Uncertainty Issues in Deterministic and Stochastic Nonlinear Systems

Diego S. Carrasco Yáñez
Ingeniero Civil Electrónico
M.Sc. in Electronics Engineering

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

March, 2014
The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University’s Digital Repository, subject to the provisions of the Copyright Act 1968.

I hereby certify that the work embodied in this thesis contains published papers of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publications.

________________________________________
Diego S. Carrasco Yáñez

March, 2014
First and foremost, I would like to thank Prof. Graham Goodwin, for without him, none of this would have been possible. I have learned many things from his passion, drive and determination for research, but the most important things I have learned are from his genuine care and worry towards the people around him. Thank you.

Secondly, I would like to thank Prof. Mario Salgado, who introduced me to the world of research. I have always appreciated his fatherly advice and care. Above all, he has always been a huge influence to me, not only as an academic, but as a person.

I would like to thank, with all my heart, my parents Gloria and Luis and my siblings Pablo and Nicolás for their constant love and support. Distance takes a toll on every relationship, but I have never felt more loved than those times we have been reunited. I cannot stop mentioning my grandparents, Olivia and Guillermo, Maria and Luis. I love you all.

Very special thanks to Rocío, Ramón, Katherine and Ricardo, for welcoming me into their families and allowing me to be part of them. I would also like to thank Laura and Antonella, for being the constant reminder that a smile, a kiss and a hug can change a person’s heart. I hope one day you can understand how much that meant to me.

To my Australian family: Emma and Luke, Jeni and Lincoln, Kash and Katie, Murat, Pierre and Elisa, Aurelio and Vi, Hope and Eduardo. Thank you for your love and companionship, time and advice when I needed it most.

I would like to thank specially Yoyi and Mauricio, for helping me stand on my own two feet when all this adventure began. I would also like to thank Boris, Alejandro, Steven and Jay for their friendship, help and advice during all these years.

Many thanks to Jayne and Dianne, for their help and advice throughout my time at the university. They made me feel at home and I could not have done it otherwise. Thank you.

To the many friends I have made during my years in Newcastle, I say thanks for making this period of my life memorable. Special thanks to Alain Yetendje, Andrés López, Daniel Quevedo, Daniel
Dolz, Damián and Bea, Erik Henriksson, Francisco Gordín, Fernando López-Caamal, Hani Abaase, He Kong, Hessam, Pouria, Juan Carlos, Marcus Reble, Robert Palma, Sonja and Edwin, Steffi and Florian, Koen Van der Mijle, Yuzhe Li, Juan Alvarez, Esteban Osella, Isabella and Denis, Cinthya and Mauricio, Elizabeth and Shane, Emma, Blake, Kate, Beth, Duncan, Brenton, Lyneece, Javier and Margarida.

To the many friends back home: Stjpe, Maria Claudia and Sebastian, Marcela and Nelson, Valeria and Pedro, Virginia, Pedro, Jaime, Sebastian, Hernán, Luna, Patricio and Daniela, Francisco, Andrés, Andrea and Tomás. I have always kept you in my heart. Thank you.

Finally, I would like to thank Eduardo Silva. You were my friend, colleague and teacher. You will be missed. Rest in peace.

I sincerely apologise to anyone I have forgotten to mention. Everyone I have ever met has played a role in my life. Thank you, from the bottom of my heart.

I acknowledge financial support from CONICYT Chile through their program “Becas Chile”.

Diego Carrasco
August 2014
Robustness issues arise in every real world control problem. The objective of any robust control strategy is to preserve closed-loop stability in situations where the real plant differs from the model used to design the controller, i.e. the real system is, in some sense, unknown. There are different ways to quantify, or describe, the uncertainty of a model. It is the amount of uncertainty, or lack of confidence in the model, that ultimately determines, and constrains, what the closed-loop can achieve.

In this thesis we address particular issues concerned with how to quantify and reduce the impact of uncertainty. To this end, the present thesis is divided in two parts:

The first part is aimed at linear systems. We propose two ideas on how to improve closed-loop performance in the face of general uncertainty, namely, (i) augmenting the control architecture with a feedforward component and (ii) augmenting the observer architecture by using the more general class of unbiased observers. We then illustrate the first strategy applied to an Artificial Pancreas problem.

The second part is aimed at nonlinear systems. A common source of uncertainty in this area is the use of approximate sampled-data models of continuous time systems, be it for control design or system identification. This is due to the fact that, contrary to the linear case, exact discretisations are not generally possible in the nonlinear case. In particular, we deal with the sampled-data scenario in both deterministic and stochastic cases and focus our attention on accuracy and related properties of sampled-data models.

We first study the accuracy properties, or error dynamics, of a particular deterministic sampled-data model, and show that it possesses an improved order of accuracy when compared to the usual Euler approximation. We then demonstrate the usefulness of having such a quantification via several applications, namely, (i) obtaining better bias-variance tradeoffs in the parameter estimation of continuous-time systems from sampled-data, (ii) obtaining a sampled-data model that depends only on input-output data that retains the improved order of accuracy, and (iii) obtaining better performance in high-gain sampled-data feedback control of nonlinear systems, via feedback
linearisation.

In addition, we extend the analysis to stochastic sampled-data nonlinear systems. In this case, we show that the error dynamics are tightly intertwined with other system properties that arise due to the sampling process. In particular, we show the existence of stochastic sampling zero dynamics that are closely related to the sampling zero dynamics associated with the deterministic case.
# Contents

Acknowledgements .......................... v

Abstract .................... vii

Contents ....................... ix

1 Introduction ..................... 1

1.1 Thesis outline ................... 3

1.2 Publications ..................... 5

I Linear Systems ................. 7

2 Feedforward Model Predictive Control .... 9

2.1 Motivation ....................... 10

2.2 Basic Definitions ................ 13

2.3 Observer Design .................. 14

2.4 Feedforward Design .............. 14

2.4.1 Constrained Nominal Design .... 15

2.4.2 Comments ..................... 15

2.5 Feedback Design ................. 15

2.5.1 Constrained Feedback Design .. 16
3 Feedforward MPC in Artificial Pancreas

3.1 Blood Glucose Regulation and Diabetes

3.1.1 Glucose Regulation

3.1.2 Glucose Regulation Disorders

3.2 Fundamental Limitations Due to Model Uncertainty in Artificial Pancreas

3.3 Type 1 Diabetes Model

3.3.1 Actuator

3.3.2 Patient

3.3.3 Sensor

3.3.4 Food absorption

3.3.5 Exercise

3.4 Model Approximation

3.5 Insights into Performance Limitations

3.6 Illustrating the advantages of Feedforward MPC

3.6.1 Robust Feedback MPC

3.6.2 Feedforward MPC (FFMPC)

3.6.3 Performance quantification
5.3 A sampled-data model for nonlinear systems ........................................... 66

5.4 Sampling Zeros .............................................................. 67

5.4.1 Euler-Frobenius Polynomials ........................................... 67

5.4.2 Deterministic Linear Systems .......................................... 68

5.4.3 Deterministic Nonlinear Systems ....................................... 69

5.4.4 Stochastic Linear Systems .............................................. 69

6 Vector Measures of Accuracy for Sampled-data Models ......................... 71

6.1 Vector Truncation Error Definitions ......................................... 72

6.2 Error Analysis of Truncated Taylor Series Sampled Data Model ............. 74

6.3 Examples ........................................................................ 78

6.4 Conclusions .................................................................... 80

7 An Input-Output Sampled Data Model ............................................. 83

7.1 An input-output model ..................................................... 83

7.2 Conclusions .................................................................... 86

8 Bias and Variance Issues in the Identification of Nonlinear Systems ....... 87

8.1 Preliminaries .................................................................... 88

8.2 Approximate Sampled Data Models ......................................... 88

8.2.1 Euler Integration ......................................................... 89

8.2.2 Truncated Taylor Series ............................................... 89

8.3 Maximum Likelihood Estimation ........................................... 89

8.4 Finite Data Length Interpretation ......................................... 92

8.5 Conclusions .................................................................... 93
A.2.1 Definition ................................................................. 138
A.2.2 Properties ............................................................... 139
A.3 Euler-Frobenius numbers .............................................. 140
A.3.1 Definition ............................................................... 140
A.3.2 Combinatorial interpretation of Eulerian numbers .............. 141
A.4 Euler-Frobenius fractions .............................................. 142
A.4.1 Definition ............................................................... 142
A.4.2 Properties ............................................................... 142
A.5 Generalised Eulerian polynomials ..................................... 143
A.5.1 Definition ............................................................... 143
A.5.2 Properties ............................................................... 144
A.6 Euler polynomials and numbers ....................................... 144
A.7 Bernoulli polynomials and numbers ................................... 145
A.8 Summary ................................................................. 146

Bibliography ................................................................. 147
Robustness issues arise in every real world control problem. The objective of any robust control strategy is to preserve closed-loop stability in situations where the real plant differs from the model used to design the controller, i.e. the real system is, in some sense, unknown. There are different ways to quantify, or describe, the uncertainty of a model. It is the amount of uncertainty, or lack of confidence in the model, that ultimately determines, and constrains, what the closed-loop can achieve.

This thesis is divided in two parts: the first part deals with how to improve performance in the face of unavoidable uncertainty. The results presented here are restricted to the linear case, but the core ideas go beyond linear systems. The second part of the thesis is focused on avoiding the uncertainty at its source. Specifically, uncertainty induced by the use of an approximate model when nonlinear sampled-data systems are considered.

The first part of this thesis is motivated by the question: How can closed-loop performance be improved in the presence of uncertainty? A quintessential example of this situation arises in the Artificial Pancreas (AP) problem, where an autonomous device is required to mimic, and ultimately replace, the blood glucose regulation functions of the pancreas. A typical AP consists of a continuous glucose monitor (CGM), a control algorithm and an insulin infusion system (IIS). Both the IIS and the CGM have uncertain dynamics, but the main source of uncertainty is the human body itself. The glucose regulatory system is affected by a myriad of factors, internal and external. The associated uncertainty severely compromises the performance that any control loop can achieve.

In this regard, we present an idea based on the use of extra information available through feedforward action and describe the performance impact in a probabilistic setup. The idea is illustrated in the linear MPC framework, but the concept applies equally to any control strategy. The novelty
lies in the use of two separate cost functions, one for feedforward control and one for feedback control.

We then develop a related idea. Specifically, we show that the complementary sensitivity function of a control loop can be written as the product of a state-feedback complementary sensitivity function and an observer complementary sensitivity function. This result shows it is possible to improve the performance of a control loop by accounting for model errors/uncertainty by using a more general class of observers, namely unbiased observers, instead of the usual Luenberger observer. By using the class of unbiased observers, extra degrees of freedom are attained in the design of the robust feedback control loop.

Both of the ideas introduced above address the robust performance issue from the controller and the observer side. The results are believed to be particularly insightful since (i) MPC is a controller architecture broadly used in practice, (ii) in the MPC framework, the observer is usually fixed and predesigned, leaving the controller to deal with every design consideration.

The second part of this thesis is based on a more general setting. The ideas stem from the fact that a common source of uncertainty arises from continuous-time systems that are controlled digitally. A strategy to design a controller is to discretise the model and then design the controller based on the discrete version of the continuous system. For linear systems, there exist exact discretisations. However, when dealing with nonlinear systems, and save for a few special cases, approximations have to be performed to obtain a discrete model. The level of such approximations are a key source of model errors and uncertainty in the digital control of continuous nonlinear systems.

This motivates the second question examined in the thesis: How can two model approximations be compared objectively? We will show that measures of accuracy taken directly from the traditional numerical analysis literature are not directly applicable to the control problem. Extensions of these measures are presented in the thesis, which allow a meaningful comparison of approximated models obtained with a control orientation in mind. In addition, we present two scenarios where the quantification of the accuracy of the model provides important insights, namely: (i) an input-output model that retains the order of accuracy and accounts for the sampling zero dynamics, and (ii) the bias-variance trade-off in the parameter estimation of continuous time nonlinear systems from sampled-data.

A third and related question is: Are there any properties linked to the degree of accuracy of an approximation? The answer here is positive and indeed an interesting one: sampling zeros dynamics. Sampling zeros are the prime byproduct of the sampling process. In this aspect, it has
been determined [24] that a specific approximation, based on the normal form of a continuous nonlinear system, captures the asymptotic sampling zero dynamics, whereas basic and simple approximations, e.g. Euler, do not. Based on this knowledge, we present a preliminary study of the impact of the sampling zeros in the high-gain sampled-data control of nonlinear systems.

All of the issues described above were first established in a deterministic framework. It would be sensible then to ask whether or not any of these properties replicate in a stochastic framework. Indeed, the notion of sampling zeros for linear stochastic systems was introduced in the late 80’s [119]. However, the existence of sampling zero dynamics for stochastic nonlinear systems has remained an open question. In this thesis, we first extend the measures of accuracy, relating to sampled-data models, from the deterministic to the stochastic case, including the notions of strong and weak convergence. It is the use of these measures of accuracy that, based on a specific structure of the stochastic nonlinear system, allows an equivalent model to be obtained. It is this equivalent model that is shown to possess sampling zero dynamics that are related to the sampling zero dynamics of the deterministic case.

### 1.1 Thesis outline

Since the span of this thesis covers many different topics, then, for convenience, a detailed review of relevant background literature is included with each individual chapter. The remainder of the thesis is organised as follows:

Chapter 2 examines the role of feedforward in model predictive control. A key point here is to compare and contrast notions of feedforward and preview. We propose a particular strategy, which we call FFMPC, based on the use of feedforward in the MPC framework. This strategy is aimed at improving performance in the presence of measurement noise and certain types of uncertainty. The strategy presented here is targeted at time-varying reference tracking (sometimes called reference feedforward).

Chapter 3 presents an application of the FFMPC strategy to the Artificial Pancreas problem. The aim is to improve the regulation of blood glucose concentration in the presence of model uncertainty. In this case, the key sources of uncertainty are the combined effect of the insulin infusion system, the human glucose regulatory system and the continuous glucose monitor. The application FFMPC is tailored to deal with disturbance rejection, e.g. food intake (disturbance feedforward).
Chapter 4 presents a relationship between control and filtering complementary sensitivity functions. By using the notion of auxiliary variable feedback, the auxiliary variable sensitivity function is shown to be the link between the first two sensitivity functions. The result is used to show that robust performance in a control loop can be improved by considering a more general class of observers, i.e. unbiased observers, rather than the common Luenberger observer.

Chapter 5 reviews background material needed for the subsequent chapters. Specifically, we review material for a particular sampled-data model for nonlinear systems, i.e. the TTS model, and review results related to sampling zeros for linear and nonlinear systems.

Chapter 6 introduces several novel measures of accuracy. Whereas the usual numerical analysis measures assign a single bound to an error vector, the new definitions assign a different bound to each element of the vector. This proves useful in control and system identification problems since certain states, or a combination of them, are usually of particular interest. We apply the new definitions to the TTS model and prove that it has an improved order of accuracy when compared to the usual Euler approximation.

The above development is followed by three short chapters which illustrate how the improved error bounds of the TTS model provide the basis for several interesting and useful results.

In particular, Chapter 7 is concerned with obtaining a sampled-data model based only on input-output data. We use the TTS model as a starting point and show that the resulting I/O model retains the order of local accuracy in the output. A byproduct of this transformation is that the sampling zero dynamics are shown explicitly in the model.

Chapter 8 is concerned with the estimation of the parameters of continuous-time nonlinear systems from sampled data. We show that when the well known problem of bias-variance trade-off is considered, then the use of the TTS model significantly reduces the bias error when finite data is used.

