading row coefficient matrix has only positive principal minors and the 'feedthrough term' D is nonnegative definite. A stronger result has been proven for linear passive complementarity systems (linear passive circuits with ideal diodes). The main contribution states that for any arbitrary sequence (and not only one special sequence) of time steps, which tends to zero, the corresponding approximations converge to the true transient solution of the network model. The same result holds for the simulation of a class of projected dynamical systems of which the defining vector field is linear.

8.1.5 Applications and generalizations

In Chapter 2 an overview of several applications of complementarity systems has been given. In spite of the special form of the complementarity conditions, many interesting classes of dynamical systems can be described by the complementarity formalism. This opens many possibilities to transfer and extend results from one subdomain of complementarity systems to another or even to the whole class.

In Chapter 6 projected dynamical systems, which are used for studying e.g. economical markets, transportation networks and international trade, have also been rewritten as complementarity systems. As an interesting bonus, we have obtained a new, and, in our opinion, more direct proof of the global existence of solutions for these *nonlinear* complementarity systems. Obviously, it is very useful to identify many classes of dynamical systems that allow a complementarity reformulation. The reason is that the available literature on these systems may serve as a potential source of knowledge to obtain analysis and design tools for complementarity systems.

8.2 Open problems and ideas for further research

The objective of the final section of the thesis is to indicate open problems, which should receive further attention in the future. Suggestions for possible starting points are also presented.

8.2.1 Nonlinear complementarity systems

This thesis is mainly concerned with *linear* complementarity systems. Only Chapter 6 deals with projected dynamical systems, which result in gradient-type complementarity systems for which the underlying state space description is in general nonlinear. Although this seems to be the only exception, the results of Chapter 3 allow a direct generalization to smooth continuations for *nonlinear* complementarity systems thereby extending Theorem 3.2 in [179]. We can formulate the following assertion.

Consider the complementarity system

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t) \tag{8.2a}$$
$$y(t) = h(x(t)) \tag{8.2b}$$

with complementarity conditions on u and y. Here, f is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$, g from $\mathbb{R}^n$ to $\mathbb{R}^{n\times k}$ and h from $\mathbb{R}^n$ to $\mathbb{R}^k$, which are sufficiently smooth.

For $x_0 \in \mathbb{R}^n$ we define the i-th leading row coefficient $\rho_i(x_0)$ as

$$\rho_i(x_0) := \inf\{j \in \mathbb{N} \setminus \{0\} \mid L_g L_f^{j-1} h_i(x_0) \neq 0\} \tag{8.3}$$

and the index set $J(x_0)$ as

$$J(x_0) := \{j \in \bar{k} \mid (h_j(x_0), \ldots, L_f^{\rho_j(x_0)-1} h_j(x_0)) = 0\}, \tag{8.4}$$

where L denotes the "Lie-derivative" (see e.g. [151]).

Theorem 8.2.1 *Consider the complementarity system (8.2) with f, g and h real-analytic. Consider $x_0 \in \mathbb{R}^n$ such that the matrix*

$$(L_{g_{\bullet j}} L_f^{\rho_i(x_0)-1} h_i(x_0))_{i,j\in J(x_0)} \tag{8.5}$$

has only positive principal minors. There exists an $\varepsilon > 0$ such that a unique real-analytic solution exists on $[0, \varepsilon)$ if and only if $(h_i(x_0), \ldots, L_f^{\rho_i(x_0)-1} h_i(x_0))$ is lexicographically nonnegative for all $i \in \bar{k}$. □

This theorem generalizes [179, Thm.3.2] in the sense that a uniform relative degree (i.e. $\rho_1(x_0) = \rho_2(x_0) = \ldots = \rho_k(x_0)$) is not required and the theorem has a 'local' character, because only a submatrix of the 'leading row coefficient matrix' needs to have positive principal minors.

However, the above result (like [179, Thm.3.2]) only deals with *smooth* continuations and this immediately touches upon one of the most essential problems in the nonlinear context. The absence of a general formulation of the re-initialization rules (impulsive motions) is a major open issue for deriving the complete dynamics. Currently, there is little known about a nonlinear equivalent for the jump space T_I, which should describe the "projection directions" of the re-initializations (see Chapter 3) for nonlinear mode dynamics of the form $f(\dot{x}, x) = 0$. This problem needs to be solved

before a suitable solution concept can be introduced. The interested reader is referred to [182] for an exposition on this problem. Of course, the study of well-posedness has to be reconsidered, although it may benefit from the ideas proposed in the thesis. The problem will naturally be more complex. One reason is the absence of a tool similar to the rational complementarity problem, which has played a crucial role in much of the well-posedness results obtained here.

8.2.2 Elastic impact rule

A key motivation for the physical relevance of the solution concept presented in Chapter 3 is the relation to the inelastic impact rule as proposed by Moreau [144] (see also [31, 139]) for unilaterally constrained mechanical systems. The projection operator P_I, i.e. the projection onto the consistent subspace V_I along the jump space T_I of mode I, corresponds to inelastic collisions. As noted in the main text, if mode I is selected from initial state x_0 (i.e. $I \in \mathcal{I}(x_0)$), x_0 is decomposed as $x_0 = v + t$ with $v \in V_I$ and $t \in T_I$. The re-initialized state is then equal to $P_I x_0 = v$. A question that arises is whether the "(partial) mirroring in V_I along T_I" defined by the operator $Q_I^e x_0 = v - et$ corresponds physically to the elastic impact case, where $0 < e < 1$ denotes the restitution coefficient (as in Newton's restitution law). In particular, is the re-initialization for the *completely elastic* impact case governed by the operator Q_I^1 ($e = 1$)?

One could even go one step further and consider the possibility to specify the restitution coefficients for every contact separately. To be specific, consider the linear complementarity system

$$\dot{x} = \underbrace{\begin{pmatrix} 0 & I \\ -M^{-1}K & -M^{-1}D \end{pmatrix}}_{A} x + \underbrace{\begin{pmatrix} 0 \\ M^{-1}E^\top \end{pmatrix}}_{B} u \tag{8.6a}$$

$$y = \underbrace{(E \;\; 0)}_{C} x \tag{8.6b}$$

$$0 \le y \perp u \ge 0 \tag{8.6c}$$

corresponding to the mechanical system

$$M\ddot{q}(t) + D\dot{q}(t) + Kq(t) = 0 \text{ subject to } Eq(t) \ge 0$$

as described in Chapter 3. Suppose that the restitution coefficient e_i is associated to the constraint $E_{i\bullet}q(t) \ge 0$. For the inelastic impact case, the re-initialization is given by $x(0+) = x_0 + Bu^0$, if the impulsive part $\mathtt{u}_{imp}$ of the initial solution $(\mathtt{u}, \mathtt{x}, \mathtt{y})$ for initial state x_0 is equal to $u^0\delta$. Since u_i^0 is the multiplier associated with the constraint $E_{i\bullet}q(t) \ge 0$, one may wonder if the re-initialization defined by $x_0 + \sum_{i\in\bar{k}}(1+e_i)B_{\bullet i}u_i^0$ makes sense in this context. Of course, in this new setting the questions of well-posedness have to be reconsidered, although it can be easily established that the 'initial well-posedness' results remain valid with this modified jump rule.

8.2.3 Global existence

The results in Chapters 3 and 4 describe mainly initial and local well-posedness results. Only in case of bimodal system, linear complementarity systems with low leading row coefficients, passive linear complementarity systems, linear relay systems (see [94, 123]) and projected dynamical systems, we have obtained global well-posedness results. Hence, a major part of the class of (linear) complementarity systems (being locally well-posed) is still not covered.

As mentioned before, the problem of extending local existence of solutions to global existence for linear complementarity systems is caused by accumulations of event times (Zeno trajectories): the durations of the smooth continuations in the successive continuous phases get smaller and smaller such that the event times accumulate and converge to a finite limit. If the state trajectory does not converge, continuation is not possible beyond the limit of the event times (using the solution concept proposed here). The proofs of the global existence results are all based on showing that the limit of the event states does exist in these circumstances. For nonlinear complementarity systems an extra phenomenon that may obstruct global existence is the occurrence of 'finite escape times' within the continuous phases.

