Linköping studies in science and technology. Dissertations.
No. 1166

Optimal Control and Model Reduction of Nonlinear DAE Models

Johan Sjöberg

Department of Electrical Engineering
Linköping University, SE–581 83 Linköping, Sweden
Linköping 2008

Linköping studies in science and technology. Dissertations.
No. 1166

Optimal Control and Model Reduction of Nonlinear DAE Models

Johan Sjöberg

johans@isy.liu.se
www.control.isy.liu.se
Division of Automatic Control
Department of Electrical Engineering
Linköping University
SE–581 83 Linköping
Sweden

ISBN 978-91-7393-964-5 ISSN 0345-7524

Printed by LiU-Tryck, Linköping, Sweden 2008

To my family!

Abstract

In this thesis, different topics for models that consist of both differential and algebraic equations are studied. The interest in such models, denoted DAE models, have increased substantially during the last years. One of the major reasons is that several modern object-oriented modeling tools used to model large physical systems yield models in this form. The DAE models will, at least locally, be assumed to be described by a decoupled set of ordinary differential equations and purely algebraic equations. In theory, this assumption is not very restrictive because index reduction techniques can be used to rewrite rather general DAE models to satisfy this assumption.

One of the topics considered in this thesis is optimal feedback control. For state-space models, it is well-known that the Hamilton-Jacobi-Bellman equation (HJB) can be used to calculate the optimal solution. For DAE models, a similar result exists where a Hamilton-Jacobi-Bellman-like equation is solved. This equation has an extra term in order to incorporate the algebraic equations, and it is investigated how the extra term must be chosen in order to obtain the same solution from the different equations.

A problem when using the HJB to find the optimal feedback law 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. For analytic nonlinear time-invariant state-space models, this problem was solved in the 1960's, and in the 1970's the time-varying case was solved as well. In both cases, the optimal solution is described by convergent power series. In this thesis, both of these results are extended to analytic DAE models.

Usually, the power series solution of the optimal feedback control problem consists of an infinite number of terms. In practice, an approximation with a finite number of terms is used. A problem is that for certain problems, the region in which the approximate solution is accurate may be small. Therefore, another parametrization of the optimal solution, namely rational functions, is studied. It is shown that for some problems, this parametrization gives a substantially better result than the power series approximation in terms of approximating the optimal cost over a larger region.

A problem with the power series method is that the computational complexity grows rapidly both in the number of states and in the order of approximation. However, for DAE models where the underlying state-space model is control-affine, the computations can be simplified. Therefore, conditions under which this property holds are derived.

Another major topic considered is how to include stochastic processes in nonlinear DAE models. Stochastic processes are used to model uncertainties and noise in physical processes, and are often an important part in for example state estimation. Therefore, conditions are presented under which noise can be introduced in a DAE model such that it becomes well-posed. For well-posed models, it is then discussed how particle filters can be implemented for estimating the time-varying variables in the model.

The final topic in the thesis is model reduction of nonlinear DAE models. The objective with model reduction is to reduce the number of states, while not affecting the input-output behavior too much. Three different approaches are studied, namely balanced truncation, balanced truncation using minimization of the co-observability function and balanced residualization. To compute the reduced model for the different approaches, a method originally derived for nonlinear state-space models is extended to DAE models.

Populärvetenskaplig sammanfattning

Huvudtemat för denna avhandling är optimal styrning av olika typer system. Optimal styrning handlar om att hitta den strategi som uppnår den bästa kompromissen mellan att uppfylla de mål som ställs och att utnyttja de resurser som finns tillgängliga. Definitionen av vad som är en bra kompromiss ges i detta fall av en matematisk funktion, benämnd kostnadsfunktion. För att kunna beräkna den optimala strategin är det nödvändigt att veta hur de olika resurserna påverkar de variabler för vilka mål har ställts upp. Detta samband beskrivs med hjälp av en matematisk modell av systemet, där de variabler som kan användas för att påverka systemet benämns insignaler. Vanligast är att systemet beskrivs av en tillståndsmodell, vilket innebär att modellen enbart innehåller differentialekvationer. I denna avhandling antas dock modellen bestå både av differentialekvationer och algebraiska ekvationer, en så kallad DAE-modell. Huvudanledningen till att man vill kunna hantera modeller sammansatta av båda typerna av ekvationer är att sådana modeller uppkommer vid objekt-orienterad modellering.

Den optimala reglerstrategin beräknas genom att lösa Hamilton-Jacobi-Bellman-ekvationen. För att lösa denna ekvation krävs att man löser en olinjär partiell differentialekvation. Lösningen till ekvationen är en funktion som talar om hur systemet ska styras baserat på vilket tillstånd det befinner sig i, det vill säga en återkoppling. Ett problem är att återkopplingen bara ges av ett slutet uttryck för speciella klasser av problem. Ett enklare problem är att istället beräkna en lokalt optimal återkoppling uttryckt som en konvergent taylorutveckling. För tillståndsmodeller löstes problemet på 1960-talet och i denna avhandling utökas metoden för att även hantera system beskrivna av DAE-modeller.

En nackdel med den lokalt optimala återkopplingen är att den för vissa problem bara är en noggrann approximation i ett ganska begränsat område. Därför har en annan typ av parametrisering av återkopplingen utvärderats, nämligen rationella funktioner. I avhandlingen visas det att denna typ av parametrisering i många fall ger en återkoppling som matchar den optimala betydligt bättre och över ett större område än lösningen uttryckt som en taylorutveckling.

Att lösa Hamilton-Jacobi-Bellman-ekvationen är oftast väldigt beräkningskrävande. Dock finns det fall då beräkningarna kan förenklas. Ett sådant fall är när styrsignalerna påverkar modellen affint. För tillståndsmodeller är det ofta relativt enkelt att se om så är fallet, medan det för DAE-modeller kan vara betydligt svårare. Därför härleds villkor i avhandlingen under vilka systemet garanterat är styrsignalaffint.

Det visar sig att affinitet i en extern signal kan användas på fler sätt. Om man vill införa vissa typer av brus i olinjära DAE-modeller måste bruset komma in affint för att man ska få en matematiskt väldefinierad modell. Anledningen till att införa brus i modellen kan vara för att modellera till exempel osäkerheter i modellen eller i mätningar som görs på systemet. Vad som mäts anges av de så kallade utsignalerna. För system som uppfyller villkoren studeras en metod, kallad partikelfiltrering, för att estimera variabler i modellen.

Det sista ämnet som behandlas i avhandlingen är modellreduktion. Modellreduktion handlar om att reducera storleken på en modell utan att påverka dess insignal-utsignalbeteende i för hög grad. Reduktion innebär i detta fall att antalet differentialekvationer minskas, vilket till exempel innebär att de optimala lösningarna ovan kan beräknas snabbare. Även i detta fall utökas några existerande metoder för tillståndsmodeller till att även hantera DAE-modeller.

Acknowledgments

First of all, I would like to thank my supervisor Professor Torkel Glad for introducing me to his interesting field of DAE models and for the skillful guidance during the work on this thesis. I have really enjoyed our cooperation. 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.

Professor Frank Allgöwer and Professor Rolf Findeisen are gratefully acknowledged for letting me visit them at University of Stuttgart. It was a really nice time, and very inspiring for the rest of the time as a PhD student! Thank you!

I would also like to thank Dr. Markus Gerdin and Lic. Henrik Tidefelt for all interesting talks about DAE models, and Markus for our cooperation about nonlinear SDAE models.

I am really grateful to the persons that have proofread various parts of the thesis: Daniel Ankelhed, Dr. Daniel Axehill, Dr. Martin Enqvist, Janne Harju, Dr. Gustaf Hendeby, Christian Lyzell, Dr. Johan Löfberg, Henrik Ohlsson, Daniel Petterson, and Lic. David Törnqvist.

Furthermore, I would like to thank Lic. Henrik Tidefelt. for all the help with the cover and with the figures in the thesis, and Dr. Gustaf Hendeby for the LaTeX-template in which this thesis is written.

There are a few guys that have followed me over the years and have been a source of both inspiration and joy. These are Johan Wahlström, Daniel Axehill, Gustaf Hendeby, Henrik Tidefelt and David Törnqvist. Thanks!

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

This work has been supported by the Swedish Research Council, and The Swedish Foundation for Strategic Research through the graduate school ECSEL, which are hereby gratefully acknowledged.

Last, but absolutely not least, I would like to thank Caroline and my family for always supporting me and being interested in what I am doing. I would never have been here without you!

Linköping, March 2008
Johan Sjöberg

Contents

Notation

Symbols and Mathematical Notation

Notation	Meaning
$\mathbb{R}^n$	The n-dimensional space of real numbers
$\mathbb{C}^n$	The n-dimensional space of complex numbers
$\in$	Belongs to
$\forall$	For all
$\otimes$	the Kronecker product
$A \subset B$	A is a subset of B
$A \cap B$	Intersection between A and B
$A \cup B$	The union of A and B
∂A	Boundary of the set A
$\mathcal{V}(A)$	Range of the matrix A
$\mathcal{N}(A)$	Null space of the matrix A
$\operatorname{rank} A$	Rank of the matrix A
$\operatorname{corank} A$	Rank deficiency of the matrix A with respect to rows (see Appendix A)
$I, (I_n)$	Identity matrix (of dimension $n \times n$)
$\operatorname{diag}(x_1, \ldots, x_n)$	Diagonal matrix with $x_1, \ldots, x_n$ as diagonal entries.
$f : \mathbb{D} \to \mathbb{Q}$	The function f maps a set $\mathbb{D}$ to a set $\mathbb{Q}$
$f \in \mathcal{C}^k(\mathbb{D}, \mathbb{Q})$	The function $f : \mathbb{D} \to \mathbb{Q}$ is k-times continuously differentiable
$f_{r;x}$	Partial derivative of f_r with respect to x
$Q \succ (\succeq)\, 0$	The matrix Q is positive (semi)definite
$Q \prec (\preceq)\, 0$	The matrix Q is negative (semi)definite
$\sigma(E, A)$	The set $\{s \in \mathbb{C} \mid \det(sE - A) = 0\}$
$\operatorname{eig}_i(A)$	The i:th eigenvalue of the matrix A

Notation	Meaning
$\mathrm{Re}\ s$	Real part of s
$\mathrm{Im}\ s$	Imaginary part of s
$\mathbb{C}^+$	Closed right half complex plane
$\mathbb{C}^-$	Open left half complex plane
$\|x\|$	$\sqrt{x^T x}$
$\min_x f(x)$	Minimization of $f(x)$ with respect to x
$\operatorname{argmin}_x f(x)$	The argument x minimizing $f(x)$
B_r	Ball of radius r (see Appendix A)
$\lfloor x \rfloor$	The floor function, which gives the largest integer less than or equal to x
$\dot{x}$	Time derivative of x
$x^{(i)}(t)$	The ith derivative of $x(t)$ with respect to t
$x^{\{i\}}(t)$	Kronecker product of x i times
$f^{[i]}(x)$	All terms in a multivariable polynomial of order i
$f^{m]}(x)$	All terms in a multivariable polynomial up to order m
$\mathcal{O}(x)^d$	$f(x) = \mathcal{O}(x)^d$ as $\|x\| \to 0$ if $f(x) = \|x\|^d B(x)$ where B is bounded near the origin
$o(h)$	$f(h) = o(h)$ as $\|h\| \to 0$ if $f(h)/\|h\| \to 0$ as $\|h\| \to 0$
$A/_B\, C$	The oblique projection of the matrix A along the space B on the space C
$E(x)$	The expected value of the stochastic variable x
$a(x) \equiv b(x)$	Identically equal, $a(x) = b(x)$, $\forall x$

Abbreviations

Abbreviation	Meaning
ARE	Algebraic Riccati Equation
BLT	Block Lower Triangular Form
DAE	Differential-Algebraic Equation
DP	Dynamic Programming
HJB	Hamilton-Jacobi-Bellman (equation)
HJI	Hamilton-Jacobi Inequality
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
PLL	Phase-Locked Loop circuit
PMP	Pontryagin Minimum Principle
RDE	Riccati Differential Equation
RMSE	Root Mean Square Error
SDAE	Stochastic Differential-Algebraic Equation
w.r.t.	with respect to

Assumptions

Assumption	Short explanation
A1	The index reduced system can be expressed in semi-explicit form (see page 32)
A2	The set, on which the implicit function is defined, is global in the control input (see page 34)
A3	The considered initial conditions are consistent (see page 60)
A4	The time-varying DAE model can locally be solved for $\dot{x}_1$ and x_3 (see page 61)
A5	The functions $\hat{F}_1$, $\hat{F}_2$, L and G in the time-varying case are analytic (see page 61)
A6	The matrices $E_1(t)$ and $A_{22}(t)$ are invertible (see page 61)
A7	The time-invariant DAE model can locally be solved for $\dot{x}_1$ and x_3 (see page 72)
A8	The functions $\hat{F}_1$, $\hat{F}_2$, L and G in the time-invariant case are analytic (see page 72)
A9	The feedback law is described by uniformly convergent power series and places the eigenvalues correctly for the DAE model (see page 73)
A10	The feedback law is described by uniformly convergent power series and places the eigenvalues correctly for the state-space model (see page 85)
A11	The linearization of the model and the cost function have the correct local properties (see page 98)
A12	The implicit ODE can locally be solved for $\dot{x}$ (see page 124)
A13	The implicit function $\mathcal{R}$ is independent of u (see page 127)
A14	The functions $\hat{F}_1$ and $\hat{F}_2$ are independent of x_3 and derivatives of u, respectively (see page 131)
A15	The feedback law is described by uniformly convergent power series and stabilizes the DAE model going backwards in time locally (see page 164)
A16	The functions F_1, F_2 and h for a semi-explicit DAE model with an explicit output equation are analytic (see page 176)
A17	The linearization of the state-space model is asymptotically stable, controllable and observable (see page 182)
A18	The eigenvalues of $G_c^{-1}G_o$ are distinct (see page 182)

Assumption	Short explanation
A19	The linearization of the DAE model is asymptotically stable, R-controllable and R-observable (see page 192)

1

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 can be used to control the temperature in a room. In order to perform this task, 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 in the room is kept at the desired level.

The very simple control strategy the thermostat uses can sometimes be sufficient, 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 the model of the system and the same system can be described by models in different forms. One such form is a model that consists of both differential and algebraic equations, denoted a DAE model. The fact that both types of equations can be included in the model opens for the possibility to model systems in an object-oriented fashion. To concretize what this means consider the modeling of a car.

Normally, the first step is to model parts of the car, for example, the engine, the gearbox, the propeller shaft and the car body as separate models. These submodels may also be modeled using smaller submodels until each submodel is sufficiently small to be easily modeled. The second step is then to connect all the separate models to obtain the model of the complete car. Typically, these connections introduce algebraic equations for example describing that the output shaft from the engine must rotate with the same angular velocity as the input shaft of the gearbox. The obtained model is then in DAE form. Other examples of systems which have been successfully modeled as DAE models are

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).

Based on a DAE model of a system, controller design is considered. More specifically, the focus will be on optimal feedback control. 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 commanded by the driver. The procedure would then yield the controller achieving the best balance between low fuel consumption and delivery of the requested torque.

A problem with optimal feedback control methods are the rather extensive computations required. The complexity of the computations also scales badly with the size of the DAE model. To reduce the required computations, two different approaches can be used. One is trying to find some structure in the model, that can be utilized to simplify the computations. Another approach is to derive an approximate model which is smaller, but still reflects the major properties of the original model. Both theses approaches are examined in this thesis.

1.1 Thesis Outline

The thesis is separated into eleven main chapters. First, some preliminary facts about nonlinear DAE models and optimal feedback control are presented in Chapter 2. Chapter 3 is the first chapter devoted to optimal feedback control of DAE models. Two different methods are investigated and some relationships between their optimal solutions are revealed. Optimal control is also the topic of Chapter 4. However, in this chapter, the optimal control problem is solved for models described by convergent power series. The method is rather general, and both time-varying and time-invariant models can be handled, but the solution is normally restricted to a neighborhood of the origin.

The method in Chapter 4 is known to give the exact optimal solution as long as the power series solution is untruncated. In practice, truncation is necessary and numerical examples have shown that the obtained approximation of the optimal solution may have some bad properties. Therefore, another parametrization of the optimal solution is considered in Chapter 5, namely rational approximants. The main advantage of this parametrization is that the approximant can be required to have the same Taylor series as the optimal solution up to some desired order, while at the same time controlling how fast the optimal solution should grow when the states tend to infinity.

The methods derived in Chapter 4 and 5 become rather computationally demanding as the number of states grows. Therefore, different structural properties that may reduce the computational complexity are studied in Chapter 6. One such case is when the DAE model has an underlying state-space model that is affine in the control signal and the cost function is quadratic in that signal as well.

The analysis of affinity in an external input signal proves to be interesting for other reasons as well. In Chapter 7, it is shown how white noise can be introduced into nonlinear DAE models in a mathematical well-posed manner. The results rely on that the underlying

state-space model becomes affine in the disturbance signals. For nonlinear DAE models with a proper introduced noise, it is then shown how particle filtering can be used to estimate the time-varying variables in the model.

Chapters 8, 9 and 10 deal with model reduction of nonlinear DAE models. The first two chapters of these introduce the controllability and observability function, respectively. These functions are used to measure the energy in the input and output signal. Based on these functions, a model reduction method for state-space models are extended to the DAE case as described in Chapter 10.

Chapter 11 summarizes the thesis with some conclusions and remarks about interesting problems for future research.

1.2 Contributions

This thesis is based on both previously published, Sjöberg and Glad (2005); Glad and Sjöberg (2006); Sjöberg (2006); Sjöberg and Glad (2006b); Gerdin and Sjöberg (2006); Sjöberg et al. (2007), and previously unpublished results, Sjöberg and Glad (2008a,b). Moreover, some completely new results are derived in the thesis. 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 for nonlinear DAE models, which can be found in Chapter 3. The presentation is based on a version of the paper:

  > T. Glad and J. Sjöberg. Hamilton-Jacobi equations for nonlinear descriptor systems. In *Proceedings of the 2006 American Control Conference*, Minneapolis, Minnesota, June 2006.

- The method for proving existence and for computing the optimal feedback law of DAE models, presented in Chapter 4. The results extend earlier developed methods for nonlinear time-invariant and time-varying analytic state-space models. Further, it is proved that a discount factor can be introduced in the cost function for the infinite horizon case and, under certain technical assumptions, the optimal solution will still exist and be time-invariant.

  The material in Chapter 4 are based upon the conference papers:

  > J. Sjöberg and T. Glad. Power series solution of the Hamilton-Jacobi-Bellman equation for descriptor systems. In *Proceedings of 44th IEEE Conference on Decision and Control and European Control Conference*, Seville, Spain, December 2005.
  >
  > J. Sjöberg and S. T. Glad. Power series solution of the Hamilton-Jacobi-Bellman equation for time-varying differential-algebraic equations. In *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, California, December 2006b.

  and the technical report:

J. Sjöberg and S. T. Glad. Power series solution of the Hamilton-Jacobi-Bellman rquation for DAE models with a discounted cost. Technical Report LiTH-ISY-R-2250, Department of Electrical Engineering, Linköpings universitet, 2008b.

- The results about how rational functions can be used to approximate the optimal return function and the optimal control law. Parts of the results are published in the following paper:

  J. Sjöberg and S. T. Glad. Rational approximation of nonlinear optimal control problems. In *Proceedings of the 17th World Congress of IFAC*, Seoul, South Korea, July 2008a. Accepted for publication.

- The discussion about under what conditions, the underlying state-space model will be affine in some external input, presented in Chapter 6. In the same chapter, it is also shown how these conditions can be used to reduce the number of equations needed to solve in Chapter 4.

- The results in Chapter 7 on how so-called white noise can be incorporated into a nonlinear DAE model such that the stochastic model becomes mathematically well-posed. For DAE models with correctly introduced noise it is also shown how the particle filtering method can be applied to estimate time-varying variables in the model. These results have been published in:

  M. Gerdin and J. Sjöberg. Nonlinear stochastic differential-algebraic equations with application to particle filtering. In *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, California, December 2006.

- The different methods in Chapter 8 and 9 for computing the controllability and observability functions, respectively. The result about the computation of the controllability function was originally presented in:

  J. Sjöberg and S. T. Glad. Computing the controllability function for nonlinear descriptor systems. In *Proceedings of the 2006 American Control Conference*, Minneapolis, Minnesota, June 2006a.

- The extension of a model reduction procedure for nonlinear state-space models to nonlinear DAE models, presented in Chapter 10. Having the model in balanced form, two different methods for reducing it are studied, namely balanced truncation and balanced residualization. The result has previously been published in:

  J. Sjöberg, K. Fujimoto, and S. T. Glad. Model reduction of nonlinear differential-algebraic equations. In *Proceedings of the 7th IFAC Symposium on Nonlinear Control Systems*, Pretoria, South Africa, August 2007.

In addition to the contributions presented in the thesis, a paper about nonlinear model predictive control has been published:

R. Sjöberg, Findeisen and F. Allgöwer. Model predictive control of continuous time nonlinear differential algebraic systems. In *Proceedings of the 7th IFAC Symposium on Nonlinear Control Systems*, Pretoria, South Africa, August 2007.

2

Preliminaries

To introduce the subject to the reader, this chapter presents some basic DAE model theory. Four key concepts: index, solvability, consistency, and stability will briefly be described. Furthermore, it will be discussed how the index of a system model can be reduced using different methods, simply denoted index reduction methods. Finally, an introduction to optimal control of state-space models will be given.

2.1 System Models

The most natural mathematical model of a system is often a set of differential and algebraic equations. However, in most literature about control theory, it is assumed that the algebraic equations can be used to eliminate some variables. The outcome is a model that only consists of differential equations, and therefore can be written in state-space form as

$$\dot{x} = F(t, x, u) \tag{2.1}$$

where $F : \mathbb{I} \times \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$, $\mathbb{I} \subseteq \mathbb{R}$ is an interval, $x \in \mathbb{R}^n$ is the state vector and $u \in \mathbb{R}^p$ are the control inputs. 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 disadvantages. For example, some systems are easier to model if both differential and algebraic equations may be used, while a reduction to a state-space model is more difficult. Another possible disadvantage occurs when the structure of the model is nice and intuitive while using both kinds of equations, but the state-space formulation loose this feature. A third disadvantage is related to object-oriented computer modeling tools, such as Dymola. Usually, the models generated by these tools are not in state-space form, but a set of both algebraic and differential equations, and the number of equations is often large. This means that reducing the DAE model to a state-space model may be almost impossible.

Therefore, the focus of this thesis is on a more general class of system descriptions, called DAE models or differential-algebraic equations (DAE). This class of mathematical models includes both differential and algebraic equations and can be formulated as

$$F(t, x, \dot{x}, u) = 0 \tag{2.2}$$

where $F : \mathbb{I} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$, $\mathbb{I} \subseteq \mathbb{R}$ is an interval, $x \in \mathbb{R}^n$ are variables and $u \in \mathbb{R}^p$ are control inputs. A conceptual difference between state-space models and DAE models is that a DAE model does not need to be solvable w.r.t. $\dot{x}$, not even numerically. A result of this fact is that all components of x will not represent a memory of the past, or with other words, be described by differential equations. This is shown in the following small example (Brenan et al., 1996).

Example 2.1: A Pendulum

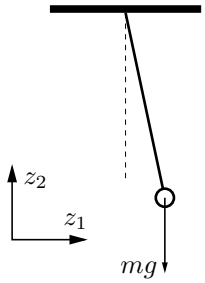


Figure 2.1: *A pendulum*

The goal is to model a pendulum. As depicted in Figure 2.1, z_1 and z_2 are the horizontal and vertical position of the pendulum. Furthermore, z_3 and z_4 are the corresponding velocities, z_5 is the tension in the pendulum, the constant b represents resistance caused by the air, g is the gravity constant, and L is the constant length of the pendulum. A model which describes the pendulum can be written as

$$\begin{aligned}
\dot{z}_1 &= z_3 \\
\dot{z}_2 &= z_4 \\
\dot{z}_3 &= -z_5 \cdot z_1 - b \cdot z_3^2 \\
\dot{z}_4 &= -z_5 \cdot z_2 - b \cdot z_4^2 - g \\
0 &= z_1^2 + z_2^2 - L^2
\end{aligned}$$

An immediate observation is that the equations are not possible to solve for $\dot{z}_5$ since it is not present in any of the equations. A first glance also indicates that the model should have four states, namely z_1 to z_4. However, as will be shown later in the thesis, it is

possible by using the constraint $0 = z_1^2 + z_2^2 - L^2$ to rewrite the model as

$$\dot{z}_1 = z_3 \tag{2.3a}$$
$$\dot{z}_3 = -z_5 \cdot z_1 - b \cdot z_3^2 \tag{2.3b}$$
$$0 = z_1^2 + z_2^2 - L^2 \tag{2.3c}$$
$$0 = z_1 z_3 + z_2 z_4 \tag{2.3d}$$
$$0 = \frac{1}{L}\big(z_3^2 + z_4^2 - b(z_1 z_3^2 + z_2 z_4^2) - z_2 g\big) - z_5 \tag{2.3e}$$

It shows that only z_1 and z_3 are really determined by differential-equations. The other variables are given by the algebraic equations (2.3c)–(2.3e).

Hence, the memory of the past is contained in z_1 and z_3, while z_2, z_4 and z_5 are algebraically determined from z_1 and z_3. Therefore, the only states in this example are z_1 and z_3. Another related property is that the number of algebraic equations has changed from one in the original set of equations to three in the equations above. The reason is that the original equations are valid over time.

In the example above, it is also possible to choose z_2 and z_4 as states and let z_1 and z_3 be determined by algebraic equations. This is an important property of DAE models that different configurations of the variables can be chosen as states. Actually, it might be necessary to change the choice of states when the solution evolves. For example, consider the pendulum above. For most positions, the choice between the two pair of states are free. However, when either z_1 or z_2 is zero, this variable and its corresponding velocity cannot be chosen as states. That is, the other pair has to be chosen.

In the reformulation above, *i.e.*, when the dynamic part was separated from the algebraic part, the function F in (2.2) was differentiated a number of times with respect to t. Therefore, an assumption made throughout the thesis is that F is sufficiently smooth, *i.e.*, sufficiently many times continuously differentiable.

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

$$F(t, x, \dot{x}) = 0 \tag{2.4}$$

The two most common reasons are either that the control input is a feedback law $u = u(t, x)$, or that $u = u(t)$ is a known signal, seen as part of the time variability. However, a third reason is that the system is modeled using a behavioral approach, see for instance Polderman and Willems (1998); Kunkel and Mehrmann (2001). Then, the control input u is simply viewed as a variable among the other, and therefore included in 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. However, when designing general purpose models, for example different electrical components in a modeling library, the input and output for a particular model may not be determined.

When modeling physical processes, the obtained model will often have more structure than the general description (2.2). One such structure is the semi-explicit form, which arises naturally for example when modeling mechanical multibody systems (Arnold et al., 2004). The model can then be expressed as

$$E\dot{x} = F(x, u) \tag{2.5}$$

where $E \in \mathbb{R}^{n\times n}$ is a possibly rank deficient matrix, *i.e.*, $\operatorname{rank} E = r \leq n$, and $F : \mathbb{R}^{n+p} \to \mathbb{R}^n$. Linear time-invariant DAE models can always be written in this form as

$$E\dot{x} = Ax + Bu \tag{2.6}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times p}$. Furthermore, the description (2.5) (and hence (2.6)) can without loss of generality be rewritten as

$$\dot{x}_1 = F_1(x_1, x_2, u) \tag{2.7a}$$
$$0 = F_2(x_1, x_2, u) \tag{2.7b}$$

where $x_1 \in \mathbb{R}^r$ and $x_2 \in \mathbb{R}^{n-r}$. It may seem like all the variables 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 the partial derivative of F_2 w.r.t. x_3, denoted $F_{2;x_2}(x_1, x_2, u)$, is nonsingular at least locally.

In some cases it might be interesting to extend the models above with an equation for an output signal y as

$$F(t, x, \dot{x}, u) = 0 \tag{2.8a}$$
$$y = h(x, u) \tag{2.8b}$$

where $y \in \mathbb{R}^q$ and $h : \mathbb{R}^{n+p} \to \mathbb{R}^q$. In general, an explicit extension of the model with an extra output equation is unnecessary for DAE models. The output equation can be included in $F(\dot{x}, x, u, y, t)$. However, in some situations, it is important to show which variables are possible to measure and then (2.8) is the best model of the system. One such case is when noise is introduced into the model. The model is then written as

$$F\big(t, x, \dot{x}, w, u\big) = 0$$
$$y = h(t, x, u) + e$$

where w is the process noise and e is the measurement noise.

2.2 Solvability and the Index Concept

Intuitively, solvability means that there exists a solution that satisfies the equations in the DAE model (2.2) for a given initial value. In this thesis, mostly two different concepts of solution will be discussed, namely classical solutions (continuously differentiable) and distributional solutions.

The definition of a classical solution is adopted from Kunkel and Mehrmann (2006) and is formulated as follows.

Definition 2.1. Consider the model (2.4) and denote the time interval for which (2.4) is defined as $\mathbb{I} \subseteq \mathbb{R}$.

1. A function $x(t) \in C^1(\mathbb{I})$ is called a solution to (2.4), if it satisfies (2.4) pointwise.

2. The function $x(t) \in C^1(\mathbb{I})$ is called a solution of the initial value problem consisting of (2.4) and
$$x(t_0) = x_0 \tag{2.10}$$
if $x(t)$ is a solution of (2.4) and satisfies (2.10).

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

4. A model is solvable if it has at least one solution.

Finding consistent initial conditions for DAE models can be rather difficult. It might seem like it is sufficient to find a point $(t_0, x_0, \dot{x}_0)$ that satisfies (2.4). The problem is, as was seen in Example 2.1, the additional algebraic conditions that may appear when reformulating the problem. These have to be fulfilled as well. For general nonlinear DAE models, the reformulation is local and in order to do the reformulation, a consistent point is required. Hence, a Catch-22 situation occurs. The most common approach to resolve the problem, is to use some method that only relies on structural information about the equations, see, for instance, Pantelides (1988); Fritzson (2004). These methods are not ensured to give the correct answer for all models but in practice they usually work well. Other approaches are to use physical insight or the method derived in (Kunkel and Mehrmann, 2006, Remark 6.10). However, note that the obtained initial points are only possible consistent initial conditions since the definition above also requires that there should exist a solution given the initial point.

For state-space models, solvability follows if the system function F in (2.1) satisfies a Lipschitz condition, see Khalil (2002). For DAE models, solvability is more intrinsic since it is not obvious which parts are dynamical and which are not. Therefore, different concepts of index have been introduced to classify how difficult a DAE model is to solve both numerically and analytically. There are a number of different index concepts depending on which property is used to measure difficulty. The common property is however that a high index model in some sense is harder to solve than a model of lower index. Though, it is important to remember that 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. A few examples of indices are differential index, strangeness index, perturbation index etc. The discussion in this thesis will focus on the two first of them.

The differential index, which is the most common index concept, measures how different the given DAE model is from a state-space model. The strangeness index on the other hand, indicates how large the difference is between the DAE model and a set of decoupled differential equations and algebraic equations. The main motivation for the latter measure, is that not only differential equations are easy to know how to solve, but also pure algebraic equations. For example, consider the set of equations $Ax = b$, which can be shown to have differential equation one, but has a strangeness index zero.

In the preceding sections, the differential index and the strangeness index, will be discussed further. More information about different concepts of index can for example be found in Campbell and Gear (1995), Kunkel and Mehrmann (2006) and the references therein. Solvability in general are discussed in Kunkel and Mehrmann (2006); Brenan et al. (1996); Campbell and Griepentrog (1995).

2.3 The Differential Index

The differential index is the most common of the index concepts. It will also be this index which, 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 model. A small example showing the idea can be found below.

Example 2.2

Consider a DAE model given in semi-explicit form

$$\dot{x}_1 = F_1(x_1, x_2, u) \tag{2.11a}$$
$$0 = F_2(x_1, x_2, u) \tag{2.11b}$$

where $x_1 \in \mathbb{R}^d$, $x_2 \in \mathbb{R}^a$ and $u \in \mathbb{R}^p$. Assume that $u = u(t) \in \mathcal{C}^1$ is known. Differentiation of the constraint equations (2.11b) w.r.t. t yields

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

If $F_{2;x_2}(x_1, x_2, u)$ is nonsingular, it is possible to rewrite the model (2.11) as

$$\dot{x}_1 = F_1(x_1, x_2, u) \tag{2.12a}$$
$$\dot{x}_2 = -F_{2;x_2}(x_1, x_2, u)^{-1}\big(F_{2;x_1}(x_1, x_2, u)F_1(x_1, x_2, u) + F_{2;u}(x_1, x_2, u)\dot{u}\big) \tag{2.12b}$$

Here, $\dot{x}$ is determined as functions of x, u and $\dot{u}$ and since one differentiation was needed to get to a state-space model, the differential index of the original DAE is one.

Now assume that $F_{2;x_2}(x_1, x_2, u)$ is singular but using algebraic manipulations, the model

$$\begin{aligned}\dot{x}_1 &= F_1(x_1, x_2, u)\\ 0 &= F_{2;x_1}(x_1, x_2, u)\dot{x}_1 + F_{2;x_2}(x_1, x_2, u)\dot{x}_2 + F_{2;u}(x_1, x_2, u)\dot{u}\end{aligned}$$

can brought to the semi-explicit form (2.7) but with other x_1 and x_2 than in the first step. Then, if it is possible to solve for $\dot{x}_2$ after a second differentiation of the constraint equation, the original model is said to have 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 2.2. The differential index is the number of times that all or parts 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 all rows in the DAE model need not be differentiated the same number of times in the definition above. This could be seen in Example 2.2, where only the constraint equations (2.11b) were differentiated.

The introduced method to compute the index is rather intuitive. However, according to Brenan et al. (1996), this method cannot be used for all solvable DAE models. The problem is the coordinate transformation needed to obtain the semi-explicit form after each iteration, which not is ensured to exist. A more general definition of the index, without the state transformation, can be formulated using the derivative array. Consider a model in the form (2.2). The derivative array is defined as

$$F_j^d(t, x, \mathbf{x}_{j+1}, u, \dot{u}, \ldots, u^{(j)}) = \begin{pmatrix} F(t, x, \dot{x}, u) \\ \frac{d}{dt}F(t, x, \dot{x}, u) \\ \vdots \\ \frac{d^j}{dt^j}F(t, x, \dot{x}, u) \end{pmatrix} \tag{2.13}$$

where

$$\mathbf{x}_{j+1} = \left(\dot{x}, \ddot{x}, \ldots, x^{(j+1)}\right)$$

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

Definition 2.3. The index ν is the smallest positive integer such that $F_\nu^d = 0$ uniquely determines the variable $\dot{x}$ as a continuous function of x, t, u and higher derivatives of u, *i.e.*,

$$\dot{x} = \eta(t, x, u, \dot{u}, \ldots, u^\nu) \tag{2.14}$$

Note that u is here considered to be a known signal, in principle, possible to include in the time variability. If u is unknown beforehand, the conditions above need to be satisfied for arbitrary u. If u cannot be seen as a known signal, the DAE model will not have a unique solution and the differential index is then undefined. In that case, it is necessary to use the concept strangeness index, see the discussion in Section 2.4.

As mentioned earlier, the index is a measure of how difficult a DAE model is to handle. Both numerical computation of the solution, see Brenan et al. (1996), and derivation of control methods become more difficult for models with a high index. For the differential index there is a major leap in difficulty between models having index zero or one and models of higher index. Index zero models are ordinary differential equations (ODE) either in explicit or implicit form. Index one models need one differentiation to be transformed to a state-space model as could be seen in Example 2.2. For initial conditions such that $F_2\big(x_1(0), x_3(0), u(0)\big) = 0$, (2.12) is equivalent to the original model. Among other things, it means that the solution will satisfy $F_2\big(x_1(t), x_3(t), u(t)\big) = 0$ on the considered time interval. Hence, the constraints the solution will satisfy are given by F itself, or in other words, the solution manifold is described by the model.

In general, all constraints may not be visible in F, but may appear after differentiation as was seen in Example 2.1. There, the solution manifold is given by (2.3c)-(2.3e), where the two latter constraints are not visible in the original model. This type of constraints are denoted implicit constraints and occur only for higher index problems. The origin of their appearance is that the equations must be valid on a time interval $\mathbb{I}$. Together, the explicit and implicit constraints define the solution manifold, see Kunkel and Mehrmann (2006).

A typical case where high index models occur is mechanical systems modeled using multibody methods, see Arnold et al. (2004). The index is then often three. It is important

to note that for time-varying linear and nonlinear DAE models, 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 DAE models to reduce the index of the closed-loop system.

It can be interesting to note that the concept of differential index is also related to different concepts in the nonlinear theory developed for state-space models. One such case is the inversion problem where the objective is to find u in terms of y and possibly x for a model

$$\begin{aligned} \dot{x} &= f(x,u) \\ y &= h(x,u) \end{aligned} \tag{2.15}$$

where it is assumed that the number of inputs and outputs are the same. The procedures for inversion typically includes some differentiations of y, until u can be recovered from y and its derivatives w.r.t. t. The number of differentiations needed is normally called the relative degree or the order. However, for a given output signal y, the model (2.15) is a DAE model in (x,u). The corresponding index of this DAE model is one higher than the relative degree.

2.3.1 Reduction of the Differential-Index

It is known that for numerical solution of DAE models, many difficulties arise if standard discretization schemes for ODE models are applied to high index DAE models. These difficulties are due to the algebraic constraints and in particular to the implicit constraints. Also from a control perspective, it is often important to know the solution manifold in order to design the feedback law. Therefore, often the high index DAE is transformed into a lower index model, with index one or zero. This process is denoted index reduction, and the rewritten model should of course have the same solutions for consistent initial conditions. The key tool for reducing the index of a model and exposing the implicit constraints is differentiation. Index reduction procedures are often the same methods used either to compute the index of a model.

Since, the numerical solvers normally use index reduction methods in order to obtain a model of at most index one, it is a well-studied area, and more information can for instance be found in Mattson and Söderlind (1993); Brenan et al. (1996); Kunkel and Mehrmann (2006).

Linear Time-Invariant DAE Models

First, index reduction of linear time-invariant DAE models

$$E\dot{x} = Ax + Bu \tag{2.16}$$

is considered. One method is the so-called Shuffle algorithm, invented by Luenberger (1978). The procedure can be described as follows. Form the matrix $(E\ A\ B)$ and 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 DAE model

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

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

$$\underbrace{\begin{pmatrix} E_1 \\ -A_2 \end{pmatrix}}_{\bar{E}} \dot{x} = \underbrace{\begin{pmatrix} A_1 \\ 0 \end{pmatrix}}_{\bar{A}} x + \underbrace{\begin{pmatrix} B_1 \\ 0 \end{pmatrix}}_{\bar{B}_0} u + \underbrace{\begin{pmatrix} 0 \\ B_2 \end{pmatrix}}_{\bar{B}_1} \dot{u}$$

If $\bar{E}$ has full rank, the following state-space model is obtained by multiplying with the inverse of $\bar{E}$ from the left

$$\dot{x} = \bar{E}^{-1}\Big(\bar{A}x + \sum_{i=0}^{1} \bar{B}_i u^{(i)}\Big) \tag{2.17}$$

Here $u^{(i)}$ denotes the i:th order derivative of $u(t)$. Otherwise, the procedure is repeated. The procedure is guaranteed to terminate if and only if $\det(sE - A) \not\equiv 0$, see Dai (1989). Concerning solvability, regularity is a key property for a DAE model, and it is therefore formalized as a definition.

Definition 2.4. Consider a linear time-invariant DAE model

$$E\dot{x} = Ax + Bu$$

where E, A and B are constant matrices. It is called regular if

$$\det(sE - A) \not\equiv 0$$

that is, the determinant is not zero for all s.

The model (2.17) is equivalent to the original DAE in the sense that both give the same solution for consistent initial conditions. However, without considering the initial conditions, the solution manifold of (2.17) is much larger than for the original DAE model, since (2.17) is a state-space model with a solution for all initial conditions. It is also possible to derive another index reduced form, which has the advantage that the dynamical and the algebraic parts are separated. By first pre-multiplying (2.16) with a constant matrix P and introducing a linear coordinate change $z = Qx$, the linear DAE model (2.16) transforms into the canonical form

$$\dot{z}_1 = A_1 z_1 + B_1 u \tag{2.18a}$$

$$N\dot{z}_2 = z_2 + B_2 u \tag{2.18b}$$

where z_1 and z_2 are the dynamical and algebraic variables, respectively, and the matrix N is a nilpotent matrix, *i.e.*, $N^k = 0$ for some integer k. It can be proved that k is the index of the model, *i.e.*, $k = \nu$ (Brenan et al., 1996). The transformation is possible for all regular linear DAE models and a computational method can be found in Gerdin (2006).

The transformed model (2.18) will have the same index as the original DAE model, and therefore no information is lost. If an index reduced model is desired, the Shuffle-algorithm can be applied and the result is

$$\dot{z}_1 = A_1 z_1 + B_1 u \tag{2.19a}$$

$$z_2 = -\sum_{i=0}^{\nu-1} N^i B_2 u^{(i)} \tag{2.19b}$$

Here, it has been assumed that only consistent initial values $x(0)$ are considered. The form (2.19) is widely used to show different properties for linear time-invariant DAE models, see Dai (1989). Note that when the dynamical and algebraical parts of the original DAE model are separated like in (2.19), 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 control input and derivatives of it. Therefore, no drift off or problems due to discretization will occur, which is a major advantage with the form (2.19).

There are also results on a nonlinear version of the canonical form (2.18), see Rouchon et al. (1992).

Quasi-Linear DAE Models

The Shuffle algorithm can be extended to also deal with nonlinear models in quasi-linear form

$$E(t, x, u)\dot{x} = A(t, x, u)$$

Using pre-multiplication with a matrix function $P(t, x, u)$, the DAE can be reformulated as

$$\begin{pmatrix} E_1(t, x, u) \\ 0 \end{pmatrix} \dot{x} = \begin{pmatrix} A_1(t, x, u) \\ A_2(t, x, u) \end{pmatrix} x$$

Differentiation of the constraints gives the model

$$\underbrace{\begin{pmatrix} E_1(t, x, u) \\ -A_{2;x}(t, x, u) \end{pmatrix}}_{\bar{E}(t,x,u)} \dot{x} = \underbrace{\begin{pmatrix} A_1(t, x, u) \\ A_{2;t}(t, x, u) + A_{2;u}(t, x, u)\dot{u} \end{pmatrix}}_{\bar{A}(t,x,u,\dot{u})}$$

The matrix $\bar{E}$ is then pre-multiplied with a new matrix function $P(t, x, u)$ to obtain zeroes in the last rows, and the process is continued until the matrix $\bar{E}$ has full rank. For a more detailed analysis, see Steinbrecher (2006) or Tidefelt (2007).

General Nonlinear Models

For general nonlinear models, the Shuffle-algorithm may be impossible to perform and other methods based on the derivative array have to be used. Hence, after some symbolical differentiations and manipulations, the result is an state-space model

$$\dot{x} = \eta(x, u, \ldots, u^{(\nu-1)}) \tag{2.20}$$

As mentioned earlier, the state-space model (2.20) is equivalent to the original DAE model in the sense that given consistent initial conditions, the models will have the same solution. For Example 2.2, it means that (2.11) and (2.12) will have the same solution if the initial conditions satisfy

$$0 = F_2\big(x_1(0), x_3(0), u(0)\big)$$

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 end, the constraints can be used in different ways.

One approach is to let the explicit and implicit constraints define the initial conditions as mentioned above. That is, the initial condition $x(t_0)$ is assumed to belong to a set Ω_0 which consists of points satisfying all the constraints. This approach can be seen as a method to deal with the constraints implicitly. Another choice is to augment the DAE model with the constraints as the index reduction procedure proceeds. The result is then an overdetermined but well-defined index one DAE model, where well-defined means that the obtained DAE model will have a solution if the original model has one. Theoretically, the choices are equivalent. However, in numerical simulation they are not.

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 Ω_0, small errors in the numerical computations result in a solution to (2.20) that diverge from the solution of the original DAE model. The reason is the larger solution manifold of (2.20) compared to the original model. A solution to this problem is to use methods known as constraint stabilization techniques (Baumgarte, 1972; Ascher et al., 1994).

For the second approach, the solution manifold is the same as for the original DAE model. However, the numerical solver discretizes the problem and according to Mattson and Söderlind (1993), an algebraically feasible point in the original DAE may then be non-feasible in the discretized problem and vice versa. To solve this problem special projection methods have been derived, see for instance the references in Mattson and Söderlind (1993).

The problem with non-feasible 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 model which instead of being overdetermined becomes well-determined. The discretized problem will then be well-determined as well.

2.4 The Strangeness Index

Another index concept is the strangeness index μ, see Kunkel and Mehrmann (2006). The main difference compared to the differential index is that the DAE model is compared with a model which is allowed to consist of one part that is an ODE and one part that is a set of purely algebraic equations. To motivate this definition consider the following example.

Example 2.3

Consider the model of a chemical reactor

$$\dot{c} = k_1(c_0 - c) - R \tag{2.21a}$$

$$\dot{T} = k_1(T_0 - T) + k_2 R - k_3(T - T_c) \tag{2.21b}$$

$$0 = R - k_3 c e^{-\frac{k_4}{T}} \tag{2.21c}$$

where c is the concentration, T is the temperature, R is the reaction rate per unit volume, c_0 is the given feed reactant concentration, T_0 is the initial temperature, k_1 to k_4 are constants and T_c is the cooling temperature which also is used as control input. Hence, the variables in this model is $x = (c, T, R)$. It can be shown that if (2.21c) is differentiated w.r.t. t once, the obtained equation can be solved for $\dot{R}$ in (c, T, R). Hence, one differentiation was required to obtain a state-space model and the differential index of the original DAE model is therefore one. However, by solving for R directly, the following model is obtained.

$$\begin{aligned}
\dot{c} &= k_1(c_0 - c) - k_3 e^{-\frac{k_4}{T}} c \\
\dot{T} &= k_1(T_0 - T) + k_2 k_3 e^{-\frac{k_4}{T}} c - k_3(T - T_c) \\
R &= k_3 c e^{-\frac{k_4}{T}}
\end{aligned}$$

Here, the model consists of a decoupled set of differential equations and algebraic equations. Since no differentiation was needed, the strangeness index is zero. This model is of course as simple, or even simpler, to solve than the index reduced model. Hence, it makes little sense to let purely algebraic equations raise the difficulty factor, *i.e.*, the index.

In the derivation of the strangeness index a central concept is invariance. The main motivation for this is that a good measure of difficulty should not be too sensitive to reformulations of the model such as variable changes and pre-multiplications with given matrix functions, see (Kunkel and Mehrmann, 2006, pp.157 – 159,182). However, also the differential index is invariant under certain reformulations, but not as generally as the strangeness index, see (Brenan et al., 1996, pp. 33). The strangeness index does also generalize the differential index in the sense that, unlike the differential index, the strangeness index is defined for over- and underdetermined DAE models. An overdetermined model is a model where the number of equations m is larger than the number of unknown variables, while the opposite holds for an underdetermined model. Normally, the unknown variables are x, but if a behavioral approach is considered, u can also be included among the unknowns. Finally, for models 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, 2006).

In the following, conditions under which a solution to (2.4) exists and is unique according to the definition in Section 2.2 will be derived. The results come from Kunkel and Mehrmann (2006) and two of their key proofs are included below. The main reason for

including the proofs is that they reveal the underlying principles of how to deal with DAE models. The main element in their theorems is a hypothesis. The hypothesis is investigated on the solution set of the derivative array (2.13) for some integer μ. The solution set is denoted $\mathbb{L}_\mu$ and is described by

$$\mathbb{L}_\mu = \{z_\mu \in \mathbb{I} \times \underbrace{\mathbb{R}^n \times \ldots \times \mathbb{R}^n}_{\mu+2} \mid F_\mu^d(z_\mu) = 0\} \tag{2.22}$$

while the hypothesis is as follows.

Hypothesis 2.1. Consider the general nonlinear DAE model (2.4), *i.e.*,

$$F(t, x, \dot{x}) = 0$$

There exist integers μ, r, a, d and v such that $\mathbb{L}_\mu$ is nonempty and such that for every $z_{\mu,0} = (t_0, x_0, \dot{x}_0, \ldots, x_0^{\mu+1}) \in \mathbb{L}_\mu$, there exists a neighborhood in which the following properties hold:

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

2. It holds that
$$\operatorname{rank} F^d_{\mu;x,\mathbf{x}_{\mu+1}} = r \tag{2.23}$$
on $\mathbb{L}_\mu$, where $\mathbf{x}_{\mu+1} = \left(\dot{x}, \ddot{x}, \ldots, x^{(\mu+1)}\right)$.

3. It holds that
$$\operatorname{corank} F^d_{\mu;x,\mathbf{x}_{\mu+1}} - \operatorname{corank} F^d_{\mu-1;x,\mathbf{x}_\mu} = v \tag{2.24}$$
on $\mathbb{L}_\mu$. Here the convention that $\operatorname{corank} F^d_{-1;x} = 0$ is used. (For a definition of the corank, see Appendix A)

4. It holds that
$$\operatorname{rank} F^d_{\mu;\mathbf{x}_{\mu+1}} = r - a \tag{2.25}$$
on $\mathbb{L}_\mu$ such that there are smooth matrix functions of pointwise full rank Z_2 and T_2 defined on $\mathbb{L}_\mu$ of size $\left((\mu+1)m, a\right)$ and $(n, n-a)$, respectively, satisfying
$$Z_2^T F^d_{\mu;\mathbf{x}_{\mu+1}} = 0, \qquad \operatorname{rank} Z_2^T F^d_{\mu;x} = a, \qquad Z_2^T F^d_{\mu;x} T_2 = 0 \tag{2.26}$$
on $\mathbb{L}_\mu$.

5. It holds that
$$\operatorname{rank} F^d_{\dot{x}} T_2 = d = m - a - v \tag{2.27}$$
on $\mathbb{L}_\mu$, such that there exists a smooth matrix function Z_1 of size $n \times d$ of pointwise full rank satisfying
$$\operatorname{rank} Z_1^T F_{\dot{x}} T_2 = d$$

Note that the different ranks appearing in the hypothesis are assumed to be constant on the manifold $\mathbb{L}_\mu$. If the hypothesis is not satisfied for a given μ, *i.e.*, if $d \neq m - a - v$, μ is increased by one and the procedure is repeated. However, it is not certain that a μ exists such that the hypothesis hold. The strangeness index is then defined as follows.

Definition 2.5. The strangeness index of (2.4) is the smallest positive integer μ such that Hypothesis 2.1 is satisfied. A model with $\mu = 0$ is denoted strangeness-free.

If there exist μ, d, a and v such that the hypothesis above is satisfied, it will imply that the system can be reduced to a model 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. The quantity v measures the number of equations in the original model (2.4) 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 model are then characterized, since $m = a + d + v$.

2.4.1 Derivation of the Reduced Model

The analysis of the implications of the hypothesis is based on the implicit function theorem. The analysis is therefore local and performed in a neighborhood of the point $z_{\mu,0} \in \mathbb{L}_\mu$. It is important to note that the variables $x_0^{(j)}$ for $j \geq 1$ in this case are seen as algebraic variables rather than as derivatives of x_0. This also means that the derivative array should be seen as formally derived, that is, $\dot{x}, \ddot{x}$ are only formal derivatives of x.

From part 1 of the hypothesis, it is known that $\mathbb{L}_\mu$ is a $(\mu + 2)n + 1 - r$ dimensional manifold. It is therefore possible to locally parameterize it using $(\mu + 2)n + 1 - r$ parameters. These parameters can be chosen from $(t, x, \mathbf{x}_{\mu+1})$ such that the rank of $F^d_{\mu;x,\mathbf{x}_{\mu+1}}(z_{\mu,0})$ is unchanged if the corresponding columns are removed. Together, parts 1 and 2 of the hypothesis give that

$$\operatorname{rank} F^d_{\mu;t,x,\mathbf{x}_{\mu+1}} = \operatorname{rank} F^d_{\mu;x,\mathbf{x}_{\mu+1}} = r$$

and therefore can t be chosen as parameter. From part 2, it follows that r variables of $(\dot{x}, \ddot{x}, \ldots, x^{(\mu+1)})$ are determined (via the implicit function theorem) by the other $(\mu + 2)n + 1 - r$ variables. From part 4, it is also known that $r - a$ of those comes from $(\dot{x}, \ddot{x}, \ldots, x^{(\mu+1)})$. These variables are denoted x_h, while the rest of $(\dot{x}, \ddot{x}, \ldots, x^{(\mu+1)})$ 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 $r - a$ of these belong to $(\dot{x}, \ddot{x}, \ldots, x^{(\mu+1)})$, the other $r - (r - a) = a$ determined variables, denoted x_3, must belong to x. Using part 4, it follows that $Z_2^T F_{\mu;x_3}$ must be nonsingular. The rest of x must then be parameters and are denoted $(x_1, x_2) \in \mathbb{R}^{n-a}$.

Hence, using the implicit function theorem (see Theorem A.1), Hypothesis 2.1 implies the existence of a neighborhood $\mathbb{V} \subseteq \mathbb{R}^{(\mu+2)n+1-r}$ of $(t_0, x_{1,0}, x_{2,0}, p_0)$, which is the part of $z_{\mu,0}$ corresponding to the selected parameters in (t, x_1, x_2, p), and a neighborhood $\tilde{\mathbb{U}} \subseteq \mathbb{R}^{(\mu+2)n+1}$ of $z_{\mu,0}$ such that

$$\mathbb{U} = \mathbb{L}_\mu \cap \tilde{\mathbb{U}} = \left\{\theta(t, x_1, x_2, p) \mid (t, x_1, x_2, p) \in \mathbb{V}\right\}$$

where $\theta : \mathbb{V} \to \mathbb{U}$ is a diffeomorphism. From this expression, it follows that $F^d_\mu(z_\mu) = 0$ if and only if $z_\mu = \theta(t, x_1, x_2, p)$ for some $(t, x_1, x_2, p) \in \mathbb{V}$. More specifically, x_3 and x_h are possible to express as

$$x_3 = \mathcal{G}(t, x_1, x_2, p) \tag{2.28}$$

$$x_h = \mathcal{H}(t, x_1, x_2, p) \tag{2.29}$$

on $\mathbb{V}$ and the equation defining the manifold $\mathbb{L}_\mu$ can be rewritten as

$$F_\mu^d\big(t, x_1, x_2, \mathcal{G}(t, x_1, x_2, p), \mathcal{H}(t, x_1, x_2, p)\big) \equiv 0 \tag{2.30}$$

on $\mathbb{V}$, where $(x_h, p) = \mathcal{H}(t, x_1, x_2, p)$.

The next step is to show that locally on $\mathbb{V}$, the implicit function $\mathcal{G}$ will only depend on x_1, x_2 and t. That is, $\mathcal{G}$ will be independent of p. Therefore, the derivative of the following expression w.r.t. p is used.

$$\begin{aligned}\frac{d}{dp}(Z_2^T F_\mu^d) &= \big(Z_{2;x_3}^T F_\mu^d + Z_2^T F_{\mu;x_3}^d\big)\mathcal{G}_p + \big(Z_{2;\mathbf{x}_{\mu+1}}^T F_\mu^d + Z_2^T F_{\mu;\mathbf{x}_{\mu+1}}^d\big)\mathcal{H}_p \\ &= Z_2^T F_{\mu;x_3}^d \mathcal{G}_p = 0\end{aligned}$$

for all $(t, x_1, x_2, p) \in \mathbb{V}$. Here, it has been used that $F_\mu^d \equiv 0$, locally, and that $Z_2^T F_{\mu;\mathbf{x}_{\mu+1}}^d = 0$ on $\mathbb{V}$. By construction, the variables x_3 were chosen such that $Z_2^T F_{\mu;x_3}^d$ is nonsingular. Hence

$$\mathcal{G}_p(t, x_1, x_2, p) = 0$$

on $\mathbb{V}$. The function $\mathcal{G}_p$ is therefore constant with respect to p, and locally there exists a function $\mathcal{R}$ such that

$$\mathcal{R}(t, x_1, x_2) = \mathcal{G}(t, x_1, x_2, p_0)$$

Using the function $\mathcal{R}$, (2.28) can be rewritten as

$$x_3 = \mathcal{R}(t, x_1, x_2) \tag{2.31}$$

and the conclusion is that x_3 is independent of derivatives of x on $\mathbb{V}$, since x_1 and x_2 only consist of terms in x. Hence, it follows that

$$F_\mu^d\big(t, x_1, x_2, \mathcal{R}(t, x_1, x_2), \mathcal{H}(t, \bar{x}, p)\big) \equiv 0$$

on $\mathbb{V}$ and if this expression is differentiated w.r.t. (x_1, x_2), the result is

$$\begin{aligned}\frac{d}{d(x_1, x_2)}(Z_2^T F_\mu^d) &= Z_{2;x_1,x_2}^T F_\mu^d + Z_2^T F_{\mu;x_1,x_2}^d + \big(Z_{2;x_3}^T F_\mu^d + Z_2^T F_{\mu;x_3}^d\big)\mathcal{R}_{x_1,x_2} \\ &\quad + \big(Z_{2;\mathbf{x}_{\mu+1}}^T F_\mu^d + Z_2^T F_{\mu;\mathbf{x}_{\mu+1}}^d\big)\mathcal{H}_{x_1,x_2} \\ &= Z_2^T F_{\mu;x_1,x_2}^d + Z_2^T F_{\mu;x_3}^d \mathcal{R}_{x_1,x_2} = Z_2^T F_{\mu;x}^d \begin{pmatrix} I_{n-a} \\ \mathcal{R}_{x_1,x_2} \end{pmatrix} = 0\end{aligned} \tag{2.32}$$

on $\mathbb{V}$. Here I_{n-a} is an identity matrix of dimension $n-a \times n-a$ and again $F_\mu^d \equiv 0$ and $Z_2^T F_{\mu;\mathbf{x}_{\mu+1}}^d = 0$ have been used. In part 4 of the hypothesis one requirement was the existence of a function T_2 such that $Z_2 F_{\mu;x}^d T_2 = 0$. Using the result in (2.32), it is possible to choose T_2 as

$$T_2(t, x_1, x_2) = \begin{pmatrix} I_{n-a} \\ \mathcal{R}(t, x_1, x_2) \end{pmatrix}$$

The matrix function Z_1 will only depend on (t, x_1, x_2) since both T_2 and F depend on these variables only. However, since the results here are local in their nature and x and $\dot{x}$ are continuous, Z_1 can even be chosen constant. Now define the model

$$\hat{F}_1(t, x, \dot{x}) = Z_1^T F(t, x, \dot{x}) \tag{2.33a}$$

$$\hat{F}_2(t, x) = Z_2^T F_\mu\big(t, x_1, x_2, x_3, \mathcal{H}(t, x_1, x_2, p_0)\big) \tag{2.33b}$$

Notice that p is here chosen constant since it was shown that it did not influence x_3 locally. The so-called reduced differential-algebraic equation is then

$$\hat{F}(t, x, \dot{x}) = \begin{pmatrix} \hat{F}_1(t, x, \dot{x}) \\ \hat{F}_2(t, x) \end{pmatrix} = 0 \tag{2.34}$$

The reduced differential-algebraic equation can be shown to satisfy Hypothesis 2.1 with $\mu = 0$, d and a. Hence, this is the model that is ensured to exist if the hypothesis is satisfied for the original DAE model. Furthermore, $\hat{F}_2$ is known to be solvable for x_3 yielding (2.31).

If the original DAE has a continuously differentiable solution, *i.e.*, $x(t) \in \mathcal{C}^1$, it is possible to differentiate (2.31) w.r.t. t and eliminate x_3 and $\dot{x}_3$ from $\hat{F}_1$ to obtain

$$\hat{F}_1(t, x_1, x_2, \mathcal{R}, \dot{x}_1, \dot{x}_2, \mathcal{R}_{x_1}\dot{x}_1 + \mathcal{R}_{x_2}\dot{x}_2 + \mathcal{R}_t) = 0$$

on $\mathbb{V}$. Differentiating the expressions for $\hat{F}_1$ w.r.t. $(\dot{x}_1, \dot{x}_2)$ yields

$$\frac{d}{d(\dot{x}_1, \dot{x}_2)}\hat{F}_1 = Z_1^T F_{\dot{x}_1, \dot{x}_2} + Z_1^T F_{\dot{x}_3}\mathcal{R}_{x_1, x_2} = Z_1^T F_{\dot{x}} T_2$$

on $\mathbb{V}$. From part 5 it is known that $\operatorname{rank} Z_1 F_{\dot{x}} T_2 = d$ and therefore d variables of $(\dot{x}_1, \dot{x}_2)$, in this case chosen as $\dot{x}_1$, are determined as a function of the other variables in $\hat{F}_1$. The other variables $\dot{x}_2$, become parameters. It means that x_1 become states, while x_2 continue to be just parameters.

Summarizing, it is known that at least locally it is possible to solve the reduced model (2.34) for $\dot{x}_1$ and x_3 yielding the model

$$\begin{aligned} \dot{x}_1 &= \mathcal{L}(t, x_1, x_2, \dot{x}_2) \\ x_3 &= \mathcal{R}(t, x_1, x_2) \end{aligned} \tag{2.35}$$

From the reasoning above, a theorem can be formulated that describes when the solution to a nonlinear DAE model (2.4) is a solution to the reduced model (2.35) as well.

Theorem 2.1
Let F in (2.4) be sufficiently smooth and satisfy Hypothesis 2.1 with some μ, a, d and v. Then every solution of (2.4) also solves the reduced problem (2.35) consisting of d differential equations and a algebraic equations.

Proof: This theorem follows immediately from the procedure above, see Theorem 4.11 in Kunkel and Mehrmann (2006). □

The procedure described has the advantage that it defines a constructive method to compute the reduced model. However, the algorithm involves rank tests and those can be rather sensitive numerically. Another important observation is given in Remark 4.15 in (Kunkel and Mehrmann, 2006, pp. 167).

Remark 2.1. In the derivation $\ddot{x}$, $\dddot{x}$ etc. appear. Therefore, it may seem like the solution is required to be smoother than continuously differentiable. However, it can be shown that for all solutions $x(t) \in \mathcal{R}^1(\mathbb{I}, \mathbb{R}^n)$ to the original DAE model, it is possible to locally prescribe a function $\mathcal{P} \in C(\mathbb{I}, R^{(\mu+1)n})$ with $\mathcal{P}[I_n\, 0, \ldots, 0] = \dot{x}(t)$ such that $F_\mu\big(t, x(t), \mathcal{P}(t)\big) \in \mathbb{L}_\mu$. These functions can then be composed to yield a global continuous parametrization $\mathcal{P}(t)$. The solutions to the original DAE is then defined as those that obtain a continuous $\mathcal{P}(t)$.

2.4.2 The Solution of the Reduced DAE Model

The next step is to show that the solution to (2.34) is also a solution to the original DAE model, locally. First consider two problems that may occur. One problem may occur, when only a point $z_{\mu,0} \in \mathbb{L}_\mu$ is known (and not the whole solution to the DAE model). Then, it is still possible to compute the reduced model (2.34), but it might fail to have a solution. A second problem is that the solution to reduced model may not solve the original DAE model. If the reduced model has a solution $x(t) \in \mathcal{C}$, it follows that x_3 is a determined variable and $\dot{x}_3$ is then the derivative of $\mathcal{R}$ w.r.t. t. Having these expressions, it is also known that $\dot{x}_1$ must satisfy the equation in (2.35). However, for the original DAE model it could be the case that $\dot{x}_1$ and $\dot{x}_3$ should be parameters (included in p) instead of being determined variables. That is, there could be a conflict in which variables that should be dynamic variables and which should not. To ensure that such situations do not happen, it is assumed that the hypothesis is satisfied for $\mu+1$ with the same d and a as for μ. From this fact, it is shown in Kunkel and Mehrmann (2006) that the following relations can be established algebraically

$$\begin{aligned} x_3 &= \mathcal{R}(t, x_1, x_3) \\ \dot{x}_3 &= \mathcal{R}_t x_1(t, x_1, x_2) + \mathcal{R}_{x_1}(t, x_1, x_2)\dot{x}_1 + \mathcal{R}_{x_2}(t, x_1, x_2)\dot{x}_2 \end{aligned}$$

and with these two expressions, it is known that $\hat{F}_1$ can be solved for $\dot{x}_1$ as

$$\dot{x}_1 = \mathcal{L}(t, x_1, x_2, \dot{x}_2)$$

Hence, neither $\dot{x}_1$ nor $\dot{x}_3$ can be part of the parameters, and the procedure to construct the solution can now be formulated as follows.