Chapter 9 is concerned with the sampled-data high-gain control of a continuous-time nonlinear system by feedback linearisation. We show that, designing a high-gain controller based on the TTS model provides better closed-loop performance on the real system than using a simple Euler approximation. These are preliminary results and deserve further study – see suggestions for future research in Chapter 11.

Chapter 10 presents extensions of Chapter 6 to the stochastic nonlinear case. We first present a stochastic approximate model analogous to the TTS model, which we will name the STTS model.
We also extend the error definitions to the stochastic case, namely considering strong and weak convergence, and prove that the STTS model has improved order of accuracy when compared to the Euler-Maruyama approximation. With this improved order of accuracy, we show that the STTS model includes stochastic sampling zero dynamics that are related to those in the deterministic case.

Chapter 11 presents a summary of the contributions of this thesis and a discussion on future research directions.

1.2 Publications

The following list presents the contributions by the author to the literature in the area during the research carried out to produce this thesis.

Book Chapter


Journal Papers


Conference Papers


Part I

Linear Systems
Feedforward Model Predictive Control

This chapter examines the role played by feedforward in model predictive control (MPC). Model Predictive Control has been a major success story in industry. There are two principal reasons for this success. Firstly, the method addresses important practical issues including constraints on inputs and states. Secondly, there exists a rich supporting theory [102, 84, 75, 86].

The main argument presented here relies on the distinction between preview and feedforward. In classical control theory [49], a distinction is made between “feedforward”, “preview” and “feedback”. Each of these tools is known to yield performance gains in certain practical problems. By way of contrast, current implementations of MPC are restricted to the use of preview and feedback. This begs the question, “Can feedforward be incorporated into MPC and, if so, under what circumstances would it be useful?” The goal of this chapter is to address this question.

We will show that the use of feedforward can significantly improve performance, e.g. reference tracking, in the presence of measurement noise and certain types of model uncertainty.

The ideas presented here have a loose connection to the ideas presented in [13, 16, 14, 15]. In the latter body of work it is argued that it may be desirable to replace absolute guarantees of performance (including stability) by high probability guarantees, in exchange for a significant improvement in performance. Our work aims at combining an absolute guarantee of robust stability with high performance reference tracking for systems “near” the nominal system. Our examples show that “near” can include large modelling errors. An important departure from the work in [13, 16, 14, 15] is that we augment the control architecture to give a richer set of trade-offs under different probabilistic scenarios rather than optimising the tuning of a given architecture.
under probabilistic conditions.

To set the framework, we first define the terms preview, feedback and feedforward. We define “Preview” as the inclusion of knowledge regarding future reference or disturbance values into the control law. Preview is helpful in control since it allows the controller to “prepare” for future changes. In particular, it is known \[88\] that the availability of preview action weakens the fundamental limitations that apply to set point tracking. We define “Feedback” as the process of observing the plant response and taking corrective action based on the difference between the measured response and the desired response. Finally, we define “Feedforward” as an open loop evaluation of an input, or part of it, without any corrections arising from observations made on the plant.

It is worth noting that the definition of feedforward used here differs from that commonly used in the MPC literature \[105, 106, 102\]. The work of \[105, 106\] mainly refers to the concept of “preview” and the influence in performance of a limited control horizon. On the other hand, in the work of \[102\] Chapter 5, the concept of feedforward has one main difference, namely the feedforward controller depends on observations made on the plant or estimated disturbances. More specifically, the nominal state \( z \), used for the feedforward controller, is initialised at every time step by the estimate of the current state of the “true” system, i.e. \( x \), making it inherently different to the approach presented here.

The remainder of this chapter is organised as follows: In Section 2.1 we motivate the idea of feedforward using a linear time invariant unconstrained control scenario. In Section 2.2 we present basic definitions of linear systems. Section 2.3 introduces the observer used to estimate the plant state (including unmeasured disturbances). Section 2.4 outlines the feedforward MPC design. Section 2.5 describes the design of the feedback component. Section 2.6 presents numerical examples. Finally, in Section 2.7 we draw conclusions.

2.1 Motivation

By way of introduction to the idea of feedforward, we consider a simple linear unconstrained control problem. Assume the plant transfer function is \( G \), either in continuous or discrete time. In the sequel, when we restrict ourselves to a specific time domain, it will be made explicit. Assume also that we have a reference signal \( r \) and a measured output disturbance \( d_n \) (arising, for example, from an up-stream process or a prediction system \[78\]). We can then sketch a linear single-input single-output control system incorporating both feedforward and feedback as in Fig. 2.1. In this figure,
2.1 Motivation

Figure 2.1: Closed loop with reference and disturbance feedforward

$y^*, e, u_{ff}, u_{fb}, u, y, d_n, d$ and $n$ denote reference, tracking error, feedforward input, feedback input, total plant input, plant output, measured disturbance, unmeasured disturbance and measurement noise respectively.

A simple calculation shows that the output error satisfies:

$$y^* - y = \frac{(1 - GC_2) \cdot y^* - (1 + GC_3) \cdot d_n - d + GC_1 \cdot n}{(1 + GC_1)} \quad (2.1)$$

If we further restrict the problem to the special case where $G$ is open loop stable, then it is known [49] that all stabilising controllers can be expressed as

$$C_1 = Q \cdot \frac{1}{1 - GQ} \quad (2.2)$$

where $Q$ is stable and biproper. Substituting (2.2) into (2.1) yields

$$y^* - y = [1 - GQ] \left\{ (1 - GC_2) \cdot y^* - (1 - GC_3) \cdot d_n - d \right\} + GQ \cdot n \quad (2.3)$$

Considering only the terms of (2.3) related to $y^*$, $d_n$, $d$, we see that all three transfer functions $Q$, $C_2$, $C_3$ share a similar design objective, namely, each should be a stable approximation to the inverse of $G$. Evaluation of a suitable inverse can be achieved by any number of design methodologies, including optimal control [49].

A distinction between the design objectives of $Q$, $C_2$ and $C_3$ arises when we consider measurement noise and plant model errors.

Regarding measurement noise, we see that there is a conflict between reducing the impact of disturbances (which requires $GQ \approx 1$) and reducing the impact of measurement noise (which requires $GQ \approx 0$). This is the well known sensitivity, complementary sensitivity trade-off [49]. However,
the transfer functions $C_2$ and $C_3$ are not compromised by this trade-off, making reference tracking and measured disturbance rejection possible over a wider bandwidth. This is clear from (2.3).

A similar argument applies to model uncertainty. To illustrate, suppose the true plant transfer function $G$ satisfies

$$G = G_o(1 + G\Delta)$$

where $G_o$ is the nominal model and $G\Delta$ denotes the multiplicative model error. In this case we replace (2.2) by

$$C_1 = \frac{Q}{1 - G_oQ}$$

and then (2.3) becomes

$$y^* - y = 1 - \frac{G_oQ}{1 + G_oG\Delta Q} \cdot \{ (1 - G_o(1 + G\Delta)C_2) \cdot y^* - (1 + G_o(1 + G\Delta)C_3) \cdot d_n - d \} + \frac{G_o(1 + G\Delta)Q}{1 + G_oG\Delta Q} \cdot n$$

We see from (2.6) that, provided $C_2$ and $C_3$ are stable, then they cannot destabilise a stable open loop plant. However, choosing $Q$ stable is insufficient to ensure robust closed loop stability. In particular, we also need to ensure that $(1 + G_oG\Delta Q)$ has no zeros outside the unit circle. This robustness requirement can further compromise the design of the feedback component $Q$.

For example, $H_\infty$ design is aimed at guaranteeing robust stability for all possible plant uncertainty [128, 127]. However, demanding closed loop stability for all (or, at least, almost all) possible plant uncertainties is likely to severely compromise the bandwidth over which $Q$ can be chosen as an (approximate) inverse to $G_o$. On the other hand, we can be much more cavalier about the choice of $C_2$ and $C_3$ since we know that use of these transfer functions cannot make an otherwise stable system unstable no matter what the plant uncertainty. This simple argument establishes a further distinction between the roles of $Q, C_2$ and $C_3$.

In the context of MPC, the “inverses” described above are typically calculated via the minimisation of a cost function. In this sense, the calculation of $C_1, C_2$ and $C_3$ are almost identical, with the important caveat that $C_1$ needs to respect the impact of measurement noise and the robust stability requirement. Hence the performance may be compromised by the noise and the worst case uncertainty (no matter how unlikely it is to occur). This suggests that, in MPC, there may be an important advantage to be gained by adding the design of $C_2$ and $C_3$ to the usual design of $C_1$. This will be the central theme of the current chapter. For simplicity we will focus primarily on model uncertainty and we will therefore refer to the output error as the tracking error.
2.2 Basic Definitions

We consider a linear time-invariant single-input single-output system:

\[ y = Gu + d_n + d \]  
\[ y_n = y + n \]  

where \( G(q) \) is a discrete time transfer function (the “true plant” transfer function) and where \( y, u, d_n, d, n, y_n \) denote unmeasured plant output, plant input, measured disturbance, unmeasured disturbance, measurement noise and measured plant output respectively. To describe plant uncertainty we write \( G \) as

\[ G = G_o(1 + G_\Delta) \]

where \( G_o \) is the nominal model and \( G_\Delta \) the multiplicative model error.

The plant dynamics are modelled in state space form by:

\[ x^+ = Ax + Bu + Bw \]  
\[ y = Cx + d_n + d \]  
\[ y_n = y + n \]  
\[ w = G_\Delta u \]  

where \( G_o = C(zI - A)^{-1}B \) and \( x^+ \) denotes the subsequent value of \( x \).

For the problem at hand, we make the following simplifying assumptions:

- \( G \) is invertible at d.c.
- \( G_o = C(zI - A)^{-1}B \) is invertible at d.c.
- \( G_\Delta \) is exponentially stable, i.e. has an impulse response \( \{h_\Delta(k)\} \) satisfying
  \[ |h_\Delta(k)| \leq \alpha \lambda^k \quad \text{for} \quad \alpha \in \mathbb{R}, \lambda < 1 \]  
  (Note that \( G_\Delta \) is typically “high pass”).
- The reference signal \( y^* \) is not constant. Nonetheless, we will include integral action in the control law so as to provide small tracking error in the event that \( y^*, d_n \) or \( d \) are constant for a long period of time.
• $y^*$ and $d_n$ can be previewed $N_p$ samples ahead of the current time, i.e. at time $k$ we assume knowledge of $y^*(j), d_n(j), j = k, \ldots, k + N_p$ (Note $N_p = 0$ corresponds to the “unpreviewed” case).

• After $N_p$ samples, we predict $y^*$ and $d_n$. One simple option is to hold the last available value, i.e. we define

\[
\hat{y}(k + i) = y(k + i) \quad i \leq N_p \quad (2.15a)
\]

\[
\hat{y}(k + i) = y(k + N_p) \quad i > N_p. \quad (2.15b)
\]

\[
\hat{d}(k + i) = d(k + i) \quad i \leq N_p \quad (2.15c)
\]

\[
\hat{d}(k + i) = d(k + N_p) \quad i > N_p. \quad (2.15d)
\]

(This is standard in contemporary formulations of MPC, where $N_p$ often takes the default value $N_p = 0$).

### 2.3 Observer Design

For the purpose of observer design, we assume that the unmeasured disturbance $d$ is constant. We then define an observer for the system state and the unmeasured disturbance:

\[
\dot{x}^+ = A \dot{x} + Bu + J_1(y - C \dot{x} - d_n - \hat{d}) \quad (2.16)
\]

\[
\dot{d}^+ = \dot{d} + J_2(y - C \dot{x} - d_n - \hat{d}) \quad (2.17)
\]

where

\[
\hat{A} = \begin{bmatrix}
A - J_1 C & -J_1 \\
-J_2 C & 1 - J_2
\end{bmatrix} \quad (2.18)
\]

is stable. The observer (2.17) will be used as a mechanism to ensure perfect tracking of constant reference signals in the presence of unmeasured disturbances and model errors.

### 2.4 Feedforward Design

As discussed previously, we require that this part of the design make no use of actual plant measurements. Hence, we use the nominal plant transfer function $G_o$. We are interested in designing the feedforward input $u_{ff}$ so that $y_{ff} = G_o \cdot u_{ff}$ tracks the given reference signal with small error (the subscript “ff” refers to the feedforward model).
Of course, we need a mechanism for restraining the bandwidth to a sensible value (otherwise impractical deadbeat type responses will ensue).

### 2.4.1 Constrained Nominal Design

Due to time invariance, we can, without loss of generality, take the current time as 0. Thus we seek to choose \( u_{ff}(0) \). We use a state-space formulation for the nominal model:

\[
\begin{align*}
    x_{ff}^+ &= Ax_{ff} + Bu_{ff} \quad (2.19a) \\
    y_{ff} &= Cx_{ff} \quad (2.19b)
\end{align*}
\]

where \( x_{ff}(0) \) is known. We also assume \( u(-1), u(-2), \ldots \) are given. We seek to minimise a cost function of the form:

\[
J_{ff}^N = \sum_{k=0}^{N} [y_{ff}(k) + \hat{d}_n(k) - \hat{y}^*(k)]^2 + \lambda_{ff} \cdot [u_{ff}(k) - u_{ff}(k-1)]^2 \quad (2.20)
\]

where \( \hat{y}^*(\cdot), \hat{d}_n(\cdot) \) were defined previously, see \( (2.15) \).

In order to make the control sequence \( \{u_{ff}(\cdot)\} \) realistic, it is important that we minimise \( (2.20) \) subject to the appropriate constraints, i.e. \( x_{ff}(k) \in \mathcal{X} \) and \( u_{ff}(k) \in \mathcal{U} \), where \( \mathcal{X}, \mathcal{U} \) are given convex sets which are consistent with the required state and input constraints.

As is usual in MPC, we implement \( u_{ff}(0) \) as the first element of the feedforward sequence. We also pass the entire sequence \( u_{ff}(0), \ldots, u_{ff}(N) \) onto the feedback design stage.

### 2.4.2 Comments

We note that it is essential that the feedforward design is carried out before the feedback design, since the feedback strategy can then robustly stabilise the system in the presence of the given, and known, input \( u_{ff} \).

### 2.5 Feedback Design

We next turn to the design of the feedback component. For this purpose, we define the total plant input as (see Fig. 2.1)

\[
u(k) = u_{fb}(k) + u_{ff}(k) \quad (2.21)
\]

Our goal here is to determine \( u_{fb}(k) \). As before, our main objective is to track \( y^* \). Here, however,
we have two other design goals, namely to achieve robust stability and to reject the (unmeasured) disturbance \( d \). Since we do not measure either the true plant state \( x \) or the disturbance \( d \), we must use the observer given earlier in (2.16), (2.17).

### 2.5.1 Constrained Feedback Design

Again, without loss of generality, we take the current time as \( k = 0 \). We use the observer to provide the current state estimates namely \( \hat{x}(0) \) and \( \hat{d}(0) \). We also use knowledge of \( u_{ff}(0), \ldots, u_{ff}(N) \) from the feedforward design.

We then choose the feedback cost function as:

\[
J_{fb}^N = \sum_{k=0}^{N} [\hat{y}_p(k) + \hat{d}_n(k) + \hat{d}_p(k) - \hat{y}^*(k)]^2 + \lambda_{fb}[u_{fb}(k) - u_{fb}(k-1)]^2
\]

(2.22)

where \( \{\hat{y}_p(\cdot)\} \) and \( \{\hat{d}_p(\cdot)\} \) are predicted quantities satisfying the following prediction model, initialised by the states provided by the observer:

\[
(\hat{x}_p)^+ = A\hat{x}_p + B(u_{fb} + u_{ff}), \quad \hat{x}_p(0) = \hat{x}(0) \quad (2.23)
\]

\[
\hat{y}_p = C\hat{x}_p
\]

\[
(\hat{d}_p)^+ = \hat{d}_p, \quad \hat{d}_p(0) = \hat{d}(0) \quad (2.25)
\]

Similar to the feedforward design, we need to optimise (2.22) subject to constraints. A simple (but nonetheless important) point is that the input constraints are on \( u(k) = u_{fb}(k) + u_{ff}(k) \). Hence, the appropriate constraints are \( \hat{x}_p(k) \in \mathcal{X} \) and \( u_{fb}(k) + u_{ff}(k) \in \mathcal{U} \).