An alternative method to prove global existence is the use of approximations (e.g. time-stepping or smoothing methods) for the system's equations. The line of reasoning to obtain global existence generally consists of proving that the approximating systems are solvable on an arbitrary time interval, that the solutions converge and that the limit is a solution to the original system. Such arguments have been used by Stewart [192, 193] to prove global existence of solutions to unilaterally constrained mechanical systems. Also in this thesis, similar arguments yield an alternative proof for global well-posedness for the class of linear passive complementarity systems. Indeed, the main result on consistency of the backward Euler time-stepping method in Chapter 7 shows global existence as a byproduct and could replace the (more direct) reasoning of Chapter 5.

8.2.4 $\mathcal{L}_2$-uniqueness

The uniqueness results obtained in the Chapters 3 and 4 state uniqueness in the sense of the solution concept of Chapter 3. This solution concept is posed in a 'forward sense' implying that solutions with left-accumulations of event times as in Example 1.4.4 are not allowed. This means for Example 1.4.4, that there is only one 'forward' solution starting from the origin. However, adopting an $\mathcal{L}_2$-solution concept as in Theorem 5.4.17 results in multiple solutions starting from the origin and consequently, $\mathcal{L}_2$-uniqueness does not hold. Hence, one should distinguish clearly between the possible concepts of uniqueness.

It is natural to consider time to be asymmetric for hybrid systems. However, a disadvantage of a system with solutions being only 'unique in a forward sense' (and not in $\mathcal{L}_2$-sense) is that we can only prove the convergence of a *subsequence* of the approximated time functions as obtained from the backward Euler time-stepping

scheme. Convergence of the whole sequence does not follow as it was the case in Chapter 7 for linear passive complementarity systems. This problem obstructs, for instance, the proof of consistency – in the sense of showing convergence of any *arbitrary* sequence of approximating functions – for the numerical scheme applied to linear complementarity systems with low leading row coefficients. $\mathcal{L}_2$-uniqueness was shown for linear passive complementarity systems and projected dynamical systems, but how to prove similar results for more general situations remains an open problem. Of course, other notions of uniqueness (in different function spaces) may be of interest as well.

8.2.5 Accumulation of event times

As seen in the previous two subsections, right-accumulations of event times may obstruct global existence of solutions, while left-accumulations may prevent that the solution trajectories are $\mathcal{L}_2$-unique. It is useful and interesting to characterize the situations when the phenomenon of accumulation of event times occurs and when it does not. In the simple case of a bimodal linear complementarity system with $D = 0$ and CB positive, the existence of accumulation of event times can be excluded by using a result from [60] (see [94]). However, we are not aware of any results in other situations.

8.2.6 Moreau's sweeping process

A widely studied dynamical system is the sweeping process of Moreau [31, 139, 140]. The sweeping process describes the motion of a particle in a moving set. The motion of a closed convex set is given by the multi-valued function $t \to C(t)$ with $C(t) \subseteq \mathbb{R}^n$. The dynamics is described by the first-order differential inclusion [140]

$$-\frac{dx}{dt} \in N_{C(t)}(x(t)), \tag{8.7}$$

where $N_{C(t)}(x(t))$ denotes the (outward) normal cone of $C(t)$ in the point $x(t)$ (see Chapter 6). The inclusion (8.7) describes that the particle is at rest when it is contained in the interior of $C(t)$ and is moving only (motion described by x) when it is caught-up by the boundary of $C(t)$.

We conjecture that if the moving set $C(t)$ is given by a finite collection of inequalities, i.e.

$$C(t) = \{z \in \mathbb{R}^n \mid c_i(z, t) \geq 0 \text{ for all } i \in \bar{k}\},$$

then there are close connections between the sweeping process and complementarity systems under certain additional conditions. Intuitively, the proofs and additional assumptions must resemble the ones given in Chapter 6 for projected dynamical systems.

The complementarity system will (probably) look like

$$\dot{x}(t) = \sum_{i=1}^{k} [\nabla c_i(x,t)]^\top u_i \tag{8.8a}$$

$$y_i(t) = c_i(x,t) \tag{8.8b}$$

with complementarity conditions on $(u_i(t), y_i(t))$. Here we use $\nabla c_i(z,t)$ to denote the gradient of $c_i(z,t)$ with respect to z, which is considered to be a row vector.

The details of this problem still have to be filled in, but it is expected that the material contained in Chapter 6 and [140] will be of great help. The consequences of such a connection have to be studied and results from complementarity systems can possibly be transferred to the sweeping process and vice versa. In particular, an alternative proof for existence of solutions (similar to the one in Chapter 6) could possibly be given for the sweeping process. Attention should also be paid to the relations between the numerical methods for the sweeping process and complementarity systems.

8.2.7 Stability

As indicated in [18], the derivation of general (computationally tractable) methods for the determination of stability is extremely complicated even for the most elementary (discrete-time) hybrid systems (see (1.1) above). Also the paper [122] indicates many open problems in the field of stability of *switched systems* and does not suggest that the problem of stability is solvable for a broad subclass of switched or hybrid systems.

It is well-known that the stability of all mode dynamics is not sufficient for the stability of the switching system. Vice versa, the instability of all the modes does not exclude the stability of the overall system. Hence, the stability problem cannot be solved by studying properties of the continuous phases separately (except in certain special cases). New methodologies are needed that incorporate both the switching logic and the continuous dynamics to access the stability for hybrid systems.

Extensions of Lyapunov methods to switched or hybrid systems are the most popular. Researchers have tried to obtain conditions for various notions of stability by using common Lyapunov functions [25, 148, 184, 185], multiple Lyapunov functions [26, 27, 101, 157], switching based on Lyapunov functions [134], piecewise quadratic Lyapunov functions [106, 107], convex homogeneous Lyapunov functions [163], and converse Lyapunov theorems [138]. Of course, the problem of explicitly constructing a suitable Lyapunov function is often extremely difficult and easily verifiable conditions are consequently not obtained in this way.

As (linear) complementarity systems have a clear additional structure, one might be optimistic and believe that efficient methods exist to answer the questions of stability. The unilaterally constrained mechanical systems and linear passive complementarity systems can be proven to be stable by using a quadratic Lyapunov functions (which is a common Lyapunov function for all linear mode dynamics and also the

re-initializations). Interestingly, this Lyapunov function satisfies both standard continuous time as well as discrete time (for the projection operators describing the re-initializations) Lyapunov inequalities. This suggests (as in [106, 107]) that parts of the question of stability of hybrid systems may be brought in the realm of the theory of linear matrix inequalities. However, the following example indicates that even for special subclasses of linear complementarity systems, where impulsive motions are absent, the problem of stability is far from trivial. The problem of determining whether there exists a matrix F such that the trajectories

$$\dot{x}(t) = Ax(t) + B \max(0, Fx(t)) \tag{8.9}$$

of x are square integrable (and consequently, $\lim_{t\to\infty} x(t) = 0$) cannot be answered by standard methodologies [95]. This problem is inspired by the control of linear systems with a positivity constraint on the attainable control values. Only the following rather simple result has been proven so far.

Theorem 8.2.2 *Suppose that (A, B) has scalar input and A has at most one pair of unstable, complex conjugate eigenvalues. There exists an F such that (8.9) has only $\mathcal{L}_2$-solution trajectories if and only if (A, B) is stabilizable (in the ordinary sense) and A has no eigenvalues contained in $\mathbb{R}_+ := [0, \infty)$.* □

Although the assumptions of this theorem are rather restrictive, it appears to be very helpful for the stabilization of surge in compressors [205]. The positivity constraint in the compressor is due to the control valve that can only attain values between zero (valve closed) and its maximal capacity (fully open).

Since the problem mentioned above falls within the category of stability problems for linear complementarity systems, it indicates that the stability problems are, in general, also difficult for this class. Of course, this renders the development of new techniques that could (partially) answer stability issues for (specific subclasses) of complementarity systems both interesting and challenging.

8.2.8 Controller synthesis

Control of hybrid systems is now widely recognized as a key area of research and investigation, both at a fundamental and at an experimental level. As mentioned in the introduction, systematic synthesis tools are currently not available for general hybrid systems. It would therefore be a giant step forwards, if structured design methodologies could be developed for the subclass of (linear) complementarity systems especially in view of the possible applications mentioned in Chapter 2.