Choose $x_2 = x_2(t)$ and $\dot{x}_2 = \dot{x}_2(t)$. Let $p = p(t)$ be arbitrary but smooth and consistent with the choice of $\dot{x}_2$ and the initial values $z_{\mu,0}$. Then, if $x_1 = x_1(t)$ and $x_3 = x_3(t)$ are chosen as the solution to

$$\begin{aligned} \dot{x}_1 &= \mathcal{L}\big(t, x_1, x_2(t), \dot{x}_2(t)\big), \quad x_1(t_0) = x_{1,0} \\ x_3 &= \mathcal{R}\big(t, x_1, x_2(t)\big) \end{aligned}$$

the equations $F^d_{\mu+1} = 0$ will be satisfied for all t in a neighborhood of t_0 and therefore $F = 0$. That is, the solution fulfills the original DAE model as well.

Hence, the following theorem can be formulated.

Theorem 2.2
Let F in (2.4) be sufficiently smooth and satisfy Hypothesis 2.1 with some μ, 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 (2.30) for $F_{\mu+1}$ include $\dot{x}_2$. Then for every function $x_2 \in \mathcal{C}^1(\mathbb{I}, \mathbb{R}^{n-a-d})$ with $x_2(t_0) = x_{p,0}$, $\dot{x}_p(t_0) = \dot{x}_{p,0}$, the reduced model (2.35) has unique solutions x_1 and x_3 satisfying $x_1(t_0) = x_{1,0}$. Moreover, together these solutions solve the original problem locally.

Proof: See Theorem 4.13 in Kunkel and Mehrmann (2006). □

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 $m = n$. Then Theorem 2.2 can be simplified since no free parameters x_2 will occur.

Corollary 2.1
Let F in (2.2) be sufficiently smooth and satisfy Hypothesis 2.1 with μ, a, d and $v = 0$ and assume that $a+d = n$. Furthermore, assume that $\mu+1$ yields the same μ, a, d and $v = 0$. For every $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$, the reduced problem (2.35) 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 2.2. Sometimes it is interesting to consider solvability only on a part of the manifold defined by

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

This is possible if $\mathbb{L}_\mu$ instead is defined as

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

where

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

and $\Omega_{x^{(i)}}$ are open sets. That is, the region on which each variable is defined is not the whole $\mathbb{R}^n$. However, if the regions are chosen inappropriate, $\mathbb{L}_\mu$ may be the empty set.

Linear Time-Invariant DAE Models

For a square linear time-invariant DAE model, *i.e.*, a model with 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 $sE - A$, i.e., $\det(sE - A) \not\equiv 0$*, and a given control signal $u \in \mathcal{C}^\nu(\mathbb{I}, \mathbb{R}^p)$. Then the model is solvable and every consistent initial condition yield a unique solution.*

Proof: See Kunkel and Mehrmann (1994) or Kunkel and Mehrmann (2006). □

So far, only solutions in classical meaning have been considered. This is a rather standard and for state-space models (2.1) with a system matrix F smooth enough, it will basically impose the control input to be continuous. In Theorem 2.3, the control input is required to be ν times continuously differentiable. This is motivated by the discussion in Section 2.3.1, where it was shown that in general, the solution may depend on derivatives up to order ν. In Section 2.4.4, it will be shown that if no derivative of order higher than k appear, it is enough to require the input to be $k+1$ times continuously differentiable. However, for linear DAE models a more general definition of a solution can be made. The solution is then defined in a distributional sense, see Dai (1989); Kunkel and Mehrmann (2006). In the distributional framework, a distributional solution exists 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).

2.4.3 Reduction of the Strangeness Index

The procedure presented in Section 2.4 for defining solvability of DAE models is also a method that can be used to reduce the strangeness index. If μ, d, a and v are found such that Hypothesis 2.1 is satisfied, it was proved the original model can be expressed in the form

$$\hat{F}_1(t, x_1, x_3, x_2, \dot{x}_1, \dot{x}_2, \dot{x}_3) = 0 \tag{2.36a}$$

$$\hat{F}_2(t, x_1, x_3, x_2) = 0 \tag{2.36b}$$

which is strangeness-free, and if the hypothesis is satisfied for $\mu + 1$ as well, it can be solved for $\dot{x}_1$ and x_3 to obtain

$$\begin{aligned} \dot{x}_1 &= \mathcal{L}(t, x_1, x_2, \dot{x}_2) \\ x_3 &= \mathcal{R}(t, x_1, x_2) \end{aligned}$$

Unfortunately, the functions $\mathcal{L}$ and $\mathcal{R}$ are defined by the implicit function theorem and may be impossible to write in closed form. On the hand, this is most often possible for the functions $\hat{F}_1$ and $\hat{F}_2$ using the system function F and possibly its derivatives (this is the purpose of the matrix functions Z_1 and Z_2 in Hypothesis 2.1). Practical aspects of the method derived by Kunkel and Mehrmann can be found in Kunkel and Mehrmann (2004) and in Arnold et al. (2004).

Note that for given μ, 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.

2.4.4 The Strangeness Index for Models with External Inputs

Until now, external inputs have not been considered explicitly. External inputs can be included in Hypothesis 2.1 either by using a behavioral approach, *i.e.*, to concatenate x

and u into one variable x, or by treating them as part of the time-variability. The latter assumes that the signals are known and does not fit our purposes, since the control inputs (or other external signals) most often are unknown at the time when the index reduced model is derived. The first approach may work if x_1, x_3 and x_2 can be chosen properly, but what does properly mean?

For control problems (2.2), it is common that the physical plant defines which variables are possible to use for controlling the system. It means that the control inputs should appear as parameters x_2 in the analysis in Section 2.4. If this happen, the standard procedure is applicable with x and u concatenated. However, as will be seen in the following example, this might not be the case. Instead, the control inputs may be categorized as algebraic variables. That is, as variables determined by the other variables.

Example 2.4

Consider a linear time-invariant DAE model. If the control signal is included in the x variable, *i.e.*, $x = (z_1,\, z_2,\, z_3,\, u)^T$, the model can be written as

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} 2 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \end{pmatrix} x$$

The result from Kunkel and Mehrmann's index reduction procedure is, with $\mu = 0$, the reduced model

$$\begin{aligned} \dot{z}_1 &= 2z_1 + u \\ \dot{z}_3 &= z_2 + 2u \\ u &= -z_3 \end{aligned}$$

Hence, the control input is seen as an algebraic variable which is determined by z_3, while the free parameter is z_2.

In the next computation u is instead seen as a known signal, which is included as part of the time variability. Hence, $x = (z_1,\, z_2,\, z_3)$, and the model can be written as

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} x + \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} u$$

After the index reduction procedure has been applied, the outcome is the following model

$$\begin{aligned} \dot{z}_1 &= 2z_1 + u \\ z_2 &= -2u - 3\dot{u} \\ z_3 &= -u \end{aligned}$$

and in this case, the strangeness index was one. Hence, by considering u as a known 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, the hypothesis needs to be slightly modified to better fit our conditions. Consider a DAE model with an external input

$$F(\dot{x}, x, u, t) = 0 \tag{2.37}$$

with the corresponding solution set

$$\begin{aligned}&\mathbb{L}_\mu = \\ &\{(t, x, \mathbf{x}_{\mu+1}, u, \mathbf{u}_\mu) \in \underbrace{\mathbb{R}^n \times \ldots \times \mathbb{R}^n}_{\mu+2} \times \underbrace{\mathbb{R}^p \times \ldots \times \mathbb{R}^p}_{\mu+1} \mid F_\mu^d(t, x, \mathbf{x}_{\mu+1}, u, \mathbf{u}_\mu) = 0\}\end{aligned} \tag{2.38}$$

where $\mathbf{x}_{\mu+1} = (\dot{x}, \ldots, x^{(\mu+1)})$ and $\mathbf{u}_\mu = (\dot{u}, \ldots, u^{(\mu)})$.

Then the modified hypothesis can be formulated as follows.

Hypothesis 2.2. The DAE model (2.37) satisfies Hypothesis 2.1 with the modifications.

1. The set $\mathbb{L}_\mu$ is given by (2.38) and forms a manifold of dimension $(\mu+2)n + 1 - r + (\mu+1)p$.
2. The rank conditions are satisfied without considering the input, *i.e.*, no partial differentiations w.r.t. $(u, \dot{u}, \ldots, u^{\mu+1})$ should be included.

That is, except for that $\mathbb{L}_\mu$ is given by (2.38), all conditions are unchanged. The extended hypothesis implies that u and its derivatives are possible to choose as parameters, but different parameters than p, since it is not possible to show that $\mathcal{R}$ is independent of them.

On the contrary, if the DAE model satisfies Hypothesis 2.2 for μ, d, a and v, the functions $\hat{F}_1$ and $\hat{F}_2$ in (2.36) become

$$\hat{F}_1(t, x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3, u) = 0 \tag{2.39a}$$

$$\hat{F}_2(t, x_1, x_2, x_3, u, \dot{u}, \ldots, u^{(\mu)}) = 0 \tag{2.39b}$$

Similarly to the standard case, it is possible to show that for a solution to the original DAE model, or if the hypothesis is satisfied with μ increased by one, the reduced DAE model (2.39) can be solved for $\dot{x}_1$ and x_3 as

$$\dot{x}_1 = \mathcal{L}(t, x_1, x_2, \dot{x}_2, u, \ldots, u^{(\mu+1)}) \tag{2.40a}$$

$$x_3 = \mathcal{R}(t, x_1, x_2, u, \dot{u}, \ldots, u^{(\mu)}) \tag{2.40b}$$

and the other way around. In the expressions above, it is necessary to require u to be sufficiently smooth, namely $\mu + 1$ times continuously differentiable, in order to have a well-defined solution. Furthermore, this assumption is necessary to be able to choose Z_1 constant.

In later chapters, models for which $\hat{F}_2$ is assumed independent of derivatives higher than k, *i.e.*,

$$\hat{F}_2(t, x, u, \dot{u}, \ldots, u^{(k)}) = 0$$

where $k \leq \mu$ will be studied. In this case, the smoothness requirement on u can be relaxed, which might not be obvious since $\ddot{u}$, $\dddot{u}$ etc. may still turn up in the analysis. This property can be shown, using the same idea as used to show that $x(t)$ need not be smoother than continuously differentiable, despite that higher derivatives of x appear in the derivative array.

For this end, consider all derivatives as formal, and do the analysis. If Hypothesis 2.2 is satisfied for μ with d, a and v, it will be possible to derive a reduced model (2.39) with Z_1 chosen constant. The fact that Z_1 can be chosen constant depends on that T_2 can be chosen as in (2.32) and will only depend on $\mathcal{R}(t, x_1, x_2, u, \ldots, u^{(k)})$ which is a continuous function. Hence $F_{\dot{x}} T_2$ will be a continuous function. If the DAE model (2.37) also satisfies Hypothesis 2.2 for $\mu + 1$ with d, a and v, it is possible to show, using a similar approach as in Section 2.4.2, that algebraically

$$\dot{x}_1 = \mathcal{L}(t, x_1, x_2, \dot{x}_2, u, \ldots, u^{(k+1)}) \tag{2.41a}$$

$$x_3 = \mathcal{R}(t, x_1, x_2, u, \dot{u}, \ldots, u^{(k)}) \tag{2.41b}$$

$$\dot{x}_3 = \mathcal{R}_t + \mathcal{R}_{x_1}\dot{x}_1 + \mathcal{R}_{x_2}\dot{x}_2 + \mathcal{R}_u \dot{u} + \ldots + \mathcal{R}_{u^{(k)}} u^{(k+1)} \tag{2.41c}$$

Hence, $\dot{x}_1$ and $\dot{x}_3$ are not part of the parametrization. Let $x_2 = x_2(t)$, $\dot{x}_2 = \dot{x}_2(t)$ and $\big(u, \dot{u}, \ldots, u^{(k)}\big) = \big(u(t), \dot{u}(t), \ldots, u^{(k)}(t)\big)$. Further, let $p = p(t)$ and $\big(u^{(k+1)}(t), \ldots, u^{(\mu+1)}(t)\big)$ be arbitrary but smooth and consistent with $\mathbb{L}_{\mu+1}$. Finally, let $x_1 = x_1(t)$ and $x_3 = x_3(t)$ be the solution to (2.41). This solution will then satisfy $F_{\mu+1} \equiv 0$ for t in a neighborhood of t_0, and then specifically

$$F\big(t, x(t), \dot{x}(t), u(t)\big) \equiv 0$$

Therefore, if the only of u up to $u^{(k)}$ appear, the smoothness requirements are reduced.

So far in this thesis, the model has had both x_2 and control inputs for generality. However, in the sequel the models are assumed to be well-determined except for the external inputs. Therefore, the parameters x_2 and $\dot{x}_2$ are left out. In some parts of this thesis, the model (2.39) is also assumed to be in semi-explicit form with system functions F_1 and F_2 given in closed form. That is, it should satisfy the following assumption.

Assumption A1. The variables $\dot{x}_1$ can be solved from (2.39a) to give

$$\dot{x}_1 = \tilde{F}_1(x_1, x_3, u, \ldots, u^{(k+1)}) \tag{2.42a}$$

$$0 = \tilde{F}_2(x_1, x_3, u, \dot{u}, \ldots, u^{(k)}) \tag{2.42b}$$

where $\tilde{F}_1$ and $\tilde{F}_2$ are possible to express in closed form.

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 $\dot{x}_3$ as

$$\dot{x}_3 = -\tilde{F}_{2;x_3}^{-1}(x_1, x_3, \mathbf{u})\big(\tilde{F}_{2;x_1}(x_1, x_3, \mathbf{u})\dot{x}_1 + \tilde{F}_{2;\mathbf{u}}(x_1, x_3, \mathbf{u})\dot{\mathbf{u}}\big)$$

where $\mathbf{u} = (u, \dot{u}, \ldots, u^{(k)})$. Using this expression, $\dot{x}_3$ 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. One such example is mechanical multibody systems which often can be written in this form, see

Kunkel and Mehrmann (2001). For control methods, it is most often undesirable to have derivatives of the control variable in the equations. One approach which can be used to avoid this is to redefine the control signal as its highest derivative and introduce an integrator chain

$$\begin{aligned} \dot{x}_{1,d+1} &= x_{1,d+2} \\ &\vdots \\ \dot{x}_{1,d+k+1} &= u^{(k+1)} \end{aligned}$$

If the integrator chain is included, the model (2.42) becomes

$$\dot{x}_1 = F_1(x_1, x_3, u) \tag{2.43a}$$
$$0 = F_2(x_1, x_3, u) \tag{2.43b}$$

where $x_1 \in \mathbb{R}^{d+(k+1)p}$, $x_3 \in \mathbb{R}^a$ and $u \in \mathbb{R}^p$. Here $u^{(k+1)}$ is denoted u in order to notationally match the sequel of this thesis.

To illustrate the method described above an example is presented.

Example 2.5

Consider a system described by the semi-explicit model (2.7) with $F_1(x_1, x_3, u) \in \mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}^a \times \mathbb{R}^p, \mathbb{R}^d)$ and $F_2(x_1, x_3, u) \in \mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}^a \times \mathbb{R}^p, \mathbb{R}^a)$. Define the set

$$\Omega = \{x_1 \in \Omega_x,\, x_3 \in \mathbb{R}^a,\, u \in \mathbb{R}^p \mid F_2(x_1, x_3, u) = 0\}$$

where Ω_x is open, and assume that Ω is nonempty. Note that Ω may not be an open set. Moreover, assume that the Jacobian matrix of the constraint equations with respect to x_3, *i.e.*, $F_{2;x_3}(x_1, x_3, u)$, is nonsingular on Ω. That is, the rank of $F_{2;x_3}$ is assumed to be constant and full on Ω.

The solvability of the semi-explicit model 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

$$\mathbb{L}_0 = \{z_0 \in \Omega_x \times \mathbb{R}^a \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^p \mid F_0^d(z_0) = 0\}$$

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

$$\dot{x}_1 = F_1(x_1, x_3, u) \tag{2.44a}$$
$$x_3 = \mathcal{R}(x_1, u) \tag{2.44b}$$

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

$$\mathbb{L}_1 = \{z_1 \in \Omega_x \times \mathbb{R}^a \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^p \mid F_1^d(z_1) = 0\}$$

and that the parameters p in F_1^d can be chosen to include $\dot{u}$. Given the initial conditions $(x_{1,0}, x_{2,0}, u_0) \in \Omega$, 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 $z_{1,0} \in \mathbb{L}_1$. 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 (2.44) such that $x_1(t_0) = x_{1,0}$. Moreover, this solution locally solves the original DAE model.

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

The implicit function only assures local solvability. It means that around a given point (x_1, x_3, u), the implicit function $\mathcal{R}(x_1, u)$ is unique and differentiable. However, in some of the theorems in this thesis, it is desired to talk about a larger region for u. The issue is that even if the Jacobian of the constraint equations w.r.t. to x_3, *i.e.*, $F_{2;x_3}$, is nonsingular in the region, it is still not ensured that $\mathcal{R}$ is continuously differentiable function there. For example for the double-cone in Figure 2.2, it is possible to choose either the upper or the lower nappe.

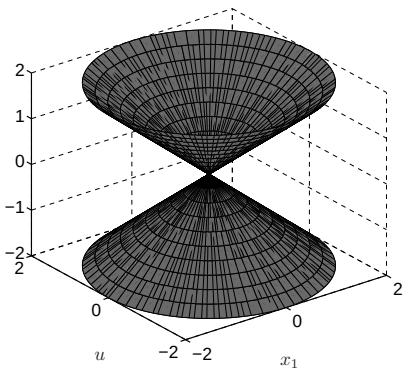


Figure 2.2: *A double-cone.*

Therefore, an assumption have to be introduced in order to select the same solution to x_3 when regions are considered for x_1 and u. For simplicity, it is assumed that the solution to the constraints in the considered region is unique.

Assumption A2. The model is strangeness-free for arbitrary $u \in \Omega_u$. Furthermore, the equation $F_2(x_1, x_3, u) = 0$ has a unique solution w.r.t. x_3 in

$$\Omega = \{x_1 \in \Omega_x,\, x_3 \in \Omega_y,\, u \in \Omega_u \mid F_2(x_1, x_3, u) = 0\}$$

where Ω_x, Ω_y and Ω_u are open and connected sets. Finally, the Jacobian matrix of the constraint equations with respect to x_3, *i.e.*, $F_{2;x_3}(x_1, x_3, u)$, is nonsingular on Ω.

The assumption above implies that the function $\mathcal{R}$ can be chosen unique and continuously differentiable. For the double cone in Figure 2.2, it means that one of the nappes should be chosen. A more general assumption could have been formulated based on global implicit function theorems, see for example, Sandberg (1981) and references therein, but the requirements in these theorems are rather involved.

2.5 DAE Solvers

In the preceding sections, different methods to reduce either the differential index or the strangeness index are presented. One reason for the reduction of the indices was that typically numerical solvers used for integrating dynamical models can only handle models that are strangeness-free, unless the model describes a specific type of process. A motivation was that algebraically, the solution manifold is not visible unless the index is low enough. This is also the main reason why the methods derived in this thesis assume that the models are strangeness-free, and mostly in the forms (2.34) or (2.43). The strangeness-free model also has the advantage that the dynamic and algebraic parts are clearly separated. However, an issue is that the computations to obtain the strangeness-free model can be difficult, for instance, because of the numerical rank tests involved in Hypotheses 2.1 or 2.2.

Therefore, the interest is turned to the kind of solvers used to simulate object-oriented models, such as MODELICA models. Such solvers are included in, *e.g.*, Dymola and OpenModelica. Normally, these solvers reduce the differential index to 1 by differentiating equations that are chosen using Pantelides's algorithm (Pantelides, 1988) and structure the equations so that large DAE models can be simulated efficiently. When the equations are reduced and structured, standard numerical methods such as Runge-Kutta schemes can be used.

As mentioned above, Pantelides's algorithm (Pantelides, 1988) is one of the principal parts in the solvers, and it decides which equations to differentiate when reducing the index of large-scale high index DAE model. The algorithm is graph-theoretical and was, as mentioned in Section 2.2, originally developed to find the conditions that consistent initial values must satisfy. Later, the algorithm has been deployed by others to find the differentiations needed to reduce the index of DAE model to at most 1 in DAE solvers. The advantage of the algorithm is that only structural information about the equations are used, instead of using rank tests. The disadvantage of this approach is that sometimes an incorrect result can be obtained, see Reißig et al. (2000). However, in most practical cases the method seems to find the correct equations to differentiate.

During the index reduction process, some of the variables $x(t)$ are selected as states. For the user, it means that the initial values can be arbitrarily chosen for these variables. The initial values of the remaining variables are computed from the initial values of the states so that the initial value is consistent. It is possible for the user to influence the state selection by indicating that some variables are preferred as states. The solver typically also structures the equations as

$$\tilde{F}_1(t, x_1, x_3, \dot{x}_1) = 0 \tag{2.45a}$$

$$\hat{F}_2(t, x_1, x_3) = 0 \tag{2.45b}$$

where x_3 can be solved from (2.45b) and $\dot{x}_1$ can be solved from (2.45a). This means that an approximation of the transformations discussed in Section 2.4.1 is computed.

2.6 Stability

This section concerns stability analysis of DAE models. In principle, stability of a DAE model means stability of a dynamical system on a manifold. The standard tool, and basically the only tool, for proving stability of 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 model. However, for physical systems such as mechanical and electrical systems, the total energy content of the system can often be used.

The stability results in this section will be focused on two classes of models, namely models that either are semi-explicit, autonomous and strangeness-free or that are linear. However, a small discussion about polynomial possibly higher index models will be presented at the end of this section. For this kind of models a computationally tractable approach, based on Lyapunov theory, has been published in Ebenbauer and Allgöwer (2004).

Consider the autonomous DAE model

$$F(\dot{x}, x) = 0 \tag{2.46}$$

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

Assume there exists an open connected set Ω of consistent initial conditions such that the solution is unique, *i.e.*, the initial value problem consisting of (2.46) together with $x(t_0) \in \Omega$ has a unique solution. Note that in the state-space case this assumption will simplify to Ω being some subset of the domain where the model 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. At the equilibrium, (2.46) 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 Ω is assumed to contain only a single equilibrium. Hence, in order to satisfy this assumption, it might be necessary to reduce Ω. However, this assumption can be relaxed using concepts of set stability, see Hill and Mareels (1990).

The definitions of stability for DAE models are natural extensions of the corresponding definitions for the state-space case.

Definition 2.6 (Stability). The equilibrium point at $(0, 0)$ of (2.46) is called stable if given an $\varepsilon > 0$, 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 (2.46) 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

$$\lim_{t\to\infty} \big|x(t)\big| = 0$$

2.6.1 Semi-Explicit Index One Models

Lyapunov stability for semi-explicit index one models 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 A1 is satisfied, also higher index models can be rewritten in semi-explicit form. Hence, consider the case when (2.46) can be expressed as

$$\dot{x}_1 = F_1(x_1, x_3) \tag{2.47a}$$
$$0 = F_2(x_1, x_3) \tag{2.47b}$$

where $x_1 \in \mathbb{R}^d$ and $x_3 \in \mathbb{R}^a$. The model is assumed to satisfy Assumption A2. This means that, on some set Ω, which in this case has the structure

$$\Omega = \{x_1 \in \Omega_x \subseteq \mathbb{R}^d,\ x_3 \in \mathbb{R}^a,\ |\ x_3 = \mathcal{R}(x_1)\} \tag{2.48}$$

the model (2.47) has index one and a unique solution for arbitrary initial conditions in Ω. It is also assumed that Ω 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 model (2.47) and let $\Omega'_x \subset \Omega_x$ be an open, connected set containing the origin. Suppose there exists a function $V \in C^1(\Omega'_x, \mathbb{R})$ such that V is positive definite and has a negative semidefinite time-derivative on Ω'_x, i.e.*,*

$$V(0) = 0 \ \textit{and}\ V(x_1) > 0,\ \forall x_1 \neq 0, \tag{2.49a}$$
$$V_{x_1}(x_1)F_1\big(x_1, \mathcal{R}(x_1)\big) \leq 0,\ \forall x_1 \tag{2.49b}$$

where $x_1 \in \Omega'_x$. Then the equilibrium $(x_1^0, x_3^0) = (0, 0)$ is stable. Moreover, if the function V is negative definite on Ω'_x, i.e.*,*

$$V_{x_1}(x_1)F_1\big(x_1, \mathcal{R}(x_1)\big) < 0,\ \forall x_1 \neq 0 \tag{2.50}$$

where $x_1 \in \Omega'_x$, then $(x_1^0, x_3^0) = (0, 0)$ is asymptotically stable.

Proof: The proof is to a large extent based on the proof for the state-space case. For $x_1 \in \Omega'_x$ it follows that $(x_1, x_3) \in \Omega' = \{x_1 \in \Omega'_x,\ x_3 \in \mathbb{R}^a \,|\, F_2(x_1, x_3) = 0\} \subset \Omega$. Then the model is given by the reduced model

$$\dot{x}_1 = F_1\big(x_1, \mathcal{R}(x_1)\big)$$

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

$$B_r = \left\{ x_1 \in \mathbb{R}^d,\, x_3 \in \mathbb{R}^a \mid \left|(x_1, x_3)\right| \leq r \right\} \subset \Omega'$$

Since $\mathcal{R}(x_1)$ is at least continuously differentiable it follows that on B_r

$$\left|(x_1, x_3)\right| \leq (1 + L)|x_1|$$

for some $L > 0$. Choose

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

Then it is known from for example Khalil (2002), that (2.49) guarantees the existence of a $\delta_{x_1} > 0$ and a corresponding set

$$B_{\delta_{x_1}} = \{x_1 \in \mathbb{R}^d \mid |x_1| \leq \delta_{x_1}\} \subset B_{r_{x_1}}$$

such that

$$x_1(t_0) \in B_{\delta_{x_1}} \Rightarrow x_1(t) \in B_{r_{x_1}},\, \forall t \geq t_0$$

Then, we can conclude that $\big(x_1(t), x_3(t)\big)$ belong to Ω' and that

$$\left|\big(x_1(t), x_3(t)\big)\right| \leq r \leq \varepsilon$$

By choosing $\delta \leq \delta_{x_1}$ it is also certain that

$$|x_1| \leq \left|(x_1, x_3)\right| \leq \delta \leq \delta_{x_1}$$

That is, by choosing δ smaller than δ_{x_1}, it follows that $|x_1| \leq \delta_{x_1}$ and stability is proved.

The discussion concerning asymptotic stability follows the same line of reasoning. In Khalil (2002), it is shown that condition (2.50) implies the existence of a $\eta_{x_1} > 0$ such that for $x_1(t_0)$ in the set

$$B_{\eta_{x_1}} = \{x_1 \in \mathbb{R}^d \mid |x_1| \leq \eta_{x_1}\} \subset \Omega'_x$$

it holds that $\left|x_1(t)\right| \to 0$ as $t \to \infty$. However, for $x_1 \in \Omega'_x$ it follows that $(x_1, x_3) \in \Omega' \subset \Omega$ and by using

$$\left|(x_1, x_3)\right| \leq (1 + L)|x_1|$$

we have that

$$\lim_{t \to \infty} \left|\big(x_1(t), x_3(t)\big)\right| = 0$$

if an $\eta \leq \eta_{x_1}$ is chosen. This concludes the proof. □

Note that the previously stated solvability conditions only guarantee that a solution exists on some time interval $\mathbb{I}$. 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 this property is ensured if there exists a Lyapunov function as required in Theorem 2.4, 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 Ω. However, the theorem above shows that this does not happen.

The result in Theorem 2.4 is often more difficult to use for DAE models than for state-space models, since the implicit function $\mathcal{R}$ is required. This means that some kind of numeric method is needed to verify the conditions, for instance, the power series method studied later in the thesis.

There are also a number of different generalizations to the results above. For example, the condition in (2.50) 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 2.4 often are difficult to use in practice, another method to prove stability is presented. This method is known as Lyapunov's indirect method. The main idea is to use the linearization of (2.47) to determine local stability of the origin. Assume that F_1 and F_2 are continuously differentiable. Then, the linearization around $(x_1, x_3) = (0, 0)$ is given by

$$\dot{x}_1 = A_{11}x_1 + A_{12}x_3 + o(x) \tag{2.51a}$$
$$0 = A_{21}x_1 + A_{22}x_3 + o(x) \tag{2.51b}$$

where

$$A_{11} = F_{1;x_1}(0,0) \qquad A_{12} = F_{1;x_3}(0,0)$$
$$A_{21} = F_{2;x_1}(0,0) \qquad A_{22} = F_{2;x_3}(0,0)$$

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

The next theorem gives conditions under which stability of the origin $(x_1, x_3) = (0, 0)$ can be concluded by investigating its stability as an equilibrium point for the linear part of (2.51).

Theorem 2.5
Consider the model (2.47) and let $\Omega'_x \subset \Omega_x$ be a neighborhood of $x_1 = 0$. Then, the origin is asymptotically stable if $\operatorname{Re}\operatorname{eig}_i(A_{11} - A_{12}A_{22}^{-1}A_{21}) < 0$ for $i = 1, 2, ..., d$.

Proof: For all $(x_1, x_3) \in \Omega' = \{x_1 \in \Omega'_x,\, x_3 \in \mathbb{R}^a \,|\, F_2(x_1, x_3) = 0\}$ it is known that $F_{2;x_3}$ is nonsingular, and since $(0, 0) \in \Omega'$ it follows that A_{22} has full rank. Hence, (2.51b) can be reformulated as

$$x_3 = -A_{22}^{-1}A_{21}x_1 + o(x)$$

Combining (2.51) and the latter equation gives

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

On a compact subset of Ω' it holds that

$$\big|(x_1, x_3)\big| \leq (1 + L)|x_1|$$

for some $L > 0$. This makes it possible to write the model as

$$\dot{x}_1 = \big(A_{11} - A_{12}A_{22}^{-1}A_{21}\big)x_1 + o(x_1) \tag{2.52}$$

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

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

Converse Lyapunov Theorem

Often in this thesis, the model (2.47) is denoted asymptotically stable on a set Ω. Then, it is assumed that the set Ω, for which (2.47) 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 Ω can always be performed.

Theorem 2.6
Consider the model (2.47) and assume that $x = 0$ is an asymptotically stable equilibrium to (2.47). Let $\Omega_x' \subset \Omega_x$ be an open, connected set containing the origin. Further, let $R_A \subset \Omega_x'$ 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 $x_1 \in R_A$ such that

$$\begin{aligned} V(0) = 0 \ \textit{and} \ W(0) = 0 \\ V(x_1) \to \infty, \ x_1 \to \partial R_A \\ V_{x_1}(x_1)F_1\big(x_1, \mathcal{R}(x_1)\big) \leq -W(x_1), \ \forall x_1 \in R_A \end{aligned}$$

where ∂R_A denotes the boundary of R_A. Furthermore, for any $c > 0$, the set $\big\{x_1 \in R_A \,|\, V(x_1) \leq c\big\}$ is a compact subset of R_A.

Proof: For $x_1 \in \Omega_x'$, it holds that the model is given by

$$\dot{x}_1 = F_1\big(x_1, \mathcal{R}(x_1)\big)$$

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

Hence, it is always possible to choose an invariant set for x_1 by choosing the set corresponding to some c. However, then a corresponding invariant set for (x_1, x_3) can be chosen since $x_3 = \mathcal{R}(x_1)$.

2.6.2 Linear Models

Consider asymptotic stability for regular linear DAE models

$$E\dot{x} = Ax \tag{2.53}$$

with consistent initial conditions. For such models, the equilibrium $x = 0$ is unique and according to Theorem 2.3, the solution is unique. For the case with $E = I$, *i.e.*, the model is a state-space model, it is well-known that stability is determined by the eigenvalues of the matrix A. Also for linear DAE models (2.53), the eigenvalues can be used to determine stability. However, since the E-matrix for a DAE model normally is rank deficient, both finite and infinite eigenvalues occur, see Dai (1989). In this case, stability is determined only by the finite eigenvalues as formulated in the following theorem.

Theorem 2.7
A regular linear DAE model (2.53) is asymptotically stable if and only if

$$\text{Re } \sigma(E, A) < 0$$

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. □

Notice that the set $\mathbb{C}$ does not contain the infinity and therefore the theorem only considers finite s. For convenience in the sequel, also consider the strangeness-free case.

Corollary 2.2
A regular linear DAE model in semi-explicit form

$$\begin{aligned} \dot{x}_1 &= A_{11}x_1 + A_{12}x_3 + B_1u \\ 0 &= A_{21}x_1 + A_{22}x_3 + B_2u \end{aligned}$$

is asymptotically stable if and only if

$$\dot{x}_1 = (A_{11} - A_{12}A_{22}^{-1}A_{21})x_1 + (B_1 - A_{12}A_{22}^{-1}B_2)u$$

is asymptotically stable.

Above, it has been assumed that the initial conditions are consistent. However, it is possible to extend the stability concept to include the more general solutions described in Section 2.4.2. The same condition as above, *i.e.*, $\text{Re } \sigma(E, A) < 0$, will still be obtained. The difference is that possible impulsive behavior at the initial time, due to the inconsistent initial values, has to be disregarded. Therefore, it is important that the definition of stability in that case does not include the initial time, that is, is formulated for $t > t_0$ instead of $t \geq t_0$.

2.6.3 A Barrier Function Method

Another interesting Lyapunov based approach to stability analysis of DAE models is presented in Ebenbauer and Allgöwer (2004). Instead of finding the reduced model (2.35) explicitly, this approach works with an implicit definition of the solution manifold. This is particularly attractive for nonlinear polynomial DAE models, since it implies that the

stability condition can be verified computationally using sum-of-squares relaxations and semidefinite programming.

To simplify the explanation, assume that the model (2.46) satisfies Hypothesis 2.1 and that the underlying model (2.35) locally solves the original model. Then the derivative array F_μ for some μ implicitly defines the solution $x(t)$ on some neighborhood $\mathbb{U}$. It turns out to be convenient to define the set

$$\mathcal{U}_x = \{x \in \Theta,\, \mathbf{x}_{\mu+1} \in \mathbb{R}^{(\mu+1)n}\}$$

where $\mathbf{x}_{\mu+1} = (\dot{x}, \ldots, x^{(\mu+1)})$ and $\Theta \subseteq \mathbb{R}^n$ is a neighborhood of the origin. A theorem for stability of a nonlinear DAE model can now be formulated.

Theorem 2.8
The equilibrium $x = 0$ of (2.46) is stable if there exist a function $V \in \mathcal{C}^1(\Omega_x, \mathbb{R})$ 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 a positive integer μ such that

$$V_x(x)\,\dot{x} \leq \left|F_\mu(x, \mathbf{x}_{\mu+1})\right|^2 \rho(x, \mathbf{x}_{\mu+1}) \tag{2.54}$$

is satisfied for $(x, \mathbf{x}_{\mu+1}) \in \mathcal{U}_x$. If (2.54) is satisfied with inequality for all nonzero $x \in \Theta$, the model is asymptotically stable.

Proof: See Ebenbauer and Allgöwer (2004). □

The assumptions on the model made in the theorem, ensure that F_μ defines the solution implicitly. These assumptions can be relaxed. For example, it is possible to handle systems without a unique solution, time-varying models etc. However, in some of these cases it is not stability but a convergence property that is proved, see Ebenbauer and Allgöwer (2004) for details.

2.7 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 dynamical system. This is a well-studied area, which goes far back in time, see Sussmann and Willems (1997). Similar to the stability analysis, optimal control for DAE models is in principle nothing but optimal control of a state-space system on a manifold. Consequently, the methods for DAE models 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.

2.7.1 Formulation and Summary of the Optimal Control Problem

An optimal control problem for a continuous-time state-space model can be written as

$$\begin{aligned} V(x_0) = \inf_{u(\cdot)} \int_0^\infty L(x,u)\,dt \\ \text{s.t.} \quad \dot{x} = F(x,u) \\ x(0) = x_0 \in \Omega_x \end{aligned} \tag{2.55}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^p$, Ω_x is a connected set and 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 $u \neq 0$.

Note that in (2.55) 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 have to be asymptotically stable. Therefore, the minimization is done with respect to all $u(\cdot)$ such that

$$\dot{x} = F\big(x, u(\cdot)\big) \tag{2.56}$$

is asymptotically stable on Ω_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 considered explicitly. Otherwise, it will always be optimal to choose $u(t) = 0$.

In (2.55), 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 $u(t) = \mu\big(x(t)\big)$ for some function μ. 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 $\mathcal{C}^1$, 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, the focus will in the sequel be on the dynamic programming approach. 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).

2.7.2 Necessary Conditions For Optimality

First a theorem is presented that yields necessary conditions. That is, given that there exist 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 (2.56) 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

$$0 = \min_u \big(L(x,u) + V_x(x) F(x,u) \big) \tag{2.57}$$

and $u_*(t)$ is the pointwise in time minimizing argument in (2.57).

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

$$\begin{aligned}
V(x_0) &= \min_{u(\cdot)} \left(\int_0^{\Delta t} L\big(x(s,x_0),u(s)\big)\, ds + V\big(x(\Delta t, x_0)\big) \right) \\
&= \min_{u(\cdot)} \left(\int_0^{\Delta t} L\big(x(s,x_0),u(s)\big)\, ds + V\big(x_0 + F\big(x_0,u(0)\big)\Delta t + o(\Delta t)\big) \right) \\
&= \min_{u(\cdot)} \left(\int_0^{\Delta t} L\big(x(s,x_0),u(s)\big)\, ds + V(x_0) + V_x(x_0) F\big(x_0,u(0)\big)\Delta t + o(\Delta t) \right)
\end{aligned}$$

where $x(s,x_0)$ is the solution to (2.56) with the initial condition x_0 and $o(\Delta t)$ is the ordo-function defined such that $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$. In the computation above, first an Euler approximation of the solution $x(\Delta t, x_0)$ around x_0 is performed and then V is Taylor expanded around x_0. Since $V(x_0)$ is independent of u, it can be subtracted from

both sides. Division by Δt and using the fact that $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$, it follows that

$$0 = \min_u \big(L(x_0, u) + V_x(x_0)F(x_0, u)\big)$$

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

$$0 = \min_u \big(L(x, u) + V_x(x)F(x, u)\big)$$

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

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).

2.7.3 Sufficient Conditions For Optimality

The next fundamental result is that the HJB (2.57) 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 (2.55) 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

$$0 = \min_u \big(L(x, u) + J_x(x)F(x, u)\big) \tag{2.58}$$

for $x \in \Omega_x$. Let

$$\mu(x) = \operatorname*{argmin}_u \big(L(x, u) + J_x(x)F(x, u)\big)$$

and assume that by using $u(t) = \mu\big(x(t)\big)$, the closed-loop system (2.56) becomes asymptotically stable on Ω_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 (2.56) is asymptotically stable on Ω_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 (2.58) that

$$J(x_0) \leq \int_0^T L\big(x(t), u(t)\big)\, dt + J\big(x(T)\big)$$

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

$$J(x_0) = \int_0^\infty L\big(x(t), \mu\big(x(t)\big)\big)\, dt \leq \int_0^\infty L\big(x(t), u(t)\big)\, dt$$

which proves optimality. □

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

Theorem 2.10 can also be seen as an algorithm to compute the optimal control law and the corresponding optimal performance criterion. The algorithm can be found in Algorithm 2.1.

Algorithm 2.1 Computation of the optimal solution.

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

$$\hat{\mu}(x, \lambda) = \underset{u}{\operatorname{argmin}} \big(L(x, u) + \lambda^T F(x, u)\big),\ x \in \Omega_x$$

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

2. Solve the partial differential equation

$$0 = L\big(x, \hat{\mu}(x, V_x(x))\big) + V_x(x)^T F\big(x, \hat{\mu}\big(x, V_x(x)\big)\big) \tag{2.59}$$

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

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

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

$$0 = L_u(x, u) + V_x(x) F_u(x, u)$$

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

Usually, the minimization involved when solving the HJB is nontrivial. As pointed out earlier, differentiation with respect to u is then often used, but it is only a necessary condition. The effect of this, is that further investigation of the different solutions is necessary to determine whether they are optimal or not.

In this section a special case, for which the minimization w.r.t. to u can be done analytically, is presented. For this end, consider a model in control-affine form

$$\dot{x} = f(x) + g(x)u \tag{2.60}$$

where f and g are smooth functions and a cost function described as

$$L(x,u) = l(x) + S(x)u + u^T R(x)u \tag{2.61}$$

where it is assumed that $L(x,u) \geq 0$ for (x,u) and $R(x) \succ 0$ for all x. In this case, Theorem 2.10 can be simplified as follows.

Corollary 2.3
Consider an optimal control problem where the cost function is chosen as (2.61) and the model is given by (2.60). Suppose there exists a positive semi-definite, continuously differentiable function $J(x)$ satisfying $J(0)$ and

$$0 = l(x) + J_x(x)f(x) - \frac{1}{4}\big(J_x(x)g(x) + S(x)\big)R^{-1}(x)\big(J_x(x)g(x) + S(x)\big)^T \tag{2.62}$$

for all $x \in \Omega_x$ such that

$$\dot{x} = f(x) - \frac{1}{2}g(x)R^{-1}(x)\big(J_x(x)g(x) + S(x)\big)^T \tag{2.63}$$

is asymptotically stable on Ω_x. Then, $V(x) = J(x)$ for $x \in \Omega_x$ and

$$\mu(x) = -\frac{1}{2}R^{-1}(x)\big(V_x(x)g(x) + S(x)\big)^T$$

is the optimal feedback law.

Proof: For the considered case, the HJB becomes

$$0 = \min_u l(x) + S(x)u + u^T R(x)u + J_x(x)\big(f(x) + g(x)u\big) \tag{2.64}$$

Using completion of squares, the right-hand side can be rewritten as follows

$$l + Su + u^T Ru + J_x(f + gu) = l + J_x f + (J_x g + S)u + u^T Ru = \\ l + J_x f - \frac{1}{4}(J_x g + S)R^{-1}(J_x g + S)^T + \\ \big(u + \frac{1}{2}R^{-1}(J_x g + S)^T\big)^T R\big(u + \frac{1}{2}R^{-1}(J_x g + S)^T\big)$$

and since $R(x)$ is positive definite for all x, it means that the minimizing $\mu(x)$ is given by

$$\mu(x) = -\frac{1}{2}R^{-1}(J_x g + S)^T$$

If this feedback law is inserted into the (2.64), equation (2.62) is obtained and the corresponding closed-loop dynamics become (2.63). Hence, if (2.62) has a sufficiently smooth solution and (2.63) becomes asymptotically stable, it follows from Theorem 2.10 that the optimal solution is found. □

2.7.4 Example

A small, but yet illustrative, example is presented below. The chosen optimal control problem fits into the conditions in Corollary 2.3, but in order to illustrate the more general procedure, Algorithm 2.1 it is used instead.

Consider the model

$$\dot{x} = \alpha x + u \tag{2.65}$$

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

$$J\big(x_0\big) = \frac{1}{2}\int_0^\infty \beta x^4 + u^2\, dt \tag{2.66}$$

is minimized, where $\beta \in \mathbb{R}$. The goal is a global feedback law and therefore $\Omega_x = \mathbb{R}$. For notational convenience, 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 $H_{uu}(x, u, \lambda) = 1/2 > 0$ for all x. Hence, the obtained feedback law must be optimal if $u = \mu(x) = \hat{\mu}\big(x, V_x(x)\big)$ makes (2.65) asymptotically stable. The partial differential equation in part 2 of Algorithm 2.1 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

$$u = -\Big(\alpha + \sqrt{\alpha^2 + \beta x^2}\Big)x, \qquad V_x(x) = \Big(\alpha + \sqrt{\alpha^2 + \beta x^2}\Big)x \tag{2.67}$$

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

$$V(x) = \frac{1}{2}\alpha x^2 + \frac{1}{3\beta}\big(\alpha^2 + \beta x^2\big)^{\frac{3}{2}} - \frac{1}{3\beta}(\alpha^2)^{\frac{3}{2}} \tag{2.68}$$

since $V(0) = 0$. Using the parameters α and β, 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. $\beta > 0$, α arbitrary:
   In this case the state is included in the cost function and (2.65) is either asymptotically stable, a pure integrator or unstable depending of the choice of α. 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 (2.68) is studied, it can be seen that a small β 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 (2.67) will in this case be
   $$u = -\Big(\alpha + \sqrt{\alpha^2}\Big)x = -2\alpha x, \qquad V_x(x) = \Big(\alpha + \sqrt{\alpha^2}\Big)x = 2\alpha x$$
   and the cost function is
   $$V(x) = \alpha x^2$$
   If α 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 (2.67) becomes

$$u = -\left(\alpha + \sqrt{\alpha^2}\right)x = 0, \qquad V_x(x) = \left(\alpha + \sqrt{\alpha^2}\right)x = 0$$

   and an 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.

3

Optimal Feedback Control of DAE Models

In principle, the problem of finding optimal feedback laws for DAE models can be solved by the theory for state-space models. The reason is that under certain regularity conditions, it was shown in Section 2.3.1 that a rather general class of DAE models can be rewritten either as a state-space model together with some algebraic equations or as a state-space model with requirements on the initial condition. In both cases, the obtained system fits into the theory presented in Section 2.7.

A problem is that the underlying state-space model is often hard or even impossible to express in closed form. Therefore, it is interesting to find methods to compute the optimal feedback law either based on a general DAE model or at least based on an index reduced model.

To the author's knowledge, no approach exists such that a general DAE model (2.2) can be treated immediately without index reduction. However, for linear time-invariant DAE models (2.6) such methods exist, see for example Bender and Laub (1987) and Mehrmann (1989). These methods are based on variational calculus which leads to a generalized eigenvalue problem to be solved, see Jonckheere (1988).

For nonlinear DAE models in semi-explicit form (2.5), results for higher index models can also be found based on the dynamic programming approach. In Xu and Mizukami (1993), an equation similar to the HJB is derived. In this chapter, this HJB-like equation will be presented and it will be investigated how its solution is related to the solution of the corresponding HJB for the underlying 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 DAE model is rewritten as its underlying state-space model 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

$$\dot{x}_1 = F_1(x_1, x_3, u) \tag{3.1a}$$
$$0 = F_2(x_1, x_3, u) \tag{3.1b}$$

which is assumed to satisfy Assumption A2 for $x_1 \in \Omega_x$ and the corresponding set for (x_1, x_3, u) is denoted Ω. The model (3.1) can be the result of the index reduction method in Section 2.4.3 in a case when Assumption A1 is satisfied.

The considered class of optimal control problems has an infinite time horizon and can be formulated as

$$V\big(x_1(0)\big) = \min_{u(\cdot)} \int_0^\infty L(x_1, x_3, u)\, dt \tag{3.2}$$

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 cost function L is assumed to be positive semidefinite and positive definite in u. 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

$$\dot{x}_1 = F_1\big(x_1, x_3, u(\cdot)\big) \tag{3.3a}$$
$$0 = F_2\big(x_1, x_3, u(\cdot)\big) \tag{3.3b}$$

is asymptotically stable, see Section 2.6. Another condition, specific for the DAE case, is that (3.3) is required to be strangeness-free. This condition is added since different feedback laws $u = k(x_1, x_3)$ may yield different indices of (3.3). The effect would be that the size of the dynamical and algebraical parts would change and some of the variables x_1 could be algebraically determined from the other variables in x_1 and x_3. 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 (3.3) is strangeness-free, x_1 will define the dynamical part and x_3 is determined algebraically from x_1. The initial conditions are then given for x_1 while the initial conditions for x_3 are assumed to be chosen consistently, *i.e.*, such that $F_2\big(x_{1,0}, x_{2,0}, u(x_{1,0})\big) = 0$. 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 DAE models, *e.g.*, Cobb (1983); Bender and Laub (1987); Jonckheere (1988); Xu and Mizukami (1993), the possibility of changing the index using feedback is utilized. The optimal feedback law is then required to yield a strangeness-free closed-loop model (3.3, even if the open loop system (3.1) has a

higher index. For linear DAE models, such a choice is ensured if the model is impulse controllable, see Dai (1989).

Notice that the requirement $L(x_1, x_3, u)$ being positive semidefinite and positive definite in u is somewhat restrictive in our case, but is made in order match the assumptions in Xu and Mizukami (1993). Since $x_3 = \mathcal{R}(x_1, u)$, which is guaranteed from Assumption A2, it follows that a more relaxed requirement is $L\big(x_1, \mathcal{R}(x_1, u), u\big)$ being positive semidefinite and positive definite in u.

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

Assumption A2 makes it possible to solve the optimal feedback control problem (3.2) as an optimal feedback control problem for a state-space system (2.55). For $x_1 \in \Omega_x$ and $u \in \mathbb{R}^p$, the considered optimal problems can be written as

$$\begin{aligned} V(x_{1,0}) = \inf_{u(\cdot)} & \int_0^\infty L\big(x_1, \mathcal{R}(x_1, u), u\big)\, dt \\ \text{s.t.} \quad & \dot{x}_1 = F_1\big(x_1, \mathcal{R}(x_1, u), u\big) \\ & x_1(0) = x_{1,0} \in \Omega_x \end{aligned}$$

The cost function L 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

$$0 = \min_u \big(L(x_1, \mathcal{R}(x_1, u), u) + V_{x_1}(x_1) F_1(x_1, \mathcal{R}(x_1, u), u)\big),\ x_1 \in \Omega_x \tag{3.4}$$

This $V(x_1)$ is then the optimal performance criterion in (3.2). Note that if (3.4) is only possible to solve on a set smaller than the set on which the DAE model can be rewritten as a state-space model, it is assumed that Ω_x is redefined as the smaller set. Furthermore, remember that $V(x_1)$ is only proved 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 $x_1(t)$ within Ω_x for $t \geq 0$, unless Ω_x is chosen to be an invariant set, see comments in Section 2.7.3.

The first-order necessary condition for optimality of (3.4) yields the set of equations

$$\begin{aligned} 0 &= L_u + V_{x_1} F_{1;u} + (L_{x_3} + V_{x_1} F_{1;x_3}) \mathcal{R}_u \\ 0 &= L + V_{x_1} F_1 \end{aligned}$$

where the quantities in the right hand sides are evaluated at $\big(x_1, \mathcal{R}(x_1, u), u\big)$. Using that

$$F_2\big(x_1, \mathcal{R}(x_1, u), u\big) = 0$$

identically in u, differentiation with respect to u gives

$$F_{2;x_3}\big(x_1, \mathcal{R}(x_1, u), u\big) \mathcal{R}_u(x_1, u) + F_{2;u}\big(x_1, \mathcal{R}(x_1, u), u\big) = 0$$

which can be solved for $\mathcal{R}_u(x_1, u)$ as

$$\mathcal{R}_u(x_1, u) = -F_{2;x_3}\big(x_1, \mathcal{R}(x_1, u), u\big)^{-1} F_{2;u}\big(x_1, \mathcal{R}(x_1, u), u\big)$$

Since $x_3 = \mathcal{R}(x_1, u)$ is the unique solution of (3.1b), it is also possible to write these equations as

$$0 = L_u + V_{x_1} F_{1;u} - (L_{x_3} + V_{x_1} F_{1;x_3}) F_{2;x_3}^{-1} F_{2;u} \tag{3.6a}$$
$$0 = L + V_{x_1} F_1 \tag{3.6b}$$
$$0 = F_2 \tag{3.6c}$$

where the right hand sides are evaluated at (x_1, x_3, u). One way of looking at (3.6) is to regard (3.6a) and (3.6c) as $p + a$ equations from which one tries to solve for u and x_3 as functions of x_1 and V_{x_1}. When these quantities are substituted into (3.6b) 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 2.7.3 can be used. It will require $\mathcal{R}_{uu}$ to be computed which is possible and the expressions can be found in Chapter 6.

3.3 The Hamilton-Jacobi-Bellman-Like Equation

In Xu and Mizukami (1993), the optimal control problem (3.2) is solved using a different approach. According to their Theorem 3.1, the optimal solution can be found by solving the Hamilton-Jacobi-Bellman-like equation

$$0 = \min_u \big(L(x_1, x_3, u) + W_1(x_1) F_1(x_1, x_3, u) + W_2(x_1, x_3) F_2(x_1, x_3, u)\big) \tag{3.7}$$

for some continuous functions $W_1(x_1)$ and $W_2(x_1, x_3)$ 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 (3.2).

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

$$0 = L_u(x_1, x_3, u) + W_1(x_1) F_{1;u}(x_1, x_3, u) + W_2(x_1, x_3) F_{2;u}(x_1, x_3, u) \tag{3.8a}$$
$$0 = L(x_1, x_3, u) + W_1(x_1) F_1(x_1, x_3, u) + W_2(x_1, x_3) F_2(x_1, x_3, u) \tag{3.8b}$$

where x_3 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 (3.6) can be obtained by restricting (3.8) to points satisfying $F_2 = 0$. The result is the following system of equations

$$0 = L_u(x_1, x_3, u) + W_1(x_1) F_{1;u}(x_1, x_3, u) + W_2(x_1, x_3) F_{2;u}(x_1, x_3, u) \tag{3.9a}$$
$$0 = L(x_1, x_3, u) + W_1(x_1) F_1(x_1, x_3, u) \tag{3.9b}$$
$$0 = F_2(x_1, x_3, u) \tag{3.9c}$$

If W_2 is considered unknown, the set of equations (3.9) is still underdetermined. Hence, more equations are needed or W_2 has to be considered as given. It should be mentioned that in Xu and Mizukami (1993), only sufficient conditions are given, *i.e.*, if a W_1 and a W_2 can be found such that (3.7) is satisfied, the optimal solution is found, and they do not mention what kind of conditions W_2 has to satisfy. This will be investigated in the next section.

3.4 Relationships Among the Solutions

The reduced Hamilton-Jacobi equation (3.4) and the Hamilton-Jacobi-like equation (3.7) 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 (3.4) *on Ω_x. Then $W_1(x_1) = V_{x_1}(x_1)$, $u = k(x_1)$ solve* (3.7) *under the constraint $F_2(x_1, x_3, u) = 0$. Moreover, with the choice*

$$W_1 = V_{x_1}, \quad W_2 = -\big(L_{x_3} + V_{x_1} F_{1;x_3}\big) F_{2;x_3}^{-1} \tag{3.10}$$

the necessary conditions for optimality (3.9) *are satisfied for $u = k(x_1)$ together with $x_3 = \mathcal{R}\big(x_1, k(x_1)\big)$.*

Proof: When $F_2 = 0$, the right hand sides of (3.7) and (3.4) coincide. Comparing (3.6) and (3.9) shows that (3.9) is satisfied for $u = k(x_1)$, $x_3 = \mathcal{R}\big(x_1, k(x_1)\big)$ with W_2 chosen as in (3.10). □

The converse relation is given by the following lemma.

Lemma 3.2
Assume that for $x_1 \in \Omega_x$[1] *it holds that:*

- (3.7) *has a solution $u = \psi(x_1, x_3)$*
- *$F_2\big(x_1, x_3, \psi(x_1, x_3)\big) = 0$ has a solution $x_3 = \eta(x_1)$*
- *$W_1(x_1) = V_{x_1}(x_1)$ for some function $V(x_1)$*

Then $V_{x_1}(x_1)$ and $u = k(x_1) = \psi\big(x_1, \eta(x_1)\big)$ solve (3.4) *for $x_1 \in \Omega_x$. Moreover, for $(x_1, x_3, u) \in \Omega$ satisfying (3.9), it follows that*

$$W_2 = -\big(L_{x_3} + W_1 F_{1;x_3}\big) F_{2;x_3}^{-1} \tag{3.11}$$

Proof: We have

$$\begin{aligned} &L\big(x_1, \eta(x_1), \psi\big(x_1, \eta(x_1)\big)\big) + V_{x_1}(x_1) F_1\big(x_1, \eta(x_1), \psi\big(x_1, \eta(x_1)\big)\big) \\ &\qquad = L\big(x_1, \mathcal{R}\big(x_1, k(x_1)\big), k(x_1)\big) + V_{x_1}(x_1) F_1\big(x_1, \mathcal{R}\big(x_1, k(x_1)\big), k(x_1)\big) = 0 \end{aligned}$$

[1] If this Ω_x is smaller than the Ω_x on which the system can be written as a state-space system, Ω_x is redefined as the smaller region.

since the minimal value in (3.7) is attained for $u = \psi(x_1, x_3)$ for all $x_1 \in \Omega_x$ and $x_3 \in \mathbb{R}^{n_2}$, and then particularly for $x_3 = \eta(x_1)$. Since $x_1 \in \Omega_x$ it is also known that $\eta(x_1) = \mathcal{R}\big(x_1, k(x_1)\big)$. According to (3.7)

$$0 \leq L(x_1, x_3, u) + V_{x_1}(x_1)F_1(x_1, x_3, u) + W_2(x_1, x_3)F_2(x_1, x_3, u)$$

for all $x_1 \in \Omega_x$, $x_3 \in \mathbb{R}^{n_2}$ and $u \in \mathbb{R}^p$. In particular, it follows that

$$0 \leq L(x_1, \mathcal{R}(x_1, u), u) + V_{x_1}(x_1)F_1\big(x_1, \mathcal{R}(x_1, u), u\big)$$

and (3.4) is satisfied.

Since a u solving (3.9a) is given by $u = \psi(x_1, x_3)$, (3.9b) and (3.9c) give

$$\begin{aligned} 0 &= L\big(x_1, x_3, \psi(x_1, x_3)\big) + W_1(x_1)F_1\big(x_1, x_3, \psi(x_1, x_3)\big) \\ 0 &= F_2\big(x_1, x_3, \psi(x_1, x_3)\big) \end{aligned}$$

Differentiation of these relations with respect to x_3 yields

$$0 = L_{x_3} + \big(L_u + W_1F_{1;u}\big)\psi_{x_3} + W_1F_{1;x_3} \tag{3.12a}$$

$$0 = F_{2;x_3} + F_{2;u}\psi_{x_3} \tag{3.12b}$$

If (3.9a) is multiplied from right by ψ_{x_3} and after (3.12) is inserted, the result is that W_2 has to satisfy

$$0 = W_1F_{1;x_3} + L_{x_3} + W_2F_{2;x_3}$$

Due to the fact that $F_{2;x_3}$ is nonsingular for $(x_1, x_3, u) \in \Omega$, it follows that

$$W_2 = -\big(L_{x_3} + W_1F_{1;x_3}\big)F_{2;x_3}^{-1} \tag{3.13}$$

□

Hence, for a system that satisfies Assumption A2 one further necessary condition for the optimal solution, namely (3.13), is obtained.

3.5 Control-Affine-like DAE Models

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

$$\dot{x}_1 = f_1(x_1, x_3) + g_1(x_1)u \tag{3.14a}$$

$$0 = f_2(x_1, x_3) + g_2(x_1)u \tag{3.14b}$$

while the cost function should be expressed in the form $L(x_1, x_3, u) = l(x_1) + \frac{1}{2}u^Tu$. Then (3.9a) can be solved explicitly in u for all x_1, x_3 since (3.9a) will become

$$0 = u^T + W_1(x_1)g_1(x_1) + W_2(x_1, x_3)g_2(x_1) \tag{3.15}$$

and from Lemma 3.2, it follows that

$$W_2(x_1, x_3) = -W_1(x_1)f_{1;x_3}(x_1, x_3)f_{2;x_3}^{-1}(x_1, x_3) \tag{3.16}$$

Note that $f_{2;x_3}(x_1, x_3)$ is nonsingular for all (x_1, x_3) such that $f_2(x_1, x_3) = 0$ is solvable since $F_{2;x_3}(x_1, x_3, u)$ is nonsingular for all $(x_1, x_3, u) \in \Omega$ and then particularly for $u = 0$. Combining (3.15) and (3.16) yields

$$u = -\hat{g}(x_1, x_3)^T W_1(x_1)^T$$

and after some manipulation the necessary conditions can be rewritten as

$$0 = l(x_1) + W_1(x_1)\hat{f}(x_1, x_3) - \frac{1}{2}W_1(x_1)\hat{g}(x_1, x_3)\hat{g}(x_1, x_3)^T W_1(x_1)^T \tag{3.17a}$$

$$0 = f_2(x_1, x_3) - g_2(x_1)\hat{g}(x_1, x_3)^T W_1(x_1)^T \tag{3.17b}$$

where

$$\hat{f}(x_1, x_3) = f_1(x_1, x_3) - f_{1;x_3}(x_1, x_3) f_{2;x_3}^{-1}(x_1, x_3) f_2(x_1, x_3)$$
$$\hat{g}(x_1, x_3) = g_1(x_1) - f_{1;x_3}(x_1, x_3) f_{2;x_3}^{-1}(x_1, x_3) g_2(x_1)$$

Note that the even though the DAE model is affine in the control input, this might not be the case for the underlying state-space model.

3.6 Example

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

Consider the simple system

$$\dot{x}_1 = x_3$$
$$0 = u - x_1^3 - x_3$$

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 (3.6) give

$$0 = u + V_{x_1} \cdot 0 + V_{x_1} \cdot 1$$
$$0 = \frac{x_1^2}{2} + \frac{u^2}{2} + V_{x_1} x_3$$
$$0 = u - x_1^3 - x_3$$

Eliminating u and x_3 gives the following equation for V_{x_1}

$$V_{x_1}^2 + 2x_1^3 V_{x_1} - x_1^2 = 0$$

To get a positive definite solution for V the solution for V_{x_1} must be chosen as

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

and the corresponding optimal feedback law then becomes

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

If instead (3.8) is used to solve the optimal control problem, it leads to the equations

$$\begin{aligned} 0 &= u + W_1 \cdot 0 + W_2 \cdot 1 \\ 0 &= \frac{x_1^2}{2} + \frac{u^2}{2} + W_1 x_3 + W_2(u - x_1^3 - x_3) \end{aligned}$$

Since the system satisfies Assumption A2, it is known from Lemma 3.2 that W_2 must satisfy (3.13), *i.e.*,

$$W_2 = W_1$$

On the solution manifold $\{x \in \mathbb{R}^n, u \in \mathbb{R}^p \mid F_2(x_1, x_3, u) = 0\}$, *i.e.*, when $0 = u - x_1^3 - x_3$, the same equation for W_1 as for V_{x_1} is obtained and therefore the solution is the same.

4

Power Series Solution of the Hamilton-Jacobi-Bellman Equation

In Chapter 3, optimal control for DAE models in semi-explicit form, which possibly could come from higher index problems, was discussed. Unfortunately, there are practical issues with these methods. One issue is that in order to explicitly and not numerically solve the optimality conditions, an expression in closed form of the implicit function $\mathcal{R}$ is needed. However, in many cases it is impossible to express the implicit function in closed form. Another problem with the methods in Chapter 3 is that even if $\mathcal{R}$ can be expressed in closed form, the optimality conditions can only be solved to obtain an explicit solution for a small class of problems. Therefore, an alternative approach to solve the optimal control problem will be presented in this chapter.

The idea is to compute the power series of the optimal solution instead. This turns out to be an easier problem. 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 obtained sequentially, by first solving a quadratic optimal control problem for the linearized system and then 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 continuously differentiable functions. 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 that have been 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 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 and Krener present 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 approximate solution. Therefore, this kind of methods are often denoted approximate methods even though the complete power series expansions of the performance criterion and feedback law yield the true optimal solution. There are other methods that 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 for finite horizon optimal control problems 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 DAE models is described in Section 4.3. In Section 4.4 the infinite horizon case is considered. Most of the assumptions in the theorems are formulated based on the reduced DAE model. However, in Section 4.5 some of these assumptions are formulated in terms of the original DAE model instead. Finally, some examples are presented in Section 4.6, and some proofs in Section 4.7.

4.1 Problem Formulation

The considered optimal control problem can be defined as to minimize the integral criterion

$$G\big(x_1(T)\big) + \int_{\tau}^{T} L(t, x_1, x_3, u) e^{\lambda t}\, dt \tag{4.1}$$

where $x = (x_1,\, x_3)$ satisfies the differential-algebraic system

$$\hat{F}_1(\dot{x}_1, x_1, x_3, u, t) = 0 \tag{4.2a}$$

$$\hat{F}_2(x_1, x_3, u, t) = 0 \tag{4.2b}$$

with some initial condition $x_1(\tau) = x_{1,0}$. The function G is the terminal cost, L is the cost function and λ is the discount factor. The optimal return function is defined as

$$V(\tau, x_{1,0}) = \inf_{u(\cdot)} \Big(G\big(x_1(T)\big) + \int_{\tau}^{T} L(t, x_1, x_3, u) e^{\lambda t}\, dt \Big) \tag{4.3}$$

The optimal control problem will be considered for τ in an interval $[T_0,\, T]$ and $x_{1,0}$ in some neighborhood of the origin. The initial condition is assumed to be consistent, that is, to satisfy the following assumption.