### 2.5.2 Discussion

An important consideration is that, generally, one would expect to choose \( \lambda_{fb} >> \lambda_{ff} \). The reason being that the feedback control uses observations from the true plant via the observer to initialise \( \hat{x}_p(0) \) and \( \hat{d}_p(0) \). Hence, if the true plant differs from the nominal model, then the feedback signal can be potentially destabilising. (Note that this is never an issue for the feedforward component since it does not rely upon plant observations). Therefore, one needs to ensure robust stability for the feedback signal based on the worst possible plant uncertainty, whereas the feedforward signal can be designed in a more optimistic fashion.

In the ideal case (no unmeasured disturbances, no noise, no uncertainty) it will be true that \( \hat{d}(0) = 0, \hat{x}(0) = x_{ff}(0) \). In this case, if \( \lambda_{ff} \) were to be chosen small, then \( y_{ff} \to y^* \). Moreover,
from \((2.23), (2.24)\), we see that \(u_{fb} = 0\) suffices for \(\hat{y}^p \rightarrow y_{ff}\). Hence, we see that \(u_{fb} \simeq 0\) will be a satisfactory choice since the feedforward signal will have “done all the work” and there will be nothing left for the feedback to do. Of course, if there is uncertainty in the form of unmeasured disturbances, noise or model error, then the feedback signal may have some “cleaning up” to do.

It is important to note that the feedforward and feedback MPC designs both have the ability to do reference and disturbance preview, as seen from \((2.20)\) and \((2.22)\). The fact that the feedback controller has to deal with more than just tracking is the reason the feedforward design can improve performance.

An important observation is that, if the feedback design does not “like” the provided sequence of \(u_{ff}\), then it can be completely removed (provided the feedback component has the same available bandwidth as the feedforward component). However, in practice, quite the opposite may occur. That is, the feedback design may decide that \(u_{ff}(k), k = 0 \ldots N\) “does the job” and, in this case, \(u_{fb}(k) = 0\) is a possible outcome.

The need for feedback arises due to the presence of unmeasured disturbances and model uncertainty. This impacts on the choice of bandwidth for the feedback control loop since it has to be “careful” about what it does to the plant. On the other hand, the plant input may have hard constraints on its change rate, which also limits bandwidth. Our proposed design takes into account these constraints, but there is a subtle difference. If the limitation on the bandwidth is mainly because of a hard constraint on the input, then our strategy will have essentially the same performance as if we had not included the feedforward component, because the total plant input change rate will be limited. On the other hand, if the bandwidth limitation is mainly because of robustness to uncertainties, then it will be beneficial to implement the feedforward strategy. This is a consequence of the fact that we separate the regulation problem from the tracking problem.

We denote the combination of feedforward, preview and feedback in MPC as feedforward MPC (FFMPC). An illustration of the strategy is presented in Fig 2.2.

2.6 Examples

In this section we present several examples to illustrate the FFMPC strategy. We utilise the feedback design input weighting \(\lambda_{fb}\) as a mechanism to ensure robustness by imposing bandwidth restrictions on the feedback component. Since reference tracking and measured disturbance rejec-
We assume a model having transfer function

\[ G = \frac{k}{z^d \cdot (z - a)} \]  
(2.26)

where \(k/(1-a)\) is the d.c. gain, \(a\) is the discrete pole and \(d\) is a pure delay. Note that this model class is used extensively in the context of control of chemical processes [109].

We consider two classes of plant uncertainty, namely (i) when the delay can take values \(\{1, 2, 3, 4, 5\}\) but \(a\) and \(k\) are known, and (ii) when the delay \(d\) and gain \(k\) are known but \(a\) can take any value in the range \([0.75, 0.95]\) with uniform probability density. For both scenarios, we firstly show how \(\lambda_{fb}\) affects the tracking dynamics, and secondly, while fixing \(\lambda_{fb}\) to yield a desirable performance, we show the effects that variations in \(\lambda_{ff}\) produce.

2.6.1 Example 1: Model Delay Mismatch

We first consider a case where the real plant has transfer function:

\[ G(z) = \frac{0.05}{z^5 \cdot (z - 0.95)} \]  
(2.27)

while the nominal plant model is

\[ G_o(z) = \frac{0.05}{z^2 \cdot (z - 0.95)} \]  
(2.28)

Note that there is an unmodelled delay of three samples. The prediction and control horizons are taken as \(N_y = N_c = 5\), and the unknown disturbance is chosen as \(d = 0.5\). We choose \(N_p = 4\) and \(\lambda_{ff} = 0\). In the figures presented below, the signals have the following interpretation:
2.6 Examples

<table>
<thead>
<tr>
<th>Sample number k</th>
<th>System output</th>
<th>y*(k)</th>
<th>y(k)</th>
<th>yfbw(k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>150</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>200</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>250</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>300</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

**Figure 2.3:** Output signal for $\lambda_{fb} = 0$ (top) and $\lambda_{fb} = 20$ (bottom)

- $y^*(k)$ is the actual reference signal.
- $y(k)$ is the output of the system with input $u(k)$.
- $y_{fbw}(k)$ is the output of the system without implementing the feedforward component (only feedback).

For the unconstrained case, Fig. 2.3 shows all the aforementioned signals for comparison, when $\lambda_{fb} = 0$ and $\lambda_{fb} = 20$ (a value found experimentally to yield acceptable feedback performance in the face of the given model uncertainty).

Fig. 2.3 clearly shows the success of our proposed MPC framework for dealing with time-varying references. For the case $\lambda_{fb} = 0$, it is clear that the inclusion of our feedforward strategy offers little advantage, since both $y(k)$ and $y_{fbw}(k)$ are essentially the same. Note that the performance of the “closed loop” reference tracking is severely compromised. This is mainly due to the bandwidth of the feedback, which dominates over the feedforward. On the other hand, if we increase the feedback
weight to $\lambda_{fb} = 20$, restricting the bandwidth of the feedback component, then the feedforward signal ensures fast reference tracking as required, whereas the feedback acts slowly and corrects for the unmeasured disturbance.

The sum of quadratic output errors for the FFMPC strategy with $\lambda_{fb} = 0$ and $\lambda_{fb} = 20$ are shown in Table 2.1, where the corresponding results for a feedback MPC controller (without the feedforward component) are also shown. This is a clear demonstration that, for this class of uncertainty, FFMPC yields significant performance improvement.

<table>
<thead>
<tr>
<th></th>
<th>With FF</th>
<th>Without FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_{fb} = 0$</td>
<td>143.1026</td>
<td>143.1026</td>
</tr>
<tr>
<td>$\lambda_{fb} = 20$</td>
<td>29.8876</td>
<td>246.6854</td>
</tr>
</tbody>
</table>

**Table 2.1: Quadratic Tracking Errors**

Behaviour under a class of uncertainty

Here we consider a more general case, where the delay can take values from a given range, namely $d = \{1, 2, 3, 4, 5\}$. We assume knowledge of $k$ and $a$, where $k = 0.05, a = 0.95$. We choose $\lambda_{fb}$ to give “good” transient response for all possible delays. This yields $\lambda_{fb} = 20$ as was found in Section 2.6.1. The reference signal used is shown by the dashed line in Fig. 2.3.

As a measure of the efficacy of utilising FFMPC, we use the following logarithmic cost ratio:

$$J = \log_{10}\left[ \frac{\sum [y_{fbw}(k) - y^*(k)]^2}{\sum [y(k) - y^*(k)]^2} \right]$$  \hspace{1cm} (2.29)

Fig. 2.4 shows the cost function $J$ as a function of $\lambda_{ff}$ for all five possible delay scenarios. The curves represent the percentage of systems that will yield better performance than the one shown, e.g., the bottom curve indicates that 100% of the cases of model uncertainty will have better performance than the one shown by the curve. The next curve (from bottom to top) indicates that 75% of the systems will achieve better performance than that indicated by the curve, and so on.

We see from Fig. 2.4 that, in this case, including feedforward action always yields a performance improvement, e.g., by choosing $\lambda_{ff} = 0$, we obtain a performance improvement of approximately 10 : 1 for all possible delays. We also see that as $\lambda_{ff}$ becomes large the performance gain tends to 1, which means the feedforward signal yields no improvement over the feedback component. This is
2.6 Examples

Figure 2.4: Logarithmic cost function for $\lambda_{fb} = 20$ and $d = \{1, 2, 3, 4, 5\}$.

consistent with the original purpose of $\lambda_{ff}$ to impose bandwidth restrictions. Overall, this scenario creates a compelling case for feedforward action.

2.6.2 Example 2: Effects of Constraints

For the case $\lambda_{fb} = 20, a = 0.95, k = 0.05, d = 5$ (as in (2.27)) and nominal model $G_o$ as in (2.28), Fig. 2.5 shows the responses now including constraints on the magnitude of the total control signal, namely $|u(k)| < 10$. Two cases are shown: (i) where the constraint is not included in the feedforward optimisation, (ii) where the constraint is taken into account. In both cases, the feedback optimisation includes knowledge of this constraint.

The results emphasise the importance of including the constraints, not only in the feedback optimisation stage, but also in the feedforward optimisation. In the top plot, because the constraints are not accounted for in the feedforward optimisation, the optimal value obtained for $u_{ff}$ may be inconsistent with the input constraints to the real plant. This means that the feedforward model will be updated with a false control signal, and hence there will be a mismatch in the subsequent control sequences. It is worth noting that the feedback signal will ultimately correct for this mismatch. In the bottom plot we see that the mismatch is corrected and most of the tracking performance is regained when the constraints are included in the feedforward optimisation stage.
Figure 2.5: Output signals for $\lambda_{fb} = 20$ with constraints $|u(k)| < 10$, (i) without accounting for it in the feedforward optimisation (top) and (ii) accounting for it (bottom).

2.6.3 Example 3: Model Pole Mismatch

In this example, the real plant is assumed to be:

$$G(z) = G_1(z) = \frac{0.15}{z^2 \cdot (z - a)}$$

(2.30)

while the nominal plant model is

$$G_o(z) = \frac{0.15}{z^2 \cdot (z - 0.85)}$$

(2.31)

The prediction and control horizons are taken as $N_p = N_c = 5$, and the unknown disturbance is chosen as $d = 0.5$. Also $N_p = 4$ and $\lambda_{ff} = 0$. No constraints are considered.

For the case $a = 0.95$, Fig. 2.6 shows the various output signals for three different scenarios, namely when increasing $\lambda_{fb}$ to improve tracking performance (as in the previous example) but in the presence of a severe mismatch between the true plant and the actual model used by both the
feedback and feedforward optimisations.

Fig. 2.6 shows that one must be careful about the decision to include this kind of reference feedforward when there is high uncertainty in the accuracy of the model. Unlike the previous example, increasing the parameter $\lambda_{fb}$ provides no enhancement to the tracking of the signal. The case presented here is extreme but representative of the repercussions that may occur. The nominal model has an open loop time constant and d.c. gain that are approximately 3 times those of the true plant.

It is important to note that, in this case, the feedback signal needs to “undo” the signal provided by the feedforward. However, due to the limitation on the feedback bandwidth, it is not possible for it to change as fast as the feedforward signal (since $\lambda_{ff} = 0$), and hence we observe increasing error in the tracking performance as $\lambda_{fb}$ increases.

In this case, the quadratic errors for the output signals (with and without feedforward) are shown in Table 2.2. This clearly shows that, in this case, the feedforward does not help to improve the tracking. This is because of the huge mismatch between the nominal and real plant model.

<table>
<thead>
<tr>
<th>$\lambda_{fb}$</th>
<th>With FF</th>
<th>Without FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>200.6503</td>
<td>104.3128</td>
</tr>
<tr>
<td>40</td>
<td>406.2332</td>
<td>178.8699</td>
</tr>
<tr>
<td>80</td>
<td>574.9992</td>
<td>194.1292</td>
</tr>
</tbody>
</table>

However, if the mismatch were to be less, then the behaviour of this feedforward strategy would not be as detrimental. To illustrate, say that the true plant pole is actually $a = 0.87$, i.e.:

$$G(z) = G_2(z) = \frac{0.15}{z^2 \cdot (z - 0.87)}$$

(2.32)

In this case, the tracking errors that arise are shown in Table 2.3. We note that feedforward action

<table>
<thead>
<tr>
<th>$\lambda_{fb}$</th>
<th>With FF</th>
<th>Without FF</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10.0493</td>
<td>57.7560</td>
</tr>
<tr>
<td>20</td>
<td>18.0209</td>
<td>152.5574</td>
</tr>
<tr>
<td>80</td>
<td>31.3044</td>
<td>216.2909</td>
</tr>
</tbody>
</table>

Table 2.2: Quadratic Tracking Errors for $G_1$

Table 2.3: Quadratic Tracking Errors for $G_2$
Feedforward Model Predictive Control

(with $\lambda_{fb} = 10$, $\lambda_{ff} = 0$) gives a 6 : 1 improvement over the use of feedback alone. This shows that the accuracy of the model can play a large role in the performance outcome of the feedforward strategy. This issue is further illustrated in the following section.

**Behaviour under a class of uncertainty**

We now consider the more general case of uncertainty where $a$ lies in the range $[0.75, 0.95]$ and $k = 0.15, d = 2$ are fixed. The nominal value of $a$ is, as before, 0.85. Again, $\lambda_{fb}$ is chosen to give acceptable transient response for all model uncertainty. This yields $\lambda_{fb} = 80$.

Fig 2.7 plots the logarithmic cost function (2.29) as a function of $\lambda_{ff}$. As previously explained, the curves represent the percentage of systems included in the model uncertainty that have better performance than the one shown by the curve. In this case, from bottom to top, the curves represent the 100%, 95%, 85%, 70% and 0% percentiles.

Unlike the example presented in Section 2.6.1 we see that for the current class of uncertainty (where both the pole location and the d.c. gain change) feedforward action does not ensure a performance gain for all possible model uncertainty.

As in Section 2.6.1 we see that, as $\lambda_{ff}$ increases, the performance ratio tends to 1. Moreover, if we insist that the performance does not deteriorate for all possible model uncertainty then the only choice is $\lambda_{ff} = \infty$. However, if we are prepared to accept a small performance loss in a small percentage of the uncertainty scenarios, then we can gain significantly in the majority of the uncertainty scenarios. Indeed, with $\lambda_{ff} = 0$, we see from Fig. 2.7 that 85% of the uncertainty scenarios experience a gain of $1.3 : 1$ or better resulting from feedforward action, 70% of the uncertainty scenarios experience a gain of $4 : 1$ or better resulting from feedforward action and if the nominal model happens to be true, then a gain of $18 : 1$ results from feedforward action.

We conclude that whether or not feedforward is advantageous depends on the type of application. In some applications, one cannot tolerate a performance loss under any circumstance. Then for the type of uncertainty considered in Section 2.6.3 feedforward would not be a preferable option. However, in other applications, a performance loss in 10% of cases is a reasonable price to pay for a substantial performance gain ($4 : 1$) in 70% of the uncertainty scenarios.
2.7 Conclusions

In this chapter we have discussed the ideas behind the use of feedforward in model predictive control. We have presented the core concepts that underlie the potential advantages of using feedforward. Also, several simulation studies have been provided to illustrate the pros and cons that one might encounter when applying the feedforward scheme.