One possible method could be based on utilizing the results on time-stepping obtained in this thesis and in [192, 193]. For smooth systems it is common practice to use sampled data control and to design a controller on the basis of a discretized version of the system. This methodology can be extended to complementarity systems (under certain conditions), because accurate discretized models can be obtained by

time-stepping techniques as shown in this thesis. Since such a discretized model can be rewritten in a discrete-time piecewise linear description (for which stabilization and control problems have already been studied before [15, 190]), this opens several possibilities to controller synthesis for complementarity systems. As a consequence, it is worthwhile to identify other classes of complementarity systems for which the backward Euler time-stepping method is consistent. Our conjecture is that for linear complementarity systems containing derivatives of Dirac impulses in the solution trajectories, the proposed numerical scheme does not generate proper output in general. The reason is that a derivative of a Dirac impulse cannot be approximated by non-negative (step) functions. The investigation of new time-stepping methods covering more general cases is therefore an interesting open problem as well. Currently, the consistency of the backward Euler time-stepping routine is being investigated for relay systems.

Other approaches for control design could be categorized as "generalizing" and "specializing" techniques. *Generalizing* techniques are related to the extension of control methods developed for subclasses of complementarity systems. For instance, the control methods developed for unilaterally constrained mechanical systems in [32] may also be applicable to electrical networks or piecewise linear systems. *Specializing* methodologies aim at applying techniques and concepts used for general hybrid systems to the class of complementarity systems.

Which methods will be successful is not clear at this moment. However, it is obvious that complementarity systems – as common meeting ground of several mature research areas – have the potential to play a major role in developing systematic methods to overcome analysis and synthesis problems in a wide range of applications. The work in this thesis forms a step in this direction, as it solves various fundamental problems, needed for setting up a general system and control theory for complementarity systems.

Bibliography

[1] R. Alur and T.A. Henzinger. Real-time logics: complexity and expressiveness. *Information and Computation*, 104:35–77, 1993.

[2] R. Alur, T.A. Henzinger, and E.D. Sontag, editors. *Hybrid Systems III.* (Proc. of the Workshop on Verification and Control of Hybrid Systems, New Brunswick, New Jersey, October 1995.), volume 1066 of *Lecture Notes in Computer Science*. Springer, 1996.

[3] B.D.O. Anderson and S. Vongpanitlerd. *Network Analysis and Synthesis. A Modern Systems Theory Approach.* Pentice-Hall, Englewood Cliffs, New Jersey, 1973.

[4] M. Andersson, S.E. Mattsson, D. Brück, and T. Schöntal. Omsim - An integrated evironment for object-oriented modelling and simulation. In *Proceedings of the IEEE/IFAC joint symposium on Computer-Aided Control System Design,* Tucson, Arizona, pages 285–290, 1994.

[5] P. Antsaklis, W. Kohn, A. Nerode, and S. Sastry, editors. *Hybrid Systems II.* (Proc. Workshop on Hybrid Systems, Cornell University, USA, October 1994.), volume 999 of *Lecture Notes in Computer Science*. Springer, 1995.

[6] P. Antsaklis, W. Kohn, A. Nerode, and S. Sastry, editors. *Hybrid Systems IV.* (Proc. of the Fourth Intern. Workshop on Hybrid Systems, Ithaca, New York, October 1996.), volume 1273 of *Lecture Notes in Computer Science*. Springer, 1997.

[7] P.J. Antsaklis and A. Nerode (guest eds.). Special issue on hybrid control systems. *IEEE Transactions on Automatic Control*, 43(4), 1998.

[8] E. Artin and O. Schreier. Algebraische Konstruktion reeller Körper. *Abh. Math. Sem. Univ. Hamburg*, 5:85–99, 1927.

[9] J.P. Aubin and A. Cellina. *Differential Inclusions.* Springer, Berlin, 1984.

[10] D.D. Bainov and P.S. Simeonov. *Systems with Impulse Effects. Stability, Theory and Applications.* Ellis Horwood Series in Mathematics and its Applications. Ellis Horwood, Chichester, 1989.

[11] P. Ballard. The dynamics of discrete mechanical systems with perfect unilateral constraints. Submitted to *Archive for Rational Mechanics and Analysis.*

[12] B. Baraff. Issues in computing contact forces for non-penetrating rigid bodies. *Algorithmica*, 10:292–352, 1993.

[13] D. Bedrosian and J. Vlach. Time-domain analysis of networks with internally controlled switches. *IEEE Transactions on Circuits and Systems-I*, 39(3):199–212, 1992.

[14] D.A. van Beek, S.H.F. Gordijn, and J.E. Rooda. Integrating continuous-time and discrete-event concepts in modelling and simulation of manufacturing machines. *Journal of simulation practice and theory*, 5:653–669, 1997.

[15] A. Bemporad and M. Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407–428, 1999.

[16] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979.

[17] C.J. Bett and M.D. Lemmon. Bounded amplitude performance of switched LPV systems with applications to hybrid systems. *Automatica*, 35(3):491–503, 1999.

[18] V.D. Blondel and J.N. Tsitsiklis. Complexity of stability and controllability of elementary hybrid systems. *Automatica*, 35(3):479–490, 1999.

[19] R.K. Boel, B. De Schutter, G. Nijsse, J.M. Schumacher, and J.H. van Schuppen. Modelling, analysis and control of hybrid systems. Submitted to *Journal A*.

[20] W.M.G. van Bokhoven. *Piecewise Linear Modelling and Analysis*. Kluwer, Deventer, the Netherlands, 1981.

[21] W.M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 1975.

[22] J.M. Borwein and M.A.H. Dempster. The linear order complementarity problem. *Math. Oper. Research*, 14(3):534–558, 1989.

[23] P.P.J. van den Bosch, H. Butler, A.R.M. Soeterboek, and M.M.W.G. Zaat. *Modelling and simulation with PSI/c*. BOZA Automatisering BV, Nuenen, The Netherlands, 1995.

[24] P.P.J. van den Bosch and W.P.M.H. Heemels. Hybrid systems: modelling embedded controllers. To appear in *Journal A*, 1999.

[25] S. Boyd and Q. Yang. Structured and simultaneous Lyapunov functions for system stability problems. *International Journal of Control*, 49(6):2215–2240, 1989.

[26] M.S. Branicky. Stability of hybrid systems: state of the art. In *Proceedings of the 36th Conference on Decision and Control,* San Diego (USA), pages 120–125, 1997.

[27] M.S. Branicky. Stability theory for hybrid dynamical systems. *IEEE Transactions on Automatic Control*, 43(4):475–482, 1998.

[28] M.S. Branicky, V.S. Borkar, and S.K. Mitter. A unified framework for hybrid control: model and optimal control theory. *IEEE Transactions on Automatic Control*, 43(1):31–45, 1998.

[29] K.E. Brenan, S.L. Campbell, and L.R. Petzhold. *Numerical solution of initial-value problems in differential-algebraic equations*, volume 14 of *Classics in Applied Mathematics*. 1996.

[30] R.W. Brockett. Hybrid models for motion control systems. In *Essays in Control* H.L. Trentelman and J.C. Willems (eds.), pages 29–53. Birkhauser, Boston, 1993.

[31] B. Brogliato. *Nonsmooth Impact Mechanics. Models, Dynamics and Control*, volume 220 of *Lecture Notes in Control and Information Sciences*. Springer, London, 1996.

[32] B. Brogliato, S.-I. Niculescu, and P. Orthant. On the control of finite-dimensional mechanical systems with unilateral constraints. *IEEE Transactions on Automatic Control*, 42(2):200–215, 1997.

[33] B. Brogliato and A. Zavala-Rio. On the control of complementary-slackness mechanical juggling systems. *IEEE Transactions on Automatic Control*, 45(3), 2000.

[34] D. de Bruin and P.P.J. van den Bosch. Measurement of the lateral vehicle position with permanent magnets. In *Proc. IFAC workshop on Intelligent Components for Vehicles (ICV'98)* Seville, Spain, pages 9–14, 1998.

[35] M.K. Çamlıbel, W.P.M.H. Heemels, and J.M. Schumacher. Dynamical analysis of linear passive networks with ideal diodes. Part II: Consistency of a time-stepping method. Submitted for publication.

[36] M.K. Çamlıbel, W.P.M.H. Heemels, and J.M. Schumacher. The nature of solutions to linear passive complementarity systems. In *38-th IEEE Conference on Decision and Control*, Phoenix (USA), 1999.

[37] M.K. Çamlıbel and J.M. Schumacher. Well-posedness of a class of piecewise-linear systems. In *Proceedings of the European Control Conference*, Karlsruhe (Germany), 1999.