Assumption A3. The initial condition satisfies $\hat{F}_2\big(x_{1,0}, x_2(\tau), u(\tau), \tau\big) = 0$.

In order to ensure that the solution manifold of the system is described by the model equations, the system will be required to satisfy the following assumption, that is, to be strangeness-free.

Assumption A4. The system equations (4.2) can be uniquely solved to give

$$\dot{x}_1 = \mathcal{L}(t, x_1, u) \tag{4.4a}$$
$$x_3 = \mathcal{R}(t, x_1, u) \tag{4.4b}$$

for all $t \in [T_0,\, T]$ and $(\dot{x}_1, x_1, x_3, u) \in \Omega$, where Ω is an open set containing the origin.

Note that although Assumption A4 requires the model to be strangeness-free, the model may still be the result of an index reduction process of a higher index model that satisfies Hypothesis 2.2 and for which an integrator chain has been introduced, see Section 2.4.4.

Assumption A4 ensures that the DAE model has an underlying ODE model as can be seen in (4.4). In principle, it means that methods for state-space models can be used. The major computational challenge is the fact that the functions $\mathcal{L}$ and $\mathcal{R}$ need not be explicit. Therefore, the focus in this section is on a method for finding optimal feedback laws that only requires the power series expansions of $\mathcal{L}$ and $\mathcal{R}$. For this end, the following assumption will be fundamental.

Assumption A5. The functions $\hat{F}_1$, $\hat{F}_2$, L and G are real analytic in some open set $\mathcal{W} \subseteq \mathbb{R}^{d+n+p}$, containing the origin $(\dot{x}_1, x_1, x_3, u) = 0$, for all $t \in [T_0,\, T]$.

Analyticity of $\hat{F}_1$, $\hat{F}_2$, L and G makes it possible to express them as power series

$$\begin{aligned}\hat{F}_1(\dot{x}_1, x_1, x_3, u, t) &= -E_1(t)\dot{x}_1 + A_{11}(t)x_1 + A_{12}(t)x_3 + B_1(t)u \\ &\quad + \hat{F}_{1h}(\dot{x}_1, x_1, x_3, u, t) \qquad (4.5a)\\ \hat{F}_2(x_1, x_3, u, t) &= A_{21}(t)x_1 + A_{22}(t)x_3 + B_2(t)u + \hat{F}_{2h}(x_1, x_3, u, t) \qquad (4.5b)\\ L(t, x, u) &= x^T Q(t)x + u^T R(t)u + 2x^T S(t)u + L_h(t, x, u) \qquad (4.5c)\\ G(x_1) &= x_1^T M x_1 + G_h(x_1) \qquad (4.5d)\end{aligned}$$

which are convergent around the origin $(\dot{x}_1, x_1, x_3, u) = 0$, uniformly for all $t \in [T_0,\, T]$. The functions $\hat{F}_{1h}$ and $\hat{F}_{2h}$ include terms beginning with order two, while the higher order terms of the performance criterion, that is L_h and G_h, are of order three at least.

For notational convenience the matrices in (4.5) will often will be concatenated and partitioned as

$$A(t) = \begin{pmatrix} A_{11}(t) & A_{12}(t) \\ A_{21}(t) & A_{22}(t) \end{pmatrix}, \qquad B(t) = \begin{pmatrix} B_1(t) \\ B_2(t) \end{pmatrix},$$
$$Q(t) = \begin{pmatrix} Q_{11}(t) & Q_{12}(t) \\ Q_{21}(t) & Q_{22}(t) \end{pmatrix}, \qquad S(t) = \begin{pmatrix} S_1(t) \\ S_2(t) \end{pmatrix}$$

where all the matrices are assumed to be continuous real matrix functions.

Assumption A6. The matrices $E_1(t)$ and $A_{22}(t)$ are invertible for all $t \in [T_0,\, T]$.

The considered class of feedback laws can be expressed as

$$u(t, x_1) = D(t)x_1 + u_h(t, x_1) \tag{4.6}$$

where $D(t)$ is a continuous matrix function. The function $u_h(t, x)$ consists of terms of order two or higher in x and is uniformly convergent in a neighborhood of the origin for $t \in [T_0,\, T]$.

The choice to just consider feedback laws based on x_1 is motivated by the fact that the dynamics of the system is described by x_1. However, for systems not satisfying Assumption A4, one could also consider to have a feedback law

$$u(t, x_1) = D_1(t)x_1 + u_h(t, x_1) + D_2(t)x_3$$

where D_2 is chosen such that the system becomes invertible around the origin.

Throughout the chapter, it will for notational reasons be assumed that Ω is covered by $\mathcal{W}$, that is, $\Omega \subset \mathcal{W}$.

4.2 State-Space Models

If the equation $\hat{F}_2 = 0$ is absent and $\hat{F}_1 = 0$ is explicit in $\dot{x}_1$, the optimal control problem can be rewritten in the ordinary form

$$\begin{aligned} V(\tau, x_0) = \inf_{u(\cdot)} \Big(G\big(x(T)\big) + \int_{\tau}^{T} L(t, x, u) e^{\lambda t}\, dt \Big) \\ \text{s.t.} \qquad \dot{x} = F(t, x, u), \\ x(\tau) = x_0 \end{aligned} \tag{4.7}$$

This problem is a standard optimal control problem and the solution is found using the Hamilton-Jacobi-Bellman equation (HJB), see Bryson and Ho (1975); Bertsekas (1995),

$$-V_t(t, x) = \min_{u} L(t, x, u)e^{\lambda t} + V_x(t, x)F(t, x, u) \tag{4.8}$$

where V_t and V_x denote the partial derivatives of V w.r.t. t and x, respectively. It follows that the optimal feedback law $u_*(t, x)$ has to satisfy the following equations.

$$\begin{aligned} 0 &= L\big(t, x, u_*(t, x)\big)e^{\lambda t} + V_t(t, x) + V_x(t, x)F\big(x, u_*(t, x)\big) \\ 0 &= L_u\big(t, x, u_*(t, x)\big)e^{\lambda t} + V_x(t, x)F_u\big(t, x, u_*(t, x)\big) \end{aligned} \tag{4.9}$$

Unfortunately, this nonlinear partial differential equation can only be solved explicitly in a few special cases. However, in Willemstein (1977), it is shown that under the given assumptions and for feedback laws described by uniformly convergent power series, the optimal control problem (4.7) has an optimal solution. Furthermore, the corresponding optimal return function becomes analytic in a neighborhood of the origin and for $t \in [T_0,\, T]$. It means that it can be written as

$$V(t, x) = x^T P(t)x + V_h(t, x) \tag{4.10}$$

where $P(t)$ is a positive definite matrix function and $V_h(t,x)$ contains terms of at least order three in x.

Based on this observation a computational method can be derived. If the unknown series expansions of V and u together with the known series expansions of F, L and G are substituted into (4.9), two polynomial equations in x are obtained, where the coefficients in the polynomials are differential equations in t. The HJB must be satisfied for all considered values of x, which in this case means that polynomial equations must be satisfied in a neighborhood of the origin. The coefficients corresponding to different orders in x will yield separate differential equations in t. Solving these differential equations then yields the optimal solution.

In the finite horizon case, the function $e^{\lambda t}$ with the discount factor λ does not introduce any extra complications, since it can be treated as any other time-variability in the cost function. However, because of the specific structure this function, the equations (4.9) can be rewritten in a way that will simplify the computations and also better explain the equations with a discount factor in the infinite horizon case.

Consider the change of optimal return function

$$\bar{V}(t,x) = V(t,x)e^{-\lambda t} \tag{4.11}$$

where $\bar{V}(t,x)$ also becomes analytic in a neighborhood of the origin and for $t \in [T_0,\, T]$. Then

$$\begin{aligned} V_t(t,x) &= \bar{V}_t(t,x)e^{\lambda t} + \bar{V}(t,x)\lambda e^{\lambda t} \\ V_x(t,x) &= \bar{V}_x(t,x)e^{\lambda t} \end{aligned}$$

Substitution of these equations into (4.9) gives

$$\begin{aligned} 0 &= L\big(t,x,u_*(t,x)\big) + \lambda\bar{V}(t,x) + \bar{V}_t(t,x) + \bar{V}_x(t,x)F\big(x,u_*(t,x)\big) \\ 0 &= L_u\big(t,x,u_*(t,x)\big) + \bar{V}_x(t,x)F_u\big(t,x,u_*(t,x)\big) \end{aligned} \tag{4.12}$$

Hence, the explicit dependence of $e^{\lambda t}$ in the equations has been removed. However, note that since this only is a change of variables, (4.12) are completely equivalent to (4.9) concerning solvability. That is, if there exist solutions $V(t,x)$ and $u_*(t,x)$ which solves (4.9), there will exist $\bar{V}(t,x)$ and $u_*(t,x)$ solving (4.12). Therefore, the equations above are used in the following theorem, which both formalizes the existence of an optimal solution and gives the expressions to compute it.

Theorem 4.1
Consider the optimal control problem (4.7). Suppose that Assumption A5 is fulfilled and that the weight matrices satisfy

$$\begin{pmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{pmatrix} \succeq 0, \quad R(t) \succ 0, \quad M \succeq 0$$

for $t \in [T_0,\, T]$. Then the optimal feedback control $u_*(t,x)$ exists and is the unique solution of (4.12) *for small $|x|$ and $t \in [T_0,\, T]$. Furthermore, the optimal return function*

and optimal feedback are of the form (4.10) *and* (4.6)*, respectively. In these expressions,* $P(t)$ *and* $D_*(t)$ *are given by*

$$\begin{aligned} 0 &= \dot{P}(t) + P(t)\big(A(t) + \tfrac{\lambda}{2}I\big) + \big(A(t) + \tfrac{\lambda}{2}I\big)^T P(t) \\ &\quad - \big(P(t)B(t) + S(t)\big)R(t)^{-1}\big(P(t)B(t) + S(t)\big)^T + Q(t) \qquad (4.13a) \\ 0 &= P(T) - M \qquad (4.13b) \end{aligned}$$

and

$$D_*(t) + R^{-1}(t)\big(S^T(t) + B(t)^T P(t)\big) = 0 \qquad (4.13c)$$

The higher order terms in (4.10) *and (4.6) can be calculated recursively from the following expressions.*

$$\begin{aligned} &\bar{V}_x^{[m]}(t,x)A_c(t)x + \bar{V}_t^{[m]}(t,x) + \lambda\bar{V}^{[m]}(t,x) = \\ &-\sum_{k=3}^{m-1} \bar{V}_x^{[k]}(t,x)B(t)u_*^{[m-k+1]}(t,x) \\ &-\sum_{k=2}^{m-1} \bar{V}_x^{[k]}(t,x)F_h^{[m-k+1]}(t,x,u_*) - L_h^{[m]}(t,x,u_*) \\ &-2\sum_{k=2}^{\lfloor\frac{m-1}{2}\rfloor} u_*^{[k]}(t,x)^T R(t)u_*^{[m-k]}(t,x) - u_*^{[m/2]}(t,x)^T R(t)u_*^{[m/2]}(t,x) \qquad (4.14a) \end{aligned}$$

$$\bar{V}^{[m]}(T,x) = e^{-\lambda T}G^{[m]}(x) \qquad (4.14b)$$

where $m = 3, 4, \ldots$, $A_c(t) = A(t) + B(t)D_*(t)$*, and*

$$\begin{aligned} u_*^{[k]}(t,x) = -\frac{1}{2}R(t)^{-1}\Big\{&\bar{V}_x^{[k+1]}(t,x)B(t) \\ &+\sum_{i=1}^{k-1} \bar{V}_x^{[k-i+1]}(t,x)F_{h;u}^{[i]}(t,x,u_*) + L_{h;u}^{[k]}(x,u_*)\Big\} \end{aligned} \qquad (4.14c)$$

for $k = 2, 3, \ldots$. *In* (4.14) *the convention that* $\sum_k^l = 0$ *for* $l < k$ *is used and the terms* $u^{[m/2]}$ *are to be omitted if* m *is odd.*

Proof: The existence of an optimal solution is proved exactly as in Willemstein (1977), with $e^{\lambda t}$ included in the cost function $L(t,x,u)$. Then, the computation of $V(t,x)$ and $u_*(t,x)$ can be done using (4.12), which yields (4.13) and (4.14). □

The fact that recursive computation is possible can be motivated by the following observations,

$$\begin{aligned} F_h^{[k]}(t,x,u_*) &= F_h^{[k]}(t,x,u_*^{[1]} + u_*^{[2]} + \ldots + u_*^{[k-1]}) \\ L_h^{[k]}(t,x,u_*) &= L_h^{[k]}(t,x,u_*^{[1]} + u_*^{[2]} + \ldots + u_*^{[k-2]}) \end{aligned}$$

and

$$F_{h;u}^{[i]}(t, x_1, u_*) = F_{h;u}^{[i]}(t, x_1, u_*^{[1]} + \ldots + u_*^{[i]})$$
$$\hat{L}_{h;u}^{[i-1]}(t, x_1, u_*) = \hat{L}_{h;u}^{[i]}(t, x_1, u_*^{[1]} \ldots + u_*^{[i-1]})$$

since $F_h(t, x, u)$ and $L_h(t, x, u)$ are power series beginning with terms of order two and three, respectively. Therefore, the right-hand side of (4.14a) will only depend on the terms

$$u_*^{[1]}, \ldots, u_*^{[m-2]},\ \bar{V}^{[2]}, \ldots, \bar{V}^{[m-1]} \tag{4.15}$$

while the right-hand side of (4.14c) only depends on

$$u_*^{[1]}, \ldots, u_*^{[k-1]},\ \bar{V}^{[2]}, \ldots, \bar{V}^{[k+1]} \tag{4.16}$$

The lowest order terms in $u_*(t, x)$ and $\bar{V}(t, x)$, *i.e.*,

$$u_*^{[1]}(t, x) = D_*(t)x, \quad \bar{V}^{[2]}(t, x) = x^T P(t)x \tag{4.17}$$

is therefore computed using the Riccati Differential Equation (RDE) in (4.13a) together with the boundary condition (4.13b) and the equation (4.13c). Having these, the higher order terms can be computed using (4.14), which together defines a linear differential equation for the coefficients of $\bar{V}^{[m]}(t, x)$, with the right hand side known and with the boundary condition $\bar{V}^{[m]}(T, x) = e^{-\lambda T} G^{[m]}(x)$. From $\bar{V}^{[i]}(t, x)$, $i = 2, \ldots$, the terms of corresponding order of $V(t, x)$ can then be found using (4.11).

Note that in the finite horizon case, there is no guarantee for a stabilizing feedback law, and thereby a stable solution, not even locally around the origin. This is one of the major reasons for studying the infinite horizon as will be seen in Section 4.4.

4.3 DAE Models

To solve the optimal control problem described by (4.2) and (4.3), it is reformulated as the following equivalent optimal control problem

$$V(x_{1,0}) = \inf_{u(\cdot)} \Big(G\big(x_1(T)\big) + \int_\tau^T \hat{L}(t, x_1, u) e^{\lambda t}\, dt \Big) \tag{4.18a}$$

subject to the dynamics

$$\dot{x}_1 = \mathcal{L}(t, x_1, u) \tag{4.18b}$$

where $\mathcal{L}$ is given by (4.4a) and

$$\hat{L}(t, x_1, u) = L\big(t, x_1, \mathcal{R}(t, x_1, u), u\big) \tag{4.18c}$$

where $\mathcal{R}$ is given by (4.4b)

Under Assumption A4, this reformulation can always be done for consistent initial values. The optimal control problem is then, in principle, a standard problem in the state variables x_1 which solution is given by

$$-V_t(t, x_1) = \min_u \hat{L}(t, x_1, u) e^{\lambda t} + V_{x_1}(t, x_1) \mathcal{L}(t, x_1, u)$$

If $\hat{F}_1$ and $\hat{F}_2$ are assumed to satisfy Assumptions A4, A5, and A6, it is ensured by the implicit function theorem, *i.e.*, Theorem A.1, that $\mathcal{L}(t, x_1, u)$ and $\mathcal{R}(t, x_1, u)$ are analytic and will therefore have convergent power series expansions. This means that if the assumption concerning positivity of the cost matrices is satisfied, Theorem 4.1 is applicable and the optimal solution can be computed using the expressions in Theorem 4.1 given that the power series expansions of $\mathcal{L}$ and $\hat{L}$ around $(x_1, u) = (0, 0)$ can be computed recursively. The procedure for doing this will be shown in the next section.

4.3.1 Power Series Expansion of the Reduced Problem

A keystone in the derived method is that the power series expansions of $\mathcal{R}(t, x_1)$ and $\mathcal{L}(t, x_1, u)$ can be computed recursively. Let

$$\dot{x}_1 = \mathcal{L}(t, x_1, u) = \mathcal{L}^{[1]}(t, x_1, u) + \mathcal{L}_h(t, x_1, u) \tag{4.19a}$$

$$x_3 = \mathcal{R}(t, x_1, u) = \mathcal{R}^{[1]}(t, x_1, u) + \mathcal{R}_h(t, x_1, u) \tag{4.19b}$$

where both $\mathcal{L}_h(t, x_1, u)$ and $\mathcal{R}_h(t, x_1, u)$ are continuous in t and contain terms in x_1 and u of order two and higher.

From (4.5b) the series expansion of $\hat{F}_2$ is given by

$$\hat{F}_2(x_1, x_3, u, t) = A_{21}(t)x_1 + A_{22}(t)x_3 + B_2(t)u + \hat{F}_{2h}(x_1, x_3, u, t)$$

By combining this equation with (4.19b), it follows that

$$\begin{aligned} 0 = A_{21}(t)x_1 + A_{22}(t)\big\{\mathcal{R}^{[1]}(t, x_1, u) + \mathcal{R}_h(t, x_1, u)\big\} \\ + B_2(t)u + \hat{F}_{2h}\big(t, x_1, \mathcal{R}^{[1]}(t, x_1, u) + \mathcal{R}_h(t, x_1, u), u\big) \end{aligned} \tag{4.20}$$

The equation above has to be satisfied for all (x_1, u) in a neighborhood of the origin. This means that the first order term of $\mathcal{R}(t, x_1, u)$ will be given by

$$\mathcal{R}^{[1]}(t, x_1, u) = -A_{22}^{-1}(t)A_{21}(t)x_1 - A_{22}^{-1}(t)B_2(t)u \tag{4.21}$$

since all other terms are of higher order than one. Furthermore, since $\hat{F}_{2h}(x_1, x_3, u)$ has an order of at least two, it follows that

$$\begin{aligned} \hat{F}_{2h}^{[m]}\big(t, x_1, \mathcal{R}(t, x_1, u), u\big) = \\ \hat{F}_{2h}^{[m]}\big(t, x_1, \mathcal{R}^{[1]}(t, x_1, u) + \ldots + \mathcal{R}^{[m-1]}(t, x_1, u), u\big) \end{aligned} \tag{4.22}$$

This makes it is possible to derive a recursive expression for a general order term of $\mathcal{R}(t, x_1, u)$ as

$$\mathcal{R}^{[m]}(t, x_1, u) = -A_{22}^{-1}(t)\hat{F}_{2h}^{[m]}\big(t, x_1, \mathcal{R}^{[1]}(t, x_1, u) + \ldots + \mathcal{R}^{[m-1]}, u\big) \tag{4.23}$$

Now consider the Taylor series of (4.5a), *i.e.*,

$$\begin{aligned} \hat{F}_1(\dot{x}_1, x_1, x_3, u, t) = \\ - E_1(t)\dot{x}_1 + A_{11}(t)x_1 + A_{12}(t)x_3 + B_1(t)u + \hat{F}_{1h}(\dot{x}_1, x_1, x_3, u, t) \end{aligned}$$

Combined with (4.19), the result is the equation

$$\begin{aligned} 0 = &-E_1(t)\mathcal{L}^{[1]}(t,x_1,u) - E_1(t)\mathcal{L}_h(t,x_1,u) + A_{11}(t)x_1 \\ &+ A_{12}(t)\big(\mathcal{R}^{[1]}(t,x_1,u) + \mathcal{R}_h(t,x_1,u)\big) + B_1(t)u \\ &+ \hat{F}_{1h}\big(\mathcal{L}^{[1]}(t,x_1,u) + \mathcal{L}_h(t,x_1,u), x_1, \mathcal{R}^{[1]}(t,x_1,u) \\ &+ \mathcal{R}_h(t,x_1,u), u, t\big) \end{aligned} \tag{4.24}$$

By assumption $E_1(t)$ is nonsingular for $t \in [T_0,\, T\,]$ and for notational reasons, it will in the sequel of this section, be assumed that it is an identity matrix. The first term in $\mathcal{L}(t,x_1,u)$ is obtained as

$$\mathcal{L}^{[1]} = A_{11}(t)x_1 + A_{12}(t)\mathcal{R}^{[1]}(t,x_1,u) + B_1(t)u = \hat{A}(t)x_1 + \hat{B}(t)u$$

where

$$\begin{aligned} \hat{A}(t) &= A_{11}(t) - A_{12}(t)A_{22}^{-1}(t)A_{21}(t) \\ \hat{B}(t) &= B_1(t) - A_{12}(t)A_{22}^{-1}(t)B_2(t) \end{aligned}$$

and the second equality is obtained using (4.21). Since $\hat{F}_{1h}(\dot{x}_1, x_1, x_3, u, t)$ contains terms of at least order two it follows that

$$\begin{aligned} \hat{F}_{1h}^{[m]}\big(\mathcal{L}(t,x_1,u), x_1, \mathcal{R}(t,x_1,u), u, t\big) = \\ \hat{F}_{1h}^{[m]}\big(\mathcal{L}^{[1]}(t,x_1,u) + \ldots + \mathcal{L}^{[m-1]}(t,x_1,u), x_1, \\ \mathcal{R}^{[1]}(t,x_1,u) + \ldots + \mathcal{R}^{[m-1]}(t,x_1,u), u, t\big) \end{aligned}$$

which shows that higher order terms in $\mathcal{L}(t,x_1,u)$ can be computed recursively using the expression

$$\begin{aligned} \mathcal{L}^{[m]}(t,x_1,u) = \\ A_{12}(t)\mathcal{R}^{[m]}(t,x_1,u) + \hat{F}_1^{[m]}\big(\mathcal{L}^{[1]}(t,x_1,u) + \ldots + \mathcal{L}^{[m-1]}(t,x_1,u), x_1, \\ \mathcal{R}^{[1]}(t,x_1,u) + \ldots + \mathcal{R}^{[m-1]}(t,x_1,u), u, t\big) \end{aligned}$$

The equations to find the coefficients of $\mathcal{R}$ and $\mathcal{L}$ will be linear in the m:th order coefficients. It means that if the equations are solved recursively, the computation can be done rather fast. However, if the number of variables in either x_1 or x_3 are large, the number of equations will grow rapidly.

For physical systems, the DAE model is often semi-explicit, *i.e.*, satisfies Assumption A1, and can be written as

$$\begin{aligned} \dot{x}_1 &= \hat{F}_1(x_1, x_3, u) \\ 0 &= \hat{F}_2(x_1, x_3, u) \end{aligned}$$

The computations above can then be simplified substantially, since the power series of $\mathcal{L}(t,x_1,u)$ is obtained, without solving any equations, as the composition of the power series of $\hat{F}_1$ and $\mathcal{R}$.

Having the power series expansions of $\mathcal{R}(t, x_1, u)$, the series expansion of (4.18c) can be computed as

$$\begin{aligned}\hat{L}(t, x_1, u) &= \begin{pmatrix} x_1 \\ u \end{pmatrix}^T \Pi^T(t) \begin{pmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{pmatrix} \Pi(t) \begin{pmatrix} x_1 \\ u \end{pmatrix} + \hat{L}_h(t, x_1, u) \\ &= \begin{pmatrix} x_1 \\ u \end{pmatrix}^T \begin{pmatrix} \hat{Q}(t) & \hat{S}(t) \\ \hat{S}^T(t) & \hat{R}(t) \end{pmatrix} \begin{pmatrix} x_1 \\ u \end{pmatrix} + \hat{L}_h(t, x_1, u)\end{aligned} \tag{4.26}$$

where

$$\Pi(t) = \begin{pmatrix} I & 0 \\ -A_{22}^{-1}(t)A_{21}(t) & -A_{22}^{-1}(t)B_2(t) \\ 0 & I \end{pmatrix} \tag{4.27}$$

and

$$\begin{aligned}&\hat{L}_h(t, x_1, u) = \\ &= L_h(t, x_1, \mathcal{R}(t, x_1, u), u) + 2x_1^T Q_{12}(t)\mathcal{R}_h(t, x_1, u) \\ &+ 2\mathcal{R}^{[1]}(t, x_1, u)Q_{22}(t)\mathcal{R}_h(t, x_1, u) + 2u^T S_2(t)\mathcal{R}_h(t, x_1, u) \\ &+ \mathcal{R}_h(x_1, u)^T Q_{22}(t)\mathcal{R}_h(t, x_1, u)\end{aligned} \tag{4.28}$$

4.3.2 Application of the Results for State-Space Models

The optimal control problem with the system model (4.19a) and the cost function (4.26) is now solvable using the method described in Section 4.4.2. Using the first order terms of the series expansions (4.19a) and (4.26), the RDE (4.13a)–(4.13b) and the expression for the first order term in the feedback (4.13c) for the DAE formulation become

$$\begin{aligned}0 &= \dot{P}(t) + P(t)\big(\hat{A}(t) + \tfrac{\lambda}{2}I\big) + \big(\hat{A} + \tfrac{\lambda}{2}I\big)^T(t)P(t) \\ &\quad - \big(P(t)\hat{B}(t) + \hat{S}(t)\big)\hat{R}^{-1}(t)\big(P(t)\hat{B}(t) + \hat{S}(t)\big)^T + \hat{Q}(t) \qquad (4.29a) \\ 0 &= P(T) - M \qquad (4.29b)\end{aligned}$$

and

$$0 = D_*(t) + \hat{R}^{-1}(t)\big(\hat{S}^T(t) + \hat{B}^T(t)P(t)\big) \tag{4.29c}$$

The higher order terms of $\bar{V}(t, x_1)$ and $u_*(t, x_1)$ are obtained from (4.14). In (4.14) only the series expansion coefficients of the different functions are included and it is therefore possible to replace these functions with the series expansion coefficients of

$\mathcal{L}(t, x_1, u)$ and $\hat{L}(t, x_1, u)$, *i.e.*,

$$\begin{aligned}
&\bar{V}_{x_1}^{[m]}(t, x_1)\hat{A}_c(t)x_1 + \bar{V}_t^{[m]}(t, x_1) + \lambda\bar{V}(t, x_1) = \\
&- \sum_{k=3}^{m-1} \bar{V}_{x_1}^{[k]}(t, x_1)\hat{B}(t)u_*^{[m-k+1]}(t, x_1) \\
&- \sum_{k=2}^{m-1} \bar{V}_{x_1}^{[k]}(t, x_1)\hat{F}_{1h}^{[m-k+1]}(t, x_1, u_*) \\
&- 2\sum_{k=2}^{\lfloor\frac{m-1}{2}\rfloor} u_*^{[k]}(t, x_1)^T\hat{R}(t)u_*^{[m-k]}(t, x_1) \\
&- u_*^{[m/2]}(t, x_1)^T\hat{R}(t)u_*^{[m/2]}(t, x_1) - \hat{L}_h^{[m]}(t, x_1, u_*) \qquad (4.30a)\\
&\bar{V}^{[m]}(T, x) = e^{-\lambda T}G^{[m]}(x) \qquad (4.30b)
\end{aligned}$$

where $m = 3, 4, \ldots$, $\hat{A}_c(t) = \hat{A}(t) + \hat{B}(t)D_*(t)$, 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

$$\begin{aligned}
u_*^{[k]}(t, x_1) = -\frac{1}{2}\hat{R}^{-1}(t)\Big\{&\bar{V}_{x_1}^{[k+1]}(t, x_1)\hat{B}(t) \\
&+ \sum_{i=1}^{k-1} \bar{V}_{x_1}^{[k-i+1]}(t, x_1)\hat{F}_{1h;u}^{[i]}(t, x_1, u_*) + \hat{L}_{h;u}^{[k]}(t, x_1, u_*)\Big\} \qquad (4.30c)
\end{aligned}$$

where $k = 2, 3, \ldots$. In (4.30), the terms $\mathcal{L}^{[i]}(t, x_1, u_*)$ and $\hat{L}_h^{[i]}(t, x_1, u_*)$ are given by the corresponding terms in (4.24) and (4.28), respectively.

Since $\mathcal{L}_h(t, x, u)$, $\mathcal{R}_h(t, x_1, u)$ and $L_h(t, x, u)$ are power series of order two, two and three, respectively, and

$$\begin{aligned}
\mathcal{L}^{[m]}(t, x_1, u_*) &= \mathcal{L}(t, x_1, u_*^{[1]} + u_*^{[2]} + \ldots + u_*^{[m]}) \\
\mathcal{R}^{[m]}(t, x_1, u_*) &= \mathcal{R}(t, x_1, u_*^{[1]} + u_*^{[2]} + \ldots + u_*^{[m]})
\end{aligned}$$

it follows that

$$\begin{aligned}
\mathcal{L}_h^{[k]}(t, x_1, u_*) &= \mathcal{L}^{[k]}(t, x_1, u_*^{[1]} + u_*^{[2]} + \ldots + u_*^{[k-1]}) \\
\hat{L}_h^{[k]}(x_1, u_*) &= \hat{L}_h^{[k]}(x_1, u_*^{[1]} + u_*^{[2]} + \ldots + u_*^{[k-2]})
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{L}_{h;u}^{[i]}(t, x_1, u_*) &= \mathcal{L}_{h;u}^{[i]}(t, x_1, u_*^{[1]} + \ldots + u_*^{[i]}) \\
\hat{L}_{h;u}^{[i-1]}(t, x_1, u_*) &= \hat{L}_{h;u}^{[i]}(t, x_1, u_*^{[1]} \ldots + u_*^{[i-1]})
\end{aligned}$$

In the same way as for the state-space case, the right-hand sides of (4.30a) and (4.30c) will then only depend on the sequences (4.15) and (4.16), respectively. So by consecutively calculating the terms of the series

$$\bar{V}^{[2]}(t, x_1),\ u_*^{[1]}(t, x_1),\ \mathcal{R}^{[1]}(t, x_1, u_*^{[1]}),\ \mathcal{L}^{[1]}(t, x_1, u_*^{[1]})\ \ldots \qquad (4.31)$$

it is possible to generate the power series for $\bar{V}(t,x_1)$, $u_*(t,x_1)$, $\mathcal{L}\big(t,x_1,u_*(t,x_1)\big)$ and $\mathcal{R}\big(t,x_1,u_*(t,x_1)\big)$. Having $\bar{V}(t,x_1)$, $V(t,x_1)$ can be computed using (4.11).

In the sequence above, it can be seen that it is unnecessary to calculate orders of $\mathcal{L}(t,x_1,u)$ and $\mathcal{R}(t,x_1,u)$ higher than the desired order of the approximation of $u_*(t,x_1)$. However, if desired for some other reason it is possible to compute arbitrarily high orders of them.

Summarizing this section the following theorem can be formulated.

Theorem 4.2
Consider the optimal control problem defined by the DAE model (4.2) and the integral criterion (4.3). Suppose that Assumptions A3 – A6 are satisfied and that

$$\begin{pmatrix} \hat{Q}(t) & \hat{S}(t) \\ \hat{S}^T(t) & \hat{R}(t) \end{pmatrix} \succeq 0, \quad \hat{R}(t) \succ 0, \quad M \succeq 0$$

for $t \in [T_0, T]$. Then, the optimal feedback control $u_*(x_1)$ exists and is the unique solution for small $|x_1|$ and $t \in [T_0, T]$ of

$$0 = L_u - V_{x_1}\hat{F}_{1;\dot{x}_1}^{-1}\hat{F}_{1;u} - \big(L_{x_3} - V_{x_1}\hat{F}_{1;\dot{x}_1}^{-1}\hat{F}_{1;x_3}\big)\hat{F}_{2;x_3}^{-1}\hat{F}_{2;u} \tag{4.32a}$$

$$0 = L + V_{x_1}\dot{x}_1 + V_t \tag{4.32b}$$

$$0 = \hat{F}_1 \tag{4.32c}$$

$$0 = \hat{F}_2 \tag{4.32d}$$

where L is evaluated in (t,x_1,x_3,u), V in (t,x_1), $\hat{F}_1$ in $(\dot{x}_1,x_1,x_3,u,t)$ and $\hat{F}_2$ in (x_1,x_3,u,t).

Furthermore, the optimal solution $V(t,x_1)$ and $u_*(t,x_1)$ can be computed from the expressions (4.29) and (4.30).

Proof: To solve the optimal control problem defined by the DAE model (4.2) and the integral criterion (4.3) is equivalent to solving the problem (4.18). Applying Theorem 4.1 on (4.18) gives

$$0 = \hat{L}(t,x_1,u) + V_t(t,x_1) + V_{x_1}(t,x_1)\mathcal{L}(t,x_1,u) \tag{4.33a}$$

$$0 = \hat{L}_u(t,x_1,u) + V_{x_1}(t,x_1)\mathcal{L}_u(t,x_1,u) \tag{4.33b}$$

Differentiation of the equations

$$\begin{aligned} 0 &= \hat{F}_1\big(\mathcal{L}(t,x_1,u),x_1,\mathcal{R}(t,x_1,u),u,t\big) \\ 0 &= \hat{F}_2\big(x_1,\mathcal{R}(t,x_1,u),u,t\big) \\ \hat{L}(t,x_1,u) &= L\big(t,x_1,\mathcal{R}(t,x_1,u),u\big) \end{aligned}$$

with respect to u gives the following expressions

$$0 = \hat{F}_{1;\dot{x}_1}\mathcal{L}_u + \hat{F}_{1;x_3}\mathcal{R}_u + \hat{F}_{1;u} \tag{4.34a}$$

$$0 = \hat{F}_{2;x_3}\mathcal{R}_u + \hat{F}_{2;u} \tag{4.34b}$$

$$\hat{L}_u = L_u + L_{x_3}\mathcal{R}_u \tag{4.34c}$$

where the unspecified arguments can be found in the equation above.

Solving (4.34b) locally around the origin for $\mathcal{R}_u$, which is possible since by assumption $F_{2;x_3}$ is nonsingular for all $t \in [T_0, T]$, and substituting into (4.34a) and (4.34c) yields

$$0 = \hat{F}_{1;\dot{x}_1}\mathcal{L}_u - \hat{F}_{1;x_3}\hat{F}_{2;x_3}^{-1}\hat{F}_{2;u} + \hat{F}_{1;u} \tag{4.35a}$$

$$\hat{L}_u = L_u - L_{x_3}\hat{F}_{2;x_3}^{-1}\hat{F}_{2;u} \tag{4.35b}$$

Around the origin the equation for $\mathcal{L}_u$ can be solved because of the assumptions and substitution into (4.33) gives the desired result (4.32). □

Note that even if it is possible to first calculate the reduced system and then formulate the optimal solution in terms of it, there are sometimes structural benefits with formulating the expression in terms of (4.32), as will be discussed in Chapter 6. Further note that the Taylor series of the inverse above can be computed efficiently. Let

$$F_{2;x_3}(t, x_1, x_3, u) = H(t, x, u) = H_0(t) - H_h(t, x, u)$$

where $H_0(t) = F_{2;x_3}(t, 0, 0, 0)$ and $H_h(t, x, u)$ are the terms of at least order one and is uniformly convergent for $t \in [T_0, T]$. From the assumptions, it is known that $H_0(t)$ is nonsingular for $t \in [T_0, T]$. Then the following expression can be used, as shown in Lancaster and Tismenetsky (1985),

$$\big(H_0(t) - H_h(t, x, u)\big)^{-1} = H_0(t)^{-1}\sum_{i=0}^{m-1}\Big(H_h(t, x, u)H_0(t)^{-1}\Big)^i \tag{4.36}$$

This expression can be computed easily to rather high orders.

4.4 The Infinite Horizon Case

The infinite horizon case has many similarities with the finite horizon case. However, there are some major differences. For example, the system is assumed time-invariant which simplifies the assumption about solvability. The infinite horizon together with the time-invariant system also means that the optimal feedback law will be time-invariant, which simplifies for instance the implementation. On the other hand, it is necessary to require the feedback law to be stabilizing, which restricts the cases that can be handled.

4.4.1 Problem Formulation

In the infinite horizon case, the considered optimal control problem is to minimize the integral criterion

$$V(x_{1,0}) = \inf_{u(\cdot)} \int_0^\infty L(x_1, x_3, u)e^{\lambda t}\, dt \tag{4.37}$$

subject to the differential-algebraic system

$$\hat{F}_1(\dot{x}_1, x_1, x_3, u) = 0 \tag{4.38a}$$
$$\hat{F}_2(x_1, x_3, u) = 0 \tag{4.38b}$$

with some initial condition $x_1(0) = x_{1,0}$. The initial condition is assumed to satisfy Assumption A3. The assumption that the DAE model should be strangeness-free can, since the system is time-invariant, be simplified as follows.

Assumption A7. It holds that $\hat{F}(0,0,0,0) = 0$. Furthermore, $\hat{F}_{1;\dot{x}_1}(0,0,0,0)$ and $\hat{F}_{2;x_3}(0,0,0)$ are nonsingular.

From the implicit function theorem, *i.e.*, Theorem A.1, it then follows that there exists a neighborhood Ω of the origin, such that for $(\dot{x}_1, x_1, x_3, u) \in \Omega$, the DAE model can be written as

$$\dot{x}_1 = \mathcal{L}(x_1, u) \tag{4.39a}$$
$$x_3 = \mathcal{R}(x_1, u) \tag{4.39b}$$

As for the finite horizon case, the assumptions above only guarantee that the DAE model has an underlying ODE description. However, the functions $\mathcal{L}$ and $\mathcal{R}$ may be hard or even impossible to express in closed form. Therefore, their power series expansions are desired and Assumption A5 is reformulated as follows.

Assumption A8. The functions $\hat{F}_1$ and $\hat{F}_2$ in (4.38), and L in (4.37) are real analytic in $\mathcal{W}$, which is a neighborhood of the origin $(\dot{x}_1, x_1, x_3, u) = 0$.

Analyticity of the functions involved makes it possible to write them as power series

$$\hat{F}_1(\dot{x}_1, x_1, x_3, u) = -E_1\dot{x}_1 + A_{11}x_1 + A_{12}x_3 + B_1u + \hat{F}_{1h}(\dot{x}_1, x_1, x_3, u) \tag{4.40a}$$
$$\hat{F}_2(x_1, x_3, u) = A_{21}x_1 + A_{22}x_3 + B_2u + \hat{F}_{2h}(x_1, x_3, u) \tag{4.40b}$$
$$L(x, u) = x^TQx + u^TRu + 2x^TSu + L_h(x, u) \tag{4.40c}$$

that are convergent in $\mathcal{W}$. The functions $\hat{F}_{1h}$ and $\hat{F}_{2h}$ include terms beginning with order two, while the higher order terms of the performance criterion, *i.e.*, L_h, is of order three at least. The linearization of the system (4.38) can easily be found as

$$\dot{x}_1 = \hat{A}x_1 + \hat{B}u \tag{4.41a}$$
$$x_3 = -A_{22}^{-1}A_{21}x_1 - A_{22}^{-1}B_2u \tag{4.41b}$$

where

$$\hat{A} = E_1^{-1}\big(A_{11} - A_{12}A_{22}^{-1}A_{21}\big) \tag{4.41c}$$
$$\hat{B} = E_1^{-1}\big(B_1 - A_{12}A_{22}^{-1}B_2\big) \tag{4.41d}$$

The objective is, as for the finite horizon case, to find the optimal feedback control locally around the origin. However, because of the infinite horizon, the considered class of feedback laws needs to satisfy some extra conditions.

Assumption A9. The considered feedback laws are described by uniformly convergent power series

$$u(x_1) = Dx_1 + u_h(x_1) \tag{4.42}$$

where $u_h(x_1)$ are terms of at least order two. Furthermore,

$$\operatorname{Re} \operatorname{eig}(\hat{A} + \hat{B}D) < \min(0, -\tfrac{\lambda}{2})$$

where $\hat{A}$ and $\hat{B}$ are given in (4.41).

The last part of the assumption is introduced for two reasons. First, it is necessary for the proof to have a feedback law that stabilizes the system, and thereby makes a neighborhood of the origin invariant. Second, the control law needs to ensure convergence of the integral criterion locally.

As can be seen above, the discount factor can be used to obtain a controller for which the linearization of the closed-loop achieves a prescribed degree of stability. That is, the poles are placed to the right of some specific limit. For linear systems, this fact was developed already in Anderson and Moore (1969) and has later been developed in numerous publications so that so that different pole placements can be chosen. In this chapter, the result in Anderson and Moore (1969) will be generalized to the nonlinear case. An interesting fact is that despite an explicitly time-varying cost function, the optimal solution will still be time-invariant. In Anderson and Moore (1971), it is shown that the function $e^{\lambda t}$ is in principle the only time varying element that can be allowed in order to have this property.

4.4.2 State-Space Models

If the DAE model has no constraints and is explicit in the states, the optimal control problem can be written as

$$\begin{aligned} V(x_0) = \inf_{u(\cdot)} & \int_0^\infty L(x,u) e^{\lambda t}\, dt \\ \text{s.t.} \quad & \dot{x} = F(x,u) \\ & x(0) = x_0 \end{aligned} \tag{4.43}$$

The solution to this optimal control problem is given by the following Hamilton-Jacobi-Bellman equation (HJB), see Bardi and Capuzzo-Dolcetta (1997),

$$0 = \min_u L(x,u) + \lambda V(x) + V_x(x) F(x,u)$$

and the optimal feedback law $u_*(x)$ must solve the equations

$$\begin{aligned} 0 &= L\big(x, u_*(x)\big) + \lambda V(x) + V_x(x) F\big(x, u_*(x)\big) \\ 0 &= L_u\big(x, u_*(x)\big) + V_x(x) F_u\big(x, u_*(x)\big) \end{aligned} \tag{4.44}$$

As for the finite horizon case, it can be shown that under the given assumptions and for analytic feedback laws, the optimal control problem (4.43) has a solution, see for the

undiscounted case Al'brekht (1961); Lee and Markus (1967); Lukes (1969) and for the discounted case the extension in Section 4.7. Furthermore, the corresponding optimal return function will be analytic in a neighborhood of the origin and can therefore be expressed as a power series. As for the finite horizon case, the equations (4.44) lead to two polynomial equations in x with coefficients including the unknown parameters in V and u_*. The following theorem states when an analytic solution exists and how it can be computed for state-space systems.

Theorem 4.3
Consider the optimal control problem (4.43), satisfying Assumptions A8 and A9. Furthermore, assume that the quadratic part of the cost function satisfies $\left(\begin{smallmatrix} Q & S \\ S^T & R \end{smallmatrix}\right) \succ 0$. *Then there exists an optimal feedback law* $u_*(x)$ *satisfying Assumption A9 if the ARE*

$$0 = (A + \tfrac{\lambda}{2}I)^T P + P(A + \tfrac{\lambda}{2}I) - (PB + S)R^{-1}(PB + S)^T + Q \tag{4.45a}$$

has a unique positive-semidefinite solution such that the matrix $A + BD$ *with* D *given by*

$$D = -R^{-1}(S^T + B^T P) \tag{4.45b}$$

satisfies

$$\operatorname{Re}\operatorname{eig}(A + BD) < \min(0, -\tfrac{\lambda}{2}) \tag{4.46}$$

The equations (4.45) also determines the lowest order terms in $V(x)$ *and* $u_*(x)$*, respectively. The higher order terms in* $V(x)$ *and* $u_*(x)$ *can be computed recursively by*

$$\begin{aligned} V_x^{[m]}(x)A_c x + \lambda V^{[m]}(x) = &- \sum_{k=3}^{m-1} V_x^{[k]}(x)Bu_*^{[m-k+1]}(x) \\ &- \sum_{k=2}^{m-1} V_x^{[k]}(x)F_h^{[m-k+1]}(x, u_*) - L_h^{[m]}(x, u_*) \\ &- 2\sum_{k=2}^{\lfloor \frac{m-1}{2} \rfloor} u_*^{[k]}(x)^T R u_*^{[m-k]}(x) - u_*^{[m/2]}(x)^T R u_*^{[m/2]}(x) \end{aligned} \tag{4.47a}$$

where $m = 3, 4, \ldots$ *and* $A_c = A + BD_*$*, and*

$$u_*^{[k]}(x) = -\frac{1}{2}R^{-1}\Big(V_x^{[k+1]}(x)B + \sum_{i=1}^{k-1} V_x^{[k-i+1]}(x)F_{h;u}^{[i]}(x, u_*) + L_{h;u}^{[k]}(x, u_*)\Big) \tag{4.47b}$$

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 (4.47) we use the conventions that $\sum_k^l = 0$ for $l < k$ and that the terms $u^{[m/2]}$ are to be omitted if m is odd.

Proof: For the case with $\lambda = 0$, the proof is found in Lukes (1969), while the general proof can be found in Section 4.7.1. That (4.47) can be solved uniquely is proved using Lemma 4.9. □

The theorem above is formulated in terms of the ARE (4.45a). The following lemma shows some typical situations when the ARE has a solution satisfying the conditions.

Lemma 4.1
Consider the ARE (4.70). Assume the assumptions in Theorem 4.5 are satisfied. Then there exists a unique positive semi-definite solution such that the eigenvalues of $A + BD$ satisfies condition (4.68) if

- $\lambda = 0$: (A, B) is stabilizable.
- $\lambda > 0$: (A, B) is controllable or $(A + \frac{1}{2}\lambda I, B)$ is stabilizable.
- $\lambda < 0$: $(A + \frac{1}{2}\lambda I, B)$ is stabilizable or (A, B) controllable, and the solution yields $\mathrm{eig}(A + BD) < 0$.

Proof: See for example Bittanti et al. (1991) and Anderson and Moore (1971). □

Note that under the assumptions about stabilizability or controllability made above, it is ensured that there exists a unique positive semi-definite solution such that with D in (4.71), the eigenvalues of $A + BD$ will have real parts less than $-\lambda/2$ (which is necessary in order to obtain a convergent performance criterion). However, in the case when $\lambda < 0$, it means that the eigenvalues need not satisfy condition (4.68) and an extra condition is therefore added in the corollary. Since the extra condition is included it is not guaranteed from the problem data that a solution exists, but at least in randomly generated problems it actually seems to happen quite often. For these cases the optimal feedback law is found.

Having the solution to the ARE, it follows using the same line of reasoning as in the finite horizon case that the optimal solution can be computed recursively. That is, first the lowest order terms are obtained as

$$u_*^{[1]}(x) = D_* x, \quad V^{[2]}(x) = x^T P x$$

and having these, the higher order terms in $V(x)$ and $u_*(x)$ are obtained uniquely from (4.47), in the sequence

$$V^{[3]}(x), u_*^{[2]}, V^{[4]}(x), u_*^{[3]}, \ldots$$

to any order.

The equations obtained from the higher order terms in (4.47) are linear in the coefficients from $V^{[m]}$ and $u_*^{[m-1]}$, both separately and simultaneously. If solved recursively, rather high orders can be computed. However, the size of the set of equations grows rather fast with the number of states as will be seen in Chapter 6, which limits the size of the problems that can be handled.

A small remark is that it is not necessary to have a system that is analytic. If the model is $\mathcal{C}^r$ it has been shown in Krener (2001) that the optimal return function up to $\mathcal{C}^{r-2}$ exists and can be computed as above.

4.4.3 DAE Models

In order to solve the optimal control problem described by (4.37) and (4.38), the problem is rewritten as the following equivalent optimal control problem

$$V(x_{1,0}) = \inf_{u(\cdot)} \int_0^\infty \hat{L}(x_1, u) e^{\lambda t}\, dt$$

subject to the dynamics

$$\dot{x}_1 = \mathcal{L}(x_1, u)$$

where $\mathcal{L}$ is given by (4.39a) and

$$\hat{L}(x_1, u) = L\big(x_1, \mathcal{R}(x_1, u), u\big) \tag{4.48}$$

This reformulation can always be done because of Assumption A7.

In principle, the optimal control problem is then a standard problem in the state variables x_1 and Theorem 4.3 is directly applicable. However, as mentioned earlier, there is a major computational barrier, namely that $\mathcal{L}$ and $\mathcal{R}$ are usually not explicitly given. However, Assumptions A7 and A8 ensure the existence of convergent power series of $\mathcal{L}(t, x_1, u)$ and $\mathcal{R}(t, x_1, u)$, and using the methods in Section 4.3.1, these power series can be derived to any degree.

Summarizing this section we have the following result.

Theorem 4.4
Consider the optimal control problem (4.37) and (4.38). Assume that it satisfies Assumptions A3, A7 and A8. Furthermore, assume that the quadratic part of the cost function satisfies $\left(\begin{smallmatrix} \hat{Q} & \hat{S} \\ \hat{S}^T & \hat{R} \end{smallmatrix}\right) \succ 0$. *Then an optimal feedback law* $u_*(x)$ *satisfying Assumption A9 exists if the ARE*

$$0 = (\hat{A} + \tfrac{\lambda}{2} I)^T P + P(\hat{A} + \tfrac{\lambda}{2} I) - (P\hat{B} + \hat{S})\hat{R}^{-1}(P\hat{B} + \hat{S})^T + \hat{Q} \tag{4.49a}$$

has a unique positive-semidefinite solution such that the matrix $\hat{A} + \hat{B}D$ *with* D *given by*

$$D = -\hat{R}^{-1}(\hat{S}^T + \hat{B}^T P) \tag{4.49b}$$

satisfies

$$\operatorname{Re} \operatorname{eig}(\hat{A} + \hat{B}D) < \min(0, -\tfrac{\lambda}{2})$$

The higher order terms are given by (4.47) with the system and cost function replaced by $\mathcal{L}(x_1, u)$ *and* $\hat{L}(x_1, u)$*, respectively.*

Proof: Follows the line of the proof of the finite horizon case, but by using Theorem 4.3 instead. □

Some cases for which the ARE (4.49a) has solutions satisfying the conditions is provided by Lemma 4.1. The computation of the higher order terms can be done similarly to the earlier shown results. That is, the optimal solution can be found recursively and the sequence is the same as in (4.31).

4.5 Conditions on the Original Model and Cost Function

The conditions in Theorems 4.2 and 4.4 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 consider the condition

$$\begin{pmatrix} \hat{Q}(t) & \hat{S}(t) \\ \hat{S}^T(t) & \hat{R}(t) \end{pmatrix} \succ 0 \tag{4.50}$$

Since the variable transformation matrix $\Pi(t)$ in (4.27) has full column rank, it follows that

$$\begin{pmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{pmatrix} \succ 0 \quad \Rightarrow \quad \begin{pmatrix} \hat{Q}(t) & \hat{S}(t) \\ \hat{S}^T(t) & \hat{R}(t) \end{pmatrix} \succ 0$$

However, note that the arrow only goes in one direction and the cost matrix (4.50) may be positive definite also for indefinite matrices in the original problem, as will be seen in Section 4.6.1.

In some cases, it is not desirable to penalize the variables x_3. In these cases, the cost matrix is given by

$$\begin{pmatrix} \hat{Q}(t) & \hat{S}(t) \\ \hat{S}^T(t) & \hat{R}(t) \end{pmatrix} = \Pi^T \begin{pmatrix} Q_{11}(t) & 0 & S_1(t) \\ 0 & 0 & 0 \\ S_1^T(t) & 0 & R(t) \end{pmatrix} \Pi(t) = \begin{pmatrix} Q_{11}(t) & S_1(t) \\ S_1^T(t) & R(t) \end{pmatrix}$$

which means that if the cost matrix for x_1 and u is positive definite, the cost matrix for the reduced system is so as well.

In 4.1 with modifications to the DAE case, there are some conditions about controllability and stabilizability. These can be shown equivalent to corresponding notions in the DAE literature. For example, stabilizability of

$$\left(\hat{A},\ \hat{B}\right) = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}),\ B_1 - A_{12}A_{22}^{-1}B_2\right) \tag{4.51}$$

is equivalent to stabilizability of the linearization of the DAE model in DAE sense. Here $E_1^{-1} = I$ is assumed for notational reasons.

Lemma 4.2
Assume that A_{22} has full rank, and that $E = \left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Then (4.51) is stabilizable if and only if

$$E\dot{x} = Ax + Bu$$

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

$$E\dot{x} = (A + BK)x \tag{4.52}$$

is asymptotically stable according to Section 2.6.2.

Proof: Dai (1989) guarantees the existence of a K such that (4.52) is stable if and only if

$$\operatorname{rank}\begin{pmatrix} sE - A & B \end{pmatrix} = n, \quad \forall s \in \mathbb{C}^+$$

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

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

which proves the lemma since A_{22} is assumed to have full rank. □

In the same way, controllability of (4.51) can be shown equivalent to R-controllability of the linearization of the DAE model, see Dai (1989).

Lemma 4.3
Assume that A_{22} has full rank, and that $E = \left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Then (4.51) is controllable if and only if

$$E\dot{x} = Ax + Bu$$

is R-controllable.

Proof: Follows immediately from Theorem 2-2.2 in Dai (1989), since $\hat{A}$ and $\hat{B}$ are the system matrices for the dynamic part of the linearization. □

However, note that in general, the linearization of the DAE need not be controllable in DAE sense, since that would require full row rank of B_2. The intuition for this is that since there is no memory connected to the algebraic variables x_3, it must be possible to place them arbitrarily using the input pointwise.

Another controllability concept for DAE models is impulse controllability, but it will not be used in this thesis. However, a linear DAE model with A_{22}^{-1} nonsingular, can be shown to be impulse controllable.

4.6 Examples

In order to illustrate the methods described in this chapter, two small examples are included here. The first example is an infinite horizon problem since this makes the notation less tedious. The second example is a finite horizon problem involving a time-varying electrical circuit.

4.6.1 A Phase-Locked Loop Circuit

Consider a Phase-Locked Loop circuit (PLL) given by the model

$$\begin{aligned}\dot{z}_1 &= z_2 \\ \dot{z}_2 &= e^{z_3} - 1 + \frac{1}{2}u \\ 0 &= z_1 - \arcsin(1 - e^{z_3} + \frac{1}{2}u)\end{aligned} \tag{4.53}$$

Algorithm 4.1 Computation of the locally optimal solution in practice.

1. Compute the power series expansions of the cost function up to the desired order m_m and the system up to order $m_m - 1$.
2. Compute $\mathcal{L}(t, x_1, u)$ and $\mathcal{R}(t, x_1, u)$ as described in Section 4.3.1 up to order $m_m - 1$.
3. Solve (4.29) to obtain $V^{[2]}(t, x_1) = x_1^T P(t) x_1$ and $u_*^{[1]}(t, x_1) = D_*(t) x_1$, or the corresponding expressions for the infinite horizon case. Use these expressions to compute $\mathcal{L}^{[1]}(t, x_1, u_*^{[1]})$ and $\mathcal{R}^{[1]}(t, x_1, u_*^{[1]})$.
4. Create parametrized solutions of $V(t, x_1)$ and $u_*(t, x_1)$ of desired order, with the lowest order terms given as above.
5. Substitute the parametrized solutions and the power series of $\mathcal{L}$ and $\mathcal{R}$ into (4.12) or (4.44). Let $m = 3$.
6. Extract the coefficients for terms of order m and $m - 1$ in the equation for V and u_*, respectively. Solve the corresponding linear set of equations, which yields $V^{[m]}$ and $u_*^{[m-1]}$. If m is equal to the m_m, then stop iterate.
7. Otherwise, substitute the solution into (4.12) or (4.44), increase m by one, and repeat from 6.

and let $z = (z_1, z_2, z_3)$. The PLL is used to control an oscillator in order to maintain a constant phase angle relative to a reference signal. The objective is to find a feedback law (4.42) which minimizes the performance criterion with the cost function chosen as

$$L(z, u) = \frac{1}{2} z_1^2 + 2 z_1 z_2 + z_2^2 + z_1(-e^{z_3} + 1 + \frac{1}{2} u) + \frac{1}{2} u^2 \tag{4.54}$$

and with the discount factor chosen as $\lambda = 0$. The cost function is chosen such that the problems has an analytic solution.

Grouping the variables as $x_1 = (z_1,\ z_2)^T$ and $x_3 = z_3$ gives a model in semi-explicit form. The power series of $\mathcal{L}(x_1, u)$ is therefore obtained simply as the composition of the power series of $\hat{F}_1$ and $\mathcal{R}$. The system is composed of elementary functions which are real analytic and Taylor series expansion of $\hat{F}_1(z, u)$ and $\hat{F}_2(z, u)$ around the origin gives the first order terms

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1/2 \\ -1/2 \end{pmatrix} \tag{4.55}$$

and the higher order terms up to order three as

$$\hat{F}_{1h}(z, u) = \begin{pmatrix} 0 \\ \frac{1}{2} z_3^2 + \frac{1}{6} z_3^3 \end{pmatrix} \tag{4.56a}$$

$$\hat{F}_{2h}(z, u) = \frac{1}{2} z_3^2 + \frac{1}{3} z_3^3 - \frac{1}{4} u z_3^2 + \frac{1}{8} u^2 z_3 - \frac{1}{48} u^3 \tag{4.56b}$$

The matrix $A_{22} = 1$ is nonsingular and Assumption A7 is therefore satisfied. The power series of $\mathcal{R}(z_1, z_2, u)$ is computed as described in Section 4.3.1 and the result is the first order terms

$$\mathcal{R}^{[1]}(z_1, z_2, u) = -z_1 + \frac{1}{2}u \tag{4.57}$$

and the higher order terms up to order three

$$\mathcal{R}_h(z_1, z_2, u) = -\frac{1}{2}z_1^2 + \frac{1}{2}z_1 u - \frac{1}{8}u^2 - \frac{1}{6}z_1^3 + \frac{1}{2}uz_1^2 - \frac{1}{4}u^2 z_1 + \frac{1}{24}u^3 \tag{4.58}$$

The system matrices for the reduced system, *i.e.*, $\hat{A}$ and $\hat{B}$, and the local state variable change Π can then be computed as

$$\hat{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \Pi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix} \tag{4.59}$$

Using the power series expansion of $\mathcal{R}$, the power series of $\hat{L}$ around the origin up to the fourth order becomes

$$L(z, u) = \begin{pmatrix} z \\ u \end{pmatrix}^T \begin{pmatrix} Q_{11} & Q_{12} & S_1 \\ Q_{12}^T & Q_{22} & S_2 \\ S_1^T & S_2^T & R \end{pmatrix} \begin{pmatrix} z \\ u \end{pmatrix} - \frac{1}{2}z_1 z_3^2 - \frac{1}{6}z_1 z_3^3 \tag{4.60}$$

where

$$Q_{11} = \begin{pmatrix} 1/2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad Q_{12} = \begin{pmatrix} -1/2 \\ 0 \end{pmatrix}, \qquad Q_{22} = 0$$
$$S_1 = \begin{pmatrix} 1/4 \\ 0 \end{pmatrix}, \qquad S_2 = 0, \qquad R = \frac{1}{2}$$

and the sizes corresponds to the size of x_1, x_3 and u, respectively. The cost matrix in (4.60) is indefinite, but after a transformation using Π, the cost matrix for the reduced model becomes

$$\begin{pmatrix} \hat{Q} & \hat{S} \\ \hat{S}^T & \hat{R} \end{pmatrix} = \begin{pmatrix} 3/2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1/2 \end{pmatrix} \tag{4.61}$$

which is positive definite.

Since $(\hat{A}, \hat{B})$ is stabilizable and the cost matrix (4.61) is positive definite, it follows from Theorem 4.4, that the optimal control problem has a unique analytic solution. The first terms in the approximation can be computed as described in (4.49) which gives the result

$$V^{[2]}(x_1) = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}^T \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \quad u_*^{[1]}(x_1) = \begin{pmatrix} -1 & -2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \tag{4.62a}$$

The corresponding closed-loop system matrix becomes

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

with the eigenvalues $\lambda = -1 \pm 1i$ and asymptotic stability is obtained. The higher order terms become

$$V^{[3]}(x_1) = 0, \quad V^{[4]}(x_1) = -\frac{1}{12}z_1^4 \tag{4.62b}$$

and

$$u_*^{[2]}(x_1) = 0, \quad u_*^{[3]} = 0 \tag{4.62c}$$

Notice that 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.

In order to validate the solutions of the power series method the explicit solution to the optimal control problem is computed, which is possible in this case. The system (4.53) can be formulated in state-space form as

$$\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= -\sin(z_1) + u \end{aligned}$$

with the cost function 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 equations (4.44), the explicit expressions are given by

$$u_*(z_1, z_2) = -z_1 - 2z_2 \tag{4.63a}$$

$$V(z_1, z_2) = 2\big(1 - \cos(z_1)\big) + z_1z_2 + z_2^2 \tag{4.63b}$$

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 power series of (4.63b) to the fourth order is computed as

$$V(z_1, z_2) = z_1^2 + z_1z_2 + z_2^2 - \frac{1}{12}z_1^4 \tag{4.64}$$

Comparison of the power series expansions of the explicit solutions (4.63) with the solutions obtained using the power series method derived in Section 4.4.3, *i.e.*, (4.62), shows that the same expressions are attained. Figure 4.1 shows a comparison between the optimal solution and a fourth order approximation of it. As can be seen, the difference is not very large in the plotted region. In Figure 4.2, a third order approximation of the optimal feedback law is depicted. Though, in this particular case, the optimal solution is of order one and the approximation will then be of order one too.

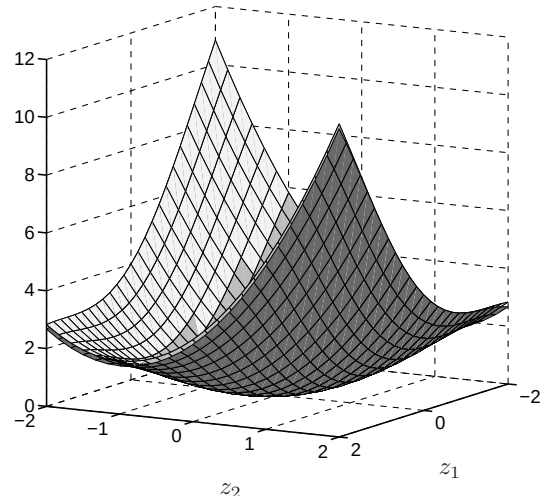


Figure 4.1: *The fourth order approximation of $V(u_C, \Phi)$ (dark grey) compared with the optimal solution $V(u_c, \Phi)$ (light grey) for the PLL example.*

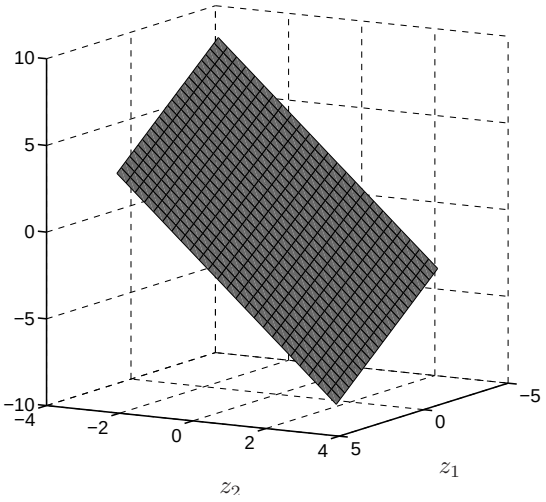


Figure 4.2: *The third order approximation of $u_*(u_C, \Phi)$ for the PLL example. In this case the approximation is of order one since that is the order of the optimal solution.*

4.6.2 An Electrical Circuit

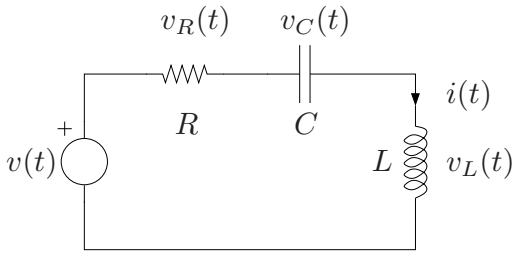


Figure 4.3: *Electrical ciruit*

The electrical circuit, depicted in Figure 4.3, consists of an ideal voltage source, an inductor with a ferromagnetic core, a capacitor and a resistor. Because of the ferromagnetic core, the flux of the inductor saturates for large currents. Furthermore, the flux is assumed to decrease with time and for large currents to model temperature dependence. The capacitor has a voltage dependent capacitance and the resistor depends both linearly and cubicly on the current. The complete model can then be written as

$$\dot{u}_C = \tfrac{i}{1+10^{-2}u_C} \tag{4.65a}$$
$$\dot{\Phi} = u_L \tag{4.65b}$$
$$0 = \Phi - \tfrac{\arctan(i)}{1+10^{-1}t+10^{-2}i^2} \tag{4.65c}$$
$$0 = u_R - i - i^3 \tag{4.65d}$$
$$0 = u - u_R - u_C - u_L \tag{4.65e}$$

where u_C is the voltage over the capacitor, Φ is the flux, u_L is the voltage over the inductor, i is the current, u_R is the voltage over the resistor and u is the voltage over the voltage source. The dynamic variables are in this case chosen as $x_1 = (u_C, \Phi)$ and the algebraic variables are $x_3 = (i, u_L, u_R)$. The control signal is the voltage over the voltage source u. This model satisfies Assumptions A4, A5, and A6. The cost functions are chosen as

$$L(u_C, \Phi, i, u_L, u_R) = i^2 + i^4 + \tfrac{1}{2}u^2$$
$$G(u_C, \Phi) = 0$$

with no discount factor and the final time is chosen as $T = 4$.