This chapter is based on work previously published by the author in [48, 22, 20].
Figure 2.6: Output signals for \( \lambda_f = 10, 40, 80 \) (top to bottom) without constraints.
Figure 2.7: Logarithmic cost function for $\lambda_{fb} = 80$ and $a = [0.75, 0.95]$. 
In the previous chapter we introduced the idea of Feedforward MPC. The main idea was to use an open loop optimisation based on selected information and a nominal model, on top of the usual feedback implementation. In this chapter we present an application of the Feedforward MPC ideas to the closed-loop control of blood glucose, using an autonomous artificial pancreas.

The available literature on artificial pancreas contains a vast array of ideas that have been proposed as candidates for the control algorithm [107, 38, 115, 7, 116, 63, 73, 60, 83]. Some of the algorithms include:

- PI/PID (with/without feedforward)
- Linear/Nonlinear MPC
- Periodic Control (run-to-run)
- Fuzzy control

An inherent difficulty in comparing all of the above strategies is that they are often implemented under different conditions. Hence it is difficult to draw definitive conclusions.

In this chapter, we first show that a major source of limitations in this application arises from model uncertainty, specifically due to the combined effect of the insulin infusion system (IIS), the continuous glucose monitor (CGM) and the human glucose regulatory system. We then argue that the uncertainty associated with each of these elements compromises the achievable closed-loop bandwidth, and, in the presence of disturbances, e.g. meal intake or exercise, the closed-loop
response will necessarily be poor. This creates the perfect environment for the Feedforward MPC strategy presented in Chapter 2.

The remainder of this chapter is organised as follows: Section 3.1 briefly reviews blood glucose regulation mechanisms and diabetes. Section 3.2 describes some of the key sources of limitation in artificial pancreas. Section 3.3 describes a realistic, yet simple, model of the blood glucose regulation system, including actuator and sensor dynamics. Section 3.4 presents a linearised form of the model. Section 3.5 develops insights into the fundamental performance limitations that arise from model uncertainty. Section 3.6 explores the use of feedforward MPC. Finally, in Section 3.7, conclusions are drawn.

3.1 Blood Glucose Regulation and Diabetes

3.1.1 Glucose Regulation

Glucose is the primary energy source for all cells in the human body. The condition of too little blood glucose (termed “hypoglycaemia”) can result in unconsciousness and death. The condition of too much blood glucose (termed “hyperglycaemia”), over the long term, can result in serious cardiovascular, renal, and ocular complications; seizures and coma (due to ketoacidosis). A correctly functioning glucose regulatory system maintains blood glucose levels within a narrow range between 80 mg/dL (4.4 mmol/L) and 110 mg/dL (6.1 mmol/L) in the fasting state. After a meal, the blood glucose level may temporarily rise to 140 mg/dL (7.8 mmol/L) or more. Normal glucose regulation returns the blood glucose to the fasting state within a few hours.

Breakdown of glucose is just one of many ways cells produce ATP (adenosine triphosphate). Red blood cells (erythrocytes) and the brain are exclusively reliant on glucose as their sole source of energy/ATP. Hence, there is a continuous process of glucose production and consumption which needs to be regulated.

It is mainly the pancreas and liver that provide the regulatory functionality. Anatomically, the pancreas and liver are closely connected via the portal vein. The proximity of these two organs implies that, in healthy individuals, there is very little lag between the sensing and actuation of regulatory commands.

The pancreas is composed two types of functional tissue. The first are groups of cells responsible for enzyme production termed “acini”. The second grouping of cells are called “islets of Langerhans”
and consist of $\alpha$, $\beta$ and $\gamma$ cells. These cells are responsible for the production of glucose regulatory hormones. The two main hormones are “insulin” and “glucagon”. Insulin is produced by the $\beta$-cells while glucagon is produced by the $\alpha$-cells. Approximately 80% of the liver consists of cells known as “hepatocytes”. These cells have special capabilities in relation to storage and production of glucose.

In a healthy individual, insulin is released in response to increased glucose levels. It binds to receptors on insulin-dependent cells allowing them to use glucose as their primary source of energy. In concert, insulin also stimulates specific liver cells (hepatocytes) and muscle cells (myocytes) to collect and store glucose in the form of “glycogen”, through a process known as “glycogenesis”. Conversely, glucagon is released in response to falling glucose levels. It binds to receptors on hepatocytes instructing them to activate the process “glycogenolysis” which breaks down the glycogen molecules into glucose and releases it into the bloodstream. We note that it is only hepatic cells that have this mechanism. Muscle cells also store glycogen but they do not have glucagon receptors nor the functionality to return any of the stored glucose to the blood stream. When glycogen stores are depleted, glucagon stimulates the process of “gluconeogenesis”. This process occurs in the liver and, to a much smaller extent, in the kidneys and intestines. As the name implies, “new glucose” is generated but, in this case, it comes from the conversion of non-carbohydrate sources.

The production of glucose may come from exogenous or endogenous mechanisms as described below.

(a) **Exogenous Mechanism**

Food is ingested and initially broken down by stomach acid and enzymes. Eventually, after further processing, the food is converted into macromolecules and absorbed into the blood by the intestines. Absorbed glucose is delivered firstly to the liver, by the portal vein, then to the rest of the body.

The amount of glucose, and the rate at which it enters the blood stream, is highly dependent on the types of food ingested. The amount of carbohydrate contained in foods determines the amount of glucose which will ultimately be absorbed. The glycaemic index (GI) reflects the rate at which glucose will be absorbed. Other significant factors influencing the rate at which glucose is absorbed are: the fat/carbohydrate/protein ratio of a meal; gastric emptying times and; blood flow to the intestine and activity level.

All of these uncertainties result in a glucose spike, of variable intensity and duration, which the regulatory system must efficiently control.
(b) **Endogenous Mechanisms**

Endogenous glucose production is the other major mechanism that introduces glucose to the bloodstream. There are two processes that re-introduce glucose into the bloodstream. Those processes, as described above, are performed by the liver and are named “glycogenolysis” and “gluconeogenesis”. Both glycogenolysis and gluconeogenesis are activated by an increased blood glucagon to insulin ratio.

From the above description of glucose production and consumption, we note that both insulin and glucagon are necessarily produced and cleared. It is the ratio of the two that determines whether the net result is a glucose increase, decrease or equilibrium (homeostasis). In the case of equilibrium, the levels of insulin and glucagon being produced are referred to as the basal rate of glucose homeostasis.

### 3.1.2 Glucose Regulation Disorders

The most common disorder is “Diabetes Mellitus” (diabetes) which manifests in two distinct types. Type 1 diabetes occurs due to an auto-immune response which attacks and kills the β-cells of the pancreas. Type 2 diabetes has been correlated to lifestyle and hereditary factors.

Type 1 diabetes renders the body incapable of producing any of the hormones normally produced by the β-cells of the pancreas. The consequences are as follows: since no insulin is produced, insulin dependent cells no longer absorb glucose and hence starve. These cells emit starvation signals which cause the body to release, or produce, more glucose, further increasing blood glucose levels. In addition, it is believed that there are intracellular mechanisms that inhibit glucagon production when insulin is being produced [80]. Hence, normal inhibition of glucagon does not occur even if insulin levels are increased via external injection.

Type 2 diabetes occurs due to the pancreas’ reduced ability to produce insulin or due to, what is commonly known as, insulin resistance. Although the exact cause is not known, cells become increasingly resistant to insulin levels so that increasing amounts of insulin are required for cells to absorb the same amount of glucose. We note that it has also been shown that type 1 diabetes are also prone to the development of insulin resistance [123].
3.2 Fundamental Limitations Due to Model Uncertainty in Artificial Pancreas

A central theme in the current chapter is that of fundamental limitations and how feedforward MPC can be used to mitigate some of these limitations. In this context, it is relevant to point to the extensive literature that exists on this topic in the systems and control area, see for example [110]. Indeed, it has been a source of major research effort over the past five decades.

An admittedly simplified summary of the fundamental limitations literature is given below. One way of thinking about the control problem is that of constructing an inverse for the plant. Hence fundamental limitations are typically ascribable to inherent difficulties involved in building an inverse [49]. Not surprisingly, pure delays and (multivariable) non-minimum phase zeros are recurring themes in the fundamental limitations literature. Unstable poles are also important since those cannot be cancelled by the controller leading to the need for interpolation constraints. Another source of limitations is model uncertainty, which again inhibits finding a single inverse that applies to all possible models.

In the artificial pancreas problem, model uncertainty is the major source of limitations. Specifically, relative model errors directly translate into relative response errors. This places a lower limit on the achievable response time and hence, on the achievable performance. Importantly, note that these limitations are unavoidable and have nothing to do with the specific control algorithm.

For the sake of simplicity, we distinguish three components in the closed-loop control system of blood glucose, namely:

- The actuator, including the insulin pump and the transportation of insulin from the injection point to the blood stream.
- The blood glucose system, including the biomechanics that link insulin that appear in the plasma to the glucose response.
- The sensor, including the glucose transportation dynamics to the tissue and the sensor dynamics. It is assumed that a continuous glucose monitor is used in the artificial pancreas.

An important observation is that each of the aforementioned components has major model uncertainty. For example,

- The action of the insulin pump can depend on a host of factors. From anything like posi-
tioning and angle of insertion of the needle to air pressure changes that modify the way the tissue absorbs the insulin [71].

- There exists extremely high inter and intra patient variability that affect the dynamics of glucose regulation. These dynamics are also affected by stress and mood. There are also external disturbances such as food and exercise.

- Continuous glucose monitors usually have their own correction algorithms embedded, which are not publicly available, i.e. the reading is actually a filtered version of the actual measurement. In addition, they need to be periodically calibrated and even body movement can affect the resulting measurement.

We will use a linear model for each of the components described above. The use of a nonlinear model is unlikely to improve the situation since the inherent variability of the parameters also applies in the nonlinear case. Thus nonlinear models give additional insight into the underlying dynamics but do not reduce the inter and intra model variability which arises due to other factors.

### 3.3 Type 1 Diabetes Model

In this section we describe a simple, yet realistic model. The model was presented in [66] and is based on Bergman’s minimal model [8]. One advantage of such a model, in the current context, is that [66] reports the identified parameters for a set of subjects, including intraday variability for some of them (see Table 3.1), hence providing a realistic framework for the characterisation of the uncertainty.

#### 3.3.1 Actuator

The following are the associated actuator dynamics:

\[
\frac{dI_{SC}(t)}{dt} = -\frac{1}{\tau_1} \cdot I_{SC}(t) + \frac{1}{\tau_1} \cdot \frac{ID(t)}{C_i} \tag{3.1}
\]

\[
\frac{dI_p(t)}{dt} = -\frac{1}{\tau_2} \cdot I_p(t) + \frac{1}{\tau_2} \cdot I_{SC}(t) \tag{3.2}
\]

where

- \(I_{SC}(t)\): subcutaneous insulin concentration
- \(I_p(t)\): plasma insulin concentration
3.3 Type 1 Diabetes Model

- $ID(t)$: subcutaneous insulin delivery (input)
- $\tau_1 [\text{min}], \tau_2 [\text{min}]$: time constants
- $C_i [\text{ml/min}]$: insulin clearance

3.3.2 Patient

The following are the associated patient dynamics:

$$\frac{dI_{EFF}(t)}{dt} = -p_2 \cdot I_{EFF}(t) + p_2 \cdot S_I \cdot I_p(t)$$  \hspace{1cm} (3.3)
$$\frac{dG(t)}{dt} = -(GEZI + I_{EFF}(t))G(t) + EGP + R_A(t)$$  \hspace{1cm} (3.4)

where

- $I_{EFF}(t) [\text{min}^{-1}]$: insulin effect on plasma glucose
- $G(t) [\text{mg/dL}]$: plasma glucose
- $R_A(t) [\text{mg/dL/min}]$: glucose rate of absorption from meals (disturbance)
- $S_I [\text{ml/µU}]$: insulin sensitivity.
- $GEZI [\text{min}^{-1}]$: glucose effect to increase glucose uptake and lower endogenous glucose production at zero insulin.
- $EGP [\text{mg/dL/min}]$: endogenous glucose production.

3.3.3 Sensor

The following are typical sensor dynamics:

$$\frac{dG_{ISF}(t)}{dt} = -\frac{1}{\tau_{SEN}} \cdot G_{ISF}(t) + \frac{1}{\tau_{SEN}} \cdot G(t)$$  \hspace{1cm} (3.5)

where

- $G_{ISF}(t)$: interstitial fluid glucose concentration.
- $\tau_{SEN} [\text{min}]$: sensor time constant.
Food absorption

The following are the associated food absorption dynamics:

\[ R_A(t) = \frac{t \cdot e^{-\frac{t}{\tau_m}}}{V_G \cdot \tau_m^2} \cdot C_H(t), \quad t \geq 0 \]  \hspace{1cm} (3.6)

where

- \( C_H(t) = C_H \delta(t) \): where \( C_H \) are the consumed carbohydrates in [mg] and \( \delta(t) \) is a unit impulse at \( t = 0 \).
- \( V_G \) [dl]: distribution volume for glucose equilibria
- \( \tau_m \) [min]: absorption peak time.

Exercise

Exercise is also an important factor but is not included in the above model. We will focus primarily on food disturbances but the core conclusions should hold when exercise is also added to the problem.

Model Approximation

The model described in Section 3.3 is linear save for equation (3.4). Assuming that \( G(t) \) is regulated close to some nominal value \( G_o \), and \( I_{EFF}(t) \) is regulated close to \( I_{EFF_o} \), then the equation can be linearised as follows:

\[ \frac{dG(t)}{dt} = A_N \cdot G(t) + B_N \cdot I_{EFF}(t) + E_N + R_A(t), \]  \hspace{1cm} (3.7)

where

\[
A_N = -(GEZI + I_{EFF_o}) \\
B_N = -G_o \\
E_N = EGP + I_{EFF_o} \cdot G_o
\]

Also, we recognise equation (3.6) as the impulse response of a simple double pole system where the pole is located at \(-1/\tau_m\). With these observations, the model can be represented in block diagram form as in Fig. 3.1.

In Fig 3.1 we use the notation,
3.5 Insights into Performance Limitations

Figure 3.1: Block diagram of approximated model

- $u(t)$: input (subcutaneous delivery rate)
- $i(t)$: intermediate variable (rate of change of plasma insulin concentration)
- $d(t)$: disturbance (rate of consumed carbohydrates)
- $E$: constant (including EGP rate)
- $g(t)$: performance variable (plasma glucose concentration)
- $y(t)$: measured output (interstitial fluid glucose concentration)

In addition, $G_A$, $G_B$, $G_D$, $G_P$, $G_S$ are five linear transfer functions which take the form:

\[
G_A = \frac{k_A}{(\tau_1 s + 1)(\tau_2 s + 1)}
\]

\[
G_B = \frac{S_I}{1/p_2 \cdot s + 1}
\]

\[
G_D = \frac{k_D}{(\tau_m s + 1)^2}
\]

\[
G_P = \frac{1}{(\tau_3 s + 1)}
\]

\[
G_S = \frac{1}{\tau_{SEN} s + 1}
\]

The range of the original parameters is given in Table 3.1 (see [66]). It is immediately evident that there is a huge range in behaviour. We will argue below that this variability renders high gain feedback impossible.