[38] M.H. Chang and E.J. Davison. Adaptive switching control of LTI MIMO systems using a family of controllers approach. *Automatica*, 35(3):453–465, 1999.

[39] Zhou Chaochen, C.A.R. Hoare, and A.P. Ravn. A calculus of durations. *Information Processing Letters*, 40(5):269–276, 1991.

[40] M.J. Chien. Existence and computation of DC solutions of nonlinear networks in a bounded set. *IEEE Transactions on Circuits and Systems*, 23(11):655–663, 1976.

[41] M.J. Chien. Piecewise-linear theory and computation of solutions of homeomorphic resistive networks. *IEEE Transactions on Circuits and Systems*, 24(3):118–127, 1977.

[42] M.J. Chien and E.S. Kuh. Solving nonlinear resistive networks using piecewise-linear analysis and simplicial subdivision. *IEEE Transactions on Circuits and Systems*, 23(6):305–317, 1977.

[43] L.O. Chua and An-Chang Deng. Canonical piecewise-linear modeling. *IEEE Transactions on Circuits and Systems*, 33(5):511–525, 1986.

[44] L.O. Chua and R.L.P. Ying. Canonical piecewise-linear analysis. *IEEE Transactions on Circuits and Systems*, 30(3):125–140, 1983.

[45] M. Ciampa, P. Terreni, and M. Poletti. Conditions for the existence and uniqueness of DC solutions of networks containing nonlinear OpAmps with ideal models. In *IEEE International Symposium on Circuits and Systems*, volume 4, pages 2498–2501, 1993.

[46] R.W. Cottle. On a problem in linear inequalities. *Journal of the London Mathematical Society*, 43:378–384, 1968.

[47] R.W. Cottle, J.-S. Pang, and R.E. Stone. *The Linear Complementarity Problem*. Academic Press, Boston, 1992.

[48] A.A. ten Dam. *Unilaterally constrained dynamical systems*. PhD-thesis Rijksuniversiteit Groningen, Dept. of Mathematics, Groningen, The Netherlands, 1997.

[49] A.A. ten Dam, K.F. Dwarshuis, and J.C. Willems. The contact problem for linear continuous-time dynamical systems: a geometric approach. *IEEE Transactions on Automatic Control*, 42(4):458–472, 1997.

[50] B. DasGupta and E.D. Sontag. A polynomial-time algorithm for an equivalence problem which arises in hybrid systems theory. In *Proceedings of the 37th IEEE Conference on Decision and Control*, Tampa, Florida, pages 1629–1634, 1998.

[51] R. David and H. Alla. Petri nets for modelling of dynamic systems - a survey. *Automatica*, 30(2):175–202, 1994.

[52] J. Davies. *Specification and proof in real-time CSP*. Cambridge University Press, 1993.

[53] B. De Moor, L. Vandenberghe, and J. Vandewalle. The generalized linear complementarity problem and an algorithm to find all its solutions. *Mathematical Programming*, 57(3):415–426, 1992.

[54] B. De Schutter and B. De Moor. The extended linear complementarity problem. *Mathematical Programming*, 71:289–325, 1995.

[55] B. De Schutter and B. De Moor. The linear dynamic complementarity problem is a special case of the extended linear complementarity problem. *Systems & Control Letters*, 34(1-2):63–75, 1998.

[56] A. Dervisoglu. State equations and initial values in active RLC networks. *IEEE Transactions on Circuit Theory*, 18:544–547, 1971.

[57] A. Deshpande, A. Göllü, and P. Varaiya. SHIFT: a formalism and a programming language for dynamic networks and hybrid automata. In *[6]*, pages 113–133, 1997.

[58] C.A. Desoer and J. Katzenelson. Nonlinear RLC networks. *The Bell System Technical Journal*, 44:161–198, 1965.

[59] J.J. DiStefano, A.R. Stubberud, and I.J. Williams. *Theory and problems of feedback and control systems.* Schaum's outline series. McGraw-Hill, 1967.

[60] A.L. Dontchev and I.V. Kolmanovsky. State constraints in the linear regulator problem: case study. *Journal of Optimization Theory and Applications*, 87(2):323–347, 1995.

[61] P. Dupuis. Large deviations analysis of reflected diffusions and constrained stochastic approximation algorithms in convex sets. *Stochastics*, 21:63–96, 1987.

[62] P. Dupuis and A. Nagurney. Dynamical systems and variational inequalities. *Annals of Operations Research*, 44:9–42, 1993.

[63] J.T.J. van Eijndhoven. Solving the linear complementarity problem in circuit simulation. *SIAM Journal on Control and Optimization*, 24(5):1050–1062, 1986.

[64] E. Feron. Quadratic stabilizability of switched systems via state and output feedback. Technical Report CICS-P-468, Center for intelligent control systems, Massachusetts Institute of Technology, Cambridge, 1996.

[65] M.C. Ferris and J.S Pang. Engineering and economic applications of complementarity problems. *SIAM Review*, 39(4):669–713, 1997.

[66] M. Fiedler and V. Pták. On matrices with non-positive off-diagonal elements and positive principal minors. *Czechoslovak Mathematical Journal*, 12:382–400, 1962.

[67] A. F. Filippov. Differential equations with discontinuous right-hand side. *Matemat. Sbornik.*, 51:99–128, 1960. In Russian. English translation: *Am. Math. Soc. Transl.* 62 (1964).

[68] A.F. Filippov. *Differential Equations with Discontinuous Righthand Sides.* Mathematics and Its Applications. Kluwer, Dordrecht, The Netherlands, 1988.

[69] T. Fujisawa and E.S. Kuh. Some results on existence and uniqueness of solutions of nonlinear networks. *IEEE Transactions on Circuit Theory*, 18(5):501–506, 1971.

[70] T. Fujisawa and E.S. Kuh. Piecewise-linear theory of nonlinear networks. *SIAM Journal on Applied Mathematics*, 22:307–328, 1972.

[71] C.W. Gear. *Numerical initial value problems in ordinary differential equations.* Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[72] F. Génot and B. Brogliato. New results on Painlevé paradoxes. *European Journal of Mechanics. A/Solids*, 18:653–678, 1998.

[73] D.N. Godbole, J. Lygeros, and S. Sastry. Hierarchical hybrid control: a case study. In *[5]*, pages 166–190, 1995.

[74] A.J. Goldman. Resolution and separation theorems for polyhedral convex sets. *Linear Inequalities and Related Systems,* Annals of Mathematics studies, No. 38, Eds. H.W. Kuhn and A.W. Tucker, pages 41–52, 1956.

[75] M. Green and A.N. Willson Jr. On the uniqueness of a circuit's DC operating point when its transistors have variable current gains. *IEEE Transactions on Circuits and Systems*, 36(12):1521–1528, 1989.

[76] F.L. Grossman, A. Nerode, A.P. Ravn, and H. Rischel, editors. *Hybrid Systems.* (Proc. Workshop on the Theory of Hybrid Systems, Lyngby, Denmark, October 1992.), volume 736 of *Lecture Notes in Computer Science.* Springer, 1993.

[77] G. Güzelis and I. Göknar. A canonical representation for piecewise linear affine maps and its application to circuit analysis. *IEEE Transactions on Circuits and Systems*, 38(11):1342–1354, 1991.

[78] I. Han and B.J. Gilmore. Multi-body impact motion with friction - analysis, simulation and experimental validation. *ASME Journal of Mechanical Design*, 115:412–422, 1993.

[79] P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Math. Progr. Ser. B*, 48:161–220, 1990.

[80] R.F. Hartl, S.P. Sethi, and R.G. Vickson. A survey of the maximum principles for optimal control problems with state constraints. *SIAM Review*, 37(2):181–218, 1995.

[81] A. Hassibi and S. Boyd. Quadratic stabilization and control of piecewise-linear systems. In *Proceedings of the American Control Conference*, Philadelphia (USA), pages 3659–3664, 1998.

[82] M.L.J. Hautus. The formal Laplace transform for smooth linear systems. *Mathematical Systems Theory, Proceedings International Symposium, Udine (Italy).* Lecture Notes in Economics and Mathematical Systems, 131:29–47, 1975.

[83] M.L.J. Hautus and L.M. Silverman. System structure and singular control. *Linear Algebra and its Applications*, 50:369–402, 1983.