Applying the method in Section 4.3 gives the third order approximation of the optimal cost function as

$$\begin{aligned} V(t, u_C, \Phi) \approx\ & p_{11}(t)u_C^2 + 2p_{12}(t)u_C\Phi + p_{22}(t)\Phi^2 \\ & + a_{30}(t)u_C^3 + a_{21}(t)u_C^2\Phi + a_{12}(t)u_C\Phi^2 + a_{03}(t)\Phi^3 \end{aligned}$$

where the coefficients can be found in Figure 4.4. The corresponding second order approximation of the optimal control feedback is

$$u_*(t, u_C, \Phi) \approx d_1(t)u_C + d_2(t)\Phi + b_{20}(t)u_C^2 + b_{11}(t)u_C\Phi + b_{02}(t)\Phi^2$$

where the coefficients can be found in Figure 4.5.

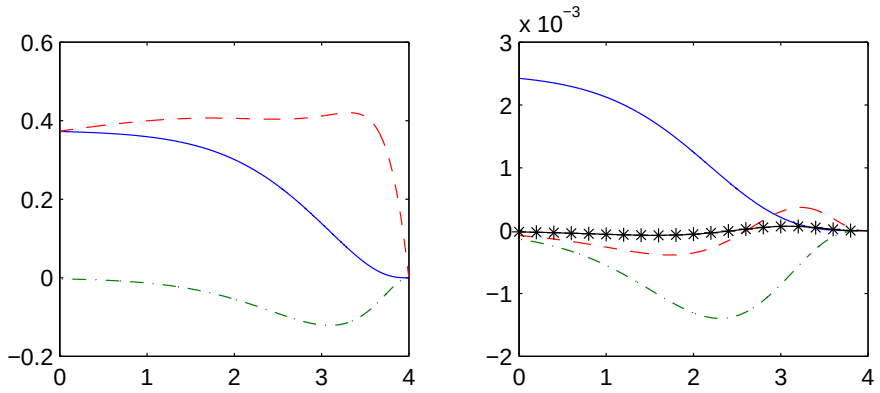


Figure 4.4: *Left: The second order terms of V – solid: $p_{11}(t)$, dash-dotted: $p_{12}(t)$, dashed: $p_{22}(t)$, Right: The third order terms of V – star-marked: $a_{30}(t)$, dashed: $a_{21}(t)$, dash-dotted: $a_{12}(t)$, solid: $a_{03}(t)$*

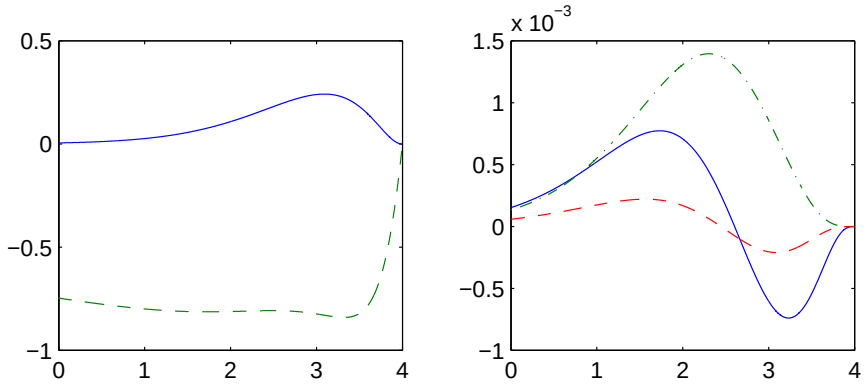


Figure 4.5: *Left: The first order terms of u_* – solid: $d_1(t)$, dashed: $d_2(t)$, Right: The second order terms of u_* – solid: $b_{20}(t)$, dash: $b_{11}(t)$, dash-dotted: $b_{02}(t)$*

4.7 Proofs

This section contains the proofs of Theorem 4.3 and Lemma 4.9

4.7.1 Proof of Theorem 4.3

The discounted case is to a large extent similar to the undiscounted case presented in Lukes (1969). Nevertheless, there are differences in most of the theorems and lemmas and their corresponding proofs. Therefore, reformulated versions of each of the theorems and lemmas are presented in this section. The proofs presented here are however most often not complete. Instead only the main idea and major differences are included to keep them short. To simplify the reading, the numbering in Lukes (1969) is presented in parentheses at the beginning of each proof.

Consider the system

$$\dot{x} = F(x, u) \tag{4.66}$$

and the performance criterion

$$J\big(x(0), u\big) = \int_0^\infty L(x, u) e^{\lambda t}\, dt \tag{4.67}$$

where $\lambda \in \mathbb{R}$ is constant and denoted the discount factor. The discount factor λ may be positive, zero or negative.

Throughout this section, the basic assumption in Assumption A9 will be made, which here is repeated with minor changes since only state-space models are considered in this section.

Assumption A10. The considered feedback laws are described by uniformly convergent power series $u(x) = Dx + u_h(x)$ such that

$$\operatorname{Re}\operatorname{eig}(A + BD) < \min(0, -\tfrac{\lambda}{2}) \tag{4.68}$$

Together with Assumption A8, it follows that the closed-loop system has the form

$$\dot{x} = F\big(x, u(x)\big) = (A + BD)x + F_h\big(x, u(x)\big) \tag{4.69}$$

where $F_h(\cdot)$ is a uniformly convergent power series around the origin, beginning with terms of order two. The solution to the closed-loop system will be denoted $x(t, x_0)$ where $x_0 \in \mathbb{R}^n$ is the initial value, that is, $x(0, x_0) = x_0$. Sometimes the initial value will be complex $z_0 \in \mathbb{C}^n$, and then complex solutions are considered.

Condition (4.68) in Assumption A10 first of all ensure that there exists an invariant neighborhood around the origin for the closed-loop system. This is essential for the proof. For a small enough neighborhood, it also ensures the following inequality

$$|x(t, x_0)| \le C_0 e^{\mu t} |x_0|$$

for some $\operatorname{Re}\operatorname{eig}(A + BD) < \mu < \min(0, -\frac{\lambda}{2})$. For cases with negative discount or a discount equal to zero, the convergence rate above will also be enough to ensure convergence of the performance criterion as will be seen later. However, for cases with positive discount the rate of convergence of $x(t, x_0)$ needs to be at least $e^{-\lambda/2t}$, which is also satisfied for control laws that satisfy (4.68).

The main theorem can then be formulated as follows.

Theorem 4.5 (Theorem 1.1 in Lukes (1969))
Consider a model (4.66) and a performance criterion (4.67) that satisfy Assumption A8. Furthermore, assume that the quadratic part of the cost function satisfies $\begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \succ 0$. *Then there exists an optimal feedback law* $u_*(x)$ *satisfying Assumption A10 if the ARE*

$$0 = (A + \tfrac{\lambda}{2}I)^T P + P(A + \tfrac{\lambda}{2}I) - (PB + S)R^{-1}(PB + S)^T + Q \tag{4.70}$$

has a unique positive-semidefinite solution such that matrix $A + BD$ *satisfies (4.68) with*

$$D = -R^{-1}(S^T + B^T P) \tag{4.71}$$

The optimal solution solves

$$0 = F_u\big(x_0, u_*(x_0)\big) J_x(x_0, u_*) + L_u\big(x_0, u_*(x_0)\big) \tag{4.72a}$$
$$0 = \lambda J(x_0, u_*) + F\big(x_0, u_*(x_0)\big) J_x(x_0, u_*) + L\big(x_0, u_*(x_0)\big) \tag{4.72b}$$

for all x_0 *near the origin and is unique in that*

1. u_* *is the unique real analytic solution.*
2. u_* *is the unique solution satisfying Assumption A10.*
3. u_* *synthesizes the unique optimal open-loop control.*

The theorem above is formulated in terms of the ARE (4.70). Lemma 4.1 shows some typical situations when the ARE has a unique solution satisfying the conditions.

The proof of the main theorem includes a number of steps and therefore a brief description of each step is given. In consecutive steps it is shown that:

- For each $u(x)$ that satisfies Assumption A10, $J(z_0, u)$ will converge locally and satisfy (4.72b).
- Equation (4.72a) can be solved for u when $J_x(\cdot)$ has been replaced by an arbitrary constant p. The result is denoted $u_*(x, p)$ (*i.e.*, algebraic solvability when p is independent of u_*, which $J_x(\cdot)$ is not).
- If there exists a $u_*(x)$ that satisfies Assumption A10 and solves (4.72a) with $J(x, u_*)$, this $u_*(x)$ is the optimal control (here $J(x, u_*)$ denotes the return function with $u_*(x)$ as feedback law).
- If the ARE (4.70) has a solution that satisfies condition (4.68) with D given by (4.71), the nonlinear Hamiltonian system (4.78) will have a n-dimensional stable manifold described by $p_*(x)$. Furthermore, with $u_*(x, p_*(x))$, the optimal return will satisfy $J_x(x_0, u_*) = p_*(x_0)$ and hence (4.72a).

The first part of the proof is to show that for feedback laws satisfying Assumption A10, the performance criterion will converge and satisfy the HJB.

Lemma 4.4 (Lemma 2.1)
For each feedback law $u(x)$ *satisfying Assumption A10 there exists an invariant neighborhood* N_u^c *around the origin in* $\mathbb{C}^n$ *in which* $J(z_0, u) = z_0^T P z_0 + J_h(z_0)$ *is analytic in* z_0.

The function $J_h(z_0)$ is a power series beginning with third order terms and converging in N_u^c. The positive definite matrix P is given by

$$P = \int_0^\infty e^{(A+BD+\lambda/2I)^T t}(Q + S^T D + D^T S^T D^T RD)e^{(A+BD+\lambda/2I)t}\, dt \tag{4.73}$$

For $z_0 \in N_u^c$, $J(z_0, u)$ satisfies

$$0 = \lambda J(z_0, u) + J_z(z_0, u)F\big(z_0, u(z_0)\big) + L\big(z_0, u(z_0)\big) \tag{4.74}$$

The performance criterion $J(x_0, u)$ is a real power series for real-valued x_0.

Proof: From the eigenvalue condition it follows that there exists a μ such that $\mathrm{Re}(A + BD) < \mu < \min(0, -\lambda/2)$. Therefore, there is a neighborhood N_u^c of the origin in $\mathbb{C}^n$ where the solution $x(t, z_0)$ to (4.69) initiated in N_u^c remains in N_u^c for all $t \geq 0$ and satisfies the inequality

$$|x(t, z_0)| \leq C_1 e^{-\mu t}|z_0|, \quad 0 \leq t < \infty$$

The neighborhood N_u^c can be chosen small enough so that $|u(z)| \leq C_2|z|$ and $|L\big(z, u(z)\big)| \leq C_3|z|^2$ for positive numbers C_1, C_2 and C_3. This yields the inequalities

$$|u\big(x(t, z_0)\big)| \leq C_1 C_2 e^{\mu t}|z_0|, \quad |L\big(x(t, z_0), u(x(t, z_0))\big)| \leq C_1 C_3 e^{2\mu t}|z_0|$$

where $2\mu < \min(0, -\lambda)$. Therefore, the integral

$$J(z_0, u) = \int_0^\infty L\big(x(t, z_0), u(x(t, z_0))\big)e^{\lambda t}\, dt \tag{4.75}$$

converges uniformly in N_u^c. Since $x(t, z_0)$ and $u\big(x(t, z_0)\big)$ are analytic in $z_0 \in N_u^c$ for each fixed $t \geq 0$ and continuous in (t, z_0) it then follows that (4.75) is analytic for $z_0 \in N_u^c$.

In the same way as in Lukes (1969), it can then be shown that termwise integration of $L\big(x(t, x_0), u(x(t, x_0))\big)$ for real valued $x_0 \in N_u^C \cap \mathbb{R}^n$ is possible, yielding the desired structure of the performance criterion, that is

$$J(x_0, u) = x_0^T P x_0 + J_h(x_0, u)$$

where P is given by (4.73) and $J_h(x_0, u)$ is convergent and begins with third order terms in x_0. Positive definiteness of P follows from the condition on the cost matrix and is not changed by the factor $e^{\lambda t}$.

The next step is to show that $J\big(z, u(z)\big)$ and $u(z)$ has to satisfy (4.74). By the uniqueness of solutions

$$x\big(s, x(t, z_0)\big) = x(s + t, z_0)$$

for small $|z_0|$ and all $t \geq 0$, $s \geq 0$, it follows that

$$J\big(x(t,z_0),u\big) = \int_0^\infty L\big(x(s+t,z_0),u(x(s+t,z_0))\big)e^{\lambda s}\,ds$$
$$= e^{-\lambda t}\int_t^\infty L\big(x(v,z_0),u(x(v,z_0))\big)e^{\lambda v}\,dv$$

Differentiation of both sides in the expression above w.r.t. t, which is possible since $J(z,u)$ is analytic, gives

$$J_x\big(x(t,z_0),u\big)F\big(x(t,z_0),u(x(t,z_0))\big) = -\lambda J\big(x(t,z_0),u\big) - L\big(x(t,z_0),u(x(t,z_0))\big)$$

and by setting $t=0$, the result is

$$0 \equiv \lambda J(z_0,u) + J_{z_0}(z_0,u)F\big(z_0,u(z_0)\big) + L\big(z_0,u(z_0)\big)$$

for small $|z_0|$. □

Lemma 4.5 (Lemma 2.2 in Lukes (1969))
Consider a model (4.66) and a performance criterion (4.67) satisfying Assumption A8. Then

$$0 = L_u(x,u) + F_u(x,u)p \tag{4.76}$$

has a unique analytic solution $u_*(x,p)$ near the origin in $\mathbb{R}^{2n}$ for which $u_*(0,0)=0$. Furthermore,

$$u_*(x,p) = -\tfrac{1}{2}R^{-1}(2S^Tx + B^Tp) + u_{*,h}(x,p) \tag{4.77}$$

where $u_{*,h}(x,p)$ is a convergent power series in a neighborhood of the origin beginning with second order terms.

Proof: See Lukes (1969). □

Lemma 4.6 (Lemma 2.4 in Lukes (1969))
Suppose there exists a feedback law $u_*(x)$ that satisfies Assumption A10 for the analytic system (4.66) and solves (4.72a), that is,

$$0 = L_u\big(x,u_*(x)\big) + F_u\big(x,u_*(x)\big)J(x,u_*)$$

for all x in a neighborhood of the origin in $\mathbb{R}^n$. Then

1. *u_* is the unique analytic solution to (4.72a).*
2. *u_* is the unique solution satisfying Assumption A10.*
3. *u_* synthesizes the unique optimal open-loop control.*

Proof: Consider the real-valued function defined near the origin in $\mathbb{R}^{n+p}$

$$Q(x,u) = \lambda J\big(x,u_*\big) + J_x\big(x,u_*\big)F\big(x,u_*(x)\big) + L\big(x,u_*(x)\big)$$

Using Lemma 4.4 and equation (4.72a), it can be shown that

$$Q\big(x,u_*(x)\big) = 0,\ Q_u\big(x,u_*(x)\big) = 0,\ \text{and } Q_{uu}\big(x,u_*(x)\big) = 2R \succ 0$$

locally around the origin. This means that there exists an $\varepsilon > 0$ such that for $|x| < \varepsilon$ and $|u_1| < \varepsilon$,

$$\begin{aligned} 0 &= \lambda J\big(x,u_*\big) + J_x\big(x,u_*\big)F\big(x,u_*(x)\big) + L\big(x,u_*(x)\big) \\ &\le \lambda J\big(x,u_*\big) + J_x\big(x,u_*\big)F\big(x,u_1\big) + L\big(x,u_1\big) \end{aligned}$$

with strict inequality for $u_1(x) \neq u_*(x)$. The constant ε is also chosen small enough to ensure that $L(x,u) \ge 0$. Let $u_1(x)$ be another control law satisfying Assumption A10 and let $x_*(t)$ and $x_1(t)$ be solutions to (4.66) for the corresponding feedback laws. Multiplying both sides with $e^{\lambda t}$ then gives

$$\begin{aligned} 0 &= \frac{d}{dt}\big(e^{\lambda t}J\big(x_*(t),u_*\big)\big) + e^{\lambda t}L\big(x_*(t),u_*(x_*(t))\big) \\ &\le \frac{d}{dt}\big(e^{\lambda t}J\big(x_1(t),u_*\big)\big) + e^{\lambda t}L\big(x_1(t),u_1(x_1(t))\big) \end{aligned}$$

for initial conditions small enough to keep the solutions within an invariant neighborhood satisfying $|x| < \varepsilon$ and $|u_1(x)| < \varepsilon$. Integrating the last part gives that

$$J(x_0,u_*) \le J(x_0,u_1) \triangleq \int_0^\infty L\big(x_1(t),u_1(x_1(t))\big)e^{\lambda t}\,dt$$

with equality if and only if $u_1(x) = u_*(x)$. Here, we have used that

$$\lim_{t\to\infty} e^{\lambda t}J\big(x_*(t),u_*\big) = \lim_{t\to\infty} e^{\lambda t}J\big(x_1(t),u_*\big) = 0$$

which follows from Assumption A10 for sufficiently small initial values. This shows that $u_*(x)$ is the unique optimal feedback control and the unique solution to (4.72a).

The last part can be proved as in Lukes (1969) using the same reformulation as above. □

The next step is to show that if the ARE (4.70) has a solution which satisfies the conditions in Theorem 4.5 there will exist a feedback law that satisfies Assumption A10 and solves (4.72a). For this end, the Hamiltonian system is studied. The Hamiltonian system for this optimal control problem is

$$\dot{x} = F\big(x,u_*(x,p)\big) \tag{4.78a}$$

$$\dot{p} = -\big(\lambda + F_x(x,u_*(x,p))\big)p - L_x\big(x,u_*(x,p)\big) \tag{4.78b}$$

where $u_*(x,p)$ is obtained from (4.77) as

$$u_*(x,p) = -\frac{1}{2}R^{-1}(2S^Tx + B^Tp) + u_{*,h}(x,p)$$

Remark 4.1. The Hamiltonian system above is obtained by studying the Hamiltonian

$$H(x, \bar{p}, u) = L(x, u)e^{\lambda t} + \bar{p}F(x, u)$$

and using the change of coordinates $\bar{p} = e^{\lambda t}p$.

It will prove useful for the nonlinear case to begin with the linear case. However, only parts needed for the nonlinear case are included. The linear Hamiltonian system in $\mathbb{R}^{2n}$ becomes

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = H \begin{pmatrix} x \\ p \end{pmatrix} \tag{4.79a}$$

where

$$H = \begin{pmatrix} A - BR^{-1}S^T & -\frac{1}{2}BR^{-1}B^T \\ -2(Q - SR^{-1}S^T) & -(A - BR^{-1}S^T + \lambda I)^T \end{pmatrix} \tag{4.79b}$$

Lemma 4.7 (Lemma 2.5 in Lukes (1969))
The linear Hamiltonian system (4.79) in $\mathbb{R}^{2n}$ transforms into

$$\begin{pmatrix} \dot{y} \\ \dot{q} \end{pmatrix} = \begin{pmatrix} A_c & 0 \\ 0 & -(A_c + \lambda I)^T \end{pmatrix} \begin{pmatrix} y \\ q \end{pmatrix}$$

by a nonsingular real linear transformation

$$\begin{pmatrix} y \\ q \end{pmatrix} = M \begin{pmatrix} x \\ p \end{pmatrix}$$

where

$$M = \begin{pmatrix} I - 2Q_*P & Q_* \\ 2P & -I \end{pmatrix}, \quad M^{-1} = \begin{pmatrix} I & Q_* \\ 2P & 2PQ_* - I \end{pmatrix}$$

The expression for A_c is given by

$$\begin{aligned} A_c &= A + BD_* \\ D_* &= -R^{-1}(S^T + B^T P) \end{aligned}$$

while Q_* and P are the positive definite solutions to

$$0 = (A + \tfrac{1}{2}\lambda I)^T P + P(A + \tfrac{1}{2}\lambda I) - (PB + S)R^{-1}(PB + S)^T + Q_* \tag{4.80a}$$

$$0 = (A_c + \tfrac{1}{2}\lambda)Q_* + Q_*(A_c + \tfrac{1}{2}\lambda)^T + \tfrac{1}{2}BR^{-1}B^T \tag{4.80b}$$

Proof: Follows by some cumbersome calculations similar to the calculations in Lukes (1969). □

Then, for linear Hamiltonian systems (4.79) where $A + BD$ becomes Hurwitz there is a linear invariant manifold in which the origin is asymptotically stable. The manifold is described by $q = 0$ or equivalently by $p = 2Px$. We will denote p on the manifold $p_*(x)$, *i.e.*, $p_*(x) = 2Px$. On this manifold the inequalities

$$|x(t, x_0)| \leq C_4 e^{\mu t}|x_0|, \quad |p(t, x_0)| \leq C_5 e^{\mu t}|x_0|$$

are obtained, where $\operatorname{Re}\operatorname{eig}(A_c) < \mu < \min(0, -\lambda/2)$. Note that, naturally, $p_*\big(x(t)\big)$ solves the Hamiltonian system as well.

Now, the results above can be used to show that there exists an analytic manifold on which the solutions $x(t)$ and $p(t)$ converge also for the nonlinear case. Moreover, on the manifold, the convergence is fast enough to obtain a well-defined performance criterion in a neighborhood of the origin in $\mathbb{R}^n$.

Theorem 4.6 (Theorem 2.7 in Lukes (1969))
Consider the nonlinear analytic Hamiltonian system (4.78). Suppose the ARE (4.70) has a unique positive semi-definite solution such that, with D given by (4.71), A_c satisfies (4.68). Then there exists a real n-dimensional analytic invariant manifold S on which the origin is asymptotically stable and the inequalities

$$|x(t, x_0)| \le C_6 e^{\mu t}|x_0|, \quad |p(t, x_0)| \le C_7 e^{\mu t}|x_0|$$

are obtained, where $\operatorname{Re}\operatorname{eig}(A_c)\mu < \min(0, -\lambda/2)$.

Proof: This theorem follows rather straightforwardly from the results about conditional stability in Coddington and Levinson (1985). In Lukes (1969), their Theorem 4.1 is reproduced, but it is only valid when the eigenvalues of the linear Hamiltonian system (4.79) are separated by the imaginary axis. When $\lambda \neq 0$ the eigenvalues are symmetric around $-\lambda/2$ which means that some other cases may happen. However, using different combinations of the theorems in Coddington and Levinson (1985), these cases can be handled as well.

Consider the nonlinear Hamiltonian system (4.78)

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = H \begin{pmatrix} x \\ p \end{pmatrix} + r(x, p) \tag{4.81}$$

where H is given by (4.79b) and

$$r(x, p) = \begin{pmatrix} Bu_{*,h} + F_h\big(x, u_*(x, p)\big) \\ -2Su_{*,h} - L_{h;x}\big(x, u_*(x, p)\big) - F_{h;x}\big(x, u_*(x, p)\big) \end{pmatrix}$$

From Lemma 4.7, it follows that a linear transformation, defined by the matrix M, exists such that the eigenvalues of H can be separated as

$$\begin{pmatrix} \dot{y} \\ \dot{q} \end{pmatrix} = \begin{pmatrix} A_c & 0 \\ 0 & -(A_c + \lambda I)^T \end{pmatrix} \begin{pmatrix} y \\ q \end{pmatrix} + r_M(y, q)$$

where

$$r_M(y, q) = Mr\big(M^{-1}(y, q)\big)$$

The function $r_M(y, q)$ is differentiable and for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $|r_M(y, q) - r_M(u, v)| < \varepsilon|(y - u, q - v)|$ for $|(y, q)| \le \delta$ and $|(u, v)| \le \delta$, since the lowest order terms are of order two. The basic conditions in Theorems 4.1, 4.3, and 4.4 in Coddington and Levinson (1985) are therefore satisfied.

Due to the structure of H, its $2n$ eigenvalues are symmetric around $-\lambda/2$. In order to use the theorems in Coddington and Levinson (1985), three different constellations of the theorems are needed.

The first case is when $\lambda < 0$. Then, it is ensured from (4.68) that there are exactly n eigenvalues in $\mathbb{C}^-$ and Theorem 4.1 can be applied straightforwardly to prove the results below. The second case is when $\lambda > 0$ and $\max \mathrm{Re}\,\mathrm{eig}(A_c) > -\lambda$. Some of the eigenvalues to $A_c + \lambda I$ will then belong to $\mathbb{C}^-$. However, only n of these correspond to a convergence that is fast enough and in Theorem 4.4 it is shown that this happens on a n-dimensional manifold. The third case is when $\lambda > 0$ and $\max \mathrm{Re}\,\mathrm{eig}(A_c) < -\lambda$. Then all the eigenvalues to $A_c + \lambda I$ belong to $\mathbb{C}^+$. In this case, Theorem 4.1 ensures that a stable manifold exists, while Theorem 4.3 guarantees that the convergence rate on the manifold is the desired.

Hence, the result from all three cases are that there exists an open n-dimensional invariant manifold S containing the origin such that for sufficiently large t, the solution $\big(x(t, x_0), p(t, x_0)\big)$ on S satisfies

$$|x(t, x_0)| \leq C_8 e^{\mu t}|x_0|, \quad |p(t, x_0))| \leq C_9 e^{\mu t}|x_0|$$

where $\mathrm{Re}\,\mathrm{eig}(A_c) < \mu < \min(0, -\lambda/2)$. Moreover, for any solution $\big(x(t), p(t)\big)$ not on S at $t = 0$, there exists an $\eta > 0$ such that any solution satisfying $|\big(x(t), p(t)\big)| < \eta$ for $t \geq 0$, satisfies

$$|x(t, x_0)| \geq C_{10} e^{\bar{\mu} t}|x_0|, \quad |p(t, x_0)| \geq C_{11} e^{\bar{\mu} t}|x_0|$$

for some $\bar{\mu} > \min(0, -\lambda/2)$.

Since r_M is analytic, it also follows that S will be analytic. □

A proof of the main theorem can now be presented.

Proof: The only part left to prove is that there exists a solution to (4.72a), *i.e.*,

$$0 = F_u\big(x_0, u_*(x_0)\big) J_x(x_0, u_*) + L_u\big(x_0, u_*(x_0)\big)$$

that satisfy the assumptions because if there is, it is known from Lemma 4.6 that $u_*(x)$ is the optimal control law.

First, define $u_*(x) = u_*\big(x, p_*(x)\big)$. In Theorem 4.6, it was shown that the motion of the nonlinear Hamiltonian (4.78) on S is described by

$$\begin{aligned} \dot{x} &= F\big(x, u_*(x)\big) \\ \dot{p}_*(x) &= -\big(\lambda + F_x\big(x, u_*(x)\big)\big) p_*(x) - L_x\big(x, u_*(x)\big) \end{aligned}$$

for small initial conditions $|x_0|$ and where $p_*(x)$ describes the manifold. Note that here, it has been used that the time derivative of $p_*(x)$ will satisfy the Hamiltonian system.

Using $u_*(x)$, it can be shown that the performance criterion $\int_0^\infty L\big(x, u_*(x)\big) e^{\lambda t}\, dt$ is uniformly convergent. To prove this fact, one can use the inequalities obtained in Theorem 4.6,

$$\left| \begin{pmatrix} x(t, x_0) \\ p_*\big(x(t, x_0)\big) \end{pmatrix} \right| \leq C_{12}|x_0| e^{\mu t}$$

for some $\mathrm{Re}\,\mathrm{eig}(A_c) < \mu < \min(0, -\lambda/2)$, together with the fact that L is at least quadratic in x and u.

The next step is to show that $p_*(x_0) = p_*\big(x(0)\big)$ is equal to $J_x(x_0, u_*)$. First note that

$$\Big|\frac{\partial x(t,x_0)}{\partial x_0}\Big| \le C_{13}e^{\mu t}$$

which can be found by studying the corresponding sensitivity ODE for which the linearization has the eigenvalues given by $\mathrm{eig}(A_c)$.

The uniform convergence together with the continuity and analyticity of the functions in the integrand permit the following differentiation when the initial conditions are real

$$\begin{aligned}
&\frac{\partial J(x_0,u_*)}{\partial x_0} = \frac{\partial}{\partial x_0}\int_0^\infty L\big(x,u_*(x,p_*(x))\big)e^{\lambda t}\,dt\\
&= \int_0^\infty \Big(\frac{\partial x}{\partial x_0}\frac{\partial L(x,u_*)}{\partial x} + \frac{\partial u_*}{\partial x_0}\frac{\partial L(x,u_*)}{\partial u_*}\Big)e^{\lambda t}\,dt\\
&= \Big[\frac{\partial L(x,u_*)}{\partial x} = -\lambda p_*(x) - \dot p_*(x) - \frac{\partial F(x,u_*)}{\partial x}p_*(x)\Big]\\
&= \int_0^\infty \frac{\partial x}{\partial x_0}e^{\frac{\lambda}{2}t}\big(-\frac{\lambda}{2}e^{\frac{\lambda}{2}t}p_*(x) - e^{\frac{\lambda}{2}t}\dot p_*(x)\big)\\
&+ \frac{\partial x}{\partial x_0}\big(-\frac{\partial F(x,u_*(x))}{\partial x}p_*(x)e^{\frac{\lambda}{2}t} - \frac{\lambda}{2}e^{\frac{\lambda}{2}t}p_*(x)\big) + \frac{\partial u_*}{\partial x_0}\frac{\partial L(x,u_*)}{\partial u_*}e^{\lambda t}\,dt\\
&= \Big[\frac{\lambda}{2}e^{\frac{\lambda}{2}t}p_*(x) + e^{\frac{\lambda}{2}t}\dot p_*(x) = \frac{d}{dt}\big(e^{\frac{\lambda}{2}t}p_*(x)\big),\ \frac{\partial L(x,u_*)}{\partial u_*} = -\frac{\partial F(x,u_*)}{\partial u_*}p_*(x)\Big]\\
&= \Big[-\frac{\partial x}{\partial x_0}e^{\frac{\lambda}{2}t}p_*(x)e^{\frac{\lambda}{2}t}\Big]_0^\infty + \int_0^\infty \big(\frac{\lambda}{2}e^{\frac{\lambda}{2}t}\frac{\partial x}{\partial x_0} + e^{\frac{\lambda}{2}t}\frac{d}{dt}\frac{\partial x}{\partial x_0}\big)p_*(x)e^{\frac{\lambda}{2}t}\\
&+ \frac{\partial x}{\partial x_0}e^{\frac{\lambda}{2}t}\big(-\frac{\partial F(x,u_*(x))}{\partial x} - \frac{\lambda}{2}\big)e^{\frac{\lambda}{2}t}p_*(x) - \frac{\partial u_*}{\partial x_0}\frac{\partial F(x,u_*(x))}{\partial x}p_*(x)e^{\lambda t}\,dt\\
&= \Big[\frac{d}{dt}\frac{\partial x}{\partial x_0} = \frac{\partial}{\partial x_0}\frac{dx}{dt} = \frac{\partial F(x,u_*)}{\partial x_0}\Big]\\
&= p_*(x_0) + \int_0^\infty \Big(\underbrace{\frac{\partial F(x,u_*)}{\partial x_0} - \frac{\partial x}{\partial x_0}\frac{\partial F(x,u_*(x))}{\partial x} - \frac{\partial u_*}{\partial x_0}\frac{\partial F(x,u_*)}{\partial u_*}}_{\equiv 0}\Big)e^{\lambda t}p_*(x)\,dt\\
&= p_*(x_0)
\end{aligned}$$

This means that with $p_*(x)$ from the nonlinear Hamiltonian and $u_*(x) = u_*\big(x, p_*(x)\big)$, $J_x(x_0,u_*)$ is equal to $p_*(x_0)$ for small $|x_0|$. The control law $u_*(x_0, p_*(x_0))$ is chosen to satisfy

$$0 = L_u\big(x_0,u_*(x_0)\big) + F_u\big(x_0,u_*(x_0)\big)p_*(x_0)$$

which since $p_*(x_0) = J_x\big(x_0,u_*(x_0)\big)$ becomes (4.72a) and the existence of a solution is proved. □

4.7.2 Proof of Lemma 4.9

In this section, Lemma 4.9 is proved. The proof will use the Kronecker product and properties for such expressions. Therefore, some short facts about Kronecker products are first presented. A more thorough description of this topic can be found in Graham (1981).

Consider the matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times q}$. The Kronecker product is then defined as

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1m}B \\ \vdots & \vdots & & \vdots \\ a_{n1}B & a_{n2}B & \dots & a_{nm}B \end{pmatrix}$$

Based on the definition some basic algebraic rules follows

$$\begin{aligned} A \otimes B \otimes C &= (A \otimes B) \otimes C = A \otimes (B \otimes C) \\ (A + B) \otimes (C + D) &= A \otimes C + A \otimes D + B \otimes C + B \otimes D \\ (A \otimes B)(C \otimes D) &= (AC) \otimes (BD) \end{aligned}$$

The following lemma will be useful in the main proof.

Lemma 4.8
Consider a set of matrices A_i, $i = 1, \dots, n$. Let the eigenvalues of matrix A_i be denoted $\lambda_{i,j}$, $i = 1, \dots, n$. Then the eigenvalues of

$$A_1 \otimes I \otimes \dots \otimes I + I \otimes A_2 \otimes \dots \otimes I + \dots + I \otimes I \otimes \dots \otimes A_n$$

are given by

$$\sum_{i=1}^{n} \{\lambda_{i,j}\}$$

where $\{\lambda_{i,j}\}$ denotes the set of all eigenvalues of matrix A_i. That is, the eigenvalues of the total matrix are all the sums that can be made by combining one arbitrary eigenvalue from each matrix.

Proof: The proof follows easily as an extension of (Graham, 1981, Result XIV, Section 2.4), by noting that

$$\begin{aligned} (A_1 \otimes I \otimes \dots I + I \otimes A_2 \otimes \dots I + \dots + I \otimes I \otimes \dots \otimes A_n)(x_1 \otimes x_2 \otimes \dots \otimes x_n) = \\ A_1x_1 \otimes x_2 \otimes \dots \otimes x_n + x_1 \otimes A_2x_2 \otimes \dots \otimes x_n \dots + x_1 \otimes x_2 \otimes \dots \otimes A_nx_n \end{aligned}$$

□

The main lemma can now be formulated.

Lemma 4.9
Let $P_m(x)$ and $Q_m(x)$ be homogeneous polynomials of order $m \geq 3$, λ be a real constant and A_c be a real square matrix of size n by n with eigenvalues eig_i, $i = 1, \dots, n$. Then an equation of the form

$$\lambda P_m(x) + P_{m;x}(x)A_cx = Q_m(x) \tag{4.82}$$

for all x in a neighborhood, can be solved uniquely for the coefficients in $P_m(x)$ if and only if

$$\lambda + \sum_{k=1}^{n} \{\text{eig}_i\} \neq 0 \tag{4.83}$$

where $\{\text{eig}_i\}$ is the set of the eigenvalues. That is, no sum of n arbitrary chosen eigenvalues are allowed to equal $-\lambda$.

In the case with $\lambda = 0$, the lemma was proved in Lyapunov (1992). However, below a general proof is presented.

Proof: Let

$$P_m(x) = v_P^T x^{\{m\}}, \quad Q_m(x) = v_Q^T x^{\{m\}}$$

where

$$x^{\{m\}} = \underbrace{x \otimes x \otimes \ldots \otimes x}_{m}$$

and v_Q is a given vector. The term $P_{m;x}(x)A_c x$ can then be obtained as

$$\begin{aligned} P_{m;x}(x)A_c x = \frac{d}{dt}P_m(x) = v_P^T \frac{d}{dt} x^{\{m\}} &= \dot{x} \otimes x \otimes \ldots \otimes x + \ldots + x \otimes x \otimes \ldots \otimes \dot{x} \\ &= A_c x \otimes x \otimes \ldots \otimes x + \ldots + x \otimes x \otimes \ldots \otimes A_c x \\ &= v_P^T (A_c \otimes I \otimes \ldots \otimes I + \ldots + I \otimes I \otimes \ldots \otimes A_c) x^{\{m\}} \end{aligned}$$

Since (4.82) must hold identically for x in a neighborhood of the origin in $\mathbb{R}^n$, an equivalent equation for the coefficients in (4.82) is

$$v_P^T \big(\text{eig}\, I + A_c \otimes I \otimes \ldots \otimes I + I \otimes I \otimes \ldots \otimes A_c \big) = v_Q^T$$

From Lemma 4.8, it follows that a unique solution for v_P exists if and only if (4.83) is satisfied. □

For the case with $\lambda = 0$, the most common sufficient condition is that all eigenvalues of A_c are Hurwitz. Then the sum will have a strictly negative real part. In the more general case, condition (4.68) will be used most.

5

Rational Approximation of Optimal Feedback Laws

In Chapter 4, optimal control was considered for systems that can be described by convergent power series. It was shown that under quite natural assumptions, the optimal solution exists. Furthermore, a computational procedure was presented based on series expansion. The advantage of the method in the latter chapter is that, theoretically, the optimal solution can be calculated. However, in practice the optimal return function and feedback law need to be truncated. This chapter deals with some issues that have been seen for these truncated approximations. One such issue is that the truncated solution tends to have bad properties for large values of $|x|$, for example that it most often grows too fast towards infinity.

The method considered in this chapter will instead approximate the optimal cost by a rational function. The advantages of using rational functions are several. One advantage is that the rational approximant can be chosen such that its Taylor series matches the Taylor series of the optimal solution to some desired order. This means that the fit with the optimal solution will be good for small $|x|$. At the same time the rational approximant can be required to have a predefined rate of growth for large $|x|$. Another advantage seen from examples is that sometimes the Taylor series of the rational approximant matches the optimal solution to a higher order than used in the computations. However, the major drawback is that it normally comes with a higher computational complexity.

The computational method derived in this section is similar to the method described in Vannelli and Vidyasagar (1985) for estimating the region of attraction for nonlinear systems. Other computational methods to find approximate solutions can be found in for example Beard et al. (1998).

The structure of the chapters is as follows. First a brief problem formulation is presented in Section 5.1. Then the main method is derived in Section 5.2. The method is based on nonconvex optimization. Therefore, another method is presented in Section 5.3 which is used to find initial guesses to the optimization. Finally, some examples and conclusions are presented in Section 5.4 and 5.5, respectively.

5.1 Problem Formulation

Consider an optimal control problem

$$\begin{aligned} V(x_0) = \inf_{u(\cdot)} & \int_0^\infty L(x,u)\,dt \\ \text{s.t.} \quad & \dot{x} = F(x,u) \\ & x(0) = x_0 \in \Omega_x \end{aligned} \tag{5.1}$$

where $x \in \mathbb{R}^d$, $u \in \mathbb{R}^p$ and Ω_x is a neighborhood of the origin. To simplify the notation, the model is assumed to be control-affine

$$\dot{x} = f(x) + g(x)u \tag{5.2}$$

and the cost function is assumed to have the structure

$$L(x,u) = l(x) + \frac{1}{2}u^T R u \tag{5.3}$$

The derived method will rely on the series expansions of f, g and l, and therefore these functions are assumed to satisfy Assumption A8. That is, to be real analytic in a neighborhood of the origin. To obtain a well-defined solution to the ARE, the following assumption is also introduced.

Assumption A11. The linearization of (5.2), *i.e.*,

$$\dot{x} = Ax + Bu$$

is stabilizable. Furthermore, the matrix Q in

$$l(x) = x^T Q x + l_h(x)$$

and the matrix R in (5.3) are positive semi-definite and positive definite, respectively.

In Section 4.4, the optimal control problem (5.1) was solved under rather natural assumptions to obtain $V(x)$ and $u_*(x)$ in a neighborhood of the origin expressed as power series. The optimal solution most often requires an infinite number of terms to be described and the solution used in practice is therefore truncated. In simulations, various drawbacks with the truncated power series solution have been noticed. First, it often tends to grow too fast compared with the optimal. Second, it is rather common that the approximation turns negative outside a quite small region which the optimal is not, since the cost function is always positive. Third, the region in which the truncated feedback law stabilize the system may be small.

The objective in this chapter is therefore to find approximate solutions that in many cases have better properties over a larger region. For that reason, another parametrization of the optimal return function is studied, namely rational functions

$$V_r(x) = \frac{R_V(x)}{1 + Q_V(x)} \tag{5.4}$$

where

$$R_V(x) = R_V^{[2]}(x) + R_{V,h}(x) = \frac{1}{2}x^T P x + R_{V,h}(x) \tag{5.5a}$$

$$Q_V(x) = Q_V^{[1]}(x) + Q_{V,h}(x) = Tx + Q_{V,h}(x) \tag{5.5b}$$

and where $R_{V,h}$ and $Q_{V,h}$ are polynomials beginning with orders three and two, respectively.

The advantage of the rational functions is that while being able to match Taylor series of the optimal solution up to some desired order, it is possible to specify the growth rate for large $|x|$ by choosing the difference between the order of the numerator and the denominator. In this chapter, the difference in the order between $R_V(x)$ and $Q_V(x)$ will always be chosen as two. The motivation for this choice is that it often gives rather good approximations. However, note that by choosing the coefficients in the polynomials, the difference in order may change which means that different growth rates can be obtained.

5.2 Rational Approximation Based on Optimization

In this section, a method is derived which relies on that the HJB is rewritten as a set of equations, whose solution is parametrized in the denominator coefficients. The advantage of this approach is that the coefficients to arbitrary high order terms of the HJB can be reduced and the rational approximant will still have the same power series up to some desired order.

5.2.1 Derivation of the Equations

The optimal solution to (5.1) is given by the HJB. For this class of optimal control problems, it was shown in Section 2.7.3, that the HJB will take the form

$$0 = H\big(x, V(x)\big) = l(x) + V_x(x)f(x) - \frac{1}{2}V_x(x)g(x)R^{-1}g(x)^T V_x(x)^T \tag{5.6a}$$

$$u(x) = -R^{-1}g(x)^T V_x(x)^T \tag{5.6b}$$

where the expression for the optimal feedback law is explicit. The objective is now to find the $V_r(x)$ that satisfies (5.6a) up to some given order. For this end, the derivative of $V_r(x)$ with respect to x is needed.

$$V_{r;x}(x) = \frac{V_{r;x,n}}{V_{r;x,d}} = \frac{(1+Q_V)R_{V;x} - R_V Q_{V;x}}{(1+Q_V)^2} \tag{5.7}$$

If (5.7) is substituted into (5.6a), the following equation is obtained

$$0 = \frac{1}{V_{r;x,d}^2}\Big(\big(\frac{1}{2}x^T Q x + l_h\big)V_{r;x,d}^2 + V_{r;x,n}(Ax + f_h)V_{r;x,d} - \frac{1}{2}V_{r;x,n}gR^{-1}g^T V_{r;x,n}^T\Big) \tag{5.8}$$

where the numerator will be denoted $\tilde{H}\big(x, V_r(x)\big)$. This equation should be satisfied for all x in a neighborhood of the origin, which is equivalent to $\tilde{H}\big(x, V_r(x)\big) = 0$ in

a neighborhood up to the given order. Since different powers of x are independent, all coefficients in $\tilde{H}$ must equal zero.

A more thorough examination of $\tilde{H}$ shows that the coefficients corresponding to the second order terms form the standard ARE

$$0 = A^T P + PA - PBR^{-1}B^T P + Q$$

while the terms of a general order $m \geq 3$ will have the structure

$$R_{V;x}^{[m]} A_c x + M_1 Q_{m-2} - R_V^{[2]} Q_{V;x}^{[m-2]} A_c x = \xi\big(R_V^{[m-1]}, \ldots, R_V^{[m-4]}, Q_V^{[m-3]}, \ldots, Q_V^{[m-5]}\big) \quad (5.9)$$

where

$$A_c = A - BR^{-1}B^T P$$
$$M_1(x) = x^T\big(P(3A - BR^{-1}B^T) + 2Q\big)x$$

and ξ is a function determined by the functions f, g, l, and R.

To study the solvability of (5.9), Lemma 4.9 can be used. Based on this lemma, the following result is easily shown.

Lemma 5.1
Assume that A_c is a Hurwitz matrix. For given values of $R_V^{[2]}$,..,$R_V^{[m-1]}$ and $Q_V^{[1]}$,..,$Q_V^{[m-3]}$, equation (5.9) *is a linear system of equations for the coefficients in $R_V^{[m]}$ and $Q_V^{[m-2]}$. The null space of the associated linear map has a dimension equal to*

$$\binom{n+m-3}{n-1} \quad (5.10)$$

In particular $R_V^{[m]}$ is uniquely determined after an arbitrary choice of Q_{m-2}.

Proof: The size of the null space corresponds to the number of coefficients in Q_V of order $m-2$. The solvability follows from Lemma 4.9. □

To understand the approximating properties of V_r, the following result can be useful.

Lemma 5.2
Assume that Assumption A8 and A11 are satisfied. Let W be an analytic function such that $W(0) = 0$, $W_x(0) = 0$ and suppose that $H\big(x, W(x)\big)$ has a series expansion beginning with terms of order $m+1$. Then W and V have identical series expansions up to and including terms of order m.

Proof: The optimal return function V has to satisfy (5.9) with $Q_V = 0$, $R_V = V$. Under Assumptions A8 and A11, it follows that $V^{[2]}, \ldots, V^{[m]}$ are uniquely determined by the requirement that terms of order up to and including m in H are zero. Since the solution is uniquely determined, W must have the same Taylor series up to the given order. □

From the lemma above, the following useful lemma can be proved.

Lemma 5.3
Let

$$R_V^{[m]}(x),\ m = 2, \ldots, m_o$$
$$Q_V^{[m]}(x),\ m = 1, \ldots, m_o - 2$$

satisfy (5.9)*. Then V_r and V have the same series expansions for terms of orders up to and including m_o.*

Proof: The expression for $H\big(x, V_r(x)\big)$ will have the structure

$$H\big(x, V_r(x)\big) = \frac{\tilde{H}\big(x, V_r(x)\big)}{\big(1 + Q_V(x)\big)^4}$$

as was seen in (5.8). By construction, the terms in $\tilde{H}$ of orders less than or equal to m_o are zero. Since the expansion of the denominator begins with 1, this is true for H as well. The lemma is then a consequence of Lemma 5.2. □

5.2.2 Choice of Denominator

It is known from Chapter 4 that the HJB (5.6) can be solved by a polynomial. The extra degrees of freedom obtained by introducing a denominator in $V_r(x)$ gave a null space. In Lemma 5.1, it was shown that for a given Q_V, the terms in R_V will be determined. It means that the denominator can be chosen arbitrarily. Let m be the order of the nominator. Then the number of free parameters, or with other words, the number of coefficients in $Q_V^{m-2]}$, becomes $\binom{m+n-2}{n} - 1$ (can be shown using mathematical induction).

The free parameters can be used for different purposes, such as reducing extra terms in (5.6a), or to obtain a $V_r(x)$ that does not tend to infinity too fast etc. In the first case, it is beneficial, at least from a reduction point of view, to have a large number of free parameters. However, the obtained minimization problem may become tough and it may therefore be advantageous to fix some parameters. In some cases, it may even be possible to obtain a reasonably good approximation even with all parameters chosen as constants. Below, a few different choices of how to choose the free parameters are discussed.

All Parameters Free

The most general choice of denominator is of course to let all coefficients in Q_V be free. In this case, all of them can be used to reduce higher order coefficients in $\tilde{H}$, but the obtained optimization problem grows rapidly with the desired order and the number of states.

Denominator with Fixed Highest Order Term Coefficients

To reduce the computational complexity, some of the parameters can instead be chosen as constant values. One such choice is to let the highest coefficients be for example $1/(m-2)!$ and let the other coefficients be free. The main motivation for this choice is that the denominator of $V_r(x)$ becomes positive for both large and small $|x|$, since for

large $|x|$ the highest order term is dominating while for small $|x|$ the constant term is dominant.

Denominator with All Coefficients Fixed

The choice which gives the easiest problem to solve is to let all coefficients in Q_V be fixed. The result is a well-determined system of equations to solve, similar to the case in Chapter 4, and no optimization is required. In principle, it also means that the obtained problem will be as simple to compute as the ordinary power series method. Despite the simplicity, this choice can sometimes give approximations that are better than the truncated power series as will be seen in Section 5.4.

One choice that may be interesting to test is for example $(x-\alpha)(x+\beta)(1+x^{m-4})$, if the cost function has limits at $x=-\alpha$ and $x=\beta$.

5.2.3 Minimization of Higher Order Terms

If not all coefficients in the denominator are chosen as constants, a minimization problem can be formulated that reduces the coefficients corresponding to terms in the HJB of higher orders than m.

Denote the higher order terms, *i.e.*, terms in $\tilde{H}\big(x, V(x)\big)$ of degree $m+1$ or higher, as $E_m(x)$. That is, if

$$V_{rm}(x) = \frac{R_V^{[2]}(x) + \ldots + R_V^{[m]}(x)}{1 + \ldots + Q_V^{[m-2]}(x)} \tag{5.11}$$

where $R_V^{[i]}$ and $Q_V^{[i-2]}$, $i = 3, \ldots, m$ satisfy (5.6a), are substituted into (5.8) the result is

$$0 = \tilde{H}\big(x, V_{rm}(x)\big) = \underbrace{\text{terms of degree} \geq m+1}_{E_m(x)}$$

The vector with the coefficients of the polynomial $E_m(x)$ will be denoted e_m.

The number of free parameters should be compared with the number of coefficients in $E_m(x)$. The number of terms in $E_m(x)$ can be very large and therefore, $E_m(x)$ is truncated at some additional order m_h. That is, $m+m_h$ is the maximal order of the terms in the HJB that is suppressed. The parameter excess C_{pe} will then be given by

$$C_{\text{pe}} = \binom{m+n}{n} - \binom{m+m_h+n}{n} \tag{5.12}$$

If C_{pe} is larger than zero, *i.e.*, if the number of free parameters is larger than the number of coefficients, and if the parameters enter the problem in an appropriate way, it is sometimes possible to zero some of the higher order coefficients exactly using an equation solver. For scalar problems of orders not too high, this approach seems to work rather well. However, for larger scalar problems and non-scalar problems, it is quite common that the equation solver requires a huge amount of time or that no solution is returned at all.

Therefore, another approach is used where the higher order coefficients in the HJB are minimized using a numerical optimization routine. The advantage of this approach is that if a set of coefficients exists such that the higher order terms are zeroed, the optimization

often finds them. On the other hand, if no such solution exists, for example in a case when the parameter excess C_{pe} is negative, *i.e.*, the number of parameters are fewer than the number of terms in $E_m(x)$, the optimization will still try to give a solution with $|e_m|^2$ as small as possible.

The recursive equation (5.9) is equivalent to the under-determined linear system of equations

$$A_m Y_m = b_m \tag{5.13}$$

where

$$\begin{aligned}
Y_m &= (y_{R,3}, y_{Q,1}, y_{R,4}, y_{Q,2}, \ldots, y_{R,m}, y_{Q,m-2})^T \\
A_m &= \begin{bmatrix} A_{m,1}(y_{R,2}) & 0 & \ldots & 0 \\ 0 & A_{m,2}(y_{R,2}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_{m,m}(y_{R,2}) \end{bmatrix}, \\
b_m &= \begin{bmatrix} b_{m,1}(y_{R,2}) \\ b_{m,2}(y_{R,2}, y_{R,3}, y_{Q,1}) \\ \vdots \\ b_{m,m}(y_{R,2}, \ldots, y_{R,m-1}, y_{Q,m-3}) \end{bmatrix}
\end{aligned}$$

and $y_{R,i}$ and $y_{Q,i}$ are the unknown coefficients in $R_V^{[i]}(x)$ and $Q_V^{[i]}(x)$, respectively. The vector $y_{R,2}$ contains the coefficients in $R_V^{[2]}$ which is P, and is therefore computed using the standard ARE. The matrices $A_{m,i}(\cdot)$ and $b_{m,i}(\cdot)$ are functions determined by the left-hand and right-hand side of (5.8), respectively.

The optimization problem is then formulated as

$$\begin{aligned}
&\min_{Y_m} |e_m(Y_m)|^2 \\
&\quad \text{s.t. } A_m Y_m = b_m
\end{aligned}$$

The optimization problem can be solved either as constrained or unconstrained. To motivate this fact, let $i = 3$. Then, if the coefficients in the denominator $y_{Q,1}$ are considered as parameters, it was shown in (5.9) that only linear equations need to be solved in order to obtain the coefficients in the nominator $y_{R,3}$ expressed in terms of $y_{Q,1}$. Furthermore, in Lemma 5.1, it was shown that the solution is unique. Repeating this procedure for the higher indices means that (5.13) is solved recursively. The coefficients $y_{R,i}$, $i = 3, \ldots, m$, will only depend on $y_{Q,j}$, $j \leq i-2$. The result is the unconstrained optimization problem

$$\min_{y_Q} \quad |e_m(y_Q)|^2 \tag{5.14}$$

where y_Q is the concatenation of $y_{Q,i}$, $i = 1, \ldots, m-2$. In this thesis, mostly the unconstrained approach has been used. However, there might be structural benefits with keeping the constraints, since simpler expressions in e_m will be obtained in that case.

The optimization problem (5.14) is polynomial, and normally it becomes nonconvex. For small m, d, and m_h, it is possible to solve the problem globally using for example

sum-of-squares relaxations and YALMIP, see Löfberg (2008). In this case, it is quite important in order to reduce the computational complexity, to rewrite the problem as

$$\min_{y_Q, Y_{\text{sos}}} |Y_{\text{sos}}|^2$$
$$\text{s.t. } e_m(y_Q) = Y_{\text{sos}}$$

where Y_{sos} are extra variables, one for each term in e_m, used to reduce the maximal order of the functions involved.

However, for medium-sized m, n and m_h, the expressions in e_m becomes large and rather involved. In this case, the global methods seem to be too computationally demanding, and it is necessary to search for a local minimum instead. Then, the initial conditions become important. The good news is that numerical experience shows that often a local optimum can be found such that the corresponding approximant match the optimal solution well (depending on the choices made in the next section). A fact that also simplifies the problem is that the unconstrained problem (5.14) is a nonlinear least-squares problem. This is a structure which in many solvers, such as the solver in Maple, can be utilized to reduce the computation time.

5.2.4 Design Choices in the Minimization

The optimization problem (5.14) involves a number of different design choices. One is the denominator choice of V_r, discussed in Section 5.2.2. Below some other of the choices are mentioned.

The Order of f, g and l

The first design choice is the order of the functions that describe the model and the cost function. It is possible to have arbitrary orders of f, g and l as long as the orders are larger than or equal to $m-1$, $m-1$ and m, respectively. Otherwise, the power series solution up to the desired order m will not be correct as was shown in Chapter 4. The standard choice in the simulations presented in this thesis, see Section 5.4, have been to truncate at order $m + m_h$.

Order of E_m

Another design parameter is the truncation degree of $E_m(x)$. From (5.8), it follows that the maximal degree that may show up in $E_m(x)$ is given by

$$\max\big(m_l + 4(m-2), m_f + 4m - 7, 2m_g + 4m - 6\big)$$

where m_f, m_g and m_l are the orders of f, g and l, respectively. This number is often quite large and therefore it is necessary to truncate $E_m(x)$ to obtain a solvable optimization problem. The truncation also implies that higher order terms in $E_m(x)$ are considered irrelevant.

The most basic choice of m_h is to choose the value for which the parameter excess switches from positive to negative. Then, the optimal e_m will be zero, *i.e.*, all higher terms are zeroed. However, in many cases it is possible to obtain better approximations

by increasing m_h even more. In this case, e_m will in general not equal zero since the number of free parameters are less than the number of coefficients, but still this choice is often good. Actually, in some cases one can gain a lot by increasing m_h, without changing m. It means that information from higher order terms are included in the lower order approximant. For example, this is the case in Section 5.4.3.

However, a small remark is that sometimes it seems like main difficulty is not the number of terms itself but the complexity of the coefficients due to the reduction of the problem to an unconstrained problem.

Initial Values

Since the optimization problem (5.14) mostly is nonconvex, the choice of initial values is important. In this thesis, the default choice is to generate a number of vectors with uniformly distributed random numbers in the interval $[-3, 3]$. The optimization problem (5.14) is then solved for each of them as initial guess, and the best solution is chosen as the optimum. The number of different vectors can be chosen, but the standard choice is two.

Another choice is to use the direct approximation, see Section 5.3, as the initial guess. In many cases, the direct approximation has proved to be better to use than the random vectors.

General Comments

Normally, many of the computed coefficients will be close to machine precision or at least small compared to the other coefficients. These terms increase the complexity of the approximant which is undesired. Therefore, coefficients smaller than some given threshold, are often truncated. Numerical computations show that even if a rather harsh truncation is performed, which will yield an slightly erroneous Taylor series compared to the optimal, the approximation will overall often be fairly good. In practice, this might be important.

Another comment concerns the issue that sometimes the denominator may be singular within the region where the approximate solution is supposed to be used. In the sum-of-squares framework, it is possible to include the constraint $Q_V(x) > 0$. In the local optimization framework, there are no simple conditions to add in order to obtain a positivity certificate. A simple solution is to include the constraints $Q_V(x_i) > 0$, where x_i are points on a grid, but of course, this does not give sufficient conditions. In the presented examples, no such constraints have been included since it was not necessary.

5.2.5 Stability

One of the major objectives for a controller is to stabilize the system. The controller obtained from the method in this chapter can be shown to yield stability at least locally in a neighborhood as shown in the following theorem.

Theorem 5.1
Consider a nonlinear system in the form (5.2) that satisfy Assumptions A8 and A11. Let $V_{rm}(x)$ in (5.11) solve (5.8) up to order m and let the corresponding control law be given

by (5.6b). Then this feedback law will stabilize the system locally in a neighborhood of the origin, and the cost function $V_{rm}(x)$ will be a Lyapunov function for the closed-loop system

$$\dot{x} = f(x) - g(x)R^{-1}g(x)^T V_{rm;x}(x)^T$$

Proof: First note that the cost function $V_{rm}(x)$ can be expanded around zero yielding

$$V_{rm}(x) = \frac{1}{2}x^T Px + V_{rm,h}(x)$$

and the time derivative of $V_{rm}(x)$ using the feedback law (5.6b) becomes

$$\dot{V}_{rm} = V_{rm;x}(f - gR^{-1}g^T V_{rm;x}^T) = -l - \frac{1}{2}V_{rm;x}gR^{-1}g^T V_{rm;x}^T + \frac{E_m}{V_{rm;x,d}^2}$$

where $V_{rm;x,d}$ is the denominator of $V_{rm;x}$. The series expansion of the first two terms in the expression above is given by

$$l + \frac{1}{2}V_{rm;x}gR^{-1}g^T V_{rm;x}^T = \frac{1}{2}x^T(Q + PBR^{-1}B^T P)x + \mathcal{O}(x)^3$$

and since $E_m(x)$ contains terms beginning with order $m+1$, it follows that for x in a neighborhood of the origin, the optimal return function will satisfy $V_{rm}(x) > 0$ and $\dot{V}_{rm}(x) < 0$. That is, the function $V_{rm}(x)$ will be a Lyapunov function for the closed-loop system and $u = -R^{-1}g(x)^T V_{rm;x}(x)^T$ is a stabilizing control law. □

Hence, the controller stabilizes the system locally around the origin, similar as for the power series approximation, see Lukes (1969). However, as for the power series method, no estimate of the region of attraction is obtained by the method. If such an estimate is desired, one has to use some other method, see for example Vannelli and Vidyasagar (1985).

5.2.6 Extension to General State-Space Models

In the earlier sections, only control-affine systems and cost functions with a quadratic dependence of the controls were considered. However, the results are also possible to extend to a more general class of optimal control problems where the system is

$$\dot{x} = F(x,u) = Ax + Bu + F_h(x,u) \tag{5.15}$$

and the cost function is

$$L(x,u) = \frac{1}{2}x^T Qx + x^T Su + \frac{1}{2}u^T Ru + L_h(x,u) \tag{5.16}$$

For notational reasons only the single-input case is considered, but the method should be possible to extend to the multi-input case as well. As in the former section, F and L are assumed to satisfy Assumption A8.

In this case, it is not possible to solve the HJB for u explicitly to get a single nonlinear PDE as in (5.6a). Instead, it is necessary to consider the system of equations

$$0 = L(x,u) + V_x(x)F(x,u) \tag{5.17a}$$
$$0 = L_u(x,u) + V_x(x)F_u(x,u) \tag{5.17b}$$

which have to be satisfied for all x in a neighborhood of the origin. Let V_r and u_r be given by

$$V_r(x) = \frac{R_V(x)}{1+Q_V(x)},\ u_r(x) = \frac{R_u(x)}{1+Q_u(x)} \tag{5.18a}$$

where

$$R_V(x) = \frac{1}{2}x^T P x + R_{V,h}(x) \tag{5.18b}$$
$$Q_V(x) = Q_V^{[1]}(x) + Q_{V,h}(x) \tag{5.18c}$$

and

$$R_u(x) = Dx + R_{u,h}(x) \tag{5.18d}$$
$$Q_u(x) = Q_u^{[1]}(x) + Q_{u,h}(x) \tag{5.18e}$$

The terms $R_{V,h}(x)$ and $R_{u,h}$ denote terms of order at least three and two, respectively, while $Q_{V,h}(x)$ and $Q_{u,h}$ are both of order at least two.

If the parameterizations in (5.18) are substituted into (5.17) and the equations are multiplied with a polynomial factor as

$$0 = \big(1+Q_u(x)\big)^{m_m}\big(1+Q_V(x)\big)^2\big(L(x,u)+V_x(x)F(x,u)\big) \tag{5.19a}$$
$$0 = \big(1+Q_u(x)\big)^{m_m-1}\big(1+Q_V(x)\big)^2\big(L_u(x,u)+V_x(x)F_u(x,u)\big) \tag{5.19b}$$

the result is two polynomial equations. Since these equations are supposed to be satisfied for all x in a neighborhood of the origin, coefficients corresponding to different orders all need to be zero. Therefore, if the second order terms are extracted from (5.19a), and the first order terms from (5.19b), we obtain the equations

$$0 = Q + (S+PB)D + D^T(S^T + B^T P) + PA + A^T P + D^T R D$$
$$0 = D + R^{-1}(S^T + B^T P)$$

which is the standard ARE. The higher order terms of V_r and u_r are solved from higher order terms in (5.19), and after rather cumbersome calculations, the expressions corresponding to the terms of order m and $m-1$, respectively, become

$$R_{V;x}^{[m]}A_c x + M_1 Q_u^{[m-2]} + M_2 Q_V^{[m-2]} - R_V^{[2]}Q_{V;x}^{[m-2]}A_c x = \xi_1\big(R_V^{m-1]}, R_u^{m-2]}, Q_V^{m-3]}, Q_u^{m-3]}\big)$$

and

$$RR_u^{[m-1]} + R_{V;x}^{[m]}B + M_3Q_u^{[m-2]} + M_4Q_V^{[m-2]} - R_V^{[2]}Q_{V;x}^{[m-2]}B = \xi_2\big(R_V^{m-1]}, R_u^{m-2]}, Q_V^{m-3]}, Q_u^{m-3]}\big)$$

where

$$\begin{aligned}
M_1(x) &= x^T\big(\tfrac{m_m}{2}Q + (m_m-1)SD + \tfrac{m_m-2}{2}D^TRD + m_mPA + (m_m-1)PBD\big)x\\
M_2(x) &= x^T\big(Q + 2SD + D^TRD + PA + PBD\big)x\\
M_3(x) &= x^T\big((m-1)S + (m-2)D^TR + (m-1)PB\big)\\
M_4(x) &= x^T\big(2S + 2D^TR + PB\big)
\end{aligned}$$

for $m = 3, \ldots, m_{\max}$ and where $A_c = A - BR^{-1}B^TP$ and ξ_i, $i = 1, 2$ are functions determined by the model and cost function.