3.5 Insights into Performance Limitations

To gain insight into the effect of model uncertainty in limiting performance, we will use a simple example as illustration. We therefore represent each of the blocks in Fig. 3.1 as a first order lag.
We assume that each transfer function has the same nominal time constant of 30 [min] with a range from 10 to 50 [min]. We note that this is consistent with the parameter variation given above. In the sequel, we focus on the total relative model error defined by $G_\Delta = (G_T - G_T^n)/G_T^n$, where $G_T$ is the composite transfer function $G_A \cdot G_B \cdot G_P \cdot G_S$ and where $G_T^n$ is its nominal value. Figure 3.2 shows the magnitude of $G_\Delta$ as a function of $\omega$ (in radians per minute) for the extreme values of model uncertainty.

From robust control theory [127, 49], we know that a sufficient condition for robust stability is that the product of the closed loop nominal complementary sensitivity function $T_o$ and the relative model error $G_\Delta$ should have magnitude less than one at all frequencies. Indeed, a margin is required so that $|T_oG_\Delta|$ should be much less than 1. For the purpose of our example, say 0.5 at all frequencies.

On the other hand, the closed loop bandwidth can be defined [49] as the frequency $\omega_{BW}$ where the magnitude of $T_o$ falls below 0.7, i.e. $|T_o(\omega_{BW})| < 0.7$. Hence, in our example, to achieve robust stability the closed loop bandwidth cannot be greater than the frequency where $|G_\Delta|$ reaches 0.7, i.e. $|T_o(\omega_{BW})G_\Delta(\omega_{BW})| < 0.5$. Based on the above considerations, we see from Fig. 3.2 that the practical closed loop bandwidth is limited to about $2 \cdot 10^{-2}$ [rad/min].

Finally, we note that, since we have assumed a nominal time constant of 30 [min] for each block, the nominal bandwidth of the disturbance model $G_D(s)$ is $\omega_d = 1/30 = 3.3 \cdot 10^{-2}$ [rad/min]. Hence, any closed loop controller that achieves robust stability for all possible parametric variations can only eliminate the disturbance up to about 2 [rad/min]. This severe limitation in performance is set in place only by the need to achieve robust stability in the presence of uncertainty.
3.6 Illustrating the advantages of Feedforward MPC

Here we utilise the linear model described in Section 3.4. We map the parameter ranges quoted in Section 3.3 to the equivalent parameters in the linear model – see Table 3.1. We then consider all different combinations of the extreme and nominal values of the parameters.

A linear observer was used to estimate the model states. The poles of the observer were placed in the range (0.8, 0.84). This yields an observer response time of approximately 20 minutes. The system was sampled with period $h = 1$ [min]. Two designs were carried out based on the nominal parameter values as follows.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Minimum Value</th>
<th>Nominal</th>
<th>Maximum Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_1$</td>
<td>41</td>
<td>85</td>
<td>131</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>10</td>
<td>40</td>
<td>70</td>
</tr>
<tr>
<td>$\tau_{SEN}$</td>
<td>10</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>$p_2$</td>
<td>$9.5 \cdot 10^{-3}$</td>
<td>$1.6 \cdot 10^{-2}$</td>
<td>$2.33 \cdot 10^{-2}$</td>
</tr>
<tr>
<td>$C_l$</td>
<td>540</td>
<td>1250</td>
<td>2010</td>
</tr>
<tr>
<td>$S_l$</td>
<td>$9.64 \cdot 10^{-5}$</td>
<td>$9 \cdot 10^{-4}$</td>
<td>$1.73 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>$E_{GP}$</td>
<td>0.6</td>
<td>2</td>
<td>3.45</td>
</tr>
<tr>
<td>$GEZI$</td>
<td>$10^{-8}$</td>
<td>$3.19 \cdot 10^{-3}$</td>
<td>$6.39 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>$V_G$</td>
<td>104</td>
<td>220</td>
<td>337</td>
</tr>
<tr>
<td>$\tau_m$</td>
<td>21</td>
<td>126</td>
<td>231</td>
</tr>
</tbody>
</table>

Table 3.1: Range of parameter values

Remark 3.5.1. The reader may wonder why this poor performance has resulted. The key issue is that the parametric uncertainty, in all four transfer functions $G_A$, $G_B$, $G_P$, $G_s$, means that high gain feedback cannot be used whilst ensuring stability for all possible parametric values. The situation would be very different if $\ddot{d}(t)$ in Fig. 3.1 could be measured. If this were possible, then high-gain feedback could be used from $\ddot{d}(t)$ to $u(t)$ which would have a major impact on the effect of the disturbance. However, the additional uncertain phase shift introduced by $G_P$ and $G_S$ makes this ideal result unachievable in practice.
3.6.1 Robust Feedback MPC

A robust MPC controller incorporating integral action was designed so that, with a food bolus of 60 \( g \), the system remains stable and the blood glucose concentration never falls below 50 \( \text{mg/dL} \) for all possible parameter values. Note that this is a reasonable lower limit to avoid the potentially catastrophic impact of hypoglycaemia. This design will be referred to as MPC in the sequel.

Remark 3.6.1. Note that the MPC algorithm used in the current study is based on a quadratic cost function. In future work, it would be sensible to use a different cost in the hypoglycaemic region to that used in the hyperglycaemic region.

3.6.2 Feedforward MPC (FFMPC)

The ideas of Chapter 2 were used to design a feedforward component, separate from the feedback MPC previously designed. Based on the nominal model, an MPC problem is solved for the feedforward component. The solution to this problem is then fed to a separate robust MPC problem for the feedback component. The feedforward controller is based only on information of the nominal model and the disturbance, whereas the feedback controller is based on steady state reference values, estimated states from output information, the nominal model, the disturbance and the control input from the feedforward controller. As in Chapter 2, the design that includes a feedback and a feedforward controller will be referred to as FFMPC.

For the example below, a very high bandwidth was assigned to the feedforward controller. On the other hand, the feedback controller bandwidth was reduced so as to achieve stability and satisfactory performance for all possible parameter values. Figure 3.3 shows simulated results.
following a food bolus of \( 60 \, [g] \) at time \( t = 700 \, [min] \). Figure 3.3 shows the output response of the nominal system for both robust feedback MPC (solid line), FFMPC (dashed line) and No Control case, i.e. the control input was constant and set to the steady state value corresponding to an output of \( 70 \, [mg/dL] \). Note that the robust MPC design is marginally better than having no feedback controller. This is due to the need to achieve stability for all parameter values.

### 3.6.3 Performance quantification

To provide a quantitative performance comparison, Table 3.2 presents the mean square error around a setpoint of \( 70 \, [mg/dL] \). The cases shown represent the nominal, best, worst and average performance over all possible combinations of the parameter values. The model used for both the controller and the observer is always based on the nominal parameters.

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
<th>Best</th>
<th>Worst</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Control</td>
<td>169.8</td>
<td>11.4</td>
<td>14936.6</td>
<td>2284.9</td>
</tr>
<tr>
<td>MPC</td>
<td>164.4</td>
<td>11.6</td>
<td>19345.1</td>
<td>1882.6</td>
</tr>
<tr>
<td>FFMPC</td>
<td>33.8</td>
<td>1.8</td>
<td>22724.8</td>
<td>1829.3</td>
</tr>
</tbody>
</table>

**Table 3.2: Mean Squared Error Performance**

It can be seen from Table 3.2 that FFMPC provides an improvement of at least 5:1 over MPC in the nominal and best performance scenarios. Moreover, Table 3.2 shows the FFMPC strategy provides, on average, an improvement in performance.

**Remark 3.6.2.** Note that the robust MPC design used above is “universal” in the sense that one controller is asked to achieve stability and satisfactory performance for all patients based on a single nominal model. Any sort of adaptive control strategy [54] would allow the nominal model to be tailored to each individual. This would undoubtedly improve the performance, specially for the extreme cases where the use of feedforward is detrimental. However, variability in the total model would still exist due to the actuator, sensor and intraday variability of the patient. Thus we anticipate that feedforward would still be of great benefit.

### 3.6.4 Probabilistic Analysis

Consider a uniform probability distribution on the set of parameters variations, i.e. every combination of parameters has the same probability of being the real system. We can then determine the
corresponding cumulative distribution of performance. The results are presented in Figures 3.4 and 3.5, where three curves are shown, namely the cumulative probability distribution when FFMPC, MPC and No Control scenarios are considered. The plots should be interpreted as follows: the pair \((x_o, y_o)\) describes the probability \(y_o\) that a system will have performance equal or better than \(x_o\), for any given curve.

We conclude from Figure 3.5 that, under uniform probabilistic uncertainty, feedforward provides better performance than feedback in 90% of cases. Moreover, even in the remainder 10% of cases, the feedforward performance is at least 1.5 times better than no control at all.

**Remark 3.6.3.** We have assumed a uniform distribution of the parameters as a worst case scenario since any distribution that has a concentration around the nominal values,ug e.g. a normal distribution, will increase the performance gain obtained by using feedforward. Figure 3.6 presents the cumulative probability function for a gaussian distribution of the parameters. The extreme values are located at two standard deviations from the nominal value (mean) of each parameter. It is then seen that FFMPC achieves 5 times better performance in 50% of the cases and 3 times better performance in 80% of the cases, when compared to an MPC strategy without feedforward.
3.6 Illustrating the advantages of Feedforward MPC

The results shown above have not considered the effect of constraints. It is sensible to expect certain degradation in the control performance depending on the tightness of the constraints. We will examine the nominal case as an illustration. Table 3.3 presents the achieved mean squared errors as the upper bound on the input is lowered. Fig. 3.7 shows the input for the case $0 < u(t) < 4 \cdot 10^4$. It is clear that the ability of the feedforward component to improve on the regulation of blood glucose decreases as tighter constraints are imposed.

**Figure 3.5:** Cumulative Probability for Uniform Distribution (zoomed)

**Figure 3.6:** Cumulative Probability for Gaussian Distribution (zoomed)

3.6.5 Adding Constraints
3. Feedforward MPC in Artificial Pancreas

<table>
<thead>
<tr>
<th>Upper Bound</th>
<th>$10^6$</th>
<th>$2 \cdot 10^5$</th>
<th>$5 \cdot 10^4$</th>
<th>$4 \cdot 10^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Control</td>
<td>16811.9</td>
<td>16811.9</td>
<td>16811.9</td>
<td>16811.9</td>
</tr>
<tr>
<td>MPC</td>
<td>14903.3</td>
<td>14903.3</td>
<td>14906.7</td>
<td>15908.4</td>
</tr>
<tr>
<td>FFMPC</td>
<td>974.0</td>
<td>3087.3</td>
<td>14079.6</td>
<td>15843.6</td>
</tr>
</tbody>
</table>

Table 3.3: MSE Performance with Constraints

![Figure 3.7](image)

Figure 3.7: Insulin delivery with constraints $0 < u(t) < 4 \cdot 10^4$

3.7 Conclusions

This chapter has presented a preliminary study of the application of FFMPC to the problem of feedback control of blood glucose. The key conclusion is that, for realistic variations in the model parameters of the actuator, plant and sensor, the achievable closed loop bandwidth is severely restricted. On the other hand, feedforward control does not suffer from this limitation. Thus feedforward provides the potential for major performance improvements. Of course, the key caveat in the case of feedforward is that measurements of the disturbances, e.g. food intake and exercise, need to be available.

Possible ways to move forward and provide a comprehensive solution of the blood glucose control problem should include (i) calibration of the models to a specific individual, so as to reduce inter patient variability and (ii) inclusion of an adaptive algorithm, to reduce the intra patient variability.

This chapter is based on work previously published by the author in [19].

Details of the cost function and the associated design parameters cannot be disclosed due to confidentiality reasons.
The ideas presented so far have addressed the issue of improving closed-loop performance in the presence of uncertainty by modifying and augmenting the controller structure. In this chapter we turn our attention to the use of observers and the role they play in improving closed-loop performance.

More specifically, in this chapter we establish a relationship between the complementary sensitivity function of output feedback control and two other sensitivity functions, namely a complementary sensitivity function associated with feedback control based on an auxiliary variable, and a complementary sensitivity function associated with an observer.

There exists a substantial body of literature on the role of sensitivity and complementary sensitivity functions in linear output feedback control, see e.g. [9, 62, 31, 49]. For example, it is well known that key performance limitations, including Bode integral constraints, apply to these functions [42, 57, 28, 41]. It is also known that the complementary sensitivity function is central to robustness considerations. More recently, a parallel theory for linear observers has been developed. As for the control problem, key performance limitations are associated with filtering sensitivity functions, including Bode-type sensitivity trade-offs [111, 110, 50, 53]. In this chapter we introduce sensitivity functions for feedback control based on an auxiliary variable and explore the relationship between the aforementioned sensitivity functions on the one hand, and the sensitivity functions associated with an observer and the sensitivity functions associated with output feedback control on the other. We restrict our analysis to the single-input single-output case.

The result presented in this chapter embellishes traditional ideas of separation in linear feedback
control. The existing literature discusses two types of separation [45, 44], namely (i) that the closed loop poles are the union of the poles associated with state feedback and those of the observer [49, Theorem 18.1] and (ii) that certainty equivalence holds for the Linear Quadratic Gaussian optimal control problem [11, Section 8.2]. The result presented here can be thought of as a third kind of separation: the complementary sensitivity of output feedback control can be separated into the product of a complementary sensitivity function associated with auxiliary variable feedback and a complementary sensitivity function associated with an observer.

We will show that this separation result can be used to improve the closed-loop performance in the face of certain types of uncertainty.

The remainder of this chapter is organised as follows: In Section 4.1 we describe a class of models and introduce the problem of interest. In Section 4.2 we introduce auxiliary variable feedback and define associated sensitivity functions. In Section 4.3 we define a class of observers. In Section 4.4 we review observer sensitivity functions. In Section 4.5 we consider output feedback control and recall the usual sensitivity functions. In Section 4.6 we present the main result linking all of the above sensitivity functions. In Section 4.7 we consider the special case of a state space model. In Section 4.8 we briefly discuss implications of the results. Finally, conclusions are presented in Section 4.9.

4.1 System Model and Problem Description

Consider a linear time-invariant single-input single-output system described in transfer function form by

\[ y^o = T^o_{yu} \cdot u + T^o_{yw} \cdot w \]  

(4.1)

where \( y^o \in \mathbb{R} \), \( u \in \mathbb{R} \), \( w \in \mathbb{R} \) denote output, input and process noise, respectively. In addition, \( T^o_{ba} \) denotes the direct linear mapping from signal \( a \) to signal \( b \).

Consider an auxiliary output \( \eta \in \mathbb{R}^p \), given by

\[ \eta = T^o_{\eta u} \cdot u + T^o_{\eta w} \cdot w, \]  

(4.2)

having the property that \( y^o = T^o_{y\eta} \cdot \eta \). The quantity \( \eta \) can be a performance variable, some suitable combination of the states, or some other variable of interest. We have the following relationships:

\[ T^o_{yu} = T^o_{y\eta} \cdot T^o_{\eta u} \]  

(4.3)

\[ T^o_{yw} = T^o_{y\eta} \cdot T^o_{\eta w} \]  

(4.4)
In the sequel, we will consider three problems:

Problem 1: The auxiliary variable $\eta$ is used for feedback control.

Problem 2: The variable $\eta$ is not directly measured but an estimate $\hat{\eta}$ is obtained via an unbiased linear observer.

Problem 3: The estimate $\hat{\eta}$ is used, in place of $\eta$, for feedback control.

**Remark 4.1.1.** Note that, in Problem 2, only the estimation problem is studied and no feedback control is considered, i.e. the control input need bear no relationship to either $\eta$ or $\hat{\eta}$.

