Some Results On Optimal Control for Nonlinear Descriptor Systems

Johan Sjöberg

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
http://www.control.isy.liu.se
johans@isy.liu.se

Linköping 2006
This is a Swedish Licentiate’s Thesis.
Swedish postgraduate education leads to a Doctor’s degree and/or a Licentiate’s degree. A Doctor’s Degree comprises 160 credits (4 years of full-time studies). A Licentiate’s degree comprises 80 credits, of which at least 40 credits constitute a Licentiate’s thesis.

Some Results On Optimal Control for Nonlinear Descriptor Systems

© 2006 Johan Sjöberg

Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping
Sweden

ISBN 91-85497-06-1 ISSN 0280-7971 LiU-TEK-LIC-2006:8

Printed by LiU-Tryck, Linköping, Sweden 2006
To Caroline!
Abstract

In this thesis, optimal feedback control for nonlinear descriptor systems is studied. A descriptor system is a mathematical description that can include both differential and algebraic equations. One of the reasons for the interest in this class of systems is that several modern object-oriented modeling tools yield system descriptions in this form. Here, it is assumed that it is possible to rewrite the descriptor system as a state-space system, at least locally. In theory, this assumption is not very restrictive because index reduction techniques can be used to rewrite rather general descriptor systems to satisfy this assumption.

The Hamilton-Jacobi-Bellman equation can be used to calculate the optimal feedback control for systems in state-space form. For descriptor systems, a similar result exists where a Hamilton-Jacobi-Bellman-like equation is solved. This equation includes an extra term in order to incorporate the algebraic equations. Since the assumptions made here make it possible to rewrite the descriptor system in state-space form, it is investigated how the extra term must be chosen in order to obtain the same solution from the different equations.

A problem when computing the optimal feedback law using the Hamilton-Jacobi-Bellman equation is that it involves solving a nonlinear partial differential equation. Often, this equation cannot be solved explicitly. An easier problem is to compute a locally optimal feedback law. This problem was solved in the 1960's for analytical systems in state-space form and the optimal solution is described using power series. In this thesis, this result is extended to also incorporate descriptor systems and it is applied to a phase-locked loop circuit.

In many situations, it is interesting to know if a certain region is reachable using some control signal. For linear time-invariant state-space systems, this information is given by the controllability gramian. For nonlinear state-space systems, the controllability function is used instead. Three methods for calculating the controllability function for descriptor systems are derived in this thesis. These methods are also applied to some examples in order to illustrate the computational steps.

Furthermore, the observability function is studied. This function reflects the amount of output energy a certain initial state corresponds to. Two methods for calculating the observability function for descriptor systems are derived. To describe one of the methods, a small example consisting of an electrical circuit is studied.
Sammanfattning

I denna avhandling studeras optimal återkopplad styrning av olinjära deskriptorsystem. Ett deskriptorsystem är en matematisk beskrivning som kan innehålla både differentialekvationer och algebraiska ekvationer. En av anledningarna till intresset för denna klass av system är att objekt-orienterade modelleringsverktyg ger systembeskrivningar på denna form. Här kommer det att antas att det, åtminstone lokalt, är möjligt att eliminera de algebraiska ekvationerna och få ett system på tillståndsform. Teoretiskt är detta inte så inskränkande för genom att använda någon indexreduktionsmetod kan ganska generella deskriptorsystem skrivas om så att de uppfyller detta antagande.


Acknowledgments

First of all, I would like to thank my supervisor Professor Torkel Glad for introducing me to the interesting field of descriptor systems and for the skillful guidance during the work on this thesis. I have really enjoyed the cooperation so far and I look forward to the continuation towards the next thesis. I would also like to thank Professor Lennart Ljung for letting me join the Automatic Control group in Linköping and for his excellent management and support when needed. Ulla Salaneck also deserves extra gratitude for all administrative help and support.

I am really grateful to the persons that have proofread various parts of the thesis. These persons are Lic. Daniel Axehill, Dr. Martin Enqvist, Lic. Markus Gerdin, Lic. Gustaf Hendeby, Henrik Tidefelt, David Törnqvist, and Johan Wahlström.

I would like to thank the whole Automatic Control group for the kind and friendly atmosphere. Being part of this group is a real pleasure.

There are some people which mean a lot also in the spare time. First, I would like thank Johan Wahlström, with whom I shared an apartment during my undergraduate studies here in Linköping, and Daniel Axehill with whom I did the Master’s Thesis Project. During these years, we have had many nice discussions and good laughters together. I would also like to especially thank Martin Enqvist, Markus Gerdin, Gustaf Hendeby, Thomas Schön, Henrik Tidefelt, David Törnqvist, and Ragnar Wallin for putting up with all kinds of questions and for being really nice friends. Gustaf Hendeby also deserves extra appreciation for all help regarding LATEX. He is also the guy who have made the nice thesis style, in which this thesis is formatted.

I would also like to sincerely acknowledge all my friends from Kopparberg. It is always a great pleasure to see all of you! Especially, I would like to thank Eva and Mikael Norwald for their kindness and generosity.

This work has been supported by the Swedish Research Council, and ECSEL (The Excellence Center in Computer Science and Systems Engineering in Linköping), which are hereby gratefully acknowledged.

Warm thanks are of course also dedicated to my parents, my brother and his wife for supporting me and always being interested in what I am doing. Finally, I would like to thank Caroline for all the encouragement, support and love you give me. I love you!

Linköping, January 2006

Johan Sjöberg
# Contents

1 **Introduction**  
1.1 Thesis Outline  
1.2 Contributions  

2 **Preliminaries**  
2.1 System Description  
2.2 System Index  
2.2.1 Differential Index  
2.2.2 Strangeness Index  
2.3 Solvability and Consistency  
2.4 Index Reduction  
2.4.1 Consecutive Differentiations  
2.4.2 Consecutive Differentiations of Linear Time-Invariant Descriptor Systems  
2.4.3 Kunkel and Mehrmann’s Method  
2.5 Stability  
2.5.1 Semi-Explicit Index One Systems  
2.5.2 Linear Systems  
2.5.3 Barrier Function Method  
2.6 Optimal Control  
2.6.1 Formulation and Summary of the Optimal Control Problem  
2.6.2 Necessary Condition Using the HJB  
2.6.3 Sufficient Condition Using HJB  
2.6.4 Example  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Thesis Outline</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Contributions</td>
<td>4</td>
</tr>
<tr>
<td>2.1 System Description</td>
<td>7</td>
</tr>
<tr>
<td>2.2 System Index</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 Differential Index</td>
<td>10</td>
</tr>
<tr>
<td>2.2.2 Strangeness Index</td>
<td>13</td>
</tr>
<tr>
<td>2.3 Solvability and Consistency</td>
<td>13</td>
</tr>
<tr>
<td>2.4 Index Reduction</td>
<td>21</td>
</tr>
<tr>
<td>2.4.1 Consecutive Differentiations</td>
<td>21</td>
</tr>
<tr>
<td>2.4.2 Consecutive Differentiations of Linear Time-Invariant Descriptor Systems</td>
<td>22</td>
</tr>
<tr>
<td>2.4.3 Kunkel and Mehrmann’s Method</td>
<td>23</td>
</tr>
<tr>
<td>2.5 Stability</td>
<td>26</td>
</tr>
<tr>
<td>2.5.1 Semi-Explicit Index One Systems</td>
<td>27</td>
</tr>
<tr>
<td>2.5.2 Linear Systems</td>
<td>31</td>
</tr>
<tr>
<td>2.5.3 Barrier Function Method</td>
<td>32</td>
</tr>
<tr>
<td>2.6 Optimal Control</td>
<td>33</td>
</tr>
<tr>
<td>2.6.1 Formulation and Summary of the Optimal Control Problem</td>
<td>33</td>
</tr>
<tr>
<td>2.6.2 Necessary Condition Using the HJB</td>
<td>34</td>
</tr>
<tr>
<td>2.6.3 Sufficient Condition Using HJB</td>
<td>36</td>
</tr>
<tr>
<td>2.6.4 Example</td>
<td>38</td>
</tr>
</tbody>
</table>
3 Optimal Feedback Control for Descriptor Systems

3.1 Optimal Feedback Control

3.2 Hamilton-Jacobi-Bellman Equation for the Reduced Problem

3.3 Hamilton-Jacobi-Bellman-Like Equation

3.4 Relationships Among the Solutions

3.5 Control-Affine Systems

3.6 Example

4 Power Series Solution of the Hamilton-Jacobi-Bellman Equation

4.1 Problem Formulation

4.2 Review of the State-Space System Case

4.3 Descriptor System Case

4.3.1 Power Series Expansion of the Reduced Problem

4.3.2 Application of the Results for State-Space Systems

4.3.3 Conditions on the Original Data

4.4 Extension

4.5 Example

5 Controllability Function

5.1 Problem Formulation

5.2 Method Based on HJB Theory

5.3 Method Based on Completion of Squares

5.4 Method to Find a Local Solution

5.4.1 Basic Assumptions and Formulations

5.4.2 Application of the Local Method

5.5 Linear Descriptor Systems

5.6 Examples

5.6.1 Rolling Disc

5.6.2 Artificial System

5.6.3 Electrical Circuit

6 Observability Function

6.1 Problem Formulation

6.2 Method Based on Partial Differential Equation

6.3 Method to Find a Local Solution

6.3.1 Basic Assumptions and Calculation of the Power Series Expansion of the Reduced System

6.3.2 Derivation of Power Series Method

6.4 Example

7 Concluding Remarks

7.1 Conclusions

7.2 Future work

A Some Facts from Calculus and Set Theory

Bibliography
## Symbols and Mathematical Notation

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{R}^n$</td>
<td>the n-dimensional space of real numbers</td>
</tr>
<tr>
<td>$\mathbb{C}^n$</td>
<td>the n-dimensional space of complex numbers</td>
</tr>
<tr>
<td>$\in$</td>
<td>belongs to</td>
</tr>
<tr>
<td>$\forall$</td>
<td>for all</td>
</tr>
<tr>
<td>$A \subset B$</td>
<td>$A$ is a subset of $B$</td>
</tr>
<tr>
<td>$A \cap B$</td>
<td>the intersection between $A$ and $B$</td>
</tr>
<tr>
<td>$A \cup B$</td>
<td>the union of $A$ and $B$</td>
</tr>
<tr>
<td>$\partial A$</td>
<td>the boundary of the set $A$</td>
</tr>
<tr>
<td>$I_n$</td>
<td>the identity matrix of dimension $n \times n$</td>
</tr>
<tr>
<td>$f : \mathbb{D} \to \mathbb{Q}$</td>
<td>the function $f$ maps a set $\mathbb{D}$ to a set $\mathbb{Q}$</td>
</tr>
<tr>
<td>$f \in C^k(\mathbb{D}, \mathbb{Q})$</td>
<td>a function $f : \mathbb{D} \to \mathbb{Q}$ is $k$-times continuously differentiable</td>
</tr>
</tbody>
</table>

- $f_{r; x}$ \(\text{the partial derivative of } f_r \text{ with respect to } x\)
- $Q \succ (\succeq) 0$ \(\text{the matrix } Q \text{ is positive (semi)definite}\)
- $Q \prec (\preceq) 0$ \(\text{the matrix } Q \text{ is negative (semi)definite}\)
- $\sigma(E, A)$ \(\text{the set } \{s \in \mathbb{C} \mid \det(sE - A) = 0\}\)
- $\lambda_i(A)$ \(\text{the } i\text{th eigenvalue of the matrix } A\)
- $\Re s$ \(\text{the real part of } s\)
- $\Im s$ \(\text{the imaginary part of } s\)
- $\mathbb{C}^+$ \(\text{the closed right half complex plane}\)
- $\mathbb{C}^-$ \(\text{the open left half complex plane}\)
- $\|x\|$ \(\sqrt{x^T x}\)
- $\min_x f(x)$ \(\text{minimization of } f(x) \text{ with respect to } x\)
- $\arg\min_x f(x)$ \(\text{the } x \text{ minimizing } f(x)\)
### Notation

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_r$</td>
<td>the ball of radius $r$ (see Appendix A)</td>
</tr>
<tr>
<td>$\lfloor x \rfloor$</td>
<td>the floor function, which gives the largest integer less than or equal to $x$</td>
</tr>
<tr>
<td>$\dot{x}$</td>
<td>time derivative of $x$</td>
</tr>
<tr>
<td>$x^{(i)}(t)$</td>
<td>the $i$th derivative of $x(t)$ with respect to $t$</td>
</tr>
<tr>
<td>$f^{[i]}(x)$</td>
<td>all terms in a multivariable polynomial of order $i$</td>
</tr>
<tr>
<td>$o(h)$</td>
<td>$f(h) = o(h)$ as $h \to 0$ if $f(h)/h \to 0$ as $h \to 0$</td>
</tr>
<tr>
<td>corank $A$</td>
<td>the rank deficiency of the matrix $A$ with respect to rows (see Appendix A)</td>
</tr>
</tbody>
</table>

### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARE</td>
<td>Algebraic Riccati Equation</td>
</tr>
<tr>
<td>DAE</td>
<td>Differential-Algebraic Equation</td>
</tr>
<tr>
<td>DP</td>
<td>Dynamic Programming</td>
</tr>
<tr>
<td>HJB</td>
<td>Hamilton-Jacobi-Bellman (equation)</td>
</tr>
<tr>
<td>HJI</td>
<td>Hamilton-Jacobi Inequality</td>
</tr>
<tr>
<td>ODE</td>
<td>Ordinary Differential Equation</td>
</tr>
<tr>
<td>PLL</td>
<td>Phase-Locked Loop circuit</td>
</tr>
<tr>
<td>PMP</td>
<td>Pontryagin Minimum Principle</td>
</tr>
</tbody>
</table>

### Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Short explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>The algebraic equations are possible to solve for the algebraic variables, i.e., an implicit function exists (see page 19)</td>
</tr>
<tr>
<td>A2</td>
<td>The set, on which the implicit function is defined, is global in the control input (see page 20)</td>
</tr>
<tr>
<td>A3</td>
<td>The index reduced system can be expressed in semi-explicit form (see page 25)</td>
</tr>
<tr>
<td>A4</td>
<td>The algebraic equations are locally possible to solve for the algebraic variables (see page 51)</td>
</tr>
<tr>
<td>A5</td>
<td>The functions $F_1$ and $F_2$ for a semi-explicit descriptor system are analytical (see page 51)</td>
</tr>
<tr>
<td>A6</td>
<td>Only feedback laws locally stabilizing a descriptor system going backwards in time is considered (see page 73).</td>
</tr>
<tr>
<td>A7</td>
<td>The functions $F_1$, $F_2$ and $h$ for a semi-explicit descriptor system with an explicit output equation are analytical (see page 90)</td>
</tr>
</tbody>
</table>
Introduction

In real life, control strategies are used almost everywhere. Often these control strategies form some kind of feedback control. This means that, based on observations, action is taken in order to obtain a certain goal. Which action to choose, given the actual observations, is decided by the so-called controller. The controller can for example be a person, a computer or a mechanical device. As an example, we can take one of the most well-known controllers, namely the thermostat. The thermostat is used to control the temperature in a room. Therefore, it measures the temperature in the room and if it is too high, the thermostat decreases the amount of hot water passing through the radiator, while if it is too low, the amount is increased instead. In this way, the temperature of the room is kept at a desired level.

This very simple control strategy can in some cases be enough, but in many situations better performance is desired. To achieve better performance it is most often necessary to take the controlled system into consideration. This can of course be done in different ways, but in this thesis it will be assumed that we have a mathematical description of the system. The mathematical description is called a model of the system and the same system can be described by models in different forms.

One such form is the descriptor system form. The advantage with this form is that it allows for both differential and algebraic equations. This fact makes it possible to model systems in a very natural way in certain cases. One such case where the descriptor form is natural to work with is when using object-oriented modeling methods. The basic idea in object-oriented modeling is to create the complete model of a system as the composition of many small models. To concretize we can use the modeling of a car as an example.

The first step is to model the engine, the gearbox, the propeller shaft, the car body etc. as separate models. The second step is to connect all the separate models to get the model of the complete car. Typically, these connections will introduce algebraic equations describing for example that the output shaft from the engine must to rotate with the same angular velocity as the input shaft of the gearbox.
Some examples where models in descriptor system form have been derived are for example chemical processes (Kumar and Daoutidis, 1999), electrical circuits (Tischendorf, 2003), multibody mechanics in general (Hahn, 2002, 2003), multibody mechanics applied to a truck (Simeon et al., 1994; Rheinboldt and Simeon, 1999) and to robotics (McClamrock, 1990).

Supplied with a model in descriptor system form we consider controller design. More specifically optimal feedback control will be studied. Optimal feedback control means that the controller is designed to minimize a performance criterion. Therefore, the performance criterion should reflect the desired behavior of the controlled system. For example, for an engine management system, the performance criterion could be a combination of the fuel consumption and the difference between the actual torque delivered by the engine and the torque wanted by the driver. The design procedure would then yield the controller achieving the best balance between low fuel consumption and delivery of the requested torque.

1.1 Thesis Outline

The thesis is separated into seven main chapters. Since the main subject of this work is descriptor systems and optimal feedback control of such systems, Chapter 2 introduces these subjects.

Chapter 3 is the first chapter devoted to optimal feedback control of descriptor systems. Two different methods are investigated and some relationships between their optimal solutions are revealed. Chapter 4 deals with the same problem as Chapter 3, but in this chapter the optimal feedback control problem is solved using series expansions. The method is very general but the solution might be restricted to a neighborhood.

The computation of the controllability function is considered in Chapter 5. The controllability function is defined as the solution to an optimal feedback control problem and therefore the methods in Chapter 3 and Chapter 4 are used to solve this problem. Chapter 6 treats the computation of the observability function. The observability function is not defined as the solution to an optimal feedback control problem, but it is still possible to use ideas similar to those presented in Chapter 5.

Chapter 7 summarizes the thesis with some conclusions and remarks about interesting problems for future research.

1.2 Contributions

The main contributions in this thesis are in the field of descriptor systems. This means that when nothing else is mentioned, some control method for state-space systems is extended to handle also descriptor systems.

A list of contributions, and the publications where these are presented, is given below.

- The analysis of the relationship among the solutions of two different methods for solving the optimal feedback control problem, which can be found in Chapter 3. The presentation there is a modified version of the technical report:
1.2 Contributions


- The method in Chapter 4 for finding a power series solution to the optimal feedback control problem. The material presented in this chapter comes from the conference paper:


  The material has then been extended as described in Section 4.4.

- The different methods to compute the controllability function, given in Chapter 5. This chapter is based on results from:


- The different methods in Chapter 6 to compute the observability function.

  In addition to the contributions mentioned above, a small survey over theory for linear descriptor systems has been published as a technical report:

To introduce the subject to the reader, we will in this chapter present some basic descriptor system theory. Four key concepts, index, solvability, consistency, and stability will be briefly described. Furthermore, it will be discussed how the index of a system description can be lowered using some kind of index reduction method. Finally, an introduction to optimal control of state-space systems will be given.

2.1 System Description

The original mathematical description of a system often consists of a set of differential and algebraic equations. However, in most literature on control theory it is assumed that the algebraic equations can be used to eliminate some variables. The result is a system description consisting only of differential equations that can be written in state-space form as

\[ \dot{x} = F(t, x, u) \]  

(2.1)

where \( x \in \mathbb{R}^n \) is the state vector and \( u \in \mathbb{R}^p \) is the control input. The state variables represent the system’s memory of its past and throughout this thesis, a variable will only be denoted state if it has this property.

The state-space form has some drawbacks. For example, some systems are easy to model if both differential and algebraic equations may be used, while a reduction to a state-space model is more difficult. Another possible drawback occurs when the structure of the system description is nice and intuitive while using both kinds of equations, but the state-space formulation loses some of these features. A third drawback is related to object-oriented computer modeling tools, such as Dymola. Usually, these tools do not yield system descriptions in state-space form, but as a set of both algebraic and differential equations. In practice, the number of equations is often large and reduction to a state-space model is then almost impossible.
Therefore, the focus of this thesis is at a more general class of model descriptions, called descriptor systems or differential-algebraic equations (DAE). This class of system descriptions includes both differential and algebraic equations and mathematically this kind of system descriptions can be formulated as

\[ F(t, x, \dot{x}, u) = 0 \]  \hspace{1cm} (2.2)

where \( x \in \mathbb{R}^n, u \in \mathbb{R}^p \) and \( F : \mathbb{D} \to \mathbb{R}^m \) for some set \( \mathbb{D} \subset \mathbb{R}^{2n+p+1} \). As for the state-space model, \( u \) is here the control input. However, not all of the variables in \( x \) need to be states. The reason is that some components of \( x \) do not represent a memory of the past, i.e., are not described by differential equations. This is shown in the following small example.

---

**Example 2.1**

Consider the system description

\[
\begin{align*}
\dot{x}_1 + x_2 + 2u &= 0 \\
       x_2 - x_1 + u &= 0 
\end{align*}
\]

By grouping the variables according to \( x = (x_1, x_2)^T \), this system description fits into (2.2). By solving the lower equation for \( x_2 \) we can obtain the description

\[
\begin{align*}
\dot{x}_1 &= -x_1 - u \\
x_2 &= x_1 - u 
\end{align*}
\]

The variable \( x_1 \) is determined by an ordinary differential equation, while \( x_2 \) is algebraically connected to \( x_1 \) and \( u \). Hence, the memory of the past is \( x_1 \), while \( x_2 \) is just a snapshot of the other variables. Therefore, the only state in this example is \( x_1 \). The fact that only parts of \( x \) are states is an important property of DAEs.

In this thesis, the function \( F \) in (2.2) will often be differentiated a number of times with respect to \( t \), in order to obtain a mathematically more tractable form. Therefore, an assumption made throughout the thesis is the function \( F \) is sufficiently smooth, i.e., \( F \) must be sufficiently many times continuously differentiable to allow for these differentiations.

In some cases, (2.2) will be viewed as an autonomous system

\[ F(t, x, \dot{x}) = 0 \]  \hspace{1cm} (2.3)

In this thesis, the two most common reasons are either that the control input is given as a feedback law \( u = u(t, x) \), or that \( u = u(t) \) is a given time signal and is seen as part of the time variability. However, a third reason is that the system is modeled using a behavioral approach, see (Polderman and Willems, 1998; Kunkel and Mehrmann, 2001). In this case, the control input \( u \) is viewed as just another variable, and it is included in the variables \( x \). The system of equations is then often underdetermined and some variables have to be chosen as inputs so that the remaining ones are uniquely defined. In engineering applications the choice of control variables is often obvious from the physical plant.
Remark 2.1. Consider a system description (2.2) with \( m = n \), i.e., with the same number of equations as variables \( x \). If the control input is a given signal, the system will have as many unknowns in \( x \) as there are equations. However, for a behavioral model where \( x \) and \( u \) are clustered and \( u \) is considered as just another variable, the system will be underdetermined.

Often when modeling physical processes, the obtained system description will get more structure than the general description (2.2). One such structure is the semi-explicit form. For example, this form naturally arises when modeling mechanical multibody systems (Arnold et al., 2004), and it can be expressed as

\[
E \dot{x} = F(x, u) \tag{2.4}
\]

where \( E \in \mathbb{R}^{n \times n} \) is a possibly rank deficient matrix, i.e., \( \text{rank} E = r \leq n \). Linear time-invariant descriptor systems can always be written in this form as

\[
E \dot{x} = Ax + Bu \tag{2.5}
\]

The description (2.4) (and hence (2.5)) can without loss of generality be written in semi-explicit form

\[
\dot{x}_1 = F_1(x_1, x_2, u) \tag{2.6a}
\]
\[
0 = F_2(x_1, x_2, u) \tag{2.6b}
\]

where \( x_1 \in \mathbb{R}^r \) and \( x_2 \in \mathbb{R}^{n-r} \). It may seem like all \( x_1 \) are states, i.e., hold information about the past. However, as will be shown in the next section this does not need to be true, unless \( F_{2; x_2}(x_1, x_2, u) \) is at least locally nonsingular.

In some cases it might be interesting to extend the system descriptions above with an equation for an output signal \( y \) as

\[
F(t, x, \dot{x}, u) = 0 \tag{2.7a}
\]
\[
y = h(x, u) \tag{2.7b}
\]

where \( y \in \mathbb{R}^q \). In general, an explicit extension of the system with an extra output equation is unnecessary for descriptor systems. Instead, the output equation can be included in \( F(\dot{x}, x, u, t) \) and the output signal \( y = y(t) \) is then seen as part of the time variability. However, in some situations it is important to show which variables that are possible to measure and in this case, (2.7) is the best system description.

### 2.2 System Index

The index is a commonly used concept in the theory of descriptor systems. Many different kinds of indeces exist, for example differential index, perturbation index, strangeness index. The common property of the different indices is that they in some sense measure how different a given descriptor system is from a state-space system. Therefore, a system description with high index will often be more difficult to handle than a description with a lower index. The index is mostly a model property and two different models in the form (2.2), modeling the same physical plant, can have different indices.
In the sequel of this section, two kinds of indices, namely the differential index and the strangeness index, will be discussed further. More information about different kinds of indices can be found in (Campbell and Gear, 1995; Kunkel and Mehrmann, 2001) and the references therein.

### 2.2.1 Differential Index

The differential index is the most common of the different index concepts. It will also be this kind of index that in this thesis is denoted only the index. Loosely speaking, the differential index is the minimum number of differentiations needed to obtain an equivalent system of ordinary differential equations, i.e., a state-space system. A small example showing the idea can be found below.

**Example 2.2**

Consider a system given in semi-explicit form

\[
\begin{align*}
\dot{x}_1 &= F_1(x_1, x_2, u) \\
0 &= F_2(x_1, x_2, u)
\end{align*}
\]

where \(x_1 \in \mathbb{R}^{n_1}, x_2 \in \mathbb{R}^{n_2}\) and \(u \in \mathbb{R}^p\). Assume that \(u = u(t)\) is given. Then differentiation of the constraint equation with respect to \(t\) yields

\[
0 = F_{2;\dot{x}_1}(x_1, x_2, u)\dot{x}_1 + F_{2;\dot{x}_2}(x_1, x_2, u)\dot{x}_2 + F_{2;u}(x_1, x_2, u)\dot{u}
\]

If \(F_{2;\dot{x}_2}(x_1, x_2, u)\) is nonsingular it is possible to rewrite the system above as

\[
\begin{align*}
\dot{x}_1 &= F_1(x_1, x_2, u) \\
\dot{x}_2 &= -F_{2;\dot{x}_2}(x_1, x_2, u)^{-1}\left(F_{2;\dot{x}_1}(x_1, x_2, u)F_1(x_1, x_2, u) + F_{2;u}(x_1, x_2, u)\dot{u}\right)
\end{align*}
\]  
(2.8a, 2.8b)

and since \(\dot{x}\) is determined as functions of \(x, u\) and \(\dot{u}\), the original system description is index one. If \(F_{2;\dot{x}_2}(x_1, x_2, u)\) is singular, suppose that with an algebraic manipulation it is possible to get the system description to the semi-explicit form (2.6) again but with another \(x_1\) and \(x_2\). If it is possible to solve for \(\dot{x}_2\) after a second differentiation of the constraint equation the original model is said to be index two. If this is not possible, the procedure is repeated and the number of differentiations will then be the index.

The example above motivates the following definition of the index, see Brenan et al. (1996).

**Definition 2.1.** The differential index is the number of times that all or part of (2.2) must be differentiated with respect to \(t\) in order to determine \(\dot{x}\) as a continuous function of \(x, u, \dot{u}\) and higher derivatives of \(u\).

Note that in the definition above, all rows in the system description need not be differentiated the same number of times.

The method described in Example 2.2 to compute the index is rather intuitive. However, according to Brenan et al. (1996), this method cannot be used for all solvable descriptor systems. The problem is the coordinate transformation needed to obtain the semi-explicit form after each iteration. However, for linear descriptor systems the coordinate
change is done using Gauss elimination. In this case, the method is called the Shuffle algorithm and was introduced by Luenberger (1978).

A more general definition of the index, without the state transformation, can be formulated using the derivative array. Assume the system to be given by (2.2). The derivative array is given by

$$
F_d^j(t, x, x_{j+1}, u, \dot{u}, \ldots, u^\nu) = \begin{pmatrix}
F(t, x, \dot{x}, u) \\
\frac{d}{dt} F(t, x, \dot{x}, u) \\
\vdots \\
\frac{d^\nu}{dt^\nu} F(t, x, \dot{x}, u)
\end{pmatrix}
$$

(2.9)

where

$$
x_j = (\dot{x}, \ddot{x}, \ldots, x^{(j)})
$$

Using the derivative array, the definition of the index may be formulated as follows (Brenan et al., 1996).

**Definition 2.2.** The index \( \nu \) is the smallest positive integer such that \( F_d^\nu \) uniquely determines the variable \( \dot{x} \) as a continuous function of \( x \), \( t \), \( u \) and higher derivatives of \( u \), i.e.,

$$
\dot{x} = \eta(t, x, u, \dot{u}, \ldots, u^\nu)
$$

(2.10)

Note that \( u \) is here considered to be a given time signal, which in principle can be included in the time variability. If \( u \) cannot be seen as a given time signal, the differential index is undefined, and it is necessary to use the concept strangeness-index, see the discussion in Section 2.2.2.

Definition 2.2 might be difficult to use directly, but the following proposition yields sufficient conditions to compute \( \nu \) (Brenan et al., 1996).

**Proposition 2.1**

Sufficient conditions for (2.9) to uniquely determine \( \dot{x} \) as a continuous function of \( x \) and \( t \) are that the Jacobian matrix of \( F_d^\nu(t, x, x_{\nu+1}, u, \dot{u}, \ldots, u^{(\nu)}) \) with respect to \( x_{\nu+1} \) is 1-full with constant rank and that there exists a point

$$
z_{\nu}^0 = (t_0, x_0, \dot{x}_0, \ddot{x}_0, \ldots, x_{0\nu}^{(\nu+1)}, u_0, \dot{u}_0, \ldots, u_{0\nu}^\nu)
$$

such that \( F_d^\nu(z_{\nu}^0) = 0 \) is satisfied.

The concept 1-full means that, by using pre-multiplication with a nonsingular time-dependent matrix \( P(t) \), it is possible to write the Jacobian matrix as

$$
P(t) \frac{\partial F_d^\nu}{\partial x_{\nu+1}} = \begin{pmatrix}
I_n & 0 \\
0 & H(t)
\end{pmatrix}
$$

That is, it must be possible to diagonalize \( F_d^\nu(x_{\nu+1}) \) and obtain an identity matrix in the upper left corner by using time-dependent row operations. Locally on some set, it is then possible to express \( \dot{x} \) as described in (2.10), i.e., without higher derivatives of \( x \).
Remark 2.2. According to Brenan et al. (1996), nonlinear coordinate transformations using pre-multiplication by a nonsingular $P(t)$ do not change the properties of constant rank or 1-fullness of the Jacobian matrix.

As was mentioned at the beginning of Section 2.2, the index is an important measure of how difficult a descriptor system is to handle. Both numerical computation of the solution, see Brenan et al. (1996), and derivation of control methods become more difficult for system descriptions of high index. It turns out that system descriptions with index zero or one are much easier to handle than descriptions with index two or higher. Index zero models are ordinary differential equations (ODEs) either in explicit or implicit form. As was seen in Example 2.2, index one descriptions need one differentiation to be transformed to a state-space model. However, all constraints are explicit for an index one model. This means that all constraints imposed on the solution are given by $F$ itself. In general, this is not the case for higher index models for which implicit constraints also may occur. Implicit constraints are constraints not visible in (2.2), but appearing since the equations must hold on a time interval, denoted $I$ in the sequel. Together, the explicit and implicit constraints define the manifold which the solution $x(t)$ belongs to. An example showing the appearance of implicit constraints is given below.