[84] W.P.M.H. Heemels, M.K. Çamlıbel, and J.M. Schumacher. Dynamical analysis of linear passive networks with ideal diodes. Part I: well-posedness. Submitted for publication.

[85] W.P.M.H. Heemels, R.J.A. Gorter, A. van Zijl, P.P.J. van den Bosch, S. Weiland, W.H.A. Hendrix, and M.R. Vonder. Asynchronous measurement and control: a case study on motor synchronisation. To appear in *Control Engineering Practice*.

[86] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Projected dynamical systems in a complementarity formalism. Submitted for publication.

[87] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Complete description of dynamics in the linear complementary-slackness class of hybrid systems. *IEEE Conference on Decision and Control '97* in San Diego (USA), pages 1243–1248, 1997.

[88] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Complementarity problems in linear complementarity systems. *Proceedings of the American Control Conference'98*, Philadelphia (USA), pages 706–710, 1998.

[89] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Dissipative systems and complementarity conditions. In *Proceedings of the 37-th IEEE Conference on Decision and Control'98*, Tampa (USA), pages 4127–4132, 1998.

[90] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Mode selection in linear complementarity systems. In *Proceedings of the IFAC Conference on System Structure and Control*, Nantes (France), pages 153–158, 1998.

[91] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Applications of complementarity systems. In Proceedings of the European Control Conference *in Karlsruhe (Germany)*, 1999.

[92] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Linear complementarity systems. To appear in *SIAM J. Appl. Math.*, 1999.

[93] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. The rational complementarity problem. *Linear Algebra and its Applications*, 294(1-3):93–135, 1999.

[94] W.P.M.H. Heemels, J.M. Schumacher, and S. Weiland. Well-posedness of linear complementarity systems. In *38-th IEEE Conference on Decision and Control,* Phoenix (USA), 1999.

[95] W.P.M.H. Heemels and A.A. Stoorvogel. Positive stabilizability of a linear continuous-time system. Technical Report 98 I/01, Eindhoven University of Technology, Dept. of Electrical Engineering, Measurement and Control Systems, Eindhoven, The Netherlands, 1998.

[96] W.P.M.H. Heemels, S.J.L. van Eijndhoven, and A.A. Stoorvogel. Linear quadratic regulator problem with positive controls. In *Proceedings of the European Control Conference,* Brussels (Belgium), 1997.

[97] W.P.M.H. Heemels, S.J.L. van Eijndhoven, and A.A. Stoorvogel. Linear quadratic regulator problem with positive controls. *International Journal of Control*, 70(4):551–578, 1998.

[98] M.F. Heertjes, M.J.G. van den Molengraft, J.J. Kok, and D.H. van Campen. Vibration reduction of a harmonically excited beam with one-sided springs using sliding computed torque control. *Dynamics and control*, 7:361–375, 1997.

[99] T.A. Henzinger and S. Sastry, editors. *Hybrid Systems: Computation and Control.* (Proc. First Internat. Workshop on Hybrid Systems: Computation and Control, Berkeley, USA, April 1998.), volume 1386 of *Lecture Notes in Computer Science*. Springer, 1998.

[100] C.A.R. Hoare. *Communicating sequential processes.* Prentice-Hall, 1985.

[101] L. Hou, A.N. Michel, and H. Ye. Stability analysis of switched systems. In *Proceedings 35th Conference on Decision and Control,* Kobe (Japan), pages 1208–1212, 1996.

[102] J.-I. Imura and A.J. van der Schaft. Characterization of well-posedness of piecewise linear systems. Memorandum 1475, University of Twente, Enschede, The Netherlands, 1998.

[103] N. Inaba and S. Mori. Chaos via torus breakdown in a piecewise-linear forced van der Pol oscillator with a diode. *IEEE Transactions on Circuits and Systems*, 38(4):398–409, 1991.

[104] M. Jirstrand. *Constructive methods for inequality constraints in control.* PhD-thesis, Linköping University, Dept. of Electrical Engineering, Linköping, Sweden, 1998.

[105] K.H. Johansson. *Relay feedback and multivariable control.* PhD-thesis, Lund Institute of Technology, Dept. of Automatic Control, Lund, Sweden, 1997.

[106] M. Johansson. *Piecewise linear control systems.* PhD-thesis, Lund Institute of Technology, Dept. of Automatic Control, Lund, Sweden, 1999.

[107] M. Johansson and A. Rantzer. Computation of piecewise quadratic Lyapunov functions for hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):555–559, 1998.

[108] C. Kahlert and L.O. Chua. A generalized canonical piecewise linear representation. *IEEE Transactions on Circuits and Systems*, 37(3):373–382, 1990.

[109] J. Katzenelson. An algorithm for solving nonlinear resistor networks. *Bell Syst. Tech. J.*, 44:1605–1620, 1965.

[110] T.A.M. Kevenaar. *PLANET, a hierarchical network simulator.* PhD-thesis Eindhoven University of Technology, Dept. of Electrical Engineering, Eindhoven, The Netherlands, 1993.

[111] T.A.M. Kevenaar and D.M.W. Leenaerts. Comparison of piecewise-linear model descriptions. *IEEE Transactions on Circuit and Systems-I*, 39(12):996–1004, 1992.

[112] A. Klarbring. A mathematical programming approach to contact problems with friction and varying contact surface. *Computers & Structures*, 30(5):1185–1198, 1986.

[113] A.J. Koivo. *Fundamentals for control of robotic manipulators.* Wiley, New York, 1989.

[114] T.C. Koopmans (editor). *Activity analysis of production and allocation.* John Wiley & Sons, 1951.

[115] H.W. Kuhn and A.W. Tucker. Nonlinear programming. In *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, 1951.

[116] M. Kuijper and J. M. Schumacher. Input-output structure of linear differential/algebraic systems. *IEEE Transactions on Automatic Control*, 38:404–414, 1993.

[117] B.C. Kuo. *Automatic Control Systems.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

[118] K.K. Lee and A. Arapostathis. On the controllability of piecewise-linear hypersurface systems. *Systems & Control Letters*, 9:89–96, 1987.

[119] D.M.W. Leenaerts. *TOPICS, a contribution to analog design automation.* PhD-thesis Eindhoven University of Technology, Dept. of Electrical Engineering, Eindhoven, The Netherlands, 1992.

[120] D.M.W. Leenaerts. On linear dynamic complementary systems. *IEEE Transactions on Circuits and Systems-I*, 46(8):1022–1026, 1999.

[121] D.M.W. Leenaerts and W.M.G. van Bokhoven. *Piecewise linear modelling and analysis.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[122] D. Liberzon and A.S. Morse. Benchmark problems in stability and design of switched systems. *Nonlinear Control Abstracts*, 10, 1999.

[123] Y.J. Lootsma, A.J. van der Schaft, and M.K. Çamlıbel. Uniqueness of solutions of relay systems. *Automatica*, 35(3):467–478, 1999.

[124] P. Lötstedt. Mechanical systems of rigid bodies subject to unilateral constraints. *SIAM Journal on Applied Mathematics*, 42(2):281–296, 1982.

[125] P. Lötstedt. Numerical simulation of time-dependent contact and friction problems in rigid body mechanics. *SIAM Journal on Scientific and Statistical Computing*, 5:370–393, 1984.

[126] J.H.A. Ludlage. *Controllability analysis of industrial processes. Towards the industrial application.* PhD-thesis Eindhoven University of Technology, Dept. of Electrical Engineering, Eindhoven, The Netherlands, 1997.

[127] D.G. Luenberger. *Optimization by Vector Space Methods.* Wiley, Chichester, 1969.

[128] H. Lütkepohl. *Handbook of Matrices.* Wiley, New York, 1996.

[129] J. Lygeros, K.H. Johansson, S. Sastry, and M. Egerstedt. On the existence and uniqueness of executions of hybrid automata.

[130] N. Lynch, R. Segala, F. Vaandrager, and H.B. Weinberg. Hybrid I/O automata. In *[2]*, pages 496–510, 1996.

[131] M. Mabrouk. On a unified variational model for the dynamics of perfect unilateral constraints. *European Journal of Mechanics. A/Solids*, 17(5):819–842, 1998.

[132] J.M. Maciejowski. *Multivariable Feedback Design.* Addison-Wesley Publishers, 1989.