**Remark 4.1.2.** The use of an unbiased observer in Problem 2 is crucial to the subsequent development.

We will define sensitivity functions for each of the three problems described above. The key result in this chapter is to develop a relationship between the aforementioned sensitivity functions. The three problems are schematically illustrated in Figures 4.1, 4.2, 4.3, respectively.

### 4.2 Auxiliary Variable Feedback

In this section we address Problem 1, as shown schematically in Figure 4.1, where $T_{\eta \eta}^o$ is a $(1 \times p)$ linear time invariant transfer function. Note that the input signal $u$ satisfies:

$$u = -z + n_w = -T_{\eta \eta}^o \cdot \eta + n_w$$  \hspace{1cm} (4.5)
4. Connecting Control and Filtering Sensitivity Functions

**Definition 4.2.1.** The auxiliary variable feedback complementary sensitivity function and sensitivity function are respectively defined to be:

\[
T_{AF}^o = \frac{T_{zu}^o}{1 + T_{zu}^o} \quad (4.6)
\]

\[
S_{AF}^o = \frac{1}{1 + T_{zu}^o} = 1 - T_{AF}^o \quad (4.7)
\]

where \( T_{zu}^o \) is the transfer function from \( u \) to \( z \), i.e., \( T_{zu}^o = T_{z\eta}^o T_{\eta u}^o \).

Consider the feedback system in Fig. 4.1 where the input variable \( u^o \) is corrupted by input noise \( n_u \). Then, the following closed loop transfer functions apply:

\[
\frac{u^o}{n_u} = -T_{AF}^o
\]

\[
\frac{u}{n_u} = S_{AF}^o
\]

Clearly, \( T_{AF}^o, S_{AF}^o \) are standard sensitivity functions which satisfy properties analogous to those found in output feedback control.

### 4.3 A Class of Observers

In this section, we deal with Problem 2. The variable \( \eta \) is not directly measured, but it is estimated by a linear observer of the following generic form:

\[
\dot{\eta} = F_y \cdot y + F_u \cdot u \quad (4.8)
\]
where $F_y$, $F_u$ are linear ($p \times 1$) stable transfer functions and where $y$ is the measured output. We assume that $y$ is related to $y^o$ by:

$$y = y^o + v$$

(4.9)

for some measurement noise process $v$.

We restrict the observer (4.8) to belong to the class of unbiased observers [110]. Observers of this class have the property that the transfer function from $u$ to $\hat{\eta}$ is the same as that from $u$ to $\eta$, i.e. the estimation error does not depend on the input [51]. Unbiased observers include, as special cases, reduced order observers (Luenberger observers) and full order observers, e.g. Kalman observers. However, unbiased observers are more general since, for instance, the degree of an unbiased observer need not be the same degree as the plant. In fact, unbiased observers can have arbitrary state dimension.

It is readily seen that a necessary and sufficient condition for (4.8) to be unbiased, i.e. $T_{\hat{\eta}u}^o = T_{\eta u}^o$, is that the following identity holds – see [110] Lemma 7.3.1:

$$F_y \cdot T_{\hat{\eta}u}^o + F_u = T_{\eta u}^o$$

(4.10)

We also define the observer error as

$$\tilde{\eta} = \hat{\eta} - \eta$$

(4.11)
Using (4.8), (4.9), and (4.1), \( \hat{\eta} \) can be written as

\[
\hat{\eta} = F_y \cdot (T_{yu}^o u + T_{yw}^o w + v) + F_u u
\]

\[
= (F_y T_{yu}^o + F_u) \cdot u + F_y T_{yw}^o w + F_y v
\]

Also, using (4.2) and (4.10), \( \eta \) can be written as

\[
\eta = (F_y T_{yu}^o + F_u) \cdot u + T_{\eta w}^o w
\]

It follows that

\[
\tilde{\eta} = T_{\tilde{\eta} v} \cdot v + T_{\tilde{\eta} w} \cdot w
\]

(4.12)

where

\[
T_{\tilde{\eta} v} = F_y
\]

(4.13)

\[
T_{\tilde{\eta} w} = F_y T_{yw}^o - T_{\eta w}^o
\]

(4.14)

4.4 Observer Sensitivity Functions

It would be tempting to use \( T_{\tilde{\eta} w} \) and \( T_{\tilde{\eta} v} \) to define observer sensitivity functions. However, we note that the units of \( \eta, w \) and \( v \) can all be different. This would negate a key property of sensitivity functions, e.g. the sum of sensitivity and complementary sensitivity is unity. In [50], a normalisation step was used to obtain dimensionless sensitivity functions. We follow this idea here and introduce “normalised” sensitivity functions as follows.

**Definition 4.4.1.** The observer complementary sensitivity is defined as

\[
M = N_1 \cdot T_{\tilde{\eta} v}
\]

(4.15)

where \( N_1 = T_{\eta y}^o \), and the observer sensitivity function is defined as

\[
P = - N_2 \cdot T_{\tilde{\eta} w}
\]

(4.16)

where \( N_2 = T_{\eta y}^o / (T_{\eta y}^o T_{\eta w}^o) \).

**Remark 4.4.1.** The filtering sensitivity \( P \) represents the relative effect of the process noise \( w \) on the estimation error \( \tilde{\eta} \), while the filtering complementary sensitivity \( M \) represents the relative effect of the measurement noise \( v \) on the estimation error \( \tilde{\eta} \). □

Several properties [50] follow immediately from Definition 4.4.1. For example, we have:
**Lemma 4.4.1.** The filtering sensitivity and complementary sensitivity functions satisfy

\[ P + M = 1 \]  
(4.17)

**Proof.** From equations (4.13), (4.14) and Definition 4.4.1 we can write

\[
M = T_{y\eta} F_y \\
P = - \frac{T_{y\eta} \cdot (F_y T_{yw} - T_{y\eta} T_{\eta w})}{T_{y\eta} T_{\eta w}}
\]

where \( T_{yw} = T_{y\eta} T_{\eta w} \), so that \( P \) becomes

\[
P = - \frac{T_{y\eta} \cdot F_y T_{y\eta} T_{\eta w}}{T_{y\eta} T_{\eta w}} + 1 \\
= - T_{y\eta} \cdot F_y + 1
\]

from which the result follows immediately.

It can also be shown that \( P \) and \( M \) satisfy Bode type integral constraints analogous to those that hold for the control sensitivity and complementary sensitivity functions – see [110, Section 8.2].

### 4.5 Output Feedback

In this section we consider Problem 3. In particular, we examine more carefully the special case where \( \hat{\eta} \) is used in an output feedback control law of the form

\[
u = - \hat{z} + n_u = - T_{z\eta} \cdot \hat{\eta} + n_u.
\]  
(4.18)

We first recall the following definition from the theory of output feedback control [110, Section 2.2]:

**Definition 4.5.1.** The nominal complementary sensitivity function of output feedback control, \( T^o \), is the transfer function relating \(-v\) to \( y^o\) under the action of (4.18), i.e.

\[
T^o = \frac{T_{z\eta} F_y T_{yu}}{1 + T_{z\eta} F_y T_{yu} + T_{z\eta} F_u}
\]  
(4.19)

which is easily obtained by considering (4.18), (4.8) and (4.9) and the fact that \( y^o = T_{yu} u \). In addition, the nominal sensitivity function \( S^o \) is defined to be

\[
S^o = 1 - T^o
\]  
(4.20)

□□□
4.6 Main Result

The key result of this chapter is to relate the various sensitivity functions introduced above. In particular, we have:

**Theorem 4.6.1.** The filtering, auxiliary-variable feedback and output feedback complementary sensitivity functions satisfy

\[ T^o = T^o_{AF} M \]  

(4.21)

**Proof.** From (4.19) and using the unbiasedness constraint (4.10) we have

\[ T^o = \frac{T^o_{z\eta} F_y T^o_{yu}}{1 + T^o_{z\eta} F_y T^o_{yu} + T^o_{z\eta} F_u} = \frac{T^o_{z\eta} F_y T^o_{yu}}{1 + T^o_{z\eta} T^o_{yu}} \]

Using the definition of \( T^o_{AF} \), given in (4.6), we obtain

\[ T^o = T^o_{AF} \cdot \frac{T^o_{z\eta} F_y T^o_{yu}}{T^o_{z\eta} T^o_{yu}} \]

By noticing that \( T^o_{yu}/T^o_{zu} = T^o_{z\eta} \), we can then write

\[ T^o = T^o_{AF} \cdot T^o_{z\eta} F_y \]

which, considering (4.13) and Definition 4.4.1 is equivalent to

\[ T^o = T^o_{AF} \cdot M \]

**Corollary 4.6.1.** The filtering, auxiliary-variable feedback and output feedback sensitivity functions satisfy:

\[ S^o = S^o_{AF} + P - S^o_{AF} P \]  

(4.22)

**Proof.** The result (4.22) follows from the complementary relationship (4.20).  

4.7 State Space Formulation

Note that a state space representation has not been used in the above development. For example, \( z \) need not be a linear combination of states but could be a filtered version of the states. Also, as
remarked earlier, the observer can have a different degree to that of the model. In this section we specialise the result to the case where a state space description is used and where \( z = Kx \).

Consider a state-space model of the following form:

\[
\begin{align*}
\rho x &= Ax + Bu + Ew \\
y^o &= Cx
\end{align*}
\]  

where \( x \in \mathbb{R}^n \) is the state, \( \rho \) is a suitable operator, e.g. Laplace or Z-transform, and where \( A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times 1}, E \in \mathbb{R}^{n \times 1}, C \in \mathbb{R}^{1 \times n} \). In addition, we consider

\[
\begin{align*}
\eta &= x \\
z &= K \cdot \eta
\end{align*}
\]

which implies

\[
\begin{align*}
T_{z\eta}^o &= K, \quad K \in \mathbb{R}^{1 \times n} \\
T_{y\eta}^o &= C, \quad C \in \mathbb{R}^{1 \times n}
\end{align*}
\]

For the special case of the state space model \((4.23), (4.25), (4.5)\) we have:

\[
\begin{align*}
T_{yu}^o &= C(\rho I - A)^{-1}B \\
T_{yw}^o &= C(\rho I - A)^{-1}E \\
T_{zu}^o &= K(\rho I - A)^{-1}B \\
T_{zw}^o &= K(\rho I - A)^{-1}E
\end{align*}
\]

As remarked earlier, the only restriction on the observer is that it be unbiased. However, a special observer, often used in connection with the state space model \((4.18), (4.23)\), is a Luenberger observer of the form

\[
\begin{align*}
\rho \hat{x} &= A\hat{x} + Bu + L(y - C\hat{x}) \\
\hat{\eta} &= \hat{x}
\end{align*}
\]

If \((4.32), (4.33)\) are adopted, then

\[
\begin{align*}
F_y &= (\rho I - A + LC)^{-1}L \\
F_u &= (\rho I - A + LC)^{-1}B
\end{align*}
\]

**Corollary 4.7.1.** *Under the special conditions imposed by \((4.28), (4.31), (4.34)\) and \((4.35)\), then:*
(i) The key sensitivity functions satisfy:

\[
T^o = K(\rho I - A + LC)^{-1}LC(\rho I - A)^{-1}B \\
\times \left[1 + K(\rho I - A + LC)^{-1}LC(\rho I - A)^{-1}B + K(\rho I - A + LC)^{-1}B \right]^{-1}
\]  

\[1 - T^o_{AF} = \frac{\det(\rho I - A)}{\det(\rho I - A + BK)} \]  

\[P = 1 - M = \frac{\det(\rho I - A)}{\det(\rho I - A + LC)} \]

(ii) The result presented in (4.21) is equivalent to the following identity

\[
K(\rho I - A + LC)^{-1}LC(\rho I - A)^{-1}B \\
\times \left[1 + K(\rho I - A + LC)^{-1}LC(\rho I - A)^{-1}B + K(\rho I - A + LC)^{-1}B \right]^{-1} \\
= \left[1 - \frac{\det(\rho I - A)}{\det(\rho I - A + BK)} \right] \times \left[1 - \frac{\det(\rho I - A)}{\det(\rho I - A + LC)} \right]
\]

**Proof.**  
(i) Equation (4.36) is obtained by direct substitution from (4.19). For equation (4.37) we have

\[
S^o_{AF} = 1 - T^o_{AF} = \frac{\det(\rho I - A)}{\det(\rho I - A + BK)}
\]

Whereas for equation (4.38) we have

\[
P = - \frac{T^o_{yy} \cdot (F^o_{yy}T^o_{yw} - T^o_{\eta w})}{T^o_{\eta y} T^o_{\eta w}} \\
= - \left\{ C(\rho I - A + LC)^{-1}LC(\rho I - A)^{-1}E \\
- C(\rho I - A)^{-1}E \right\} \times (C(\rho I - A)^{-1}E)^{-1} \\
= 1 - C(\rho I - A + LC)^{-1}L \\
= \frac{\det((\rho I - A + LC)^{-1}LC)}{\det(\rho I - A + LC)} \\
= \frac{\det(\rho I - A + LC) - LC}{\det(\rho I - A + LC)} \\
= \frac{\det(\rho I - A)}{\det(\rho I - A + LC)}
\]
4.8 Comments and Implications of the Result

(ii) By direct substitution. The result can also be directly verified via simple algebra.

\[\text{Remark 4.7.1.} \quad \text{Notice that the left hand side of (4.39) depends upon the feedback gain } K \text{ and the observer gain } L \text{ in a convoluted fashion, whereas on the right hand side the dependance on } K \text{ and } L \text{ has been factored into two independent terms, which depend only on } K \text{ and } L \text{ respectively.} \]

\[\square\]

\[\text{Remark 4.7.2.} \quad \text{Note that the result presented in Theorem 4.6.1 is more general than the special case treated in Corollary 4.7.1.} \]

\[\square\]

4.8 Comments and Implications of the Result

It follows from (4.21) that \( T^0 \) is the product of the observer complementary sensitivity function and the complementary sensitivity of auxiliary variable feedback. Hence, a poorly designed observer, e.g. one resulting in a sensitivity peak in \( M \), will need to be compensated by changing \( T^0_{AF} \) to retain robust stability. This simple fact is often overlooked. For example, in contemporary literature on robust model predictive control, see e.g. [82], the observer is often chosen arbitrarily. However, the above argument suggests that the choice of observer is certainly not arbitrary. Indeed, if the observer results in a poor function \( M \), then it is clear that the state feedback designed by MPC will need to be changed if robust stability is to be retained. Conversely, the observer can be designed to alleviate robustness issues without necessarily impacting the state feedback design. This is illustrated by the following simple example.

4.8.1 Nominal Design

Consider the following nominal system

\[ G_0 = \frac{1}{s+1} \quad (4.40) \]

A Luenberger observer and state feedback gain are designed to stabilise the nominal system, using \( K = L = 9 \). This, in turn, leads to

\[ T^0_{AF} = \frac{K}{s+1+K} = \frac{9}{s+10} \quad (4.41) \]

\[ M = \frac{L}{s+1+L} = \frac{9}{s+10} \quad (4.42) \]

The nominal output response resulting from an initial condition \( x(0) = 1 \), is presented in Fig. 4.4.
4.8.2 Unmodelled Sensor Dynamics

Next, consider the situation where there are unmodelled dynamics, e.g. a non-ideal sensor whose dynamics are described by $G_1$. In this case, the measured plant output is related to the input via the transfer function $G = G_0G_1$. To illustrate, we consider

$$
G_1 = \frac{s^2 + 2\xi_n\omega_0 s + \omega_0^2}{s^2 + 2\xi_n\omega_0 s + \omega_0^2} e^{-ds}
$$

(4.43)

To be specific, we assume $0.02 \leq \xi_d \leq 0.2$, $\xi_n = 0.2$, $\omega_0 = 10$, $d = 0.05$, where $\omega_0$ denotes a resonant frequency and $d$ a small delay. The frequency response of $G_1$ is shown in Fig. 4.5 for different values of $\xi_d$. It can be seen that the peak of the frequency response of $G_1$ lies within the nominal closed loop bandwidth and is potentially destabilising. The same observer and state estimate feedback are used as previously. The true output (not the measured one), resulting from an initial condition $x(0) = 1$ in $G_0$ with different sensor dynamics, are presented in Fig. 4.6. Clearly, the closed loop system is unstable for some sensor realisations.