---

**Example 2.3**

Consider a nonlinear semi-explicit descriptor system of index two. The system is given by

\begin{align}
\dot{x}_1 &= f(x_1, x_2) \quad (2.11a) \\
0 &= g(x_1) \quad (2.11b)
\end{align}

where $x_1$ and $x_2$ are scalars and $g_{x_1}(x_1)f_{x_2}(x_1, x_2)$ is nonsingular. At first sight it might look like $g(x_1) = 0$ is the only constraint. However, differentiating (2.11b) with respect to $t$ gives

\[0 = g_{x_1}(x_1)f(x_1, x_2) \quad (2.12)\]

and one further differentiation with respect to $t$ yields

\[\dot{x}_2 = -\left(g_{x_1}f_{x_2}\right)^{-1}\left(g_{x_1}x_1f^2 + g_{x_1}f_{x_1}f\right)\]

where the arguments have been left out for notational clarity. Hence, the solution $x(t) = (x_1(t), x_2(t))^T$ must not only satisfy the explicit constraint (2.11b) but also the implicit constraint (2.12).

System models of index higher than one will be denoted higher index models. A typical case where high index models occur is when mechanical systems are modeled since mechanical multibody systems often have index three (Arnold et al., 2004). It is important to note that for time-varying linear and nonlinear descriptor systems, the index can vary in time and space. In particular, different feedback laws may yield different indices of the model. This fact has been used for feedback control of descriptor systems to reduce the index of the closed loop system.
In some cases, the concept of differential index plays an important role also for state-space systems. One such case is the inversion problem where the objective is to find $u$ in terms of $y$ and possibly $x$ for a system
\begin{align*}
\dot{x} &= f(x, u) \\
y &= h(x, u)
\end{align*}
(2.13)
where it is assumed that the number of inputs and outputs are the same. The procedures for inversion typically includes some differentiations, sometimes using Lie-bracket notation, until $u$ can be recovered. The number of differentiations needed, is normally called the relative degree or order. However, for a given output signal $y$, the system (2.13) is a descriptor system in $(x, u)$. The corresponding index of this descriptor system is the relative degree plus one.

### 2.2.2 Strangeness Index

Another index concept is the strangeness index $\mu$, which for example is described in Kunkel and Mehrmann (2001). The definition of the strangeness index will be presented in the next section when solvability of descriptor systems is considered.

The strangeness index is a generalization of the differential index in the sense that some rank conditions are relaxed. Furthermore, unlike the differential index the strangeness index is defined for over- and underdetermined system descriptions. However, for system descriptions where both the strangeness index and the differential index are well-defined the relation is, in principle, $\mu = \max\{0, \nu - 1\}$. For a more thorough discussion about this relationship the reader is referred to Kunkel and Mehrmann (1996). A system with $\mu = 0$ is denoted strangeness-free.

### 2.3 Solvability and Consistency

Intuitively, solvability means that the descriptor system (2.2) possesses a well-behaved solution. Well-behaved in this context means unique and sufficiently smooth, for example continuously differentiable. For state-space systems, solvability follows if the system satisfies a Lipschitz condition, see Khalil (2002). For descriptor systems, the solvability problem is somewhat more intrinsic.

In Section 2.2, it was shown that the solution of a descriptor system is, in principle, defined by the derivative array. If the index is finite, the derivative array can be solved for $\dot{x}$ and a solution can be computed by integration. However, for some systems the index does not exist. Another complicating fact is the possibility of choosing initial condition $x(t_0)$ such that the constraints are not satisfied. A third problem occurs when the control signal is not smooth enough and the solution contains derivatives of the control input.

The solvability definitions and theorems in this section are based on the results in Kunkel and Mehrmann (1994, 1998, 2001). Other results on solvability for descriptor systems can be found in (Brenan et al., 1996; Campbell and Gear, 1995; Campbell and Griepentrog, 1995). The method presented by Kunkel and Mehrmann (2001) also handles under- or overdetermined systems etc. An overdetermined system is a system where the number of equations $m$ is larger than the number of unknowns, while the opposite
holds for an underdetermined system. Normally, the unknowns are $x$, but if a behavioral approach is considered also $u$ can be seen as unknown.

First, the definition of a solution for an autonomous systems (2.3) from Kunkel and Mehrmann (2001) will be presented. This definition also includes the case when the system has a control input either using a behavioral approach or by seeing $u = u(t)$ as given and therefore part of the time variability.

**Definition 2.3.** Consider the system (2.3) and denote the time interval for which (2.3) is defined as $\mathbb{I} \subset \mathbb{R}$.

A function $x(t)$ is called a solution to (2.3) if $x(t) \in C^1(\mathbb{I})$ and $x(t)$ satisfies (2.3) pointwise. The function is called a solution of the initial value problem consisting of (2.3) and

$$x(t_0) = x_0$$

if $x(t)$ is a solution of (2.3) and satisfies (2.14).

We also define what is meant with a consistent initial condition.

**Definition 2.4.** An initial condition $(t_0, x_0)$ is called consistent if the corresponding initial value problem has at least one solution.

Note that a necessary condition for a point $(t_0, x_0)$ to be a consistent initial condition for (2.3) is algebraic allowance. That is, it must be possible to choose $q$ such that $F^d_{\nu}(t_0, x_0, q) = 0$, or with other words, the point must satisfy both the explicit but also the implicit constraints. Note that in this case, $F^d_{\nu}$ is just a function of $t, x$ and higher derivatives of $x$, since $u$ is included in $x$. The problem of finding such points have been studied in for example (Pantelides, 1988; Campbell et al., 1996). However, notice that in these references, algebraically allowed initial conditions are called consistent.

To derive conditions under which a solution to (2.3) exists and is unique according to the definitions above, Kunkel and Mehrmann (2001) use a hypothesis. The hypothesis is investigated on the solution set of the derivative array (2.9) for some integer $\mu$. The solution set is denoted $\mathbb{L}_\mu$ and is described by

$$\mathbb{L}_\mu = \left\{ z_\mu \in \mathbb{I} \times \mathbb{R}^n \times \ldots \times \mathbb{R}^n \mid F^d_{\mu}(z_\mu) = 0 \right\}$$

while the hypothesis is as follows.

**Hypothesis 2.1.** Consider the general nonlinear descriptor system (2.3). There exist integers $\mu, r, a, d$ and $v$ such that $\mathbb{L}_\mu$ is not empty, and the following properties hold:

1. The set $\mathbb{L}_\mu \subset \mathbb{R}^{(\mu+2)n+1}$ forms a manifold of dimension $(\mu + 2)n + 1 - r$.

2. It holds that

$$\text{rank } F^d_{\mu;x,x_{\mu+1}} = r$$

on $\mathbb{L}_\mu$ where $x_{\mu+1} = (\dot{x}, \ddot{x}, \ldots, x^{(\mu+1)})$.

3. It holds that

$$\text{corank } F^d_{\mu;x,x_{\mu+1}} - \text{corank } F^d_{\mu-1;x,x_{\mu}} = v$$

on $\mathbb{L}_\mu$. Here the convention that $\text{corank } F^d_{-1;x} = 0$ is used. (For a definition of the corank, see Appendix A)
4. It holds that
\[ \text{rank } F^d_{\mu;x_{\mu+1}} = r - a \] (2.18)
on \mathbb{L}_\mu \text{ such that there are smooth full rank matrix functions } Z_2 \text{ and } T_2 \text{ defined on } \mathbb{L}_\mu \text{ of size } ((\mu + 1)m, a) \text{ and } (n, n - a), \text{ respectively, satisfying}
\[ Z_2^T F^d_{\mu;x_{\mu+1}} = 0, \quad \text{rank } Z_2^T F^d_{\mu;x} = a, \quad Z_2^T F^d_{\mu;x} T_2 = 0 \] (2.19)
on \mathbb{L}_\mu.

5. It holds that
\[ \text{rank } F^d_d = d = m - a - v \] (2.20)
on \mathbb{L}_\mu.

Note that the different ranks appearing in the hypothesis are assumed to be constant on the manifold \( \mathbb{L}_\mu \).

If there exist \( \mu, d, a \) and \( v \) such that the hypothesis above holds, it will imply that the system can be reduced to a system consisting of an implicit ODE and some algebraic equations. The implicit ODE forms \( d \) differential equations, while the number of algebraic equations are \( a \). The motivation and procedure are described below. If the hypothesis is not satisfied for a given \( \mu \), i.e., if \( d \neq m - a - v \), \( \mu \) is increased by one and the procedure is repeated. However, it is not certain that a \( \mu \) exists such that the hypothesis hold.

The quantity \( v \) needs to be described. It measures the number of equations in the original system (2.3) resulting in trivial equations \( 0 = 0 \), i.e., \( v \) measures the number of redundant equations. Together with the numbers \( a \) and \( d \), all \( m \) equations in the original system are then characterized, since \( m = a + d + v \).

The earlier mentioned strangeness index is also defined using Hypothesis 2.1. The definition is as follows.

**Definition 2.5.** The strangeness index of (2.3) is the smallest positive integer \( \mu \) such that Hypothesis 2.1 is satisfied.

The analysis of what the hypothesis implies is local and is done close to the point \( z_\mu^0 = (t_0, x_0, x_{\mu+1,0}) \in \mathbb{L}_\mu \) where \( x_{\mu+1,0} = (\dot{x}_0, \ldots, x_0^{(\mu+1)}) \). The variables \( x_0^{(j)} \) where \( j \geq 1 \) are in this case seen as algebraic variables rather than as derivatives of \( x_0 \). From part 1 of the hypothesis it is known that \( \mathbb{L}_\mu \) is a \((\mu + 2)n + 1 - r\) dimensional manifold. Thus it is possible to locally parameterize it using \((\mu + 2)n + 1 - r\) parameters. These parameters can be chosen from \((t, x, x_{\mu+1})\) such that the rank of (2.21) is unchanged if the corresponding columns in
\[ F^d_{\mu;x,x_{\mu+1}}(t_0, x_0, x_{\mu+1,0}) \] (2.21)
are removed. Together, parts 1 and 2 of the hypothesis give that
\[ \text{rank } F^d_{\mu;x,x_{\mu+1}} = \text{rank } F^d_{\mu;x,x_{\mu+1}} = r \]
and hence \( t \) can be chosen as parameter. From part 2 it is also known that \( r \) variables of \((x, x_{\mu+1})\) are determined (via the implicit function theorem) by the other \((\mu + 2)n + 1 - r\) variables. From part 4, we have that \( r - a \) variables of \( x_{\mu+1} \) are determined. We
denote these variables \( x_h \), while the rest of \( x_{\mu+1} \) must be parameters and are denoted \( p \in \mathbb{R}^{(\mu+1)n+a-r} \).

Since \( r \) variables are implicitly determined by the rest and only \( a \) of these belong to \( x_{\mu+1} \), the other \( r - (r - a) = a \) determined variables must belong to \( x \). We denote these variables \( x_2 \in \mathbb{R}^a \) and using part 4 it follows that \( Z_2^T F_{\mu;x_2} \) must be nonsingular. The rest of \( x \) must then be parameters and are denoted \( \bar{x} \in \mathbb{R}^{n-a} \).

Hence, using the implicit function theorem (see Theorem A.1), Hypothesis 2.1 implies the existence of a diffeomorphism \( \zeta \) defined on a neighborhood \( U \subset \mathbb{R}^{(\mu+2)n+1-r} \) of \((t_0, \bar{x}_0, p_0)\), which is the part of \( z_\mu^0 \) corresponding to the selected parameters in \((t, \bar{x}, p)\), and a neighborhood \( \mathcal{V}(\mu+2)^n+1 \) of \( z_\mu^0 \) such that

\[
\mathcal{L}_\mu \cap \mathcal{V} = \{ \zeta(t, \bar{x}, p) \mid (t, \bar{x}, p) \in U \}
\]

From this expression follows that \( F_{\mu}^d(z_\mu) = 0 \) if and only if \( z_\mu = \zeta(t, \bar{x}, p) \) for some \((t, \bar{x}, p) \in U\). More specifically, \( x_2 \) and \( x_h \) are possible to express as

\[
\begin{align*}
x_2 &= \mathcal{G}(t, \bar{x}, p) \quad (2.22) \\
x_h &= \mathcal{H}(t, \bar{x}, p) \quad (2.23)
\end{align*}
\]

and on \( U \), the equation defining the manifold \( \mathcal{L}_\mu \) can be rewritten as

\[
F_{\mu}^d(t, \bar{x}, \mathcal{G}(t, \bar{x}, p), \mathcal{H}(t, \bar{x}, p)) = 0 \quad (2.24)
\]

The next step is to show that \( \mathcal{G} \) locally on \( U \) only depends on \( \bar{x} \) and \( t \), and not on \( p \). On \( U \) we define

\[
\hat{F}_2 = Z_2^T F_{\mu}^d
\]

where \( Z_2 \) is given by part 4 in Hypothesis 2.1. That is, \( \hat{F}_2 \) is formed from linear combinations of the rows of \( F \) and derivatives of \( F \). The result is

\[
\hat{F}_2(t, \bar{x}, \mathcal{G}(t, \bar{x}, p), \mathcal{H}(t, \bar{x}, p)) = 0 \quad (2.25)
\]

on \( U \). Differentiation of (2.25) with respect to \( p \) yields

\[
\frac{d}{dp} \hat{F}_2 = (Z_{2;x_2}^T F_{\mu}^d + Z_2^T F_{\mu;x_2}) \mathcal{G}_p + (Z_{2;x_{\mu+1}}^T F_{\mu}^d + Z_2^T F_{\mu;x_{\mu+1}}) \mathcal{H}_p = Z_2^T F_{\mu;x_2}^d \mathcal{G}_p = 0
\]

for all \((t, \bar{x}, p) \in U\). Here, we have used that on the neighborhood \( U \), it is known that \( F_{\mu}^d \equiv 0 \) and that \( Z_2^T F_{\mu;x_{\mu+1}}^d = 0 \). By construction, the variables \( x_2 \) were chosen such that \( Z_2^T F_{\mu;x_2}^d \) is nonsingular. Hence

\[
\mathcal{G}_p(t, \bar{x}, p) = 0
\]

on \( U \). The function \( \mathcal{G}_p \) is therefore constant with respect to \( p \), and locally there exists a function \( \varphi \) such that

\[
\varphi(t, \bar{x}) = \mathcal{G}(t, \bar{x}, p_0)
\]

Using the function \( \varphi \), (2.22) can be rewritten as

\[
x_2 = \varphi(t, \bar{x}) \quad (2.26)
\]
and the conclusion is that on \( \bar{U} \), \( x_2 \) does not depend on derivatives of \( x \), since \( \bar{x} \) only consists of terms in \( x \).

Differentiating (2.25) where (2.26) has replaced (2.22), i.e.,

\[
\hat{F}_2(t, \bar{x}, \varphi(t, \bar{x}), H(t, \bar{x}, p)) = 0
\]

with respect to \( \bar{x} \) yields

\[
\frac{d}{d\bar{x}} \hat{F}_2 = Z_{2,\bar{x}} F_\mu + Z_{2,\bar{x}x} F_\mu \bar{x} + (Z_{2,\bar{x}x} F_\mu + Z_{2,0} F_{\mu,x}) \varphi_{\bar{x}}
\]

\[
+ (Z_{2,\mu+1} F_\mu + Z_{2,\mu+1} F_{\mu,x+1}) H_{\bar{x}}
\]

\[
= Z_{2,\bar{x}} F_\mu \bar{x} + Z_{2,\bar{x}x} F_\mu \varphi_{\bar{x}} = Z_{2,\bar{x}} F_\mu \varphi_{\bar{x}} (I_{n-a}) = 0
\]

on \( \bar{U} \). Here \( I_{n-a} \) is an identity matrix of dimension \( n-a \times n-a \) and again we have used that \( F_\mu \equiv 0 \) and that \( Z_{2,\bar{x}} F_\mu \bar{x} = 0 \). In part 4 of the hypothesis one requirement was the existence of a function \( T_2 \) such that \( Z_{2,\bar{x}} F_\mu = 0 \). Using the result in (2.27), it is possible to choose \( T_2 \) as

\[
T_2(t, \bar{x}) = \left( I_{n-a} \varphi(t, \bar{x}) \right)
\]

This choice of \( T_2 \) makes it possible to interpret the condition in part 5. First notice that

\[
F_{\bar{x}T_2} = F_{\bar{x}} + F_{\bar{x}\mu} \varphi(t, \bar{x}) = \frac{d}{d\bar{x}} F(t, \bar{x}, \varphi(t, \bar{x}), \hat{x}, \varphi_{\bar{x}}(t, \bar{x})\hat{x})
\]

on \( \bar{U} \). From part 5 it is known that rank \( F_{\bar{x}}T_2 = d \) and thus \( d \) variables of \( \bar{x} \), denoted \( x_1 \), have a derivative \( \hat{x}_1 \) that is determined as a function of the other variables. The other variables in \( \bar{x} \), continue to be parameters. These variables are denoted \( x_p \in \mathbb{R}^{n-a-d} \). Part 5 also implies that there exists a matrix function \( Z_1 \in \mathbb{R}^{m+d} \) with full rank such that

\[
\text{rank } Z_1 T_2 F_{\bar{x}} = d
\]

on \( \bar{U} \). Since the rank was \( d \) without \( Z_1 \), it is possible to choose \( Z_1 \) constant.

Summarizing the construction up to now, Hypothesis 2.1 implies that the original system locally on \( \bar{U} \) can be rewritten as a reduced system (in the original variables) given by

\[
\hat{F}_1(t, x_1, x_p, \varphi(t, x_1, x_p), \hat{x}_1, \hat{x}_p, \varphi(t, x_1, x_p)) = 0
\]

\[
x_2 - \varphi(t, x_1, x_p) = 0
\]

(2.28a)

(2.28b)

where we have used the definition

\[
\hat{F}_1 = Z_1^T F
\]

and \( \varphi(t, x_1, x_p) \) is

\[
\varphi(t, x_1, x_p) = \varphi(t, x_1, x_p) + \varphi_{x_1}(t, x_1, x_p)\hat{x}_1 + \varphi_{x_p}(t, x_1, x_p)\hat{x}_p
\]
From the discussion above, it is known that at least locally it is possible to solve (2.28a) for \( \dot{x}_1 \) yielding the system

\[
\dot{x}_1 = \mathcal{F}(t, x_1, x_p, \dot{x}_p) \\
x_2 = \varphi(t, x_1, x_p)
\]  

(2.29)

Based on Hypothesis 2.1, theorems describing when a nonlinear descriptor system is solvable can be formulated. First, a theorem is presented stating when a solution to the descriptor system (2.3) also solves the reduced system (2.29).

**Theorem 2.1**

Let \( F \) in (2.3) be sufficiently smooth and satisfy Hypothesis 2.1 with some \( \mu, a, d \) and \( v \). Then every solution of (2.3) also solves the reduced problem (2.29) consisting of \( d \) differential equations and \( a \) algebraic equations.

**Proof:** This theorem follows immediately from the procedure above, see Kunkel and Mehrmann (2001).

Notice that the procedure yields a constructive method to compute the reduced system. We also formulate a theorem giving sufficient conditions for the reduced system (2.29) to yield the solution of the original description (2.3), at least locally.

**Theorem 2.2**

Let \( F \) in (2.3) be sufficiently smooth and satisfy Hypothesis 2.1 with some \( \mu, a, d \) and \( v \). Further let \( \mu + 1 \) give the same \( a, d \) and \( v \). Assume \( z_{\mu+1}^0 \in \mathbb{L}_{\mu+1} \) to be given and let \( p \) in (2.24) for \( F_{\mu+1} \) include \( \dot{x}_p \). Then for every function \( x_p \in C^1(\mathbb{I}, \mathbb{R}^{n-a-d}) \) with \( x_p(t_0) = x_{p,0}, \dot{x}_p(t_0) = \dot{x}_{p,0}, \) the reduced system (2.29) has unique solutions \( x_1 \) and \( x_2 \) satisfying \( x_1(t_0) = x_{1,0}. \) Moreover, together these solutions solve the original problem locally.

**Proof:** See Kunkel and Mehrmann (2001).

Often, the considered physical processes are well-behaved in the sense that no equations are redundant and the number of components in \( x \) is the same as the number of rows in \( F \). Then \( v = 0 \) and \( m = n \). Furthermore, the control input is assumed to be either a known time signal which is sufficiently smooth and handled separately as a time variability, or a feedback law. Then Theorem 2.2 can be simplified since no free parameters \( x_p \) will occur.

**Corollary 2.1**

Let \( F \) in (2.2) be sufficiently smooth and satisfy Hypothesis 2.1 with \( \mu, a, d \) and \( v = 0 \) and assume that \( a + d = n \). Furthermore, assume that \( \mu + 1 \) yields the same \( \mu, a, d \) and \( v = 0 \). For every \( z_{\mu+1}^0 \in \mathbb{L}_{\mu+1} \), the reduced problem (2.29) has a unique solution satisfying the initial condition given by \( z_{\mu+1}^0 \). Furthermore, this solution solves the original problem locally.

**Proof:** See Kunkel and Mehrmann (2001).
Remark 2.3. Sometimes it is interesting only to consider solvability on some part of the manifold defined by

$$L_\mu = \{ z_\mu \in I \times \mathbb{R}^n \times \ldots \times \mathbb{R}^n \mid F_\mu(z_\mu) = 0 \}$$

This is possible if $L_\mu$ instead is defined as

$$L_\mu = \{ z_\mu \in I \times \Omega_x \times \ldots \times \Omega_{x^{\mu+1}} \mid F_\mu(z_\mu) = 0 \}$$

where

$$\Omega_{x^{(i)}} \subset \mathbb{R}^n, \ i = 0, \ldots, \mu + 1$$

and $\Omega_{x^{(i)}}$ are open sets. That is, the region on which each variable is defined is not the whole $\mathbb{R}^n$.

To illustrate the method described above an example is presented.

Example 2.4

Consider a system described by the semi-explicit description (2.6) with $F_1(x_1, x_2, u) \in C^1(\mathbb{D} \times \mathbb{R}^{n_2} \times \mathbb{R}^p, \mathbb{R}^{n_1})$ and $F_2(x_1, x_2, u) \in C^1(\mathbb{D} \times \mathbb{R}^{n_2} \times \mathbb{R}^p, \mathbb{R}^{n_2} \times \mathbb{R}^p)$ and where $\mathbb{D} \subset \mathbb{R}^{n_1}$ is an open set. The system is assumed to satisfy the assumption below.

Assumption A1. Assume there exists an open set $\tilde{\Omega}_x \subset \mathbb{D}$ such that for all $(\tilde{x}_1, \tilde{u}) \in \tilde{\Omega}_{x,u} = \{ x_1 \in \tilde{\Omega}_x, u \in \mathbb{R}^p \}$ it is possible to solve $F_2(\tilde{x}_1, \tilde{x}_2, \tilde{u}) = 0$ for $\tilde{x}_2$. We define the corresponding solution manifold as

$$\tilde{\Omega} = \{ x_1 \in \tilde{\Omega}_x, x_2 \in \mathbb{R}^{n_2}, u \in \mathbb{R}^p \mid F_2(x_1, x_2, u) = 0 \}$$

which not necessarily will be an open set. Further assume that the Jacobian matrix of the constraint equations with respect to $x_2$, i.e., $F_{2,x_2}(\tilde{x}_1, \tilde{x}_2, \tilde{u})$, is nonsingular for $(\tilde{x}_1, \tilde{x}_2, \tilde{u}) \in \tilde{\Omega}$. That is, the rank of $F_{2,x_2}$ is assumed to be constant and full on the solution manifold.

Using the implicit function theorem, see Theorem A.1, the assumption tells us that for every point $(\tilde{x}_1, \tilde{u}) \in \tilde{\Omega}_{x_1,u}$ there exist a neighborhood $\tilde{\Omega}_{\tilde{x}_1,u}$ of $(\tilde{x}_1, \tilde{u})$ and a corresponding neighborhood $\tilde{\Omega}_{\tilde{x}_2}$ of $\tilde{x}_2$ such that for each point $(x_1, u) \in \tilde{\Omega}_{\tilde{x}_1,u}$ a unique solution $x_2 \in \tilde{\Omega}_{\tilde{x}_2}$ exists and the solution can be given as

$$x_2 = \varphi_{\tilde{x}_1, \tilde{u}}(x_1, u) \quad (2.30)$$

where the subscript $\tilde{x}_1, \tilde{u}$ is included to clarify that the implicit function is only local.

The solvability of the semi-explicit system can now be investigated. In a behavioral manner, $x$ and $u$ are concatenated to a vector, and it can be shown that Hypothesis 2.1 is satisfied on

$$L_0 = \{ z_0 \in \tilde{\Omega}_x \times \mathbb{R}^{n_2} \times \mathbb{R}^p \times \mathbb{R}^a \times \mathbb{R}^p \mid F_0^d(z_0) = 0 \}$$

with $\mu = 0$, $d = n_1$, $a = n_2$ and $v = 0$ and the resulting reduced system is given by

$$\begin{align*}
\dot{x}_1 &= F_1(x_1, x_2, u) \quad (2.31a) \\
x_2 &= \varphi_{x_1,0, u_0}(x_1, u) \quad (2.31b)
\end{align*}$$
in some neighborhood of \( x_{1,0} \) and \( u_0 \) which both belong to \( \mathbb{L}_0 \). Furthermore, it can be shown that the same \( d, a \) and \( v \) satisfy the hypothesis for \( \mu = 1 \) on
\[
\mathbb{L}_1 = \{ z_1 \in \tilde{\Omega}_x \times \mathbb{R}^{n_2} \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^p \mid F_1^d(z_1) = 0 \}
\]
and that the parameters \( p \) in \( F_1^d \) can be chosen to include \( \dot{u} \). Given the initial conditions \((x_{1,0}, x_{2,0}, u_0) \in \tilde{\Omega} \) the initial conditions \((\ddot{x}_{1,0}, \ddot{x}_{2,0}, \dot{x}_{1,0}, \dot{x}_{2,0}, \dot{u}_0) \) are possible to choose such that \( z_{1,0}^0 \in \mathbb{L}_1 \). From Theorem 2.2 it then follows that for every continuously differentiable \( u(t) \) with \( u(t_0) = u_0 \), a unique solution exists for (2.31) such that \( x_1(t) = x_{1,0} \). Moreover, this solution locally solves the original system description.

Note that no \( \dot{u} \) appear in the reduced system and therefore no initial condition \( \dot{u}(t_0) = \dot{u}_0 \) need to be specified when solving the system in practice.

In the sequel of this thesis, Assumption A1 will most often be combined with a second assumption. Therefore, a combined assumption is formulated below.

**Assumption A2.** Assume that Assumption A1 is satisfied. Furthermore, assume that one of the sets \( \Omega_{\tilde{x}_1, \tilde{u}} \) is global in \( u \) in the sense that it can be expressed as
\[
\Omega_{\tilde{x}_1, \tilde{u}} = \{ x_1 \in \Omega_x, u \in \mathbb{R}^p \}
\]  
(2.32)
where \( \Omega_x \) is a neighborhood of \( x_1 = 0 \).

For notational convenience in the sequel of the thesis, a corresponding set \( \Omega \) is defined as
\[
\Omega = \{ x_1 \in \Omega_x, x_2 \in \mathbb{R}^{n_2}, u \in \mathbb{R}^p \mid F_2(x_1, x_2, u) = 0 \}
\]  
(2.33)
on which the implicit function solving \( F_2(x_1, x_2, u) = 0 \) is denoted
\[
x_2 = \varphi(x_1, u), \ \forall x_1 \in \Omega_x, u \in \mathbb{R}^p
\]  
(2.34)

For a linear descriptor system that is square, i.e., has as many equations as variables \( x \), the solvability conditions will reduce to the following theorem.

**Theorem 2.3 (Solvability)**
Consider a linear time-invariant DAE
\[
E \dot{x} = Ax + Bu
\]
with regular \( sE - A \), that is \( \text{det}(sE - A) \neq 0 \), and a given control signal \( u \in C^\nu(\mathbb{I}, \mathbb{R}^p) \).
Then the system is solvable and every consistent initial condition yield a unique solution.

**Proof:** See Kunkel and Mehrmann (1994). \( \square \)

To simplify the notation in the sequel of the thesis, a linear time-invariant descriptor with regular \( sE - A \) is denoted a regular system.

The definition of a solution for a general possibly nonlinear descriptor system requires \( x(t) \) to be continuously differentiable. This is the classical requirement and for state-space systems (2.1) with a smooth enough system matrix \( F \), it will basically impose the control input to be continuous.
However, a continuous control input to a descriptor system will in certain cases not lead to a continuously differentiable solution. Even if the solution of the descriptor system does not depend on derivatives of the control input, it is still a fact that the solution to the algebraic part, i.e., $x_2 = \varphi(x_1, u)$, can be only continuous since $u$ can influence the solution directly. Therefore, a more natural requirement on the solution might be to require the solution to the dynamical part to be continuously differentiable while the solution to the algebraic part is allowed only to be continuous. In this case, continuous control inputs can be used, if no derivatives of them appear.

One further generalization of the solution to state-space systems is that the solution only needs to be piecewise continuously differentiable. Then the control inputs are allowed to be only piecewise continuous, which is a rather common case, e.g., when using computer generated signals. A natural extension of the solvability definition for descriptor systems would then be to require piecewise continuous differentiability of the solution $x_1$ to the dynamic part, while the solution $x_2$ to the algebraic part only would need to be piecewise continuous.

For linear time-invariant systems it is possible to relax the requirements even more and define the solution in a distributional sense, see Dai (1989). With this framework, there exists a distributional solution even when the initial condition does not satisfy the explicit and implicit constraints or when the control input is not sufficiently differentiable. For a more thorough discussion about distributional solutions, the reader is referred to Dai (1989), or the original works by Verghese (1978) and Cobb (1980).

### 2.4 Index Reduction

Index reduction is a procedure which takes a high index problem and rewrites it as a lower index description, often index one or zero. Of course, the objective is to obtain a description that is easier to handle, and often also to reveal the manifold which the solution must belong to. The key tool for lowering the index of a description and exposing the implicit constraints is differentiation. Index reduction procedures are often the same methods that are used either to compute the index of a system description or to show solvability.

Numerical solvers for descriptor systems normally use index reduction methods in order to obtain a system description of at most index one. The reason is that many numerical solvers are designed for index one descriptions (Brenan et al., 1996). Therefore, index reduction is a well-studied area, see, for example (Mattson and Söderlind, 1993; Kunkel and Mehrmann, 2004; Brenan et al., 1996) and the references therein.

#### 2.4.1 Consecutive Differentiations

The most basic procedure for reducing the index of a system is to use the methods for finding the differential index. The first method can be found in Example 2.4 and the other is to use the derivative array. Hence, after some symbolical differentiations and manipulations, the result is an ODE description

$$\dot{x} = \mathcal{R}(x, u, \ldots, u^{(\nu-1)})$$

(2.35)
The description (2.35) is equivalent to the original DAE in the sense that they yield the same solution given consistent initial conditions. However, without considering the initial conditions, the solution manifold of (2.35) is much larger than for the original DAE. To reduce the solution manifold and regain the same size as for the original problem the explicit and implicit constraints, obtained in the index reduction procedure, need to be considered. For this purpose, the constraints can be used in different ways.