This means that Lemma 5.1 can be generalized to general state-space models as follows.

Lemma 5.4
Assume that A_c is a Hurwitz matrix and that R is positive definite. For given expressions of R_V, Q_V, R_u and Q_u up to orders $m-1$, $m-3$, $m-2$ and $m-3$, respectively, equation (5.9) *is a linear system of equations for the coefficients in $R_V^{[m]}$, $Q_V^{[m-2]}$, $R_u^{[m-1]}$ and $Q_u^{[m-2]}$. The null space of the associated linear map has a dimension equal to*

$$2\binom{n+m-3}{n-1} \tag{5.21}$$

In particular $R_V^{[m]}$ and $R_u^{[m]}$ are uniquely determined after an arbitrary choice of $Q_V^{[m-2]}$ and $Q_u^{[m-2]}$.

Proof: The size of the null space corresponds to the number of coefficients in $Q_V^{[m-2]}$ and $Q_u^{[m-2]}$. The solvability follows from Lemma 4.9. □

The conclusion is therefore that also the general case can be solved in the same way. However, the number of free parameters is larger than before, and thereby a tougher optimization problem need to be solved.

5.3 Direct Approximation

Another method that can be used to compute a rational approximation is Padé approximation, see Chisholm (1973); Cuyt and Wuytack (1987); Guillaume et al. (1998); Guillaume and Huard (2000). The idea is to find a rational function that matches the power series solution up to some desired order. Let the series expansion of the optimal return function be computed using the method in Chapter 4, and form the following equation

$$V(x)\big(1 + Q_V(x)\big) = R_V(x) \tag{5.22}$$

where $R_V(x)$ and $Q_V(x)$ are given by the expressions in (5.5). In standard multivariate Padé approximation, the number of equations obtained from (5.22) should equal the number of parameters in R_V and Q_V.

In this thesis, a similar approach has been used, mainly to generate initial guesses to the minimization (5.14). Let R_V and Q_V be chosen as in (5.5), with orders m and $m-2$, respectively. This choice is made to ensure that the solution has the same structure as used in the optimization. The order of V is then a design parameter chosen as $m+m_h$ and the equations in (5.22) are truncated at the same order.

$$\Big\{V(x)\big(1+Q_V(x)\big)-R_V(x)\Big\}^{[k]}=0,\ k=2,\dots,m+m_h \tag{5.23}$$

The obtained set of equations will be linear in the parameters in R_V and Q_V, and can be written as

$$\sum_{i=1}^{k-2} V^{[k-i]}(x)Q_V^{[i]}(x)+V^{[k]}(x)-R_V^{[k]}(x)=0,\ k=2,\dots,m$$

$$\sum_{i=1}^{m-2} V^{[k-i]}(x)Q_V^{[i]}(x)+V^{[k]}(x)=0,\ k=m+1,\dots,m+m_h$$

The set of equations above is solved in a least-squared sense, because it is not certain that a solution exists. First, the set of equations can be under-, well- or over-determined, depending on the choice of m_h. Second, it is assumed that $Q_V^{[0]}(x)=1$, because it gives a rational approximation that is well-behaved locally around the origin. This assumption leads to that even for the under-determined case, there may be no solution to the equations, unless $m_h=0$.

However, if the residuals obtained from the least-squares solution become small, the obtained rational approximation seems to be a very good initial guess, and the computational complexity is fairly small. Moreover, if the residuals become zero, the rational approximation will have the same Taylor series as V up to order $m+m_h$.

5.4 Examples

In this section three examples are presented. The first example is a scalar problem which comes from Navasca (1996). In this example the cost function includes a barrier function on the state. The second and third examples are multivariable problems. The second one is a physical system, namely a nonlinear phase-locked loop, while the third example is a purely mathematical one. The special feature of the third example is, similarly to the scalar problem, that the optimal return function tends to infinity for finite x, *i.e.*, the cost function is a kind of a barrier function.

5.4.1 A Scalar Problem

The considered system is given by

$$\dot{x} = (1 + x)u$$

which is a stabilizable system around the origin. The cost function is chosen as

$$l(x) = \ln(1 + x)^2$$

The corresponding optimal control problem can be solved explicitly and the optimal cost function becomes

$$V(x) = \frac{\sqrt{2}}{2} \ln(1 + x)^2$$

while the optimal feedback law is given by

$$u(x) = -\sqrt{2} \ln(1 + x)$$

In the scalar case, it is most often possible to solve for extra coefficients in the HJB exactly. This fact has been exploited in this example, where a fifth order rational approximation has been computed. By using the three extra terms in the denominator, three additional terms in the HJB have been zeroed. The obtained solution will therefore have the same Taylor series as the optimal solution up to order eight. Actually, the series expansions are the same with three decimals accuracy up to the 14:th order. The functions f, g and l are truncated after the eighth degree. The result can be seen in Figure 5.1. The same figure also shows the rational approximation with a denominator where the highest order term is fixed. As can be seen the difference is rather small.

In Figure 5.2, an ordinary truncated power series solution of order five and a rational approximation with fixed denominator have also been included in the comparison. As can be seen, these solutions are substantially worse than the earlier rational approximations. However, the rational approximation with fixed coefficients is better than the truncated power series solution.

Concerning stability it can also be shown that the rational approximation is substantially better than the power series solution. The region in which the rational approximation with free denominator is stabilizing the system is $x_0 \in [-0.99, 21]$, while the truncated power series solution only stabilizes the system in the region $x_0 \in [-0.99, 0.8]$.

In the last two figures another advantage of the rational approximations is illustrated. Here a higher order approximation of order 8 plus the 6 free parameters in the denominator. For the rational approximation a higher order often give a better approximation in a larger region than a lower one, which is the case for this example as can be seen in Figure 5.3. However, for truncated power series a higher order often only yields a better fit with the optimal solution locally around the origin and outside this region an even worse fit is obtained as can be seen in Figure 5.4.

It can be noticed that no plots of the corresponding feedback laws are shown. The reason is the explicit relation between the optimal return and the feedback law given by (5.6b). Due to this relation more or less the same improvement is obtained for the optimal return function and the feedback law, at least for the examples studied in this thesis,.

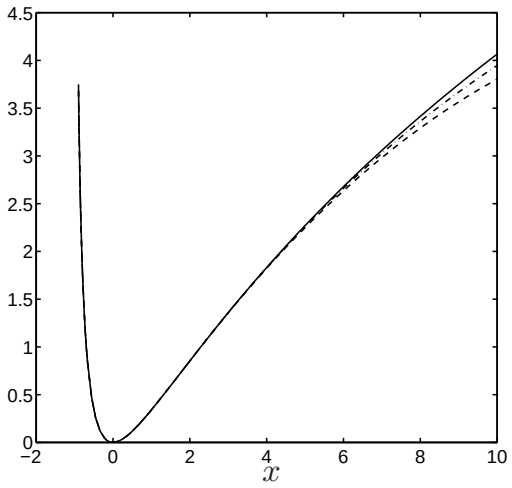


Figure 5.1: *A comparison between V (solid) and two rational approximations. The dash-dotted line corresponds to the approximation with free denominator and the dashed line has fixed highest order term in the denominator.*

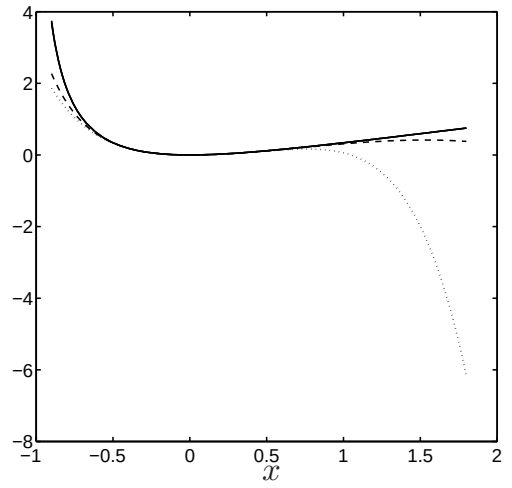


Figure 5.2: *A comparison between the optimal cost (solid), the three different rational approximations and the truncated power series. The rational approximation with free denominator and with fixed highest degree term are indistinguishable from the optimal solution. The dashed line is the rational with fixed denominator and the dotted line is the power series.*

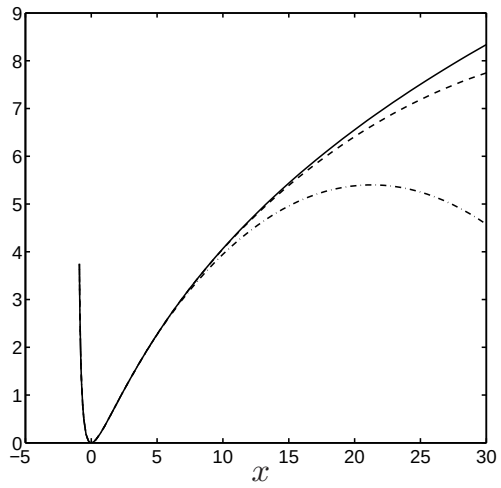


Figure 5.3: *A comparison between the optimal cost (solid) and two rational approximations of different order. The dashed line is a eighth order approximation and the dashed-dotted line is a fifth order one.*

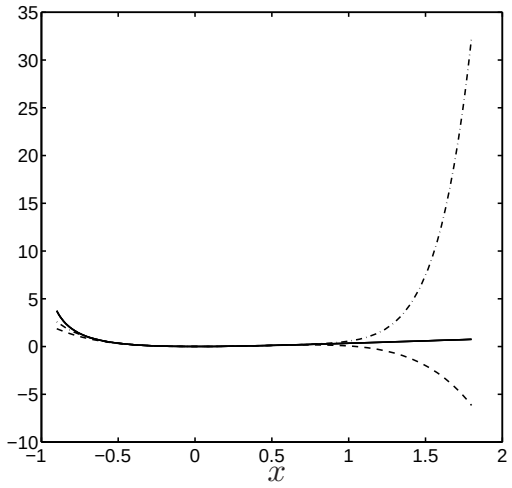


Figure 5.4: *A comparison between the optimal cost (solid) and four different approximations. Two of them are rational with free denominators but with different order (*4 *or* 8*). They are hidden behind the optimal solution. The other approximations are truncated power series, one of order* 5 *(dashed) and order* 8 *(dash-dotted).*

5.4.2 A Phase Lock Loop Circuit

Consider a model for a nonlinear phase lock loop circuit (PLL). The dynamics for the system can be written as

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

The cost function $l(x)$ is chosen as

$$l(x_1, x_2) = \frac{1}{2}x_1^2 + 2x_1x_2 + x_2^2 + x_1\sin(x_1)$$

which makes it possible to find an explicit solution as

$$V(x_1, x_2) = 2(1 - \cos(x_1)) + x_1x_2 + x_2^2$$

For this optimal control problem, the fourth order rational approximation is computed. As the comparison in Figure 5.5 shows, the rational approximation describes the optimal solution rather well. In this example, the terms of order six and below of the power series of f, g and l are included in $E_m(x)$ and the HJB is also truncated at order six, *i.e.*, two orders higher than m. The corresponding value of e_m^2 became $3 \cdot 10^{-13}$.

In the same figure, also a truncated power series solution of order four is presented. The improvement by using the rational approximation is quite large, which is even more clear in Figure 5.6 where the error of the rational approximation is compared with the error for the truncated power series solution.

In Figure 5.7, another comparison of errors is shown. The error that bends upwards and which has the smallest amplitude corresponds to a rational approximation of order 4 with the denominator chosen as

$$Q_V(x_1, x_2) = 1 + \tfrac{1}{6}x_1^2 + \tfrac{1}{6}x_1x_2 + \tfrac{1}{6}x_2^2$$

The other error corresponds to a truncated power series. As can be seen, the rational approximation is still better than the power series but worse than the rational approximation with free denominator (which could be seen in the Figure 5.6).

Figure 5.8 shows the error for two higher order approximations. The order of the rational approximation has been increased to six and the truncated power series approximation is of order eight. For the rational approximation, $m_h = 8$ has been used and the functions are also truncated at $m + 8$. It means that the order of the rational function is not that high, but information about the model and the cost function up to order 14 is included. In this plot, the intervals for the variables are increased, because in a smaller region both approximations are good.

Finally, Figure 5.9 shows the improvement of using the rational approximation of the higher order compared with the lower order approximation. As can be seen, the difference is between 4 and 5 times in this particular example.

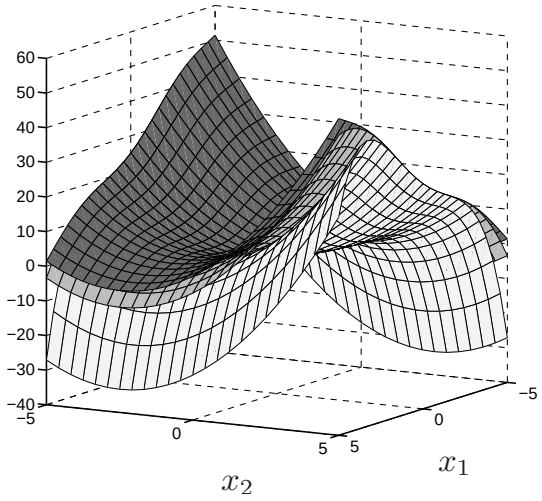


Figure 5.5: *A comparison between V (dark), the rational approximation (medium dark) and truncated power series (light) the applied to the PLL example.*

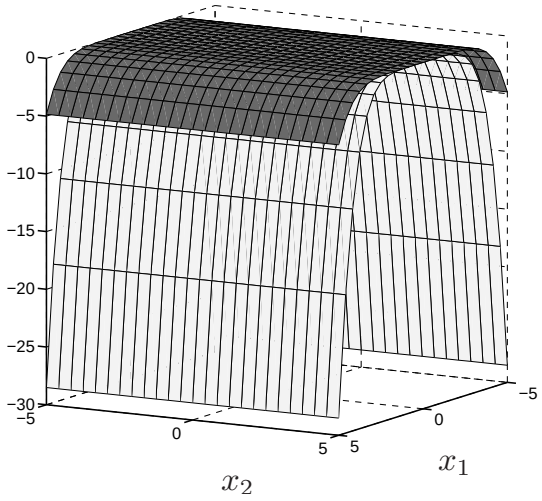


Figure 5.6: *A comparison between the errors of the rational approximation (dark) and of the truncated power series (light) for the PLL system.*

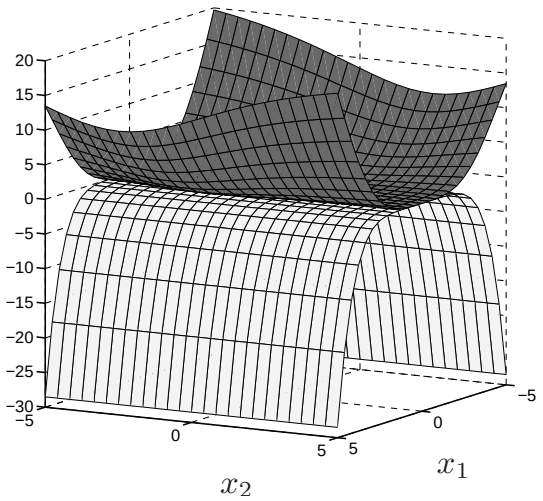


Figure 5.7: *A comparison between the errors of the rational approximation with fixed denominator (dark) and of the truncated power series (light) for the PLL system.*

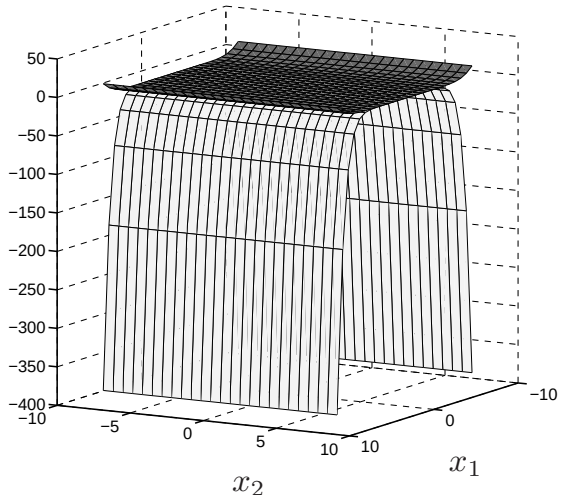


Figure 5.8: *A comparison between the errors of the rational approximation of order* 6 *(dark) and of the truncated power series of order* 8 *(light) for the PLL system.*

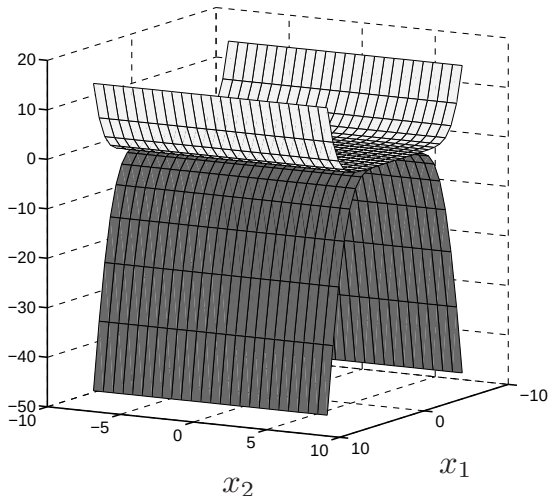


Figure 5.9: *A comparison between the errors of the rational approximation of two different orders. An fourth order (dark) and a sixth order (light).*

5.4.3 A Barrier Example

Consider the model

$$\dot{x} = \begin{pmatrix} -2x_1 + x_2 + x_3^2 \\ x_3 + x_1^2 \\ 0 \end{pmatrix} + \begin{pmatrix} x_2 \\ 1 + x_3^2 \\ 1 \end{pmatrix} u$$

and let the cost function be

$$L(x, u) = l(x) + u^2$$

where

$$\begin{aligned} l(x) = \Big(& (4x_3x_2x_1 + 2x_3x_2 + 2x_3^2x_2^2 + x_2^2 + x_3^2 + 4x_1x_2^2 + 4x_2^2x_1x_3^2 + 4x_2^2x_1^2 + x_2^2x_3^4 \\ & + 2x_2x_3^3 - 2(2x_2x_1 + x_3x_2 + 2x_3^2x_1 + x_1^2x_2 - 4x_1^2)\cos(2x_1^2 + x_2^2 + x_3^2)^2 \Big) \Big/ \\ & \cos(2x_1^2 + x_2^2 + x_3^2)^4 \end{aligned}$$

As can be seen $l(x)$ tends to infinity when $|2x_1^2 + x_2^2 + x_3^2| \to \frac{\pi}{2}$, because of the cos-term in the denominator. Hence, $l(x)$ is a barrier function, even though not as obvious as for $l(x)$ in Section 5.4.1. In this case, the system and the cost function are chosen such that the corresponding optimal control problem has an explicit optimal return function $V(x)$ given by

$$V(x) = \tan(2x_1^2 + x_2^2 + x_3^2) \tag{5.24}$$

which for $|2x_1^2 + x_2^2 + x_3^2|$ close to $\frac{\pi}{2}$, approaches infinity (as expected since the cost function does so).

The plots requires four dimensions to be visualized with all coordinates at once. Therefore, one state variable will be set to zero. First, the optimal return function with $x_1 = 0$ is shown in Figure 5.10. The following figure, *i.e.*, Figure 5.11, includes two plots with the errors for the two approximants. In Figure 5.11a, the error between the optimal solution and a truncated power series solution of order 10 is shown, while Figure 5.11b shows the error for a rational approximation of order 6. The rational approximation has been computed with $m_h = 4$ and with the model and the cost function truncated at order 10. In this case, the improvement by using the rational function is quite substantial.

Plots with the same configurations but with $x_3 = 0$ are shown in Figure 5.12 and 5.13, respectively. Also in this case, the improvement is noticeable.

5.5 Conclusions

In this chapter a new method to find approximate solutions to nonlinear optimal control problems has been derived. The result is a rational approximation. The method is to a large extent influenced by the power series method. To compute the approximation first an ARE is solved, then some linear equations and finally an optimization problem.

As for the power series method, the obtained rational approximation is shown to have the same power series as the optimal solution around the origin. This means that locally around the origin, the methods produce comparable approximations, but at least in three examples, the rational approximation is shown to give better performance in a larger region.

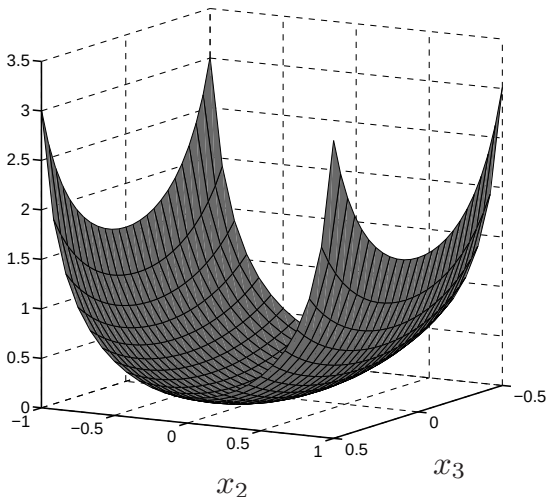


Figure 5.10: *The exact optimal return function when* $x_1 = 0$.

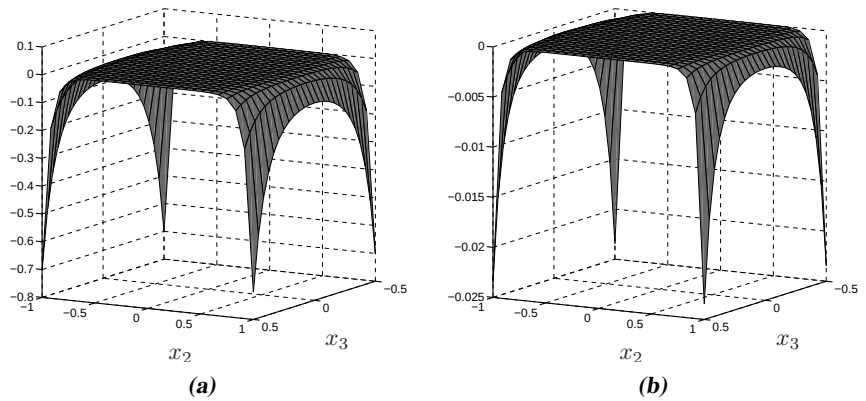


Figure 5.11: *The errors between the optimal solution and a) the truncated power series of order* 10*, and b) the rational approximation of order* 6.

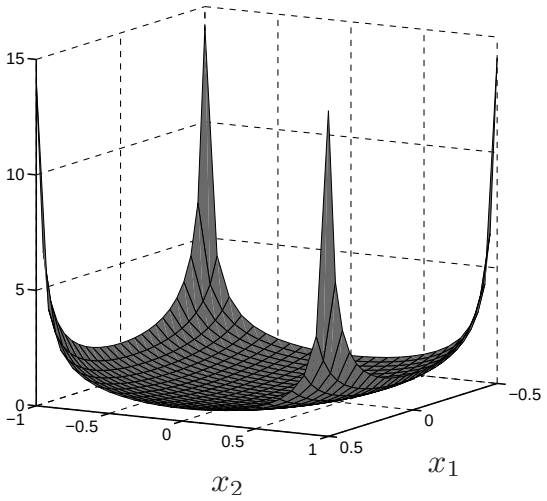
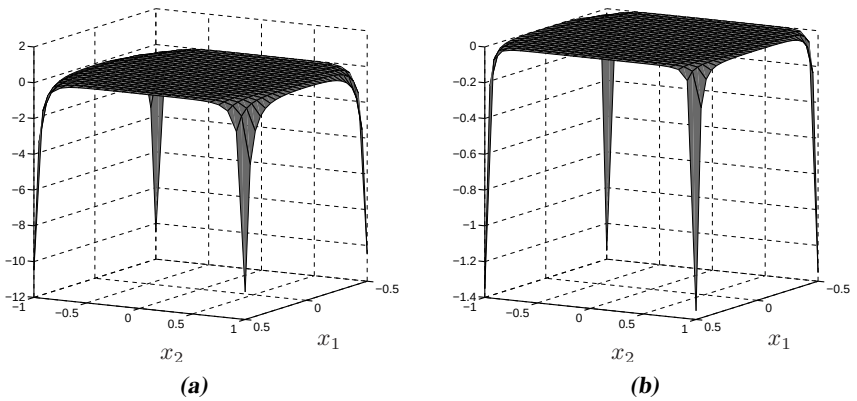


Figure 5.12: *The exact optimal return function when* $x_3 = 0$.



(a) ***(b)***

Figure 5.13: *The errors between the optimal solution and a) the truncated power series of order* 10*, and b) the rational approximation of order* 6.

6

Utilization of Structure and Control Affinity

A problem with the methods described in Chapter 4 and 5 is the computational complexity. The part that consumes most of the time is the solution for the coefficients in the approximant. Therefore, it would be interesting to reduce the number of equations that need to be solved. For control-affine systems $\dot{x} = f(x) + g(x)u$ with cost functions quadratic in u, it was shown in Section 2.7.3 that the equation being used to determine the optimal feedback law had an explicit solution. It means the feedback law need not be parametrized which reduces the size of the system of equations that is solved in the power series method. Moreover, the number of other computations required is reduced as well, which shortens the computation time even more.

For state-space models control affinity is recognized by just looking at the equations. For DAE models it might however be harder. Therefore, some conditions are derived under which the DAE model is equivalent to a control-affine system. It is also shown that the DAE model is required to have a certain structure in order to obtain affinity in the control input. Given the conditions, some theorems under which the optimal control problem can be simplified are presented.

6.1 Introduction

The standard way to solve the HJB for nonlinear DAE models using the power series expansions leads to the solution of the following four polynomial equations

$$0 = L_u - V_{x_1}\hat{F}_{1;\dot{x}_1}^{-1}\hat{F}_{1;u} - \left(L_{x_3} - V_{x_1}\hat{F}_{1;\dot{x}_1}^{-1}\hat{F}_{1;x_3}\right)\hat{F}_{2;x_3}^{-1}\hat{F}_{2;u} \tag{6.1a}$$

$$0 = L + \lambda V(x_1) + V_{x_1}\dot{x}_1 \tag{6.1b}$$

$$0 = \hat{F}_1 \tag{6.1c}$$

$$0 = \hat{F}_2 \tag{6.1d}$$

where V is evaluated at x_1, $\hat{F}_2$ and L at (x_1, x_3, u), and $\hat{F}_1$ at $(\dot{x}_1, x_1, x_3, u)$.

In Chapter 4, the procedure used to show that (6.1) has a solution is to first solve (6.1c) and (6.1d) using the power series method. The outcome is the power series of the underlying state-space model (4.39). Having this, the cost function $\hat{L}$ in (4.48) can be expressed as a power series as well. Finally, the optimal solution is obtained from the HJB for the state-space model.

$$0 = \hat{L}_u + V_{x_1}(x_1)\mathcal{L}_u \tag{6.2a}$$

$$0 = \hat{L} + \lambda V(x_1) + V_{x_1}(x_1)\mathcal{L} \tag{6.2b}$$

$$\tag{6.2c}$$

Hence, to find the terms in $V(x_1)$ up to order m, and the optimal control law $u_*(x_1)$ up to order $m-1$, the total number of equations to solve will be

$$\underbrace{\binom{d+m}{d} - (1+d)}_{V^{[m]}} + \underbrace{n\binom{d+p+m-1}{d+p} - n}_{\mathcal{L}^{[m-1]} \text{ and } \mathcal{R}^{[m-1]}} + \underbrace{p\binom{d+m-1}{d} - p}_{u_*^{[m-1]}} \tag{6.3}$$

where we have used that there are $\Sigma_{\text{m}}(n,m)$ terms of order m in an n variable polynomial, and the accumulated number of terms in the same polynomial up to order m is given by $\Sigma_{\text{m,accum}}(n,m)$.

$$\Sigma_{\text{m}}(n,m) = \binom{n+m-1}{n-1}, \quad \Sigma_{\text{m,accum}}(n,m) = \binom{n+m}{n} \tag{6.4}$$

The approach to first compute the Taylor series of the reduced system and then solve the problem in a state-space framework using (6.2), is mostly used because it makes it easier to refer to earlier results such as Lukes (1969). A better approach is to solve (6.1) simultaneously, *i.e.*, to solve for x_3 and $\dot{x}_1$ at the same time as for V and u_*. In that way, it is possible to exploit the property that x_3 and $\dot{x}_1$ only depend on x_1 and not on u, as in the derivation. The number of equations to solve will then be (6.3) but without the term p in the middle term. If either the desired order m or the number of inputs is large, the reduction may be quite substantial.

However, if the system and the cost function have certain structure, it is possible to further reduce the number of equations that need to be solved by power series expansion. The idea is to find the solution for u_* explicitly. One important case for which this is possible is described in Chapter 3. The restriction on the cost function can be relaxed to $l(x_1, x_3)$. Note that this structure also contains models for which the underlying state-space model is not control-affine.

Another important case of optimal control problems for which the computations can be simplified, is when the underlying state-space model is control-affine, *i.e.*, has the structure

$$\dot{x}_1 = \mathcal{L}(x_1, u) = \mathcal{L}_1(x_1) + \mathcal{L}_2(x_1)u \tag{6.5}$$

and the cost function is chosen such that

$$\hat{L}(x_1, u) = \mathcal{Q}(x_1) + \mathcal{S}(x_1)u + u^T\mathcal{R}_c u \tag{6.6}$$

According to Section 2.7.3, the optimal return function is then found by solving

$$0 = \mathcal{Q} + V_{x_1}\mathcal{L}_1 - \frac{1}{4}(V_{x_1}\mathcal{L}_2 + \mathcal{S})\mathcal{R}_c^{-1}(V_{x_1}\mathcal{L}_2 + \mathcal{S})^T$$

and the corresponding optimal control law is given by

$$u_* = -\frac{1}{2}\mathcal{R}_c^{-1}(V_{x_1}\mathcal{L}_2 + \mathcal{S})^T$$

Hence, only the equation for $V(x_1)$, *i.e.*, (6.2b), needs to be solved in order to obtain the optimal solution. The solution above is described in terms of several implicitly known functions for which the power series can computed by first solving (6.1c) and (6.1d). However, it is also interesting to study how the expression for u_* is defined in terms of the original functions, *i.e.*, $\hat{F}_1$, $\hat{F}_2$ and L.

Assume the model is described by (6.5). Since the model is described by (6.5), it follows that $\mathcal{L}_u$ will be $\mathcal{L}_2$ on the solution manifold $\mathbb{L}_\mu$. From (4.35a), the relation

$$\mathcal{L}_u = \hat{F}_{1;\dot{x}_1}^{-1}\Big(\hat{F}_{1;x_3}\hat{F}_{2;x_3}^{-1}\hat{F}_{2;u} - \hat{F}_{1;u}\Big)$$

is obtained on $\mathbb{L}_\mu$. With the same motivation, the expression for $\mathcal{S}(x_1)$ can be obtained from $\hat{L}_u$, defined in (4.35b), with $u = 0$ as

$$\mathcal{S} = \hat{L}_u(x_1, 0) = L_u(x_1, x_3, 0) - L_{x_3}(x_1, x_3, 0)\hat{F}_{2;x_3}^{-1}(x_1, x_3, 0)\hat{F}_{2;u}(x_1, x_3, 0)$$

on $\mathbb{L}_\mu$. Finally, the expression for $\mathcal{R}_c(x_1)$ is obtained as the Hessian of $\hat{L}$ w.r.t. u as can be seen in (6.6). The expression for the Hessian in the original functions is

$$\hat{L}_{uu}(x_1, u) = \mathcal{R}_u^T L_{x_3x_3}\mathcal{R}_u + 2L_{x_3u}\mathcal{R}_u + L_{uu} + \sum_{i=1}^{a} L_{x_{3i}}\mathcal{R}_{i;uu}$$

where $\mathcal{R}_{i;uu}$ is the Hessian of $\mathcal{R}_i$, which is the implicit function corresponding to x_{3i}, and the terms derived from L and $\mathcal{R}$ are evaluated at $\big(x_1, \mathcal{R}(x_1, u), u\big)$ and (x_1, u), respectively. On the solution manifold $\mathbb{L}_\mu$, it is possible to formulate all expressions in terms of the original coordinates and functions, *i.e.*, $(x_1, x_3, u) = \big(x_1, \mathcal{R}(x_1, u), u\big)$ and

$$\mathcal{R}_u(x_1, u) = -\hat{F}_{2;x_3}^{-1}F_{2;u}, \quad \mathcal{R}_{uu}(x_1, u) = -\frac{\partial}{\partial u}\hat{F}_{2;x_3}^{-1}F_{2;u} \tag{6.7}$$

Let the elements in $\mathcal{R}_u$ be denoted $c_{i,j}$ where $i = 1, \ldots, a$ and $j = 1, \ldots, p$. These elements can be written as

$$c_{i,j}\big(x_1, u\big) = \sum_{r=1}^{a} a_{i,r}\big(x_1, \mathcal{R}(x_1, u), u\big)b_{r,j}\big(x_1, \mathcal{R}(x_1, u), u\big)$$

where $a_{i,r}$ and $b_{r,j}$ are the elements in $F_{2;x_3}^{-1}$ and $F_{2;u}$, respectively. The elements in $\mathcal{R}_{uu}$ are computed by differentiating $c_{i,j}$ w.r.t. u. The result is

$$c_{i,j;u}\big(x_1, u\big) = \sum_{r=1}^{a} b_{r,j}(a_{i,r;x_3}\mathcal{R}_u + a_{i,r;u}) + a_{i,r}(b_{r,j;x_3}\mathcal{R}_u + b_{r,j;u})$$

on $\mathbb{L}_\mu$, using R_u in (6.7), and where $a_{i,r}$ and $b_{r,j}$ are evaluated at (x_1, x_3, u). The derivatives of $a_{i,r}$, *i.e.*, the coefficients of $F_{2;x_3}^{-1}$, w.r.t. to x_3 and u can be obtained without differentiating the inverse as

$$\frac{\partial \hat{F}_{2;x_3}^{-1}}{\partial z} = -\hat{F}_{2;x_3}^{-1} \frac{\partial \hat{F}_{2;x_3}}{\partial z} \hat{F}_{2;x_3}^{-1}$$

where z can be either x_{3i}, $i = 1, \ldots, a$ or u_i, $i = 1, \ldots, p$. Note that for all functions in this expression, the Taylor series can be computed easily, without doing a symbolic matrix inverse. In the case when x_3 is scalar, the expressions can be simplified to

$$\mathcal{R}_{uu} = -\hat{F}_{2;x_3}^{-1} \left(\mathcal{R}_u^T \hat{F}_{2;x_3x_3} \mathcal{R}_u + 2\hat{F}_{2;x_3u} \mathcal{R}_u + \hat{F}_{2;uu} \right)$$

To conclude, the following lemma is shown.

Lemma 6.1
Consider a DAE model for which the underlying state-space model is in the form (6.5) and for which the reduced cost $\hat{L}$ has the structure in (6.6). Then, on the solution manifold $\mathbb{L}_\mu$, the optimal feedback law u_* can be written in terms of the original functions as

$$u_* = -\frac{1}{2} \hat{L}_{uu}(x_1, u_*)^{-1} \Big(V_{x_1}(x_1) \mathcal{L}_u + \hat{L}_u(x_1, 0) \Big) \tag{6.8}$$

with $\mathcal{L}_u$, $\hat{L}_u$ and $\hat{L}_{uu}$ defined as above.

Proof: Follows from the discussion above. □

This is one motivation to why it is interesting to study when the underlying state-space model becomes affine in some external input signal. In addition, it is also interesting from a structural point of view.

6.2 Conditions for Control Affinity

The following sections describe conditions under which the underlying state-space model of either an implicit ODE model or a DAE model, satisfying Hypothesis 2.2, becomes affine in an external signal u. It is also shown, that if the underlying state-space model is affine in u, the implicit ODE model or the DAE model must be equivalent to a model with a certain structure.

6.2.1 Implicit ODE Models

Consider

$$F(\dot{x}, x, u) = 0 \tag{6.9}$$

where x and u are n- and p-vectors, respectively, and F is a continuously differentiable function from $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p$ to $\mathbb{R}^n$.

Assumption A12. There exists a point (p_o, x_o, v_o) such that

$$F(p_o, x_o, u_o) = 0, \quad F_{\dot{x}}(p_o, x_o, u_o) \text{ nonsingular} \tag{6.10}$$

Theorem 6.1
Suppose that Assumption A12 is satisfied. Then the implicit ODE (6.9) *is equivalent to an explicit ODE, affine in* u*,*

$$\dot{x} = f_1(x) + f_2(x)u \tag{6.11}$$

in a neighborhood of (x_o, u_o)*, if and only if there exists an* $n \times m$ *matrix function* $M(x)$ *such that*

$$F_{\dot{x}}(\dot{x}, x, u)M(x) + F_u(\dot{x}, x, u) = 0 \tag{6.12}$$

in a neighborhood of (x_o, u_o) *on* $\mathbb{L}_0 = \big\{(\dot{x}, x, u) \in \mathbb{R}^{2n+p} \mid F(\dot{x}, x, u) = 0\big\}$.

Note that it is sufficient that the condition is satisfied on the manifold, but in practice it is sometimes easier to verify the condition on a larger region.

Proof: *Necessity.*
If (6.11) holds then

$$F(f_1(x) + f_2(x)u, x, u) = 0$$

identically for x and u in the neighborhood of (x_0, u_0). It follows that

$$F_{\dot{x}}\big(f_1(x) + f_2(x)u, x, u\big)f_2(x) + F_u\big(f_1(x) + f_2(x)u, x, u\big) = 0$$

which is (6.12) with $M(x) = f_2(x)$.
Sufficiency.
First, the implications of Assumption A12 are studied. Application of the implicit function theorem on (6.9) shows that locally around (x_o, u_o)

$$\dot{x} = f(x, u)$$

for some function f. Differentiating the relation

$$F(f(x, u), x, u) = 0$$

with respect to u gives

$$F_{\dot{x}} f_u(x, u) + F_u = 0 \tag{6.13}$$

where the terms without arguments are evaluated at $(f(x, u), x, u)$. Now assume that there exist an $M(x)$ such that (6.12) is satisfied. A comparison of (6.13) and (6.12) then shows that on $\mathbb{L}_0$ in the neighborhood of (x_0, u_0) it must hold that

$$F_{\dot{x}}(f_u - M(x)) = 0$$

or, since $F_{\dot{x}}$ is nonsingular,

$$f_u(x, u) = M(x)$$

Since the Jacobian of f with respect to u is independent of u, the function f has to be $f(x, u) = f_1(x) + M(x)u$, *i.e.*, it is affine with respect to u. □

The next theorem shows that the structure of F can also be used to reach the same conclusion about affinity.

Theorem 6.2
The implicit ODE (6.9) *is equivalent to an explicit ODE, affine in* u*,*

$$\dot{x} = f_1(x) + f_2(x)u$$

in a neighborhood of (x_o, u_o)*, if and only if there exists an* $n \times m$ *matrix function* $M(x)$ *and a function* $\bar{F}(p, x)$ *such that*

$$F(\dot{x}, x, u) = \bar{F}(\dot{x} - M(x)u, x) \tag{6.14}$$

in a neighborhood of (x_o, u_o) *on* $\mathbb{L}_0 = \big\{(\dot{x}, x, u) \in \mathbb{R}^{2n+p} \mid F(\dot{x}, x, u) = 0\big\}$.

Proof: It is known from Theorem 6.1 that (6.9) is equivalent to an affine system if and only if there is a matrix function $M(x)$ such that (6.12) is satisfied. Therefore, it is sufficient to prove that condition (6.14) is equivalent to condition (6.12).

Let there exist a matrix $M(x)$ satisfying (6.12) in a neighborhood of a point (x_0, u_0) on $\mathbb{L}_0$. Introduce the coordinate change

$$\dot{x} = \dot{y} + M(y)w, \quad x = y, \quad u = w$$

and define a function $\tilde{F}(\dot{y}, y, w)$ as

$$F(\dot{x}, x, u) = F(\dot{y} - M(y)w, y, w) = \tilde{F}(\dot{y}, y, w) \tag{6.15}$$

where $\dot{y} = \dot{y}(\dot{x}, x, u)$, $y = y(x)$ and $w = w(u)$. Using the relations above, if follows that

$$\begin{aligned} F_{\dot{x}}(\dot{x}, x, u) &= \tilde{F}_{\dot{y}}(\dot{y}, y, w)\dot{y}_{\dot{x}}(\dot{x}, x, u), \\ F_u(\dot{x}, x, u) &= \tilde{F}_{\dot{y}}(\dot{y}, y, w)\dot{y}_u(\dot{x}, x, u) + \tilde{F}_w(\dot{y}, y, w)w_u(u) \end{aligned}$$

For $(\dot{x}, x, u) \in \mathbb{L}_0$ in the neighborhood of (x_0, u_0), condition (6.12) transforms to

$$\begin{aligned} 0 &= F_{\dot{x}}(\dot{x}, x, u)M(x) + F_u(\dot{x}, x, u) \\ &= \tilde{F}_{\dot{y}}(\dot{y}, y, w)M(x) - \tilde{F}_{\dot{y}}(\dot{y}, y, w)M(x) + \tilde{F}_w(\dot{y}, y, w) = \tilde{F}_w(\dot{y}, y, w) \end{aligned}$$

Since the partial derivative of $\tilde{F}$ w.r.t. to w is zero in an open interval (corresponding to the open interval for u), it means that a function $\bar{F}$ exists such that

$$\tilde{F}(\dot{y}, y, w) = \bar{F}(\dot{y}, y) = \bar{F}(\dot{x} - M(x)u, x)$$

which is condition (6.14) with the same $M(x)$ as in condition (6.12).

In the other direction, it follows by straightforward calculations. □

6.2.2 DAE Models with Algebraic Equations Independent of the External Input

Consider the DAE model (6.9). Assume that it satisfies Hypothesis 2.2, *i.e.*, assume there exist matrices Z_1, Z_2 and a partitioning $x = (x_1, x_3)$ of the variables such that the system (6.9) can be written as

$$\hat{F}_1(\dot{x}_1, \dot{x}_3, x_1, x_3, u) = 0 \tag{6.16a}$$

$$\hat{F}_2(x_1, x_3, \mathbf{u}_\mu) = 0 \tag{6.16b}$$

where $\hat{F}_1 = Z_1^T F$, $\hat{F}_2 = Z_2^T F_\mu$ and $\mathbf{u}_\mu = (u, \dot{u}, \ldots, u^{(\mu)})$. The hypothesis guarantees that (6.9) can be rewritten as (6.16) for every

$$(\mathbf{x}_{\mu+1,0}, \mathbf{u}_{\mu+1,0}) = (x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}, u_0, \dot{u}_0, \ldots, u_0^{(\mu+1)})$$

on

$$\mathbb{L}_\mu = \left\{(x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}, u_0, \dot{u}_0, \ldots, u_0^{(\mu+1)}) \in \mathbb{R}^{(\mu+2)(n+p)+p} \mid F_\mu = 0\right\}$$

However, here only one point in $\mathbb{L}_\mu$ is considered in order to have fixed Z_1 and Z_2. From Hypothesis 2.2, it also follows that the Jacobian of $\hat{F}_2$ with respect to x_3 is nonsingular on $\mathbb{L}_\mu$. It means that

$$x_3 = \mathcal{R}(x_1, \mathbf{u}_\mu) \tag{6.17}$$

for some function $\mathcal{R}$. Likewise, the Jacobian of

$$\hat{F}_1(\dot{x}_1, \dot{\mathcal{R}}(x_1, \mathbf{u}_\mu), x_1, \mathcal{R}(x_1, \mathbf{u}_\mu), u)$$

with respect to $\dot{x}_1$ is nonsingular.

Introduce the following assumption.

Assumption A13. $Z_2^T F_{\mu;u,\dot{u},\ldots,u^{(\mu)}} = 0$ on $\mathbb{L}_\mu$.

This assumption will ensure that the algebraic equations are independent of u and its derivatives. It may seem like a restrictive assumption, but sometimes it is physically motivated as will be seen in later sections. Under this assumption, the following result can be shown.

Theorem 6.3
Suppose the DAE model (6.9) satisfies Assumption A13. Then, it is equivalent to the following model, i.e.*, an explicit ODE that is affine in u and some static relations,*

$$\dot{x}_1 = f_1(x_1) + f_2(x_1)u \tag{6.18a}$$
$$x_3 = \mathcal{R}(x_1) \tag{6.18b}$$

in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$, if and only if there exists a $d \times m$ matrix function $M(x_1)$ such that

$$\left(\hat{F}_{1;\dot{x}_1}(\dot{x}, x, u) - \hat{F}_{1;\dot{x}_3}(\dot{x}, x, u)\hat{F}_{2;x_3}^{-1}(x, \mathbf{u}_\mu)\hat{F}_{2;x_1}(x, \mathbf{u}_\mu)\right)M(x_1) + \hat{F}_{1;u}(\dot{x}, x, u) = 0 \tag{6.19}$$

on $\mathbb{L}_\mu$ in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu+1,0})$.

In the condition, the matrix M can also be allowed to explicitly depend not only on x_1 but also on x_3, since on $\mathbb{L}_\mu$, x_3 will be a function of x_1. This fact may sometimes simplify the computations.

Proof: *Necessity.*
If (6.18) holds then

$$\hat{F}_2\big(x_1, \mathcal{R}(x_1), \mathbf{u}_\mu\big) = 0$$

identically in x_1, u ,..., $u^{(\mu)}$. Differentiating the equation w.r.t. x_1 gives

$$\mathcal{R}_{x_1}(x_1) = -\hat{F}_{2;x_3}^{-1}\big(x_1, \mathcal{R}(x_1), \mathbf{u}_\mu\big)\hat{F}_{2;x_1}\big(x_1, \mathcal{R}(x_1), \mathbf{u}_\mu\big) \tag{6.20}$$

Equation (6.19) can, if (6.18) is valid, be written as

$$\hat{F}_1\Big(f_1(x_1) + f_2(x_1)u, \mathcal{R}_{x_1}(x_1)\big(f_1(x_1) + f_2(x_1)u\big), x_1, \mathcal{R}(x_1), u\Big) = 0$$

identically in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$. Differentiation w.r.t. u will give that

$$\big(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}\big)f_2 + \hat{F}_{1;u} = 0$$

and by using (6.20), it follows that (6.12) is obtained with $M = f_2$.
Sufficiency.
First, the implications of the assumptions are studied. From the fact that (6.9) satisfies Hypothesis 2.2, it follows that

$$\begin{aligned}\dot{x}_1 &= \mathcal{L}(x_1, \mathbf{u}_{\mu+1})\\ x_3 &= \mathcal{R}(x_1, \mathbf{u}_\mu)\end{aligned}$$

on $\mathbb{L}_\mu$ in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu+1,0})$. Differentiating the relation

$$\hat{F}_2\big(x_1, \mathcal{R}(x_1, \mathbf{u}_\mu), \mathbf{u}_\mu\big) = 0 \tag{6.21}$$

with respect to u gives

$$\hat{F}_{2;x_3}\mathcal{R}_u + \hat{F}_{2;u} = 0$$

which since $\hat{F}_{2;x_3}$ has full rank on $\mathbb{L}_\mu$ can be solved for $\mathcal{R}_u$ as

$$\mathcal{R}_u = -\hat{F}_{2;x_3}^{-1}\hat{F}_{2;u}$$

However,

$$\hat{F}_{2;u} = \frac{\partial}{\partial u}(Z_2^T F_\mu) = Z_{2;u}^T F_\mu + Z_2^T F_{\mu;u} = 0$$

since $F_\mu = 0$ for solutions and $Z_2^T F_{\mu;u} = 0$ from Assumption A13. It means that $\mathcal{R}_u = 0$ and $\mathcal{R}$ is therefore independent of u. Repeating the process for $\dot{u}, \ldots, u^{(\mu)}$, yields that $\mathcal{R}$ is also independent of those and can be written as (6.18b). Finally, differentiating (6.21) w.r.t. x_1 gives

$$\mathcal{R}_{x_1}(x_1) = -\hat{F}_{2;x_3}^{-1}(x, \mathbf{u}_\mu)\hat{F}_{2;x_1}(x, \mathbf{u}_\mu) \tag{6.22}$$

on $\mathbb{L}_\mu$ in the neighborhood of $(x_{1,0}, \mathbf{u}_0)$. The derivative of x_3 w.r.t. t can be found as

$$\dot{x}_3 = \mathcal{R}_{x_1}(x_1)\dot{x}_1 = \mathcal{R}_{x_1}(x_1)\mathcal{L}(x_1, \mathbf{u}_{\mu+1})$$

which substituted into $\hat{F}_1$ gives

$$\hat{F}_1\big(\mathcal{L}(x_1, \mathbf{u}_{\mu+1}), \mathcal{R}_{x_1}(x_1)\mathcal{L}(x_1, \mathbf{u}_{\mu+1}), x_1, \mathcal{R}(x_1), u\big) = 0 \tag{6.23}$$

which holds identically in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu+1,0})$. The derivative of this expression with respect to $\dot{u}$ becomes

$$\big(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}\big)\mathcal{L}_{\dot{u}} = 0 \Rightarrow \mathcal{L}_{\dot{u}} = 0$$

and by the same argument also $\mathcal{L}_{\ddot{u}}, \ldots, \mathcal{L}_{u^{(\mu+1)}}$ become zero. Hence, $\mathcal{L}$ does not depend on derivatives of u. Differentiation of (6.23) with respect to u yields

$$\big(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}\big)\mathcal{L}_u + \hat{F}_{1;u} = 0 \tag{6.24}$$

Now, let there exist a matrix function $M(x_1)$ such that (6.19) is satisfied in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$. On this neighborhood, the condition (6.19) can be written as

$$\big(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}\big)M(x_1) + F_{1;u} = 0$$

using (6.20). A comparison between this expression and (6.24) shows that

$$\big(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}\big)(\mathcal{L}_u - M(x_1)) = 0$$

or that

$$\mathcal{L}_u(x, u) = M(x_1)$$

since $\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}$ is nonsingular on $\mathbb{L}_\mu$. Hence, $\mathcal{L}(x_1, u) = \mathcal{L}_1(x_1) + M(x_1)u$ which means that it is affine in u. □

The next theorem shows that also for DAE models the affinity can be deduced to a certain structure of the functions involved.

Theorem 6.4
Suppose Assumption A13 is satisfied. Then the DAE (6.9) is equivalent to an explicit ODE, affine in u as in (6.18) in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu+1,0})$ on $\mathbb{L}_\mu$, if and only if there exist an $d \times p$ matrix function $M(x_1)$ and a function $\bar{F}_1(y, u)$ such that

$$\hat{F}_1(\dot{x}, x, u) = \bar{F}_1\big(\dot{x}_1 - M(x_1)u, x_1\big) \tag{6.25}$$

in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$.

Proof: It is known that the first statement is satisfied if and only if there is a matrix functions $M(x_1)$ such that (6.19) is satisfied in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu+1,0})$ on $\mathbb{L}_\mu$. Therefore, it is sufficient to prove that (6.25) is equivalent to (6.19).
(6.25) ⇒ (6.19).
Let there exist a matrix $M(x_1)$ and a function $\bar{F}_1(\dot{y}, y)$ satisfying (6.25) in a neighborhood of $(x_0, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$. Under the given assumptions, it is known that $x_3 = \mathcal{R}(x_1)$. Condition (6.25) can therefore be written as

$$\hat{F}_1\big(\dot{x}_1, \mathcal{R}_{x_1}(x_1)\dot{x}_1, x_1, \mathcal{R}(x_1), u\big) = \bar{F}_1\big(\dot{x}_1 - M(x_1)u, x_1\big) \tag{6.26}$$

in a neighborhood of $(x_0, \mathbf{u}_{\mu,0})$. Differentiation of the equation above w.r.t. to $\dot{x}_1$ and u yields

$$\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1} = \bar{F}_{1;\dot{y}}, \quad \hat{F}_{1;\dot{x}_1} = -\bar{F}_{1;\dot{y}}M(x_1)$$

where the terms are evaluated in the arguments as in (6.26). It means that

$$\big(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}\big)M(x_1) + \hat{F}_{1;u} = 0 \tag{6.27}$$

In the neighborhood of $(x_{10}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$, (6.27) can be rewritten as (6.19) using $x_3 = \mathcal{R}(x_1)$ and (6.22).
(6.19) ⇒ (6.25).
Let there exist a matrix $M(x_1)$ satisfying (6.19) in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu+1,0})$ on $\mathbb{L}_\mu$. On this set, it is known that $x_3 = \mathcal{R}(x_1)$ and the derivative of $\mathcal{R}$ w.r.t. x_1 is given by (6.20). Hence, (6.19) can be written as

$$\big(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1}(x_1)\big)M(x_1) + \hat{F}_{1;u} = 0 \tag{6.28}$$

where the functions without arguments are evaluated in
$(\dot{x}_1, \mathcal{R}_{x_1}(x_1)\dot{x}_1, x_1, \mathcal{R}(x_1), u)$.

Introduce the change of coordinates

$$\begin{aligned} \dot{x}_1 &= \dot{y} + M(y)w, & x_1 &= y & u &= w \\ \dot{x}_3 &= \mathcal{R}_{x_1}(y)\big(\dot{y} + M(y)w\big), & x_3 &= \mathcal{R}(y) & & \end{aligned}$$

Using this coordinate change, it follows that a function $\tilde{F}_1(\dot{y}, y, w)$ can be defined as

$$\begin{aligned} \hat{F}_1(\dot{x}_1, \dot{x}_3, x_1, x_3, u) &= \hat{F}_1\big(\dot{y} + M(y)w, \mathcal{R}_{x_1}(y)\big(\dot{y} + M(y)w\big), y, \mathcal{R}(y), u\big) \\ &= \tilde{F}_1(\dot{y}, y, w) = \tilde{F}_1\big(\dot{y}(\dot{x}_1, x_1, u), y(x_1), w(u)\big) \end{aligned}$$

in the neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$ and the following relations for the derivatives can be obtained

$$\begin{aligned} \hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1} &= \tilde{F}_{1;\dot{y}}\dot{y}_{\dot{x}_1} = \tilde{F}_{1;\dot{y}} \\ \hat{F}_{1;u} &= \tilde{F}_{1;\dot{y}}\dot{y}_u + \tilde{F}_{1;w}w_u = -\tilde{F}_{1;\dot{y}}M(x_1) + \tilde{F}_{1;w} \end{aligned}$$

where the transformed variables $\dot{y}$, y and w are considered at points that correspond to the neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$. Using the relations above together with the reformulated version of (6.19), that is, condition (6.28), it follows that

$$0 = (\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_3}\mathcal{R}_{x_1})M(x_1) + \hat{F}_{1;u} = \tilde{F}_{1;\dot{y}}M(x_1) - \tilde{F}_{1;\dot{y}}M(x_1) + \tilde{F}_{1;w} = \tilde{F}_{1;w}$$

on $\mathbb{L}_\mu$ around $(x_{1,0}, \mathbf{u}_\mu)$. Hence

$$\hat{F}_1(\dot{x}_1, x_1, u) = \tilde{F}_1(\dot{y}, y, w) = \bar{F}_1(\dot{y}, y) = \bar{F}_1(\dot{x}_1 - M(x_1)u, x_1)$$

□

6.2.3 DAE Models with Algebraic Equations Affine in the External Input

Earlier, it has been assumed that $Z_2^T F_{\mu;u,\dot{u},\ldots,u^{(\mu)}} = 0$. The reason was to ensure that $\mathcal{R}$ is independent of u which sometimes is physically motivated. However, in other cases, the static relations may depend on u and then Assumption A13 is too restrictive.

Instead, the following assumption is introduced.

Assumption A14. $\hat{F}_{1;\dot{x}_3} = 0$ and $\hat{F}_{2;\dot{u},\ldots,u^{(\mu)}} = 0$

Note that since it is assumed that $\hat{F}_{1;\dot{x}_3} = 0$, it means that $\hat{F}_{1;\dot{x}_1}$ will be nonsingular in order to satisfy the conditions in Hypothesis 2.2.

Then, the following result can be shown.

Theorem 6.5
The DAE (6.9) is equivalent to an explicit ODE, affine in u, and static relations

$$\dot{x}_1 = f_1(x_1) + f_2(x_1)u \tag{6.29a}$$
$$x_3 = \mathcal{R}_1(x_1) + \mathcal{R}_2(x_1)u \tag{6.29b}$$

in a neighborhood of $(x_{1,0}, \mathbf{u}_\mu)$, if and only if there exist matrix functions $M_1(x_1)$ and $M_2(x_1)$ of dimensions $d \times d$ and $d \times a$, respectively, such that

$$\hat{F}_{2;x_3}(x, \mathbf{u}_\mu)M_2(x_1) + \hat{F}_{2;u}(x, \mathbf{u}_\mu) = 0 \tag{6.30a}$$
$$\hat{F}_{1;\dot{x}_1}(\dot{x}, x, u)M_1(x_1) + \hat{F}_{1;x_3}(\dot{x}, x, u)M_2(x_1) + \hat{F}_{1;u}(\dot{x}, x, u) = 0 \tag{6.30b}$$

in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$.

In this case, the matrix functions M_1 and M_2 are only allowed to depend on x_1 in the general case. The reason is of course that x_3 may depend on u. However, if some of the variables x_3 are known to be independent of x_1, these x_3 can also be included in the matrices.

Proof: *Necessity.*
Assume that (6.16) can be written as (6.29). Then

$$\hat{F}_2(x_1, \mathcal{R}_1(x_1) + \mathcal{R}_2(x_1)u, \mathbf{u}_\mu) = 0$$

identically on the neighborhood, which differentiated w.r.t. u gives

$$F_{2;x_3}\mathcal{R}_2 + \hat{F}_{2;u} = 0$$

Hence, the same relationship as in (6.30a) is satisfied with $M_2 = \mathcal{R}_2$ on a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$.

The equation for $\hat{F}_1$ becomes

$$\hat{F}_1\big(f_1 + f_2u, \dot{x}_3, x_1, \mathcal{R}_1 + \mathcal{R}_2u, u\big) = 0$$

where $\dot{x}_3$ is left as it is, since $\hat{F}_{1;\dot{x}_3} = 0$ and the exact expression is not needed. Differentiation w.r.t. u gives

$$\hat{F}_{1;\dot{x}_1}f_2 + \hat{F}_{1;\dot{x}_3}(\cdot) + \hat{F}_{1;x_3}\mathcal{R}_2 + \hat{F}_{1;u} = \hat{F}_{1;\dot{x}_1}f_2 + \hat{F}_{1;x_3}\mathcal{R}_2 + \hat{F}_{1;u} = 0$$

which is the same as (6.30b) with $M_1 = f_2$ and $M_2 = \mathcal{R}_2$.
Sufficiency.
The first step is to analyze what the assumptions ensure. Since the model is assumed to satisfy Hypothesis 2.2, it is known that

$$\dot{x}_1 = \mathcal{L}(x_1, \mathbf{u}_{\mu+1})$$
$$x_3 = \mathcal{R}(x_1, \mathbf{u}_\mu)$$

and that

$$\hat{F}_2\big(x_1, \mathcal{R}(x_1, \mathbf{u}_\mu), \mathbf{u}_\mu\big) = 0 \tag{6.31}$$

identically in the variables. Differentiation w.r.t. $\dot{u}$ gives

$$\hat{F}_{2;x_3}\mathcal{R}_{\dot{u}} + \hat{F}_{2;\dot{u}} = 0$$

which from Assumption A14 and that $\hat{F}_{2;x_3}$ has full rank implies that $\mathcal{R}$ is independent of $\dot{u}$. Repeating the same procedure for the higher derivatives of u shows that $\mathcal{R}$ is independent of them all, *i.e.*, $\mathcal{R}(x_1, u)$. Finally, if (6.31) is differentiated w.r.t. u the result is

$$\hat{F}_{2;x_3}\big(x_1, \mathcal{R}(x_1, \mathbf{u}_\mu), \mathbf{u}_\mu\big)\mathcal{R}_u(x_1, u) + \hat{F}_{2;u}\big(x_1, \mathcal{R}(x_1, \mathbf{u}_\mu), \mathbf{u}_\mu\big) = 0 \tag{6.32}$$

Now assume that there exist matrix function $M_2(x_1)$ such that (6.30a) is satisfied. A comparison between (6.32) and (6.30a) yields that $\mathcal{R}_u(x_1, u) = M_2(x_1)$, which means that $\mathcal{R}(x_1, u) = \mathcal{R}_1(x_1) + M_2(x_1)u$, *i.e.*, $\mathcal{R}$ is affine in u.

Now consider $\mathcal{L}$. Using the obtained expression for x_3, the expression for $\hat{F}_1$ becomes

$$\hat{F}_1(\mathcal{L}, \dot{x}_3, x_1, \mathcal{R}_1 + M_2 u, u) = 0$$

where $\dot{x}_3$ is left unevaluated, since it will not be needed in the calculations. The derivative of the equation above w.r.t. $\dot{u}$ becomes

$$\hat{F}_{1;\dot{x}_1}\mathcal{L}_{\dot{u}} + \hat{F}_{1;\dot{x}_3}(\cdot) = 0$$

and from Assumption A14 together with that $\hat{F}_{1;\dot{x}_1}$ is nonsingular, it follows that $\mathcal{L}_{\dot{u}} = 0$. That is, $\mathcal{L}$ is independent of $\dot{u}$. The same procedure can be performed for the higher derivatives of u with the same result.

Differentiation of $\hat{F}_1$ w.r.t. u instead, gives

$$\hat{F}_{1;\dot{x}_1}\mathcal{L}_u + \hat{F}_{1;\dot{x}_3}(\cdot) + \hat{F}_{1;x_3}M_2 + \hat{F}_{1;u} = \hat{F}_{1;\dot{x}_1}\mathcal{L}_u + \hat{F}_{1;x_3}M_2 + \hat{F}_{1;u} = 0$$

If a $M_1(x_1)$ exists such that (6.30b) is satisfied, a comparison of the obtained expression with (6.30b) yields that

$$\hat{F}_{1;\dot{x}_1}(\mathcal{L}_u - M_1) = 0$$

and hence $\mathcal{L}(x_1) = \mathcal{L}_1(x_1) + M_1(x_1)u$. □

The structure of a DAE model satisfying the conditions must also be in a certain way as proven in the following theorem.

Theorem 6.6

Suppose Assumption A14 is satisfied. Then the DAE (6.9) is equivalent to an explicit ODE, affine in u as in (6.29) in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu+1,0})$ on $\mathbb{L}_\mu$, if and only if there exist matrix functions $M_1(x_1)$ and $M_2(x_1)$ of dimensions $d \times d$ and $d \times a$, respectively, and functions $\bar{F}_1(y, u)$ and $\bar{F}_2(y, u)$ such that

$$\hat{F}_1(\dot{x}, x, u) = \bar{F}_1\big(\dot{x}_1 - M_1(x_1)u, x_1\big) \tag{6.33a}$$

$$\hat{F}_2(x, u) = \bar{F}_2\big(x_1, x_3 - M_2(x_1)u\big) \tag{6.33b}$$

in a neighborhood of $(x_{1,0}, \mathbf{u}_{\mu,0})$ on $\mathbb{L}_\mu$.

Proof: The proof follows the same line as the proof of Theorem 6.3. □

Two Special Cases

In this section, two rather common special cases are presented. The structure in the first case is, except for affine in u, also affine in x_3.