Based on Theorem 4.6.1, especially (4.21), it is clear that we can remedy the problem by either redesigning the observer or state feedback.

4.8.3 Modified Observer

We first explore redesigning the observer by considering an alternative unbiased observer. We choose:

$$
F_y = \frac{s^2 + 2\xi_d\omega_0 s + \omega_0^2}{s^2 + 2\xi_n\omega_0 s + \omega_0^2}
$$

(4.44)
4.8 Comments and Implications of the Result

and

\[ F_u = \frac{1}{s + 1} \cdot \frac{2(\xi_n - \xi_d)\omega_0 s}{s^2 + 2\xi_n\omega_0 s + \omega_0^2} \]  

(4.45)

where \( \xi_d \) is a parameter to be chosen. From (4.44) and (4.45), the observer is seen to be

\[ \hat{\eta} = \frac{\rho^2 + 2\xi_d\omega_0\rho + \omega_0^2}{\rho^2 + 2\xi_n\omega_0\rho + \omega_0^2} \cdot y + \frac{1}{\rho + 1} \cdot \frac{2(\xi_n - \xi_d)\omega_0 \rho}{\rho^2 + 2\xi_n\omega_0 \rho + \omega_0^2} \cdot u \]

where \( \rho \) is the differential operator.

It is easily checked that this is an unbiased observer for \( \eta = x \). It is also readily seen that

\[ M = \frac{s^2 + 2\xi_n\omega_0 s + \omega_0^2}{s^2 + 2\xi_n\omega_0 s + \omega_0^2} \]

(4.46)

Finally, based on the result presented in Theorem 4.6.1 we choose \( \xi_d = 0.02 \), which is the worst case for the sensor dynamics, i.e. the largest peak in the frequency response.

Utilising this new observer, Fig. 4.7 presents the true output resulting from an initial condition \( x(0) = 1 \) in \( G_0 \), with different sensor dynamics. It can be seen from Fig. 4.7 that robust stability has been achieved by modifying only the observer and maintaining the previously designed state feedback gain as \( K = 9 \).
4.8.4 Luenberger observer with redesigned state feedback

Here we compare the above result with a common strategy to deal with robustness issues, i.e. by reducing the closed loop bandwidth by adjusting $K$. Figure 4.8 presents the resulting true output when the original Luenberger observer is used with $L = 9$, and when $K = 2.4$. The latter value was chosen as the gain which provides the least $L_2$ cost under the worst model error, i.e. $\xi_d = 0.02$.

Table 4.1 compares the mean squared errors achieved for the modified observer and redesigned state feedback over a period of $T = 10[s]$. From Figures 4.7 and 4.8 and Table 4.1 it is clear that decreasing the state feedback gain to achieve robust stability compromises performance. Indeed, redesigning the observer beyond the Luenberger structure gives a performance gain of at least 2:1 in all cases.

<table>
<thead>
<tr>
<th>$\xi_d$</th>
<th>0.02</th>
<th>0.1</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified Observer</td>
<td>0.0083</td>
<td>0.0089</td>
<td>0.0097</td>
</tr>
<tr>
<td>Redesigned SF Gain</td>
<td>0.0234</td>
<td>0.0217</td>
<td>0.0223</td>
</tr>
</tbody>
</table>

Table 4.1: Mean Squared Errors
4.9 Conclusions

In this chapter we have presented a relationship between output feedback sensitivity functions, filtering sensitivity functions and auxiliary variable feedback sensitivity functions. The main result is that the output feedback complementary sensitivity function can be decomposed into the product of the other two complementary sensitivity functions, i.e. the filtering and the auxiliary variable feedback complementary sensitivity functions. The result has potential implications in several areas. For example, we have shown that robust stability can be achieved, without severely compromising performance, by redesigning the observer beyond the Luenberger structure to the more general class of unbiased observers.

This chapter is based on work previously described by the author in [23].
Figure 4.8: True output with $K = 2.4$ and Luenberger observer.
Part II

Nonlinear Systems
Background

In the first part of the thesis, the focus was on how to improve performance in the presence of uncertainty. In the second part of the thesis we turn our focus to how to modify the model so as to reduce uncertainty due to model errors. In particular, we do this in the context of using approximate sampled-data models for continuous-time nonlinear systems.

In this chapter we present preliminary results needed for the correct understanding and development of the following chapters. In particular, we review a particular sampled-data model for deterministic nonlinear systems that is central to the results obtained hereafter. In addition, we review the concept of sampling zeros and sampling zero dynamics, which, as we will see, are a recurrent topic in this part of the thesis.

5.1 Overview

The topic of approximate sampled-data models for nonlinear dynamic systems has a long history in the numerical analysis literature [11]. Results are available on local and global truncation errors for various classes of sampled data models.

An issue of importance is that the corresponding sampled data model has extra zero dynamics with associated “sampling zeros”, which arise from the folding induced by the sampling process.

The presence of these sampling zeros is a well studied problem for linear systems. The work on this topic began with the seminal work of Åström, Stenberg and Hagander [5]. Other contributions can be found in [119, 59, 120]. A key conclusion from this body of work is that discrete models are generically of relative degree one [5, 91]. In the linear case, the continuous system poles $p_i, i = 1, \ldots, n$ map to $e^{p_i \Delta}$. The situation with respect to zeros is more complex. It is known [5, 120] that the discrete zeros fall into two classes: there are $m$ zeros which converge to 1 as $e^{z_i \Delta}$, where $z_i$ are
the continuous time intrinsic zeros. In addition, there are \( n - m - 1 \) extra zeros, sometimes called the “sampling zeros”. The latter zeros have no continuous counterpart. However, their location can be asymptotically characterised as a function of the relative degree \( r = n - m \). Moreover, these extra zeros play a key role in the accuracy of sampled data models. For example, it has been shown in [55] that it is crucial that these extra zeros be included in the approximate model if one wants to have a model for which the relative errors converge to zero as \( \Delta \to 0 \), where \( \Delta \) is the sampling period. Moreover, we note that relative errors are of direct relevance to inverse problems such as robust control [49].

The nonlinear case is more difficult. In recent work [91, 124], it has been shown that sampling zero dynamics also arise in the sampling of nonlinear continuous systems. Remarkably, the asymptotic sampling zero dynamics are the same as for the linear case and depend only on the nonlinear relative degree [124].

For example, it has been shown in [124] that, by using a particular form of truncated Taylor series, an (approximate) sampled-data model for continuous time nonlinear systems can be obtained. The resultant sampled data model has three key properties, namely:

(i) the sampled data model has extra zero dynamics which are identical to the sampling zero dynamics for a linear system of the same relative degree,

(ii) the model has local (one-step) truncation error of order \( \Delta^{r+1} \) in the output, where \( r \) is the relative degree of the system,

(iii) the model depends on the underlying continuous model in a simple fashion.

One of the goals here is to extend property (ii) to multi-step truncation errors. This will require that we go significantly beyond the work in [124].

5.2 Nonlinear Systems

We consider a nonlinear system described as follows [64]:

\[
\dot{x}(t) = f(x(t)) + g(x(t))u(t) \tag{5.1}
\]

\[
y(t) = h(x(t)) \tag{5.2}
\]

where \( x(t) \) is the state evolving in an open subset \( \mathcal{M} \subset \mathbb{R}^n \), and where the vector fields \( f(\cdot), g(\cdot) \), and the output function \( h(\cdot) \) are sufficiently smooth. Also, motivated by control applications we
Definition 5.2.1 (Relative Degree). The nonlinear system \((5.1)-(5.2)\) is said to have relative degree \(r\) at a point \(x_0\) if

(i) \(L_g L_k^j f h(x) = 0\) for all \(x\) in a neighbourhood of \(x_0\) and all \(k < r - 1\).

(ii) \(L_g L^{r-1} f h(x_0) \neq 0\).

where \(L f h(x) = \partial h/\partial x \cdot f(x)\) is the Lie Derivative of \(h\) with respect to \(f\).

Definition 5.2.2 (Local Coordinate Transformation). Suppose the system \((5.1)-(5.2)\) has relative degree \(r\) at \(x_0\) \((r \leq n)\). Then a local coordinate transformation \(\Phi(x) = [\phi^T_1(\cdot), \ldots, \phi^T_n(\cdot)]^T\) in a neighbourhood of \(x_0\) is given by

\[
\begin{align*}
z_1 &= \phi_1(x) = h(x) \\
z_2 &= \phi_2(x) = L_f h(x) \\
&\vdots \\
z_r &= \phi_r(x) = L^{r-1}_f h(x)
\end{align*}
\]

It is well known (see e.g. [64]), that if \(r < n\), then it is always possible to find \(n - r\) functions \(z_{r+1} = \phi_{r+1}(x), \ldots, z_n = \phi_n(x)\) such that \(z = [z_1, \ldots, z_n]^T = [\phi_1(x), \ldots, \phi_n(x)]^T = \Phi(x)\) has a nonsingular Jacobian at \(x_0\) and, in addition, \(L_g \phi_i(x) = 0\) in a neighbourhood of \(x_0\) for all \(i = r + 1, \ldots, n\). This is the basis for the following definition.

Definition 5.2.3 (Normal Form). The normal form of the nonlinear system \((5.1)-(5.2)\) is given by the state space description in the new coordinates \(\Phi(x)\) as described in Definition 5.2.2. The model can be written as

\[
\begin{align*}
\dot{z}_1 &= z_2 \\
&\vdots \\
\dot{z}_{r-1} &= z_r \\
\dot{z}_r &= b(\zeta, \eta) + a(\zeta, \eta) \cdot u(t) \\
\dot{\eta} &= c(\zeta, \eta)
\end{align*}
\]
where \( y = z_1, \), \( \zeta = \begin{bmatrix} z_1 & \ldots & z_r \end{bmatrix}^T, \), \( \eta = \begin{bmatrix} z_{r+1} & \ldots & z_n \end{bmatrix}^T, \) and

\[
\begin{align*}
    a(\zeta, \eta) &= L_g L_f^{r-1} h(\Phi^{-1}(z)) \\
    b(\zeta, \eta) &= L_f^r h(\Phi^{-1}(z)) \\
    c(\zeta, \eta) &= \begin{bmatrix} L_f \phi_{r+1}(\Phi^{-1}(z)) \\ \vdots \\ L_f \phi_n(\Phi^{-1}(z)) \end{bmatrix}
\end{align*}
\]

with \( z = \begin{bmatrix} \zeta^T & \eta^T \end{bmatrix}^T. \)

It is easily seen that, since \( y = z_1, \) the vector \( \zeta \) contains the output and its first \( r - 1 \) derivatives.

### 5.3 A sampled-data model for nonlinear systems

The following two definitions correspond to a Taylor series expansion of the normal form description (5.6)–(5.9), giving rise to an exact and an approximate sampled data model respectively. We assume a zero-order hold (ZOH) input, i.e. \( u(t) = u_k, k\Delta \leq t < k\Delta + \Delta. \)

**Remark 5.3.1.** If an anti-aliasing filter is used, then we assume that this has been incorporated into the original model. Note that this will typically increase the relative degree.

**Definition 5.3.1 (Exact Sampled Data Model).** Define \( z_i = z_i(k\Delta) = z_i[k]\) and \( z_i^+ = z_i(k\Delta + \Delta) = z_i[k+1]\) for \( i = 1, \ldots, n. \) Then, the evolution of the state variables is exactly described by

\[
\begin{align*}
    z_1^+ &= z_1 + \Delta z_2 + \frac{\Delta^2}{2} z_3 + \ldots + \frac{\Delta^{r-1}}{(r-1)!} z_r + \Delta^r F_1 \\
    z_2^+ &= z_2 + \Delta z_3 + \ldots + \frac{\Delta^{r-2}}{(r-2)!} z_r + \Delta^{r-1} F_2 \\
    & \vdots \\
    z_{r-1}^+ &= z_{r-1} + \Delta z_r + \frac{\Delta^2}{2} F_{r-1} \\
    z_r^+ &= z_r + \Delta F_r \\
    \eta^+ &= \eta + \Delta c(\zeta, \eta) \big|_{t = \xi_{r+1}}
\end{align*}
\]

where \( F_\ell = \left\{ b(\zeta, \eta) + a(\zeta, \eta) \cdot u_k \right\} \big|_{t = \xi_\ell} \) for some appropriately chosen time instants \( k\Delta < \xi_\ell < k\Delta + \Delta, \) \( \ell = 1, \ldots, r + 1. \)

Key features of the above model are: (i) the use of different order Taylor’s series for each state
and (ii) the evaluation of the final term at an intermediate point on the integration interval. The latter condition is relaxed in the following.

Definition 5.3.2 (Truncated Taylor Series Sampled Data Model). An approximate discrete-time model with states $\hat{z} = \begin{bmatrix} \hat{\zeta}^T & \hat{\eta}^T \end{bmatrix}^T$ is obtained by replacing the unknown time instants $\xi_\ell$ in Definition 5.3.1 by $k\Delta$, leading to the following approximate sampled data model:

\begin{align*}
\hat{z}_1^+ &= \hat{z}_1 + \Delta \hat{z}_2 + \frac{\Delta^2}{2} \hat{z}_3 + \ldots + \frac{\Delta^{r-1}}{(r-1)!} \hat{z}_r + \frac{\Delta^r}{r!} \hat{F} \\
\hat{z}_2^+ &= \hat{z}_2 + \Delta \hat{z}_3 + \ldots + \frac{\Delta^{r-2}}{(r-2)!} \hat{z}_r + \frac{\Delta^{r-1}}{(r-1)!} \hat{F} \\
&\vdots \\
\hat{z}_{r-1}^+ &= \hat{z}_{r-1} + \Delta \hat{z}_r + \frac{\Delta^2}{2} \hat{F} \\
\hat{z}_r^+ &= \hat{z}_r + \Delta \hat{F} \\
\hat{\eta}^+ &= \hat{\eta} + \Delta \xi(\hat{\zeta}, \hat{\eta}) \bigg|_{t=k\Delta}
\end{align*}

where $\hat{F} = \left\{ b(\hat{\zeta}, \hat{\eta}) + a(\hat{\zeta}, \hat{\eta}) \cdot u_k \right\} \bigg|_{t=k\Delta}$ and the output is $\hat{y}_k = \hat{z}_{1,k}$.

In the sequel, we will refer to the model presented in Definition 5.3.2 as the TTS model. The TTS model was first described in [124].

5.4 Sampling Zeros

5.4.1 Euler-Frobenius Polynomials

Throughout the remainder of the thesis, the so called Euler-Frobenius polynomials will play a central role. We therefore provide here a brief definition for the sake of completeness. A more complete summary of the definitions and properties of the Euler-Frobenius Polynomials is given in Appendix A of this thesis.