One way is to do as described above. That is, the constraints are used to define a set \( \Omega_0 \) and the initial condition \( x(t_0) \) is then assumed to belong to this set, i.e., \( x(t_0) \in \Omega_0 \). This way can be seen as a method to deal with the constraints implicitly. Another choice is to augment the system description with the constraints as the index reduction procedure proceeds. The result is then an overdetermined but well-defined index one descriptor system. Theoretically, the choices are equivalent. However, in numerical simulation the methods have some differences.

A drawback with the first method is that it suffers from drift off, which often leads to numerical instability. It means that even if the initial condition is chosen in \( \Omega_0 \), small errors in the numerical computations result in a solution to (2.35) which diverge from the solution of the original descriptor system. This is a result of the larger solution set of (2.35) compared to the original system description. A solution to this problem is to use methods known as constraint stabilization techniques (Baumgarte, 1972; Ascher et al., 1994).

For the second choice, the solution manifold is the same as for the original DAE. However, the numerical solver discretizes the problem and then according to Mattson and Söderlind (1993), an algebraically allowed point in the original DAE may be non-allowed in the discretized problem and vice versa. This problem can be handled using special projection methods, see references in Mattson and Söderlind (1993).

The problem with non-allowed points occurs because of the overdeterminedness obtained when all equations are augmented. Therefore, Mattson and Söderlind (1993) present another method where dummy derivatives are introduced. Extra variables are added to the augmented system which instead of being overdetermined becomes determined. The discretized problem will then be well-defined.

### 2.4.2 Consecutive Differentiations of Linear Time-Invariant Descriptor Systems

Index reduction of linear time-invariant descriptor systems (2.5) is an important special case. One method is to use the Shuffle algorithm, described in Example 2.2. The Shuffle algorithm applied to (2.5) results in a system in the form

\[
\dot{x} = \bar{E}^{-1} \left( \bar{A}x + \sum_{i=0}^{\nu} \bar{B}_i u^{(i)} \right) \tag{2.36}
\]

The steps for obtaining the matrices \( \bar{E}, \bar{A} \) and \( \bar{B}_i \) for \( i = 0, \ldots, \nu \) will now be described.

Form the matrix \( \begin{pmatrix} E & A & B \end{pmatrix} \). Use Gauss-elimination to obtain the new matrix

\[
\begin{pmatrix}
  E_1 & A_1 & B_1 \\
  0 & A_2 & B_2
\end{pmatrix}
\]
where $E_1$ is nonsingular. This matrix corresponds to the descriptor system

$$\begin{pmatrix} E_1 \\ 0 \end{pmatrix} \dot{x} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} x + \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} u$$

Differentiation of the constraint equation, i.e., the lower row, yields the description

$$\begin{pmatrix} E_1 \\ -A_2 \end{pmatrix} \dot{x} = \begin{pmatrix} A_1 \\ 0 \end{pmatrix} x + \begin{pmatrix} B_1 \\ 0 \end{pmatrix} u + \begin{pmatrix} 0 \\ B_2 \end{pmatrix} \dot{u}$$

If $\bar{E}$ has full rank, the description (2.36) is obtained by multiplying with the inverse of $\bar{E}$ from the left. Otherwise, the procedure is repeated. The procedure is guaranteed to terminate if and only if the system is regular, see Dai (1989).

Another method with the advantage that it separates the dynamical and the algebraic parts is to use the canonical form. For linear descriptor systems (2.5), the canonical form is

\[ \begin{array}{l}
\dot{x}_1 = A_1 x_1 + B_1 u \\
N \dot{x}_2 = x_2 + B_2 u 
\end{array} \tag{2.37a} \tag{2.37b} \]

The matrix $N$ is a nilpotent matrix, i.e., $N^k = 0$ for some integer $k$, and it can be proven that this $k$ is the index of the system, i.e., $k = \nu$ (Brenan et al., 1996). A system with $\det(sE - A) \neq 0$, can always be rewritten on this form and a computational method to achieve this form can be found in Gerdin (2004).

Consecutive differentiations of the second row in the canonical form (2.37) will lead to that another form

\[ \begin{array}{l}
\dot{x}_1 = A_1 x_1 + B_1 u \\
x_2 = -\sum_{i=0}^{\nu-1} N^i B_2 u^{(i)}(t) 
\end{array} \tag{2.38a} \tag{2.38b} \]

can be obtained. Here, it has been assumed that only consistent initial values $x(0)$ are considered. The form (2.38) is widely used to show different properties for linear time-invariant descriptor systems, see Dai (1989).

Note that when the dynamical and algebraical parts of the original descriptor system are separated like in (2.38), the numerical simulation becomes very simple. Only the dynamical part needs to be solved using an ODE solver, while the algebraic part is given by the states and the control input. Therefore, no drift off or problems due to discretization will occur. This is a major advantage with this system description.

Results on a nonlinear version of the canonical form (2.37) can be found in Rouchon et al. (1992).

### 2.4.3 Kunkel and Mehrmann’s Method

The procedure presented in Section 2.3 when defining solvability for descriptor systems can also be seen as an index reduction method. If $\mu$, $d$, $a$ and $v$ are found such that
Hypothesis 2.1 is satisfied, it is in principle possible to express the original description in the form (2.29), i.e.,

\[ \dot{x}_1 = F(t, x_1, x_p, \dot{x}_p) \]
\[ x_2 = \varphi(t, x_1, x_p) \]

which has strangeness index zero. Unfortunately this description is only valid locally in some neighborhood, and it might be impossible to find explicit expressions for the functions \( F \) and \( \varphi \). However, the form above is obtained by solving the description

\[ \hat{F}_1(t, x_1, x_2, x_p, \dot{x}_1, \dot{x}_p, \dot{x}_2) = 0 \] (2.39a)
\[ \hat{F}_2(t, x_1, x_2, x_p) = 0 \] (2.39b)

with respect to \( \dot{x}_1 \) and \( x_2 \), and (2.39) has also strangeness index zero. Unlike \( F \) and \( \varphi \), it is often possible to express the functions \( \hat{F}_1 \) and \( \hat{F}_2 \) explicitly using the system function \( F \) and possibly its derivatives (this is what the matrix functions \( Z_1 \) and \( Z_2 \) in Hypothesis 2.1 do).

More practical aspects of the method described in this section can be found in Kunkel and Mehrmann (2004) and in Arnold et al. (2004).

**Remark 2.4.** For given \( \mu, d, a \) and \( v \) the index reduction process is performed in one step. Hence, no rank assumptions on intermediate steps are necessary. This may be an advantage compared to other index reduction procedures.

The behavioral approach to handle the control signals \( u \) used in Kunkel and Mehrmann (2001), i.e., to group \( x \) and \( u \) into one big \( x \), might not fit our purposes. For control problems (2.2) the signals possible to use for control are often given by the physical plant. This means that at least some of the parameters \( x_p \) are given by the physical context. However, if \( u \) is handled in a behavioral manner, the index reduction procedure may yield undesired results as shown in the example below.

**Example 2.5**

Consider a linear time-invariant descriptor system. First the control signal is included in the \( x \) variable, i.e., \( x = (z_1, z_2, z_3, u)^T \). The dynamics are then described by

\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
\dot{x}_3
\end{pmatrix}
= 
\begin{pmatrix}
2 & 0 & 0 & 1 \\
0 & 1 & 0 & 2 \\
0 & 0 & 1 & 3
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3
\end{pmatrix}
\]

The result from Kunkel and Mehrmann’s index reduction procedure is a reduced system, with strangeness index equal to \( \mu = 0 \), given by

\[ \dot{z}_1 = 2z_1 - z_3 \]
\[ \dot{z}_3 = -2z_3 + z_2 \]
\[ u = -z_3 \]

Hence, the control input is seen as an algebraic variable which is given by \( z_3 \), while the free parameter is \( z_2 \).
In the next computation, \( u \) is instead seen as a given signal, which is included in the time variability. Hence, \( x = (z_1, z_2, z_3) \), and the system description can be written as

\[
\begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0
\end{pmatrix}
\dot{x} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} x + \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} u
\]

Applying the index reduction procedure on this description yields a dynamic part

\[
\begin{align*}
\dot{z}_1 &= 2z_1 + u \\
z_2 &= -2u - 3\dot{u} \\
z_3 &= -3u
\end{align*}
\]

For this description, \( \mu = 1 \). Hence, by considering \( u \) as a given time signal, the strangeness index has increased and \( \dot{u} \) has appeared as a parameter.

The example above clearly shows that this index reduction procedure does not necessarily choose the control input \( u \) as parameter.

Therefore, we will not use the behavioral approach and if the system (2.2) is handled without grouping \( x \) and \( u \), and there exist \( \mu, d, a \) and \( v \) such that Hypothesis 2.1 is satisfied, the functions \( \tilde{F}_1 \) and \( \tilde{F}_2 \) in (2.39) become

\[
\begin{align*}
\tilde{F}_1(t, x_1, x_2, \dot{x}_1, \dot{x}_2, u) &= 0 \\
\tilde{F}_2(x_1, x_2, u, \dot{u}, \ldots, u^{(\mu)}) &= 0
\end{align*}
\]

Since we assume that Hypothesis 2.1 is satisfied it is in principle possible, at least locally, to solve (2.40) for \( \dot{x}_1 \) and \( x_2 \) to obtain

\[
\begin{align*}
\dot{x}_1 &= \mathcal{F}(x_1, u, \ldots, u^{(\mu)}) \\
x_2 &= \varphi(x_1, u, \dot{u}, \ldots, u^{(\mu)})
\end{align*}
\]

The system above is in semi-explicit form (2.6), but as mentioned earlier it may be impossible to find explicit expressions for \( \mathcal{F} \) and \( \varphi \). However, often in this thesis, it will be assumed that the system (2.40) can be written in semi-explicit form with system functions \( F_1 \) and \( F_2 \) given in closed form. We formulate this assumption more formally.

**Assumption A3.** The variables \( \dot{x}_1 \) can be solved from (2.40a) to give

\[
\begin{align*}
\dot{x}_1 &= \tilde{F}_1(x_1, x_2, u, \ldots, u^{(\mu + 1)}) \\
0 &= \tilde{F}_2(x_1, x_2, u, \dot{u}, \ldots, u^{(\mu)})
\end{align*}
\]

where \( \tilde{F}_1 \) and \( \tilde{F}_2 \) are possible to express explicitly.

It may seem strange that \( \dot{x}_2 \) has disappeared in \( \tilde{F}_1 \). However, differentiation of (2.42b) makes it is possible to get an expression for \( \dot{x}_2 \) as

\[
\dot{x}_2 = -\tilde{F}^{-1}_{2; x_2}(x_1, x_2, u)(\tilde{F}_{2; x_1}(x_1, x_2, u)\dot{x}_1 + \tilde{F}_{2; u}(x_1, x_2, u)\dot{u})
\]
where \( u = (u, \dot{u}, \ldots, u^{(\mu)}) \). Using this expression, \( \dot{x}_2 \) can be eliminated from \( \tilde{F}_1 \).

The class of applications where \( \hat{F}_1 \) actually is affine in \( \dot{x}_1 \) seems to be rather large. For example mechanical multibody systems can in many cases be written in this form, see Kunkel and Mehrmann (2001).

One complication is the possible presence of derivatives of the control variable (originating from differentiations of the equations). If the procedure is allowed to choose the input signal, or with other words the parameter \( x_p \), freely the highest possible derivative is \( \dot{x}_p \). However, if the control input is chosen by the physical plant, the highest possible derivative becomes \( u^{(\mu+1)} \). For linear systems, it is possible to make transformations removing the input derivatives from the differential equations and just have the derivatives in the algebraic part as can be seen in (2.38). This might not be possible in the nonlinear case. In that case, it could be necessary to redefine the control signal so that its highest derivative becomes a new control variable and the lower order derivatives become state variables. This procedure introduces an integrator chain

\[
\dot{x}_{1,n_1+1} = x_{1,n_1+2} \\
\vdots \\
\dot{x}_{1,n_1+\mu+1} = u^{(\mu+1)}
\]

If the integrator chain is included, the system description (2.42) becomes

\[
\dot{x}_1 = F_1(x_1, x_2, u) \\ 0 = F_2(x_1, x_2, u)
\]

where \( x_1 \in \mathbb{R}^{n_1+\mu+1} \), \( x_2 \in \mathbb{R}^{n_2} \) and \( u \in \mathbb{R}^p \). Here \( u^{(\mu+1)} \) is denoted \( u \) in order to notationally match the sequel of this thesis.

### 2.5 Stability

This section concerns stability analysis of descriptor systems. In principle, stability of a descriptor system means stability of a dynamical system on a manifold. The standard tool, and basically the only tool, for proving stability for nonlinear systems is Lyapunov theory. The main concept in the Lyapunov theory is the use of a Lyapunov function, see Lyapunov (1992). The Lyapunov function is in some sense a distance measure between the variables \( x \) and an equilibrium point. If this distance measure decreases or at least is constant, the state is not diverging from the equilibrium and stability can be concluded.

A practical problem with Lyapunov theory is that in many cases, a Lyapunov function can be difficult to find for a general nonlinear system. However, for mechanical and electrical systems, often the total energy content of the system can be used.

The stability results will be focused on two system descriptions, the semi-explicit, autonomous, index one case and the linear case. These two cases will be the most important for the forthcoming chapters. However, a small discussion about polynomial possibly higher index systems will be presented at the end of this section. For this kind of systems a computationally tractable approach, based on Lyapunov theory, has been published in Ebenbauer and Allgöwer (2004).
Consider the autonomous descriptor system

\[ F(\dot{x}, x) = 0 \]  \hspace{1cm} (2.44)

where \( x \in \mathbb{R}^n \). This system can be thought of as either a system without control input or as a closed loop system with feedback \( u = u(x) \).

Assume that there exists an open connected set \( \Omega \) of consistent initial conditions such that the solution is unique, i.e., the initial value problem consisting of (2.44) together with \( x(t_0) \in \Omega \) has a unique solution. Note that in the state-space case this assumption will simplify to \( \Omega \) being some subset of the domain where the system satisfies a Lipschitz condition.

Stability is studied and characterized with respect to some equilibrium. Therefore, it is assumed that the system has an equilibrium \( x^0 \in \Omega \). Without loss of generality the equilibrium can be assumed to be the origin, since if \( x^0 \neq 0 \), the change of variables \( z = x - x^0 \) can be used. In the equilibrium, (2.44) gives

\[ 0 = F(0, x^0) = \bar{F}(0, 0) \]

where \( \bar{F}(\dot{z}, z) = F(\dot{x}, x) \). Hence, in the new variables \( z \), the equilibrium has been shifted to the origin.

Finally, the set \( \Omega \) is assumed to contain only a single equilibrium. Hence, in order to satisfy this assumption, it might be necessary to reduce \( \Omega \). However, this assumption can be relaxed using concepts of set stability, see Hill and Mareels (1990).

The definitions of stability for descriptor systems are natural extensions of the corresponding definitions for the state-space case.

**Definition 2.6 (Stability).** The equilibrium point at \((0, 0)\) of (2.44) is called stable if given a \( \varepsilon > 0 \), there exists a \( \delta(\varepsilon) > 0 \) such that for all \( x(t_0) \in \Omega \cap B_\delta \) it follows that \( x(t) \in \Omega \cap B_\varepsilon, \forall t > 0 \).

**Definition 2.7 (Asymptotic stability).** The equilibrium point at \((0, 0)\) of (2.44) is called asymptotically stable if it is stable and there exists a \( \eta > 0 \) such that for all \( x(t_0) \in \Omega \cap B_\eta \) it follows that

\[ \lim_{t \to \infty} \| x(t) \| = 0 \]

### 2.5.1 Semi-Explicit Index One Systems

Lyapunov stability for semi-explicit index one systems is a rather well-studied area, see for example Hill and Mareels (1990), Wu and Mizukami (1994) and Wang et al. (2002). In many cases, using the index reduction method in Section 2.4.3 and by assuming that Assumption A3 is satisfied, also higher index descriptions can be rewritten in semi-explicit form. Hence, we consider the case when (2.44) can be expressed as

\[ \dot{x}_1 = F_1(x_1, x_2) \]  \hspace{1cm} (2.45a)
\[ 0 = F_2(x_1, x_2) \]  \hspace{1cm} (2.45b)

where \( x_1 \in \mathbb{R}^{n_1} \) and \( x_2 \in \mathbb{R}^{n_2} \). The system is assumed to satisfy Assumption A2. This means that, on some set \( \Omega \), which in this case has the structure

\[ \Omega = \{ x_1 \in \Omega_x \subset \mathbb{R}^{n_1}, x_2 \in \mathbb{R}^{n_2}, \ \mid x_2 = \varphi(x_1) \} \]  \hspace{1cm} (2.46)
the system (2.45) has index one and a unique solution for arbitrary initial conditions in $\Omega$. We also assume that $\Omega$ is connected and contains the origin.

Lyapunov’s Direct Method

The method known as Lyapunov’s direct method is described in the following theorem, which is also called Lyapunov’s stability theorem.

**Theorem 2.4**

Consider the system (2.45) and let $\Omega' \subset \Omega$ be an open, connected set containing the origin. Suppose there exists a function $V \in C^1(\Omega', \mathbb{R})$ such that $V$ is positive definite and has a negative semidefinite time-derivative on $\Omega'$, i.e.,

$$
V(0) = 0 \quad \text{and} \quad V(x_1) > 0, \quad \forall x_1 \neq 0, \quad \text{(2.47a)}
$$

$$
V_{x_1}(x_1)F_1(x_1, \varphi(x_1)) \leq 0, \quad \forall x_1 \quad \text{(2.47b)}
$$

where $x_1 \in \Omega'$. Then the equilibrium $(x_0^1, x_0^2) = (0, 0)$ is stable. Moreover, if the function $V$ is negative definite on $\Omega'$, i.e.,

$$
V_{x_1}(x_1)F_1(x_1, \varphi(x_1)) < 0, \quad \forall x_1 \neq 0 \quad \text{(2.48)}
$$

where $x_1 \in \Omega'$, then $(x_0^1, x_0^2) = (0, 0)$ is asymptotically stable.

**Proof:** The proof is to a large extent based on the proof for the state-space case. For $x_1 \in \Omega'$ it follows that $(x_1, x_2) \in \Omega' = \{x_1 \in \Omega'_x, \ x_2 \in \mathbb{R}^{n_2} | F_2(x_1, x_2) = 0 \} \subset \Omega$. Then the system is given by the reduced system

$$
\dot{x}_1 = F_1(x_1, \varphi(x_1))
$$

Given an $\varepsilon$, choose $r \in (0, \varepsilon]$ such that

$$
B_r = \{x_1 \in \mathbb{R}^{n_1}, \ x_2 \in \mathbb{R}^{n_2} | \|(x_1, x_2)\| \leq r \} \subset \Omega'
$$

Since $\varphi(x_1)$ is at least continuously differentiable it follows that on $B_r$

$$
\|(x_1, x_2)\| \leq (1 + L)\|x_1\|
$$

for some $L > 0$. We choose

$$
B_{\|x_1\|} = \left\{ x_1 \in \mathbb{R}^{n_1} | \|x_1\| \leq \frac{r}{1 + L} \right\} \subset \Omega'_x
$$

Then it is known from for example Khalil (2002), that (2.47) guarantees the existence of a $\delta_{x_1} > 0$ and a corresponding set

$$
B_{\delta_{x_1}} = \{x_1 \in \mathbb{R}^{n_1} | \|x_1\| \leq \delta_{x_1} \} \subset B_{\|x_1\|}
$$

such that

$$
x_1(t_0) \in B_{\delta_{x_1}} \Rightarrow x_1(t) \in B_{\|x_1\|}, \quad \forall t \geq t_0
$$
We can then conclude that \((x_1(t), x_2(t))\) belong to \(\Omega'\) and that
\[
\left\| (x_1(t), x_2(t)) \right\| \leq r \leq \varepsilon
\]
By choosing \(\delta \leq \delta_{x_1}\) it is also certain that
\[
\left\| x_1 \right\| \leq \left\| (x_1, x_2) \right\| \leq \delta \leq \delta_{x_1}
\]
That is, by choosing \(\delta\) smaller than \(\delta_{x_1}\) is guaranteed that \(\left\| x_1 \right\|\) is smaller than \(\delta_{x_1}\) and stability is proven.

The discussion concerning asymptotic stability is similar. From Khalil (2002), it is known that the harder condition (2.48) implies the existence of a \(\eta_{x_1} > 0\) such that for \(x_1(t_0)\) in the set
\[
B_{\eta_{x_1}} = \{x_1 \in \mathbb{R}^{n_1} \mid \left\| x_1 \right\| \leq \eta_{x_1} \} \subset \Omega_{x_1}'
\]
it holds that \(\left\| x_1(t) \right\| \to 0\) as \(t \to \infty\). However, for \(x_1 \in \Omega_{x_1}'\) it follows that \((x_1, x_2) \in \Omega_{x_1}' \subset \Omega\) and by using
\[
\left\| (x_1, x_2) \right\| \leq (1 + L)\left\| x_1 \right\|
\]
we have that
\[
\lim_{t \to \infty} \left\| (x_1(t), x_2(t)) \right\| = 0
\]
if an \(\eta \leq \eta_{x_1}\) is chosen. This concludes the proof. \(\square\)

**Remark 2.5.** The previously stated solvability conditions guaranteed only the existence of a solution on some time interval \(I\). However, Definitions 2.6 and 2.7 require global existence of a solution, i.e., a solution for all \(t > t_0\). It can be shown that if there exists a Lyapunov function as required in Theorem 2.4, this property is ensured, see Khalil (2002). Briefly, the idea is that if the solution fails to exist for some \(t = T < \infty\), it must leave any compact set of \(\Omega\). However, the theorem above shows that this does not happen.

Contrary to the state-space case, the result in Theorem 2.4 is mainly of theoretical interest for descriptor systems, since the implicit function \(\varphi\) is used. Furthermore, the use of the implicit function theorem also implies that a fundamental limitation in whether this method can show global or local asymptotic stability is whether the inverse holds globally or not.

Some generalizations can be made to the results above. For example, the condition in (2.48) can be relaxed to provide a counterpart to the LaSalle Invariance Principle (Hill and Mareels, 1990; Khalil, 2002). Another generalization is the incorporation of systems with solutions exhibiting certain jump discontinuities, see Mahony and Mareels (1995).

**Lyapunov’s Indirect Method**

Since the conditions in Theorem 2.5 often are difficult to use in practice another method for proving stability is presented. This method is known as Lyapunov’s indirect method. The main idea is to use the linearization of (2.45) to determine local stability of the origin. Assume that \(F_1\) and \(F_2\) are continuously differentiable. Then, the linearization around \((x_1, x_2) = (0, 0)\) is given by
\[
\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + o(\|x\|) \quad (2.49a)
\]
\[
0 = A_{21}x_1 + A_{22}x_2 + o(\|x\|) \quad (2.49b)
\]
where

\[
A_{11} = F_{1;x_1}(0, 0) \\
A_{21} = F_{2;x_1}(0, 0)
\]

and \( o(||x||)/||x|| \to 0, ||x|| \to 0. \)

The next theorem gives conditions under which stability of the origin \((x_1, x_2) = (0, 0)\) can be concluded by investigating its stability as an equilibrium point for the linear part of (2.49).

**Theorem 2.5**

Consider the system (2.45) and let \( \Omega' \subset \Omega \) be a neighborhood of \( x_1 = 0 \). Then, the origin is asymptotically stable if \( \Re \lambda_i (A_{11} - A_{12}A_{22}^{-1}A_{21}) < 0 \) for \( i = 1, 2, \ldots, n_1 \).

**Proof:** For all \( (x_1, x_2) \in \Omega' = \{x_1 \in \Omega', x_2 \in \mathbb{R}^{n_2} | F_2(x_1, x_2) = 0\} \) it is known that \( F_{2;x_2} \) is nonsingular, and since \((0, 0) \in \Omega'\) it follows that \( A_{22} \) has full rank. Hence, (2.49b) can be reformulated as

\[
x_2 = -A_{22}^{-1}A_{21}x_1 + o(||x||)
\]

Combining (2.49) and the latter equation gives

\[
\dot{x}_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})x_1 + o(||x||)
\]

On a compact subset of \( \Omega' \) it holds that

\[
||(x_1, x_2)|| \leq (1 + L)||x_1||
\]

for some \( L > 0 \). This makes it possible to write the system as

\[
\dot{x}_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})x_1 + o(||x_1||) \tag{2.50}
\]

Hence, the linearization of the reduced system is obtained and for this system Theorem 4.7 in Khalil (2002) can be used to prove the statements. \( \Box \)

For more information about linearization of descriptor systems, the reader is referred to Campbell (1995).

**Converse Lyapunov Theorem**

Often in this these, the system (2.45) will be said to be asymptotically stable on a set \( \Omega \). Then, it is assumed that the set \( \Omega \), for which (2.45) has a unique solution and in which only one equilibrium exists, has been reduced to an invariant set. That is, for all \( x(t_0) \in \Omega \), it follows that \( x(t) \in \Omega \) for all \( t \geq t_0 \) and moreover that \( x(t) \to 0 \) as \( t \to \infty \).

The following theorem proves that such a reduction of \( \Omega \) can always be performed.

**Theorem 2.6**

Consider the system (2.45) and let \( x = 0 \) be an asymptotically stable equilibrium. Let \( \Omega' \subset \Omega \) be open, connected set containing the origin. Further, let \( R_A \subset \Omega' \) be a part of the region of attraction of \( x_1 = 0 \), where the region of attraction is defined as the set of
all points \( x_1(0) = x_{1,0} \) such that \( x_1(t) \to 0 \) as \( t \to \infty \). Then, there is a smooth, positive definite function \( V(x_1) \) and a continuous, positive definite function \( W(x_1) \) both defined for all \( x_1 \in \mathcal{R}_A \) such that

\[
V(0) = 0 \quad \text{and} \quad W(0) = 0
\]

\[
V(x_1) \to \infty, \ x_1 \to \partial \mathcal{R}_A
\]

\[
V_{x_1}(x_1)F_1(x_1, \varphi(x_1)) \leq -W(x_1), \ \forall x_1 \in \mathcal{R}_A
\]

where \( \partial \mathcal{R}_A \) denotes the boundary of \( \mathcal{R}_A \). Furthermore, for any \( c > 0 \), the set \( \{ x_1 \in \mathcal{R}_A \mid V(x_1) \leq c \} \) is a compact subset of \( \mathcal{R}_A \).

**Proof:** For \( x_1 \in \Omega'_x \), it holds that the system is given by

\[
\dot{x}_1 = F_1(x_1, \varphi(x_1))
\]

and the result then follows from Theorem 4.17 in Khalil (2002). \( \square \)

Hence, it is always possible to choose an invariant set for \( x_1 \) by choosing the set corresponding to some \( c \). However, then we can choose a corresponding set for \( (x_1, x_2) \) which also will be invariant since \( x_2 = \varphi(x_1) \).

### 2.5.2 Linear Systems

Stability for a linear state-space system

\[
\dot{x} = Ax
\]

is determined by the eigenvalues of the matrix \( A \). Also for linear descriptor systems the eigenvalues can be used to determine stability. However, since the matrix in front of \( \dot{x} \) in the descriptor system case normally is rank deficient, both finite and infinite eigenvalues occur, see Dai (1989), and stability is determined by the finite eigenvalues only as can be seen below.

We only consider asymptotic stability for regular linear descriptor systems

\[
E\dot{x} = Ax
\]

(2.51)

with consistent initial conditions. For such systems, the equilibrium \( x = 0 \) is unique and according to Theorem 2.3 the solution is unique. The following theorem can be used to show asymptotic stability.

**Theorem 2.7**

A regular linear descriptor system (2.51) is asymptotically stable if and only if

\[
\Re \sigma(E, A) < 0
\]

where \( \sigma(E, A) = \{ s \in \mathbb{C} \mid \det(se - A) = 0 \} \).

**Proof:** See Dai (1989). However, notice that in Dai (1989) the notion stable is used for the property we denote asymptotic stability. \( \square \)
Notice that the set $\mathbb{C}$ does not contain the infinity and therefore the theorem only considers finite $s$.

Using slightly different definitions of stability, where the initial condition does not need to be consistent, it is possible to extend the stability concept to include the more general solutions described in Section 2.3. The same condition as above, i.e., $\Re \sigma(E,A) < 0$, will still be obtained. To disregard possible impulsive behavior at the initial time, due to inconsistent initial values, it is important that the definition of stability does not include the initial time. It means that the stability definitions should be formulated for $t > t_0$ instead of $t \geq t_0$.

### 2.5.3 Barrier Function Method

Another interesting Lyapunov based approach to stability analysis of a descriptor system is presented in Ebenbauer and Allgöwer (2004). Instead of finding the reduced description (2.29) explicitly, this approach works with an implicit definition of the solution manifold. This is particularly attractive for nonlinear polynomial descriptor systems, since it implies that the stability condition can be computationally verified using semidefinite programming and sum of squares decomposition.

To simplify the explanation the system (2.44) is assumed to satisfy the conditions in Hypothesis 2.1 and it is also assumed that the underlying system (2.29) solves the original system locally. Then the derivative array $F_\mu$ for some $\mu$ implicitly defines the solution $x(t)$ on some neighborhood $U$.

It turns out to be convenient to define the set

$$
U_x = \{ x \in \Theta, x_{\mu+1} \in \mathbb{R}^{(\mu+1)n} \}
$$

where $x_{\mu+1} = (\dot{x}, \ldots, x^{(\mu+1)})$ and $\Theta \subset \mathbb{R}^n$ is a neighborhood of the origin. A theorem for stability of a nonlinear descriptor system can now be stated.

**Theorem 2.8**

The equilibrium $x = 0$ of (2.44) is stable if there exist a function $V \in C^1(\Omega_x, \mathbb{R})$ such that $V$ is positive definite and $V(x) \to \infty$ when $\|x\| \to \infty$, a function $\rho : \mathbb{R}^{(\mu+2)n} \to \mathbb{R} \cup \{+\infty\}$, and an positive integer $\mu$ such that

$$
V_x(x) \dot{x} \leq \|F_\mu(x, x_{\mu+1})\|^2 \rho(x, x_{\mu+1}) \quad (2.52)
$$

is satisfied for $(x, x_{\mu+1}) \in U_x$. If (2.52) is satisfied with inequality for all nonzero $x \in \Theta$, the system is asymptotically stable.

**Proof:** See Ebenbauer and Allgöwer (2004).
2.6 Optimal Control

Optimal control is the theory of finding a control input such that a certain measure, or performance criterion, is minimized for a particular a dynamical system. This is a well-studied area, which goes far back in time, see Sussmann and Willems (1997).

As for the stability analysis, optimal control for descriptor systems is in principle nothing but optimal control of a state-space system on a manifold. Consequently, the methods for descriptor systems will to a large extent rely on results for the state-space case. Therefore, a short summary of these results is presented using a notation matching the rest of this thesis.

2.6.1 Formulation and Summary of the Optimal Control Problem

Consider a continuous-time state-space system

\[ \dot{x} = F(x, u) \]  (2.53)

where \( x \in \mathbb{R}^n \) and \( u \in \mathbb{R}^p \). The initial condition is chosen as

\[ x(0) = x_0 \in \Omega_x \]  (2.54)

where \( \Omega_x \subset \mathbb{R}^n \) is a connected set. The considered class of performance criteria is

\[ J = \int_0^\infty L(x, u) \, dt \]  (2.55)

where the cost function \( L(x, u) \) is assumed positive semidefinite in \( x \) and positive definite in \( u \). That is, \( L(x, u) \geq 0 \) for all \((x, u) \in \mathbb{R}^{n+p}\) and \( L(x, u) > 0 \) when \( u \neq 0 \).