[133] O. Maler, editor. *Hybrid and Real-Time Systems.* (Proc. Intern. Workshop HART'97, Grenoble, France, March 1997.), volume 1201 of *Lecture Notes in Computer Science*, Berlin, 1997. Springer.

[134] J. Malmborg. *Analysis and design of hybrid control systems.* PhD-thesis , Lund Institute of Technology, Dept. of Automatic Control, Lund, Sweden, 1998.

[135] O.L. Mangasarian. *Nonlinear Programming.* McGraw-Hill, New York, 1969.

[136] A. Massarini, U. Reggiani, and K. Kazimierczuk. Analysis of networks with ideal switches by state equations. *IEEE Transactions on Circuits and Systems-I*, 44(8):692–697, 1997.

[137] S.E. Mattsson, H. Elmqvist, and J.F. Broenink. Modelica: an international effort to design the next generation modelling language. *Journal A*, 38(3):16–19, 1997.

[138] A.P. Molchanov and Y.S. Pyatnitskiy. Criteria of absolute stability of differential and difference inclusions encountered in control theory. *System and Control Letters*, 13:59–64, 1989.

[139] M.D.P. Monteiro Marques. *Differential Inclusions in Nonsmooth Mechanical Problems. Shocks and Dry Friction.* Progress in Nonlinear Differential Equations and their Applications. Birkhäuser, Basel, 1993.

[140] J.J. Moreau. Numerical aspects of the sweeping process. *Preprint.*

[141] J.J. Moreau. Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement pôlaires. *C.R. Académie des Sciences Paris*, 255:238–240, 1962.

[142] J.J. Moreau. Les liaisons unilatérales et le principe de Gauss. *C.R. Académie des Sciences Paris*, 256:871–874, 1963.

[143] J.J. Moreau. Approximation en graphe d'une évolution discontinue. *R.A.I.R.O. Analyse numérique/Numerical Analysis*, 12:75–84, 1978.

[144] J.J. Moreau. Unilateral contact and dry friction in finite freedom dynamics. In *Nonsmooth mechanics and applications (*Eds. J.J. Moreau and P.D. Panagiotopoulos), Springer. International centre for mechanical sciences, Courses and Lectures 302, pages 1–82, 1988.

[145] A.S. Morse, S.S. Pantelides, S. Sastry, and J.M. Schumacher (guest eds.). A special issue on hybrid systems. *Automatica*, 35(3), 1999.

[146] Y. Murakami. A method for the formulation and solution of circuits composed of switches and linear RLC networks. *IEEE Transactions on Circuits and Systems*, 34:496–509, 1987.

[147] A. Nagurney and D. Zhang. *Projected Dynamical Systems and Variational Inequalities with Applications*. Kluwer, Boston, 1996.

[148] K.S. Narendra and J. Balakrishnan. A common Lyapunov function for stable LTI systems with commuting A-matrices. *IEEE Transactions on Automatic Control*, 39(12):2469–2471, 1994.

[149] D. Nešić, E. Skafidas, I.M.Y. Mareels, and R.J. Evans. On the use of switched linear controllers for stabilizability of implicit recursive equations. *Proceedings of the American Control Conference'98*, Philadelphia (USA), pages 3639–3643, 1998.

[150] R.O. Nielsen and A.N. Willson Jr. Topological criteria for establishing the uniqueness of solutions. *IEEE Transactions on Circuits and Systems*, 24(7):349–362, 1977.

[151] H. Nijmeijer and A.J. van der Schaft. *Nonlinear dynamical control systems*. Springer, 1990.

[152] T. Ohtsuki, T. Fujisawa, and S. Kumagai. Existence theorems and a solution algorithm for piecewise-linear resistor networks. *SIAM Journal on Mathematical Analysis*, 8(1):69–99, 1977.

[153] T. Ohtsuki and H. Watanabe. State-variable analysis of RLC networks containing nonlinear coupling elements. *IEEE Transactions on Circuit Theory*, 18(1):26–38, 1969.

[154] A. Opal and J. Vlach. Consistent initial conditions of nonlinear networks with switches. *IEEE Transactions on Circuits and Systems*, 38(7):698–710, 1991.

[155] L. Paoli and M. Schatzman. Schéma numérique pour un modèle de vibrations avec constraintes unilatérales et perte d'énergie aux impacts, en dimension finie. *C.R. Académie des Sciences Paris Sér. I Math.*, 317:211–215, 1993.

[156] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Springer-Verlag, New York, 1983.

[157] P. Peleties and R.A. DeCarlo. Asymptotic stability of m-switched systems using Lyapunov-like functions. In *Proceedings of the American Control Conference*, Boston (USA), pages 1679–1684, 1991.

[158] D. Percivale. Uniqueness in the elastic bounce problem. *Journal of Differential Equations*, 56:206–215, 1985.

[159] A.L. Peressini. *Ordered Topological Vector Spaces*. Harper & Row Publishers, 1967.

[160] F. Pfeiffer and C. Glocker. *Multibody Dynamics with Unilateral Contacts*. Wiley, Chichester, 1996.

[161] A. Pnueli. The temporal logic of programs. In *Proceedings of the 18th Annual Symposium on the Foundations of Computer Science*, pages 46–57. IEEE Computer Science Press, New York, 1977.

[162] A. Pnueli and J. Sifakis (guest eds.). Special issue on hybrid systems. *Theoretical Computer Science* 138, 1995.

[163] A. Yu. Pogromsky, M. Jirstrand, and P. Spångéus. On stability and passivity of a class of hybrid systems. In *Proceedings of the 37-th IEEE Conference on Decision and Control'98*, Tampa (USA), pages 3705–3710, 1998.

[164] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mishchenko. *The mathematical theory of optimal processes*. John Wiley & Sons, 1962.

[165] V.C. Prasad. Necessary and sufficient condition for uniqueness of solutions of certain piecewise linear resistive networks containing transistors and diodes. *Electronic Letters*, 28(13), 1992.

[166] V.C. Prasad and V.P. Prakash. Existence, uniqueness and determination of solutions of certain piecewise linear resistive networks. In *IEEE International Symposium on Circuits and Systems*, volume 2, pages 1470–1473, 1990.

[167] A. Prestel. *Lectures on Formally Real Fields*, volume 1093 of *Lecture Notes in Mathematics*. Springer-Verlag, 1984.

[168] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[169] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1976.

[170] W. Rudin. *Functional Analysis*. McGraw-Hill, New York, 1977.

[171] H. Samelson, R.M. Thrall, and O. Wesler. A partition theorem for Euclidean n-space. *Proc. Amer. Math. Soc.*, 9:805–807, 1958.

[172] I.W. Sandberg. Theorems on the computation of the transient response of nonlinear networks containing transistors and diodes. *Bell System Technical Journal*, 49:1739–1776, 1970.

[173] I.W. Sandberg and A.N. Willson Jr. Some theorems on the properties of DC equations of nonlinear networks. *Bell System Technical Journal*, 48:1–35, 1969.

[174] I.W. Sandberg and A.N. Willson Jr. Existence and uniqueness of solutions for the equations of nonlinear DC networks. *SIAM Journal on Applied Mathematics*, 22:173–186, 1972.

[175] S.S. Sastry and C.A. Desoer. Jump behaviour of circuits and systems. *IEEE Transactions on Circuits and Systems*, 28(12):1109–1124, 1981.

[176] A.J. van der Schaft. *System Theoretic Descriptions of Physical Systems*. CWI Tract 3. CWI, Amsterdam, 1984.

[177] A.J. van der Schaft and J.M. Schumacher. The complementary-slackness class of hybrid systems. *Mathematics of Control, Signals and Systems*, 9:266–301, 1996.

[178] A.J. van der Schaft and J.M. Schumacher. Hybrid systems described by the complementarity formalism. In *[133]*, pages 403–408, 1997.

[179] A.J. van der Schaft and J.M. Schumacher. Complementarity modelling of hybrid systems. *IEEE Transactions on Automatic Control*, 43(4):483–490, 1998.

[180] A.J. van der Schaft and J.M. Schumacher. *An introduction to hybrid dynamical systems*, volume 251 of *Lecture Notes in Control and Information Sciences*. Springer, London, 1999.

[181] M. Schatzman. A class of nonlinear differential equations of second order in time. *Nonlinear Analysis, Theory, Methods & Applications*, 2(3):355–373, 1978.