Corollary 6.1
Consider the following DAE model

$$0 = \bar{F}_1\big(\dot{x}_1 - \sigma_1(x_1)u - \sigma_2(x_1)x_3, x_1\big) \tag{6.34a}$$
$$0 = \bar{F}_2\big(x_1, x_3 - \mathcal{R}_2(x_1)u\big) \tag{6.34b}$$

Suppose it satisfies the conditions in Hypothesis 2.2. Then it is equivalent to the control-affine model

$$\begin{aligned}\dot{x}_1 &= \mathcal{L}_1(x_1) + \sigma_2(x_1)\mathcal{R}_1(x_1) + \big(\sigma_1(x_1) + \sigma_2(x_1)\mathcal{R}_2(x_1)\big)u\\ x_3 &= \mathcal{R}_1(x_1) + \mathcal{R}_2(x_1)u\end{aligned}$$

Another common case is when the algebraic constraints are known to have one part that depend on u and one part that does not. Furthermore, the part that not contains u should also be solvable for one part of x_3, denoted $x_{3,1}$.

Corollary 6.2
Consider the following DAE model

$$0 = \bar{F}_1\big(\dot{x}_1 - \sigma_1(x_1, x_{3,1})u - \sigma_2(x_1, x_{3,1})x_{3,2}, x_1, x_{3,1}\big) \tag{6.35a}$$
$$0 = \bar{F}_{21}(x_1, x_{3,1}) \tag{6.35b}$$
$$0 = \bar{F}_{22}\big(x_1, x_{3,1}, x_{3,2} - \mathcal{R}_2(x_1, x_{3,1})u\big) \tag{6.35c}$$

Suppose it satisfies the conditions in Hypothesis 2.2. Then it is equivalent to the control-affine model

$$\begin{aligned}\dot{x}_1 &= \mathcal{L}_1(x_1) + \sigma_2(x_1)\mathcal{R}_1(x_1) + \big(\sigma_1(x_1) + \sigma_2(x_1)\mathcal{R}_2(x_1)\big)u\\ x_3 &= \mathcal{R}_1(x_1) + \mathcal{R}_2(x_1)u\end{aligned}$$

6.2.4 Conditions on the Original DAE Model

All requirements are given at the strangeness-free level. The main reason is that algebraically the solution manifold is not visible for a higher strangeness index. That is, there are more algebraic connections between the variables than visible in the equations. However, there are possible to give some simple relations on the original expressions.

The main idea is to use the definitions of $\hat{F}_1$ and $\hat{F}_2$. Equation (6.30) can then be written as

$$\begin{aligned}\frac{\partial}{\partial x_3}(Z_2^T F_\mu)M_2 + \frac{\partial}{\partial u}(Z_2^T F_\mu) &= 0\\ \frac{\partial}{\partial \dot{x}_1}(Z_1^T F)M_1 + \frac{\partial}{\partial x_3}(Z_1^T F)M_2 + \frac{\partial}{\partial u}(Z_1^T F) &= 0\end{aligned}$$

which should be satisfied on the solution manifold in a neighborhood of some point. Since Z_1 is constant and $F_\mu = 0$ on the solution manifold, it is possible to move the derivatives through Z_1 and Z_2, respectively, and obtain

$$Z_2^T \left(F_{\mu;x_3} M_2 + F_{\mu;u} \right) = 0$$
$$Z_1^T \left(F_{\dot{x}_1} M_1 + F_{x_3} M_2 + F_u \right) = 0$$

which then should be valid on $\mathbb{L}_\mu$ in the considered neighborhood. These expressions do not contain $\hat{F}_1$ and $\hat{F}_2$, but they still depend on Z_1, Z_2, and the classification of the variables in x as either dynamic or algebraic. However, in some cases this information can be obtained from the physical process.

6.2.5 Test of the Conditions

The conditions above can for many DAE models be hard to verify. As usual, the reason is that the conditions should be valid for points on the solution manifold. If the solution manifold is defined by implicit functions, the matrices M_1 and M_2 will normally be implicitly defined as well, at least in x_1. However, a feature for some models is that the conditions actually are satisfied on a larger region than just on the manifold. This fact can simplify the test of the conditions to some extent. If the conditions only are satisfied on the manifold and the manifold is implicitly defined, different approximate methods can be used, such as power series solution, but then the simpler test in the next section is probably preferred.

6.2.6 Basic Tests Indicating Control Affinity

The conditions above are exact and yields both necessity and sufficiency. However, they may be hard to compute analytically. Therefore, consider a control-affine system,

$$\dot{x} = f(x) + g(x)u \tag{6.36}$$

For a given x, it means that $\dot{x}$ is affine in u. This fact can of course be utilized, if a numerical solver is used. For $u = 0$, the value of $\dot{x}$ will be $f(x_0)$, where x_0 is the fixed point. If (6.36) is rearranged as

$$\dot{x} - f(x_0) = g(x_0)u$$

a practical test would then be to choose the control signal as

$$\ldots, u_0, 2u_0, -u_0, -2u_0, \ldots$$

If the corresponding values of $\dot{x} - f(x_0)$ changes nonlinearly, the system is not control-affine for that x_0. If x_0 is chosen in the interesting region, the system cannot be assumed to be control-affine. However, if it changes linearly there is a chance that the system is affine. If the test is performed for a large number of x_0 and a large number of u_0, in the specific region, it is a strong indication that methods that rely on control affinity can be used.

For state-space models, the test above is not very useful since most often the structure can be seen immediately from the equations. However, for large DAE-models, this might not be the case. If a numerical solver which computes $\dot{x}_1$ is used, it is possible to use the same test.

If the models are analytic, it is also possible to use the methods in Section 4.3.1. That is, the power series of $\dot{x}_1$ and x_3 can be computed. If the model is control-affine around the considered point, all coefficients that correspond to higher order terms in u should then be zero. This test is of course also only necessary and not sufficient.

6.3 Optimal Control

In this section, the results about control affinity are merged with conditions of the cost function. The easiest case, which in practice is quite common, is the following.

Corollary 6.3
Consider an optimal control problem with the model (6.34) and the cost function

$$L(x_1, u) = l(x_1) + S_1(x_1)^T u + u^T R(x_1) u \tag{6.37}$$

Then, locally, the optimal return and control law can be found by the equations

$$0 = L(x_1, x_3, u_*) + V_{x_1}(x_1)\dot{x}_1 \tag{6.38a}$$
$$0 = \bar{F}_1\big(\dot{x}_1 - \sigma_1(x_1)u_* - \sigma_2(x_1)x_3, x_1\big) \tag{6.38b}$$
$$0 = \bar{F}_2\big(x_1, x_3 - \mathcal{R}_2(x_1)u_*\big) \tag{6.38c}$$

where

$$u_* = -\frac{1}{2}R^{-1}(x_1)\Big(V_{x_1}(x_1)\big(\sigma_1(x_1) + \sigma_2(x_1)\mathcal{R}_2(x_1)\big) + S(x_1)\Big)^T$$

Proof: The model (6.34) and the cost function (6.37) match the structure described in (6.5) and (6.6) on the solution manifold, and the expression for u_* in (6.8) then gives the result. □

The number equations to solve will then be reduced to

$$\underbrace{\binom{d+m}{d} - (d+1)}_{V^{[m]}} + \underbrace{n\binom{d+p+m-1}{d+p} - n}_{\mathcal{L}^{[m-1]} \text{ and } \mathcal{R}^{[m-1]}} \tag{6.39}$$

This number is substantially smaller than (6.3) if either the number of variables or the desired order is large. If $\hat{F}_1$ can be solved explicitly for $\dot{x}_1$ the equations in (6.38) can be simplified even more since then $\mathcal{L}$ does not need to be solved for with power series yielding that the number of equations becomes

$$\underbrace{\binom{d+m}{d} - (d+1)}_{V^{[m]}} + \underbrace{a\binom{d+p+m-1}{d+p} - a}_{\mathcal{R}^{[m-1]}} \tag{6.40}$$

Two observations can be mentioned here. First, the inverse of $R(x_1, x_{3,1})$ requires a computation as described in (4.36). This computation is fairly easy, but since $R(x_1, x_{3,1})$ is a design choice, the computational burden can be decreased if it is chosen simple, for example, as a diagonal or a constant matrix function, or by defining the inverse. Second, the more general case with (6.35) and the cost function

$$L(x_1, x_{3,1}, x_{3,2}) = \\ l(x_1, x_{3,1}) + x_1 S_1(x_1, x_{3,1})^T u + x_{3,2}^T W(x_1, x_{3,1}) x_{3,2} + \\ x_{3,2}^T S_2(x_1, x_{3,1}) u + u^T R(x_1, x_{3,1}) u$$

can be handled in the same way.

6.4 Structure in the Equations

If some of the equations in $\hat{F}_1$ or $\hat{F}_2$ are easy to solve explicitly for $\dot{x}_1$ or x_3, this fact should of course be exploited. Another property which can simplify the computations substantially is if the variables are ordered such that they can be solved as separately as possible, *i.e.*, without having to solve one big set of equations. This property may not be satisfied in the original setting but there are methods, based on structural analysis (graph theory), that orders the variables in so-called block-lower-triangular (BLT) form which is a form with this property. A reference which discusses the BLT form but not in the case of equation solving is Duff and Reid (1978).

As an example, consider a set of algebraic equations in the form

$$0 = \hat{F}_2(x_1, x_3, u) = \begin{pmatrix} F_{21}(x_{3,3}, x_{3,4}, y_1) \\ F_{22}(x_{3,2}, y_2) \\ F_{23}(x_{3,2}, x_{3,3}, x_{25}, y_3) \\ F_{24}(x_{3,1}, x_{3,2}, y_4) \\ F_{25}(x_{3,1}, x_{3,3}, x_{3,5}, y_5) \end{pmatrix}$$

where y_i, $i = 1, \ldots, 5$ denotes the variables in x_1 and u present in the corresponding equation. If a BLT transformationis performed, the result will be

$$0 = \hat{F}_2(x_1, x_3, u) = \begin{pmatrix} F_{22}(\underline{x}_{3,2}, y_2) \\ F_{24}(\underline{x}_{21}, x_{3,2}, y_4) \\ F_{23}(x_{22}, \underline{x}_{3,3}, \underline{x}_{3.5}, y_3) \\ F_{25}(x_{21}, \underline{x}_{3,3}, \underline{x}_{3,5}, y_5) \\ F_{21}(x_{23}, \underline{x}_{3,4}, y_1) \end{pmatrix}$$

where the underlined variables denotes the variables to solve for in that equation. In the case above, it means that first F_{22} should be solved w.r.t. $x_{3,2}$. Having $x_{3,2}$, F_{24} can be solved for $x_{3,1}$ and so forth. In F_{23} and F_{25}, two variables are underlined. It means that those correspond to a so-called strong component. Strong components are groups that need to be solved simultaneous and the BLT transformationwill give that information as well.

The most obvious approach to find the power series of $x_{3,i}$, $i = 1, \ldots, 5$, would be to let $x_{3,2}$ depend on the variables in y_2 and then find the power series in the ordinary way. The next step could then be to let $x_{3,1}$ depend on $y_4 \cup y_1$ and so on. However, eventually $x_{3,i}$ will then in most cases depend on all variables in x_1 and u. This is of course true, but not efficient. A better way is to use $x_{3,i}$, $i = 1, \ldots, 5$ as parameters. It means that $x_{3,2}$ is computed as a function of y_2, $x_{3,1}$ is computed as a function of $x_{3,2}$ and y_4 and so forth. Finally, the power series of, for example, $x_{3,2}$ is obtained as the composition of its series expansion in $x_{3,2}$ and y_4, and the series expansion of $x_{3,2}$. Note that in the beginning of the solution process, it might still be better not to use the variables $x_{3,i}$ as parameters. For example, if $y_2 \subseteq y_4$ in the example above, it is better to parametrize in just y_4, since then the number of parameters will be one less and the computations at the end can be avoided.

Since it often seems like equations have structures like the one above, and especially equations generated in an object-oriented fashion, the method above can reduce the time rather substantially.

6.5 Mechanical Systems

Mechanical systems can be shown to satisfy the conditions in Section 6.2.3 for quite general configurations. Consider holonomic multibody systems expressed in first order form (Kunkel and Mehrmann, 2006)

$$\dot{p} = q \tag{6.41a}$$
$$M(p)\dot{q} = F(p, q, u) + \lambda^T G_p(p) \tag{6.41b}$$
$$G(p) = 0 \tag{6.41c}$$

where p are the positions, q are the velocities, $M(p)$ is the mass matrix, $G(p)$ describes the constraints, and λ are the associated Lagrange multipliers. If the constraints are such that $G_p(p)$ is nonsingular, that is, has full row rank, and $\mathbb{L}_\mu \neq 0$ it can be shown that (6.41) satisfies Hypothesis 2.2 with $\mu = 2$, $d = 2(n_p - n_\lambda)$, $a = 3n_\lambda$ and $\nu = 0$. Divide p into (p_1, p_2) such that $G(p_1, p_2) = 0$ can be solved for p_2, and q accordingly. Then reduced model (2.34) becomes

$$\dot{p}_1 = q_1 \tag{6.42a}$$
$$\dot{q}_1 = \begin{pmatrix} I & 0 \end{pmatrix} M(p)^{-1} \big(F(p, q, u) + G_p(p)^T \lambda \big) \tag{6.42b}$$
$$0 = G(p) \tag{6.42c}$$
$$0 = G_p(p)q \tag{6.42d}$$
$$0 = G_{pp}(p)(q, q)G_p(p)M(p)^{-1} \big(F(p, q, u) + G_p(p)^T \lambda \big) \tag{6.42e}$$

It can be seen that this model satisfies the condition in Section 6.2.3 given that

$$F(p, q, u) = \bar{F}(p, q) + \bar{G}(p, q)u$$

For mechanical systems, it can also be seen that the constraint equations has a lot of structure. The first constraint only involves p, the next only involves q and p, while the last one involves all variables. The method in Section 6.4, then suggests that in first step,

the first equation can be solved for p_2 in terms of p_1. The second step is to solve the second algebraic equation for q_2 expressed in p_1 and q_1. And the last step is to solve the last equation for λ in terms of p_1, q_1 and u. In this case, one does not gain by solving the equations in terms of the variables in each equation as described in Section 6.4 instead of the variables x_1 and u, since x_1 and u may be present in all equations.

6.6 Example

In this section, a small example is studied only to give a hint about the differences in time required to compute an approximate solution. The example is obtained from Kunkel and Mehrmann (2001) and describes a cart-pendulum system in DAE form as

$$\begin{aligned}
\dot{p}_1 &= q_1 \\
\dot{p}_2 &= q_2 \\
\dot{p}_3 &= q_3 \\
\dot{q}_1 &= -2\lambda(p_1 - p_3) \\
\dot{q}_2 &= -2\lambda p_2 - g \\
\dot{q}_3 &= 2\lambda(p_1 - p_3) + u \\
0 &= (p_1 - p_3)^2 + p_2^2 - l^2
\end{aligned}$$

If the model above is rewritten in the form (6.41) using index reduction, it is possible to find an optimal control law. For this case, the strangeness-free model becomes

$$\begin{aligned}
\dot{p}_1 &= q_1 \\
\dot{p}_3 &= q_3 \\
\dot{q}_1 &= -2\lambda(p_1 - p_3) \\
\dot{q}_3 &= 2\lambda(p_1 - p_3) + u \\
0 &= (p_1 - p_3)^2 + p_2^2 - l^2 \\
0 &= (p_1 - p_3)(q_1 - q_3) + p_2 q_2 \\
0 &= (q_1 - q_3)^2 + q_2^2 - 2\lambda\big(2(p_1 - p_3)^2 + p_2^2\big) - p_2 g - (p_1 - p_3)u
\end{aligned}$$

The cost function L is chosen as

$$L(x_1, u) = p_1^2 + p_3^2 + q_1^2 + q_3^2 + u^2$$

The algorithm to find the optimal solution is implemented in Maple 11 on a PC running Linux. The computation times for four different setups are presented. The first and second column show the computation times for $\mathcal{R}$ up to orders 8 and 12, respectively. In the first column, the structure has been considered while in column 2 it has not. The third and fourth columns present the times required to compute V up to the given orders, with and without computing the equation for u_*.

Table 6.1: *The solution times for the computation of the optimal solution for the cart-pendulum system, when the structure of the DAE model has been considered to different extents.*

Order	$\mathcal{R}$ w/o struct.	$\mathcal{R}$ w/ struct.	V w/o control affinity	V w/ control affinity
8	11.5	3.5	3.4	2.1
12	220	49	28	13

6.7 Conclusions

In this chapter different methods have been presented of how the structure of a nonlinear DAE model can be utilized to reduce the computational complexity for certain control optimal control problems. One method was to use that for models in control-affine form and if the cost function is quadratic in the control signal, one of the equations corresponding to the HJB can be solved explicitly. To test for control affinity, some conditions were derived. It was also shown that a control-affine DAE model must have an equivalent model in which $\dot{x}_1$ and u enters in a specific way. Another presented method to reduce the complexity rather much is to first transform the model to the BLT form before solving the problem.

7

Well-Posedness of SDAE Models

When modeling physical systems, it is usually impossible to predict the exact behavior of the system. This can have several explanations. One common situation is that it is known that external signals affect the systems, but these signals neither can be measured nor chosen. A common choice is then to model them as stochastic processes.

Another common situation is that certain signals in the system are measured, but there are imperfections in the measurements. For example, a sensor may have an unknown offset or produce measurements with a time-varying error. This is denoted measurement noise and can also be modeled as a stochastic process.

A third possibility is that a model has imperfections, which cannot be classified as unmeasured external signals or measurement imperfections. This is denoted process noise, and is normally modeled as a stochastic process. Hence, above three cases are presented in which it might be appropriate to include stochastic processes when modeling a physical system.

This chapter deals with the problem of how to deal with stochastic processes for DAE models, that is, the objective is to incorporate process noise $w(t)$ and measurement noise $e(t_k)$ in such a model. In the general case, this would result in the stochastic DAE (SDAE) model

$$\begin{aligned} F\big(\dot{x}(t), x(t), w(t), u(t), t\big) &= 0 \\ y(t_k) &= h\big(x(t_k)\big) + e(t_k) \end{aligned}$$

where $u \in \mathbb{R}^p$ is the control input and $y \in \mathbb{R}^q$ is a measured output.

The conditions will be very similar to the control-affinity conditions posed in Chapter 6. Having the conditions for well-posedness, it will also be discussed how the variables x can be estimated using particle filters (Gordon et al., 1993; Doucet et al., 2001; Ristic et al., 2004).

7.1 Literature Overview

The question whether the state estimation problem for DAE models is well-defined has been discussed by, e.g., Schein and Denk (1998), Winkler (2004), Darouach et al. (1997), Kučera (1986), Germani et al. (2002), and Becerra et al. (2001). In Schein and Denk (1998), linear SDAE models are treated, and it is guaranteed that the noise is not differentiated by assuming that the system has differential index 1 (see for example Chapter 2). The assumption that the system has differential index 1 is more restrictive than necessary, and rules out some applications such as many mechanics systems. This assumption will not be made here. Schein and Denk (1998) also note that some internal variables actually may be so-called generalized stochastic processes, that is a time-continuous white noise process. Winkler (2004) makes the same assumption as Schein and Denk (1998), but also treats a class of nonlinear DAE models.

Darouach et al. (1997) deals with linear DAE models with differential index 1, and a Kalman filter is constructed. However, in the estimation procedure the authors seem to overlook the fact that some variables may have infinite variance. In Kučera (1986), the original linear SDAE system specification may actually specify derivatives of white noise, but a controller is designed that removes any derivatives. In Germani et al. (2002) restrictive assumptions are made that guarantee that no derivatives appear in the linear SDAE, although this is not stated explicitly. Finally, in Becerra et al. (2001) nonlinear semi-explicit DAE models (e.g., Brenan et al., 1996) are discussed. Here well-posedness is guaranteed by only adding noise to the state-space part of the system.

7.2 Background and Motivation

As mentioned above, the considered class of system can be written as

$$F\big(\dot{x}(t), x(t), w(t), u(t), t\big) = 0 \tag{7.2a}$$

$$y(t_k) = h\big(x(t_k)\big) + e(t_k) \tag{7.2b}$$

where w is process noise and e is measurement noise. The discussion will only include the case when $w(t)$ is a Gaussian second order stationary process with spectrum $\phi_w(\omega)$. The spectrum is assumed to be rational in ω with pole excess $2p_w$, which means that (Gerdin, 2006; Åström, 1970)

$$\lim_{\omega\to\infty} \omega^{2p_w}\phi_w(\omega) = C$$
$$0 < C < \infty$$

An important property of DAE models is that the variables $x(t)$ may depend on derivatives of the inputs to the model as was seen in for example Section 2.4.4. This is one of the main issues when discussing noise for DAE models. Since $w(t)$ occurs as an external signal in the DAE equations (7.2), one or more of its derivatives with respect to time may affect the variables $x(t)$. This is an issue, since time derivatives of a Gaussian second order stationary process may not have finite variance. Actually, $w(t)$ can be differentiated at most $p_w - 1$ times, since it has pole excess $2p_w$, see Gerdin (2006).

Example 7.1: Noise modeling difficulties

Consider the DAE

$$\begin{pmatrix} \dot{x}_1(t) - x_2(t) \\ \dot{x}_2(t) - x_2(t) \\ x_1^2(t) + x_3^2(t) - 1 - w(t) \end{pmatrix} = 0$$

where a stochastic process has been added to the last equation to model an unmeasured disturbance. Differentiating the last equation w.r.t. time gives

$$2x_1(t)\dot{x}_1(t) + 2x_3(t)\dot{x}_3(t) - \dot{w}(t) = 0$$

Eliminating $\dot{x}_1(t)$ and $\dot{x}_3(t)$ using the first two equations of the DAE and solving for $x_2(t)$ gives

$$x_2(t) = \frac{\dot{w}(t)}{2x_1(t) + 2x_3(t)}$$

If the spectrum of $w(t)$ has pole excess 2, this is questionable since $\dot{w}(t)$ then has infinite variance. However, if the pole excess is 3 or higher, the involved signals have finite variance.

As seen in the example above, it is essential to examine how many derivatives of $w(t)$ that affect the variables. For this end, the method in Section 2.4.4 is used. Therefore, consider the DAE model

$$F\big(\dot{x}(t), x(t), w(t), u(t), t\big) = 0$$

and let it satisfy Hypothesis 2.2 for μ and $\mu + 1$ with the same d, a and v. Then, it is known that there exist matrices Z_1 and Z_2 such that

$$\hat{F}_1(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3, u, w, t) = Z_1^T F \tag{7.3a}$$

$$\hat{F}_2(x_1, x_2, x_3, u, \dot{u}, \dots, u^{(\mu)}, w, \dot{w}, \dots, w^{(\mu)}, t) = Z_2^T \begin{pmatrix} F \\ \frac{d}{dt}F \\ \vdots \\ \frac{d^\mu}{dt^\mu}F \end{pmatrix} \tag{7.3b}$$

where the derivatives are considered formal. From Section 2.4.4, it is known that $\hat{F}_2 = 0$ can be solved for x_3, and after using that equation to eliminate $\dot{x}_3$ and x_3 in $\hat{F}_1$, the equation $\hat{F}_1 = 0$ can be solved for $\dot{x}_1$.

Now, let $w(t)$ be a stochastic process which has a spectrum with pole excess $2p_w$. Then, it can be differentiated at most $p_w - 1$ times. If it is differentiated p_w times or more, the resulting signal has infinite variance. This means that a sufficient condition for the signals x in the DAE to have finite variance is that no derivatives of w higher than $p_w - 1$ occur in $\hat{F}_2$ in (7.3b). As before, only DAE models for which all variables are determined by the model, except for u and w, are considered. This means that no x_2 appear. This discussion leads to the following result.

Lemma 7.1
Consider the SDAE model

$$F\big(\dot{x}(t), x(t), u(t), w(t), t\big) = 0$$

where $w(t)$ is a Gaussian second order stationary process with spectrum $\phi_w(\omega)$, which is rational in ω with pole excess $2p_w$. Assume that the SDAE fulfills Hypothesis 2.2 with both u and w seen as external signals. The signals $x(t)$ then have finite variance provided that $\hat{F}_2$ can be written as

$$\hat{F}_2 = \bar{F}_2(x_1, x_3, u, \dot{u}, \ldots, u^{(k)}, w, \dot{w}, \ldots, w^{(l)}, t)$$

where $l \leq p_w - 1$, $\bar{F}_2$ is $\hat{F}_2$ with the special structure and $\hat{F}_2$ is defined by (7.3b)*.*

The above discussion shows how it can be examined if a noise process $w(t)$ is differentiated too many times so that the resulting equations include signals with infinite variance. However, it would be nice to be able to discuss solutions to stochastic DAE models in terms of stochastic differential equations. The approach in this work will be to convert the SDAE to the state-space form

$$\dot{x}(t) = f\big(x(t), t\big) + \sigma\big(x(t), t\big)v(t) \tag{7.4}$$

where the noise enters affinely into the equations. Of course, this is a special case of a more general model structure where the noise enters through a general nonlinear function. However, the general case is less treated in the literature. Since our goal is to extend existing results for state-space models to DAE models, the discussion is limited to the special case (7.4). Note that (7.4) must be handled as a stochastic integral (Åström, 1970). To point this out, (7.4) is often written as

$$dx = f\big(x(t), t\big)dt + \sigma\big(x(t), t\big)\, dv$$

where $v(t)$ is a Wiener process.

The model (7.4) requires the noise process to be white noise, but in this chapter so far only noise $w(t)$ with finite variance has been discussed. However, as $w(t)$ is assumed to be a Gaussian second order stationary process, it can be seen as white noise filtered through a linear filter, see Åström (1970); Gerdin (2006). The filter can for example be written in state-space form,

$$\dot{x}_w(t) = Ax_w(t) + Bv(t) \tag{7.5a}$$
$$w(t) = Cx_w(t) \tag{7.5b}$$

where $v(t)$ is white noise. Combining the SDAE model (7.2a) and the noise model (7.5) gives

$$\begin{aligned} F\big(\dot{x}(t), x(t), Cx_w(t), u(t)\big) &= 0 \\ \dot{x}_w(t) &= Ax_w(t) + Bv(t) \end{aligned}$$

This can be seen as the single SDAE model,

$$G\big(\dot{z}(t), z(t), v(t), u(t)\big) = 0$$

where $v(t)$ is white noise and

$$z(t) = \begin{pmatrix} x(t) \\ x_w(t) \end{pmatrix}$$

When the SDAE model contains white noise terms, additional restrictions apply. Not only is it not allowed to differentiate the white noise signal, but it must also be integrated in the affine form (7.4). As shown in the following example, this is not ensured for a general SDAE model.

Example 7.2: White noise modeling difficulties

Consider the nonlinear DAE

$$\begin{aligned} \dot{x}_1(t) - x_2^2(t) &= 0 \\ x_2(t) - v(t) &= 0 \end{aligned}$$

where $v(t)$ is white noise. The second equation states that $x_2(t)$ is equal to a time-continuous white noise process. Since such processes have infinite variance, this is questionable if $x_2(t)$ represents a physical quantity. The first equation states that

$$\dot{x}_1(t) = v^2(t)$$

which also is questionable since nonlinear operations on white noise cannot be handled in the standard framework of stochastic integrals (Åström, 1970; Gerdin, 2006).

The main topics of this chapter concern how noise can be included in DAE models without introducing problems such as those discussed in the example and how particle filters can be implemented for DAE models with white noise inputs.

7.3 Well-Posedness for Linear SDAE Models

First, the linear case is considered, which has been studied in Gerdin (2006). Let the linear SDAE be given by

$$E\dot{x}(t) = Ax(t) + Bu(t) + \sum_{l=1}^{n_w} J_l w_l(t) \tag{7.7a}$$

$$y(t_k) = Cx(t_k) + Du(t_k) + e(t_k) \tag{7.7b}$$

where $w_l(t)$ is a Gaussian second order stationary process with a spectrum ϕ_{w_l} which is rational and has pole excess $2p_l$. Note that if $w_l(t)$ is a white noise, p_l is zero.

In the linear case, it is possible to require only a part of the variables to have finite variance. This is interesting in cases where some variables are used as internal variables in the model and may have no interest in themselves. The variables that need to have finite variance are denoted $\bar{x}(t)$ and are selected as

$$\bar{x}(t) = Mx(t)$$

for some rectangular matrix M.

In order to formulate the result, first recall the definition of an oblique projection of a matrix A along the space B on the space C,

$$A/_B\, C = \begin{pmatrix} 0 & \bar{C} \end{pmatrix} \begin{pmatrix} \bar{B} & \bar{C} \end{pmatrix}^{-1} A$$

where $\bar{B}$ and $\bar{C}$ are bases for B and C, respectively. The result can now be formulated as follows (Gerdin, 2006).

Theorem 7.1
Consider the linear SDAE model (7.7) and assume that it is regular. Let λ be a scalar such that $(\lambda E + A)$ is invertible, and define

$$\bar{E} = (\lambda E + A)^{-1}$$

Then the variables $\bar{x}(t)$ and the measured output $y(t_k)$ will have finite variances if and only if

$$\left(\bar{E}^j(\lambda E + A)^{-1}J_l\right)\big/_{\mathcal{V}(\bar{E}^n)}\mathcal{N}(\bar{E}^n) \in \mathcal{N}\left(\begin{pmatrix} M \\ H \end{pmatrix}\right),\ j \geq p_l,\ \forall l$$

where $\mathcal{N}$ denotes the null space and $\mathcal{V}$ the range.

In the nonlinear case, all variables will be required to have finite variance, which in the linear case would correspond to the case with M chosen as an identity matrix of size $d \times d$. This means that the corresponding null space is given by the zero matrix.

7.4 Well-Posedness for Nonlinear SDAE Models

Now, consider the nonlinear case. In this case, a well-posed SDAE model is required to have the following properties.

Definition 7.1. The SDAE model (7.2) is well-posed if the underlying state-space model has a solution that is well-defined in the standard framework of stochastic integrals, and if all variables have finite variance.

It has been motivated in Examples 7.1 and 7.2 that the nonlinear model must not differentiate the noise input. Furthermore, the noise input must enter the model affinely. The following theorem gives sufficient conditions to satisfy these requirements. To simplify the notation, $u(t)$ is included into t.

Theorem 7.2
Consider the nonlinear SDAE model

$$F\big(\dot{x}(t), x(t), v(t), t\big) = 0 \tag{7.8}$$

and assume that it satisfies Theorem 2.2 with Hypothesis 2.2 with $v(t)$ as the external input. Let $\hat{F}_1$, $\hat{F}_2$, x_1, x_2, and x_3 be defined as in Section 2 and assume that x_2 is of size zero.

Then there exists a well-defined solution $x(t)$ in terms of stochastic differential equations to (7.8) with $v(t)$ considered as white noise provided that $\hat{F}_1$ and $\hat{F}_2$ can be written as

$$\hat{F}_1 = \bar{F}_1\big(t, x_1, x_3, \dot{x}_1 - \sigma(x_1, x_3)v, \dot{x}_3 + \hat{F}_{2;x_3}^{-1}\hat{F}_{2;x_1}\sigma(x_1, x_3)v\big) \tag{7.9a}$$

$$\hat{F}_2 = \bar{F}_2\big(t, x_1, x_3\big) \tag{7.9b}$$

for some function $\sigma(x_1, x_3)$.

Proof: Differentiating (7.9b) w.r.t. time yields

$$\bar{F}_{2;t} + \bar{F}_{2;x_1}\dot{x}_1 + \bar{F}_{2;x_3}\dot{x}_3 = 0$$

From the assumptions in Theorem 2.2 it follows that $F_{2;x_3}$ is invertible, and that $\bar{F}_2$ is locally solvable for x_3. This means that $\dot{x}_3$ can be written as

$$\dot{x}_3 = -\bar{F}_{2;x_3}^{-1}(\bar{F}_{2;t} + \bar{F}_{2;x_1}\dot{x}_1)$$

The constraints (7.9b) can also be locally solved for x_3 to give

$$x_3 = \mathcal{R}(t, x_1) \tag{7.10}$$

for some function $\mathcal{R}$. Inserting this into (7.9a) gives

$$\bar{F}_1\big(t, x_1, \mathcal{R}, \dot{x}_1 - \sigma(x_1, \mathcal{R})v, -\bar{F}_{2;x_3}^{-1}(\bar{F}_{2;t} + \bar{F}_{2;x_1}\dot{x}_1) + \bar{F}_{2;x_3}^{-1}\bar{F}_{2;x_1}\sigma(x_1, \mathcal{R})v\big)$$

The equation $\bar{F}_1 = 0$ now takes the form

$$\bar{F}_1\big(t, x_1, \mathcal{R}, \dot{x}_1 - \sigma(x_1, \mathcal{R})v, -\bar{F}_{2;x_3}^{-1}\bar{F}_{2;t} - \bar{F}_{2;x_3}^{-1}\bar{F}_{2;x_1}(\dot{x}_1 - \sigma(x_1, \mathcal{R})v)\big) = 0$$

Since Theorem 2.2 is fulfilled, this equation can be solved for $\dot{x}_1$. Since $-\sigma(x_1, \mathcal{R})v$ enters the equations in the same way as $\dot{x}_1$, the solution can be written as

$$\dot{x}_1 - \sigma(x_1, \mathcal{R})v = \mathcal{L}(t, x_1)$$

for some function $\mathcal{L}$. This can be interpreted as the stochastic differential equation

$$dx_1 = \mathcal{L}(t, x_1)dt + \sigma(x_1, \mathcal{R})\, dv$$

which means that x_1 has a well-defined solution. A solution for x_3 is then defined through (7.10). □

Note that in the derivation of the reduced DAE (and the underlying ODE), derivatives of the noise may show up. However, as shown in Section 2.4.4, such derivatives can be neglected if they disappear in the final expressions. Further note that (7.9) can also be shown to satisfy the conditions in Theorem 6.3 with $M(x_1, x_3) = \sigma(x_1, x_3)$.

If noise has been added to a DAE model using physical insight or for other reasons in a predefined way, the theorem above gives conditions for the system to be well-posed using a transformed version of the system. It may also be interesting to be able to see if the SDAE is well-posed already in the original equations. As discussed in the theorem above, the SDAE model is well-posed if the equations $\hat{F}_1 = 0$ and $\hat{F}_2 = 0$ take the form

$$\begin{aligned}\bar{F}_1\big(t, x_1, x_3, \dot{x}_1 - \sigma(x_1, x_3)v, \dot{x}_3 + \bar{F}_{2;x_3}^{-1}\bar{F}_{2;x_1}\sigma(x_1, x_3)v\big) &= 0\\ \bar{F}_2\big(t, x_1, x_3\big) &= 0.\end{aligned}$$

In the original equations, this can typically be seen as adding noise according to

$$F\left(\begin{pmatrix}\dot{x}_1 - \sigma(x_1, x_3)v\\ \dot{x}_3 + \hat{F}_{2;x_3}^{-1}\hat{F}_{2;x_1}\sigma(x_1, x_3)v\end{pmatrix}, \begin{pmatrix}x_1\\ x_3\end{pmatrix}, t\right) = 0. \tag{7.12}$$

One common situation when it is easy to see how white noise can be added is for semi-explicit index one DAE models (Brenan et al., 1996). This is considered in the following example.

Example 7.3: Noise modeling: semi-explicit index 1 DAE

Consider a semi-explicit index one DAE model

$$\dot{x}_a = f(x_a, x_b) \tag{7.13a}$$
$$0 = g(x_a, x_b) \tag{7.13b}$$

Locally, x_b can be solved from (7.13b), so these equations correspond to $\hat{F}_1 = 0$ and $\hat{F}_2 = 0$ respectively. Noise can thus be added according to

$$\begin{aligned} \dot{x}_a &= f(x_a, x_b) + \sigma(x_a, x_b)v \\ 0 &= g(x_a, x_b) \end{aligned}$$

7.5 Particle Filtering

An important aspect of uncertain models is state estimation and prediction. For nonlinear systems this is a difficult problem, see for instance Ristic et al. (2004); Andrieu et al. (2004); Schön (2006). Therefore, it is necessary to resort to approximate methods. One approximate method for nonlinear state estimation is the particle filter, see for example Gordon et al. (1993); Doucet et al. (2001); Ristic et al. (2004). In this section, the problem of how particle filter methods can be extended for use with SDAE models will be discussed.

To be able to describe how existing particle filtering algorithms can be extended to DAE models, first a brief summary of how particle filtering can be implemented for state-space models is presented. For a more thorough treatment, see *e.g.*, Gordon et al. (1993); Doucet et al. (2001); Ristic et al. (2004). Existing particle filtering methods may allow other model structures than state-space models, but here the discussion is limited to this class since it is enough in order to extend particle filtering methods to SDAE models.

Consider a nonlinear discrete-time state-space model,

$$x(t_{k+1}) = f\big(x(t_k), u(t_k), w(t_k)\big) \tag{7.15a}$$
$$y(t_k) = h\big(x(t_k)\big) + e(t_k) \tag{7.15b}$$

where x is the state vector, u is a known input, y is a measured output, and w and e are stochastic processes with known probability density functions. The particle filter is based on estimating the probability density function of the state $x(t_k)$, given the measurements

$$Z^N = \{u(t_0), y(t_0), ..., u(t_N), y(t_N)\}. \tag{7.16}$$

The goal is therefore to compute the probability density function

$$p\big(x(t_k)|Z^N\big) \tag{7.17}$$

Depending on if $k < N$, $k = N$, or $k > N$ the type of the problem is either a smoothing problem, a filtering problem, or a prediction problem, respectively. In this thesis only the filtering problem and the one-step-ahead prediction problem are considered, which means that $N = k$ or $N = k - 1$.

Once (the estimate of) the probability density function has been computed, it can be used to estimate the value of $x(t)$. One possibility is to use the expected value of $x(t_k)$ given Z^N, another is to use the maximum a posteriori estimate, that is the $x(t_k)$ that maximizes $p\big(x(t_k)|Z^N\big)$.

In the particle filter, the probability density function (7.17), here with $N = k - 1$, is approximated by a sum of Dirac functions,

$$p\big(x(t_k)|Z^{k-1}\big) \approx \sum_{i=1}^{M} q^{(i)}_{t_k|t_{k-1}} \delta\big(x(t_k) - x^{(i)}_{t_k|t_{k-1}}\big)$$

This means that the density function is approximated using M particles

$$\big\{x^{(i)}_{t_k|t_{k-1}}\big\}_{i=1}^{M}$$

with associated weights,

$$\big\{q^{(i)}_{t_k|t_{k-1}}\big\}_{i=1}^{M}.$$

Since the approximation is made using Dirac functions, it is not an approximation at each point x. Instead, the approximation is valid for integrals of p. For example, the mean value of $x(t_k)$ can be estimated as

$$\mathrm{E}\left(x(t_k)|Z^{k-1}\right) = \int x \cdot p\big(x(t_k)|Z^{k-1}\big)dx \approx \sum_{i=1}^{M} q^{(i)}_{t_k|t_{k-1}} x^{(i)}_{t_k|t_{k-1}}$$

Now assume that a new measurement $\{y(t_k), u(t_k)\}$ is obtained. Using Bayes's rule, the probability density function $p\big(x(t_k)|Z^{k-1}\big)$ should be updated according to

$$p\big(x(t_k)|Z^{k}\big) = \frac{p\big(y(t_k)|x(t_k)\big)\, p\big(x(t_k)|Z^{k-1}\big)}{p\big(y(t_k)|Z^{k-1}\big)}.$$

Since $p\big(y(t_k)|Z^{k-1}\big)$ does not depend on x, the approximation of the probability density function is updated by the particle filter by updating the weights $\{q^{(i)}_{t_k|t_{k-1}}\}_{i=1}^{M}$ as

$$q^{(i)}_{t_k|t_k} = \frac{p\big(y(t_k)|x^{(i)}_{t_k|t_{k-1}}\big)q^{(i)}_{t_k|t_{k-1}}}{\sum_{j=1}^{M} p\big(y(t_k)|x^{(j)}_{t_k|t_{k-1}}\big)q^{(j)}_{t_k|t_{k-1}}}, \quad i = 1, \ldots, M.$$

For the state-space model (7.15), it follows that

$$p\left(y(t_k)\,\middle|\,x^{(i)}_{t_k|t_{k-1}}\right) = p_e\left(y(t_k) - h(x^{(i)}_{t_k|t_{k-1}})\right)$$

where p_e is the probability density function of $e(t_k)$.

After this step, called the measurement update, the resampling step takes place. The resampling step redistributes the particles to avoid degeneration of the filter. It does not introduce additional information (actually, information is lost). The method used here is sampling importance resampling. For other alternatives, see the Gordon et al. (1993); Doucet et al. (2001); Ristic et al. (2004). In the resampling step the M particles are replaced by M new particles. This is performed by drawing M particles with replacement from the old particles. The probability to draw particle i is proportional to its weight $q^{(i)}_{t_k|t_k}$. The new particles $x^{(i)}_{t_k|t_k}$ are thus chosen according to

$$\Pr\left(x^{(i)}_{t_k|t_k} = x^{(j)}_{t_k|t_{k-1}}\right) = q^{(j)}_{t_k|t_k} \quad i = 1, \ldots, M.$$

The weights are changed to

$$q^{(i)}_{t_k|t_k} = \frac{1}{M} \quad i = 1, \ldots, M$$

so that the approximation of the probability density function is, approximately, left unchanged.

After the resampling step, the time update step takes place. This means that $x(t_{k+1})$ is predicted using available information about $x(t_k)$. For the particle filter and the state-space model (7.15), this is done by drawing M independent samples $w^{(i)}(t_k)$, $i = 1, \ldots, M$, of $w(t_k)$, according to its probability density function p_w. The particles are then updated according to

$$x^{(i)}_{t_{k+1}|t_k} = f\big(x^{(i)}_{t_k|t_k}, u(t_k), w^{(i)}(t_k)\big), \quad i = 1, \ldots, M.$$

In general, this can be seen as drawing new particles according to their conditional distribution,

$$x^{(i)}_{t_{k+1}|t_k} \sim p\left(x_{t_{k+1}|t_k} \,\middle|\, x^{(i)}_{t_k|t_k}\right), \quad i = 1, \ldots, M.$$

The weights are unchanged, $q^{(i)}_{t_{k+1}|t_k} = q^{(i)}_{t_k|t_k} = \frac{1}{M}$. Note that a more general version of the time update equation is available, see the references. After this step, a new measurement is obtained and the filter is restarted from the measurement update step.

When starting a filter, the particles should be initialized according to available information about the initial value, $x(t_0)$. If the probability density function of $x(t_0)$ is p_{x_0}, the particles are initially chosen according to that distribution. We can write this as

$$x^{(i)}_{t_0|t_{-1}} \sim p_{x_0}(x_0), \quad i = 1, \ldots, M$$

and we get

$$q^{(i)}_{t_0|t_{-1}} = \frac{1}{M}, \quad i = 1, \ldots, M.$$

To conclude, the particle filter algorithm can be written as follows.

1. Initialize the M particles,

$$x^{(i)}_{t_0|t_{-1}} \sim p_{x_0}(x_0), \quad i = 1, \ldots, M$$

and

$$q_{t_0|t_{-1}}^{(i)} = \frac{1}{M}, \quad i = 1, \ldots, M.$$

Set $k := 0$.

2. Measurement update: calculate weights $\{q_{t_k|t_k}^{(i)}\}_{i=1}^{M}$ according to

$$q_{t_k|t_k}^{(i)} = \frac{p\big(y(t_k)|x_{t_k|t_{k-1}}^{(i)}\big)q_{t_k|t_{k-1}}^{(i)}}{\sum_{j=1}^{M} p\big(y(t_k)|x_{t_k|t_{k-1}}^{(j)}\big)q_{t_k|t_{k-1}}^{(j)}}, \quad i = 1, \ldots, M.$$

3. Resampling: draw M particles, with replacement, according to

$$\Pr\left(x_{t_k|t_k}^{(i)} = x_{t_k|t_{k-1}}^{(j)}\right) = q_{t_k|t_k}^{(j)} \quad i = 1, \ldots, M$$

and set

$$q_{t_{k+1}|t_k}^{(i)} = \frac{1}{M} \quad i = 1, \ldots, M.$$

4. Time update: predict new particles according to

$$x_{t_{k+1}|t_k}^{(i)} \sim p\left(x_{t_{k+1}|t_k} \,\middle|\, x_{t_k|t_k}^{(i)}\right), \quad i = 1, \ldots, M.$$

5. Set $k := k + 1$ and iterate from step 2.

To examine how the implementation for DAE models can be done, consider a SDAE model in the form (7.2),

$$\begin{aligned} G\big(\dot{z}(t), z(t), w(t), t\big) &= 0 \\ y(t_k) &= h\big(z(t_k)\big) + e(t_k) \end{aligned}$$

In order to use the methods for stochastic simulation with white noise inputs, the stochastic process $w(t)$ is realized as white noise $v(t)$ filtered through a linear filter as discussed in Section 7.2, and the following model is obtained

$$F\big(\dot{x}(t), x(t), v(t), t\big) = 0 \tag{7.19a}$$

$$y(t_k) = h\big(x(t_k)\big) + e(t_k) \tag{7.19b}$$

Now, consider a SDAE model (7.19a) that fulfills the conditions in Theorem 7.2. Then, it follows that the model has the structure

$$\bar{F}_1\big(t, u, x_1, x_3, \dot{x}_1 - \sigma(x_1, x_3)v, \dot{x}_3 + \bar{F}_{2;x_3}^{-1}\bar{F}_{2;x_1}\sigma(x_1, x_3)v\big) = 0 \tag{7.20a}$$

$$\bar{F}_2\big(t, u, \dot{u}, \ldots, u^{(\mu)}, x_1, x_3\big) = 0 \tag{7.20b}$$

$$y(t_k) = h\big(x(t_k)\big) + e(t_k) \tag{7.20c}$$

Since $\bar{F}_1$ and $\bar{F}_2$ are the result of the transformations discussed in Section 2.4, $\bar{F}_2$ can be solved locally for x_3 as

$$x_3 = \mathcal{R}(t, x_1, u, \dot{u}, \ldots, u^{(\mu)}) \tag{7.21}$$

After using (7.21) to eliminate x_3 and $\dot{x}_3$ in $\bar{F}_1$, $\bar{F}_1$ can be solved for $\dot{x}_1$ as

$$\dot{x}_1 = \mathcal{L}(t, u, \dot{u}, \ldots, u^{(\mu+1)}, x_1) + \sigma(x_1, \mathcal{R})v \tag{7.22}$$

Combining (7.20)–(7.22) gives

$$\dot{x}_1 = \mathcal{L}(t, u, \dot{u}, \ldots, u^{\mu+1}, x_1) + \sigma(x_1, \mathcal{R})v \tag{7.23a}$$

$$y(t_k) = h\big(x_1(t_k), \mathcal{R}(u(t_k), \dot{u}(t_k), \ldots, u^{(\mu)}x_1(t_k)\big) + e(t_k) \tag{7.23b}$$

The state-space model (7.23) can be used to implement a particle filter for estimation of x_1. After estimating x_1, estimates of x_3 can be computed using (7.21).

Since it is usually not possible to solve for $\dot{x}_1$ and x_3 explicitly, numerical implementation methods will be discussed in the following section. Furthermore, the state equation need to be discretized. This can be done using for example a numerical solver for stochastic differential equations. The time update in step 4 in the particle filtering algorithm is thus performed by solving (7.23a) for one time step. The measurement update in step 2 of the particle filtering algorithm is performed using the measurement equation (7.23b).

7.6 Implementation Issues

As mentioned in earlier sections, the transformation which brings the model into the form (7.20) may be difficult to compute in practice. Furthermore, in order to run the particle filter, the equations (7.20) need to be solved numerically for $\dot{x}_1$ and x_3 which may be an issue. Therefore, approximate implementations can be considered. One approach to do this is to use the type of DAE solver that is included in modeling environments for object-oriented modeling such as Dymola (Mattsson et al., 1998).

As discussed in Section 2.5, DAE solvers that solve models obtained from object-oriented modeling tools compute an approximation in the form

$$\begin{aligned} \check{F}_1(t, x_1, x_3, \dot{x}_1) &= 0 \\ \hat{F}_2(t, x_1, x_3) &= 0 \end{aligned}$$

where $\check{F}_1$ is $\hat{F}_1$ with $\dot{x}_3$ eliminated. This model can be used to examine if a DAE with a noise model satisfies the conditions of Theorem 7.2. The most straightforward way to check if a given noise model is correct, is to examine if the transformed system is of the form

$$\check{F}_1 = \tilde{F}_1\big(t, x_1, x_3, \dot{x}_1 - \sigma(x_1, x_3)v\big) = 0 \tag{7.25a}$$

$$\hat{F}_2(t, x_1, x_3) = 0 \tag{7.25b}$$

where $\tilde{F}_1$ is $\check{F}_1$ with the desired structure of the arguments. The check can either be done by just looking at the equations or by using the methods in Chapter 6. If v appears in incorrect positions (so that the transformed system is not of the form (7.25)), one way to handle the situation would be to remove $v(t)$ from these incorrect locations in $\tilde{F}_1$ and $\hat{F}_2$, and assume that noise is added to the original equations so that this is achieved.

The solvers can also be used for approximate implementation of particle filters for DAE models. The idea behind this is that the transformation to the form

$$\dot{x}_1 = \mathcal{L}(t, x_1) + \sigma(x_1, \mathcal{R})v \tag{7.26a}$$
$$x_3 = \mathcal{R}(t, x_1) \tag{7.26b}$$

can be made by solving $\tilde{F}_1$ and $\hat{F}_2$ numerically at each time step using a DAE solver. This means that given values of x_1 and v the solver can give $\dot{x}_1$ and x_3. The state equation (7.26a) can then be used to estimate x_1, and x_3 can be computed from (7.26b).

To summarize, the following procedure can be used when modeling noise in DAE models and implementing a particle filter. First, a DAE model without noise is derived by writing down equations, or from component-based modeling. This DAE model is then entered into a DAE solver to determine which variables are states. Using physical insight, noise is then added to the original equations and the equations are transformed into $\tilde{F}_1$ and $\hat{F}_2$. If noise terms appear at incorrect positions, these are removed so that the equations are in the form (7.25). For this form a particle filter is then implemented by solving for $\dot{x}_1$ and x_3 using the DAE solver.

7.7 Example: Dymola Assisted Modeling and Particle Filtering

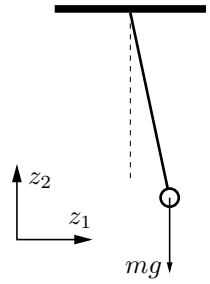


Figure 7.1: *A pendulum.*

In this section a DAE model of a pendulum is studied. First noise is added, and then a particle filter is implemented to estimate the internal variables of the pendulum. The example is slightly modified example from Brenan et al. (1996). As shown in Figure 7.1, z_1 and z_2 are the horizontal and vertical position of the pendulum. Furthermore, z_3 and z_4 are the respective velocities, z_5 is the tension in the pendulum, the constant b represents resistance caused by the air, g is the gravity constant, and L is the constant length of the

pendulum. The equations describing the pendulum are

$$\dot{z}_1 = z_3 \tag{7.27a}$$
$$\dot{z}_2 = z_4 \tag{7.27b}$$
$$\dot{z}_3 = -z_5 \cdot z_1 - b \cdot z_3^2 \tag{7.27c}$$
$$\dot{z}_4 = -z_5 \cdot z_2 - b \cdot z_4^2 - g \tag{7.27d}$$
$$0 = z_1^2 + z_2^2 - L^2 \tag{7.27e}$$

We will use the approximate methods discussed in Section 7.6, so the equations are entered into the DAE solver in Dymola. The first step in the noise modeling is to let Dymola select which variables are states. There are several possible ways to select states for these equations, but here z_1 and z_3 are selected, which gives

$$x_1 = \begin{pmatrix} z_1 \\ z_3 \end{pmatrix}, \; x_3 = \begin{pmatrix} z_2 \\ z_4 \\ z_5 \end{pmatrix}$$

It means that $\hat{F}_1$ can be chosen as

$$\hat{F}_1 = \begin{pmatrix} \dot{z}_1 - z_3 \\ \dot{z}_3 - (-z_5 \cdot z_1 - b \cdot z_3^2) \end{pmatrix}$$

corresponding to (7.27a) and (7.27c). White noise could thus be added to the states z_1 and z_3. A choice in this case, is to only add noise to $\dot{z}_3$ which should model disturbances caused by, *e.g.*, turbulence. The equations (7.27a) and (7.27c) then take the form

$$\dot{z}_1 = z_3 \tag{7.28a}$$
$$\dot{z}_3 = -z_5 \cdot z_1 - b \cdot z_3^2 + v \tag{7.28b}$$

where v is white noise. This corresponds to

$$\sigma = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

in (7.9). The next step in the noise modeling is to transform these equations together with the remaining noise-free equations into $\tilde{F}_1$ and $\hat{F}_2$ in (7.25). Doing this reveals that $\tilde{F}_1$, which is available as C code from Dymola, is of the desired form

$$\tilde{F}_1\big(t, x_1, x_3, \dot{x}_1 - \sigma(x_1, x_2)v\big)$$

that is, the noise term only occurs in affine form and together with $\dot{x}_1$. However, $\hat{F}_2$ includes the noise term v which is not allowed. To solve this problem, occurrences of v in $\hat{F}_2$ are deleted before it is used for particle filtering. Removing the noise from F_2 can typically be seen as adding noise in the original equations, but a user does not need to consider the exact form of this. (For illustration, a short discussion of the exact form for the pendulum case is included is anyway below.)

Next, a particle filter is implemented to estimate the internal variables of the system. To generate data for the estimation experiment, the model is inserted into the Simulink

environment using the Dymola-Simulink interface available with Dymola. The purpose of this experiment is not to demonstrate the performance of a filtering algorithm, but rather to show how DAE models can be used in a direct way when constructing particle filters. Therefore it is sufficient to use simulated data for the experiment. The constants were chosen as $L = 1$, $b = 0.05$ and $g = 9.81$. Process noise was generated with the *Band-Limited White Noise*-block in Simulink with noise power 0.01. The initial values of the states were $z_1 = 0.5$ and $z_3 = -0.1$. The measured variable is the tension in the pendulum z_5,

$$y(t_k) = z_5(t_k) + e(t_k)$$

Measurements with noise variance 0.1 were collected at the sampling interval 0.05 s.

After generating the data, a particle filter was implemented using the algorithm in Section 7.5 to estimate the internal variables z_1, z_2, z_3, z_4, and z_5. Since the selected states are z_1 and z_3, these are the variables that are estimated directly by the particle filter. The remaining variables are then computed by Dymola using $\hat{F}_2$.

The particle filter was implemented in MATLAB with the time updates being performed by simulating the model using the Dymola-Simulink interface. The initial particles were spread between $z_1 = 0.1$ and $z_1 = 0.6$ and between $z_3 = -0.2$ and $z_3 = 0.2$. Only positive values of z_1 were used since the symmetry in the system makes it impossible to distinguish between positive and negative z_1 using only measurements of z_5. The particle filter was tuned to use noise power 0.1 for the process noise and variance 0.2 for the measurement noise to simulate the situation were the noise characteristics are not exactly known. A typical result of an estimation is shown in Figure 7.2 where an estimation of z_1 is plotted together with the true value.

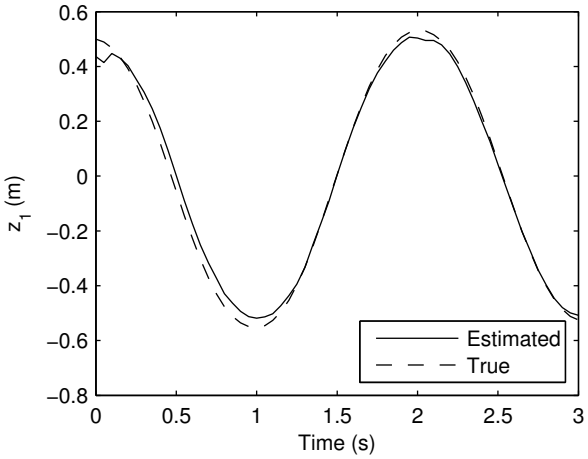


Figure 7.2: *Typical result of particle filtering.*

To examine the reliability of the filtering algorithm, 100 Monte Carlo runs were made.

Then the RMSE value was calculated according to

$$\mathrm{RMSE}(t) = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\big(x(t) - \hat{x}_j(t)\big)^2}$$

where M is the number of runs, here $M = 100$, $x(t)$ is the true state value and $\hat{x}_j(t)$ is the estimated state value in run j. The result is shown in Figure 7.3. The estimation error

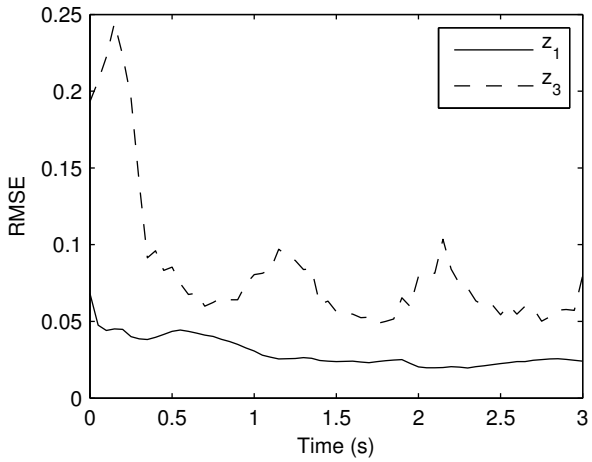


Figure 7.3: *RMSE for the estimations of z_1 and z_3 for 100 Monte Carlo runs.*

in the velocity z_3 is larger when the pendulum changes direction, which could mean that it is more difficult to estimate the velocity there.

Noise Modeling Details

When adding noise to a DAE, a user can only add noise so that it enters through a function σ. This was done in equation (7.28) above. However, noise must also be added according to the term $\hat{F}_{2;x_3}^{-1}\hat{F}_{2;x_1}\sigma(x_1, x_3)v$ in (7.12) to make all variables well-defined (otherwise the conditions of Theorem 7.2 will not be satisfied).

To compute $\hat{F}_2$, consider the pendulum equations (7.27). The constraints that defines the circle is

$$0 = z_1^2 + z_2^2 - L^2. \tag{7.29}$$

Differentiating (7.29) w.r.t. time gives

$$0 = 2z_1\dot{z}_1 + 2z_2\dot{z}_2.$$

Inserting (7.27a) and (7.27b) gives

$$0 = 2z_1z_3 + 2z_2z_4 \tag{7.30}$$

which after differentiation gives

$$0 = 2\dot{z}_1 z_3 + 2z_1 \dot{z}_3 + 2\dot{z}_2 z_4 + 2z_2 \dot{z}_4$$

Inserting the expressions for the derivatives gives

$$0 = 2z_3^2 + 2z_1(-z_5 \cdot z_1 - bz_3^2) + 2z_4^2 + 2z_2(-z_5 \cdot z_2 - bz_4^2 - g) \tag{7.31}$$

The equations (7.29), (7.30), and (7.31) together define one possible selection of $\hat{F}_2$. These can be used to compute

$$\hat{F}_{2;x_3}^{-1} \hat{F}_{2;x_1} \sigma(x_1, x_3) v = \hat{F}_{2;x_3}^{-1} \hat{F}_{2;x_1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} v = \begin{pmatrix} 0 \\ \frac{z_1}{z_2} \\ * \end{pmatrix} v \tag{7.32}$$

where the last term $*$ is unimportant since $\dot{z}_5$ does not occur in the equations. It can be realized that noise should be added to $\dot{z}_4$ according to

$$\dot{z}_4 + \frac{z_1}{z_2} v = -z_5 \cdot z_2 - b \cdot z_4^2 - g \tag{7.33}$$

to satisfy the conditions of Theorem 7.2.

7.8 Conclusions

In this chapter, a theoretical basis for introduction of noise processes in DAE models has been presented. The exact conditions that is obtained can be hard to use in practice, for instance since it requires rank tests. Therefore, an approximate solution was proposed. This solution uses the type of DAE solvers included in modeling environments for object-oriented modeling. Typically, these solvers produce an approximation of the transformation that is necessary to include noise in a feasible way.

It was also discussed how particle filtering can be implemented for DAE models, and an example which shows the required steps was presented. The results were similar to what could be expected from an implementation using a regular state-space model.

8

The Controllability Function

In this chapter the controllability function for nonlinear DAE 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 requires a lot of input effort 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 for example Scherpen (1994); Newman and Krishnaprasad (2000). Scherpen shows that for linear time-invariant state-space models, 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, for example Scherpen and Gray (2000).

The controllability function for regular time-invariant linear DAE models with consistent initial conditions, has been considered in Stykel (2004). The method suggested by Stykel also handles models that are not strangeness-free, without first using index reduction.

For nonlinear DAE systems, Lin and Ahmed (1991) consider controllability using the maximum principle. However, instead of formulating the controllability function, they solve the problem, using optimal control, as a feasibility problem.

In this chapter, the optimal control problem for the controllability function is solved using the results obtained in Chapter 3 and Chapter 4. Three different methods are derived. In Section 8.2.1, necessary conditions are derived based on the necessary conditions given in Chapter 3. The second method is described in Section 8.2.2 and is similar to a method derived in Scherpen (1994). It is based on completion of squares and gives sufficient conditions. These two methods find the controllability function on some set $x_1 \in \Omega_x$. The third method, presented in Section 8.3, finds a local solution, *i.e.*, a controllability function valid in a neighborhood of the origin. In practice, the local solution is truncated and this method will then give an approximation of the controllability function.

8.1 Problem Formulation

Basically, a general controllability function should 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

$$J_c = \int_{-\infty}^{0} m\big(u(t)\big)\, dt = \frac{1}{2} \int_{-\infty}^{0} u(t)^T u(t)\, dt \tag{8.1}$$

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

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

$$\begin{aligned} L_c\big(x_{1,0}\big) = \min_{u(\cdot)}\ & J_c \\ \text{s.t.} \quad \dot{x}_1 &= F_1(x_1, x_3, u) \\ 0 &= F_2(x_1, x_3, u) \\ x_1(0) &= x_{1,0} \in \Omega_x \\ 0 &= \lim_{t \to -\infty} x_1(t) \end{aligned} \tag{8.2}$$

The DAE model is for notational reasons assumed semi-explicit, but as was seen in Chapter 4 more general definitions are possible to include. The model is also assumed to have an equilibrium at the origin and to be strangeness-free. However, the region in which the model must be strangeness-free will vary between the sections. It means that either the assumption is required locally around a point or in a whole region.

Since the origin is an equilibrium, no control effort is required to stay there. This means that controllability function must satisfy $L_c(0) = 0$. Moreover, $L_c(x_{1,0})$ is defined as infinite 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.

The boundary conditions implies that the feedback law $u(\cdot)$ must give a closed-loop system asymptotically anti-stable for $x_{1,0} \in \Omega_x$. That is, if the time is reversed and considered as going from 0 to $-\infty$, the system must be asymptotically stable.

Throughout the chapter, only points in the set

$$\mathcal{N} = \{x_1 \in \Omega_x,\, x_3 \in \mathbb{R}^{n_2} \,|\, x_3 = \mathcal{R}(x_1, u),\, u \in \Omega_u\}$$

are considered, where Ω_u is either $\mathbb{R}^p$ or a neighborhood of $u = 0$ depending on which assumption that is used. That is, the controllability function is only computed for points not violating the constraints. This is ensured by choosing the final state $x_{1,0}$ in Ω_x. Note that if all $x_{1,0} \in \Omega_x$ can be reached, 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 $x_3(0) = x_{3,0}$, which will leave the value of the controllability function unchanged.

8.2 Methods Based on HJB Theory

In this section, two different methods are derived. The first method relies on the conditions derived in Chapter 3, while the second method uses completion of squares to find sufficient conditions for optimality.

8.2.1 Necessary Conditions

Consider a DAE model that satisfies Assumption A2, that is the model is strangeness-free for $u \in \mathbb{R}^p$. The optimal control problem (8.2) can then be seen as a special case of the case studied in Chapter 3 with the cost function L chosen as the squared input.

A difference compared to the standard optimal control problem is that the final state, and not the initial state, is specified. This means that the time can be considered as going backwards compared to the standard case, and the HJB (3.7) becomes

$$0 = \min_u \big(\frac{1}{2}u^T u - W_1(x_1)F_1(x_1, x_3, u) - W_2(x_1, x_3)F_2(x_1, x_3, u)\big) \tag{8.3}$$

where $W_1(x_1)$ and $W_2(x_1, x_3)$ 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 that corresponds to (8.3) can be found in Section 3.3. However, due to the structure of the cost function, the conditions can be simplified.