The optimal control problem is then formulated as

\[ V(x_0) = \min_{u(\cdot)} J \]  (2.56)

subject to the dynamics (2.53) and the initial condition (2.54). Note that in (2.56) the minimization is done with respect to the function \( u(\cdot) \). The notation \( u(\cdot) \) is used to indicate that it is not yet decided which structure the optimal control input will have, i.e., if it is to be interpreted as a time signal \( u(t) \) or as a feedback law \( u(x) \).

A common requirement in control theory is that the closed loop system obtained using the optimal control input be asymptotically stable. Therefore, the minimization is done with respect to all \( u(\cdot) \) such that

\[ \dot{x} = F(x, u(\cdot)) \]  (2.57)

is asymptotically stable on \( \Omega_x \). If \( L \) is positive definite also in \( x \), i.e., \( L(x, u) > 0 \) for \((x, u) \neq 0\), the requirement of asymptotic stability is implicitly included by the infinite time-horizon. The reason is that if \( x(t) \not\to 0 \) as \( t \to \infty \), the performance criterion cannot converge and the corresponding control law cannot be optimal. However, for a positive semidefinite \( L \), for example \( L = u^2 \), the requirement of stability must be handled explicitly. Otherwise, it will always be optimal to choose \( u(t) = 0 \).
In (2.56), it can be seen that the optimal performance criterion \( V(x_0) \) only depends on the initial condition and neither the time nor the state at another time instant than the initial time. This is a result of the infinite horizon together with the assumption that the system has a unique solution.

There are two different approaches to solve an optimal control problem. One approach is dynamic programming (DP). This approach can be used to find the optimal solution for all initial conditions in a set. Another approach is the Pontryagin Minimum Principle (PMP), which works for a single initial condition. The dynamic programming approach can be used in the latter case as well, by choosing the set as a single initial condition. According to Jönsson et al. (2002), the different approaches have some characteristics. These characteristics are presented below.

The dynamic programming approach can be summarized as follows:

+ It gives sufficient conditions for optimality.
+ The optimal control is obtained as a feedback \( u(t) = \mu(x(t)) \) for some function \( \mu \). Therefore, this approach is often called optimal feedback control.

- The optimal control is obtained by solving a possibly nonlinear partial differential equation, known as the Hamilton-Jacobi-Bellman equation (HJB) (or just the Bellman equation).
- It requires the performance criterion to be sufficiently smooth, normally \( C^1 \), which is not always the case.

The PMP approach also has some advantages and disadvantages:

+ It can be used in cases where the dynamic programming approach fails due to lack of smoothness of the optimal performance criterion.
+ It gives optimality conditions that in general are easier to verify than solving the partial differential equation obtained in the dynamic programming approach.

- It only gives necessary conditions for optimality. Hence, only candidates for optimality are obtained, which must be further investigated.

In this thesis, the objective is to find optimal feedback laws. Therefore, we will focus on the dynamic programming approach in the sequel. There are many references on dynamic programming. The classical book on this subject is Bellman (1957). However, this book treats the discrete time case. Early works on the continuous time case are (Kalman, 1963; Isaacs, 1965) and more recent works are (Bryson and Ho, 1975; Leitmann, 1981; Bertsekas, 1995; Jönsson et al., 2002) etc. An interesting paper about dynamic programming is also the historical overview by Pesch and Bulirsch (1994).

### 2.6.2 Necessary Condition Using the HJB

First a theorem is presented that yields necessary conditions. That is, given that there exists a sufficiently smooth \( V \) and a corresponding optimal feedback law, they must satisfy the HJB.
Theorem 2.9

Assume that there exists an optimal control \( u^*(\cdot) \) such that (2.57) is asymptotically stable and that the optimal value of the performance criterion \( V(x) \) is continuously differentiable. Then \( V(x) \) solves the Hamilton-Jacobi-Bellman equation

\[
0 = \min_u \left( L(x, u) + V_x(x)F(x, u) \right)
\]

(2.58)

and \( u^*(t) \) is the pointwise in time minimizing argument in (2.58).

Proof: Assume that there exists an optimal control \( u^*(\cdot) \) such that the closed loop system is asymptotically stable. The corresponding state trajectory is denoted \( x^* \). For an arbitrary initial condition \( x_0 \in \Omega_x \) and future time \( \Delta t \) we have

\[
V(x_0) = \min_{u(\cdot)} \left( \int_0^{\Delta t} L(x(s), u(s)) \, ds + V(x(\Delta t)) \right)
\]

\[
= \min_{u(\cdot)} \left( \int_0^{\Delta t} L(x(s), u(s)) \, ds + V(x_0 + F(x_0, u(0)) \Delta t + o(\Delta t) \right)
\]

\[
= \min_{u(\cdot)} \left( \int_0^{\Delta t} L(x(s), u(s)) \, ds + V(x_0) + V_x(x_0)F(x_0, u(0)) \Delta t + o(\Delta t) \right)
\]

where \( o(\Delta t) \) is the ordo-function and has the property that \( o(\Delta t)/\Delta t \to 0 \) as \( \Delta t \to 0 \). In the first step of the calculation above we have used an Euler approximation of the solution and in the second step that \( V \) is continuously differentiable and therefore has a Taylor expansion around \( x_0 \). Since \( V(x_0) \) is independent of \( u \), it can be subtracted from both sides. Division by \( \Delta t \) and using the fact that \( o(\Delta t)/\Delta t \to 0 \) as \( \Delta t \to 0 \) it follows that

\[
0 = \min_u \left( L(x_0, u) + V_x(x_0)F(x_0, u) \right)
\]

where we have used that \( 1/\Delta t \int_0^{\Delta t} f(x(s)) \, ds \to f(x(0)) \) as \( \Delta t \to 0 \) and the minimization is performed pointwise in time. However, since \( x_0 \) was arbitrary in \( \Omega_x \) the result above can also be formulated as

\[
0 = \min_u \left( L(x, u) + V_x(x)F(x, u) \right)
\]

for \( x \in \Omega_x \), and the Hamilton-Jacobi-Bellman equation is obtained. The pointwise optimized control input is given as \( u^*(t) = \mu(x^*(t)) \).

The proof shows how the minimization is transformed from a minimization of the complete control signal \( u(\cdot) \) to a minimization performed pointwise in time, where \( u \) is seen as a variable.

In the theorem above and also later in this section, \( V \) is required to be continuously differentiable. This is an assumption made in most references on optimal feedback control. The reason is that \( V_x \) is supposed to have the ordinary interpretation as the gradient of \( V \). However, even for some rather simple examples, \( V \) does not satisfy this condition. In many cases, this problem is possible to handle using viscosity solutions where \( V_x \) is interpreted as a subgradient, see Bardi and Capuzzo-Dolcetta (1997).
2.6.3 Sufficient Condition Using HJB

The next fundamental result is that the HJB (2.58) also yields sufficient conditions for optimality. Hence, if a continuously differentiable function $J$ is found and a corresponding feedback law $\mu(x)$, together satisfying the HJB, the solution to the optimal control problem (2.56) is found. More formally, this is formulated as a theorem.

**Theorem 2.10**

Suppose there exists a positive semidefinite, continuously differentiable function $J(x)$ satisfying $J(0) = 0$ and

$$0 = \min_u \left( L(x, u) + J_x(x) F(x, u) \right) \tag{2.59}$$

for $x \in \Omega_x$. Let

$$\mu(x) = \arg\min_u \left( L(x, u) + J_x(x) F(x, u) \right)$$

and assume that by using $u(t) = \mu(x(t))$, the closed loop system (2.57) becomes asymptotically stable on $\Omega_x$. Then

$$V(x) = J(x), \ x \in \Omega_x$$

and $\mu(x)$ is an optimal feedback control law.

**Proof:** Consider initial conditions $x_0 \in \Omega_x$. For all $u(\cdot)$ such that the closed loop system (2.57) is asymptotically stable on $\Omega_x$, it holds that $x_0 \in \Omega_x \Rightarrow x(t) \in \Omega_x$ and $x(t) \to 0$ as $t \to \infty$. Then, it follows by integration of (2.59) that

$$J(x_0) \leq \int_{0}^{T} L(x(t), u(t)) \, dt + J(x(T))$$

with equality for $u(t) = \mu(x(t))$. If we let $T \to \infty$ and use that $x(T) \to 0 \Rightarrow J(x(T)) \to 0$ (since the considered feedback laws are stabilizing) the result is

$$J(x_0) = \int_{0}^{\infty} L(x(t), \mu(x(t))) \, dt \leq \int_{0}^{\infty} L(x(t), u(t)) \, dt$$

which proves optimality.

The obtained feedback is the optimal feedback among the feedback laws keeping the state $x(t)$ in $\Omega_x$ and driving it towards the origin. Another common formulation of the theorem is to introduce $\bar{\Omega}_x \subset \Omega_x$ and let $\bar{\Omega}_x$ denote the initial states for which the trajectories belong to $\Omega_x$. Then the optimal solution only holds on $\bar{\Omega}_x$ instead.

Theorem 2.10 also defines an algorithm. The result is presented in Algorithm 2.6.3, where it is described how the optimal control law and the corresponding optimal performance criterion are computed.

The dynamic programming approach yields sufficient conditions, but often the minimization in the HJB is done using the first-order necessary conditions

$$0 = L_u(x, u) + V_x(x) F_u(x, u)$$
Algorithm 2.1 Hamilton-Jacobi-Bellman equation

1. Define the function \( \hat{\mu} \) by pointwise optimization over \( u \).

\[
\hat{\mu}(x, \lambda) = \text{argmin}_u \left( L(x, u) + \lambda^T F(x, u) \right), \quad x \in \Omega_x
\]

Here, \( \lambda \in \mathbb{R}^n \) is a parameter vector.

2. Solve the partial differential equation

\[
0 = L(x, \hat{\mu}(x, V_x(x))) + V_x(x)^T F(x, \hat{\mu}(x, V_x(x))) \tag{2.60}
\]

to obtain the optimal performance criterion \( V(x) \).

3. The optimal feedback law is obtained as \( \mu(x) = \hat{\mu}(x, V_x(x)) \).

for \( x \in \Omega_x \). Then, the sufficiency part is lost and it is necessary to show that the obtained feedback law and performance criterion are optimal. This verification can be done in several ways. One standard approach for proving optimality is to use the second order sufficiency condition. Optimality is then concluded if the second derivative with respect to \( u \) is positive definite, i.e.,

\[
L_{uu}(x, u) + V_x(x) F_{uu}(x, u) \succ 0
\]

for \( x \in \Omega_x \).

Control-affine Systems and Quadratic Cost Function

Often, the minimization in the HJB is nontrivial. As pointed out earlier, differentiation with respect to \( u \) is mostly used and the sufficiency property is then lost. This has the effect that further investigation of the different solutions is necessary to decide which of them that is optimal.

In this section a special case is considered, for which the optimal control problem can be solved using a slightly different approach preserving the sufficiency property. The system is assumed to be given in control-affine form

\[
\dot{x} = f(x) + g(x)u \tag{2.61}
\]

where \( f \) and \( g \) are smooth functions and the cost function is \( L(x, u) = \frac{1}{2} u^T u \). In this case, the optimal control problem can be solved by the following theorem (Scherpen, 1994).

**Theorem 2.11**

Assume that there exists a smooth function satisfying \( J(0) = 0 \) and

\[
0 = J_x(x) f(x) + \frac{1}{2} J_x(x) g(x) g(x)^T J_x(x)^T \tag{2.62}
\]

for all \( x \in W \) such that

\[
\dot{x} = -(f(x) + g(x) g(x)^T J_x(x)^T) \tag{2.63}
\]
is asymptotically stable on $\Omega_x \subset W$. Then, $V(x)$ is the solution to (2.62) such that (2.63) is asymptotically stable for all $x \in \Omega_x$.

**Proof:** Assume that (2.62) has a smooth solution $J(x)$ on $\Omega_x$. Then for all $u(\cdot)$ such that the solution to the closed loop system satisfies $x(t) \in \Omega_x$ for all $t \leq 0$ and $x(t) \to 0$ as $t \to -\infty$, we have

$$J(x(0)) = \int_{-\infty}^{0} \frac{d}{dt} J(x(t)) \, dt = \int_{-\infty}^{0} \frac{1}{2} J_x(x) g(x) g^T J_x(x)^T \, dt$$

$$= \frac{1}{2} \int_{-\infty}^{0} \|u\|^2 \, dt - \frac{1}{2} \int_{-\infty}^{0} \|u - g(x)^T J_x(x)^T\|^2 \, dt \leq \frac{1}{2} \int_{-\infty}^{0} \|u\|^2 \, dt$$

for all $x(0) \in \Omega_x$ and hence $J(x)$ is a lower bound for $\frac{1}{2} \int_{-\infty}^{0} \|u\|^2 \, dt$. It is also clear that this bound is obtained by using the feedback law $u(x) = g(x)^T J_x(x)^T$. By the asymptotic stability of (2.63) on $\Omega_x$, the latter feedback law is such that $x(t) \to 0$ as $t \to -\infty$ and the optimal performance criterion $V(x)$ is therefore found as

$$V(x) = J(x), \quad x \in \Omega_x$$

The proof is illustrative, because it shows how an optimal control problem for a state-space system can be solved if the system is affine in the control input and has a quadratic cost function. We will use similar reasoning in Section 5.3 to compute the controllability function for descriptor systems.

### 2.6.4 Example

A small, but yet illustrative, example is presented below. It shows how Algorithm 2.6.3 can be used and some different behaviors that may occur. Consider the system

$$\dot{x} = \alpha x + u$$

(2.64)

where $x$ and $\alpha$ both belong to $\mathbb{R}$. The objective is to find a stabilizing feedback law such that

$$J = \frac{1}{2} \int_{0}^{\infty} \beta x^4 + u^2 \, dt$$

(2.65)

is minimized, where $\beta \in \mathbb{R}$. We are interested in a global feedback law and therefore $\Omega_x = \mathbb{R}$. For notational convenience we define

$$H(x, u, \lambda) = \frac{1}{2} \beta x^4 + \frac{1}{2} u^2 + \lambda(\alpha x + u)$$

where $\lambda \in \mathbb{R}$. The HJB for the given problem can be formulated as

$$0 = \min_u H(x, u, \lambda)$$
The first order necessary condition for optimality, i.e., \( H_u(x, u, \lambda) = 0 \), yields the feedback law

\[
\hat{\mu}(x, \lambda) = -\lambda
\]

and the second order sufficient condition becomes \( H_{uu}(x, u, \lambda) = \frac{1}{2} > 0 \) for all \( x \). Hence, the obtained feedback law must be optimal if \( u = \mu(x) = \hat{\mu}(x, V_x(x)) \) makes (2.64) asymptotically stable. The partial differential equation in part 2 of Algorithm 2.6.3 becomes

\[
0 = V_x(x)^2 - 2\alpha x - \beta x^4
\]

which has the solutions

\[
V_x(x) = \alpha x \pm |x|\sqrt{\alpha^2 + \beta x^2}
\]

The optimal feedback law then has the form

\[
u = -\alpha x \mp |x|\sqrt{\alpha^2 + \beta x^2}
\]

and the corresponding closed loop system is

\[
\dot{x} = \mp |x|\sqrt{\alpha^2 + \beta x^2}
\]

Since an asymptotically stable closed loop system is desired, the optimal feedback law and \( V_x(x) \) can be written as

\[
u = -\left(\alpha + \sqrt{\alpha^2 + \beta x^2}\right)x, \quad V_x(x) = \left(\alpha + \sqrt{\alpha^2 + \beta x^2}\right)x
\]

and by integration of \( V_x(x) \), the optimal cost is obtained as

\[
V(x) = \frac{1}{2}\alpha x^2 + \frac{1}{3\beta}\left(\alpha^2 + \beta x^2\right)^{\frac{3}{2}} - \frac{1}{3\beta}\left(\alpha^2\right)^{\frac{3}{2}} \tag{2.67}
\]

since \( V(0) = 0 \). Using the parameters \( \alpha \) and \( \beta \), the behavior of the optimal control problem can be changed. To get a well-posed problem, \( \beta \geq 0 \) is assumed. Three different cases will be investigated:

1. \( \beta > 0, \alpha \) arbitrary:
   In this case the state is included in the cost function and (2.64) is either asymptotically stable, a pure integrator or unstable depending of the choice of \( \alpha \). For \( \beta > 0 \) it follows that

   \[
   \alpha + \sqrt{\alpha^2 + \beta x^2} > 0
   \]

   for all \( x \neq 0 \), but if \( x = 0 \) the system is at the equilibrium. If (2.67) is studied, it can be seen that a small \( \beta \) yields a small cost and vice versa. Furthermore, it can be realized that if \( \alpha < 0 \), i.e., the undriven system is asymptotically stable, a smaller cost is obtained than if \( \alpha > 0 \). These observations coincide with the intuition, since in first case the undriven system helps the feedback law to reach the origin.

2. \( \beta = 0, \alpha > 0 \):
   In this case, the state is not included in the performance criterion and the system is unstable. The expressions in (2.66) will in this case be

   \[
u = -\left(\alpha + \sqrt{\alpha^2}\right)x = -2\alpha x, \quad V_x(x) = \left(\alpha + \sqrt{\alpha^2}\right)x = 2\alpha x
\]
and the cost function is

\[ V(x) = \alpha x^2 \]

If \( \alpha \) is large, i.e., if the divergence is fast, a larger performance criterion is obtained which corresponds to the intuition. Note the comments in the beginning of this section about \( x \) being present in \( L \). It is possible to choose \( V_x(x) = 0 \) and satisfy the HJB. However, the closed loop system is then unstable, and it is necessary to explicitly choose the stabilizing control law.

3. \( \beta = 0, \alpha < 0 \):

In this case, the state is not included in the cost function and the system is asymptotically stable. The expressions corresponding to (2.66) becomes

\[ u = -\left( \alpha + \sqrt{\alpha^2} \right) x = 0, \quad V_x(x) = \left( \alpha + \sqrt{\alpha^2} \right) x = 0 \]

and a optimal performance criterion \( V(x) = 0 \). This is natural, since it does not cost anything to have a nonzero state and the state goes towards the origin without using the control signal. Of course, it is cheapest not to use the control signal in this case.
In principle, the problem of finding optimal feedback laws for descriptor systems can be solved by the theory for state-space systems. The reason is that given certain regularity conditions we know from Section 2.4 that a rather general class of descriptor system can be reduced to either a reduced size state-space system and some algebraic equations or to a state-space system together with conditions on the initial conditions. In both cases, the obtained system can be handled by the theory presented in Section 2.6.

A problem is that it is often not possible to compute the underlying state-space system explicitly because no explicit inverse exists. Therefore, it is interesting to find methods to compute the optimal feedback law either based on a general descriptor system or at least based on an index reduced description.

As far as we know, no approach exists such that a general descriptor system (2.2) can be treated immediately without index reduction. However, for linear time-invariant descriptor systems (2.5) such methods exist, see for example (Bender and Laub, 1987; Mehrmann, 1989). These methods are based on variational calculus which leads to a generalized eigenvalue problem to be solved, see Jonckheere (1988). For nonlinear descriptor systems in semi-explicit form (2.4), there are also results based on the dynamic programming approach. In Xu and Mizukami (1993) an equation similar to the HJB is derived. We will in this chapter present this HJB-like equation and investigate how its solution is related to the solution of the corresponding HJB for a reduced state-space model describing the same system.

The chapter is organized as follows. The optimal feedback control problem is formulated in Section 3.1. In Section 3.2, the descriptor system is rewritten as a state-space system and the optimal control problem is solved using the ordinary HJB equation. In Section 3.3, the HJB-like equation is used instead to find the optimal solution and Section 3.4 shows how the different methods relate to each other. A special case, for which the first order conditions for optimality become simple, is described in Section 3.5. Finally, a small example is presented in Section 3.6.
3.1 Optimal Feedback Control

Consider a semi-explicit system description

\[ \dot{x}_1 = F_1(x_1, x_2, u) \]  
\[ 0 = F_2(x_1, x_2, u) \]

which is assumed to satisfy Assumption A2 for \( x_1 \in \Omega_x \) and the corresponding set for \( (x_1, x_2, u) \) is denoted \( \Omega \). The system description (3.1) can be the result of the index reduction method in Section 2.4.3 in a case when Assumption A3 is satisfied.

The considered class of performance criterions is

\[ J = \int_{0}^{\infty} L(x_1, x_2, u) \, dt \]

with an infinite time horizon and with a cost function \( L \) that is positive semidefinite and positive definite in \( u \). The optimal control problem can then be formulated as

\[ V(x_1(0)) = \min_{u(\cdot)} J \]

subject to the dynamics (3.1) and the boundary conditions

\[ x_1(0) = x_{1,0} \in \Omega_x \]
\[ \lim_{t \to \infty} x_1(t) = 0 \]

The minimization is done with respect to all \( u(\cdot) \) such that two conditions are satisfied. One condition is that the obtained closed loop system

\[ \dot{x}_1 = F_1(x_1, x_2, u(\cdot)) \]
\[ 0 = F_2(x_1, x_2, u(\cdot)) \]

is asymptotically stable, see Section 2.5. Another condition, specific for the descriptor system case, is that (3.4) is required to have index one. This condition is added since different feedback laws \( u = k(x_1, x_2) \) may yield different indices of (3.4). The effect would be that the size of dynamical and algebraical part would change and that some of the variables \( x_1 \) could be algebraically determined by the other variables in \( x_1 \) and \( x_2 \). It would also make the problem harder to analyze, since implicit constraints could occur. However, in the case when the system is given in semi-explicit form satisfying Assumption A2 it will be shown that the index is automatically preserved for the optimal control input \( u(\cdot) \). Therefore since the closed loop system (3.4) has index one, it is known that \( x_1 \) defines the dynamical part and \( x_2 \) is algebraically connected to \( x_1 \). The initial conditions are then given for \( x_1 \) while the initial conditions for \( x_2 \) are assumed to be chosen consistently, i.e., such that \( F_2(x_{1,0}, x_{2,0}, u(x_{1,0})) = 0 \). Assumption A2 together with the fact that only consistent initial conditions are considered also yields that the closed loop system will have a unique solution. This is an important fact when it comes to \( V \) being a function of the initial conditions only.
In some articles about optimal control for descriptor systems, e.g., (Cobb, 1983; Jonckheere, 1988; Xu and Mizukami, 1993), the possibility of changing the index using feedback is utilized. They require the optimal feedback law to be such that the closed loop system \( (3.4) \) has index one even if the open loop system \( (3.1) \) with the control input \( u(t) \) has a higher index.

Notice that the requirement \( L(x_1, x_2, u) \) being positive semidefinite and positive definite in \( u \) is somewhat restrictive in our case. However, it is done in order match the assumptions in Xu and Mizukami (1993). From Assumption A2 we have that \( x_2 = \varphi(x_1, u) \) and the true requirement then is \( L(x_1, \varphi(x_1, u), u) \) being positive semidefinite and positive definite in \( u \).

### 3.2 Hamilton-Jacobi-Bellman Equation for the Reduced Problem

Assumption A2 makes it possible to solve the optimal feedback control problem \( (3.3) \) as an optimal feedback control problem for a state-space system \( (2.56) \). For \( x_1 \in \Omega_x \) and \( u \in \mathbb{R}^p \), \( (3.1a) \) and \( (3.2) \) can be rewritten as the performance criterion

\[
J = \int_0^\infty L(x_1, \varphi(x_1, u), u) \, dt
\]

and the state-space model

\[
\dot{x}_1 = F_1(x_1, \varphi(x_1, u), u)
\]

respectively. The cost function \( L \) in \( J \) is positive semidefinite and positive definite in \( u \). From Theorem 2.10 it follows that the optimal control problem is solved by finding a positive definite, continuously differentiable \( V(x_1) \) satisfying the HJB

\[
0 = \min_u \left( L(x_1, \varphi(x_1, u), u) + V_{x_1}(x_1) F_1(x_1, \varphi(x_1, u), u) \right), \quad x_1 \in \Omega_x
\]

This \( V(x_1) \) is then the optimal performance criterion in \( (3.3) \). Note that if \( (3.5) \) is only possible to solve on a set smaller than the set on which the descriptor system can be rewritten as a state-space system, it is assumed that \( \Omega_x \) is redefined as the smaller set. Furthermore, remember that \( V(x_1) \) is only proven to be optimal on some set \( \Omega'_x \subset \Omega_x \) for which \( x_{1,0} \in \Omega'_x \) is such that the obtained feedback law \( u = \mu(x_1) \) gives an asymptotically stable closed loop system and keeps \( x_1(t) \) in \( \Omega_x \) for \( t \geq 0 \), unless \( \Omega_x \) is chosen to be an invariant set, see comments in Section 2.6.3.

The first-order necessary condition for optimality of \( (3.5) \) yields the set of equations

\[
\begin{align*}
0 &= L_{u} + V_{x_1} F_{1;u} + (L_{x_2} + V_{x_1} F_{1;x_2}) \varphi_{u} \\
0 &= L + V_{x_1} F_1
\end{align*}
\]

where the quantities in the right hand sides are evaluated at \((x_1, \varphi(x_1, u), u)\). Using that

\[
F_2(x_1, \varphi(x_1, u), u) = 0
\]
identically in $u$, differentiation with respect to $u$ gives

$$F_{2,x_2}(x_1, \varphi(x_1, u), u) \varphi_u(x_1, u) + F_{2,u}(x_1, \varphi(x_1, u), u) = 0$$

which can be solved for $\varphi_u(x_1, u)$ as

$$\varphi_u(x_1, u) = -F_{2,x_2}(x_1, \varphi(x_1, u), u)^{-1}F_{2,u}(x_1, \varphi(x_1, u), u)$$

Since $x_2 = \varphi(x_1, u)$ is the unique solution of (3.1b), it is also possible to write these equations according to

$$0 = L_u + V_{x_1} F_{1;u} - (L_{x_2} + V_{x_1} F_{1;x_2}) F_{2,x_2}^{-1} F_{2,u}$$

(3.7a)

$$0 = L + V_{x_1} F_1$$

(3.7b)

$$0 = F_2$$

(3.7c)

where the right hand sides are evaluated at $x_1, x_2, u$. One way of looking at (3.7) is to regard (3.7a) and (3.7c) as $p + n_2$ equations from which one tries to solve for $u$ and $x_2$ as functions of $x_1$ and $V_{x_1}$. When these quantities are substituted into (3.7b) the result is a first order partial differential equation for $V$ as a function of $x_1$. When this partial differential equation is solved the result can be substituted back into the expression for $u$ to give the optimal feedback law.

To ensure that the optimal solution is found, the second order sufficient condition mentioned in Section 2.6.3 can be used. It will require $\varphi_{uu}$ to be computed which is possible but the expressions become rather involved.

### 3.3 Hamilton-Jacobi-Bellman-Like Equation

In Xu and Mizukami (1993) the optimal control problem (3.3) is solved in a different manner. According to their Theorem 3.1 it is possible to find the optimal solution by solving the Hamilton-Jacobi-Bellman-like equation

$$0 = \min_u (L(x_1, x_2, u) + W_1(x_1) F_1(x_1, x_2, u) + W_2(x_1, x_2) F_2(x_1, x_2, u))$$

(3.8)

for some continuous functions $W_1(x_1)$ and $W_2(x_1, x_2)$ such that $W_1(x_1)$ is a gradient of some continuously differentiable function $V(x_1)$. This $V(x_1)$ is then the optimal cost in (3.3).

Using the first-order necessary condition for optimality, the control is defined by the following set of equations

$$0 = L_u(x_1, x_2, u) + W_1(x_1) F_{1;u}(x_1, x_2, u) + W_2(x_1, x_2) F_{2;u}(x_1, x_2, u)$$

(3.9a)

$$0 = L(x_1, x_2, u) + W_1(x_1) F_{1;u}(x_1, x_2, u) + W_2(x_1, x_2) F_{2}(x_1, x_2, u)$$

(3.9b)

where $x_2$ is considered to be independent of $u$ when differentiating with respect to $u$. From these equations it is not immediately obvious how to obtain a relation from which $W_1$ can be computed. Similar equations to (3.7) can be obtained by restricting (3.9) to
points satisfying \( F_2 = 0 \). Then the result is the following system

\[
0 = L_u(x_1, x_2, u) + W_1(x_1)F_{1;u}(x_1, x_2, u) + W_2(x_1, x_2)F_{2;u}(x_1, x_2, u) \tag{3.10a}
\]
\[
0 = L(x_1, x_2, u) + W_1(x_1)F_1(x_1, x_2, u) \tag{3.10b}
\]
\[
0 = F_2(x_1, x_2, u) \tag{3.10c}
\]

If \( W_2 \) is considered unknown (3.10) is still underdetermined. Hence more equations are needed or \( W_2 \) is to be considered as given. However, Xu and Mizukami (1993) only gives sufficient conditions, i.e., if \( aW_1 \) and \( W_2 \) can be found such that (3.8) is satisfied, the optimal solution is found, and they do not mention what kind of conditions that \( W_2 \) have to satisfy. This will be investigated in the next section.

### 3.4 Relationships Among the Solutions

The reduced Hamilton-Jacobi equation (3.5) and the Hamilton-Jacobi-like equation (3.8) solve the same underlying optimal control problem. Therefore it is natural that the functions \( V, W_1 \) and \( W_2 \) are related and below these relationships are investigated.

#### Lemma 3.1

Suppose there exist a function \( V(x_1) \) and a feedback \( u = k(x_1) \) solving (3.5) on \( \Omega_x \). Then \( W_1(x_1) = V_{x_1}(x_1), u = k(x_1) \) solve (3.8) under the constraint \( F_2(x_1, x_2, u) = 0 \). Moreover, with the choice

\[
W_1(x_1) = V_{x_1}(x_1), \quad W_2 = -\left(L_{x_2} + V_{x_1}F_{1;x_2}\right)F_{2;x_2}^{-1} \tag{3.11}
\]

the necessary conditions for optimality (3.10) are satisfied for \( u = k(x_1) \) together with \( x_2 = \varphi(x_1, k(x_1)) \).

**Proof:** When \( F_2 = 0 \) the right hand sides of (3.8) and (3.5) coincide. Comparing (3.7) and (3.10) shows that (3.10) is satisfied for \( u = k(x_1), x_2 = \varphi(x_1, k(x_1)) \). \( \square \)

The converse relation is given by the following lemma.