[182] J.M. Schumacher. Re-initialization in discontinuous systems. *Open problems in mathematical systems and control theory* Eds. V.D. Blondel, E.D. Sontag, M. Vidyasagar and J.C. Willems, pages 203–209, 1999.

[183] L. Schwartz. *Théorie des Distributions*. Hermann, Paris, 1951.

[184] H. Shim, D.J. Noh, and J.H. Seo. Common Lyapunov function for exponentially stable nonlinear systems. In *4th SIAM Conference on Control and its Apllications*, 1998.

[185] R.N. Shorten and K.S. Narendra. A sufficient condition for the existence of a common Lyapunov function for two second order linear systems. In *Proceedings of the 36th Conference on Decision and Control*, San Diego (USA), pages 3521–3522, 1997.

[186] R.N. Shorten and K.S. Narendra. On the existence of a common Lyapunov function for linear stable switching systems. In *Proceedings of the 10th Yale Workshop on Adaptive and Learning Systems*, pages 130–140, 1998.

[187] R.W. Sierenberg and O.B. de Gans. Personal Prosim: a fully integrated simulation environment. In *Proceedings of the 1992 European simulation symposium*, Dresden (Germany), pages 167–173.

[188] A.V. Skorokhod. Stochastic equations for diffusions in a bounded region. *Theory of Probability and its Applications*, 6:264–274, 1961.

[189] U. Söderman and J.-E. Strömberg. Switched bond graphs: Multiport switches, mathematical characterization and systematic composition of computational models. Technical report, Linköping University, Dept. of Computer and Information Science, Sweden, available at `http://www.ida.liu.se/ext/pur/enter/`, 1995.

[190] E.D. Sontag. Nonlinear regulation: the piecewise linear approach. *IEEE Transactions on Automatic Control*, 26(2):346–357, 1981.

[191] S.N. Stevens and P.-M. Lin. Analysis of piecewise-linear resistive networks using complementary pivot theory. *IEEE Transactions on Circuits and Systems*, 28(5):429–441, 1981.

[192] D.E. Stewart. Convergence of a time-stepping scheme for rigid body dynamics and resolution of Painlevé's problem. *Archive for Rational Mechanics and Analysis*, 145(3):215–260, 1998.

[193] D.E. Stewart. Time-stepping methods and the mathematics of rigid body dynamics. Chapter 1 of *Impact and Friction,* A. Guran, J.A.C. Martins and A. Klarbring (eds.), Birkhäuser, 1999.

[194] D.E. Stewart and J.C. Trinkle. An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and Coulomb friction. *Int. Journal for Numerical Methods in Engineering*, 39:2673–2691, 1996.

[195] L. Tavernini. Differential automata and their discrete simulators. *Nonlinear Analysis, Theory, Methods & Applications*, 11(6):665–683, 1987.

[196] P. Terwiesch, E. Scheiben, A.J. Petersen, and T. Keller. A digital real-time simulator for rail-vehicle control system testing. In *[133]*, pages 199–212, 1997.

[197] J. Tolsa and M. Salichs. Analysis of linear networks with inconsistent initial conditions. *IEEE Transactions on Circuits and Systems-I*, 40(12):885–894, 1993.

[198] C. Tomlin, G. Pappas, J. Lygeros, D. Godbole, and S. Sastry. Hybrid models of next generation of air traffic management. In *[6]*, pages 378–404, 1997.

[199] J.C. Trinkle, J.-S. Pang, S. Sudarsky, and G. Lo. On dynamic multi-rigid-body contact problems with Coulomb friction. *Zeitschrift für Angewandte Mathematik und Mechanik*, 77(4):267–279, 1997.

[200] V.I. Utkin. Variable structure systems with sliding modes. *IEEE Transactions on Automatic Control*, 22(1):31–45, 1977.

[201] L. Vandenberghe, B.L. De Moor, and J. Vandewalle. The generalized linear complementarity problem applied to the complete analysis of resistive piecewise-linear circuits. *IEEE Transactions on Circuits and Systems*, 36(11):1382–1391, 1989.

[202] V.M. Veliov and M.I. Krastanov. Controllability of piecewise linear systems. *Systems & Control Letters*, 7:335–341, 1986.

[203] J. Vlach, J.M. Wojciechowski, and A. Opal. Analysis of nonlinear networks with inconsistent initial conditions. *IEEE Transactions on Circuits and Systems-I*, 42(4):195–200, 1995.

[204] M.A. Wicks, P. Peleties, and R.A. DeCarlo. Construction of piecewise Lyapunov functions for hybrid systems. In *Proceedings of the 33rd IEEE Conference on Decision and Control*, Lake Buena Vista, Florida, pages 3492–3497, 1994.

[205] F. Willems, M. Heemels, B. de Jager, and A. Stoorvogel. Positive feedback stabilization of compressor surge. In *38-th IEEE Conference on Decision and Control*, Phoenix (USA), 1999.

[206] J.C. Willems. Dissipative dynamical systems. *Archive for Rational Mechanics and Analysis*, 45:321–393, 1972.

[207] J.L. Willems. *Stability Theory of Dynamical Systems*. Thomas Nelson and Sons Ltd., 1970.

[208] H.S. Witsenhausen. A class of hybrid-state continuous-time dynamic systems. *IEEE Transactions on Automatic Control*, 11(2):161–167, 1966.

[209] K. Yosida. *Functional Analysis*. Springer, 1980.

Samenvatting

Door technologische ontwikkelingen is belangstelling ontstaan voor de analyse en synthese van systemen met zowel een discreet (digitaal) als een continu (analoog) karakter. Deze zogenaamde "hybride systemen" ontstaan o.a. door tijd-continuë processen te koppelen aan tijd-asynchrone digitale regelsystemen. Veel gebruiksartikelen (auto's, magnetrons, wasmachines, enzovoorts) worden aangestuurd door digitale "embedded software," waardoor het gehele proces een systeem is met hybride dynamica. Ook fysische systemen zijn vaak hybride van aard: de beschrijving van mechanische objecten hangt sterk af van het actief zijn van bepaalde contacten, wrijvingsmodellen maken duidelijk verschil tussen "slip" en "stick" fasen en elektrische schakelaars als diodes kunnen zowel geleidend als blokkerend zijn.

De algemeenheid van de bovenstaande voorbeelden maakt duidelijk dat de studie van hybride systemen vanuit een (te) generiek kader weinig specifieke uitspraken zal opleveren over de individuele elementen in de modelklasse. Dientengevolge is het verstandig een deelklasse te bestuderen met een additionele structuur, die analyse en regelaarontwerp wel mogelijk maakt. De keuze dient echter zodanig gemaakt te worden dat vele praktisch interessante systemen binnen de modelstructuur vallen. De klasse van (lineaire) complementariteitssystemen voldoet aan de twee genoemde eigenschappen en vormt dan ook het onderwerp van dit proefschrift. Complementariteitssystemen bestaan uit differentiaalvergelijkingen, ongelijkheden en logische uitdrukkingen en kunnen beschouwd worden als de dynamische generalisaties van het "Linear Complementarity Problem" (LCP) van de mathematische programmering.

Het bestuderen van complementariteitssystemen wordt in hoofdstuk 2 gemotiveerd door een breed spectrum aan mogelijke applicaties: mechanische systemen met ongelijkheidsnevenvoorwaarden, Coulomb wrijving of eenzijdige veren; elektrische netwerken met diodes; regelsystemen met saturatie-verschijnselen en dode zones; stuksgewijs lineaire systemen; "variable structure systems;" relay systemen; hydraulische processen met kleppen, die stromingen slechts in één richting toelaten; verzamelingen van vergelijkingen afkomstig van optimale besturingsproblemen met toestandsbeperkingen; enzovoort. In hoofdstuk 6 wordt tevens aangetoond, dat de klasse van geprojecteerde dynamische systemen in het complementariteitsformalisme past.

De eerste essentiële stappen voor het opzetten van een goed gefundeerde theorie voor een klasse van (discontinuë) dynamische systemen zijn het definiëren van een fysisch relevant oplossingsconcept en de beantwoording van de klassieke vragen betreffende existentie en uniciteit van oplossingen. Het introduceren van een oplossingsconcept voor complementariteitssystemen is een niet-triviale aangelegenheid als gevolg van de sprongverschijnselen en de verschillende werkgebieden (ook wel "configuraties" of "discrete toestanden" genoemd), die elk hun eigen karakteristieke bewegingsvergelijkingen hebben. De definitie van de oplossingstrajecten is dan ook gebaseerd op de combinatie van een distributioneel en een hybrid kader. De praktische

waarde van het voorgestelde oplossingsconcept wordt aangetoond door te laten zien dat de oplossingstrajecten aan algemeen geaccepteerde regels voldoen als gespecificeerd in de literatuur over o.a. mechanische systemen met botsingsverschijnselen en schakelende elektrische netwerken.