Corollary 8.1
Consider the optimal control problem (8.2). Then, the optimal solution must satisfy

$$0 = u^T - W_1(x_1)F_{1;u}(x_1, x_3, u) - W_2(x_1, x_3)F_{2;u}(x_1, x_3, u) \tag{8.4a}$$
$$0 = \frac{1}{2}u^T u - W_1(x_1)F_1(x_1, x_3, u) \tag{8.4b}$$
$$0 = F_2(x_1, x_3, u) \tag{8.4c}$$
$$0 = W_2(x_1, x_3) + W_1(x_1)F_{1;x_3}(x_1, x_3, u)F_{2;x_3}^{-1}(x_1, x_3, u) \tag{8.4d}$$

for $x_1 \in \Omega_x$.

In the corollary above, (8.4d) is included since the system is assumed to satisfy Assumption A2. Note that if the model has the control-affine-like structure discussed in Section 3.5, it is possible to simplify the equations even more.

8.2.2 Sufficient Conditions

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, the sufficiency is lost and only necessary conditions are obtained.

Therefore, consider the class of models

$$E\dot{x} = f(x) + g(x)u \tag{8.5}$$

where $E = \left(\begin{smallmatrix} I_{n_1} & 0 \\ 0 & 0 \end{smallmatrix}\right)$. For models with this structure, another approach can be used 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. The advantage is that sufficient conditions for optimality are obtained. The result is stated in the following theorem.

Theorem 8.1
Suppose there exist continuous functions $W_1(x_1) = V_{x_1}(x_1)$ and $W_2(x_1, x_3)$ such that $\tilde{L}_c(x) = \big(W_1(x_1),\, W_2(x_1, x_3)\big)$ fulfills

$$0 = \tilde{L}_c(x)f(x) + \frac{1}{2}\tilde{L}_c(x)g(x)g(x)^T\tilde{L}_c(x)^T \tag{8.6}$$

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

$$u = g(x)^T\tilde{L}_c(x)^T \tag{8.7}$$

the system (8.5) can be solved backwards in time from $t = 0$, with $x(t) \to 0, t \to -\infty$. Under these conditions, $L_c(x_1) = V(x_1)$ on Ω_x 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 (8.5) fulfills $x(t) \to 0$ as $t \to -\infty$ it follows that

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

where $V(x_1)$, $W_2(x_1, x_3)$ 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 (8.6) is satisfied. It can be realized that $V(x_{1,0})$ is a lower bound for the integral in (8.1). 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 $x_{1,0} \in \Omega_x$

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

□

The proof require Ω_x to be an invariant set for the closed-loop system. However, since the closed loop system is asymptotically anti-stable, such a choice of Ω_x is always possible, as shown in Section 2.6.1.

The requirement that the closed-loop model with (8.7) must be asymptotically stable going backwards in time for $x \in \mathcal{N}$ is equivalent to

$$E\dot{\tilde{x}}(s) = -\big(f(\tilde{x}(s)) + g(\tilde{x}(s))g(\tilde{x}(s))^T \tilde{L}_c(\tilde{x}(s))^T\big) \tag{8.8}$$

being asymptotically stable on Ω_x, where $\tilde{x}(s) = x(-s)$ and $s = -t$. To verify that (8.8) is asymptotically stable, the methods described in Section 2.7 can be used.

Similarly to the method by Xu and Mizukami (1993), Theorem 8.1 is primarily intended to verify optimality and not to calculate W_1 and W_2, since additional equations are needed in order to have the same number of unknowns as equations, see Chapter 3. Therefore, the standard procedure in cases when the model match the structure (8.5) is to combine the sufficient and necessary conditions in Section 8.2. That is, first candidate solutions are found using the necessary conditions and the optimal solution is then chosen using the sufficient conditions. The approach is illustrated in Section 8.4.1.

8.3 Existence and Computation of a Local Solution

In Chapter 4, it was proved that under certain conditions, an optimal control problem is ensured to have real analytic solution. Moreover, a computational procedure to obtain the solutions as Taylor expansions was presented. In this section, the same approach will be applied on the optimal control problem that defines the controllability function. However, minor changes of the assumptions on the cost function are needed.

In Chapter 4, and more specifically Theorem 4.4, the cost function is assumed to be positive definite in both x_1 and u locally around the origin. In the controllability function case, x_1 and x_3 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.

8.3.1 Basic Assumptions and Formulations

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

$$L_c(x_1) = \frac{1}{2}x_1^T G_c x_1 + L_{ch}(x_1) \tag{8.9}$$

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

For this end, the model is assumed to satisfy Assumption A8, *i.e.*, being real analytic on some set $\mathcal{W}$, which means that $F(x_1, x_3, u) = (F_1(x_1, x_3, u)^T,\ F_2(x_1, x_3, u)^T)^T$ can be expressed as convergent Taylor series as in (4.40). Since the boundary condition on x_1 is the final state and not the initial state, time is considered as going backwards compared to the optimal control problem (4.37). It means that the underlying state-space model will be

$$\dot{x}_1 = -\hat{A}x_1 - \hat{B}u - \hat{F}_{1h}(x_1, u) \tag{8.10}$$

where $\hat{A} = A_{11} - A_{12}A_{22}^{-1}A_{21}$, $\hat{B} = B_1 - A_{12}A_{22}^{-1}B_2$ and $\hat{F}_{1h}(x_1, u)$ are terms of at least order 2.

The cost function is already a convergent power series, since

$$L(x_1, x_3, u) = \frac{1}{2}u^T u \tag{8.11}$$

In this case, the reduced cost function $\hat{L}$ becomes very simple, since L depends neither on x_1 nor on x_3, and can be written as

$$\hat{L}(u) = \frac{1}{2}u^T u$$

which yields the cost matrix

$$\begin{pmatrix} \hat{Q} & \hat{S} \\ \hat{S}^T & \hat{R} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1/2I \end{pmatrix} \tag{8.12}$$

and no higher order terms, *i.e.*, $\hat{L}_h(u) = 0$. The lack of higher order terms also leads to $\hat{L}_{h;u}(u) = 0$.

As in Chapter 4, only feedback laws given by uniformly convergent power series are considered. However, because of the time change, Assumption A9 is reformulated as follows.

Assumption A15. The feedback laws are described by uniformly convergent power series

$$u(x_1) = Dx_1 + u_h(x_1) \tag{8.13}$$

where $u_h(x_1)$ are terms of at least order two. Furthermore, D must satisfy

$$\operatorname{Re}\operatorname{eig}(-\hat{A} - \hat{B}D) < 0$$

8.3.2 Existence of a Local Solution

From Theorem 4.4, it can be realized that a fundamental assumption in Chapter 4 was a positive definite cost matrix (8.12). However, this requirement is not satisfied in the calculation of the controllability function. The results in Chapter 4 are based on the proof in Lukes (1969) (see also Section 4.7). Careful examination of the proof shows that most parts are still valid 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, positive definiteness of (8.12) enters in two places of the proof. First, it is used to prove that the second order term in the optimal performance criterion is positive definite. This fact is used to determine which solution is optimal. Second, the equations corresponding to the lowest order terms of the optimal cost and optimal feedback law, *i.e.*, ARE (4.49a) and the matrix D_* (4.49b), need to have a stabilizing solution such that the condition above is satisfied.

In this case, (4.49a) and (4.49b) become

$$G_c\hat{A} + \hat{A}^T G_c + G_c\hat{B}\hat{B}^T G_c = 0 \tag{8.14a}$$

$$D_* - \hat{B}^T G_c = 0 \tag{8.14b}$$

and hence there must exist a solution G_c such that the eigenvalues to $-\hat{A} - \hat{B}\hat{B}^T G_c$ are all in $\mathbb{C}^-$.

Therefore, it is necessary to study what properties G_c will have when the cost function is given by (8.12), and under which conditions the ARE (8.14a) has a stabilizing solution such that the necessary properties of G_c are satisfied.

The following lemma answers the first question.

Lemma 8.1
Given an asymptotically stable matrix $\hat{A}$ and a matrix D such that $-\hat{A}_c = -\hat{A} - \hat{B}D$ is Hurwitz, the matrix G_c in the controllability function (8.9) is positive definite.

Proof: Consider a model (8.10) and feedback laws (8.13) such that $-\hat{A}_c$ is Hurwitz. Then if follows from Lukes (1969), that for a general optimal control problem, as in Chapter 4, the matrix P is given by

$$P = \int_0^\infty e^{-\hat{A}_c^T t}\big(\hat{Q} + \hat{S}D + D^T\hat{S}^T + D^T D\big)e^{-\hat{A}_c t}\,dt$$

In the controllability function case, where the cost matrix is given by (8.12), it follows that

$$G_c = \int_0^\infty e^{-\hat{A}_c^T t} D^T D e^{-\hat{A}_c t}\,dt$$

which obviously is positive semidefinite. For G_c to be zero, it requires that $De^{-\hat{A}_c t}x_0 \equiv 0$. Differentiation of this expression yields that

$$\underbrace{\begin{pmatrix} D \\ D(-\hat{A}_c) \\ \vdots \\ D(-\hat{A}_c)^{d-1} \end{pmatrix}}_{\mathcal{DA}} e^{-\hat{A}_c t}x_0 = 0, \forall t$$

The property above, can only hold if $\mathcal{DA}$ does not have full column rank. Therefore, the rank properties of this matrix is studied. The PBH-test, see Kailath et al. (2000), states that the matrix $\mathcal{DA}$ will have full column rank if and only if $Dv \neq 0$ for all non-trivial v such that $-\hat{A}_c v = v\lambda$.

Therefore, let λ and v satisfy the equality and assume that $Dv = 0$ for a non-trivial v. Then it follows that

$$0 = (-\hat{A} - \hat{B}D)v - v\lambda = -\hat{A}v - \lambda v$$

which means that λ must be an eigenvalue to both $-\hat{A}_c$ and $-\hat{A}$. This is impossible, since their eigenvalues are strictly separated by the imaginary axis. Hence, there is no λ and v and the theorem is proved. □

The lemma below shows when the considered ARE can be expected to have a solution such that G_c becomes positive definite and D_* becomes stabilizing.

Lemma 8.2
Assume that $\hat{A}$ is Hurwitz and $(\hat{A}, \hat{B})$ is controllable. Then the ARE (8.14a) has a unique positive definite solution G_c such that $\hat{A}_{c,b} = -\hat{A} - \hat{B}D_* = -\hat{A} - \hat{B}\hat{B}^T G_c$ is Hurwitz.

Proof: Since only $G_c \succ 0$ are considered, the ARE (8.14a) can be rewritten as the following Lyapunov equation

$$0 = \hat{A}G_c^{-1} + G_c^{-1}\hat{A}^T + \hat{B}\hat{B}^T \tag{8.15}$$

It is well-known that (8.15) 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).

For the solution above, it follows algebraically that

$$G_c^{-1}\hat{A}_{c,b}^T + \hat{A}_{c,b}G_c^{-1} = -\hat{B}\hat{B}^T$$

since G_c is symmetric. A linear feedback law does not change the controllability and therefore $(\hat{A}_{c,b}, \hat{B})$ is controllable. Then it follows as a standard result that $\hat{A}_{c,b}$ is Hurwitz, see for example Khalil (2002). □

The results in this section is summarized in the following theorem.

Theorem 8.2
Consider the optimal control problem (8.2). Assume the DAE model satisfies Assumptions A7, A8 and A15. Furthermore, assume that $\hat{A}$ is Hurwitz and $(\hat{A}, \hat{B})$ is controllable. Then locally the controllability function $L_c(x_1)$ exists and is the unique solution to

$$0 = u_*^T(x_1) - L_{c;x_1}(x_1)\big(F_{1;u} - F_{1;x_3}F_{2;x_3}^{-1}F_{2;u}\big) \tag{8.16a}$$

$$0 = \tfrac{1}{2}u_*^T(x_1)u_*(x_1) - L_{c;x_1}(x_1)F_1 \tag{8.16b}$$

$$0 = F_2 \tag{8.16c}$$

where F_1 and F_2 are evaluated in (x_1, x_3, u).

Proof: First, using the same approach as in Theorem 4.2, it is possible to show that the system of equations (8.16) is equivalent to formulating the problem in terms of the underlying state-space model (8.10) and the cost function $\hat{L}$ in (8.11). For this problem, it was motivated in the discussion that only two parts were left to prove. Both of these parts dealt with choosing the right solution to the ARE (8.14a).

From the assumptions, it is given that $\hat{A}$ is Hurwitz and D must be such that $-\hat{A}-\hat{B}D$ is Hurwitz. Lemma 8.1, shows that these facts imply that the optimal solution G_c have to be positive definite. From Lemma 8.2 it then follows that, under the given assumptions, there exists a unique positive definite solution G_c such that $\hat{A}_{c,b}$ is Hurwitz. Since $\hat{A}_{c,b}$ is the closed-loop system obtained if D_* in (8.14b) is used, it means that G_c is the unique solution that has the required properties. The existence of the controllability function then follows from Lukes (1969) or Section 4.7.1. □

Note that also fully implicit DAE models can be included as can be seen in Chapter 4.

As in Chapter 4 the conditions in Theorem 8.2 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 proved in a very similar way to the proof regarding stabilizability, see Section 4.5 or Dai (1989).

8.3.3 A Computational Algorithm

In the former section, it was proved that the controllability function exists. In this section, a computational algorithm is given. Based on the discussion in Chapter 6, it is known that it is favorable to solve the set of equations (8.16) at once, *i.e.*, without deriving the power series of $\mathcal{R}$. However, of course the expressions for the higher order terms in Chapter 4 can be modified to fit the controllability function problem.

First the lower order terms are obtained from (8.14) and then the higher order terms of $L_c(x_1)$ are, similarly as in Chapter 4, obtained from the expressions

$$\begin{aligned} &L_{c;x_1}^{[m]}(x_1)\hat{A}_c x_1 = \\ &-\sum_{k=3}^{m-1} L_{c;x_1}^{[k]}(x_1)\hat{B}u_*^{[m-k+1]}(x_1) - \sum_{k=2}^{m-1} L_{c;x_1}^{[k]}(x_1)\hat{F}_{1h}^{[m-k+1]}(x_1,u_*) \\ &+\sum_{k=2}^{\lfloor\frac{m-1}{2}\rfloor} u_*^{[k]}(x_1)^T u_*^{[m-k]}(x_1) + \frac{1}{2}u_*^{[m/2]}(x_1)^T u_*^{[m/2]}(x_1) \end{aligned} \tag{8.17a}$$

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 equation for the series expansion of the feedback law is obtained as

$$u_*^{[k]}(x_1) = L_{c;x_1}^{[k+1]}(x_1)\hat{B} + \sum_{i=1}^{k-1} L_{c;x_1}^{[k-i+1]}(x_1)\hat{F}_{1h,u}^{[i]}(x_1,u_*) \tag{8.17b}$$

where $k = 2, 3, \ldots$.

The equations in (8.17) are very similar to the original equations in Chapter 4, however, the computation of the equations above are less involved, since $\hat{L}_h$ and $\hat{L}_{h;u}$ are zero.

8.4 Examples

In order to illustrate the methods for computing the controllability function, two different examples will be presented.

8.4.1 A Rolling Disc

Consider a DAE model given by the set of differential and algebraic equations

$$\dot{z}_1 = z_2 \tag{8.18a}$$

$$\dot{z}_2 = -\frac{k_1}{m}z_1 - \frac{k_2}{m}z_1^3 - \frac{b}{m}z_2 + \frac{1}{m}\lambda \tag{8.18b}$$

$$\dot{z}_3 = -\frac{r}{J}\lambda + \frac{1}{J}u \tag{8.18c}$$

$$0 = z_2 - rz_3 \tag{8.18d}$$

The model describes a disc, rolling on a surface without slipping, see Figure 8.1. The disc is connected to a fixed wall with a nonlinear spring and a linear damper. The spring

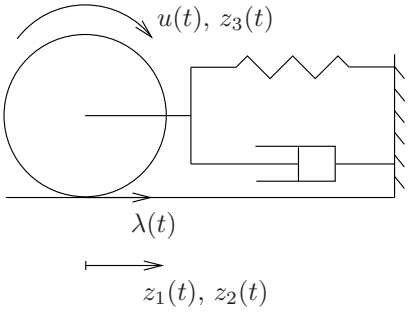


Figure 8.1: *A disc, which rolls on a surface without slipping. The disc is affected by a nonlinear spring and a linear damper.*

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, λ is the contact force between the disc and the surface.

This model has strangeness index one and before the methods in this chapter are applied, index reduction is needed. If the method in Section 2.4 is used, (8.18) can be rewritten as the strangeness-free model

$$\dot{z}_1 = z_2 \tag{8.19a}$$

$$\dot{z}_2 = -\frac{k_1}{m} z_1 - \frac{k_2}{m} z_1^3 - \frac{b}{m} z_2 + \frac{1}{m}\lambda \tag{8.19b}$$

$$0 = z_2 - r z_3 \tag{8.19c}$$

$$0 = -\frac{k_2}{m} z_1^3 - \frac{k_1}{m} z_1 - \frac{b}{m} z_2 + \left(\frac{r^2}{J} + \frac{1}{m}\right)\lambda + \frac{-r}{J} u \tag{8.19d}$$

The variables will be denoted $x = (z_1,\ z_2,\ z_3,\ \lambda)^T$ and it can be seen that $x_1 = (z_1,\ z_2)^T$ and $x_3 = (z_3,\ \lambda)^T$.

From (8.4d), it follows that W_2 must satisfy

$$\begin{aligned} W_2(x_1, x_3) &= -V_{x_1}(x_1) F_{1;x_3}(x_1, x_3, u) F_{2;x_3}^{-1}(x_1, x_3, u) \\ &= -V_{x_1}(x_1) \begin{pmatrix} 0 & 0 \\ 0 & \frac{J}{J+mr^2} \end{pmatrix} \end{aligned} \tag{8.20}$$

Since F_1 is independent of u, (8.4a) becomes

$$u = F_{2,u}(x_1, x_3, u)^T W_2(x_1, x_3)^T = \begin{pmatrix} 0 & \frac{r}{J+mr^2} \end{pmatrix} V_{x_1}(x_1)^T \tag{8.21}$$

For (8.19), it is possible to compute $x_3 = \mathcal{R}(x_1, u)$ explicitly using the last two rows as

$$\begin{aligned} z_3 &= \frac{1}{r} z_2 \\ \lambda &= \left(\frac{r^2}{J} + \frac{1}{m} \right)^{-1} \left(\frac{k_1}{m} z_1 + \frac{k_2}{m} z_1^3 + \frac{b}{m} z_2 + \frac{r}{J} u \right) \end{aligned} \tag{8.22}$$

If (8.21) and (8.22) are substituted into (8.4b) and if the performance criterion is assumed to have the structure $V(x_1) = a_1 z_1^2 + a_2 z_1^4 + a_3 z_2^2$. Then (8.4b) can be solved for the unknowns a_1, a_2 and a_3 and the solutions for $V(x_1)$ become either

$$V(x_1) = b k_1 r^2 z_1^2 + \frac{1}{2} b k_2 r^2 z_1^4 + b\big(J + m r^2\big) z_2^2 \tag{8.23}$$

or the trivial solution $V(x_1) = 0$. Back-substitution of (8.23) into (8.20) and (8.21) yields

$$W_2(x_1, x_3) = \begin{pmatrix} 0 & -2bJz_2 \end{pmatrix}, \quad u(x_1) = 2brz_2 \tag{8.24}$$

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 (8.19) with the control choice (8.24). Instead, stability is proved using the closed-loop reduced system when the time is considered as going backwards, *i.e.*,

$$\dot{x}_1 = -F_{red,cl}(x_1)$$

where

$$F_{red,cl}(x_1) = \begin{pmatrix} z_2 \\ -\frac{k_1}{\frac{J}{r^2}+m} z_1 - \frac{k_2}{\frac{J}{r^2}+m} z_1^3 - \frac{b}{\frac{J}{r^2}+m} z_2 + \frac{1}{(\frac{J}{r^2}+m)r} 2brz_2 \end{pmatrix}$$

For this system, $V(x_1)$ is a Lyapunov function since

$$\begin{aligned} V(x_1) = b k_1 r^2 z_1^2 + \frac{1}{2} b k_2 r^2 z_1^4 + b\big(J + m r^2\big) z_2^2 > 0 \\ -V_{x_1}(x_1) F_{red,cl}(x_1) = -2b^2 r^2 z_2^2 < 0 \end{aligned}$$

for all $x_1 \neq 0$. The motivation is that if $V_{x_1}(x_1) F_{red,cl}(x_1) = 0$ it requires that $z_2 = 0$, but then $z_1 = 0$ because k_1 and k_2 are positive. Therefore, the conditions in Theorem 8.1 are fulfilled, yielding

$$L_c(x_1) = V(x_1)$$

for all $x_1 \in \mathbb{R}^2$ with $u(x_1)$ chosen as (8.24).

Note that since the controllability function is polynomial it is also possible to find the solution using the method in Section 8.3.

8.4.2 An Artificial System

This example considers a completely artificial model, but illustrates an advantage of the methods in Section 8.2.

The models is

$$\dot{z}_1 = -z_1 + z_2 + \frac{1}{2}z_2^2$$
$$0 = z_2 - u$$

where $x_1 = z_1$ and $x_3 = z_2$, and fits into the affine structure (8.5). The interesting feature is that the underlying 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 model fits within the structure (3.14), the necessary conditions in Corollary 8.1 reduce to

$$0 = W_1(x_1)(-z_1 - \frac{1}{2}z_2^2) + \frac{1}{2}W_1(x_1)^2(1+z_2)^2 \tag{8.25a}$$
$$0 = z_2 - (1+z_2)W_1(x_1) \tag{8.25b}$$

where we have used that

$$\hat{f}(x_1, x_3) = -z_1 + z_2 + \frac{1}{2}z_2^2 - (1+z_2)z_2 = -z_1 - \frac{1}{2}z_2^2$$
$$\hat{g}(x_1, x_3) = 1 + z_2$$

The expressions for u and $W_2(x_1, x_3)$ become

$$u = (1+z_2)W_1(x_1), \quad W_2(x_1, x_3) = -(1+z_2)W_1(x_1)$$

From (8.25b), 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 (8.25a) yields

$$0 = \frac{z_2}{(1+z_2)}\Big(-z_1 + \frac{1}{2}z_2\Big)$$

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 $W_2(x_1, x_3)$ become

$$u = 0, \quad W_1(x_1) = 0, \quad W_2(x_1, x_3) = 0$$

while the second solution, $z_2 = 2z_1$, gives

$$u = 2z_1, \quad W_1(x_1) = \frac{2z_1}{1+2z_1}, \quad W_2(x_1, x_3) = -2z_1$$

Hence, two different solutions to the necessary conditions in Corollary 8.1 are obtained. Since the system has control-affine structure, Theorem 8.1 can be used to determine which of these solutions that is optimal. The first solution solves (8.6) on $\mathcal{N} = \{z_1 \in \mathbb{R},\, z_2 \in \mathbb{R} \,|\, z_2 = 0\}$, while the second solution solves (8.6) 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

$$L_c(x_1) = z_1 - \frac{1}{2}\ln(2z_1 + 1) \tag{8.26}$$

Figure 8.2 shows the controllability function. 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.

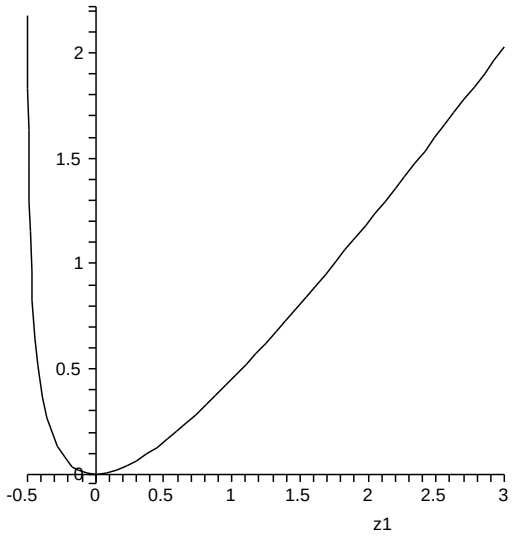


Figure 8.2: *The controllability function $L(x_1)$ for the artificial example.*

9

The Observability Function

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 model 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 models 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 model, 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, and is based on that the solution is found recursively. 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 DAE models with consistent initial conditions, see Stykel (2004). The method presented by Stykel can also handle DAE models of higher index without using index reduction.

In this chapter, two different methods to calculate the observability function for nonlinear DAE models are presented. In Section 9.2, 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 in Section 9.2, the power series method is presented and extended.

9.1 Problem Formulation

The observability function should reflect the energy in the output signal when the model 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

$$L_o\big(x_1(0)\big) = \frac{1}{2}\int_0^\infty y(t)^T y(t)\,dt \tag{9.1}$$

subject to

$$\begin{aligned} x_1(0) &= x_{1,0} \in \Omega_x \\ u(t) &= 0,\ 0 \le t < \infty \end{aligned}$$

and a DAE model. In this chapter, the DAE model is assumed to be in the form

$$\dot{x}_1 = F_1(x_1, x_2) \tag{9.2a}$$
$$0 = F_2(x_1, x_2) \tag{9.2b}$$
$$y = h(x_1, x_2) \tag{9.2c}$$

Hence, an output equation is added explicitly and it is assumed that $h(0,0) = 0$. The DAE model in (9.2) is also assumed to have an equilibrium at the origin.

Similar to Chapter 8, different assumptions on the implicit function is used in the different sections. In Section 9.2, Assumption A2 is made while in Section 9.3, Assumption A7 is used instead. In both cases, it is known that on a set Ω_x, it holds that $x_2 = \mathcal{R}(x_1)$. The corresponding set of points (x_1, x_2) satisfying the constraints will be denoted $\mathcal{N}$, *i.e.*,

$$\mathcal{N} = \{x_1 \in \Omega_x,\, x_2 \in \mathbb{R}^{n_2} \,|\, x_2 = \mathcal{R}(x_1)\} \tag{9.3}$$

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 model, it is also necessary that (9.2) 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 Ω_x' is defined as

$$\mathcal{N}' = \{x_1 \in \Omega_x',\, x_2 \in \mathbb{R}^{n_2} \,|\, x_2 = \mathcal{R}(x_1)\}$$

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 model, *i.e.*, $x(t)$, for a given initial condition and then integrate the square of the corresponding output.

Remark 9.1. 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 (9.1), the energy measure can be some positive definite function $m\big(y(t)\big)$.

9.2 A Method Based on Partial Differential Equation

Solving the DAE model (9.2) 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 9.1
Suppose the model (9.2) *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*

$$0 = \frac{1}{2}h(x_1, x_2)^T h(x_1, x_2) + V_{x_1}(x_1)F_1(x_1, x_2) \tag{9.4}$$

for all $(x_1, x_2) \in \mathcal{N}'$. *Then for all* $x_{1,0} \in \Omega_x'$, *it holds that*

$$L_o(x_{1,0}) = V(x_{1,0}) \tag{9.5}$$

Proof: Assume that only $x_{1,0} \in \Omega_x'$ are considered. Then, for any solution to (9.2) it follows that

$$V\big(x_1(0)\big) = \int_0^\infty \frac{dV\big(x(t)\big)}{dt}\,dt = \int_0^\infty -V_{x_1}F_1\,dt = \int_0^\infty y^T y\,dt \tag{9.6}$$

provided (9.4) is satisfied and $V(x_1)$ is a sufficiently smooth function. Therefore, for all $x_{1,0} \in \Omega_x'$ it follows that

$$L_o(x_{1,0}) = \frac{1}{2}\int_0^\infty y^T y\,dt = V(x_{1,0})$$

□

The set $\mathcal{N}'$ defines, in a rather implicit manner, that (9.4) 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 (9.4) can be reformulated as

$$\begin{aligned} 0 &= \frac{1}{2}h(x_1, x_2)^T h(x_1, x_2) + V_{x_1}(x_1)F_1(x_1, x_2) \\ 0 &= F_2(x_1, x_2) \end{aligned}$$

which must hold for $x_1 \in \Omega_x'$. Another reformulation is to use that for on $\mathcal{N}'$, it is known that $x_2 = \mathcal{R}(x_1)$ and the result becomes

$$0 = \frac{1}{2}h\big(x_1, \mathcal{R}(x_1)\big)^T h\big(x_1, \mathcal{R}(x_1)\big) + V_{x_1}(x_1)F_1\big(x_1, \mathcal{R}(x_1)\big) \tag{9.7}$$

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 models.

9.3 A 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 $\mathcal{R}(x_1)$ 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.

9.3.1 Power Series Expansion of the Reduced Model

Consider DAE models (9.2) which satisfy Assumption A7. Similar to in Chapter 4, another assumption is also made.

Assumption A16. The functions F_1, F_2 and h are analytic on a set $\mathcal{W}$, which is a neighborhood of the origin $(x_1, x_3) = 0$

For notional reasons, $\mathcal{W}$ is assumed to be large enough to cover the region in which the $\mathcal{R}$ is defined. Based on Assumption A16 it follows that F_1, F_2 and h in (9.2) can be expanded in convergent power series as

$$\begin{pmatrix} F_1(x_1, x_2) \\ F_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} F_{1h}(x_1, x_2) \\ F_{2h}(x_1, x_2) \end{pmatrix}$$
$$h(x_1, x_2) = \begin{pmatrix} C_1 & C_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + h_h(x_1, x_2)$$

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 order two. From Assumption A7, it is known that A_{22} has full rank, since $F_{2;x_2}(0, 0) = A_{22}$ is nonsingular.

The model (9.2) can be written as the reduced state-space model

$$\dot{x}_1 = \hat{F}_1(x_1) = F_1\big(x_1, \mathcal{R}(x_1)\big)$$
$$y = \hat{h}(x_1) = h\big(x_1, \mathcal{R}(x_1)\big)$$

and by making a series expansion of the implicit function $\mathcal{R}(x_1)$ 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

$$\hat{F}_1(x_1) = \hat{A}x_1 + \hat{F}_{1h}(x_1), \qquad \hat{h}(x_1) = \hat{C}x_1 + \hat{h}_h(x_1) \tag{9.8}$$

where

$$\hat{A} = A_{11} - A_{12}A_{22}^{-1}A_{21}, \qquad \hat{C} = C_1 - C_2A_{22}^{-1}A_{21}$$

and the higher order terms of $\hat{F}_{1h}(x_1)$ and $\hat{h}_h(x_1)$ can be obtained as

$$\hat{F}_{1h}^{[m]}(x_1) = F_{1h}\big(x_1, \mathcal{R}^{[1]}(x_1) + \ldots \mathcal{R}^{[m-1]}(x_1)\big) + A_{12}\mathcal{R}_h^{[m]}(x_1)$$
$$\hat{h}_h^{[m]}(x_1) = h_{1h}\big(x_1, \mathcal{R}^{[1]}(x_1) + \ldots + \mathcal{R}^{[m-1]}(x_1)\big) + C_2\mathcal{R}_h^{[m]}(x_1)$$

for $m = 2, 3, \ldots$.

9.3.2 Existence and Computation of a Local Solution

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 (9.2) 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

$$L_o(x_1) = \frac{1}{2} x_1^T G_o x_1 + L_{oh}(x_1) \tag{9.9}$$

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 9.2, it is known that the observability function can be found by solving (9.7). If (9.8) and (9.9) are inserted into (9.7) we obtain

$$\begin{aligned} 0 &= \frac{1}{2}\big(\hat{C}x_1 + \hat{h}_h(x_1)\big)^T\big(\hat{C}x_1 + \hat{h}_h(x_1)\big) + \big(x_1^T G_o + L_{oh;x_1}(x_1)\big)\big(\hat{A}x_1 + \hat{F}_{1h}(x_1)\big) \\ &= \frac{1}{2} x_1^T\Big(\hat{C}^T\hat{C} + G_o\hat{A} + \hat{A}^T G_o\Big)x_1 \\ &\quad + L_{oh;x_1}(x_1)\hat{A}x_1 + L_{o;x_1}(x_1)\hat{F}_h(x_1) + x_1^T\hat{C}^T\hat{h}_h(x_1) + \frac{1}{2}\hat{h}_h(x_1)^T\hat{h}_h(x_1) \end{aligned}$$

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

$$0 = G_o\hat{A} + \hat{A}^T G_o + \hat{C}^T\hat{C} \tag{9.10a}$$

and the higher order terms in $L_o(x_1)$, *i.e.*, $L_{oh}(x_1)$ must satisfy

$$\begin{aligned} L_{o;x_1}^{[m]}(x_1)\hat{A}x_1 = &- \sum_{k=2}^{m-1} L_{o;x_1}^{[k]}(x_1)\hat{F}_{1h}^{[m+1-k]}(x_1) - x_1^T\hat{C}^T\hat{h}_h^{[m-1]}(x_1) \\ &- 2\sum_{k=2}^{\lfloor\frac{m-1}{2}\rfloor} \hat{h}_h^{[k]}(x_1)^T\hat{h}_h^{[m-k]}(x_1) - \hat{h}_h^{[m/2]}(x_1)^T\hat{h}_h^{[m/2]}(x_1) \end{aligned} \tag{9.10b}$$

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 $\sum_k^l = 0$ for $l < k$.

The second order term in $L_o(x_1)$ is given by a Lyapunov equation (9.10a). A lemma presenting conditions under which the Lyapunov equation will have a solution is formulated below.

Lemma 9.1
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 (9.10b). 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 (9.8). 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 (9.10b) 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 $L_o(x_1)$.

The results are summarized in the following theorem.

Theorem 9.2
Consider a DAE model given in the form (9.2). Assume that it satisfies Assumptions A7 and A16. Furthermore, assume $\hat{A}$ is Hurwitz. Then, a local observability function, given in the form (9.9), 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 (9.10b). 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 from the discussion in the section. The second part, *i.e.*, that $L_o(x_1)$ is positive definite locally in a neighborhood of the origin when $(\hat{A}, \hat{C})$ is observable, follows since $G_o \succ 0$, see Lemma 9.1. □

Remark 9.2. 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 9.3. As in the computation of the controllability function, more general model descriptions, not being semi-explicit, can be treated using the results presented in Section 5.2.6.

10

Model Reduction

In many engineering situations, the model obtained becomes rather complex. To simplify the analysis and control design, it is desirable to reduce the order of the model without affecting the accuracy too much. In this context, accuracy most often refers to the input-output behavior. The basic idea, invented for linear state-space system by Moore (1981), is to analyze the system and find a linear coordinate change such that the transformed system reveals which coordinate directions that are most important for the input-output behavior. The analysis is done by measuring the energy a certain state corresponds to in the input and output signals, respectively, using the controllability and observability gramians. The coordinate change is chosen such that the gramians are simultaneously diagonalized and equal, and the diagonal terms are the squared Hankel singular values. The system is, after this coordinate change has been applied, denoted balanced.

The interpretation is then that a small Hankel singular value means that the particular direction is hard to control, *i.e.*, requires a large amount of control effort, while its contribution to the output energy is small. Therefore, the given direction does not influence the input-output behavior as much as a direction with a larger singular value. In Moore (1981), the reduction was then accomplished by removing the states with small Hankel singular values.

In Scherpen (1994), the earlier results for linear state-space models were partially extended to nonlinear state-space models. Scherpen used the nonlinear variants of the controllability and observability functions, mentioned in the earlier chapters, to measure the input and output energies. She showed that, using a state transformation, it is possible to transform the controllability function to one half the sum of the squares of the new states variables, while the observability function is diagonalized. The diagonal terms, denoted singular value functions, are in this case state dependent, and depend normally on all states. The model reduction was then, similar to Moore, done by removing states with small singular value functions.

In the structure obtained by Scherpen the different directions are not really separated from each other. Therefore, an enhanced form denoted input-normal/output-diagonal,

is introduced in Fujimoto and Scherpen (2003a,b). This form is also studied in Krener (2008). He shows how the controllability and observability functions can be written in so-called input normal form of degree m, which means that the contributions from different coordinate directions are separated up to some desired order. There are two differences compared with the nonlinear reduction methods described by Scherpen and Fujimoto. First, this method does not simply pick out the part of the system corresponding to the largest singular value functions as done in their papers. Instead, a minimization is done to create a reduced order model which hopefully fits the true model better. Second, this work includes error bounds, which the earlier mentioned nonlinear methods do not. As usual for nonlinear systems, the error bounds depend on the input signals, and in his work the input signals are those obtained when solving for the controllability function.

The results in Krener (2008) are based on power series computations, which make them possible to use computationally. The same idea was also derived in Fujimoto and Tsubakino (2007) for the standard truncation method.

The methods described above take into account that the model is nonlinear. However, it means that the computational complexity is rather high. Therefore, another approximate computational approach is to use empirical measures of the energy instead. The method based on empirical gramians assume that the input consists of impulses and computes an estimate of them, see Lall et al. (1999, 2002); Hahn and Edgar (2002). The estimate is exact if Dirac impulses are used as inputs and the system is linear. Later, the concept is extended in Hahn et al. (2003), where the input need not be impulses but can be step signals or signals closer to those that the model will be used for. The gramians are then denoted controllability and observability covariances, respectively.

A method which also aims at finding the controllability functions without solving the corresponding partial differential equation is described in Newman and Krishnaprasad (1998, 2000). The method relies on theory for stochastically excited systems, and is exact for linear systems and approximate for nonlinear systems.

The area of model reduction for general DAE models is not very large. For linear DAE models, Stykel (2004) is a good reference where higher index problems are studied as well. For nonlinear DAE models, the papers Hahn and Edgar (2002) and Sun and Hahn (2005) are based on the covariance measures mentioned above. For model reduction of chemical systems, see Vora and Daoutidis (2001).

10.1 Model Reduction of State-Space Models

In this section, a short introduction to model reduction of state-space systems is given. For a more thorough discussion see Scherpen (1994); Fujimoto and Scherpen (2005); Krener (2008); Skogestad and Postlethwaite (2001). There are two major steps in model reduction. First the system needs to be transformed to a form which reveals which parts are most important. Second, the model has to be approximated in some sense.

10.1.1 Revealing the Important Parts of the System

Linear Systems

The first step in model reduction is to extract the parts of the system that are most important for the input-output behavior. First consider the linear case

$$\dot{x} = Ax + Bu \tag{10.1a}$$
$$y = Cx + Du \tag{10.1b}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^q$ and $u \in \mathbb{R}^p$.

From Chapters 8 and 9, it is known that the controllability and observability functions for the linear case can be written as

$$L_c(x) = \frac{1}{2}x^T G_c x$$
$$L_o(x) = \frac{1}{2}x^T G_o x$$

where G_c^{-1} and G_o are the controllability and observability gramians, respectively. Using a linear change of coordinates, *i.e.*, $x = Tz$, it is possible to simultaneously diagonalize both G_c^{-1} and G_o as

$$\Sigma \triangleq G_c^{-1} = G_o = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix} \tag{10.2}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$, see Moore (1981). The values σ_i, $i = 1, \ldots, n$ are denoted the Hankel singular values. A representation with a diagonal Σ is called balanced. Another representation possible to obtain from (10.2) using a linear coordinate change $z = \Sigma^{\frac{1}{2}} q$ is the so-called input normal form where $G_c^{-1} = I$ and $G_o = \Sigma^2$.

The name Hankel singular values stems from the fact that σ_1 is the Hankel norm of the system, defined as

$$\|\Sigma\|_H \triangleq \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{L_o(x)}{L_c(x)} = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{x^T G_o x}{x^T G_c x} = \sup_{\substack{q \in \mathbb{R}^n \\ q \neq 0}} \frac{q^T \Sigma^2 q}{q^T q} = \sigma_1^2 \tag{10.3}$$

The basic idea is that for small σ_i, the amount of control energy required to reach the state $z = (0, \ldots, 0, z_i, 0, \ldots, 0)$ is large while the output energy generated by the same state is small. It means that if there is a major gap between two Hankel singular values, *i.e.*, if $\sigma_k \gg \sigma_{k+1}$ for some k, the last state components x_{k+1} to x_n will be less important from an energy point of view and can therefore be approximated without affecting the input-output behavior too much.

Note that even though the Hankel singular values are similarity invariant, *i.e.*, independent of the choice of state coordinates, it is important that the inputs and outputs are scaled such that their different components match in size.

Nonlinear Systems

Now consider a general nonlinear model

$$\dot{x} = F(x, u) \tag{10.4a}$$
$$y = h(x) \tag{10.4b}$$

where $x \in \mathbb{R}^d$, $y \in \mathbb{R}^q$, $u \in \mathbb{R}^p$, $F : \mathbb{R}^{d+p} \to \mathbb{R}^d$ and $h : \mathbb{R}^d \to \mathbb{R}^q$.

Motivated by the case for linear systems, the objective is to find a structure of the controllability and observability function, respectively, which separates the coordinates in a way such that their relative importance is revealed. This problem has been studied in numerous works. One of the first is presented in Scherpen (1994). However, first two assumptions are needed.

Assumption A17. The linearization of (10.4) is asymptotically stable, controllable and observable.

Assumption A18. The eigenvalues of $G_c^{-1}G_o$ are distinct.

Under these assumptions, it is possible to prove the following result given that F and h are smooth, see Scherpen (1994).

Theorem 10.1
Consider the system (10.4) and assume that it satisfies Assumption A17 and A18. Then there exist a neighborhood U and a smooth coordinate transformation $x = \Phi(z)$, $\Phi(0) = 0$ on U, which converts the system (10.4) into an input normal/output diagonal form, where

$$L_c\big(\Phi(z)\big) = \frac{1}{2} z^T z \tag{10.5a}$$
$$L_o\big(\Phi(z)\big) = \frac{1}{2} z^T \begin{pmatrix} \tau_1(z) & & 0 \\ & \ddots & \\ 0 & & \tau_n(z) \end{pmatrix} z \tag{10.5b}$$

with $\tau_1(z) \geq \ldots \geq \tau_n(z)$ being the so-called smooth singular value functions on U.

Proof: See (Scherpen, 1994, pp. 38 and 45). □

The form above has three drawbacks. First, the different axes are not completely separated. It means that it can be hard to decide whether a state is important or not. Second, as pointed out in Gray and Scherpen (2001), the singular value functions $\tau_i(z)$, $i = 1, \ldots, n$ in (10.5) are not unique except for $x = 0$, where they coincide with the squared singular values as defined in (10.2). Third, except for $x = 0$, the squared singular value functions $\tau_i(z)$ are not equal to the Hankel norm of the nonlinear system.

To solve the first and the third issue, an additional coordinate transformation is performed in Fujimoto and Scherpen (2005).

Theorem 10.2
Consider a nonlinear system (10.4). Suppose that Assumptions A17 and A18 hold. Then

there exist a neighborhood U of the origin and a coordinate transformation $x = \Phi(z)$, $\Phi(0) = 0$ on U converting the system into input normal/output diagonal form (10.5) with the following properties:

$$z_i = 0 \Leftrightarrow L_{c;z_i}\big(\Phi(z)\big) = 0 \Leftrightarrow L_{o;z_i}\big(\Phi(z)\big) = 0 \tag{10.6a}$$

$$\tau_i(0,\ldots,0,z_i,0,\ldots,0) = \rho_i^2(z_i) \tag{10.6b}$$

$$\tau_{i;z}(0,\ldots,0,z_i,0,\ldots,0) = \Big(0,\ldots,0,\tfrac{d\rho_i^2(z_i)}{dz_i},0,\ldots,0\Big) \tag{10.6c}$$

for $i = 1,\ldots,n$. Here, $\rho_i(z_i)$ are the axis singular value functions which are the gain functions to the Hankel operator. In particular, if $U = \mathbb{R}^n$, then

$$\|\Sigma\|_H = \sup_{z_1\in\mathbb{R}} \sqrt{\tau_1(z_1,0,\ldots,0)}$$

Proof: See Fujimoto and Scherpen (2005). Note that in Fujimoto and Scherpen (2005), the prerequisites are formulated in terms of existence of the controllability operator $\mathcal{C}$, the pseudo-inverse of this operator and the observability operator. All these can be proved to exist under Assumption A17. □

The last theorems show what can be obtained using smooth transformations. However, in this chapter, the goal is to obtain an algorithm that finds the reduced order model for a nonlinear DAE model order-by-order. For this purpose, the form defined in Krener (2008) is introduced.

Theorem 10.3
Consider the nonlinear system (10.4) and assume that it fulfills Assumption A17 and A18. Then there exists at least one coordinate transformation

$$x = \Phi^{d]}(z), \quad \Phi(0) = 0 \tag{10.7}$$

converting the system into input normal form of degree m for which

$$L_c\big(\Phi(z)\big) = \frac{1}{2}z^T z + \mathcal{O}(z_i)^{m+2} \tag{10.8a}$$

$$L_o\big(\Phi(z)\big) = \frac{1}{2}\sum_{i=1}^{n} \tau_i^{m-1]}(z_i) z_i^2 + \mathcal{O}(z_i)^{m+2} \tag{10.8b}$$

where $\tau_i^{m-1]}(z_i) = \tau_{i0} + \tau_{ih}^{m-1]}(z_i)$ are polynomials in z_i with terms of order 0 through $m-1$. They are called squared singular value polynomials of order $m-1$.

Furthermore, for $m \leq 6$ (or if the system is odd $m \leq 12$), the observability and observability functions will be unique. However, the system and change of coordinates that achieves input normal form of degree m are not necessarily unique even to order m.

Proof: See Krener (2008). □

In Fujimoto and Tsubakino (2006), which together with the work by Krener, is the foundation of the work presented here, another structure is used instead. However, as shown below both of these structures have the same computational properties.

Lemma 10.1
Suppose (10.4) satisfies Assumptions A17 and A18. Then there exists a coordinate transformation

$$x = \Phi^{d]}(z), \quad \Phi(0) = 0$$

in a neighborhood of the origin that brings the system to a form for which

$$L_c\big(\Phi(z)\big) = \frac{1}{2} z^T z + \mathcal{O}(z)^{m+2} \tag{10.9a}$$

$$L_o\big(\Phi(z)\big) = \frac{1}{2} \sum_{i=1}^{n} z_i^2 \rho_i^{m-1]}(z_i)^2 + \mathcal{O}(z)^{m+2} \tag{10.9b}$$

where $\rho_i(z_i)$ are polynomials, such that for $i = \{1, 2, \ldots, n\}$.

Proof: The property to prove is that for given squared singular values $\tau_i^{m-1]}(z_i)$, $i = 1, \ldots, n$, the singular value functions $\rho_i^{m-1]}(z_i)$ are uniquely determined from the fact that (10.8) should equal (10.9) up to order $m + 1$. To prove uniqueness, the following equations are analyzed

$$\tau_i^{m-1]}(z_i) = \big(\rho_i^{m-1]}(z_i)\big)^2 + \mathcal{O}(z_i)^m$$

Let each $\rho_i^{m-1]}(z_i)$ be a polynomial of maximal order $m - 1$ and assume that $\rho_{0,i} > 0$, *i.e.*, the constant term in each $\rho_i^{m-1]}(z_i)$ should be positive. Then it follows that

$$\rho_0 = \sqrt{\tau}, \qquad \rho_i^{[m]}(z_i) = \frac{1}{2\rho_0} \tau_i^{[m]}(z_i), \ i = 1, \ldots, n$$

for m up to the given order. □

The expressions for L_c and L_o in (10.8) or (10.9) will normally be used to determine which states are most important. However, compared with the linear case they are not completely balanced. Though, if desired, it is possible to obtain a balanced realization using an additional change of coordinates as described by the following lemma. (The lemma is a minor extension of the lemma in Fujimoto and Tsubakino (2007) in the sense that it is proved that the solution exists and can be computed recursively.)

Lemma 10.2
Consider a model (10.4) which is assumed to be in input normal form of degree m. Then, using an m:th order coordinate transformation

$$z = \Psi^{m]}(q) = \big(\psi(q_1), \psi_2(q_2), \ldots, \psi_n(q_n)\big), \ \Psi(0) = 0 \tag{10.10}$$

where ψ_i, $i = 1, \ldots, n$ is given as the solution to

$$q_i = \psi_i(q_i)\Big(\tau_i^{m-1]}\big(\psi_i(q_i)\big)\Big)^{\frac{1}{4}}, \ \forall q_i \in \Omega \tag{10.11}$$

the controllability function and observability function can be written as

$$L_c\big(\Psi^{m]}(q)\big) = \frac{1}{2}\sum_{i=1}^{n}\frac{q_i^2}{\sigma_i^{m-1]}(q_i)} + \mathcal{O}(q)^{m+2} \tag{10.12a}$$

$$L_o\big(\Psi^{m]}(q)\big) = \frac{1}{2}\sum_{i=1}^{n}\sigma_i^{m-1]}(q_i)q_i^2 + \mathcal{O}(q)^{m+2} \tag{10.12b}$$

where the singular value functions are defined as

$$\sigma_i(q_i) \triangleq \sqrt{\tau_i^{m-1]}\big(\psi_i(q_i)\big)} \tag{10.13}$$

Proof: First, by substituting the expressions above into the controllability and observability functions in (10.8), it can be verified that the expressions in (10.12) are obtained.

Second, it needs to be shown that (10.11) has a well-defined solution. Consider the equation

$$0 = q_i - z_i\Big(\tau_i^{m-1]}(z_i)\Big)^{\frac{1}{4}},\ i = 1,\dots,d \tag{10.14}$$

It can be shown that (10.14) is satisfied for $(q_i, z_i) = 0$, and that the derivative of its right-hand side w.r.t. z_i is nonsingular (due to that $\tau_0 \neq 0$). Moreover, the right-hand side is a real analytic function. Therefore, the implicit function theorem ensures that (10.14) has a solution $z_i = \psi_i(q_i)$, where each ψ_i is described by a convergent power series.

Third, the solution can be computed recursively. Let $\psi_i(q_i) = \psi_i^{[1]}(q_i) + \psi_{ih}(q_i)$, $\tau_i^{d-1]}(z_i) = \tau_{i0} + \tau_{ih}^{d-1]}(z_i)$ and

$$(a+x)^{\frac{1}{4}} = a^{\frac{1}{4}} + \xi(a,x)$$

where $a > 0$ and ξ denote terms of at least order one. The equations (10.11) can then be written as

$$0 = q_i - \big(\psi_i^{[1]}(q_i) + \psi_{ih}(q_i)\big)\big(\tau_{i0}^{\frac{1}{4}} + \xi\big(\tau_{i0}, \tau_{ih}^{d-1]}(\psi_i(q_i))\big)\big),\ i = 1,\dots,n$$

The composite function $\xi(\cdot)$ is convergent, since it is the composition of three real analytic functions. From the terms of order one in the equations above, it follows that

$$\tau_{i0}^{\frac{1}{4}}\psi_i^{[1]}(q_i) = q_i$$

while a general term of order m gives the equations

$$\tau_{i0}^{\frac{1}{4}}\psi_{ih}^{[m]}(q_i) = -\psi_i^{[1]}(q_i)\left\{\xi(\tau_{i0}, \tau_{ih}^{d-1]})\right\}^{[m-1]} - \sum_{j=1}^{m-1}\psi_h^{[j]}(q_i)\left\{\xi(\tau_{i0}, \tau_{ih}^{d-1]})\right\}^{[m-j]}$$

where

$$\left\{\xi(\tau_{i0}, \tau_{ih}^{d-1]})\right\}^{[m-1]} = \xi^{[m-1]}\big(\tau_{i0}, \tau_{ih}^{m-1]}(\psi_i^{m-1]}(q_i))\big)$$

for $m = 1, 2, \dots$ up to the order of interest. Hence, the terms in the right-hand side only depend on lower orders of $\psi_i(q_i)$ which makes recursive computation possible.

Note that since the lowest order terms in $\psi_i(q)$ are of order 1, the order of the $\mathcal{O}$-terms in (10.8) will be unchanged.

Fourth, even though terms of ψ_i and σ_i up to an arbitrary order can be computed only those of order 1 through $m-1$ are needed. In L_o it is easily seen since σ_i is multiplied by q_i^2. However, also in L_c, where σ_i appears in the denominator, it is possible to motivate the choice of order. If the terms $1/\sigma_i$ are Taylor expanded, terms in σ_i of higher order than $m-1$ will end up in the $\mathcal{O}$-term. □

10.1.2 Approximation of the Model

Given a change of coordinates bringing the system to input normal form of degree m, the next step is to extract the parts of the system that contribute the most to the input-output behavior. There are a number of different choices.

Balanced Truncation

The easiest and most common method to approximate a model is to remove the part that corresponds to the smallest squared singular value functions (Scherpen, 1994; Fujimoto and Scherpen, 2001).

Therefore, let $x = \Phi^{m]}(z)$ be a coordinate transformation such that the transformed system is in input normal form of order m. Then under the given assumptions $\Phi^{m]}(z)$ is invertible, at least locally, and the transformed system can be written as

$$\dot{z} = \tilde{F}(z,u) \tag{10.15a}$$
$$y = \tilde{h}(z,u) \tag{10.15b}$$

where

$$\tilde{F}(z,u) = \Big(\frac{\partial \Phi^{m]}(z)}{\partial z}\Big)^{-1} F\big(\Phi^{m]}(z),u\big)$$
$$\tilde{h}(z,u) = h\big(\Phi^{m]}(z),u\big)$$

Partition the transformed system (10.15) into two parts

$$\tilde{F}(z,u) = \begin{pmatrix} \tilde{F}_a(z,u) \\ \tilde{F}_b(z,u) \end{pmatrix}$$

where $z = (z_a, z_b)$ with $z_a = (z_1, \ldots, z_k)$ and $z_b = (z_{k+1}, \ldots, z_n)$. The structure of $\tilde{F}(z,u)$ is

$$\tilde{F}(z,u) = \begin{pmatrix} A_{a,1} & A_{a,2} \\ A_{b,2} & A_{b,2} \end{pmatrix} \begin{pmatrix} z_a \\ z_b \end{pmatrix} + \begin{pmatrix} B_a \\ B_b \end{pmatrix} u + \tilde{F}_h(z,u)$$
$$\tilde{h}(z,u) = \begin{pmatrix} C_a & C_b \end{pmatrix} \begin{pmatrix} z_a \\ z_b \end{pmatrix} + Du + \tilde{h}_h(z,u)$$

where $\tilde{F}_h$ and $\tilde{h}_h$ denote terms of at least order two.

Now, assume that the squared singular value functions are in order, *i.e.*,

$$\min_{z_i\in[-c,c]} \tau_i(z_i) > \max_{z_i\in[-c,c]} \tau_{i+1}(z_i), \quad i = 1, 2, \ldots, n$$

where c determines the range of states of interest, and that τ_k is substantially larger than τ_{k+1} for some k. Then

$$\dot{z}_a = \tilde{F}(z_a, 0, u), \quad y = \tilde{h}(z_a, 0, 0) \tag{10.16}$$

is a k:th order reduced model of (10.4).

The reduced order model locally preserves several important properties of the original system, such as controllability, observability and stability as proved in the following lemma.

Lemma 10.3
Consider the state-space model (10.4) and assume that it satisfies Assumption A17 and A18. Let $x = \Phi^{d]}(z) = Tz + \Phi_h^{d]}(z)$ be a change of coordinates that brings (10.4) to input normal form of degree m. Then the following properties hold

- *$A_{a,1}$ and $A_{b,2}$ are Hurwitz matrices.*
- *$(A_{a,1}, B_a)$ and $(A_{b,2}, B_b)$ are controllable.*
- *$(A_{a,1}, C_a)$ and $(A_{b,2}, C_b)$ are observable.*

independently of the partition.

Proof: The proof follows easily by first noting that T is determined by the linearization of (10.4) and then using the results in Pernebo and Silverman (1982). □

Note that the properties above are not true for general full rank matrices T. Even though the transformed system (10.15) will have the properties mentioned above as a whole, the parts will not necessarily have them.

Let the controllability and observability functions, computed for the reduced model, be denoted $L_{ca}(z_a)$ and $L_{oa}(z_a)$. Then these will be the same as if $z = (z_a, 0)$ is substituted into $L_c\big(\Phi(z)\big)$ and $L_o\big(\Phi(z)\big)$. The same holds symmetrically for the z_b. This is formalized in the following lemma which is an extension of a result in Fujimoto and Tsubakino (2006) to the case when only finite series are considered.

Lemma 10.4
Consider the system. Then

$$\begin{aligned}
L_{ca}(z_a) &= L_c\big(\Phi^{m]}(z_a, 0)\big) + \mathcal{O}(z_a)^{m+2}, & L_{oa}(z_a) &= L_o\big(\Phi^{m]}(z_a, 0)\big) + \mathcal{O}(z_a)^{m+2} \\
L_{cb}(z_b) &= L_c\big(\Phi^{m]}(0, z_b)\big) + \mathcal{O}(z_b)^{m+2}, & L_{ob}(z_b) &= L_o\big(\Phi^{m]}(0, z_b)\big)\mathcal{O}(z_b)^{m+2} \\
u_{*a}(z_a) &= u_*\big(\Phi(z_a, 0)\big) + \mathcal{O}(z_a)^{m+1}, & u_{*b}(z_b) &= u_*\big(\Phi(0, z_b)\big) + \mathcal{O}(z_b)^{m+1}
\end{aligned}$$

That is, the controllability and observability functions for the reduced system are the same up to order $m + 1$ as for the original system with the coordinate transformation $\Phi^{m]}(x)$ which transforms the system to input normal form of degree m.

Proof: First consider the observability function case. The key property of the observability function in input normal form of degree m is that

$$L_{o;z_i}\big(\Phi^{m]}(z)\big) = \tfrac{1}{2}\tau_{i;z_i}^{m-1]}(z_i)z_i^2 + \tau_i^{m-1]}(z_i)z_i + \tfrac{\partial}{\partial z_i}\mathcal{O}(z)^{m+2},\ i = 1,\dots,d$$

In a sufficiently small neighborhood $\mathbb{U}$ of the origin, the term with z_i is dominating and since $\tau_i^{m-1]}(z_i)$ has a non-zero constant term, the conclusion is that

$$L_{o;z_i}\big(\Phi^{m]}(z)\big) = \mathcal{O}(z)^{m+1}, \quad z_i \to 0 \tag{10.17}$$

Otherwise, it is $\mathcal{O}(z)$.

For the non-transformed system, the observability function is given as the solution to

$$0 = L_{o;x}(x)F(x,0) + \tfrac{1}{2}h(x,0)^T h(x,0) \tag{10.18}$$

The equation above has a solution for arbitrary $x \in \Omega$ and then specifically for $x = \Phi^{m]}(z)$ where $z \in \Omega_z$. If this fact is used together with the following relation, obtained using the chain rule,

$$\frac{\partial L_o(\Phi^{m]}(z))}{\partial z} = \frac{\partial L_o(x)}{\partial x}(\Phi^{m]}(z))\frac{\partial \Phi^{m]}(z)}{\partial z}$$

it follows that

$$0 = L_{o;z}\big(\Phi^{m]}(z)\big)\tilde{F}(z,0) + \tfrac{1}{2}\tilde{h}(z,0)^T h(z,0) \tag{10.19}$$

Assume that (10.19) is evaluated in $z = (z_a, 0)$. Then, using (10.17), it follows that

$$0 = \frac{\partial L_o(\Phi^{m]}(z))}{\partial z_a}(z_a,0)\tilde{F}_a(z_a,0,0) + \tfrac{1}{2}\big|h(\Phi^{m]}(z_a,0),0)\big|^2 + \mathcal{O}(z_a)^{m+2} \tag{10.20}$$

Note that the expression above also holds symmetrically for $z = (0, z_b)$.

The reduced model is obtained by letting $z = (z_a, 0)$ and picking out the upper part described by $\tilde{F}(z_a, 0, 0)$. The equation for calculating the observability function for the reduced system becomes

$$0 = L_{oa;z_a}(z_a)\tilde{F}_a(z_a,0,0) + \tfrac{1}{2}h(\Phi^{m]}(z_a,0),0)^T h(\Phi^{m]}(z_a,0),0) \tag{10.21}$$

This is the same equation as (10.20) up to order $m+1$. From Lemma 10.3, it is known that the local properties of the reduced system is such that (10.21) has a unique solution, and therefore it must hold that

$$L_{oa}(z_a) = L_o\big(\Phi(z_a,0)\big) + \mathcal{O}(z_a)^{m+2}$$

The same property can then be shown for $L_{ob}(z_b)$.

For the controllability function, the same steps can be performed, but in this case with the equations

$$0 = L_{c;x}(x)F(x,u_*(x)) - \tfrac{1}{2}u_*(x)^T u_*(x) \tag{10.22a}$$

$$0 = L_{c;x}(x)F_u\big(x,u_*(x)\big) - \tfrac{1}{2}u_*(x)^T u_*(x) \tag{10.22b}$$

□

Balanced Truncation based on the Co-Observability Function

In the last section the most common method for approximating the system was described. However, in Krener (2008) another approach is presented. The main motivation starts with the question: What is minimized with balanced truncation? To explain this, introduce a projection defined by the submersion ϕ and the embedding ψ as

$$q = \phi(z), \quad z = \psi(q)$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^k$ and $\psi : \mathbb{R}^k \to \mathbb{R}^d$ such that $\phi\big(\psi(q)\big) = q$ and $(\psi \circ \phi)^2(z) = \psi \circ \phi(z)$.

The corresponding reduced order model is then given by

$$\dot{q} = \bar{f}(q, u) \tag{10.23a}$$
$$y = \bar{h}(q, u) \tag{10.23b}$$

where

$$\bar{f}(q, u) = \frac{\partial \phi(z)}{\partial z}(\psi(q))\tilde{f}\big(\psi(q), u\big)$$
$$\bar{h}(q, u) = \tilde{h}\big(\psi(q), u\big)$$

First, consider the linear case. In balanced truncation, the projection is given by $\phi(z) = T_\phi z$ and $\psi(q) = T_\psi q$, where

$$T_\phi = T_\psi^T = \begin{pmatrix} I & 0 \end{pmatrix} \tag{10.24}$$

Intuitively, to obtain a good reduced model of order k, T_ψ should be chosen such that the states in its range maximize $L_o(z)$ when $L_c(z) = c$ with c given. This means that if the Hankel singular values are sorted in descending order, the range should satisfy $z_k = z_{k+1} = \ldots = z_d = 0$. A convenient choice is then (10.24).

The choice of T_ϕ can be motivated by studying the norm of the difference between the outputs starting from z and $T_\psi T_\phi z$. Therefore, define the co-observability function for a general nonlinear model (10.15) as

$$L_{oo}(z, q) = \frac{1}{2} \int_0^\infty (y(t) - \bar{y}(t))^T (y(t) - \bar{y}(t))\, dt \tag{10.25}$$

where y and $\bar{y}$ are the outputs of the nonlinear system (10.15) starting from z and $\bar{z}$, respectively, with $u(t) = 0$.

The co-observability is calculated from the equation (cf. with the ordinary observability function)

$$0 = \begin{pmatrix} \frac{\partial L_{oo}(z,\bar{z})}{\partial z} & \frac{\partial L_{oo}(z,\bar{z})}{\partial z} \end{pmatrix} \begin{pmatrix} \tilde{F}(z, 0) \\ \tilde{F}(\bar{z}, 0) \end{pmatrix} + \frac{1}{2} \big\| h(z, 0) - h(\bar{z}, 0) \big\|^2 \tag{10.26}$$

The following lemma gives conditions under which (10.26) has a unique solution.

Lemma 10.5
Consider a nonlinear state-space model (10.15) in input normal form of degree m. Assume that its linearization is asymptotically stable and observable. Then there exists a unique solution to (10.26) given by

$$L_{oo}(z,\bar{z}) = \frac{1}{2}\sum_{i=1}^{k}\tau_{i0}(z-\bar{z})^2 + L_{oo,h}^{m]}(z,\bar{z}) + \mathcal{O}(z,\bar{z})^{d+1} \tag{10.27}$$

where $L_{oo,h}(z,\bar{z})$ denotes terms of at least order three.

Note that it is not necessary for the existence and uniqueness of a solution to (10.26) that the system is in input normal form of degree m. However, then the second order terms in (10.27), which in the general case is given by $\frac{1}{2}(z-\bar{z})^T Q(z-\bar{z})$ where

$$A^T Q + QA = -C^T C$$

do not need to be diagonal and have the squared Hankel singular values on the diagonal. In the linear case, $L_{oo}(z,\bar{z})$ will be quadratic and the choice of T_ϕ that minimizes the norm of the difference between the two different outputs is therefore given by (10.24). Hence, the projection is chosen such that the extra energy in the output that is obtained by starting outside the projected space is minimized.