#### Lemma 3.2

Assume that for \( x_1 \in \Omega_x \)\(^1\) it holds that:

- (3.8) has a solution \( u = \psi(x_1, x_2) \)
- \( F_2(x_1, x_2, \psi(x_1, x_2)) = 0 \) has a solution \( x_2 = \eta(x_1) \)
- \( W_1(x_1) = V_{x_1}(x_1) \) for some function \( V(x_1) \)

Then \( V_{x_1}(x_1) \) and \( u = k(x_1) = \psi(x_1, \eta(x_1)) \) solve (3.5) for \( x_1 \in \Omega_x \). Moreover, for \( (x_1, x_2, u) \in \Omega \) satisfying (3.10) it follows that

\[
W_2 = -\left(L_{x_2} + W_1F_{1;x_2}\right)F_{2;x_2}^{-1} \tag{3.12}
\]

\(^1\)If this \( \Omega_x \) is smaller than the \( \Omega_x \) on which the system can be written as a state-space system, \( \Omega_x \) is redefined as the smaller region.
Proof: We have

\[ L(x_1, \eta(x_1), \psi(x_1, \eta(x_1))) + V_{x_1}(x_1)F_1(x_1, \eta(x_1), \psi(x_1, \eta(x_1))) \]
\[ = L(x_1, \varphi(x_1, k(x_1)), k(x_1)) + V_{x_1}(x_1)F_1(x_1, \varphi(x_1, k(x_1)), k(x_1)) = 0 \]

since the minimal value in (3.8) is attained for \( u = \psi(x_1, x_2) \) for all \( x_1 \in \Omega_x \) and \( x_2 \in \mathbb{R}^n \), and then particularly for \( x_2 = \eta(x_1) \). Since \( x_1 \in \Omega_x \) it is also known that \( \eta(x_1) = \varphi(x_1, k(x_1)) \). According to (3.8) we have

\[ 0 \leq L(x_1, x_2, u) + V_{x_1}(x_1)F_1(x_1, x_2, u) + W_2(x_1, x_2)F_2(x_1, x_2, u) \]

for all \( x_1 \in \Omega_x, x_2 \in \mathbb{R}^n \) and \( u \in \mathbb{R}^p \). In particular we have

\[ 0 \leq L(x_1, \varphi(x_1, u), u) + V_{x_1}(x_1)F_1(x_1, \varphi(x_1, u), u) \]

and (3.5) is thus satisfied.

Since a \( u \) solving (3.10a) is given by \( u = \psi(x_1, x_2) \), (3.10b) and (3.10c) give

\[ 0 = L(x_1, x_2, \psi(x_1, x_2)) + W_1(x_1)F(x_1, x_2, \psi(x_1, x_2)) \]
\[ 0 = F_2(x_1, x_2, \psi(x_1, x_2)) \]

Differentiation of these relations with respect to \( x_2 \) yields

\[ 0 = L_{x_2} + (L_u + W_1F_{1;u})\psi_{x_2} + W_1F_{1;x_2} \]  \hspace{1cm} (3.13a)
\[ 0 = F_{2;x_2} + F_{2;u}\psi_{x_2} \] \hspace{1cm} (3.13b)

If (3.10a) is multiplied from right by \( \psi_{x_2} \) and after (3.13) is inserted, the result is that \( W_2 \) is given by

\[ 0 = W_1F_{1;x_2} + L_{x_2} + W_2F_{2;x_2} \]

Due to the fact that \( F_{2;x_2} \) is nonsingular for \( (x_1, x_2, u) \in \Omega \), it follows that

\[ W_2 = -\left( L_{x_2} + W_1F_{1;x_2} \right)F_{2;x_2}^{-1} \] \hspace{1cm} (3.14)

Hence, for a system that satisfies Assumption A2 one further necessary condition for the optimal solution, namely (3.14), is obtained.

### 3.5 Control-Affine Systems

In this section, a special class of problems is considered for which the necessary conditions becomes simple. The system description in the class should be possible to write as

\[ \dot{x}_1 = f_1(x_1, x_2) + g_1(x_1)u \] \hspace{1cm} (3.15a)
\[ 0 = f_2(x_1, x_2) + g_2(x_1)u \] \hspace{1cm} (3.15b)
while the cost function should be expressed in the form \( L(x_1, x_2, u) = l(x_1) + \frac{1}{2} u^T u \). Then (3.10a) can be solved explicitly in \( u \) for all \( x_1, x_2 \) since (3.10a) will become
\[
0 = u^T - W_1(x_1)g_1(x_1) - W_2(x_1, x_2)g_2(x_1)
\]
(3.16)
and from Lemma 3.2, we have
\[
W_2(x_1, x_2) = -W_1(x_1)f_{1;x_2}(x_1, x_2)f_{2;x_2}^{-1}(x_1, x_2)
\]
(3.17)
Note that \( f_{2;x_2}(x_1, x_2) \) is nonsingular for all \( (x_1, x_2) \) such that \( f_{2}(x_1, x_2) = 0 \) is solvable since \( F_{2;x_2}(x_1, x_2, u) \) is nonsingular for all \( (x_1, x_2, u) \in \Omega \) and then particularly for \( u = 0 \). Combining (3.16) and (3.17) yields
\[
u = -\hat{g}(x_1, x_2)^T W_1(x_1)^T
\]
and after some manipulation the necessary conditions can be rewritten as
\[
0 = l(x_1) + W_1(x_1)\hat{f}(x_1, x_2) - \frac{1}{2} W_1(x_1)\hat{g}(x_1, x_2)\hat{g}(x_1, x_2)^T W_1(x_1)^T
\]
(3.18a)
\[
0 = f_2(x_1, x_2) + g_2(x_1)\hat{g}(x_1, x_2)^T W_1(x_1)^T
\]
(3.18b)
where
\[
\hat{f}(x_1, x_2) = f_1(x_1, x_2) - f_{1;x_2}(x_1, x_2)f_{2;x_2}^{-1}(x_1, x_2)f_2(x_1, x_2)
\]
\[
\hat{g}(x_1, x_2) = g_1(x_1) - f_{1;x_2}(x_1, x_2)f_{2;x_2}^{-1}(x_1, x_2)g_2(x_1)
\]

### 3.6 Example

In this section a small example showing the different methods is presented.

Consider the simple system
\[
\begin{align*}
\dot{x}_1 &= x_2 \\
0 &= u - x_1^3 - x_2
\end{align*}
\]
which satisfies Assumption A2 with performance criterion
\[
J = \int_0^\infty \left( \frac{x_1^2}{2} + \frac{u^2}{2} \right) dt
\]
The necessary conditions (3.7) give
\[
\begin{align*}
0 &= u + V_{x_1} \cdot 0 + V_{x_1} \cdot 1 \\
0 &= \frac{x_1^2}{2} + \frac{u^2}{2} + V_{x_1}x_2 \\
0 &= u - x_1^3 - x_2
\end{align*}
\]
Eliminating \( u \) and \( x_2 \) gives the following equation for \( V_{x_1} \)
\[
V_{x_1}^2 + 2x_1^3V_{x_1} - x_1^2 = 0
\]
To get a positive definite solution for $V$ the solution for $V_{x_1}$ must be chosen as

$$V_{x_1} = x_1(\sqrt{1 + x_1^4} - x_1^2)$$

and the corresponding optimal feedback law then becomes

$$u = -V_{x_1} = -x_1(\sqrt{1 + x_1^4} - x_1^2)$$

If the problem is instead solved using (3.9) we get the equations

$$0 = u + W_1 \cdot 0 + W_2 \cdot 1$$
$$0 = \frac{x_2^2}{2} + \frac{u^2}{2} + W_1 x_2 + W_2(u - x_1^3 - x_2)$$

Since the system satisfies Assumption A2, it is known from Lemma 3.2 that $W_2$ must satisfy (3.14), i.e.,

$$W_2 = W_1$$

When the equation $0 = u - x_1^3 - x_2$ is satisfied this gives the same equation for $W_1$ as for $V_{x_1}$ so the solution is the same.
In Chapter 3 we discussed optimal control for descriptor systems on semi-explicit form, which possibly could come from higher index problems. Unfortunately, there are practical issues with the methods presented there. One issue is that in order to explicitly and not numerically, solve the optimality conditions an explicit expression for the implicit function $\varphi$ is needed. However, in many cases the implicit function is not possible to express on closed form. Another problem with the methods in Chapter 3 is that even if $\varphi$ can be expressed explicitly, it is only for a small class of problems the optimality conditions can be solved to obtain an explicit solution. Therefore, another approach to solve the optimal control problem will be presented in this chapter.

The idea is to instead compute the optimal solution, i.e., the optimal performance criterion and the corresponding feedback law, expressed as power series expansions. For state-space models this idea was first considered by Al’brekht (1961). He shows that the terms in the power series expansions can be sequentially obtained through solution of a quadratic optimal control problem for the linearized system and subsequent solution of a series of linear partial differential equations. Further, a formal proof of the convergence of the power series is presented in the case when the input signal is scalar and the system has the form $\dot{x} = f(x) + Bu$. In Lee and Markus (1967) these results are extended to general state-space systems, $\dot{x} = f(x, u)$, and this work is extended even more in Lukes (1969).

In the earlier works (Al’brekht, 1961; Lee and Markus, 1967), the functions involved are required to be analytic functions around the origin. In Lukes (1969), this requirement is relaxed to twice differentiability. An alternative proof to the one presented in Lukes (1969) is given in van der Schaft (1991) where the requirements on the cost function are relaxed. Krener (1992) studied the case when the dynamics of an external signal generator are included and in Krener (2001) he also investigated the case when the system is not stabilizable or not detectable. The latter reference also considers the Hamilton-Jacobi inequality.

Other classes of problems also studied are for example the case with bilinear system
dynamics and quadratic cost which can be found in Cebuhar and Costanza (1984) and the extension to discrete time systems described in Navasca (1996). The case of finite time optimal control is found in Willemstein (1977) and Yoshida and Loparo (1989) use the Carleman linearization to study both finite and infinite time optimal control problems.

A possible problem with the methods based on power series expansion is that validity of the optimal solution can only be guaranteed locally. Therefore, Navasca presents a method that uses power series solutions around extremals to enlarge the region where the solution is optimal, see Navasca and Krener (2000).

In practice, the series solution needs to be truncated and the result is an approximative solution. Therefore, these kind of methods are often denoted approximative methods even though the complete power series expansions of the performance criterion and feedback law yield the true optimal solution. There are other methods which theoretically describe the exact optimal solution but in practice are truncated, see Beard et al. (1998) and references therein.

This chapter is organized as follows. In Section 4.1, the problem formulation is presented. Section 4.2 shows how a locally optimal solution is computed when the system is described in state-space form. The extension to nonlinear semi-explicit descriptor systems is described in Section 4.3, while Section 4.4 shows how the assumption concerning semi-explicitness can be relaxed. Finally, a small example is presented in Section 4.5.

### 4.1 Problem Formulation

Consider again the optimal control problem in Chapter 3, i.e.,

\begin{align}
\dot{x}_1 &= F_1(x_1, x_2, u) \\
0 &= F_2(x_1, x_2, u)
\end{align}  

(4.1a)  

(4.1b)

and the performance criterion

\[ J = \int_0^\infty L(x_1, x_2, u) \, dt \]

The objective in this chapter is to find the optimal control input locally around the origin, \((x, u) = (0, 0)\), and we consider the class of feedback laws that can be expressed as real analytic power series expansions

\[ u(x) = Dx + u_h(x) \]

(4.3)
where \( u_h(x) \) consists of terms of at least degree two.

Mostly in this thesis it is assumed that the system satisfies Assumption A2, but in this case a more convenient assumption is as follows.

**Assumption A4.** It holds that \( F_2(0, 0, 0) = 0 \) and that \( F_{2,x_2}(0, 0, 0) \) is nonsingular.

Assumption A4 together with Theorem A.1 yield existence of a neighborhood \( \Omega_{x_1,u} \) of the origin such that for \((x_1, u) \in \Omega_{x_1,u}\)

\[
x_2 = \varphi(x_1, u)
\]

Furthermore, we assume that the system has an equilibrium at the origin, which also yields that \( \varphi(0, 0) = 0 \). We also make an assumption regarding the system functions.

**Assumption A5.** The functions \( F(x, u) \) and \( L(x, u) \) are real analytic functions in some open set

\[
\mathcal{W} = \{(x_1, u) \in \mathcal{W}_{x_1,u}, \quad x_2 \in \mathcal{W}_{x_2} \supset \varphi(\mathcal{W}_{x_1,u})\}
\] (4.4)

where \( \mathcal{W}_{x_1,u} \) and \( \mathcal{W}_{x_2} \) are neighborhoods of \((x_1, u) = 0 \) and \( x_2 = 0 \), respectively.

Note that since \( \mathcal{W}_{x_2} \) is a neighborhood it is open, which not \( \varphi(\mathcal{W}_{x_1,u}) \) necessarily is despite the fact that \( \mathcal{W}_{x_1,u} \) is a neighborhood and therefore is open.

The analyticity of \( F \) and \( L \) yields that they can be expanded in convergent power series

\[
F(x, u) = Ax + Bu + F_h(x, u)
\] (4.5a)

\[
L(x, u) = x^TQx + 2x^TSu + u^TRu + L_h(x, u)
\] (4.5b)

where the matrices \( A, B, Q \) and \( S \) are partitioned as

\[
A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}
\]

and \( F_h(x, u) \) and \( L_h(x, u) \) contain higher order terms of at least degree two and three, respectively. The matrix \( A_{22} \) will be nonsingular since it was assumed in Assumption A4 that \( F_{2,x_2}(0, 0, 0) = A_{22} \) is nonsingular.

In the computations it will be assumed that the analysis is done in a neighborhood where both Assumption A4 and Assumption A5 are satisfied. To lessen the notational complexity, \( \Omega_{x_1,u} \) is assumed to be covered by \( \mathcal{W}_{x_1,u} \), i.e., \( \Omega_{x_1,u} \subset \mathcal{W}_{x_1,u} \).

As mentioned in Section 3.1 the optimization is done with respect to control inputs such that the closed loop system is asymptotically stable and has index one locally around the origin. Under the assumption that the initial conditions are chosen close enough to the origin, it is known from Theorem 2.5, i.e., Lyapunov’s indirect method, that a sufficient condition is asymptotic stability of the linearization of the closed loop system.

### 4.2 Review of the State-Space System Case

First a short review is presented describing how to obtain a power series solution to an optimal control problem in the case when the system is described by a state-space model.
In the state-space case it is known from Section 2.6 that the optimal control problem mentioned in Section 4.1 is associated with the HJB equation

\[ 0 = \min_{u(x)} L(x, u) + V_x(x) F(x, u) \]  

(4.6)

which corresponds to first-order necessary conditions for optimality

\[ 0 = L(x, u_*(x)) + V_x(x) F(x, u_*(x)) \]
\[ 0 = L_u(x, u_*(x)) + V_x(x) F_{u}(x, u_*(x)) \]  

(4.7)

for the optimal performance criterion \( V(x) \) and the optimal feedback law \( u_*(x) \).

The idea is that if \( L(x, u), F(x, u) \) and \( u(x) \) are real analytical near the origin and if \( u(x) \) is chosen such that it stabilizes the closed loop system locally around \( (x, u) = 0 \), the performance criterion \( V(x) \) will be analytical in a neighborhood of the origin, see Lukes (1969). This guarantees that \( V(x) \) can be expressed as a convergent power series

\[ V(x) = x^T P x + V_h(x) \]  

(4.8)

where \( V_h(x) \) contains the terms of order three and higher. Lukes (1969) also shows that \( P \) is a positive definite matrix if only cost matrices satisfying

\[ \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \succ 0 \]  

(4.9)

is considered. Positive definiteness of \( P \) will be important when determining which of the solutions to the necessary conditions (4.7) is optimal.

If (4.3), (4.5) and (4.8) are inserted into (4.7), the result is two polynomial equations in \( x \). These polynomial equations are supposed to hold for all \( x \) in a neighborhood of the origin and different orders of \( x \) will then yield separate equations in the corresponding coefficients. According to Lukes (1969), the lowest order terms of the polynomial equations yield the equations

\[ P \bar{A} + \bar{A}^T P - PBR^{-1}B^T P + Q - SR^{-1}S^T = 0 \]
\[ D_u + R^{-1}(S^T + B^T P) = 0 \]  

(4.10a)

(4.10b)

where \( \bar{A} = A - BR^{-1}S^T \), while the higher order terms in \( x \) will be

\[ V_{x}^{[m]}(x) A_c x = - \sum_{k=3}^{m-1} V_{x}^{[k]}(x) B u_*^{[m-k+1]}(x) \]
\[ - \sum_{k=2}^{m-1} V_{x}^{[k]}(x) F_{h}^{[m-k+1]}(x, u_*) - L_{h}^{[m]}(x, u_*) \]
\[ - 2 \sum_{k=2}^{m/2} u_*^{[k]}(x) R u_*^{[m-k]}(x) - u_*^{[m/2]}(x) R u_*^{[m/2]}(x) \]  

(4.11a)

where \( m = 3, 4, \ldots \) and \( A_c = A + BD_u \), and

\[ u_*^{[k]}(x) = - \frac{1}{2} R^{-1} \left( V_{x}^{[k+1]}(x) B + \sum_{i=1}^{k-1} V_{x}^{[k-i+1]}(x) F_{h;i}^{[i]}(x, u_*) + L_{h;i}^{[i]}(x, u_*) \right) \]  

(4.11b)
for $k = 2, 3, \ldots$. In the equations above, $f^{[i]}$ denotes the $i$th order terms of $f$ and $\lfloor i \rfloor$ denotes the floor function, which gives the largest integer less than or equal to $i$. Moreover, in (4.11) we use the conventions that $\sum_{k}^l = 0$ for $l < k$ and that the terms $u_{l/m/2}$ are to be omitted if $m$ is odd.

The equations (4.10) and (4.11) are only necessary conditions for optimality and if several solutions exist it is necessary to decide which is optimal. The second order sufficient condition, i.e., $L_{uu}(x, u) + V(x)F_{uu}(x, u)$, evaluated in $(x, u) = 0$ is positive definite because $R \succ 0$. Therefore, it is known that locally in some neighborhood of $x$, the necessary conditions also yield the optimal solution. However, the optimal solution must also satisfy the requirements in the problem formulation. That is, $u^*_s(x)$ must stabilize the closed loop system locally around the origin and therefore $A_c = A + BD_s$ must be Hurwitz. If only cost matrices satisfying (4.9) are considered we also know that $P$ must be positive definite.

The matrices $P$ and $D_s$ are given by (4.10), where (4.10a) is an algebraic Riccati equation (ARE). Hence, the ARE must have a positive definite solution $P$ such that $D_s$ from (4.10b) is stabilizing. A criteria for when such a solution exists is given by the following lemma.

**Lemma 4.1**

Consider the ARE (4.10a). Assume that the cost matrix satisfies $\begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \succ 0$. Then the following statements are equivalent:

1. $(A, B)$ is stabilizable.

2. The ARE (4.10a) has a unique positive definite stabilizing solution $P$, i.e., a solution such that $A + BD_s$ is Hurwitz, where $D_s$ is given by (4.10b).

**Proof:** See Lukes (1968).

Note that under the assumptions given in Lemma 4.1, the positive definite solution $P$ is the only possible semidefinite solution as well (Bittanti et al., 1991).

The first order terms of $u_s(x)$ and $V(x)$ are then unique and given by the expressions

$$u_s^{[1]}(x) = D_s x, \quad V^{[2]}(x) = x^T P x$$

(4.12)

In order to obtain higher order approximations of the optimal feedback law $u_s(x)$ and the corresponding performance criterion $V(x)$, we need to solve (4.11). First observe that

$$F_h^{[k]}(x, u) = F_h^{[k]}(x, u_s^{[1]} + u_s^{[2]} + \ldots + u_s^{[k-1]})$$

$$L_h^{[k]}(x, u) = L_h^{[k]}(x, u_s^{[1]} + u_s^{[2]} + \ldots + u_s^{[k-2]})$$

since $F_h(x, u)$ and $L_h(x, u)$ are power series beginning with terms of order two and three, respectively. Based on this it can be seen that the right-hand side of (4.11a) depends only on the terms

$$u_s^{[1]}, \ldots, u_s^{[m-2]}, V^{[2]}, \ldots, V^{[m-1]}$$

(4.13)

while the right-hand side of (4.11b) depends only on

$$u_s^{[1]}, \ldots, u_s^{[k-1]}, V^{[2]}, \ldots, V^{[k+1]}$$

(4.14)
Since $A_c = A + BD_*$ is Hurwitz, it is shown in for example Lyapunov (1992), that the partial differential equation (4.11a) is uniquely solvable for a given right-hand side. Therefore, by starting with $u_*^{[1]}(x) = D_+x$ and $V^{[2]}(x) = x^TPx$ it is possible to recursively compute the terms
\[ V^{[3]}(x), \ u_*^{[2]}(x), \ V^{[4]}(x), \ u_*^{[3]}(x), \ldots \]
and thereby generating power series for $u_*(x)$ and $V(x)$.

The method shows how arbitrarily high orders of $V(x)$ and $u_*(x)$ can be computed. However, in practice an important fact is that the computational complexity is exponential in the number of variables and the desired approximation order.

### 4.3 Descriptor System Case

To solve the optimal control problem (4.2) we use the same technique as in Section 3.2, i.e., reduce the descriptor system to a state-space system. For $(x_1, u) \in \Omega_{x_1,u}$, the dynamical part is given by
\[
\dot{x}_1 = \hat{F}_1(x_1, u) = F_1(x_1, \varphi(x_1, u), u) \tag{4.15a}
\]
and the cost function reduces to
\[
\hat{L}(x_1, u) = L(x_1, \varphi(x_1, u), u) \tag{4.15b}
\]

The optimal control problem (4.2) is then a state-space optimal control problem which can be solved using the Hamilton-Jacobi-Bellman equation
\[
0 = \min_u \left( \hat{L}(x_1, u) + V_{x_1}(x_1)\hat{F}_1(x_1, u) \right)
\]
Finding power series solutions can in this case done using the method discussed in Section 4.2.

However, a problem is that in many cases it is hard or even impossible to find an explicit expression for $\varphi(x_1, u)$. The remedy to this problem is to discover that in order to compute the power series solutions of $V(x)$ and $u_*(x)$, using the method described in Section 4.2, only the series expansions of the functions involved, i.e., $F(x, u)$ and $L(x, u)$, are needed. Thus, in order to determine a power series solution to the descriptor problem only the series expansions of $\hat{F}(x_1, u)$ and $\hat{L}(x_1, u)$ around $(x_1, u) = (0, 0)$, have to be computed.

#### 4.3.1 Power Series Expansion of the Reduced Problem

The idea, used in this thesis, is to utilize that the series expansions of the composite functions $\hat{F}_1(x_1, u)$ and $\hat{L}(x_1, u)$ can be computed as the compositions of the series expansions of $F_1(x_1, x_2, u)$, $L(x_1, x_2, u)$ and $\varphi(x_1, u)$. The composite power series expansions are known to converge if the functions involved are analytical, see Brown and Churchill (1996). This condition is satisfied since $F_1(x_1, x_2, u)$, $L(x_1, x_2, u)$ are assumed to be
analytical in the problem formulation and \( \varphi(x_1, u) \) is analytical when \( F_2(x_1, x_2, u) \) is analytical, according to Theorem A.1.

Before the series expansions of \( \hat{F}(x_1, u) \) and \( \hat{L}(x_1, u) \) can be calculated, we need to calculate the series expansion of \( \varphi(x_1, u) \). This can be done recursively using the power series expansion of \( F_2(x_1, x_2, u) \) as shown below.

Assume

\[
x_2 = \varphi(x_1, u) = \varphi^{[1]}(x_1, u) + \varphi_h(x_1, u) \tag{4.16}
\]

where \( \varphi_h(x_1, u) \) contains terms of degree two and higher. From (4.5) we have that the series expansion of \( F_2(x_1, x_2, u) \) is given by

\[
0 = F_2(x_1, x_2, u) = A_{21}x_1 + A_{22}x_2 + B_2u + F_{2h}(x_1, x_2, u) \tag{4.17}
\]

If (4.16) is combined with (4.17) the expression obtained is

\[
0 = A_{21}x_1 + A_{22}(\varphi^{[1]}(x_1, u) + \varphi_h(x_1, u)) + B_2u + F_{2h}(x_1, \varphi^{[1]}(x_1, u) + \varphi_h(x_1, u), u) \tag{4.18}
\]

Since (4.18) must hold for all \((x_1, u)\) in a neighborhood of the origin, the first order term of \( \varphi(x_1, u) \) will be given by

\[
\varphi^{[1]}(x_1, u) = -A_{22}^{-1}A_{21}x_1 - A_{22}^{-1}B_2u \tag{4.19}
\]

since all other terms have degrees higher than one. Furthermore, since \( F_{2h}(x_1, x_2, u) \) only contains terms with at least degree two, we have that

\[
F_{2h}^{[m]}(x_1, \varphi(x_1, u), u) = F_{2h}^{[m]}(x_1, \varphi^{[1]}(x_1, u) + \ldots + \varphi^{[m-1]}(x_1, u), u) \tag{4.20}
\]

The fact that the \( m: \)th order term of \( F_{2h}(x_1, \varphi(x_1, u)) \) only depends on terms of \( \varphi(x_1, u) \) of degree \( m - 1 \) or less, makes it possible to derive a recursive expression for a general degree term of \( \varphi(x_1, u) \) as

\[
\varphi^{[m]}(x_1, u) = -A_{22}^{-1}F_{2h}^{[m]}(x_1, \varphi^{[1]}(x_1, u) + \ldots + \varphi^{[m-1]}(x_1, u), u) \tag{4.21}
\]

using (4.18).

For later convenience the first order approximation of \( \varphi(x_1, u) \), i.e., (4.19), is used to define the variable transformation

\[
\begin{pmatrix}
x_1 \\
\varphi^{[1]}(x_1, u) \\
u
\end{pmatrix} = \Pi \begin{pmatrix}
x_1 \\
u
\end{pmatrix} = \begin{pmatrix}
I & 0 \\
-A_{22}^{-1}A_{21} & -A_{22}^{-1}B_2
\end{pmatrix} \begin{pmatrix}
x_1 \\
u
\end{pmatrix} \tag{4.22}
\]

The variable transformation will mainly be used compute the second order term of the reduced cost function \( \hat{L}(x_1, u) \).

To compute the series expansion of \( \hat{F}_1(x_1, u) \), (4.16) is inserted into (4.5a). The system (4.15a) can then be written as

\[
\dot{x}_1 = \hat{A}x_1 + \hat{B}u + \hat{F}_{1h}(x_1, u) \tag{4.23}
\]
where $\hat{A} = A_{11} - A_{12}A_{22}^{-1}A_{21}$, $\hat{B} = B_{1} - A_{12}A_{22}^{-1}B_{2}$ and

$$\hat{F}_{1h}(x_1, u) = F_{1h}(x_1, \varphi(x_1, u), u) + A_{12}\varphi_h(x_1, u)$$

(4.24)

In the same manner, the series expansion of (4.15b) is obtained as

$$\hat{L}(x_1, u) = \begin{pmatrix} x_1 \\ u \end{pmatrix}^T \Pi^T \begin{pmatrix} Q & S \\ ST & R \end{pmatrix} \Pi \begin{pmatrix} x_1 \\ u \end{pmatrix} + \hat{L}_h(x_1, u)$$

(4.25)

where

$$\hat{L}_h(x_1, u) = L_h(x_1, \varphi(x_1, u), u) + 2x_1^TQ_{12}\varphi_h(x_1, u)$$

$$+ 2\varphi^{[1]}(x_1, u)Q_{22}\varphi_h(x_1, u) + \varphi_h(x_1, u)^TQ_{22}\varphi_h(x_1, u)$$

$$+ 2u^TS_{22}\varphi_h(x_1, u)$$

(4.26)

### 4.3.2 Application of the Results for State-Space Systems

The system description (4.23) and the description of the cost function (4.25) is now possible to incorporate in the method described in Section 4.2. Using the first order terms of the series expansions (4.23) and (4.25), the ARE (4.10a) and the expression for the first order term in the feedback (4.10b) become

$$P\hat{A} + \hat{A}^TP - P\hat{B}\hat{R}^{-1}\hat{B}^TP + \hat{Q} - \hat{S}\hat{R}^{-1}\hat{S}^T = 0$$

(4.27a)

$$D_* + \hat{R}^{-1}\left(\hat{S}^T + \hat{B}^TP\right) = 0$$

(4.27b)

where $\hat{A} = \hat{A} - \hat{B}\hat{R}^{-1}\hat{S}^T$.

To compute the higher order terms of $V(x_1)$ and $u_*(x_1)$ we make use of (4.11). In (4.11) the series expansion coefficients of $F_h(x, u)$ and $L_h(x, u)$ are replaced by the corresponding coefficients of $\hat{F}_{1h}(x_1, u)$ and $\hat{L}_h(x_1, u)$, respectively. The result is

$$V_{x_1}^{[m]}(x_1)\hat{A}_c x_1 = -\sum_{k=3}^{m-1} V_{x_1}^{[k]}(x_1)\hat{B}u_*^{[m-k+1]}(x_1)$$

$$- \sum_{k=2}^{m-1} V_{x_1}^{[k]}(x_1)\hat{F}_{1h}^{[m-k+1]}(x_1, u_*)$$

(4.28a)

$$- 2 \sum_{k=2}^{\left\lfloor \frac{m-1}{2} \right\rfloor} u_*^{[k]}(x_1)^T\hat{R}u_*^{[m-k]}(x_1)$$

$$- u_*^{[m/2]}(x_1)^T\hat{R}u_*^{[m/2]}(x_1) - \hat{L}_h^{[m]}(x_1, u_*)$$

where $m = 3, 4, \ldots$, $\hat{A}_c = \hat{A} + \hat{B}D_*$, and the terms $u_*^{[m/2]}$ are to be omitted if $m$ is odd.
The corresponding expression for the series expansion of the feedback law is

\[ u_*^{[k]}(x_1) = -\frac{1}{2} \hat{R}^{-1} \left( V_{\hat{h}}^{[k+1]}(x_1) \hat{B} \right) + \sum_{i=1}^{k-1} V_{\hat{x}}^{[k-i+1]}(x_1) \hat{F}_{1h;u}^{[i]}(x_1, u_*) + \hat{L}_{1h;u}^{[k]}(x_1, u_*) \]  \hspace{1cm} (4.28b)

where for \( k = 2, 3, \ldots \).

In (4.28a) the terms \( \hat{F}_{1h}^{[i]}(x_1, u_*) \) and \( \hat{L}_{h}^{[i]}(x_1, u_*) \) are given by the corresponding terms in (4.24) and (4.26) and can therefore be expressed in terms of the series expansions of the original functions as

\[ \hat{F}_{1h}^{[i]}(x_1, u_*) = F_{1h}^{[i]}(x_1, \varphi_*, u_*) + A_{12} \varphi_{h;u,*}^{[i]} \]  \hspace{1cm} (4.29a)

and

\[ \hat{L}_{h}^{[i]}(x_1, u_*) = L_{h}^{[i]}(x_1, \varphi_*, u_*) + 2x_1^T Q_{12} \varphi_{h;u,*}^{[i-1]} + 2 \sum_{k=2}^{[(i-1)/2]} (\varphi_{h,*}^{[k-1]})^T Q_{22} \varphi_{h;u,*}^{[i-k]} + 2 \sum_{k=1}^{i-2} (u_*^{[k]})^T S_2 \varphi_{h;u,*}^{[i-k]} \]  \hspace{1cm} (4.29b)

where \( \varphi_* = \varphi(x_1, u_*) \) and \( \varphi_{h,*} = \varphi_h(x_1, u_*) \).