Een belangrijke kwestie betreft de existentie en uniciteit van oplossingen gegeven een beginconditie (zgn. goedgesteldheid). Het is verrassend te moeten constateren, dat dit fundamentele probleem nauwelijks aandacht krijgt in de hybride systeemtheorie. In dit proefschrift, zullen dan ook verifieerbare condities voor goedgesteldheid van lineaire complementariteitssystemen afgeleid worden. Goedgesteldheid vormt een eerste verificatie van het gebruikte model en is dus van onafhankelijk belang. Bovendien geven zowel het oplossingsconcept als goedgesteldheid de noodzakelijke inzichten, die belangrijk zijn om vragen over bestuurbaarheid, stabiliteit en regelaarontwerp te beantwoorden.

In hoofdstuk 4 wordt aangetoond dat de existentie en uniciteit van "initiële oplossingen" van een lineair complementariteitssysteem (gegeven een beginconditie) equivalent is aan de existentie en uniciteit van oplossingen van een familie van statische LCP's. Dit verband is gebaseerd op het gebruik van het "Rational Complementarity Problem," een generalisatie van het LCP voor rationale functies. De kracht van deze equivalentie ligt in de uitgebreide literatuur beschikbaar voor LCP's, die nu aangewend kan worden om goedgesteldheidsresultaten voor lineaire complementariteitssystemen te verkrijgen. Dit wordt in het proefschrift geïllustreerd aan de hand van mechanische systemen onder ongelijkheidsnevenvoorwaarden, lineaire relay systemen en lineaire elektrische netwerken met diodes.

Omdat existentie van initiële oplossingen een oneindig aantal re-initialisaties op één tijdstip niet uitsluit, kan locale existentie van oplossingen op een tijdsinterval met positieve lengte niet gegarandeerd worden op grond van de bovengenoemde resultaten. In hoofdstuk 3 worden dan ook voldoende voorwaarden voor locale existentie en uniciteit van oplossingen afgeleid in termen van de positiviteit van de hoofdminoren van de rijgewijze- en kolomsgewijze kopcoëfficientenmatrices. Deze condities zijn gebaseerd op het gebruik van een andere variant van het LCP, het zogenaamde "Linear Dynamic Complementarity Problem." Een interessant tussenresultaat karakteriseert de toestanden waarvoor geen re-initialisatie vereist is voordat een gladde voortzetting kan plaats vinden (zgn. reguliere toestanden). Voor lineaire complementariteitssystemen met één complementariteitspaar en dus twee discrete toestanden (zgn. bimodale systemen), en lineaire complementariteitssystemen met leidende rij indices gelijk aan nul of een zijn deze "initiële" en "locale" resultaten uitgebreid tot "globale" existentie van oplossingen.

In hoofdstuk 5 wordt aandacht besteed aan lineaire complementariteitssystemen waarvoor de onderliggende toestandsmodellen passief zijn, waardoor de systemen corresponderen met lineaire passieve elektrische netwerken met ideale diodes. De passiviteitsconditie maakt het mogelijk specifieke eigenschappen van de oplossingstrajecten te bewijzen. Verder wordt de set van reguliere toestanden expliciet beschreven m.b.v. de duale kegel van de oplossingsruimte van een homogeen LCP. Het testen of

een toestand regulier is kan nu geschieden door het bepalen of een zeker LCP oplosbaar is. Het combineren van al deze resultaten resulteert in globale existentie en uniciteit van oplossingen in een sterkere zin dan voor algemene lineaire complementariteitssystemen.

Naast het formuleren van een oplossingsconcept en de resultaten betreffende goedgesteldheid, wordt in dit proefschrift ook aandacht geschonken aan numerieke simulatie methoden. Dit proefschrift presenteert bijdragen voor o.a. "event-driven" methodieken, die het simulatie-interval beschouwen als de vereniging van disjuncte subintervallen, waarin de discrete toestand (de actieve beperkingen) niet veranderen. Op een subinterval wordt het systeem beschreven door differentiaal- en algebraïsche vergelijkingen, die met standaard integratie routines opgelost kunnen worden (DAE-simulatie). Gedurende de integratie dienen indicatoren, die het einde bepalen van een subinterval, gecontroleerd te worden (event-detectie). Vervolgens, moet een nieuwe discrete toestand (mode-selectie) en de re-initialisatie van de continue toestand bepaald worden. Het voorgestelde oplossingsconcept in dit proefschrift is nauw verbonden met deze event-driven methodiek en het werk op het gebied van goedgesteldheid heeft dan ook directe consequenties voor mode-selectie en re-initialisatie.

Een alternatief voor de event-driven methode is de "time-stepping" techniek, die de systeemvergelijkingen door discrete equivalenten vervangt. Numerieke integratie formules worden gebruikt om de afgeleiden te benaderen en alle algebraïsche condities worden voor elke tijdstap opgelegd. Een regelmatig toegepaste methode is gebaseerd op de "backward Euler" integratie formule, die voor lineaire complementariteitssystemen resulteert in het oplossen van een LCP op elke tijdstap. Een fundamentele basis voor deze "backward Euler time-stepping" methode blijkt noodzakelijk te zijn, omdat er voorbeelden van lineaire complementariteitssystemen bestaan, waarvoor convergentie van de benaderingen niet geldt. Een voorbeeld wordt beschreven in hoofdstuk 7. Ook het feit, dat de tijden, waarop de overgangen tussen de discrete toestanden plaats vinden, niet exact getraceerd worden, maakt het onduidelijk of de methodiek consistent is. Het is dus onverstandig deze methode voor algemene lineaire complementariteitssystemen toe te passen zonder enige verificatie vooraf. In hoofdstuk 7 wordt dan ook voor passieve elektrische netwerken met ideale diodes aangetoond dat de benaderingen naar de echte oplossing van het netwerkmodel convergeren.

Tijdens het verkrijgen van de beschreven resultaten en in het overzicht van mogelijke applicaties in hoofdstuk 2 worden relaties tussen de deelklassen van complementariteitssystemen aangegeven. Enerzijds, biedt het vaststellen van de gemeenschappelijke structuur voor de verscheidene toepassingsgebieden mogelijkheden voor generalisatie of transformatie van resultaten van het ene domein naar het andere. Anderzijds, hebben complementariteitssystemen de potentie om een belangrijke rol te vervullen in het ontwikkelen van systematische technieken, die analyse en synthese mogelijk moeten maken voor een breed scala aan toepassingen. Het werk in dit proefschrift vormt een stap in die richting, daar het verschillende fundamentele vragen beantwoordt, die noodzakelijk zijn voor het opzetten van een algemene systeem- en regeltheorie voor complementariteitssystemen.

Curriculum Vitae

1972 Born in St. Odiliënberg, the Netherlands

1984-1991 Secondary school: Atheneum-B, Bisschoppelijk College Schöndeln, Roermond, the Netherlands.

1991-1995 Master student at the Eindhoven University of Technology, Department of Mathematics and Computing Science, Systems and Control Group. The master project entitled "*Optimal positive control and optimal control with positive state entries*" was carried out under supervision of Prof. Dr. Ir. M.L.J. Hautus, Dr. A.A. Stoorvogel and Dr. Ir. S.J.L. van Eijndhoven.

1995-1999 Ph.D. research at the Eindhoven University of Technology, Department of Electrical Engineering, Measurement and Control Systems and the Tilburg University, Department of Economics. The Ph.D. project entitled "*Linear Complementarity Systems: A Study in Hybrid Dynamics*" was carried out under supervision of Prof. Dr. Ir. P.P.J. van den Bosch, Prof. Dr. J.M. Schumacher and Dr. S. Weiland

1996 Received the certificate of the "Dutch Institute of Systems and Control" for successfully finishing the courses "system identification," "mathematical models of systems," "system and control theory of nonlinear systems," "design methods for control systems" and "control theory of linear systems."

This page is not part of the thesis.