Inspired by the linear case, a nonlinear extension is derived in Krener (2008). Again, one would like to choose ψ such that its range is the k-dimensional submanifold on which $L_o(z)$ is maximized when $L_c(z) = c$ for given values of c. However, this is not a well-defined submanifold unless $k = 1$ (*i.e.*, the same number as c).

However, an approximate optimal choice is $z_{k+1} = \ldots = z_d = 0$, that is,

$$z = \psi(q) = (q_1, \ldots, q_k, 0, \ldots, 0) \tag{10.28}$$

The submersion ϕ is chosen to minimize $L_{oo}\big(z, \psi(\phi(z))\big)$ which gives

$$\phi_i(z_i) = z_i - \frac{1}{\tau_i}\frac{\partial L_{oo,h}(z,\bar{z})}{\partial \bar{z}_i}(z, (\phi_i(z_i), 0)), \quad i = 1, \ldots, k \tag{10.29}$$

In many cases, this method may give reduced order models with an input-output behavior more similar to the original model than the standard balanced truncation method. However, there are three issues. First, there is not guarantee that the obtained controllability and observability functions are the restricted version of the original model, as in standard balanced truncation. Second, more computations are required. Third, the static gain need not be preserved, which is a disadvantage this method shares with standard balanced truncation.

Residualization

In residualization, the time derivatives of the less important states are assumed to be zero. The main advantage of this method is that it gives better low frequency behavior, for example, the correct static gain.

Again, consider a model in input normal form of degree m (10.15) and assume there is a gap in the squared singular value functions for some k. Then, by letting the last $n-k$ derivatives of the states be zero, the reduced model will be the following nonlinear DAE model.

$$\begin{pmatrix} \dot{z}_a \\ 0 \end{pmatrix} = \begin{pmatrix} A_{a,1} & A_{a,2} \\ A_{b,1} & A_{b,2} \end{pmatrix} \begin{pmatrix} z_a \\ z_b \end{pmatrix} + \begin{pmatrix} B_a \\ B_b \end{pmatrix} u + \tilde{F}_h(z_a, z_b, u) \tag{10.30a}$$

$$y = \begin{pmatrix} C_a & C_b \end{pmatrix} \begin{pmatrix} z_a \\ z_b \end{pmatrix} + Du + \tilde{h}_h(z_a, z_b, u) \tag{10.30b}$$

Nonlinear residualization is a rather well-studied method since it is the limit case for singular perturbation, see Kokotović et al. (1986), but model reduction of such systems is mostly studied for particular classes of systems such as chemical system, see Vora and Daoutidis (2001); Schneider and Wilhelm (2000). Results for more general systems are more unusual but some can for example be found in Hahn (2003); Sun and Hahn (2005). The obtained model will in this case have the same number of equations as the original model. The only difference is that some of the differential equations have been converted into algebraic equations.

For the reduced model (10.30) a similar lemma to Lemma 10.3 can be proved. However, except for the standard properties, an important feature concerning the solvability is also shown.

Lemma 10.6
Consider the state-space model (10.4) and assume that it satisfies Assumption A17 and A18. Let $x = \Phi^{m]}(z) = Tz + \Phi_h^{m]}(z)$ be a coordinate change obtaining input normal form of degree m.

Then, the reduced model (10.30) will satisfy Assumption A7. Furthermore, its linearization will be asymptotically stable, R-controllable and R-observable.

Proof: The linear part of the transformation $\Phi^{m]}(z)$, *i.e.*, T, is determined by the linearization of (10.4). The linear part of (10.30) will therefore be the same as obtained in Liu and Anderson (1989), where the three latter properties are shown.

To prove the first property, first note that $\tilde{F}(0,0) = 0$, since $\Phi^{m](0)} = 0$ and $F(0,0) = 0$. It is also known from Liu and Anderson (1989) that $A_{b,2}$ is Hurwitz, which means it is nonsingular as well, and the conditions in Assumption A7 are satisfied. □

A general definition of R-controllability and R-observability can be found in Dai (1989), but for the linearization of a strangeness-free semi-explicit model as in (10.30), it simplifies to $(\hat{A}, \hat{B}) = (A_{11} - A_{12}A_{22}^{-1}A_{21},\ B_1 - A_{12}A_{22}^{-1}B_2)$ being controllable and observable.

An interesting feature, which at least numerical computations seems to confirm, is that the controllability and observability functions for the residualized model become the restricted versions of the corresponding functions for the full system. If so, it is in accordance with the linear case as can be seen in for example Liu and Anderson (1989); Skogestad and Postlethwaite (2001).

10.2 Model Reduction of DAE Models

Now the focus is turned to nonlinear DAE models

$$\hat{F}_1(\dot{x}_1, x_1, x_3, u) = 0 \tag{10.31a}$$
$$\hat{F}_2(x_1, x_3, u) = 0 \tag{10.31b}$$
$$y - h(x_1, x_3, u) = 0 \tag{10.31c}$$

From earlier chapters, it is known that if the considered model satisfies Assumption A7, there is a neighborhood around the origin in which the DAE model is equivalent to

$$\dot{x}_1 = \mathcal{L}(x_1, u) \tag{10.32a}$$
$$y = \hat{h}(x_1, u) = h\big(x_1, \mathcal{R}(x_1, u), u\big) \tag{10.32b}$$

Since this is a state-space model, it is from a theoretical point of view possible to use the results in Section 10.1. However, as in the former chapters, the challenge comes from the fact that $\mathcal{L}$ and $\mathcal{R}$ normally cannot be written in closed form. Also for nonlinear state-space models, it is normally hard or even impossible to find solutions in closed form to the equations in Section 10.1.

Therefore, a similar computational method as shown earlier which only relies on power series expansions will be used find the transformation and reduce the DAE model. The method is based on the methods derived in Fujimoto and Tsubakino (2006) and Krener (2008).

10.2.1 Computing the Input Normal Form of Order m

The first step is to compute the balanced form. To obtain a well-posed problem the following assumption will be needed.

Assumption A19. The linearization of the DAE model (10.31) is asymptotically stable, R-controllable and R-observable.

The first step in order to find the input normal form of degree m is to compute the controllability and observability functions up to some desired order. The key point is that even though closed expressions for $\hat{F}$ and $\hat{h}$ are not known, their series expansions can be computed locally in a neighborhood of the origin and the functions will be expressed as power series, see Chapter 8 and 9, respectively, for details. If the DAE model (10.31) satisfies the required assumptions, it is known from Theorem 8.2 and 9.2 that these functions can be computed to an arbitrary order as

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

The next step is to let the transformation and the squared singular value functions be expressed as power series of order m and $m-1$, respectively,

$$x_1 = \Phi^{m]}(z) = T_\Phi z + \Phi_h^{m]}(z) \tag{10.33a}$$

$$\tau_i^{m]}(z_i) = \tau_{i,0} + \tau_{i,h}(z_i)^{m-1]}, \quad i = 1, \ldots, d \tag{10.33b}$$

If $\Phi^{m]}(z)$ and $\tau_i^{m-1]}(z_i)$ in (10.33) are substituted into the controllability and observability functions the result should be the input normal form in (10.8). The equations formed then become

$$\begin{aligned} 0 &= z^T(T_\Phi^T G_c T_\Phi - I)z + z^T T_\Phi^T G_c \Phi_h^{m]}(z) + \Phi_h^{m]}(z)^T G_c \Phi^{m]}(z) \\ &\quad + 2L_c\big(\Phi^{m]}(z)\big) + \mathcal{O}(z)^{m+2} \end{aligned} \tag{10.34a}$$

$$\begin{aligned} 0 &= z^T\big(T_\Phi^T G_o T_\Phi - \tau(z)\big)z + z^T T_\Phi^T G_o \Phi_h^{m]}(z) + \Phi_h^{m]}(z)^T G_o \Phi^{m]}(z) \\ &\quad + 2L_o\big(\Phi^{m]}(z)\big) + \mathcal{O}(z)^{m+2} \end{aligned} \tag{10.34b}$$

where

$$\tau(z) = \operatorname{diag}\big(\tau_1(z_1), \tau_2(z_2), \ldots, \tau_d(z_d)\big)$$

Since these equations are supposed to be valid for all z in a neighborhood, the coefficients corresponding to different orders in z must equal zero. The second order terms yield the equations

$$G_c^{-1} G_o T_\Phi = T_\Phi \operatorname{diag}(\tau_{1,0}, \tau_{2,0}, \ldots, \tau_{d,0}) \tag{10.35a}$$

$$T_\Phi^T G_c T_\Phi = I \tag{10.35b}$$

and solving these equations then give the zeroth order terms of $\tau_i^{m-1]}$ and the first order terms of $\Phi^{m]}$. The obtained $\tau_{i,0}$, $i = 1, \ldots, d$ and T_Φ become unique for a given representation of a system. However, T_Φ is not similarity invariant, *i.e.*, it may change if a different coordinate system is used.

The higher order terms of $\Phi(z)$ and $\tau_i(z_i)$ of order m and $m-1$, respectively, are obtained from the terms in (10.34) of order $m+1$ as

$$\Phi_h^{[m]}(z)^T G_c T_\Phi z = -\frac{1}{2}\sum_{i=1}^{m} \Phi_h^{[m-i]}(z)^T G_c \Phi_h^{[i]}(z) - \Big\{ L_{ch}^{m+1]}\big(\Phi^{m-1]}(z)\big)\Big\}^{[m]} \tag{10.36a}$$

and

$$\begin{aligned} z^T T_\Phi^T G_o \Phi_h^{m]}(z) + z^T \tau_h^{m-1]}(z) z &= -\sum_{i=1}^{m} \Phi_h^{[m-i]}(z)^T G_c \Phi_h^{[i]}(z) \\ &\quad - 2\Big\{ L_{oh}^{m+1]}\big(\Phi^{m-1]}(z)\big)\Big\}^{[m]} - \frac{1}{2} z^T \tau_h^{m-1]}(z) z \end{aligned} \tag{10.36b}$$

where

$$\tau_h(z) = \operatorname{diag}\big(\tau_{1,h}(z_1), \ldots, \tau_{d,h}(z_d)\big)$$

As can be seen from (10.36), the system of equations will be linear in the unknown coefficients if the equations are solved recursively one order at a time.

In Krener (2008), it is shown that the system of equations (10.36) has at least one solution, and in most cases there will be many because the set of equations is under-determined. However, in Krener (2008) it is also shown that (10.36) will have a unique solution for m up to 6 (or 12 if the system is odd, *i.e.*, $f(-x,-u) = -f(x,u)$ and $h(-x,-u) = -h(x,u)$). Another case that also yields unique solutions is for $d \leq 2$, as shown in Fujimoto and Tsubakino (2007).

Remark 10.1. It might seem like Φ_h is uniquely determined by (10.36), since $G_c T_\Phi$ is nonsingular (compare with the discussion in for example Lukes (1969)). However, this is not the case since it will be the m:th order terms of (10.36) that determine the terms in Φ of order $m-1$. Therefore, the system of equations will normally be under-determined.

If a balanced form of order m is wanted, the next step is to obtain the coordinate change $\Psi^{m]}$. The computations are described in the proof of Lemma 10.2. The terms in Ψ are computed recursively using (10.11), and the singular value functions $\sigma_i(q_i)$ can then be computed from (10.13) up to order $m-1$.

10.2.2 Balanced Truncation

Given the change of coordinates, it is now possible to express the reduced nonlinear DAE model. Consider the semi-explicit model

$$\dot{x}_1 = F_1(x_1, x_3, u) \tag{10.37a}$$
$$0 = F_2(x_1, x_3, u) \tag{10.37b}$$
$$y = h(x_1, x_3, u) \tag{10.37c}$$

and assume that the first k squared singular value functions are significantly larger than the other $d-k$. Divide the states z into two subspaces z_a and z_b, where the former corresponds to the larger singular values and form the subsystems

$$\dot{z}_a = \bar{F}_{1a}(z_a, 0, x_3, u) \tag{10.38a}$$
$$0 = \bar{F}_2(z_a, 0, x_3, u) \tag{10.38b}$$
$$y_a = \bar{h}(z_a, 0, x_3) \tag{10.38c}$$

and

$$\dot{z}_b = \bar{F}_{1b}(0, z_b, x_3, u) \tag{10.39a}$$
$$0 = \bar{F}_2(0, z_b, x_3, u) \tag{10.39b}$$
$$y_b = \bar{h}(0, z_b, x_3) \tag{10.39c}$$

where $\bar{F}_{1a}$ and $\bar{F}_{1b}$ correspond to the first k rows and the last $d-k$ rows, respectively, of

$$\bar{F}_1(z_{1a}, z_{1b}, x_3, u) = \left(\frac{\partial \Phi^{m]}(z)}{\partial z}\right)^{-1} F_1\big(\Phi^{m]}(z), x_3, u\big)$$

and $\bar{F}_2$ and $\bar{h}$ are given by

$$\bar{F}_2(z_{1a}, z_{1b}, x_3, u) = F_2\big(\Phi^{m]}(z), x_3, u\big)$$
$$\bar{h}(z_{1a}, z_{1b}, x_3) = h\big(\Phi^{m]}(z), x_3\big)$$

A reduced order model of order k of the DAE model (10.37) is then given by (10.38). The subsystems will both locally retain some of the important properties of the original DAE model.

Lemma 10.7
Consider a nonlinear DAE model (10.37) and assume that it satisfies Assumption A7 and A19. Let $x_1 = \Phi^{d]}(z)$ be obtained using the procedure in Section 10.2.1. Then both (10.38) and (10.39) will satisfy Assumption A7 and A19.

Proof: The subsystems (10.38) and (10.39) are calculated from

$$\dot{z}_a = \hat{\bar{F}}_{1a}(z_a, u) = \bar{F}_{1a}\big(z_a, 0, \bar{\mathcal{R}}(z_a, 0, u), u\big) \tag{10.40a}$$
$$y_a = \hat{\bar{h}}(z_a, u) = \bar{h}\big(z_a, 0, \bar{\mathcal{R}}(z_a, 0, u), u\big) \tag{10.40b}$$
$$\tag{10.40c}$$

and

$$\dot{z}_b = \hat{\bar{F}}_{1b}(z_b, u) = \bar{F}_{1b}\big(0, z_b, \bar{\mathcal{R}}(0, z_b, u), u\big) \tag{10.40d}$$
$$y_b = \hat{\bar{h}}(z_b, u) = \bar{h}\big(0, z_b, \bar{\mathcal{R}}(0, z_b, u), u\big) \tag{10.40e}$$

where $\bar{\mathcal{R}}(z, u) = \mathcal{R}\big(\Phi^{m]}(z), u\big)$. The system in (10.40) is a state-space model. The linear part of the transformation, *i.e.*, T_Φ, is determined from its linearization and will make the linear part of each subsystem in (10.40) controllable, observable and asymptotically stable. Since the solvability properties of the constraints (10.37b) are not changed locally by the change of coordinates, $\mathcal{R}\big(\Phi(z), u\big)$ can be replaced by (10.38b) and (10.39b). □

Using the same discussion as in the last lemma, the following theorem can be proved.

Theorem 10.4
Consider a nonlinear DAE model (10.37), satisfying Assumptions A7 and A18. Then it holds that

$$L_{ca}(z_a) = L_c\big(\Phi^{m]}(z_a, 0)\big) + \mathcal{O}(z_a)^{m+2}, \quad L_{oa}(z_a) = L_o\big(\Phi^{m]}(z_a, 0)\big) + \mathcal{O}(z_a)^{m+2}$$
$$L_{cb}(z_b) = L_c\big(\Phi^{m]}(0, z_b)\big) + \mathcal{O}(z_b)^{m+2}, \quad L_{ob}(z_b) = L_o\big(\Phi^{m]}(0, z_b)\big) + \mathcal{O}(z_b)^{m+2}$$

Proof: Follows similarly as in Lemma 10.4. □

Balanced Truncation based on the Co-Observability Function

The method in Section 10.1.2 can be applied straightforwardly by expressing the system in terms of the series expansion of the underlying state-space model.

10.2.3 Residualization

Consider the residualization method instead. Assume that the DAE model is given by the general model (10.31). Then a k:th order reduced model for the DAE model can be

written as

$$0 = \hat{F}_1(\Phi^{m]}_{z_a}(z)z_a, 0, \Phi^{m]}(z), x_3, u) \tag{10.41a}$$
$$0 = \hat{F}_2(\Phi^{m]}(z), x_3, u) \tag{10.41b}$$
$$y = h(\Phi^{m]}(z), x_3, u) \tag{10.41c}$$

For this model the following lemma can be proved.

Lemma 10.8
Consider the reduced model (10.41). This system is locally asymptotically stable and satisfies Assumption A7 and A19.

Proof: The proof follows the same line as the proof for Lemma 10.7. The matrix T_Φ is derived based on the underlying state-space model. The solvability properties do not change when $\dot{z}_b$ is set to zero, which proves the result. □

The main advantage of the residualized model is the improved low frequency properties as will be seen in Section 10.3. However, for some models, the approximation and the simulation properties are better if the model, on which the coordinate transformations are applied, is the exact model. That is, the coordinate change $\Phi^{m]}(z)$ is computed from a truncated version of the system, but $\Phi^{m]}(z)$ is then substituted into the non-truncated expressions. If the system then is fully implicit, *i.e.*, also in $\dot{x}_1$, the residualized model is the best choice.

In summary, the steps need to be performed in order to obtain the reduced order model can be seen in Algorithm 10.1.

10.3 Example

In this section the methods are exemplified on a model of a double pendulum. The input torque τ is applied to the top joint and the output is the horizontal displacement of the bottom. The mass of the pendulums are m_1 and m_2, respectively, and the mass of each pendulum is concentrated to the ball. That is, the rod of each pendulum is massless. The length of the pendulums are L_1 and L_2, respectively. To make the system passive, a viscous damping with coefficient b_1 and b_2, respectively, are introduced in the two joints.

The coordinates $x_i, i = 1, \ldots, 4$ are defined as

$$x_1 = L_1 \sin(\theta_1) \qquad x_2 = L_1 - L_1 \cos(\theta_1)$$
$$x_3 = L_2 \sin(\theta_1 + \theta_2) \qquad x_4 = L_2 - L_2 \cos(\theta_1 + \theta_2)$$

where θ_1 is the angle from the vertical axis to pendulum 1 and $\theta_1 + \theta_2$ is the angle from the same axis to pendulum 2 (see Figure 10.1). The choice of putting the origin in the point furthest down follows from the fact that the origin should be a stationary point. The output signals are chosen as the x and y position of the second mass, *i.e.*,

$$y_1 = x_1 + x_3$$
$$y_2 = x_2 + x_4$$

Algorithm 10.1 Computation of the reduced order model.

1. Compute the controllability and observability functions up to order $m+1$ by solving the equations in Section 8.3 and 9.3, respectively. (In this step, the power series expansion of the DAE model is also obtained).
2. Solve the set of equations (10.35) and (10.36) to obtain $\Phi^{m]}$ and $\tau^{m-1]}$, to transform the system into input normal form of degree m. (The eigenvalue problem is solved by first solving (10.35a) and then normalize T_Φ to satisfy (10.35b))
3. Examine the squared singular value functions $\tau_i^{d-1]}(z_i)$ to see if there is a gap for some k over the range of interest of the states x_1.
4. Choose approximation method.
   - Standard balanced truncation:
     Remove the $d-k$ last states as seen in (10.38).
   - Balanced truncation based on the co-observability function:
     Define the embedding ψ as in (10.28). Compute the submersion ϕ by solving (10.26) up to order m and formulate the approximate model as (10.23).
   - Balanced Residualization:
     Set the derivatives of the last $d-k$ states to zero, as can be seen in (10.41).

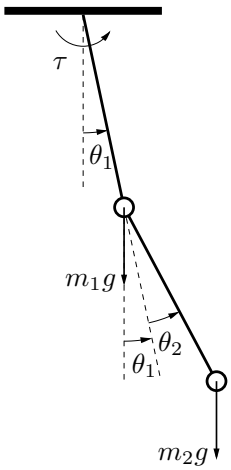


Figure 10.1: *A double pendulum with friction in the joints and a torque in the first joint as control input.*

The pendulum is modeled using Euler-Lagrange theory, see Goldstein (1980). The output from such a modeling process is naturally a DAE model which fits into the structure described in Chapter 6. First, the kinetic energy of the double pendulum is expressed as

$$T = \frac{m_1}{2}(\dot{x}_1^2 + \dot{x}_2^2) + \frac{m_2}{2}\big((\dot{x}_1 + \dot{x}_3)^2 + (\dot{x}_2 + \dot{x}_4)^2\big) \tag{10.42}$$

Second, the potential energy becomes

$$V = m_1 g(x_2 + L_2) + m_2 g(x_2 + x_4) \tag{10.43}$$

where the zero-level has been placed at the lowest vertical position. That is, both pendulums hanging straight down. Third, the friction terms are modeled using the Rayleigh function

$$F_{\text{rayleigh}} = \frac{1}{2}b_1\Big(\frac{\dot{x}_1}{L_1 - x_2}\Big)^2 + \frac{1}{2}b_2\Big(\frac{\dot{x}_3}{L_2 - x_4} - \frac{\dot{x}_1}{L_1 - x_2}\Big)^2 \tag{10.44}$$

The two terms correspond to the energy dissipated in the first and second joint, respectively. In this case, the constraints are given by

$$G(x) = \begin{pmatrix} x_1^2 + (L_1 - x_2)^2 - L_1^2 \\ x_3^2 + (L_2 - x_4)^2 - L_2^2 \end{pmatrix} \tag{10.45}$$

since x_i,$i = 1, \ldots, 4$ are local coordinates for each pendulum.

From the equations above, the Lagrangian is formulated as follows

$$L = T - V - (\lambda_1 + \frac{(m_1 + m_2)g}{2L_1})G_1(x) - (\lambda_2 + \frac{m_2 g}{2L_2})G_2(x) \tag{10.46}$$

where $G_1(x)$ and $G_2(x)$ denote the first and second row of $G(x)$, respectively, and λ_1 and λ_2 are the corresponding Lagrange multipliers. The DAE model is then found from the Euler-Lagrange equation (Goldstein, 1980)

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}_i} - \frac{\partial L}{\partial x_i} + \frac{\partial F_{\text{rayleigh}}}{\partial \dot{x}_i} = f_{\text{ext}}(x, u),\ i = 1, \ldots, 4 \tag{10.47}$$

where f_{ext} corresponds to external forces and torques, *i.e.*, the input torque τ, that act on the system. In this case, f_{ext} becomes

$$f_{\text{ext}}(x, u) = \begin{pmatrix} \frac{1}{L_1\sqrt{1 + \frac{x_1^2}{(L_1 - x_2)^2}}}u \\ \frac{x_1}{L_1(L_1 - x_2)\sqrt{1 + \frac{x_1^2}{(L_1 - x_2)^2}}}u \\ 0 \\ 0 \end{pmatrix}$$

The result of (10.47) is then a model in the multibody form (6.41),

$$\begin{aligned} \dot{p} &= q \\ M(p)\dot{q} &= f_{\text{ext}}(p, u) + \lambda^T g_p(p) \\ G(p) &= 0 \end{aligned}$$

where $p = x$ and q are the corresponding velocities. The partial derivatives of the constraints, *i.e.*, $G_p(p)$, are nonsingular, and (6.41) can then be shown to satisfy Hypothesis 2.2 with $\mu = 2$, $d = 4$, $a = 6$ and $\nu = 0$. Let x be divided as $p_1 = (x_1, x_3)$ and $p_2 = (x_2, x_4)$, which means that p_2 can be solved from $G(p) = 0$ (note that another possible choice is $p_1 = (x_2, x_4)$ and $p_2 = (x_1, x_3)$). The strangeness-free model locally becomes

$$\begin{aligned}
\dot{p}_1 &= q_1 \\
\dot{q}_1 &= \begin{pmatrix} I & 0 \end{pmatrix} M(p)^{-1} \big(f_{\text{ext}}(p,u) + G_p(p)^T \lambda \big) \\
0 &= G(p) \\
0 &= G_p(p) q \\
0 &= G_{pp}(p)(q,q) G_p(p) + G_p(p) M(p)^{-1} \big(f_{\text{ext}}(p,u) + G_p(p)^T \lambda \big)
\end{aligned}$$

with $\lambda = (\lambda_1, \lambda_2)$ and with the constants chosen as

$$g = 9.8,\ m_1 = 1,\ m_2 = 2,\ L_1 = 1,\ L_2 = 1,\ b_1 = 1,\ b_2 = 2$$

Different reduced order models are derived using the procedure described in Algorithm 10.1. In all cases except for the plot showing the true outputs, the nonlinear DAE model has been described by power series of order 5. This makes it possible to write the DAE model as a state-space model. When the truncation methods are used it means that the reduced order model will be a state-space model, which can be solved using Maple. However, when residualization is used the resulting model is a DAE model (note that even in this case an extra step could have been performed to obtain a state-space model in power series form), which then is solved using Dymola.

In Figure 10.2, the singular value functions are shown. The figure shows that the two first singular value functions are substantially larger than the other two. Therefore, a reduced order model of order two is appropriate.

Figure 10.3 shows the first output y_1 with the input signal $u = \sin(2.5t)$ for five different systems. The five different systems are the following: the true model, a linear and a fifth order approximation obtained using standard balanced truncation, a fifth order balanced truncation approximation where the co-observability function has been minimized and finally a residualized model of order 5. As can be seen all the different models match the true output well.

The next figure, *i.e.*, Figure 10.4, shows the output signal y_2 for the same set of models with the same input signal as in the last figure. The main feature to see in this figure is that the output for the linear approximation is identically zero, while all the nonlinear approximations reproduce the true output rather accurately.

In Figure 10.5, output y_1 is shown again for all the models but with the input chosen as a unit step starting at $t = 0.1$. A feature with this input signal is that it tests the accuracy in stationarity. The five models are clustered into two groups. The two models obtained using balanced truncation and the linear approximation are all very similar and form the upper of the two visible curves, while the true model and the residualized model coincides and form the lower curve.

The same tendency as in Figure 10.5 can be seen in the last figure, *i.e.*, Figure 10.6. In this figure, the y_2 signal is shown for all the models but with the unit step as input signal.

The output of the linear model is similarly to Figure 10.4 equal to zero, while the other signals are divided into two groups. The lower curve is the two models obtained using balanced truncation, while the upper curve is the output from true and the residualized model.

The two latter figures show that also for nonlinear models (obtained as a truncated power series) the residualized model often has better low frequency behavior.

It might be good with two final remarks. In this case, the DAE model was expressed as truncated power series also in the simulations since the truncation of the DAE model did not change the output too much. However, experience shows that for the residualization method it is often quite good to let the system be non-approximated. That is, to substitute $x_1 = \Phi^{m](z)}$ into the DAE model without Taylor expanding the model first.

It is also very important that the basis, *i.e.*, the set of coordinates, is chosen in an appropriate way. This fact can change the behavior of the approximated model substantially. An intuitive explanation is if the coordinates are chosen such that the original model includes a lot of terms with infinite series expansions, the truncation of the system may remove a lot of information. In the pendulum case, a more natural set of coordinates is to choose the angles, but the improvements in the simulations are in this case small.

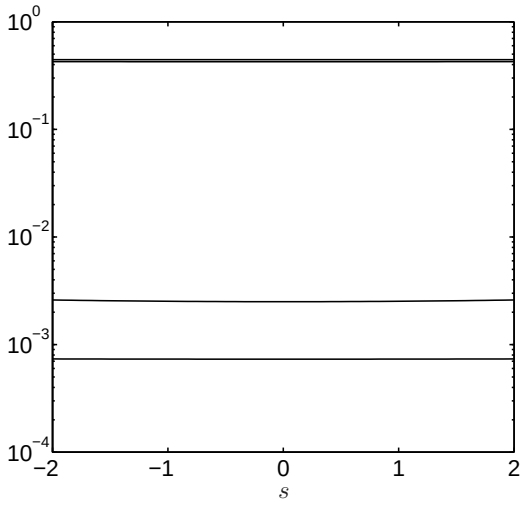


Figure 10.2: *The Hankel singular values σ_i, $i = 1, 2, 3, 4$, for the double pendulum.*

10.4 Conclusion

In this chapter it has been shown that for quite a large class of nonlinear DAE models it is possible to find a reduced model computationally. Further, it has been shown in an double pendulum example that the reduced nonlinear model can re-create the behavior of the true model not shown by the reduced model obtained from the linearization. However, it comes with a price. First, the computational complexity grows rather rapidly with the

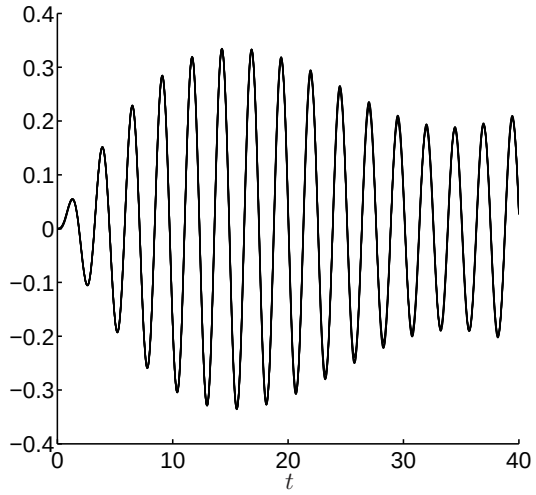


Figure 10.3: *The output y_1 corresponding to the x-position of the lower mass, when the input is the sinusoid $u(t) = \sin(2.5t)$. In the figure, both the true solution and all approximations are shown.*

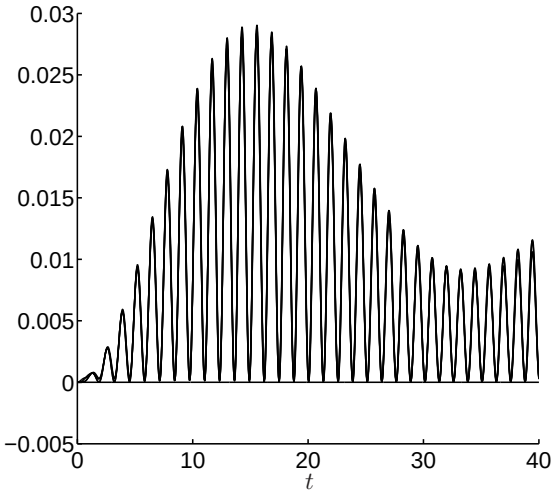


Figure 10.4: *The output y_2 corresponding to the y-position of the lower mass, when the input is the sinusoid $u(t) = \sin(2.5t)$. In the figure, both the true solution and all approximations are included.*

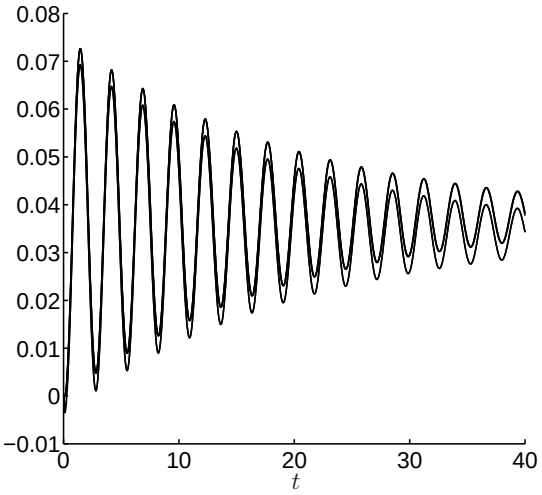


Figure 10.5: *The output y_1 corresponding to the x-position of the lower mass, when the input has been a unit step starting at $t = 0.1s$. In the figure, both the true solution and all approximations are included.*

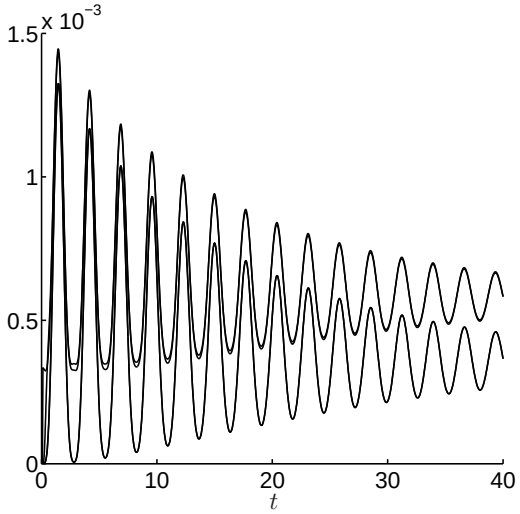


Figure 10.6: *The output y_2 corresponding to the y-position of the lower mass, when the input has been a unit step starting at $t = 0.1s$. In the figure, both the true solution and all approximations are included.*

number of variables in the model. This is a large issue. Second, the reduced model is expressed in terms of power series which means the reduced model will only be a good approximant locally around the expansion point.

11

Concluding Remarks

In this chapter some conclusions and some possible research openings found during the work are presented.

11.1 Conclusions

In this thesis, three major topics are studied. The foremost is optimal feedback control of nonlinear DAE models. It is shown that one possible method to deal with nonlinear DAE models is to use Taylor series expansions. Both time-invariant and time-varying models have been considered. Furthermore, it is shown that a discount factor can be introduced in both cases and the solution will remain to be a convergent series. However, a disadvantage with the power series solutions is that, when truncated, the obtained approximants tend to be accurate in a region that is rather small. Therefore, another approach based on rational approximants is presented. At least in a number of examples, these approximants approximate the optimal solution with higher accuracy over a larger region. A common issue with both the methods above is that the computational complexity is high and it increases rapidly with the size of the problem. Therefore, some conditions are derived under which the system has a structure that can be used to reduce the complexity.

The second topic that is covered is how white noise can be introduced into a nonlinear DAE model in a mathematical well-posed way. Some conditions are derived and for models that satisfy the conditions, it is shown how the variables in the model can be estimated.

The final topic is model reduction based on energy analysis. The main idea is to use the framework developed for optimal control, *i.e.*, power series. First, the controllability and observability functions are studied, since these functions are central in nonlinear model reduction. Then, it is shown how some methods for state-space models can be adapted to the DAE case.

11.2 Future Work

During the work on this thesis some possible research ideas have appeared. Most of them concern the power series method and the rational approximation method, described in Chapter 4 and 5, respectively. The reason is the lack of an explicit expression for the implicit function. However, one interesting question about the method in Section 3.2 is what conditions W_2 must satisfy in cases when the DAE model is not strangeness-free. The answer would probably increase the usability of the theorem in Xu and Mizukami (1993).

For the methods that rely on power series expansion there are numerous extensions. For example, it would be interesting to handle discrete-time DAE models or stochastic state-space and DAE models. Another extension would be to find the power series solution of the optimal feedback control problem for systems of higher strangeness index. In this case, the controller must be such that the closed-loop system becomes strangeness-free.

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 models other methods to find approximate optimal solutions exist, for instance, succesive Galerkin approximations. It would be interesting to study if these methods can be extended to handle DAE models as well.

Concerning model reduction, it would be nice to prove that the residualized model obtain the restricted versions of the controllability and observability functions for the full system. Furthermore, in the methods above, the dynamic part is reduced. However, it would be nice to reduce the number of algebraic equations as well.

Finally, it would be interesting to implement some of the methods in real life applications.

Appendices

A

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_δ is an open set around some point p with radius δ, *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 φ is k-times continuously differentiable as well. The function φ is often called a diffeomorphism. It means that φ is a continuously differentiable function between manifolds, with a continuously differentiable inverse. However, in some references the continuous differentiability is strengthened to $\mathcal{C}^\infty$.

Bibliography

E. G. Al'brekht. On the optimal stabilization of nonlinear systems. *Journal of Applied Mathematics and Mechanics*, 25(5):1254–1266, 1961.

B. D. O. Anderson and J. B. Moore. Linear system optimization with prescribed degree of stability. *IEE*, 116(12):2083–2087, 1969.

B. D. O. Anderson and J. B. Moore. *Linear Optimal Control*. Prentice Hall, Englewood Cliffs, New Jersey, 1971.

C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadić. Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, 92(3):423–438, March 2004.

M. Arnold, V. Mehrmann, and A. Steinbrecher. Index reduction in industrial multibody system simulation. Preprint 146, MATHEON, DFG Research Center, Berlin, 2004.

U. M. Ascher, C. Hongsheng, and S. Reich. Stabilization of DAE's and invariant manifolds. *Numerische Mathematik*, 67(2):131–149, 1994.

K. J. Åström. *Introduction to Stochastic Control Theory*. Mathematics in Science and Engineering. Academic Press, New York and London, 1970.

M. Bardi and I. Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Systems & Control: Foundations & Applications. Birkhäuser, Boston, 1997.

J. Baumgarte. Stabilization of constraints and integrals of motion in dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 1:1–16, 1972.

R. W. Beard, G. N. Saridis, and J. T. Wen. Approximative solutions to the time-invariant Hamilton-Jacobi-Bellman equation. *Journal of Optimization Theory and Applications*, 96(3):589–626, March 1998.

V. M. Becerra, P. D. Roberts, and G. W. Griffiths. Applying the extended Kalman filter to systems described by nonlinear differential-algebraic equations. *Control Engineering Practice*, 9:267–281, 2001.

R. E. Bellman. *Dynamic Programming*. Princeton University Press, New Jersey, 1957.

D. J. Bender and A. J. Laub. The linear-quadratic optimal regulator for descriptor systems. *IEEE Transactions on Automatic Control*, AC-32(8):672–688, 1987.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, Belmont, USA, 1995.

S. Bittanti, A. J. Laub, and J. C. Willems, editors. *The Riccati Equation*. Communications and Control Engineering. Springer-Verlag, Berlin, 1991.

K. E. Brenan, S.L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, New York, 1996.

A. E. Bryson and Y.-C. Ho. *Applied Optimal Control — Optimization, Estimation and Control*. Hemisphere Pub. Corp., Washington, 1975. Revised Printing.

S. L. Campbell. Linearization of DAE's along trajectories. *Zeitschrift für angewandte Mathematik und Physik*, 46(1):70–84, 1995.

S. L. Campbell and C. W. Gear. The index of general nonlinear DAEs. *Numerische Mathematik*, 72:173–196, 1995.

S. L. Campbell and E. Griepentrog. Solvability of general differential algebraic equations. *SIAM Journal on Scientific Computing*, 16(2):257–270, March 1995.

W. A. Cebuhar and V. Costanza. Approximation procedures for the optimal control of bilinear and nonlinear systems. *Journal of Optimization Theory and Applications*, 43 (4):615–627, 1984.

J. S. R. Chisholm. Rational approximants defined from double power series. *Mathematics of Computation*, 27(124):841–848, 1973.

J. D. Cobb. *Descriptor Variable and Generalized Singularly Perturbed Systems: A Geometric Approach*. PhD thesis, University of Illinois, Urbana, Illinois, 1980.

J. D. Cobb. Descriptor variable systems and optimal state regulation. *IEEE Transactions on Automatic Control*, AC-28(5):601–611, 1983.

E. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York, 1985.

A. Cuyt and L. Wuytack. *Nonlinear Methods in Numerical Analysis*. Elsevier Science Publishers, North-Holland — Asterdam, 1987.

L. Dai. *Singular Control Systems*. Lecture Notes in Control and Information Sciences. Springer-Verlag, Berlin, 1989.

M. Darouach, M. Boutayeb, and M. Zasadzinski. Kalman filtering for continuous descriptor systems. In *Proceedings of the 1997 American Control Conference*, pages 2108–2112, Albuquerque, New Mexico, June 1997. AACC.

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo methods in Practice*. Springer-Verlag, New York, 2001.

I. S. Duff and J. K. Reid. An implementation of Tarjan's algorithm for the block triangularization of a matrix. *ACM Transactions on Mathematical Software*, 4(2):137–147, June 1978.

C. Ebenbauer and F. Allgöwer. Computer-aided stability analysis of differential-algebraic equations. In *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems*, pages 1025–1029, Stuttgart, Germany, 2004.

P. Fritzson. *Principles of Object-Oriented Modeling and Simulation with Modelica 2.1*. Wiley-IEEE, New York, 2004.

K. Fujimoto and J. M. A. Scherpen. Model reduction for nonlinear systems based on the differential eigenstructure of Hankel operators. In *Proceedings of the 40th IEEE Conference on Decision and Control*, Orlando, Florida, December 2001.

K. Fujimoto and J. M. A. Scherpen. Singular value analysis of Hankel operators for general nonlinear systems. In *Proceedings of 7th European Control Conference*, Cambridge, UK, September 2003a.

K. Fujimoto and J. M. A. Scherpen. Nonlinear balanced realization based on singular value analysis of Hankel operators. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, Hawaii, December 2003b.

K. Fujimoto and J. M. A. Scherpen. Nonlinear input-normal realizations based on the differential eigenstructure of Hankel operators. *IEEE Transactions on Automatic Control*, 50(1):2–18, 2005.

K. Fujimoto and D. Tsubakino. On computation of nonlinear balanced realization and model reduction. In *Proceedings of the 2006 American Control Conference*, Minneapolis, Minnesota, June 2006.

K. Fujimoto and D. Tsubakino. Computation of nonlinear balanced realization and model reduction based on Taylor series expansion. *Systems & Control Letters*, 2007. Available online: doi:10.1016/j.sysconle.2007.08.015.

M. Gerdin. *Identification and Estimation for Models Described by Differential-Algebraic Equations*. PhD thesis, Department of Electrical Engineering, Linköpings universitet, nov 2006.

M. Gerdin and J. Sjöberg. Nonlinear stochastic differential-algebraic equations with application to particle filtering. In *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, California, December 2006.

A. Germani, C. Manes, and P. Palumbo. Kalman-Bucy filtering for singular stochastic differential systems. In *Proceedings of the 15th World Congress of IFAC*, Barcelona, Spain, July 2002.

T. Glad and J. Sjöberg. Hamilton-Jacobi equations for nonlinear descriptor systems. In *Proceedings of the 2006 American Control Conference*, Minneapolis, Minnesota, June 2006.

H. Goldstein. *Classical Mechanics*. Addison-Wesley, Reading, Massachusetts, 2nd edition, 1980.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 107–113, April 1993.

A. Graham. *Kronecker Products and Matrix Calculus: with Applications*. Ellis Horwood, New York, 1981.

W. S. Gray and J. P. Mesko. Observability functions for linear and nonlinear systems. *Systems & Control Letters*, 38(2):99–113, 1999.

W. S. Gray and J. M. A. Scherpen. On the nonuniqueness of singular value functions and balanced nonlinear realizations. *Systems & Control Letters*, 44(3):219–232, 2001.

P. Guillaume and A. Huard. Multivariate Padé approximation. *Journal of the ACM*, 121 (1–2):197–219, 2000.

P. Guillaume, A. Huard, and V. Robin. Generalized multivariate Padé approximants. *Journal of Approximation Theory*, 95(2):203–214, 1998.

H. Hahn. *Rigid Body Dynamics of Mechanics 1 — Theoretical Basis*. Springer-Verlag, Berlin, 2002.

H. Hahn. *Rigid Body Dynamics of Mechanics 2 — Applications*. Springer-Verlag, Berlin, 2003.

J. Hahn and T. F. Edgar. An improved method for nonlinear model reduction using balancing of empirical gramians. *Computer & Chemical Engineering*, 26(10):1379–1397, 2002.

J. Hahn, T. F. Edgar, and W. Marquardt. Controllability and observability covariance matrices for the analysis and order reduction of stable nonlinear systems. *Journal of Process Control*, 13(2):115–127, 2003.

D. J. Hill and I. M. Y. Mareels. Stability theory for differential algebraic systems with application to power systems. *IEEE Transactions on Circuits and Systems*, 37(11): 1416–1423, 1990.

L. Hörmander. *An Introduction to Complex Analysis in Several Variables*. The University Series in Higher Mathematics. D. Van Nostrand, Princeton, New Jersey, 1966.

R. Isaacs. *Differential games : a mathematical theory with applications to warfare and pursuit, control and optimization.* Wiley, New York, 1965.

A. Isidori. *Nonlinear Control Systems.* Springer-Verlag, London, 3rd edition, 1995.

E. A. Jonckheere. Variational calculus for descriptor systems. *IEEE Transactions on Automatic Control*, AC-33(5):491–495, 1988.

U. Jönsson, C. Trygger, and P. Ögren. Lectures in optimal control. Kungliga Tekniska högskolan, Stockholm, 2002.

T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation.* Prentice Hall, Upper Saddle River, New Jersey, 2000.

R. E. Kalman. The theory of optimal control and the calculus of variations. In R. Bellman, editor, *Mathematical optimization techniques*, pages 309–331. University of California Press, Berkeley, California, 1963.

H. K. Khalil. *Nonlinear Systems.* Prentice Hall, Upper Saddle River, New Jersey, 3rd edition, 2002.

P. Kokotović, H. K. Khalil, and J. O'Reilly. *Singular Perturbation Methods in Control: Analysis and Design.* Academic Press, London, 1986.

A. J. Krener. The local solvability of a Hamilton-Jacobi-Bellman PDE around a nonhyperbolic critical point. *SIAM Journal on Control and Optimization*, 39(5):1461–1484, 2001.

A. J. Krener. Reduced order modeling of nonlinear control systems. In A. Astolfi and L. Marconi, editors, *Analysis and Design of Nonlinear Control Systems — In Honor of Alberto Isidori*, pages 41–62. Springer-Verlag, 2008.

A. J. Krener. The construction of optimal linear and nonlinear regulators. In A. Isidori and T. J. Tarn, editors, *Systems, Models and Feedback: Theory and Applications*, pages 301–322. Birkhäuser, Boston, 1992.

A. Kumar and P. Daoutidis. *Control of nonlinear differential algebraic equation systems.* Chapman & Hall CRC, 1999.

P. Kunkel and V. Mehrmann. Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems. *Mathematics of Control, Signals, and Systems*, 14:233–256, 2001.

P. Kunkel and V. Mehrmann. Index reduction for differential-algebraic equations by minimal extension. *Zeitschrift für Angewandte Mathematik und Mechanik*, 84(9):579–597, 2004.

P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations — Analysis and Numerical Solution.* Textbooks in Mathematics. European Mathematical Society, Zürich, Schweiz, 2006.

P. Kunkel and V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. *Journal of Computational and Applied Mathematics*, 56(3): 225–251, December 1994.

P. Kunkel and V. Mehrmann. Local and global invariants of linear differential-algebraic equations and their relation. *Electronic Transactions on Numerical Analysis*, 4:138–157, 1996.

V. Kučera. Stationary LQG control of singular systems. *IEEE Transactions on Automatic Control*, AC-31:31–39, 1986.

S. Lall, J. E. Marsden, and S. Glavaški. Empirical model reduction of controlled nonlinear systems. In *Proceedings of the 14th World Congress of IFAC*, pages 473–478, Beijing, China, July 1999.

S. Lall, J. E. Marsden, and S. Glavaški. A subspace approach to balanced truncation for model reduction of nonlinear control systems. *International Journal of Robust and Nonlinear Control*, 12(6):519–535, 2002.

P. Lancaster and M. Tismenetsky. *The Theory of Matrices — With Applications*. Academic Press, San Diego, USA, 2 edition, 1985.

E. B. Lee and L. Markus. *Foundations of Optimal Control Theory*. Wiley, New York, 1967.

G. Leitmann. *The Calculus of Variations and Optimal Control*. Plenum Press, New York, 1981.

J.-Y. Lin and N. U. Ahmed. Approach to controllability problems for singular systems. *International Journal of Systems Science*, 22(4):675–690, 1991.

Yi Liu and B. D. O. Anderson. Singular perturbation approximation of balanced systems. In *Proceedings of the 28th IEEE Conference on Decision and Control*, pages 1355–1360, Tampa, Florida, December 1989.

J. Löfberg. Pre- and post-processing sum-of-squares programs in practice. *IEEE Transactions on Automatic Control*, Accepted for publication, 2008.

D. G. Luenberger. Time-invariant descriptor systems. *Automatica*, 14(5):473–480, 1978.

D. L. Lukes. Optimal regulation of nonlinear dynamical systems. *SIAM Journal on Control*, 7(1):75–100, February 1969.

A. M. Lyapunov. *The General Problem of the Stability of Motion*. Taylor & Francis, London, 1992.

R. E. Mahony and I. M. Mareels. Global solutions for differential/algebraic systems and implications for Lyapunov direct stability methods. *Journal of Mathematical Systems, Estimation, and Control*, 5(4):1–26, 1995.

S. E. Mattson and G. Söderlind. Index reduction in differential-algebraic equations using dummy derivatives. *SIAM Journal on Scientific Computing*, 14(3):677–692, 1993.

S. E. Mattsson, H. Elmqvist, and M. Otter. Physical system modeling with Modelica. *Control Engineering Practice*, 6:501–510, 1998.

V. Mehrmann. Existence, uniqueness, and stability of solutions to singular linear quadratic optimal control problems. *Linear Algebra and Its Applications*, 121:291–331, 1989.

B. C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26:17–32, 1981.

C. L. Navasca. *Local Solutions of the Dynamic Programming Equations and the Hamilton-Jacobi-Bellman PDE.* PhD thesis, University of California, Davis, 1996.

C. L. Navasca and A. J. Krener. Solution of Hamilton-Jacobi-Bellman equations. In *Proceedings of the 39th IEEE Conference on Decision and Control*, pages 570–574, Sydney, Australia, December 2000.

A. J. Newman and P. S. Krishnaprasad. Computing balanced realizations for nonlinear systems. In *Proceedings of the 14th International Symposium on Mathematical Theory of Networks and Systems*, Perpignan, France, June 2000.

A. J. Newman and P. S. Krishnaprasad. Computation for nonlinear balancing. In *Proceedings of the 37th IEEE Conference on Decision and Control*, pages 4103–4104, Tampa, Florida, December 1998.

C. C. Pantelides. The consistent initialization of differential-algebraic systems. *SIAM Journal on Scientific Computing*, 9(2):213 – 231, March 1988.

L. Pernebo and L. M. Silverman. Model reduction via balanced state space representations. *IEEE Transactions on Automatic Control*, 27(2):382–387, April 1982.

H. J. Pesch and R. Bulirsch. The Maximum Principle, Bellman's Equation and Carathéodory's work. *Journal of Optimization Theory and Applications*, 80(2):199–225, February 1994.

J. W. Polderman and J. C. Willems. *Introduction to mathematical systems theory: a behavioral approach.* Springer-Verlag, New York, 1998.

G. Reißig, W. S. Martinson, and P. I. Barton. Differential-algebraic equations of index 1 may have an arbitrarily high structural index. *SIAM Journal on Scientific Computing*, 21(6):1987–1990, 2000.

W. C. Rheinboldt and B. Simeon. On computing smooth solutions of DAEs for elastic multibody systems. *Computers & Mathematics with Applications*, 37(6):69–83, 1999.

B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: particle filters for tracking applications.* Artech House, Boston, Mass., London, 2004.

P. Rouchon, M. Fliess, and J. Lévine. Kronecker's canonical form for nonlinear implicit differential systems. In *2nd IFAC Workshop on System Structure and Control*, Prague, Czech Republic, September 1992.

I. W. Sandberg. Global implicit function theorems. *IEEE Transactions on Circuits and Systems*, 28(2):145–149, 1981.

O. Schein and G. Denk. Numerical solution of stochastic differential-algebraic equations with applications to transient noise simulation of microelectronic circuits. *Journal of Computational and Applied Mathematics*, 100(1):77–92, November 1998.

J. M. A. Scherpen. *Balancing for Nonlinear Systems*. PhD thesis, University of Twente, The Netherlands, 1994.

J. M. A. Scherpen and W. S. Gray. Minimality and local state decompositions of a nonlinear state space realization using energy functions. *IEEE Transactions on Automatic Control*, AC-45(11):2079–2086, 2000.

K. R. Schneider and T. Wilhelm. Model reduction by extended quasi-steady state approximation. *Journal of Mathematical Biology*, 40(5):443–450, 2000.

T. B. Schön. *Estimation of Nonlinear Systems: Theory and Applications*. PhD thesis, Department of Electrical Engineering, Linköpings universitet, February 2006.

B. Simeon, F. Grupp, C. Führer, and P. Rentrop. A nonlinear truck model and its treatment as a multibody system. *Journal of Computational and Applied Mathematics*, 50:523–532, 1994.

J. Sjöberg. *Some Results On Optimal Control for Nonlinear Descriptor Systems*. Licentiate thesis no 1227, Department of Electrical Engineering, Linköpings universitet, January 2006.

J. Sjöberg and S. T. Glad. Computing the controllability function for nonlinear descriptor systems. In *Proceedings of the 2006 American Control Conference*, Minneapolis, Minnesota, June 2006a.

J. Sjöberg and S. T. Glad. Power series solution of the Hamilton-Jacobi-Bellman equation for time-varying differential-algebraic equations. In *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, California, December 2006b.

J. Sjöberg and S. T. Glad. Rational approximation of nonlinear optimal control problems. In *Proceedings of the 17th World Congress of IFAC*, Seoul, South Korea, July 2008a. Accepted for publication.

J. Sjöberg and S. T. Glad. Power series solution of the Hamilton-Jacobi-Bellman rquation for DAE models with a discounted cost. Technical Report LiTH-ISY-R-2250, Department of Electrical Engineering, Linköpings universitet, 2008b.

J. Sjöberg and T. Glad. Power series solution of the Hamilton-Jacobi-Bellman equation for descriptor systems. In *Proceedings of 44th IEEE Conference on Decision and Control and European Control Conference*, Seville, Spain, December 2005.

J. Sjöberg, K. Fujimoto, and S. T. Glad. Model reduction of nonlinear differential-algebraic equations. In *Proceedings of the 7th IFAC Symposium on Nonlinear Control Systems*, Pretoria, South Africa, August 2007.

R. Sjöberg, Findeisen and F. Allgöwer. Model predictive control of continuous time nonlinear differential algebraic systems. In *Proceedings of the 7th IFAC Symposium on Nonlinear Control Systems*, Pretoria, South Africa, August 2007.

S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control — Analysis and Design*. Wiley, Chichester, 2001.

A. Steinbrecher. *Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems*. PhD thesis, Technischen Universität Berlin, 2006.

T. Stykel. Gramian-based model reduction for descriptor systems. *Mathematics of Control, Signals, and Systems*, 16:297–319, 2004.

C. Sun and J. Hahn. Reduction of stable differential-algebraic equation systems via projections and system identification. *Journal of Process Control*, 15(6):639–650, 2005.

H. J. Sussmann and J. C. Willems. 300 years of optimal control: From the brachystochrone to the maximum principle. *IEEE Control Systems Magazine*, 17(3):32–44, 1997.

H. Tidefelt. *Structural algorithms and pertubations in differential-algebraic equations*. Licentiate thesis no 1318, Department of Electrical Engineering, Linköpings universitet, May 2007.

C. Tischendorf. Coupled systems of differential algebraic and partial differential equations in circuit and device simulation – modeling and numerical analysis. Habilitationsschrift, Institute of Mathematics, Humboldt-Universität, Berlin, Germany, 2003.

A. J. van der Schaft. Relations between H_∞ optimal control of a nonlinear system and its linearization. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pages 1807–1808, Brighton, England, December 1991.

A. Vannelli and M. Vidyasagar. Maximal Lyapunov functions and domains of attraction for autonomous nonlinear systems. *Automatica*, 21(1):69–80, 1985.

G. C. Verghese. *Infinite-frequency behavior in generalized dynamical systems*. PhD thesis, Stanford University, California, 1978.

N. Vora and P. Daoutidis. Nonlinear model reduction of chemical reaction systems. *AIChE Journal*, 47(10):2320 – 2332, 2001.

H. Wang, C. Fai Yung, and F.-R. Chang. H_∞ control for nonlinear descriptor systems. *IEEE Transactions on Automatic Control*, AC-47(11):1919–1925, 2002.

A. P. Willemstein. Optimal regulation of nonlinear dynamical systems on a finite interval. *SIAM Journal on Control and Optimization*, 15(6):1050–1069, 1977.

R. Winkler. Stochastic differential algebraic equations of index 1 and applications in circuit simulation. *Journal of Computational and Applied Mathematics*, 163(2):435–463, February 2004.

H. Wu and K. Mizukami. Stability and robust stability of nonlinear descriptor systems with uncertainties. In *Proceedings of the 33rd IEEE Conference on Decision and Control*, pages 2772–2777, Lake Buena Vista, Florida, December 1994.

H. Xu and K. Mizukami. Hamilton-Jacobi equation for descriptor systems. *Systems & Control Letters*, 21:321–327, 1993.

T. Yoshida and K. A. Loparo. Quadratic regulatory theory for analytic non-linear systems with additive control. *Automatica*, 25(4):531–544, 1989.

PhD Dissertations
Division of Automatic Control
Linköping University

M. Millnert: Identification and control of systems subject to abrupt changes. Thesis No. 82, 1982. ISBN 91-7372-542-0.
A. J. M. van Overbeek: On-line structure selection for the identification of multivariable systems. Thesis No. 86, 1982. ISBN 91-7372-586-2.
B. Bengtsson: On some control problems for queues. Thesis No. 87, 1982. ISBN 91-7372-593-5.
S. Ljung: Fast algorithms for integral equations and least squares identification problems. Thesis No. 93, 1983. ISBN 91-7372-641-9.
H. Jonson: A Newton method for solving non-linear optimal control problems with general constraints. Thesis No. 104, 1983. ISBN 91-7372-718-0.
E. Trulsson: Adaptive control based on explicit criterion minimization. Thesis No. 106, 1983. ISBN 91-7372-728-8.
K. Nordström: Uncertainty, robustness and sensitivity reduction in the design of single input control systems. Thesis No. 162, 1987. ISBN 91-7870-170-8.
B. Wahlberg: On the identification and approximation of linear systems. Thesis No. 163, 1987. ISBN 91-7870-175-9.
S. Gunnarsson: Frequency domain aspects of modeling and control in adaptive systems. Thesis No. 194, 1988. ISBN 91-7870-380-8.
A. Isaksson: On system identification in one and two dimensions with signal processing applications. Thesis No. 196, 1988. ISBN 91-7870-383-2.
M. Viberg: Subspace fitting concepts in sensor array processing. Thesis No. 217, 1989. ISBN 91-7870-529-0.
K. Forsman: Constructive commutative algebra in nonlinear control theory. Thesis No. 261, 1991. ISBN 91-7870-827-3.
F. Gustafsson: Estimation of discrete parameters in linear systems. Thesis No. 271, 1992. ISBN 91-7870-876-1.
P. Nagy: Tools for knowledge-based signal processing with applications to system identification. Thesis No. 280, 1992. ISBN 91-7870-962-8.
T. Svensson: Mathematical tools and software for analysis and design of nonlinear control systems. Thesis No. 285, 1992. ISBN 91-7870-989-X.
S. Andersson: On dimension reduction in sensor array signal processing. Thesis No. 290, 1992. ISBN 91-7871-015-4.
H. Hjalmarsson: Aspects on incomplete modeling in system identification. Thesis No. 298, 1993. ISBN 91-7871-070-7.
I. Klein: Automatic synthesis of sequential control schemes. Thesis No. 305, 1993. ISBN 91-7871-090-1.
J.-E. Strömberg: A mode switching modelling philosophy. Thesis No. 353, 1994. ISBN 91-7871-430-3.
K. Wang Chen: Transformation and symbolic calculations in filtering and control. Thesis No. 361, 1994. ISBN 91-7871-467-2.
T. McKelvey: Identification of state-space models from time and frequency data. Thesis No. 380, 1995. ISBN 91-7871-531-8.
J. Sjöberg: Non-linear system identification with neural networks. Thesis No. 381, 1995. ISBN 91-7871-534-2.
R. Germundsson: Symbolic systems – theory, computation and applications. Thesis No. 389, 1995. ISBN 91-7871-578-4.
P. Pucar: Modeling and segmentation using multiple models. Thesis No. 405, 1995. ISBN 91-7871-627-6.
H. Fortell: Algebraic approaches to normal forms and zero dynamics. Thesis No. 407, 1995. ISBN 91-7871-629-2.

A. Helmersson: Methods for robust gain scheduling. Thesis No. 406, 1995. ISBN 91-7871-628-4.
P. Lindskog: Methods, algorithms and tools for system identification based on prior knowledge. Thesis No. 436, 1996. ISBN 91-7871-424-8.
J. Gunnarsson: Symbolic methods and tools for discrete event dynamic systems. Thesis No. 477, 1997. ISBN 91-7871-917-8.
M. Jirstrand: Constructive methods for inequality constraints in control. Thesis No. 527, 1998. ISBN 91-7219-187-2.
U. Forssell: Closed-loop identification: Methods, theory, and applications. Thesis No. 566, 1999. ISBN 91-7219-432-4.
A. Stenman: Model on demand: Algorithms, analysis and applications. Thesis No. 571, 1999. ISBN 91-7219-450-2.
N. Bergman: Recursive Bayesian estimation: Navigation and tracking applications. Thesis No. 579, 1999. ISBN 91-7219-473-1.
K. Edström: Switched bond graphs: Simulation and analysis. Thesis No. 586, 1999. ISBN 91-7219-493-6.
M. Larsson: Behavioral and structural model based approaches to discrete diagnosis. Thesis No. 608, 1999. ISBN 91-7219-615-5.
F. Gunnarsson: Power control in cellular radio systems: Analysis, design and estimation. Thesis No. 623, 2000. ISBN 91-7219-689-0.
V. Einarsson: Model checking methods for mode switching systems. Thesis No. 652, 2000. ISBN 91-7219-836-2.
M. Norrlöf: Iterative learning control: Analysis, design, and experiments. Thesis No. 653, 2000. ISBN 91-7219-837-0.
F. Tjärnström: Variance expressions and model reduction in system identification. Thesis No. 730, 2002. ISBN 91-7373-253-2.
J. Löfberg: Minimax approaches to robust model predictive control. Thesis No. 812, 2003. ISBN 91-7373-622-8.
J. Roll: Local and piecewise affine approaches to system identification. Thesis No. 802, 2003. ISBN 91-7373-608-2.
J. Elbornsson: Analysis, estimation and compensation of mismatch effects in A/D converters. Thesis No. 811, 2003. ISBN 91-7373-621-X.
O. Härkegård: Backstepping and control allocation with applications to flight control. Thesis No. 820, 2003. ISBN 91-7373-647-3.
R. Wallin: Optimization algorithms for system analysis and identification. Thesis No. 919, 2004. ISBN 91-85297-19-4.
D. Lindgren: Projection methods for classification and identification. Thesis No. 915, 2005. ISBN 91-85297-06-2.
R. Karlsson: Particle Filtering for Positioning and Tracking Applications. Thesis No. 924, 2005. ISBN 91-85297-34-8.
J. Jansson: Collision Avoidance Theory with Applications to Automotive Collision Mitigation. Thesis No. 950, 2005. ISBN 91-85299-45-6.
E. Geijer Lundin: Uplink Load in CDMA Cellular Radio Systems. Thesis No. 977, 2005. ISBN 91-85457-49-3.
M. Enqvist: Linear Models of Nonlinear Systems. Thesis No. 985, 2005. ISBN 91-85457-64-7.
T. B. Schön: Estimation of Nonlinear Dynamic Systems — Theory and Applications. Thesis No. 998, 2006. ISBN 91-85497-03-7.
I. Lind: Regressor and Structure Selection — Uses of ANOVA in System Identification. Thesis No. 1012, 2006. ISBN 91-85523-98-4.
J. Gillberg: Frequency Domain Identification of Continuous-Time Systems Reconstruction and Robustness. Thesis No. 1031, 2006. ISBN 91-85523-34-8.
M. Gerdin: Identification and Estimation for Models Described by Differential-Algebraic Equations. Thesis No. 1046, 2006. ISBN 91-85643-87-4.

C. Grönwall: Ground Object Recognition using Laser Radar Data – Geometric Fitting, Performance Analysis, and Applications. Thesis No. 1055, 2006. ISBN 91-85643-53-X.
A. Eidehall: Tracking and threat assessment for automotive collision avoidance. Thesis No. 1066, 2007. ISBN 91-85643-10-6.
F. Eng: Non-Uniform Sampling in Statistical Signal Processing. Thesis No. 1082, 2007. ISBN 978-91-85715-49-7.
E. Wernholt: Multivariable Frequency-Domain Identification of Industrial Robots. Thesis No. 1138, 2007. ISBN 978-91-85895-72-4.
D. Axehill: Integer Quadratic Programming for Control and Communication. Thesis No. 1158, 2008. ISBN 978-91-85523-03-0.
G. Hendeby: Performance and Implementation Aspects of Nonlinear Filtering. Thesis No. 1161, 2008. ISBN 978-91-7393-979-9.