The series expansion coefficients of the functions \( \hat{F}_{1h} \) and \( \hat{L}_h \) were obtained as (4.29a) and (4.29b), respectively. In the expression for the optimal control signal (4.28b), the derivatives of \( \hat{F}_1(x_1, u) \) and \( \hat{L}(x_1, u) \) with respect to \( u \) are also needed. These can be found as

\[ \hat{F}_{1h;u}^{[i]}(x_1, u_*) = F_{1h;u}^{[i]}(x_1, \varphi_*, u_*) \]
\[ + \sum_{j=1}^{i} F_{1h;x2}^{[j]}(x_1, \varphi_*, u_*) \varphi_{u;u,*}^{[i-j]} + A_{12} \varphi_{h;u,*}^{[i]} \]  \hspace{1cm} (4.30a)

where \( \varphi_{h;u,*}^{[i]} = \varphi_{h;u}^{[i]}(x_1, u_*) \), \( \varphi_{u;u,*}^{[i]} = \varphi_{u}^{[i]}(x_1, u_*) \) and

\[ \hat{L}_{h;u}^{[k]}(x_1, u_*) = L_{h;u}^{[k]}(x_1, \varphi_*, u_*) + \sum_{j=2}^{k} L_{h;x2}^{[j]}(x_1, \varphi_*, u_*) \varphi_{u;u,*}^{[k-j]} \]
\[ + 2x_1^T Q_{12} \varphi_{h;u,*}^{[k-1]} + 2 \varphi_*^{[1]} Q_{22} \varphi_{h;u,*}^{[k-1]} - 2B_2^T A_{22} Q_{22} \varphi_{h;u,*}^{[k]} \]
\[ + \sum_{j=1}^{k-2} (\varphi_{h;u,*}^{[j]})^T Q_{22} \varphi_{h;u,*}^{[k-j]} + \sum_{j=1}^{k-2} (\varphi_{h;u,*}^{[k-j]})^T Q_{22} \varphi_{h;u,*}^{[j]} \]
\[ + 2S_2 \varphi_{h;u,*}^{[k]} + 2 \sum_{j=1}^{k-1} u_*^{[j]} S_2 \varphi_{h;u,*}^{[k-j]} \]  \hspace{1cm} (4.30b)
Since $F_1(x, u)$, $\varphi_1(x, u)$ and $L(x, u)$ are power series of degree two, two and three, respectively, and
$$\varphi^{[i]}(x_1, u_*) = \varphi(x_1, u_*[1] + u_*[2] + \ldots + u_*[i])$$
we know that
$$\hat{F}_1^{[i]}(x_1, u_*) = \hat{F}_1^{[i]}(x_1, u_*[1] + u_*[2] + \ldots + u_*[i-1])$$
$$\hat{L}^{[i]}(x_1, u_*) = \hat{L}^{[i]}(x_1, u_*[1] + u_*[2] + \ldots + u_*[i-2])$$
and
$$\hat{F}^{[i]}_{1; u}(x_1, u_*) = \hat{F}^{[i]}_{1; u}(x_1, u_*[1] + \ldots + u_*[i])$$
$$\hat{L}^{[i-1]}_{h; u}(x_1, u_*) = \hat{L}^{[i-1]}_{h; u}(x_1, u_*[1] \ldots + u_*[i-1])$$
Hence, as for the state-space case, the right-hand side of (4.28a) depends only on the sequence (4.13), while the right-hand side of (4.28b) only depends on (4.14). So by recursively compute the terms of the sequence
$$V^{[2]}(x_1), u_*^{[1]}(x_1), \varphi^{[1]}(x_1, u_*^{[1]}), \ldots$$
it is possible to generate the power series for $V(x_1), u_*(x_1)$ and $\varphi(x_1, u_*(x_1))$.

As was seen in (4.19) and (4.21), arbitrarily high orders of the series expansion of $\varphi(x, u)$ can be computed in advance, based only on the series expansion of $F_2$. However, from a computational point of view, it can be noted from the sequence above, that the highest needed order of the series expansion of $\varphi(x, u)$ is the desired order of the approximation of $u_*(x_1)$.

We summarize the result in this section in Theorem 4.1 and an algorithm to compute the power series solution is found in Algorithm 4.1.

**Theorem 4.1**
Consider the optimal control problem (4.2) with all assumptions in Section 4.1 satisfied. Furthermore, assume that the cost matrix $\begin{pmatrix} \hat{Q} & \hat{S} \\ \hat{S}^T & \hat{R} \end{pmatrix} \succ 0$. Then, the following statements are equivalent

1. $(\hat{A}, \hat{B})$ is stabilizable.
2. The optimal control problem has a unique local solution for $x_1$ in a neighborhood of the origin, and $x_2$ chosen consistently.

**Proof:** Follows directly from the discussion above. \qed

In the problem formulation we also had a requirement that the closed loop system must have index one. This requirement is automatically satisfied as a result of two facts. First, the optimal feedback law only depends on $x_1$, i.e., $u = k(x_1)$. Second, from Assumption A2 we know that $F_{2;x_2}(x_1, x_2, u)$ is nonsingular for all $u$. Combined, these facts impose full rank of $F_{2;x_2}(x_1, x_2, u(x_1))$ and the closed loop system has index one for the optimal feedback law.
4.3.3 Conditions on the Original Data

The conditions in Theorem 4.1 are expressed in terms of the reduced optimal control problem, e.g., \( \hat{A}, \hat{B} \) and \( \hat{Q} \). However, in some cases these conditions can be translated to conditions on the original data. First, we have the condition

\[
\begin{pmatrix}
\hat{Q} \\
\hat{S}^T \\
R
\end{pmatrix} \succ 0
\tag{4.31}
\]

Since the variable transformation matrix \( \Pi \) in (4.22) has full column rank, it follows that

\[
\begin{pmatrix}
Q \\
S^T \\
R
\end{pmatrix} \succ 0 \implies \begin{pmatrix}
\hat{Q} \\
\hat{S}^T \\
\hat{R}
\end{pmatrix} \succ 0
\]

In some cases, it is not desired to penalize the variables \( x_2 \). In these cases the cost matrix is given by

\[
\begin{pmatrix}
\hat{Q} \\
\hat{S}^T \\
\hat{R}
\end{pmatrix} = \Pi^T \begin{pmatrix}
Q_{11} & 0 & S_1 \\
0 & 0 & 0 \\
S_1^T & 0 & R
\end{pmatrix} \Pi = \begin{pmatrix}
Q_{11} & S_1 \\
S_1^T & R
\end{pmatrix}
\]

which means that if the cost matrix for \( x_1 \) and \( u \) is positive definite, the cost matrix for the composite system (4.23) - (4.27) will also be so.

It is also possible to express the stabilizability condition

\[
(\hat{A}, \hat{B}) = (A_{11} - A_{12}A_{22}^{-1}A_{21}, B_1 - A_{12}A_{22}^{-1}B_2)
\tag{4.32}
\]

in terms of the descriptor system. This condition can be proven equivalent to the linearization of the descriptor system being stabilizable in descriptor sense. We formulate the result as a lemma.

**Lemma 4.2**

Assume that \( A_{22} \) has full rank, and that \( E = (I \ 0 \ 0) \). Then (4.32) is stabilizable if and only if

\[
E \dot{x} = Ax + Bu
\]

is stabilizable in descriptor sense, that is there exists a matrix \( K \in \mathbb{R}^{p \times n} \) such that

\[
E \dot{x} = (A + BK)x
\tag{4.33}
\]

is asymptotically stable according to Section 2.5.2.

**Proof:** Dai (1989) guarantees the existence of a \( K \) such that (4.33) is stable if and only if

\[
\text{rank}(sE - A \ B) = n, \quad \forall s \in \mathbb{C}^+
\]

where \( \mathbb{C}^+ \) denotes the closed right half complex plane. Note that \( \mathbb{C}^+ \) does not include infinity and therefore only finite \( s \) are considered. Pre-multiplication with a full rank matrix gives

\[
\text{rank}(sE - A \ B) = \text{rank}\left(\begin{pmatrix}
I & -A_{12}A_{22}^{-1} \\
0 & I
\end{pmatrix} \left(\begin{pmatrix}
sI - A_{11} & -A_{12} \\
-A_{21} & -A_{22}
\end{pmatrix} \begin{pmatrix}
B_1 \\
B_2
\end{pmatrix}
\right)\right)
\]

\[
= \text{rank}\left(\begin{pmatrix}
sI - A_{11} + A_{12}A_{22}^{-1}A_{21} & 0 \\
-A_{21} & -A_{22}
\end{pmatrix} \begin{pmatrix}
B_1 - A_{12}A_{22}^{-1}B_2 \\
B_2
\end{pmatrix}
\right)
\]

which proves the lemma since \( A_{22} \) is assumed have full rank. \( \square \)
Algorithm 4.1 Computation of the locally optimal solution

1. Compute the series expansion of the system and performance criterion using (4.5).

2. Compute \( \varphi^{[1]}(x_1, u) \) as described in (4.19) to obtain the expressions for \( \hat{A} \) and \( \hat{B} \).

3. Solve (4.27) to obtain \( V^{[2]}(x_1) = x_1^T P x_1 \) and \( u^{[1]}(x_1) = D^* x_1 \). Use these expressions to compute \( \varphi^{[1]}(x_1, u^{[1]}_*) \). Let \( m = 3 \).

4. Compute
\[
\hat{F}^{[m-1]}_{1h}(x_1, u_*) = \hat{F}^{[m-2]}_{1h}(x_1, u^{[1]}_* + \ldots + u^{[m-2]}_*) \\
\hat{L}^{[m]}_h(x_1, u_*) = \hat{L}^{[m-1]}_h(x_1, u^{[1]}_* + \ldots + u^{[m-2]}_*)
\]
using (4.29) and solve (4.28a) to obtain \( V^{[m]}(x_1) \).

5. Having \( V^{[m]}(x_1) \) and after computation of
\[
\hat{F}^{[m-2]}_{1h;u}(x_1, u_*) = \hat{F}^{[m-2]}_{1h;u}(x_1, u^{[1]}_* + \ldots + u^{[m-2]}_*) \\
\hat{F}^{[m-1]}_{h;u}(x_1, u_*) = \hat{F}^{[m-1]}_{h;u}(x_1, u^{[1]}_* + \ldots + u^{[m-2]}_*)
\]
using (4.30), \( u^{[m-1]}_*(x_1) \) can be obtained from (4.28b).

6. Compute \( \varphi^{[m-1]}(x_1, u) \) from (4.21) and use it to compute
\[
\varphi^{[m-1]}(x_1, u_*) = \varphi^{[m-1]}(x_1, u^{[1]}_* + \ldots + u^{[m-1]}_*)
\]

7. If a higher order approximation is desired, increase \( m \) by one and repeat from step 4.
4.4 Extension

For notational convenience is assumed throughout this chapter that Assumption A3 is satisfied. As a result of this assumption we only need to compute the series expansion of $\varphi$ in order to find the series expansions of the reduced state-space system and solve the optimal control problem. However, if Assumption A3 is not satisfied the method in Section 2.4.3 yields a description in the form (2.40) which, if an integrator chain is introduced, can be written as

\begin{align}
\hat{F}_1(x_1, x_2, u) &= 0 \quad (4.34a) \\
\hat{F}_2(x_2, u) &= 0 \quad (4.34b)
\end{align}

We assume that $\hat{F}_1$ and $\hat{F}_2$ are both analytical in a neighborhood of the origin. If $x_2$ in (4.34a) is replaced by $\varphi(x_1, u)$ the result is

$$
\hat{F}_1(x_1, \varphi(x_1, u), u) = 0 \quad (4.35)
$$

which from the construction in Section 2.4.3 locally can be solved for $\dot{x}_1$ as $\dot{x}_1 = \xi(x_1, u)$. The neighborhood in which (4.35) can be solved for $\dot{x}_1$ is assumed to be $(x_1, u) \in W_{x_1, u}$.

The series expansion of $\xi(x_1, u)$ can be computed recursively, in the same manner as $\varphi(x_1, u)$. Let

$$
\dot{x}_1 = \xi(x_1, u) = \xi^{[1]}(x_1, u) + \xi_h(x_1, u) \quad (4.36)
$$

where $\xi_h(x_1, u)$ contains terms beginning with degree two. From the assumption there is an equilibrium point at $(x_1, x_2, u) = (0, 0, 0)$. It follows that $\hat{F}_1(0, 0, 0, 0) = 0$ and therefore no constant term appears in the series expansion of (4.34). The series expansion of (4.34a) can be written as

$$
0 = -\dot{x}_1 + A_{11}x_1 + A_{12}x_2 + B_1u + \hat{F}_{1h}(\dot{x}_1, x_1, x_2, u) \quad (4.37)
$$

where it is assumed that the matrix in front of $\dot{x}_1$ is a negative identity matrix. This assumption introduces no loss of generality, because from the construction it is known that $\hat{F}_{1h}(0, 0, 0, 0)$ must have full rank. Therefore, any matrix in front of $\dot{x}_1$ can be eliminated by multiplication from left by its inverse.

If (4.37) is combined with (4.16) and (4.36) it leads to the equation

$$
0 = -\xi^{[1]}(x_1, u) - \xi_h(x_1, u) + A_{11}x_1 + A_{12}\varphi^{[1]}(x_1, u) + \varphi_h(x_1, u) + B_1u + \hat{F}_{1h}(\hat{\xi}^{[1]}(x_1, u) + \xi_h(x_1, u), x_1, x_2, u) \quad (4.38)
$$

and the first term in $\xi(x_1, u)$ is obtained as

$$
\xi^{[1]} = A_{11}x_1 + A_{12}\varphi^{[1]}(x_1, u) + B_1u = (A_{11} - A_{12}A_{21}^{-1}A_{22})x_1 + (B_1 - A_{12}A_{22}^{-1}B_2)u
$$

where (4.19) is used to obtain the second equality. Since $\hat{F}_{1h}(\dot{x}_1, x_1, x_2, u)$ contains terms of at least order two it follows that

$$
\hat{F}_{1h}^{[m]}(\xi(x_1, u), \varphi(x_1, u), u) = \hat{F}_{1h}^{[m]}(\xi(x_1, u) + \ldots + \xi^{[m-1]}(x_1, u), \varphi^{[1]}(x_1, u) + \ldots + \varphi^{[m-1]}(x_1, u), u)
$$
Hence, it is possible to compute $\xi(x_1, u)$ recursively. The series expansion obtained fits into earlier described results, and the computation order for the optimal control problem becomes

$$V^{[2]}(x_1), u_1^{[1]}(x_1), \varphi^{[1]}(x_1, u_1^{[1]}), \xi^{[1]}(x_1, u_1^{[1]}) \ldots$$

As a result, it is possible to compute the optimal solution for systems which is given on the more general form (4.34), i.e., not satisfying Assumption A3, with the method derived in this chapter.

### 4.5 Example

In order to illustrate the method we will study a small example. The system dynamics are given by

$$\dot{z}_1 = z_2$$

$$\dot{z}_2 = e^{z_3} - 1 + \frac{1}{2} u$$

$$0 = z_1 - \arcsin(1 - e^{z_3} + \frac{1}{2} u)$$

(4.39)

consisting of both differential and algebraic equations. The physical interpretation of (4.39) is a Phase-Locked Loop circuit (PLL) which is used to control an oscillator so that it maintains a constant phase angle relative to a reference signal. The objective is to find the feedback law (4.3) which minimizes a performance criterion with the cost function chosen as

$$L(z_1, z_2, z_3, u) = \frac{1}{2} z_2^2 + 2z_1 z_2 + z_2^2 + z_1(-e^{z_3} + 1 + \frac{1}{2} u) + \frac{1}{2} u^2$$

(4.40)

When the variables are grouped according to $x_1 = (z_1, z_2)^T$ and $x_2 = z_3$, the system is in semi-explicit form, with

$$F_1(z_1, z_2, z_3, u) = \left(\begin{array}{c} e^{z_3} - 1 + \frac{1}{2} u \\
\end{array}\right)$$

$$F_2(z_1, z_2, z_3, u) = z_1 - \arcsin(1 - e^{z_3} + \frac{1}{2} u)$$

and $E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Expressing the descriptor system in the form (4.5a) by computing the power series expansion of $F_1(z_1, z_2, z_3, u)$ and $F_2(z_1, z_2, z_3, u)$ around the origin yields the first order terms

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1/2 \\ -1/2 \end{pmatrix}$$

(4.41)

and the higher order terms of order up till three as

$$F_{1h}(z_1, z_2, z_3, u) = \begin{pmatrix} 0 \\ 1/2 z_3^2 + 1/6 z_3^3 \end{pmatrix}$$

(4.42a)

$$F_{2h}(z_1, z_2, z_3, u) = 1/2 z_3^2 + 1/3 z_3^3 - 1/4 u z_3^2 + 1/8 u^2 z_3 - 1/48 u^3$$

(4.42b)
As can be seen in (4.41) the matrix \( A_{22} = 1 \), i.e., it is nonsingular. This guarantees that it is possible to compute the function \( \varphi(z_1, z_2, u) \) using the implicit function theorem. From (4.19) the first order term of \( \varphi(z_1, z_2, u) \) is obtained as

\[
\varphi^{[1]}(x_1, u) = -z_1 + \frac{1}{2} u \tag{4.43}
\]

and the higher order terms of \( \varphi_h(z_1, z_2, u) \) to order three are then recursively computed using (4.21) as

\[
\varphi_h(z_1, z_2, u) = -\frac{1}{2} z_1^2 + \frac{1}{2} z_1 u - \frac{1}{8} u^2 - \frac{1}{6} z_1^3 + \frac{1}{2} u z_1^2 - \frac{1}{4} u^2 z_1 + \frac{1}{24} u^3 \tag{4.44}
\]

The system matrices for the composite system (4.23), i.e., \( \hat{A} \) and \( \hat{B} \), and the the local state variable change \( \Pi \) can then be computed as

\[
\hat{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \Pi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.45}
\]

We also need the power series expansion of the cost function (4.40). Around the origin this can be computed to the fourth order as

\[
L(z_1, z_2, z_3, u) = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ u \end{pmatrix}^T \begin{pmatrix} Q_{11} & Q_{12} & S_1 \\ Q_{12}^T & Q_{22} & S_2 \\ S_1^T & S_2^T & R \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ u \end{pmatrix} - \frac{1}{2} z_1 z_2^2 - \frac{1}{6} z_1 z_3^3 \tag{4.46}
\]

where

\[
Q_{11} = \begin{pmatrix} 1/2 & 1 \\ 1 & 1 \end{pmatrix}, \quad Q_{12} = \begin{pmatrix} -1/2 \\ 0 \end{pmatrix}, \quad Q_{22} = 0
\]

\[
S_1 = \begin{pmatrix} 1/4 \\ 0 \end{pmatrix}, \quad S_2 = 0, \quad R = \frac{1}{2}
\]

The cost matrix in (4.46) is not positive definite but indefinite. However, if \( \Pi \) in (4.45) is used to transform the cost matrix according to in (4.25), the cost matrix for the composite system (4.15a) becomes

\[
\begin{pmatrix} \hat{Q} & \hat{S} \\ \hat{S}^T & \hat{R} \end{pmatrix} = \begin{pmatrix} 3/2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \tag{4.47}
\]

which is positive definite.

From Section 4.3 we know that the computations is done in the order defined by the sequence

\[
V^{[2]}(x_1), \ u_*^{[1]}(x_1), \ \varphi_*^{[1]}, \ V^{[3]}(x_1), \ldots
\]

Since \( (\hat{A}, \hat{B}) \) is stabilizable and the cost matrix (4.47) is positive definite, it follows from Lemma 4.1 that the ARE (4.27a) has a unique stabilizing positive definite solution.
The first terms in the approximation can then be computed as described in (4.12) whereby we obtain

\[ V^{[2]}(x_1) = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}^T \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \]

(4.48)

\[ u_s^{[1]}(x_1) = \begin{pmatrix} -1 \\ -2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \]

The corresponding closed loop system matrix becomes

\[ \hat{A}_c = \begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} \]

(4.49)

with the eigenvalues \( \lambda = -1 \pm 1i \). Now, it is possible to compute \( \varphi_s^{[1]}(z_1, z_2) \) and \( \varphi_s^{[2]}(z_1, z_2) \) as

\[ \varphi_s^{[1]} = -\frac{3}{2} z_1 - z_2, \quad \varphi_s^{[2]} = -\frac{9}{8} z_1^2 - \frac{3}{2} z_1 z_2 - \frac{1}{2} z_2^2 \]

In order to compute the third order term in the approximation of \( V(x_1) \), (4.28) is used. For \( m = 3 \), the expression becomes

\[ V_{x_1}^{[3]}(x_1) \hat{A}_c x_1 = -V_{x_1}^{[2]}(x_1) \hat{F}_{1h}^{[2]}(x_1, u_*) - \hat{L}_{h}^{[3]}(x_1, u_*) = 0 \]

(4.50)

since from (4.29)

\[ \hat{F}_{1h}^{[2]}(x_1, u_*^{[1]}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \hat{L}_{h}^{[3]}(x_1, u_*^{[1]}) = 0 \]

(4.51)

Solving (4.50) then results in

\[ V^{[3]}(x_1) = 0 \]

(4.52)

**Remark 4.1.** The problem is symmetric in the sense that the same dynamics, except for the sign of the control input, are obtained if \( x_1 \) is replaced by \( -x_1 \). Therefore, the same optimal performance criterion should be obtained using the same change of variables. Thus, it is natural that \( V^{[k]}(x_1) = 0 \) for odd \( k \).

The next step is to solve for \( u_s^{[2]}(x_1) \). Letting \( k = 2 \) in (4.28) results in

\[ u_s^{[2]}(x_1) = -\frac{1}{2} \hat{R}^{-1} \left( V_{x_1}^{[3]}(x_1) \hat{B} + V_{x_1}^{[2]}(x_1) \hat{F}_{1h;u}^{[1]}(x_1, u_*) + \hat{L}_{h;u}^{[2]}(x_1, u_*) \right) \]

(4.53)

The unknown terms in (4.53) are then computed utilizing (4.30) and we obtain

\[ \hat{F}_{1h;u}^{[1]}(x_1, u_*^{[1]}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \hat{L}_{h}^{[2]}(x_1, u_*^{[1]}) = 0 \]

(4.54)

which combined with \( V^{[3]}(x_1) = 0 \) yields that \( u_s^{[2]}(x_1) = 0 \) as well.
We will also calculate $V^{[4]}(x_1)$ for which $\varphi^{[2]}_x$ and $\varphi^{[3]}_{h,*}$ are required. Because $u^{[2]}_x = 0$, $\varphi^{[2]}_x$ will equal $\varphi^{[2]}_{h,*}$, while

$$\varphi^{[3]}_{h,*} = -\frac{23}{24}z_1^3 - \frac{9}{4}z_1^2z_2 - \frac{3}{2}z_1z_2^2 - \frac{1}{3}z_2^3$$

Again using (4.28a) leads to the expression

$$V^{[4]}_{x_1}(x_1)\hat{A}_c x_1 = -V^{[3]}_{x_1}(x_1)\hat{B} u^{[2]}_x - V^{[2]}_{x_1}(x_1)\hat{F}^{[3]}_{1h}(x_1, u_*) - V^{[3]}_{x_1}(x_1)\hat{F}^{[2]}_{1h}(x_1, u_*) - \hat{L}^{[4]}_{h}(x_1, u_*)$$

(4.55)

where

$$\hat{L}^{[4]}_{h}(x_1, u^{[1]}_x + u^{[2]}_x) = -\frac{1}{6}z_1^4$$

(4.56)

and the other terms in (4.55) equal zero. The solution to (4.55) becomes

$$V^{[4]}(x_1) = -\frac{1}{12}z_1^4$$

(4.57)

The details of the calculation of $u^{[3]}_x$ are omitted, but the result is $u^{[3]}_x = 0$.

In order to validate the solutions of the power series method we need the series expansions of the explicit solutions to the optimal control problem, $u_*(z_1, z_2)$ and $V(z_1, z_2)$, to compare with. The system (4.39) can be formulated in state space form as

$$\dot{z}_1 = z_2$$

$$\dot{z}_2 = -\sin(z_1) + u$$

(4.58)

with the cost function (4.40) given as

$$L(z_1, z_2, u) = \frac{1}{2}z_1^2 + 2z_1z_2 + z_2^2 + z_1\sin(z_1) + \frac{1}{2}u^2$$

By solving the HJB (4.6), the explicit expressions are given by

$$u_*(z_1, z_2) = -z_1 - 2z_2$$

(4.59)

$$V(z_1, z_2) = 2(1 - \cos(z_1)) + z_1z_2 + z_2^2$$

(4.60)

For $u_*(z_1, z_2)$ truncation is unnecessary, since the exact solution is a polynomial in $z_1$ and $z_2$ of order one. However, for $V(z_1, z_2)$ the series expansions of (4.60) to the fourth order is computed as

$$V(z_1, z_2) = z_1^2 + z_1z_2 + z_2^2 - \frac{1}{12}z_1^4$$

(4.61)

Comparison of the power series expansions of the explicit solutions (4.59) and (4.61) with the solutions obtained using the power series method (4.48), (4.52) and (4.57) shows that the same expressions are attained. Plots showing the fourth order approximation of $V$ and the optimal feedback law can be found in Figures 4.1 and 4.2, respectively.

Hence, if (4.39) had been a more natural model of the PLL, or if it would have been impossible to rewrite this model as a state-space system, the method derived in this chapter can still be used to compute the optimal solution.
Figure 4.1: The fourth order approximation of $V(x_1)$.

Figure 4.2: The optimal feedback law, $u_*(x_1)$. 
In this chapter the controllability function for nonlinear descriptor systems is considered. The controllability function describes the minimum amount of control energy required to reach a specific state in infinite time. That is, a large value means that the specific state is difficult to reach. As the formulation suggests, the controllability function is defined as the solution to an optimal control problem where the performance criterion is an energy measure of the control input.

For state-space systems the controllability function is studied in Scherpen (1994) and references therein. Scherpen shows that for linear time-invariant state-space systems the controllability function is given by the inverse of the controllability gramian multiplied from the left and right by the state. The connection between a finite, nonzero controllability function and different concepts of controllability for control-affine nonlinear systems is to some extent studied in, e.g., Scherpen and Gray (2000).

The controllability function for regular time-invariant linear descriptor systems with consistent initial conditions, has been considered in Stykel (2004). The method suggested by Stykel can also be used to handle systems that for a given input $u(t)$ have higher index, without first using index reduction.

For nonlinear descriptor systems, Lin and Ahmed (1991) consider controllability using the maximum principle. However, they do not formulate the controllability function, but instead formulate the controllability problem as a feasibility problem which is solved using optimal control.

The optimal control problem for the controllability function is in this chapter solved using the results obtained in Chapter 3 and Chapter 4. Three different methods are derived. In Section 5.2, a method is derived based on the necessary conditions for optimality given by the Hamilton-Jacobi-Bellman theory in Chapter 3. The second method can be found in Section 5.3. It is based on an approach similar to the one used by Scherpen (1994). It uses completion of squares and is applicable because the performance criterion only includes the term $u^T u$. These two methods find the controllability function on some
set \( x_1 \in \Omega_x \). The third method, presented in Section 5.4, finds a local controllability function, i.e., a controllability function valid in a neighborhood of the origin. The last method can also be used to find an approximation for the controllability function.

### 5.1 Problem Formulation

A general controllability function should basically measure the minimal amount of energy in the control signal \( u(t) \) required to reach a specific state \( x \). Therefore, it is necessary to define a measure of the control signal energy. The most common energy measure, see for example Scherpen (1994), and the energy measure used in this thesis is

\[
J_c = \int_{-\infty}^{0} m(u(t)) \, dt = \frac{1}{2} \int_{-\infty}^{0} u(t)^T u(t) \, dt
\]  

(5.1)

It would be possible to use a more general \( m(u(t)) \), but in order to get a nice interpretation it has to be positive definite, i.e., satisfying \( m(u(t)) > 0 \) for all nonzero \( u(t) \).

The controllability function \( L_c(x_1) \) is defined as the solution to the optimal control problem

\[
L_c(x_1(0)) = \min_{u(\cdot)} J_c
\]  

(5.2)

subject to the system dynamics

\[
\dot{x}_1 = F_1(x_1, x_2, u)
\]

(5.3a)

\[
0 = F_2(x_1, x_2, u)
\]

(5.3b)

and the boundary conditions

\[
x_1(0) = x_{1,0} \in \Omega_x \\
\lim_{t \to -\infty} x_1(t) = 0
\]

The system dynamics are, as in the other chapters, assumed to be in semi-explicit form and to have an equilibrium at the origin. However, different assumptions regarding the implicit function will be made in different sections. In Section 5.2 and Section 5.3, Assumption A2 is made, while in Section 5.4, Assumption A4 is used instead. In both cases it is known that \( x_2 = \varphi(x_1, u) \) for \( x_1 \in \Omega_x \) and the difference is the assumption on the region for \( u \).

The controllability function must satisfy the condition \( L_c(0) = 0 \). The reason is the interpretation of \( L_c(x_{1,0}) \) as the minimum amount of input energy required to drive the system from zero at \( t = -\infty \) to \( x_1(0) = x_{1,0} \) at \( t = 0 \). According to the assumptions, the system (5.3) has an equilibrium in the origin and since no control effort is needed to keep \( x = 0 \) the energy then equals zero. We also use a convention that if \( x_0 \) cannot be asymptotically reached from 0, i.e., if no control input such that \( x_1(-\infty) = 0 \) and \( x_1(0) = x_{1,0} \) exists, \( L_c(x_{1,0}) \) is infinite.
The boundary conditions on \( x_1 \) implies that the control input \( u(\cdot) \) must make the closed loop system asymptotically anti-stable for \( x_{1,0} \in \Omega_x \). That is, if the time is reversed and instead is considered as going from 0 towards \(-\infty\), the closed loop system must be asymptotically stable.

We will only consider controllability on the set \( \mathcal{N} = \{ x_1 \in \Omega_x, x_2 \in \mathbb{R}^{n_2} | x_2 = \varphi(x_1, u), u \in \Omega_u \} \)

where \( \Omega_u \) is either \( \mathbb{R}^p \) or a neighborhood of \( u = 0 \) depending on the assumption made. This means that, only controllability on the set of points not violating the constraints are considered. This is assured by choosing the final state \( x_{1,0} \) in \( \Omega_x \). Note that if it is possible to reach all \( x_{1,0} \in \Omega_x \), it is also possible to reach all \( (x_{1,0}, x_{2,0}) \in \mathcal{N} \). The reason is that it is possible to use some \( u(t) \) for \(-\infty < t < 0 \) and then at \( t = 0 \) use \( u(0) \) to obtain \( x_2(0) = x_{2,0} \). This does not change the value of the performance criterion.

### 5.2 Method Based on HJB Theory

In this section the system is assumed to satisfy Assumption A2, that is \( \Omega_u = \mathbb{R}^p \). The optimal control problem (5.2) can then be seen as a special case of the problems considered in Chapter 3 with the cost function

\[
L(x_1, x_2, u) = \frac{1}{2} u^T u
\]

(5.4)

However, since the final state and not the initial state is specified, the time in (5.2) can be considered as going backwards compared to (3.3). This fact leads to the change of some signs in the Hamilton-Jacobi-like equation (3.8), and the result is the equation

\[
0 = \min_u \left( \frac{1}{2} u^T u - W_1(x_1) F_1(x_1, x_2, u) - W_2(x_1, x_2) F_2(x_1, x_2, u) \right)
\]

(5.5)

where \( W_1(x_1) \) and \( W_2(x_1, x_2) \) are continuous functions such that \( W_1(x_1) \) is the gradient of some continuously differentiable function \( V(x_1) \).

The necessary conditions for optimality corresponding to (5.5) can be found in Section 3.3. However, due to the structure of the cost function, the conditions can be somewhat simplified. We formulate the conditions as a proposition.

**Proposition 5.1**

Consider the optimal control problem (5.2) under Assumption A2. Then the optimal solution must satisfy

\[
0 = u^T - W_1(x_1) F_{1;u}(x_1, x_2, u) - W_2(x_1, x_2) F_{2;u}(x_1, x_2, u)
\]

(5.6a)

\[
0 = \frac{1}{2} u^T u - W_1(x_1) F_1(x_1, x_2, u)
\]

(5.6b)

\[
0 = F_2(x_1, x_2, u)
\]

(5.6c)

\[
0 = W_2(x_1, x_2) + W_1(x_1) F_{1;x_2}(x_1, x_2, u) F_{2;x_2}^{-1}(x_1, x_2, u)
\]

(5.6d)

for \( x_1 \in \Omega_x \).
Proof: Follows immediately from the conditions in Section 3.3.

In Proposition 5.1, (5.6d) is included since the system is assumed to satisfy Assumption A2.

Remark 5.1. In Section 3.5 a special case was considered where the system is given by
\[
\dot{x}_1 = f_1(x_1, x_2) + g_1(x_1)u \quad (5.7a) \\
0 = f_2(x_1, x_2) + g_2(x_1)u \quad (5.7b)
\]
In this case, for which the necessary conditions became extra simple. In the controllability function setup the equations obtained in Section 3.5 can be even more simplified. First notice that \( f_2(x_1, x_2) \) is nonsingular for all \((x_1, x_2)\) such that \( f_2(x_1, x_2) = 0 \) is solvable since \( F_{2;x_2}(x_1, x_2, u) \) is nonsingular for all \((x_1, x_2, u) \in \Omega \) and then particularly for \( u = 0 \).

Using (5.6d) an expression for \( W_2(x_1, x_2) \) can be formulated as
\[
W_2(x_1, x_2) = -W_1(x_1) f_{1;x_2}(x_1, x_2) f_{2;x_2}^{-1}(x_1, x_2)
\]
Combining this expression with (5.6a) yields
\[
u = \hat{g}(x_1, x_2)^T W_1(x_1)^T
\]
and after some more manipulations the necessary conditions can be rewritten as
\[
0 = W_1(x_1) \hat{f}(x_1, x_2) + \frac{1}{2} W_1(x_1) \hat{g}(x_1, x_2) \hat{g}(x_1, x_2)^T W_1(x_1)^T \quad (5.8a)
0 = f_2(x_1, x_2) + g_2(x_1) \hat{g}(x_1, x_2)^T W_1(x_1)^T \quad (5.8b)
\]
where
\[
\hat{f}(x_1, x_2) = f_1(x_1, x_2) - f_{1;x_2}(x_1, x_2) f_{2;x_2}^{-1}(x_1, x_2) f_2(x_1, x_2)
\]
\[
\hat{g}(x_1, x_2) = g_1(x_1) - f_{1;x_2}(x_1, x_2) f_{2;x_2}^{-1}(x_1, x_2) g_2(x_1)
\]
Hence, the original four equations with four unknowns are reduced to the two equations (5.8) and the two unknowns \( W_1(x_1) \) and \( x_2 = \eta(x_1) \).

5.3 Method Based on Completion of Squares

The HJB equation yields both necessary and sufficient conditions for optimality. However, when differentiation with respect to \( u \) is used to find the optimal control, sufficiency is lost and only necessary conditions are obtained.

In this section the class of systems considered is limited to systems in the form
\[
E \dot{x} = f(x) + g(x)u \quad (5.9)
\]
where \( E = \begin{pmatrix} I_n & 0 \\ 0 & 0 \end{pmatrix} \). For the considered class of systems it is possible to use another approach to show optimality. The approach is to a large extent similar to the approach in Scherpen (1994) and uses the fact that the performance criterion only depends on the squared control signal, i.e., \( u^T u \). The advantage of this approach is that sufficient conditions for optimality are obtained. The result is stated in Theorem 5.1.
5.3 Method Based on Completion of Squares

Theorem 5.1
Suppose there exist continuous functions $W_1(x_1) = V_1(x_1)$ and $W_2(x_1, x_2)$ such that $\tilde{L}_c(x) = (W_1(x_1), W_2(x_1, x_2))$ fulfills

$$0 = \tilde{L}_c(x) f(x) + \frac{1}{2} \tilde{L}_c(x) g(x) g(x)^T \tilde{L}_c(x)^T$$

(5.10)

for all $x \in \mathbb{N}$. Furthermore, assume that for the control choice

$$u = g(x)^T \tilde{L}_c(x)^T$$

(5.11)

the system (5.9) can be solved backwards in time from $t = 0$, with $x(t) \to 0$, $t \to -\infty$. Under these conditions, $L_c(x_1) = V(x_1)$ and the corresponding $u$ is the optimal control law.

Proof: Assume that $x_{1,0} \in \Omega_x$. For any control signal $u(\cdot)$ such that the solution to (5.9) fulfills $x(t) \to 0$ as $t \to -\infty$ it follows that

$$\frac{1}{2} \int_{-\infty}^{0} u^T u \, dt = V(x_1(0)) + \int_{-\infty}^{0} \left( \frac{1}{2} u^T u - V_{x_1} (f_1 + g_1 u) - W_2(f_2 + g_2 u) \right) \, dt$$

where $V(x_1)$, $W_2(x_1, x_2)$ are arbitrary sufficiently smooth functions. Completing the squares gives

$$\frac{1}{2} \int_{-\infty}^{0} u^T u \, dt = V(x_1(0)) + \int_{-\infty}^{0} \frac{1}{2} \| u - g(x)^T \tilde{L}_c(x)^T \|^2 \, dt$$

provided (5.10) is satisfied. It can be realized that $V(x_1(0))$ is a lower bound for the integral in (5.1). By choosing $u = g(x)^T \tilde{L}_c(x)^T$ this lower bound is obtained and since this control choice is such that the closed loop system can be solved backwards in time and $x(-\infty) = 0$, it is optimal. Therefore, for all $x_{1,0} \in \Omega_x$

$$L_c(x_{1,0}) = \min_{u(\cdot)} \frac{1}{2} \int_{-\infty}^{0} u^T u \, dt = V(x_{1,0})$$

Note that the proof above assumes that $\Omega_x$ is an invariant set for the closed loop system. However, since the closed loop system is asymptotically anti-stable such a choice of $\Omega_x$ can always be done, as was shown in Section 2.5.1.

The requirement that the closed loop system, using (5.11) must be asymptotically stable going backwards in time for $x \in \mathbb{N}$ is equivalent to

$$E \ddot{x} = -(f(\tilde{x}) + g(\tilde{x}) g(\tilde{x})^T \tilde{L}_c(\tilde{x})^T)$$

(5.12)

being asymptotically stable on $\Omega_x$, where $\tilde{x}(s) = x(-s)$ and $s = -t$. To verify that (5.12) is asymptotically stable, the methods described in Section 2.6 can be used.
Theorem 5.1 is primarily intended to verify optimality and not to calculate $W_1$ and $W_2$. This can be realized since it is assumed that $W_1$ and $W_2$ are known beforehand. Therefore, a good idea if in the case when the system fits into (5.9) is to combine the method in this section with the method in Section 5.2. First candidate solutions are found using the necessary conditions and the optimal solution is then chosen using Theorem 5.1. The approach is illustrated in Example 5.6.1.

5.4 Method to Find a Local Solution

In Chapter 4 it was shown how to compute a local solution to an optimal control problem. The solution was expressed as a power series expansion and as long as it was not truncated, it was the true optimal solution. In this section, we use the same approach for the controllability function computation. However, minor changes of the assumptions on the cost function are needed. In Chapter 4, and more specifically Theorem 4.1, the cost function $\hat{L}(x_1,u)$ is assumed to be positive definite in both $x_1$ and $u$ locally around the origin. In the calculation of the controllability function, $x_1$ and $x_2$ do not appear in the cost function and the cost function can therefore not be positive definite. However, using slightly different requirements, it is still possible to guarantee the existence of a local optimal solution in a neighborhood of the origin, i.e., a local controllability function.

5.4.1 Basic Assumptions and Formulations

The objective is to find the controllability function expressed as a convergent power series expansion

$$L_c(x_1) = \frac{1}{2} x_1^T G_c x_1 + L_{ch}(x_1)$$

(5.13)

where $L_{ch}(x_1)$ contains higher order terms at least of order three.

The considered class of control inputs can also be written as a convergent power series

$$u(x_1) = D x_1 + u_h(x_1)$$

(5.14)

where $u_h(x_1)$ is higher order terms of order two or higher. The feedback law only depends on $x_1$, which is known since the system (5.3) is assumed to satisfy Assumption A4 (cf. Section 4.3). To be able to express $F = (F_1^T, F_2^T)^T$ as a convergent series expansions, it is assumed that $F$ satisfies Assumption A5, i.e., is analytical on some set $\mathcal{W}$, where $\mathcal{W}$ is given in (4.4). Under this assumption, $F(x_1, x_2, u)$ can be expressed as described in (4.5a).

The cost function is already expressed as a convergent power series, since

$$L(x_1, x_2, u) = \frac{1}{2} u^T u$$

(5.15)

The corresponding cost matrix is

$$\begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1/2I \end{pmatrix}$$

and no higher order terms exist, i.e., $L_h(x_1, x_2, u) = 0$. 
The series expansion of the reduced system can be found in the same manner as in Chapter 4. However, since the boundary condition on $x_1$ is the final state and not the initial state, time must be considered as going backwards compared to the optimal control problem (4.2). Therefore, when using the results in Chapter 4, the system (4.23) will be

$$\dot{x}_1 = -\hat{A}x_1 - \hat{B}u - \hat{F}_{1h}(x_1)$$

(5.16)

where $\hat{A} = A_{11} - A_{12}A_{22}^{-1}A_{21}$, $\hat{B} = B_1 - A_{12}A_{22}^{-1}B_2$ and $\hat{F}_{1h}(x_1)$ is found in (4.24). That is, the signs have changed.

The reduced cost function $\hat{L}$ will in this case be very simple since $L$ depends on neither $x_1$ nor $x_2$. Therefore, $\hat{L}$ is given by

$$\hat{L}(u) = \frac{1}{2}u^T u$$

(5.17)

which yields the cost matrix

$$
\begin{pmatrix}
\hat{Q} & \hat{S}^T \\
\hat{S} & \hat{R}
\end{pmatrix} = \begin{pmatrix}
0 & 0 \\
0 & 1/2I
\end{pmatrix}
$$

(5.18)

and no higher order terms, i.e., $\hat{L}_{h}(u) = 0$. The lack of higher order terms also leads to $\hat{L}_{h;u}(u) = 0$.

One assumption in the problem formulation, presented in Section 5.1, is that the control input must be chosen such that the closed loop system is made asymptotically anti-stable. This is equivalent to that the closed loop system in backward time is asymptotically stable. In this case, the system going backwards in time is given by (5.16) and since only local properties are considered, a sufficient condition for the control input (5.14) is that $-\hat{A} - \hat{B}D$ is Hurwitz, i.e., $\Re\lambda_i(-\hat{A} - \hat{B}D) < 0$ for $i = 1, \ldots, n_1$. We formulate this as an assumption.

**Assumption A6.** Only matrices $D$ such that $-\hat{A} - \hat{B}D$ is Hurwitz are considered.

### 5.4.2 Application of the Local Method

Using the information in the previous section, it is possible to modify the results in Chapter 4 so that the results can be used to find the controllability function. In this section we derive the new equations and a new existence theorem.

Using (4.27), the equations for the lowest order terms of (5.13) and (5.14), i.e., $G_c$ and $D_*$, respectively, are obtained as

$$G_c\hat{A} + \hat{A}^T G_c + G_c\hat{B}\hat{B}^T G_c = 0$$

(5.19a)

$$D_* - \hat{B}^T G_c = 0$$

(5.19b)

The equations for the higher order terms of $L_c(x_1)$ are obtained from (4.28a) as

$$L_{c;x_1}(x_1)\hat{A}c x_1 =$$

$$- \sum_{k=3}^{m-1} L_{c;x_1}^{[k]}(x_1)\hat{B}u_*^{[m-k+1]}(x_1) - \sum_{k=2}^{m-1} L_{c;x_1}^{[k]}(x_1)\hat{F}_{1h}^{[m-k+1]}(x_1, u_*)$$

(5.20a)

$$+ \sum_{k=2}^{[m/2]} u_*^{[k]}(x_1) T u_*^{[m-k]}(x_1) + \frac{1}{2} u_*^{[m/2]}(x_1) T u_*^{[m/2]}(x_1)$$
where \( m = 3, 4, \ldots, \hat{A}_c = \hat{A} + \hat{B}D_s \), and the terms \( u^{[m/2]} \) are to be omitted if \( m \) is odd. The corresponding equation for the series expansion of the feedback law is obtained from (4.28b) as

\[
u^*[k](x_1) = \sum \sum \sum L^{[k]}(x_1) \hat{F}^*[i\in h, u](x_1, u)
\]

where \( k = 2, 3, \ldots \).

The equations in (5.20) are very similar to the original equations in Chapter 4, however, the computation of the equations above are less involved. The reason lies in the expressions for \( \hat{L}_h \) and \( \hat{L}_{h; u} \). In Chapter 4, these expressions are given by (4.29b) and (4.30b), respectively, which are rather involved, while in this chapter both of these are zero.

As can be seen in Theorem 4.1, a fundamental assumption in Chapter 4 was a positive definite cost matrix (5.18). However, this requirement is not satisfied in calculation of the controllability function. The results in Chapter 4 are based on the proof in Lukes (1969). Careful examination of the proof shows that most parts still hold when \( x \) is not present in the cost function as long as \( \hat{R} \) is positive definite, which is satisfied in the computation of the controllability function. However, two statements rely on positive definiteness of (5.18). First, the argument that the second order term in the optimal performance criterion is positive definite. This fact is used to determine which solution to the ARE (4.10a) is optimal. Second, the assertion that the ARE (4.10a) has a unique positive definite solution, which moreover locally stabilizes the closed loop system. We therefore need to investigate which properties \( G_c \) will have when the cost function is given by (5.18), and under which conditions the ARE (5.19a) has a stabilizing solution such that the properties of \( G_c \) are satisfied.

The two parts that does obviously hold in the proof by Lukes (1969) is reformulated in the following lemmas.

**Lemma 5.1**
Given an asymptotically stable matrix \( \hat{A} \) and a matrix \( D \) such that \( \hat{A} - \hat{B}D \) is Hurwitz, the matrix \( G_c \) in the controllability function (5.13) is positive semidefinite.

**Proof:** From Lukes (1969) it is known that if we consider systems given by (5.16) and feedback laws (5.14) such that \( \hat{A} - \hat{B}D \) is Hurwitz, the matrix \( P \) in a general optimal control problem, described in Chapter 4, is given by

\[
P = \int_0^{\infty} e^{(-\hat{A}-\hat{B}D)^T t} (\hat{Q} + \hat{S}D + D^T \hat{S}^T + D^T D) e^{(-\hat{A}-\hat{B}D)^T t} \, dt
\]

In the controllability function case, where the cost matrix is given by (5.18), it follows that

\[
G_c = \int_0^{\infty} e^{(-\hat{A}-\hat{B}D)^T t} D^T D e^{(-\hat{A}-\hat{B}D)^T t} \, dt
\]

which obviously is positive semidefinite. \(\square\)
Lemma 5.1 is rather simple, but is formulated to show the difference compared to Lukes (1969). If necessary, Lemma 5.1 can be strengthened to guarantee $G_c \succcurlyeq 0$, however this is not needed to determine which of the solutions to the necessary conditions for optimality that is optimal.

The lemma below states when the ARE (5.19a) can be expected to have a solution such that $G_c$ becomes at least positive semidefinite and $D_s$ becomes stabilizing.

**Lemma 5.2**

Assume that $\hat{A}$ is Hurwitz and $(\hat{A}, \hat{B})$ is controllable. Then the ARE

$$0 = \hat{A}^T G_c + G_c \hat{A} + G_c \hat{B} \hat{B}^T G_c$$

(5.21)

has a unique positive definite solution $G_c$ such that $\hat{A}_{c,b} = -\hat{A} - \hat{B} \hat{B}^T G_c$ is Hurwitz. Furthermore, $G_c$ is the only positive semidefinite solution such that $\hat{A}_{c,b}$ is Hurwitz.

**Proof:** The ARE in (5.21) is somewhat special since it has no constant term. Therefore, $G_c = 0$ is always a solution. However, for the given assumptions $G_c = 0$ can never make $\hat{A}_{c,b}$ Hurwitz, since $\hat{A}$ is Hurwitz.

From the assumptions it follows that: a) $(-\hat{A}, -\hat{B})$ is stabilizable, because $(\hat{A}, \hat{B})$ is controllable, b) $(-\hat{A}, 0)$ has no undetectable modes on the imaginary axis, because $\hat{A}$ is asymptotic stable. According to Bittanti et al. (1991), a) and b) combined will yield that (5.21) has a unique maximal positive definite solution, $G_c$. Moreover, this solution is such that $\hat{A}_{c,b}$ is Hurwitz.

The stabilizing solution $G_c$, is also the only positive definite solution. To show this, first note that for all $G_c \succcurlyeq 0$, (5.21) can be reformulated as a Lyapunov equation

$$0 = \hat{A} G_c^{-1} + G_c^{-1} \hat{A}^T + \hat{B} \hat{B}^T$$

Hence, all positive definite solutions to the ARE are positive definite solutions to the Lyapunov equation and vice versa. It is well-known that the Lyapunov equation has a unique positive definite solution if $\hat{A}$ is asymptotically stable and $(\hat{A}, \hat{B})$ is controllable, see for example Kailath et al. (2000). Therefore, $G_c \succcurlyeq 0$ must be unique.

Finally, according to Bittanti et al. (1991) all other positive semidefinite solutions to (5.21) are such that $\hat{A}_{c,b}$ will have some eigenvalues with positive real part, which proves the last statement in the theorem.

The last statement, i.e., that $G_c$ is the only stabilizing solution which is at least positive semidefinite, is important. The reason is that we have not showed that $G_c$ must be positive definite in order to be optimal, and instead the stability requirement is used to select the optimal solution.

The results in this section is summarized in the following theorem.

**Theorem 5.2**

Assume that the system (5.3) satisfies Assumptions A4, A5 and A6. Furthermore, assume that $\hat{A}$ is Hurwitz and $(\hat{A}, \hat{B})$ is controllable. Then the system has a local controllability function given by

$$L_c(x_1) = \frac{1}{2} x_1^T G_c x_1 + L_{ch}(x_1)$$
where $G_c$ is the unique positive definite solution to the ARE (5.19a) such that $D_*$, given by (5.19b), makes $-\hat{A} - \hat{B}D_*$ Hurwitz.

The higher order terms $L_{ch}(x_1)$ and $u_{h,*}$ are given as the unique solutions to (5.20).

**Proof:** According to the assumptions, $\hat{A}$ is Hurwitz and only matrices $D$ such that $-\hat{A} - \hat{B}D$ is Hurwitz are considered. Lemma 5.1 then guarantees that the optimal solution $G_c$ will be positive semidefinite.

With $\hat{A}$ Hurwitz and $(\hat{A}, \hat{B})$ controllable, it follows from Lemma 5.2 that a unique positive definite solution $G_c$ exists such that all eigenvalues to $-\hat{A} - \hat{B}\hat{B}^T G_c$ have strictly negative real parts. Moreover, all other positive semidefinite solutions to the ARE (5.19a) are such that at least some eigenvalue to $-\hat{A} - \hat{B}\hat{B}^T G_c$ have a nonnegative real part.

Hence, using $D_*$ from (5.19b) the only positive semidefinite solution such that Assumption A6 is satisfied, i.e., such that $-\hat{A} - \hat{B}D_*$ is Hurwitz, is $G_c$. The first terms in the optimal solution must therefore be $G_c$ and $D_*$. When $-\hat{A} - \hat{B}D_*$ is Hurwitz, (5.20a) is uniquely solvable for a given right-hand side, see for example Lyapunov (1992). Therefore, by choosing the unique positive definite solution to the ARE, the controllability function is found.

**Remark 5.2.** The local controllability function satisfies $L_c(x_1) > 0$ for all nonzero $x_1$ in some neighborhood of the origin.

**Remark 5.3.** The extension in Section 4.4 can be used also in the computation of the controllability function to handle a more general class of systems.

As in Chapter 4 the conditions in Theorem 5.2 are formulated in terms of the reduced system. Controllability of $(\hat{A}, \hat{B})$ is equivalent to R-controllability of $(A, B)$. This can be proven in a very similar way to the proof regarding stabilizability, see Section 4.3.3 or Dai (1989).

### 5.5 Linear Descriptor Systems

In this section the methods described in Sections 5.1 and 5.3, are applied to linear descriptor systems. It should be pointed out that the purpose is just to show these methods. The theory and a method to compute the controllability function for linear descriptor systems can both be found in (Stykel, 2004) and be considered as a special case of the method in Section 5.4.2, without any higher order terms.

Suppose we have a linear descriptor system

\[
\begin{pmatrix}
I & 0 \\
0 & 0
\end{pmatrix}
\begin{pmatrix}
\dot{x}_1 \\
\dot{x}_2
\end{pmatrix} =
\begin{pmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix} +
\begin{pmatrix}
B_1 \\
B_2
\end{pmatrix} u
\]

(5.22)

satisfying Assumption A2. Since $F_{2; x_2}(x_1, x_2, u) = A_{22}$, the assumption implies $A_{22}$ being nonsingular. In this case,

$$x_2 = \varphi(x_1, u) = -A_{22}^{-1}A_{21}x_1 - A_{22}^{-1}B_2u$$

for all $x_1 \in \mathbb{R}^{n_1}$ and all $u \in \mathbb{R}^p$. The reduced system then becomes

\[
\dot{x}_1 = \hat{A}x_1 + \hat{B}u
\]

(5.23)
\[ \hat{A} = A_{11} - A_{12}A_{22}^{-1}A_{21} \quad \hat{B} = B_1 - A_{12}A_{22}^{-1}B_2 \]

where \( \hat{A} \) and \( \hat{B} \) are defined as above.

It will be assumed that (5.23) is controllable and asymptotically stable with \( u(t) \equiv 0 \).

In order to compute the controllability function for the linear descriptor system (5.22), the method described in Section 5.2 is applied. According to Proposition 5.1, the optimal feedback and performance criterion, respectively, must satisfy

\[ 0 = u^T - V_{x_1}(x_1)B_1 - W_2(x_1, x_2)B_2 \]  \hspace{1cm} (5.24a)

\[ 0 = \frac{1}{2}u^Tu - V_{x_1}(x_1) \left( A_{11}x_1 + A_{12}x_2 + B_1u \right) \]  \hspace{1cm} (5.24b)

\[ 0 = A_{21}x_1 + A_{22}x_2 + B_2u \]  \hspace{1cm} (5.24c)

\[ 0 = V_{x_1}(x_1)A_{12} + W_2(x_1, x_2)A_{22} \]  \hspace{1cm} (5.24d)

We assume that \( V(x_1) = \frac{1}{2}x_1^T G_c x_1 \), where \( G_c \) is positive semidefinite. After some manipulation it follows that (5.24) has a solution for all \( x_1 \), i.e., \( \Omega_x = \mathbb{R}^{n_1} \), if and only if

\[ 0 = \hat{A}^T G_c + G_c \hat{A} + G_c \hat{B} \hat{B}^T G_c \]  \hspace{1cm} (5.25)

has a solution. The corresponding feedback law is given by

\[ u = \hat{B}^T V_{x_1}(x_1) = \hat{B}^T G_c x_1 \]  \hspace{1cm} (5.26)

and \( W_2(x_1, x_2) \) is found as

\[ W_2(x_1, x_2) = x_1^T G_c A_{12}A_{22}^{-1} \]

Above only necessary conditions for optimality are considered. However, since (5.22) also fits into the control-affine form described in Section 5.3, sufficient conditions for optimality can be obtained from Theorem 5.1. If \( G_c \) is chosen as a solution to (5.25), each \( W_1(x_1) = V_{x_1} \) and \( W_2(x_1, x_2) \) solving the necessary conditions (5.24), will solve (5.10). The control law (5.26) is exactly (5.11), and the corresponding closed loop system in backward time becomes

\[ \dot{x}_1 = -(\hat{A} + \hat{B} \hat{B}^T G_c)x_1 \]  \hspace{1cm} (5.27a)

\[ x_2 = -A_{22}^{-1}(A_{21} + B_2 \hat{B}^T G_c)x_1 \]  \hspace{1cm} (5.27b)

which in order for Theorem 5.1 to be applicable, must be asymptotically stable. This fact is satisfied if (5.27a) is asymptotically stable, as can be seen in Section 2.5.2. Note that, if (5.27a) is not asymptotically stable, it is impossible to start in \( x_1(0) \) and end up in \( x(-\infty) = 0 \), and therefore asymptotic stability of (5.27a) is crucial in this section.

Therefore, a necessary and sufficient condition for the existence of an optimal solution is that the ARE (5.21) has a positive semidefinite solution such that (5.27a) is asymptotically stable. The ARE is the same as in Section 5.4 and under the assumptions above, Lemma 5.2 guarantees that there exists a unique positive definite solution \( G_c \) such that the closed loop system (5.27a) is asymptotically stable. Then it follows that the controllability function is given by

\[ L_c(x_1) = V(x_1) = \frac{1}{2}x_1^T G_c x_1 \]

for all \( x_1 \in \mathbb{R}^{n_1} \).
5 Controllability Function

\[ u(t), z_3(t), \lambda(t), z_1(t), z_2(t) \]

Figure 5.1: A disc, which rolls on a surface without slipping. The disc is affected by a nonlinear spring and a linear damper.

5.6 Examples

In order to illustrate the methods for computing the controllability function, three different examples will be presented.

5.6.1 Rolling Disc

Consider a descriptor system given by the set of differential and algebraic equations

\[ \dot{z}_1 = z_2 \quad \text{(5.28a)} \]
\[ \dot{z}_2 = -\frac{k_1}{m} z_1 - \frac{k_2}{m} z_3 - \frac{b}{m} z_2 + \frac{1}{m} \lambda \quad \text{(5.28b)} \]
\[ \dot{z}_3 = -\frac{r}{J} \lambda + \frac{1}{J} u \quad \text{(5.28c)} \]
\[ 0 = z_2 - rz_3 \quad \text{(5.28d)} \]

The descriptor system describes a disc, rolling on a surface without slipping, see Figure 5.1. The disc is connected to a fixed wall with a nonlinear spring and a linear damper. The spring has the coefficients \(k_1\) and \(k_2\), which both are positive. The damping coefficient of the damper is \(b\) which is also positive. The radius of the disc is \(r\), its inertia is given by \(J\) and the mass of the disc is \(m\). The position of the center of the disc along the surface is given by \(z_1\), while \(z_2\) the translational velocity of the same point. The angular velocity of the disc is denoted \(z_3\). The control input is denoted \(u\) and is a torque applied at the center of the disc. Finally, \(\lambda\) is the contact force between the disc and the surface.

This system description has an index equal to two and before the methods in this chapter are applied, the index need to be reduced. Using the method in Kunkel and
Mehrmann (2001), (5.28) can be rewritten as the following index one description

\[
\dot{z}_1 = z_2 \\
\dot{z}_2 = -\frac{k_1}{m}z_1 - \frac{k_2}{m}z_3^2 - \frac{b}{m}z_2 + \frac{1}{m}\lambda \\
0 = z_2 - rz_3 \\
0 = -\frac{k_2}{m}z_3^3 - \frac{k_1}{m}z_1 - \frac{b}{m}z_2 + \left(\frac{r^2}{J} + \frac{1}{m}\right)\lambda + -\frac{r}{J}u
\]

We will use the abbreviation \(x = (z_1, z_2, z_3, \lambda)^T\) and it can be seen that \(x_1 = (z_1, z_2)^T\) and \(x_2 = (z_3, \lambda)^T\).

From (5.6d), it follows that \(W_2\) must satisfy

\[
W_2(x_1, x_2) = -V_{x_1}(x_1)F_{1;x_2}(x_1, x_2, u)F_{2;x_2}^{-1}(x_1, x_2, u)
\]

Since \(F_1\) is independent of \(u\), (5.6a) becomes

\[
u = F_{2,u}(x_1, x_2, u)^T W_2(x_1, x_2)^T = (0 \quad 0 \quad J_m r z_2) V_{x_1}(x_1)^T
\]

For (5.29), it is possible to compute \(x_2 = \varphi(x_1, u)\) explicitly using the last two rows as

\[
z_3 = \frac{1}{r}z_2 \\
\lambda = \left(\frac{r^2}{J} + \frac{1}{m}\right)^{-1} \left(\frac{k_1}{m}z_1 + \frac{k_2}{m}z_3^2 + \frac{b}{m}z_2 + \frac{r}{J}u\right)
\]

Inserting (5.31) and (5.32) into (5.6b) and letting \(V(x_1) = a_1z_1^2 + a_2z_3^4 + a_3z_2^2\) makes it possible to solve (5.6b) for the unknowns \(a_1, a_2\) and \(a_3\). The in this way obtained solutions for \(V(x_1)\) become either

\[
V(x_1) = bk_1r^2z_1^2 + \frac{1}{2}bk_2r^2z_1^4 + b(J + mr^2)z_2^2
\]

or the trivial solution \(V(x_1) = 0\). Back-substitution of (5.33) into (5.30) and (5.31) yields

\[
W_2(x_1, x_2) = \left(0 -2b J z_2\right), \quad u(x_1) = 2brz_2
\]

The system is polynomial, and for given values of the parameters it would be possible to use the method in (Ebenbauer and Allgöwer, 2004) to show asymptotic anti-stability of (5.29) with the control choice (5.34). Instead, stability is proven using the closed loop reduced system when the time is considered as going backwards, i.e.,

\[
\dot{x} = -F_{red,cl}(x_1)
\]

where

\[
F_{red,cl}(x_1) = \left(-\frac{k_1}{r^2 + m}z_1 - \frac{k_2}{r^2 + m}z_3^2 - \frac{b}{r^2 + m}z_2 + \frac{1}{(r^2 + m)r}2brz_2\right)
\]
For this system, $V(x_1)$ is a Lyapunov function since

$$V(x_1) = bk_1 r^2 z_1^2 + \frac{1}{2} bk_2 r^2 z_1^4 + b(J + mr^2)z_2^2 > 0$$

$$-V_{x_1}(x_1) F_{red,cl}(x_1) = -2b^2 r^2 z_2^2 < 0$$

for all $x_1 \neq 0$, since if the $V_{x_1}(x_1) F_{red,cl}(x_1) = 0$ it requires that $z_2 = 0$, but then $z_1 = 0$ because $k_1$ and $k_2$ being positive. Therefore, the conditions in Theorem 5.1 are fulfilled, yielding

$$L_c(x_1) = V(x_1)$$

for all $x_1 \in \mathbb{R}^2$ with $u(x_1)$ chosen as (5.34).

### 5.6.2 Artificial System

The system in this section is completely artificial, but illustrates an advantage of the methods in Sections 5.2 and 5.3.

The system dynamics are given by

$$\dot{z}_1 = -z_1 + z_2 + \frac{1}{2} z_2^2$$

$$0 = z_2 - u$$

where $x_1 = z_1$ and $x_2 = z_2$. This system fits into the control affine structure (5.9) or the simplified structure (5.7) as well. The interesting feature is that the corresponding state-space model

$$\dot{z}_1 = -z_1 + u + \frac{1}{2} u^2$$

is not control affine, and it would therefore not be possible to handle using the results in Scherpen (1994). It can also be realized that the smallest reachable state is $z_1 = -\frac{1}{2}$, since $u + \frac{1}{2} u^2 > -\frac{1}{2}$.

Since the system fits into (5.7) the necessary conditions in Proposition 5.1 reduce to the conditions (5.8) which in this case become

$$0 = W_1(x_1)(-z_1 - \frac{1}{2} z_2^2) + \frac{1}{2} W_1(x_1)^2 (1 + z_2)^2$$

$$0 = z_2 - (1 + z_2)W_1(x_1)$$

(5.35a) (5.35b)

where we have used that

$$\hat{f}(x_1, x_2) = -z_1 + z_2 + \frac{1}{2} z_2^2 - (1 + z_2)z_2 = -z_1 - \frac{1}{2} z_2^2$$

$$\hat{g}(x_1, x_2) = 1 + z_2$$

The equation for the control law is

$$u = (1 + z_2)W_1(x_1)$$

and $W_2(x_1, x_2)$ is given by

$$W_2(x_1, x_2) = -(1 + z_2)W_1(x_1)$$
From (5.35b) it follows that

\[ W_1(x_1) = \frac{z_2}{1 + z_2} \]

where it is assumed that \( z_2 \neq -1 \). Combining this equation with (5.35a) yields

\[ 0 = \frac{z_2}{(1 + z_2)} \left( -z_1 + \frac{1}{2} z_2 \right) \]

which has the solutions

\[ z_2 = 0, \quad z_2 = 2z_1 \]

For the first solution, \( z_2 = 0 \), the variables \( u, W_1(x_1) \) and \( W_2(x_1, x_2) \) become

\[ u = 0, \quad W_1(x_1) = 0, \quad W_2(x_1, x_2) = 0 \]

while the second solution, \( z_2 = 2z_1 \), gives

\[ u = 2z_1, \quad W_1(x_1) = \frac{2z_1}{1 + 2z_1}, \quad W_2(x_1, x_2) = -2z_1 \]

Hence, two different solutions to the necessary conditions in Proposition 5.1 are obtained. Since the system has control-affine structure, Theorem 5.1 can be used to determine which of these solutions is optimal. The first solution solves (5.10) on \( \mathcal{N} = \{ z_1 \in \mathbb{R}, z_2 \in \mathbb{R} | z_2 = 0 \} \), while the second solution solves (5.10) on \( \mathcal{N} = \{ z_1 \in \mathbb{R}, z_2 \in \mathbb{R} | z_1 > -\frac{1}{2}, z_2 = 2z_1 \} \). The solution with \( z_1 < -\frac{1}{2} \) has been omitted, since the set for \( z_1 \) must contain the origin.

For the first solution the closed loop dynamics are given by \( \dot{z}_1 = -z_1 \), which is asymptotically stable. Therefore, this solution cannot correspond to the controllability function.

For the second solution the closed loop system \( \dot{z}_1 = z_1(1 + 2z_1) \) is asymptotically anti-stable on \( \mathcal{N} \). Hence, this solution corresponds to the controllability function, which in this case for \( z_1 > -\frac{1}{2} \) becomes

\[ L_c(x_1) = z_1 - \frac{1}{2} \ln(2z_1 + 1) \]

(5.36)

In Figure 5.2, the controllability function is plotted. As can be seen the energy for reaching states close to \( z_1 = -\frac{1}{2} \) goes towards infinity, which agrees with the discussion earlier.

### 5.6.3 Electrical Circuit

Finally, a small example is presented where the method in Section 5.4.2 is used. The studied process is the electrical circuit in Figure 5.3. The circuit consists of a voltage source, a resistor, a capacitor and an inductor. The voltage source is assumed to be ideal, i.e., having no internal resistance and giving the voltage \( u \). The inductor is assumed to have a ferromagnetic core resulting in a saturated magnetic flux \( \Phi \) for large currents \( i \).
Figure 5.2: The controllability function $L(x_1)$ for the artificial example.

Figure 5.3: An electrical circuit consisting of an ideal voltage source and three nonlinear components, namely a resistor, a capacitor and an inductor.
By modifying a general inductor model from Hasler and Neirynck (1986), the inductor is modeled as
\[ \dot{\Phi} = u_L \]
\[ \Phi = \arctan(i) \]
where \( u_L \) is the voltage over the inductor. The capacitor is assumed to have a voltage dependent capacitance \( C \) and is therefore modeled as
\[ (1 + 10^{-2}u_C) \dot{u}_C = i \]
where \( u_C \) is the voltage over the capacitance. Finally, the resistor is described by
\[ u_R = i + i^3 \]
where \( u_R \) is the voltage over the resistor. That is, it is assume that \( u_R \) except for a linear dependence of the current, also depend on the cube of the current.

If the notation \( z_1 = u_C, z_2 = \Phi, z_3 = i \) and \( z_4 = u_L \) is used, a model for the complete system is
\[
\begin{align*}
\dot{z}_1 &= \frac{z_3}{1 + 10^{-2}z_1} \quad (5.37a) \\
\dot{z}_2 &= z_4 \quad (5.37b) \\
0 &= z_2 - \arctan(z_3) \quad (5.37c) \\
0 &= -z_1 - z_3 - z_3^3 - z_4 + u \quad (5.37d)
\end{align*}
\]
where the last equation comes from Kirchhoff’s voltage law. The model (5.37) is a descriptor system where \( x_1 = (z_1, z_2)^T \) and \( x_2 = (z_3, z_4)^T \).

It can be verified that (5.37) satisfies Assumption A4, since
\[
F_{2; x_2}(0, 0, 0) = \begin{pmatrix} -1 & 0 \\ -1 & -1 \end{pmatrix}
\]
The functions in (5.37) are analytical for \( z_1 > -100 \) and Assumption A5 is therefore also satisfied. The series expansion of the implicit function up to order four becomes
\[
\varphi(z_1, z_2, u) \approx \begin{pmatrix} z_2 + \frac{1}{3}z_2^3 \\ -z_1 - z_2 - \frac{1}{3}z_2^3 + u \end{pmatrix}
\]
(5.38)
The matrices \( \hat{A} \) and \( \hat{B} \) can then be computed as
\[
\hat{A} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}
\]
and it can verified that \( \hat{A} \) is Hurwitz and \( (\hat{A}, \hat{B}) \) is controllable. Theorem 5.2 now guarantees the existence of a local controllability function (5.13) and a corresponding control signal (5.14).
By first solving the ARE (5.19a) and the equation for matrix $D$ (5.19b) we obtain

$$G_c = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad D_* = \begin{pmatrix} 0 & 2 \end{pmatrix}$$

Using these, it is possible to recursively solve the equations (5.20) to obtain $L_c(x_1)$ and $u_*(x_1)$. The terms of $L_c(x_1)$ up to order four and of $u_*(x_1)$ up to order three respectively, are found to be

$$L_c(x_1) \approx z_1^2 + z_2^2 + \frac{1}{150} z_1^3 + \frac{2}{7} z_1^4 + \frac{4}{f} z_2^4 + \frac{4}{f} z_1^2 z_2^2 + \frac{8}{21} z_1 z_2^3$$

$$u_*(x_1) \approx 2z_2 + \frac{8}{7} z_2 z_1^2 + \frac{16}{7} z_2^3 z_1 + \frac{8}{7} z_1^2 z_2$$

In Figure 5.4, the fourth order approximation of $L_c(x_1)$ is shown, while in Figure 5.5 shows the third order approximation of $u_*(x_1)$.

**Figure 5.4:** The fourth order approximation of $L_c(x_1)$. 
Figure 5.5: The third order approximation of $u_*(x_1)$. 

}\clearpage
In the previous chapter, the controllability function was investigated and some different methods to compute it were derived. In this chapter, the observability function is considered instead. The observability function measures the energy in the output signal when the system is released from a given state and the control input is equal to zero. The basic idea is that if a state is observable, the energy in the output signal will be nonzero.

For nonlinear state-space systems in control-affine form, the computation of the observability function has been studied in for example (Scherpen, 1994; Gray and Mesko, 1999). Both these references also show that for a time-invariant linear state-space system, the observability function equals the observability gramian post- and pre-multiplied by the state. To find the observability function, a linear partial differential equation needs to be solved. In practice, an explicit solution can be hard to find. Therefore, numerical methods to compute the observability function have been studied. One such method is described in Scherpen (1994) and yields a local observability function expressed as a power series. The computations are very similar to the computations in Chapter 4. Another kind of methods are the empirical methods, based on stochastics, see Newman and Krishnaprasad (1998, 2000).

The observability function has also been studied for regular linear time-invariant descriptor systems with consistent initial conditions, see Stykel (2004). The method presented by Stykel can also handle descriptor systems of higher index without using index reduction.

In this chapter two different methods to calculate the observability function for nonlinear descriptor systems are presented. In Section 6.2, the approach based on the explicit solution of the first order linear partial differential equation is extended. As earlier mentioned, it can in many cases be hard to find an explicit solution, and therefore the approach based on power series is also extended. It means that a local observability function, valid only in a neighborhood of the origin, is found by recursive solution of the linear partial differential equation from Section 6.2.
6.1 Problem Formulation

The observability function should reflect the energy in the output signal when the system is released from a certain initial state. It is only the energy corresponding to the initial state that is of interest and therefore the control signal is set to zero. The observability function $L_o(x_1)$ is then defined as

$$L_o(x_1(0)) = \frac{1}{2} \int_0^\infty y(t)^T y(t) \, dt$$

subject to

$$x_1(0) = x_{1,0} \in \Omega_x$$
$$u(t) = 0, \quad 0 \leq t < \infty$$

and a descriptor system. In this chapter, the descriptor system is assumed to be in the form

$$\dot{x}_1 = F_1(x_1, x_2)$$
$$0 = F_2(x_1, x_2)$$
$$y = h(x_1, x_2)$$

Hence, an output equation is added explicitly and it is assumed that $h(0,0) = 0$. The descriptor system (6.2) is also assumed to have an equilibrium at the origin.

Similar to Chapter 5, different assumptions on the implicit function is used in the different sections. In Section 6.2, Assumption A2 is made while in Section 6.3, Assumption A4 is used instead. In both cases, it is known that on a set $\Omega_x$, it holds that $x_2 = \varphi(x_1)$. The corresponding set of points $(x_1, x_2)$ satisfying the constraints will be denoted $\mathcal{N}$, i.e.,

$$\mathcal{N} = \{ x_1 \in \Omega_x, \ x_2 \in \mathbb{R}^{n_2} \mid x_2 = \varphi(x_1) \}$$

Throughout this chapter, it is assumed that $x_2(0) = x_{2,0}$ is chosen such that $(x_{1,0}, x_{2,0}) \in \mathcal{N}$, i.e., only consistent initial values are considered.

Since the control input cannot be used to stabilize the system, it is also necessary that (6.2) is asymptotically stable, at least locally on some set $\Omega'_x \subset \Omega_x$ around the origin. Otherwise, $L_o(x_{1,0})$ might become infinite. For notational convenience the consistent states corresponding to $\Omega'_x$ is defined as

$$\mathcal{N}' = \{ x_1 \in \Omega'_x, \ x_2 \in \mathbb{R}^{n_2} \mid x_2 = \varphi(x_1) \}$$

A boundary condition for the observability function is $L_o(0) = 0$, since the origin is an equilibrium and $h(0,0) = 0$.

A small but perhaps clarifying note is that in contrast to the controllability function computation, the observability function computation does not include optimization. It is just a matter of finding the solution to the system, i.e., $x(t)$, for a given initial condition and then integrating the square of the corresponding output.

Remark 6.1. As in the controllability function case it is possible to consider more general energy measures. That is, instead of using $\frac{1}{2} y(t)^T y(t)$ in (6.1), the energy measure can be some positive definite function $m(y(t))$. 
6.2 Method Based on Partial Differential Equation

Solving the descriptor system (6.2) in order to obtain an explicit solution for \( y(t) \), which can be squared and integrated, is typically very hard. Therefore, other methods need to be derived.

One such method is based on a first-order linear partial differential equation. The method is presented in the following theorem and is an extension of a result in Scherpen (1994).

**Theorem 6.1**

Suppose the system (6.2) is asymptotically stable for \( x_{1,0} \in \Omega_x' \). Further, assume there exists a continuously differentiable positive semidefinite function \( V(x_1) \) satisfying \( V(0) = 0 \) and

\[
0 = \frac{1}{2} h(x_1, x_2)^T h(x_1, x_2) + V_{x_1}(x_1) F_1(x_1, x_2) \tag{6.4}
\]

for all \((x_1, x_2) \in N'\). Then for all \( x_{1,0} \in \Omega_x' \), it holds that

\[
L_0(x_{1,0}) = V(x_{1,0}) \tag{6.5}
\]

**Proof:** Assume that only \( x_{1,0} \in \Omega_x' \) are considered. Then, for any solution to (6.2) it follows that

\[
V(x_1(0)) = \int_0^\infty \frac{dV(x(t))}{dt} dt = \int_0^\infty -V_{x_1} F_1 dt = \int_0^\infty y^T y dt \tag{6.6}
\]

provided (6.4) is satisfied and \( V(x_1) \) is a sufficiently smooth function. Therefore, for all \( x_{1,0} \in \Omega_x' \) it follows that

\[
L_0(x_{1,0}) = \frac{1}{2} \int_0^\infty y^T y dt = V(x_{1,0})
\]

The set \( N' \) defines, in a rather implicit manner, that (6.4) only needs to be satisfied for \((x_1, x_2)\) satisfying the constraint equation. However, this dependence can be expressed more explicitly by including the constraint equation as part of the condition as well. Then, the condition (6.4) can be reformulated as

\[
0 = \frac{1}{2} h(x_1, x_2)^T h(x_1, x_2) + V_{x_1}(x_1) F_1(x_1, x_2)
\]

\[
0 = F_2(x_1, x_2)
\]

which must hold for \( x_1 \in \Omega_x' \). Another reformulation is to use that for \((x_1, x_2) \in N'\) it is known that \( x_2 = \varphi(x_1) \) and the result then becomes

\[
0 = \frac{1}{2} h(x_1, \varphi(x_1))^T h(x_1, \varphi(x_1)) + V_{x_1}(x_1) F_1(x_1, \varphi(x_1)) \tag{6.7}
\]

which must hold for \( x_1 \in \Omega_x' \). The last equation clearly shows that the implicit function must be known explicitly to compute the observability function in this way. This is a major drawback for many systems.
6.3 Method to Find a Local Solution

In earlier chapters, it has been shown that one method to overcome the problem of not knowing an explicit expression for the implicit function \( \varphi(x_1) \) is to solve the problems locally in some neighborhood of the origin. In that case, a power series expansion of the implicit function is enough and if certain assumptions are made, it is well-known that such a power series can be calculated. In this section these assumptions are formulated, and a local observability function valid in an neighborhood of the origin, is derived.

6.3.1 Basic Assumptions and Calculation of the Power Series Expansion of the Reduced System

Consider descriptor systems (6.2) which satisfies Assumption A4. Similar to in Chapter 4, another assumption is also made.

**Assumption A7.** The functions \( F_1, F_2 \) and \( h \) are analytical on a set \( \mathcal{W} \) defined as

\[
\mathcal{W} = \{ x_1 \in \mathcal{W}_{x_1} \subset \Omega, x_2 \in \mathcal{W}_{x_2} \supset \varphi(\mathcal{W}_{x_1}) \} \tag{6.8}
\]

where \( \mathcal{W}_{x_1} \) and \( \mathcal{W}_{x_2} \) are neighborhoods of \( x_1 = 0 \) and \( x_2 = 0 \), respectively.

Here, \( \Omega \) is the neighborhood on which the implicit function is defined. Based on Assumption A7 it follows that \( F_1, F_2 \) and \( h \) in (6.2) can be expanded in convergent power series as

\[
\begin{pmatrix}
F_1(x_1, x_2) \\
F_2(x_1, x_2)
\end{pmatrix} =
\begin{pmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
+ \begin{pmatrix}
F_{1h}(x_1, x_2) \\
F_{2h}(x_1, x_2)
\end{pmatrix}
\]

\[
h(x_1, x_2) = (C_1 C_2)
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix}
+ h_h(x_1, x_2)
\]

where \( F_{ih}(x_1, x_2) \) for \( i = 1, 2 \) and \( h_h(x_1, x_2) \) contains higher order terms of at least degree two. From Assumption A4 it is known that \( A_{22} \) has full rank, since \( F_{2;x_2}(0, 0) = A_{22} \) is nonsingular.

The system (6.2) can be written as the reduced state-space system

\[
\dot{x}_1 = \hat{F}_1(x_1) = F_1(x_1, \varphi(x_1))
\]

\[
y = \hat{h}(x_1) = h(x_1, \varphi(x_1))
\]

and by making a series expansion of the implicit function \( \varphi(x_1) \) using the method in Section 4.3.1, the power series expansions of the composite functions \( \hat{F}_1 \) and \( \hat{h} \) can be expressed as

\[
\hat{F}_1(x_1) = \hat{A} x_1 + \hat{F}_{1h}(x_1), \quad \hat{h}(x_1) = \hat{C} x_1 + \hat{h}_h(x_1) \tag{6.9}
\]

where

\[
\hat{A} = A_{11} - A_{12} A_{22}^{-1} A_{21}, \quad \hat{C} = C_1 - C_2 A_{22}^{-1} A_{21}
\]
and the higher order terms of $\hat{F}_{1h}(x_1)$ and $\hat{h}_h(x_1)$ can be obtained as

$$\hat{F}^{|m|}_{1h}(x_1) = F_{1h}(x_1, \varphi^{[1]}(x_1) + \ldots + \varphi^{[m-1]}(x_1)) + A_{12}\varphi^{|m|}_h(x_1)$$

$$\hat{h}^{|m|}_h(x_1) = h_{1h}(x_1, \varphi^{[1]}(x_1) + \ldots + \varphi^{[m-1]}(x_1)) + C_2\varphi^{|m|}_h(x_1)$$

for $m = 2, 3, \ldots$.

### 6.3.2 Derivation of Power Series Method

Assume that $\hat{A}$ is Hurwitz. Then in some neighborhood of $x_1 = 0$, it is known that for $x_{1,0}$ in that neighborhood the solution to (6.2) will converge exponentially towards the origin. Using similar methods to those presented in Lukes (1969), it is possible to show that the local solution for $L_o(x_1)$ will have the form

$$L_o(x_1) = \frac{1}{2}x_1^T G_o x_1 + L_{oh}(x_1)$$

(6.10)

where $L_{oh}(x_1)$ is a convergent power series on some neighborhood of $x_1 = 0$ containing terms of order three or higher. It can also be shown that $G_o$ must be at least positive semidefinite (or even positive definite) under the given assumptions.

From Section 6.2, it is known that the observability function can be found by solving (6.7). If (6.9) and (6.10) are inserted into (6.7) we obtain

$$0 = \frac{1}{2}(\hat{C}x_1 + \hat{h}_h(x_1))^T (\hat{C}x_1 + \hat{h}_h(x_1)) + (x_1^T G_o + L_{oh; x_1}(x_1)) (\hat{A} x_1 + \hat{F}_{1h}(x_1))$$

$$= \frac{1}{2}x_1^T (\hat{C}^T \hat{C} + G_o \hat{A} + \hat{A}^T G_o) x_1$$

$$+ L_{oh; x_1}(x_1) \hat{A} x_1 + L_{oh; x_1}(x_1) \hat{F}_{1h}(x_1) + x_1^T \hat{C}^T \hat{h}_h(x_1) + \frac{1}{2} \hat{h}_h(x_1)^T \hat{h}_h(x_1)$$

which is supposed to hold for $x_1$ in a neighborhood of the origin. The coefficient for each power of $x_1$ must then equal zero, leading to that $G_o$ must satisfy

$$0 = G_o \hat{A} + \hat{A}^T G_o + \hat{C}^T \hat{C}$$

(6.11a)

and the higher order terms in $L_o(x_1)$, i.e., $L_{oh}(x_1)$ must satisfy

$$L_{o; x_1}^{[m]}(x_1) \hat{A} x_1 = - \sum_{k=2}^{m-1} L_{o; x_1}^{[k]}(x_1) \hat{F}_{1h}^{[m+1-k]}(x_1) - x_1^T \hat{C} \hat{h}_h^{[m-1]}(x_1)$$

$$- 2 \sum_{k=2}^{[m/2]} \hat{h}_h^{[k]}(x_1)^T \hat{h}_h^{[m-k]}(x_1) - \hat{h}_h^{[m/2]}(x_1)^T \hat{h}_h^{[m/2]}(x_1)$$

(6.11b)

where $m = 3, 4, \ldots$. The terms $\hat{h}_h^{[m/2]}$ are to be omitted for odd $m$ and we use the convention that $\sum_{k=2}^{l} = 0$ for $l < k$.

The second order term in $L_o(x_1)$ is given by a Lyapunov equation (6.11a). A lemma presenting conditions under which the Lyapunov equation will have a solution is formulated below.
Lemma 6.1
Assume that \( \hat{A} \) is Hurwitz. Then the Lyapunov equation

\[
0 = G_o \hat{A} + \hat{A}^T G_o + \hat{C}^T \hat{C}
\]

has a unique positive semidefinite solution. If in addition \((\hat{A}, \hat{C})\) is observable, the solution is positive definite.

Proof: See for example Kailath et al. (2000).

The higher order terms are given by (6.11b). The right-hand-side is determined by the sequence

\[
L_o^{[2]}(x_1), L_o^{[3]}(x_1), \ldots, L_o^{[m-1]}(x_1)
\]

and expressions known from (6.9). Hence, when computing \(L_o^{[m]}(x_1)\) only terms of \(L_o(x_1)\) up to order \(m - 1\) are needed. Since \(\hat{A}\) is assumed Hurwitz it is known that (6.11b) has a unique solution, see for example Lyapunov (1992). Therefore, by starting with the \(L_o^{[2]}(x_1) = \frac{1}{2} x_1^T G_o x_1\), where \(G_o\) is the solution to the Lyapunov function, it is possible to recursively compute \(L_o(x_1)\).

We summarize the results above in Theorem 6.2.

Theorem 6.2
Consider a descriptor system given in the form (6.2). Assume that it satisfies Assumptions A4 and A7. Furthermore, assume \(\hat{A}\) is Hurwitz. Then, a local observability function, given in the form (6.10), exists.

The first term \(G_o\), is given as the positive semidefinite solution to

\[
0 = G_o \hat{A} + \hat{A}^T G_o + \hat{C}^T \hat{C}
\]

and higher order terms in \(L_o(x_1)\) can recursively be computed using (6.11b). If in addition \((\hat{A}, \hat{C})\) is observable, \(L_o(x_1) > 0\) for \(x_1\) in a neighborhood of the origin.

Proof: The first part follows immediately discussion in the section. The second part, i.e., that \(L_o(x_1)\) becomes positive definite locally in a neighborhood of the origin when \((\hat{A}, \hat{C})\) is observable, follows as a result of that in this case \(G_o \succ 0\), see Lemma 6.1.

Remark 6.2. Similar to the stabilizability case and the controllability case, it is possible to show that observability of \((\hat{A}, \hat{C})\) is equivalent to that \((A, C)\) is R-observable.

Remark 6.3. As in the computation of the controllability function, more general system descriptions, not being semi-explicit, can be treated using the results presented in Section 4.4.

6.4 Example

Again consider the electric circuit in Section 5.6.3. In this example a power series solution of the observability function is computed.
In the observability case an equation describing the output signal needs to be included. We assume the current \( i \) to be measured, and expressed in the notation used in Section 5.6.3, the output equation becomes

\[
y = z_3
\]

Note that \( z_3 \) is one of the algebraic variables, which is defined by the constraint equations.

Since the system description (5.37) is unchanged, the power series expansion of the implicit function \( \varphi \) equals (5.38) and the matrix \( \hat{A} \) will also be the same. The matrices \( \hat{A} \) and \( \hat{C} \) in (6.9) will then be

\[
\hat{A} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}, \quad \hat{C} = (0 \ 1)
\]

The matrix \( \hat{A} \) is, as mentioned in Section 5.6.3, Hurwitz and \( (\hat{A}, \hat{C}) \) is observable. Then it follows from Theorem 6.2 that a local observability function exists and is possible to compute.

The solution to the Lyapunov equation (6.11a) is given by

\[
G_o = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}
\]

which yields that \( L_o^{[2]}(z_1, z_2) = \frac{1}{4} z_1^2 + \frac{1}{4} z_2^2 \). Consecutive calculation of higher order terms in \( L_o(x_1) \) using (6.11b) then yields the fourth order approximation as

\[
L_o(z_1, z_2) \approx \frac{1}{4} z_1^2 + \frac{1}{4} z_2^2 + \frac{1}{600} z_1^3 - \frac{1}{28} z_1^4 - \frac{1}{14} z_2^4 - \frac{1}{14} z_2^3 z_1^2 + \frac{1}{21} z_2^3 z_1 \quad (6.12)
\]

In Figure 6.1, the approximation of \( L_o(z_1, z_2) \) is shown on a small set. On this set, the approximation is positive definite, and it looks like one can expect an observability function to do. However, if the plotted set is extended, the result is different as can be seen in Figure 6.2. Here, the approximation becomes negative for large values of \( z_1 \) and \( z_2 \), which can be seen immediately in (6.12), too. However, it enlightens two important facts. Firstly, the approximation of course introduces errors. Secondly, even without the approximation, the observability function will only be valid on a neighborhood of the origin.
Figure 6.1: The fourth order approximation of $L_o(z_1, z_2)$, plotted on a small region.

Figure 6.2: The fourth order approximation of $L_o(z_1, z_2)$, plotted on a large region.
In this chapter we present some conclusions and some possible research openings found during the work.

7.1 Conclusions

In this thesis we have mainly studied optimal feedback control and computation of the controllability function and observability function, respectively. The methods presented for solving these problems can be divided into two groups, namely those needing an explicit expression for the implicit function and those that only need its power series expansion.

Unfortunately for the first group, an explicit expression for the implicit function is known only for certain classes of systems. Moreover, even when the implicit function can be expressed explicitly the solution of the nonlinear partial differential equation obtained in the optimal control problems has to be found. Therefore, this kind of methods have limited usability in real life applications.

The methods based on power series expansions do not suffer from these problems, and can rather easily be computed for moderately sized nonlinear descriptor systems using a symbolic tool such as Maple or MATHEMATICA. It has been showed how in principle the algorithm presented in Chapter 4 can be applied to an index reduced descriptor system almost without any reformulations or assumptions. This fact opens a lot of possibilities. A possible drawback is that the computed feedback law is optimal only locally. However, in many applications a local result can be enough. Another drawback with methods based on power series expansions is the possible difficulty to know in which region the solution actually is optimal.
7.2 Future work

There are a lot of possible research openings related to the work in this thesis. Most of the them are associated with the power series expansion method described in Chapter 4. The reason is the lack of an explicit expression for the implicit function which most often is a complicating fact. However, a few interesting questions can also be found in the case when an explicit expression is required. One such question is which conditions $W_2$ in Section 3.2 must satisfy in cases when the descriptor system does not satisfy Assumption A2. The answer would probably increase the usability of the theorem in Xu and Mizukami (1993).

For the power series method there are numerous extensions. For example, it would be interesting to extend the power series method in Chapter 4 to handle discrete-time descriptor systems and performance criterions with finite time-horizon as well. Another possible development could be to extend the method for computation of a Lyapunov function using series expansion. A third interesting extension would be to find the power series solution of the optimal feedback control problem without assuming that the system satisfies Assumption A4. In this case, one further requirement on the controller is to make the closed loop system index one.

A problem with the methods based on power series expansions is that it can often be difficult to determine in what region the optimal solution is obtained. For state-space systems other methods to find approximative optimal solutions exist, for example where the HJB is approximated using successive Galerkin approximations etc. It would be interesting to study if these methods can be extended to handle descriptor systems as well.

In the area of controllability and observability it would be nice to extend the notion of the natural observability function appearing in (Gray and Mesko, 1999; Scherpen and Gray, 2000). It is defined similarly to the ordinary observability function but also includes a maximization.

The reason for Scherpen and Gray to study the controllability function and observability function in the state-space case are model reduction. For linear time-invariant systems the extension to descriptor systems is already done, but perhaps also nonlinear descriptor systems can be handled.

Finally, it would be interesting to try some of the methods in real life applications.
Some Facts from Calculus and Set Theory

This appendix provides the reader with simple definitions of some frequently used results from mathematics. For more rigorous mathematical definitions of the different concepts the reader is referred to (Isidori, 1995; Khalil, 2002) and references therein.

Manifolds

A \( k \)-dimensional manifold in \( \mathbb{R}^n \) can be thought of as the solution of the equation

\[
\eta(x) = 0
\]

where \( \eta : \mathbb{R}^n \to \mathbb{R}^{n-k} \) is sufficiently smooth, i.e., sufficiently many times continuously differentiable. A small example of a manifold is the \( n \)-dimensional unit sphere

\[
\left\{ x \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i^2 = 1 \right\}
\]

which is a \( (n - 1) \)-dimensional manifold in \( \mathbb{R}^n \).

Sets

Two topological concepts frequently used in this thesis are neighborhoods and balls. A neighborhood of a point \( p \) is any open set which contains \( p \). A ball \( B_\delta \) is an open set around some point \( p \) with radius \( \delta \), i.e.,

\[
B_\delta = \{ x \in \mathbb{R}^n \mid \| x - p \| < \delta, \ 0 < \delta \}
\]

where the norm can be any norm, but in this thesis always the Euclidian norm, i.e., \( \| x \| = \sqrt{x^T x} \), will be used. Since a neighborhood is an open set, it is always possible to place
a ball inside it and vice versa. It is therefore common that neighborhoods are assumed to be balls.

In many cases, such as in the stability analysis, the considered sets are assumed open and connected. A set is connected if every pair of points in the set can be joined by an arc lying in the set.

**Rank of Matrix-Valued Functions**

The rank of a matrix-valued function $F : \mathbb{R}^n \to \mathbb{R}^m$ at $x^0 \in \mathbb{R}^n$, is the rank of the matrix obtained if $F$ is evaluated at $x^0$. Sometimes the rank of $F$ on a manifold is considered. The rank of $F$ is then evaluated pointwise for all points belonging to the manifold.

The corank of a matrix-valued function $F$ in $x^0$ is the rank deficiency of $F(x)$ with respect to rows. That is, it is the rank of the corange, which is the nullspace of $F(x)^T$. These concepts can be studied on a manifold as well. A small example is the function

$$F(x_1, x_2) = \begin{pmatrix} x_1 & 0 \\ 0 & x_2 \\ x_1^2 & x_2^2 \end{pmatrix}$$

which at $(x_1^0, x_2^0) = (1, 0)$ has rank one and corank two, while for $(x_1^0, x_2^0) = (1, 1)$ has rank two and corank one.

**Implicit Function Theorem**

One of the most used theorems in this thesis is the implicit function theorem, Theorem A.1.

**Theorem A.1**

*Let $F : \mathbb{C}^m \times \mathbb{C}^n \to \mathbb{C}^m$ be an analytic function of $(x, y)$ in a neighborhood of a point $(x^0, y^0)$. Assume that $F(x^0, y^0) = 0$ and that the matrix $F_x(x^0, y^0)$ is nonsingular. Then, the equation $F(x, y) = 0$ has a uniquely determined analytic solution $x = \varphi(y)$ in a neighborhood of $y^0$, such that $\varphi(y^0) = x^0$.***

**Proof:** See Hörmander (1966). ☐

If $F$ is $k$-times continuously differentiable instead, the implicit function $\varphi$ is $k$-times continuously differentiable as well. The function $\varphi$ is often called a diffeomorphism. It means that $\varphi$ is a continuously differentiable function between manifolds, with a continuously differentiable inverse. However, in some references the continuous differentiability is strengthened to $C^\infty$. 