Definition 5.4.1 (Euler-Frobenius Polynomials). The Euler-Frobenius Polynomials are defined by the following exponential generating function

$$
\sum_{n=0}^{\infty} B_n(q) \cdot \frac{x^n}{n!} = \frac{q - 1}{q - e^{(q-1)x}}
$$

(5.15)
The first few polynomials are given by:

\[ B_0(q) = 1 \quad (5.16a) \]
\[ B_1(q) = 1 \quad (5.16b) \]
\[ B_2(q) = q + 1 \quad (5.16c) \]
\[ B_3(q) = q^2 + 4q + 1 \quad (5.16d) \]
\[ B_4(q) = q^3 + 11q^2 + 11q + 1 \quad (5.16e) \]

See [29, Section 1.13] for generating functions and Appendix A for multiple definitions of the polynomials and related properties.

The following definition will be used in Chapter 7.

**Definition 5.4.2 (Euler-Fröbenius polynomials).** The polynomials \( p_n(\Delta \delta) \) are defined by the following recursion:

\[
p_0(q - 1) \triangleq 1 \\
\]
\[
p_n(q - 1) = \sum_{\ell=1}^{n} \frac{(q - 1)^{\ell-1}}{\ell!} \cdot p_{n-\ell}(q - 1), \quad n \geq 1
\]

where \( \delta \) is the known delta operator [89, 37]. These polynomials express the usual Euler-Fröbenius polynomials [120] in terms of the delta operator (see [124, Lemma 2] for further details).

**Remark 5.4.1.** Note that \( p_n(\Delta \delta) = p_n(q - 1) \). In Chapter 7 we will use the latter notation for clarity of exposition, since it allows to identify explicitly the truncation error as a power of the sampling period.

### 5.4.2 Deterministic Linear Systems

Consider a single-input single-output continuous-time linear system

\[
G(s) = \frac{K(s - z_1) \cdots (s - z_m)}{(s - p_1) \cdots (s - p_n)} \quad (5.17)
\]

with relative degree \( r = n - m \).

**Theorem 5.4.1.** The discrete time model that corresponds to sampling system (5.17) with sampling period \( \Delta \) and ZOH input, is given by

\[
G(z) = \frac{K(z - \bar{z}_1) \cdots (z - \bar{z}_m)P_{n-m}(z)}{(z - \bar{p}_1) \cdots (z - \bar{p}_n)} \quad (5.18)
\]
where $G(z)$ generically has relative degree one and, as $\Delta \to 0$, then

\[ \bar{p}_i \to 1 \text{ as } e^{p_i \Delta}, \quad i = 1, \ldots, n \]  

\[ \bar{z}_i \to 1 \text{ as } e^{z_i \Delta}, \quad i = 1, \ldots, m \]  

\[ P_{n-m}(z) \to B_{n-m}(z) \]  

where $B_k(z)$ is the Euler-Frobenius polynomial of order $k$.

Proof. See [5, 120] and Appendix A.

5.4.3 Deterministic Nonlinear Systems

We can then have the following result which characterises the asymptotic sampling zero dynamics in the nonlinear case.

**Theorem 5.4.2.** The sampled-data model (5.14) generically has relative degree 1, with respect to the output $y = z_1$. Furthermore, the discrete-time zero dynamics of this model are given by two subsystems:

(i) The sampled counterpart of the continuous-time zero dynamics

\[ \dot{\eta}^+ = \dot{\eta} + \Delta \cdot c(0, \dot{z}_2, \ldots, \dot{z}_r, \dot{\eta}) \]  

(ii) A linear subsystem of dimension $r - 1$

\[ \tilde{z}_{2,r} = \tilde{z}_{2,r} + \Delta Q_{22} \tilde{z}_{2,r} \]  

where the eigenvalues of the matrix $Q_{22}$ are the same sampling zeros that appear in the asymptotic linear case, namely, the roots of the Euler-Frobenius polynomials and where $z_{1,r} = [\tilde{z}_1, \ldots, \tilde{z}_r]^T$ and $\tilde{z}_{1,r} = Tz_{1,r}$, where $T$ is a linear transformation defined in [124], Eq. 66 and Appendix IV.

Proof. See Theorem 2 in [124].

5.4.4 Stochastic Linear Systems

Consider a continuous-time stationary process $v(t)$, with covariance function

\[ R(\tau) = E\{v(t)v(t + \tau)\} \]  

(5.24)
The associated power spectrum is given by
\[ \Phi(s) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R(\tau) \cdot e^{-s\tau} d\tau \] (5.25)

Consider the sequence \( v_k = v(k\Delta) \), obtained by the instantaneous sampling of \( v(t) \) with sampling period \( \Delta \), then the covariance function of \( v_k \) is given by
\[ R_\Delta[k] = R(k\Delta) \] (5.26)

The discrete time power spectrum is then defined as
\[ \Phi_\Delta(z) = \frac{\Delta}{2\pi} \sum_{k=-\infty}^{\infty} R_\Delta[k] z^{-k} \] (5.27)

Consider that \( v(t) = H(p)e(t) \), where \( e(t) \) is continuous-time white noise and \( p \) is the differential operator. Then we have the following result:

**Theorem 5.4.3.** Consider the instantaneous sampling of the process \( v(t) \). Then we have
\[ \Phi_\Delta(e^{s\Delta}) \to \Phi(s) \] (5.28)

as \( \Delta \to 0 \), uniformly in \( s \), on compact subsets. Furthermore, let \( \pm z_i, i = 1, \ldots, m \), be the 2m zeros of \( \Phi(s) \), and \( \pm p_i, i = 1, \ldots, n \), be the 2n poles of \( \Phi(s) \). Then

- 2m zeros of \( \Phi_\Delta(z) \) will converge to 1 as \( e^{\pm z_i\Delta} \) and
- the remaining \( 2(n - m) - 1 \) zeros of \( \Phi_\Delta(z) \) will converge to the zeros of \( z \cdot B_{2(n-m)-1}(z) \) as \( \Delta \to 0 \).
- the 2n poles of \( \Phi_\Delta(z) \) equal \( e^{\pm p_i\Delta} \), and will hence go to 1 as \( \Delta \to 0 \).

**Proof.** See Theorem 3.1 in [119].
Vector Measures of Accuracy for Sampled Data Models of Nonlinear Systems

The problem of obtaining an exact sampled-data model for a continuous time nonlinear system is usually an intractable one [99]. Hence some form of numerical integration is typically used to obtain approximate solutions [11, 70], e.g. Euler or Runge-Kutta methods. Under these conditions, the accuracy of the approximate model is of importance. However, this raises the question, “Accuracy in what sense?”.

The accuracy measures in current use for sampled-data models treat all states on an equal basis [11]. On the other hand, there are applications where some (combination of) states are more important than others. For example, in control theory, the output has a special status. In this regard, the concept of relative degree is crucial since it defines the smoothness of the input-output relationship for dynamical systems. In nonlinear systems, this concept leads to state space models expressed in what is known as the normal form – see discussion in Chapter 5.

In this chapter we argue that the definitions for both local and global truncation error used in numerical analysis (see for example [11]) can be improved to better account for control applications. Specifically, we propose three novel definitions of truncation errors. We argue that these new measures of truncation error are more appropriate when dealing with specific problems in control theory and system identification. In this sense, the results presented in this chapter can be thought of as intersecting numerical analysis with nonlinear systems theory.

In addition, we apply these new truncation error measures to the TTS sampled-data model (Defi-
nition 5.3.2) and show that this leads to new insights regarding the model.

The remainder of this chapter is organised as follows: In Section 6.1, new definitions for local and global truncation errors are given. In Section 6.2, the new definitions are applied to the TTS sampled-data model introduced in Section 5.3. A numerical example is presented in Section 6.3. Finally, conclusions are drawn in Section 6.4.

6.1 Vector Truncation Error Definitions

The usual definitions of local and global truncation error used in numerical analysis [11] do not distinguish between elements of the state vector. The new definitions given below (see Definitions 6.1.3, 6.1.4, and 6.1.5), which are motivated by control applications, assign a different error bound to each element of the state vector. We first introduce the following:

Definition 6.1.1. We define \( T = N \cdot \Delta \) as the time horizon over which the approximate model will be used, where \( \Delta \) is the length of the time discretisation and \( N \) is the number of steps.

Definition 6.1.2 (Big-O). For asymptotic analysis, as \( \Delta \to 0 \), we define the notation (see, e.g. [74])

\[
\text{if } f(\Delta) \in O(g(\Delta)) \quad (6.1)
\]

for two functions \( f(\cdot) \) and \( g(\cdot) \) if and only if

\[
|f(\Delta)| \leq C \cdot |g(\Delta)| \quad (6.2)
\]

for all \( \Delta < \Delta_0 \) and where \( C \) is a constant.

When (6.1) holds, we say that \( f(\Delta) \) is of the order of \( g(\Delta) \). In the sequel we will use \( f[k] \) to denote \( f(k\Delta) \). We then introduce the following definitions of truncation error:

Definition 6.1.3 (Local Vector Truncation Error). Consider a dynamical system with states \( (x_1, \ldots, x_n) \) and an associated approximate model with states \( (\hat{x}_1, \ldots, \hat{x}_n) \). Then, the local vector truncation error of the approximate model is said to be of the order of \( (\Delta^{m_1}, \ldots, \Delta^{m_n}) \) if and only if

\[
\begin{align*}
\hat{x}_1[k] - x_1[k] &= 0, & \hat{x}_1[k + 1] - x_1[k + 1] &\in O(\Delta^{m_1}) \\
&\vdots & &\vdots \\
\hat{x}_n[k] - x_n[k] &= 0, & \hat{x}_n[k + 1] - x_n[k + 1] &\in O(\Delta^{m_n})
\end{align*}
\]

where \( m_1, \ldots, m_n \) are integers.
Remark 6.1.1. Definition 6.1.3 embeds the usual definition for local truncation error given in [1], when \( m_1, \ldots, m_n \) are all chosen as the same integer \( m \). □

Definition 6.1.4 (Local Vector Fixed Step Truncation Error). The local vector fixed step truncation error of an approximate model is said to be of the order of \((\Delta^{m_1}, \ldots, \Delta^{m_n})\) if and only if, for initial state errors

\[
\hat{x}_1[k] - x_1[k] \in O(\Delta^{\bar{m}_1}) \\
\vdots \\
\hat{x}_n[k] - x_n[k] \in O(\Delta^{\bar{m}_n})
\]

for any \( \bar{m}_i \geq m_i, i = 1, \ldots, n \) then after \( N \) steps, where \( N \) is a finite fixed number, we have that

\[
\hat{x}_1[k+N] - x_1[k+N] \in O(\Delta^{m_1}) \\
\vdots \\
\hat{x}_n[k+N] - x_n[k+N] \in O(\Delta^{m_n})
\] □

Remark 6.1.2. The motivation behind Definition 6.1.4 is that the order of the error bounds (i.e. the exponent \( m_i \)) may decrease when one takes finitely many steps. This follows the idea in numerical analysis where, given the usual definition of local truncation error, the global truncation error can be of one order less in terms of powers of the discretisation step. □

Remark 6.1.3. Definition 6.1.4 has practical relevance to applications where an algorithm requires a fixed number of iterations of the model, e.g. model predictive control or system identification (see, e.g., [46] or [123] [27], respectively). For example, in model predictive control, one typically needs to iterate the model over a finite prediction horizon. □

Definition 6.1.5 (Global Vector Truncation Error). The global vector (fixed time) truncation error of an approximate model is said to be of the order of \((\Delta^\bar{m}_1, \ldots, \Delta^\bar{m}_n)\) if and only if, for initial state errors

\[
\hat{x}_1[k] - x_1[k] \in O(\Delta^{\bar{m}_1}) \\
\vdots \\
\hat{x}_n[k] - x_n[k] \in O(\Delta^{\bar{m}_n})
\]
for any $\bar{m}_i \geq \tilde{m}_i$, $i = 1, \ldots, n$ then after a fixed (continuous) time $T$, i.e. after $N = \lfloor T/\Delta \rfloor$ steps, we have that

$$\hat{x}_1[k+N] - x_1[k+N] \in O(\Delta^{\bar{m}_1})$$

$$\vdots$$

$$\hat{x}_n[k+N] - x_n[k+N] \in O(\Delta^{\bar{m}_n})$$

Remark 6.1.4. The key difference between Definition 6.1.4 and Definition 6.1.5 is that, in the latter, the number of steps is a function of the sampling time $\Delta$. As $\Delta$ decreases, the number of iteration steps $N$ increases, since $T$ remains fixed. This is in contrast with Definition 6.1.4, where, independent of the sampling time, the error bound is always measured after a fixed number of iterations.

6.2 Error Analysis of Truncated Taylor Series Sampled Data Model

In this section we apply the definitions introduced in Section 6.1 to quantify the local and global truncation errors of the approximate model described in Definition 5.3.2.

Theorem 6.2.1 (Local Vector Truncation Error). Consider a continuous-time nonlinear system (5.1)–(5.2) having uniform relative degree $r \leq n$ in an open subset $\mathcal{M} \subset \mathbb{R}^n$ and having a ZOH input. Then, the approximate sampled data model given in Definition 5.3.2 with states $(\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{r-1}, \hat{z}_r, \hat{\eta})$, has a Local Vector Truncation Error of the order of $(\Delta^{r+1}, \Delta^r, \ldots, \Delta^3, \Delta^2, \Delta^2)$.

Proof. From the definition of local truncation error we assume that

$$\hat{z}_1[k] - z_1[k] = 0$$

$$\hat{z}_2[k] - z_2[k] = 0$$

$$\vdots$$

$$\hat{z}_{r-1}[k] - z_{r-1}[k] = 0$$

$$\hat{z}_r[k] - z_r[k] = 0$$

$$\hat{\eta}[k] - \eta[k] = 0$$
and, because of this, the approximate model states satisfy \((\hat{\zeta}, \hat{\eta}) = (\zeta, \eta)\). Therefore
\[
\hat{F} = F = \{b(\zeta, \eta) + a(\zeta, \eta) \cdot u_k\}_{t=k\Delta}
\]

Hence, the 1-step ahead differences are
\[
\hat{z}_1^+ - z_1^+ = \frac{\Delta^r}{r!} \cdot (F - F_1)
\]
\[
\hat{z}_2^+ - z_2^+ = \frac{\Delta^{r-1}}{(r-1)!} \cdot (F - F_2)
\]
\[
\vdots
\]
\[
\hat{z}_{r-1}^+ - z_{r-1}^+ = \frac{\Delta^2}{2} \cdot (F - F_{r-1})
\]
\[
\hat{z}_r^+ - z_r^+ = \Delta \cdot (F - F_r)
\]
\[
\hat{\eta}^+ - \eta^+ = \Delta \cdot \left( c(\zeta, \eta) \big|_{t=k\Delta} - c(\zeta, \eta) \big|_{t=\xi_{r+1}} \right)
\]

We define the errors \(e_i[p] = \hat{z}_i[p] - z_i[p], i = 1, \ldots, r\) and \(e_\eta[p] = \hat{\eta}[p] - \eta[p]\). In order to show that the approximate sampled data model given in Definition 5.3.2 has a Local Vector Truncation Error of the order of \((\Delta^{r+1}, \Delta^r, \ldots, \Delta^3, \Delta^2, \Delta^2)\), it suffices to show that
\[
|F - F_i| \in \mathcal{O}(\Delta) \quad i = 1, \ldots, r \tag{6.12}
\]
\[
|c(\zeta, \eta) \big|_{t=k\Delta} - c(\zeta, \eta) \big|_{t=\xi_{r+1}}| \in \mathcal{O}(\Delta) \tag{6.13}
\]

From Definition 5.3.2 we have that
\[
|F - F_i| = \left| \left\{b(\zeta, \eta) + a(\zeta, \eta) \cdot u_k\right\} \big|_{t=k\Delta} - \left\{b(\zeta, \eta) + a(\zeta, \eta) \cdot u_k\right\} \big|_{t=\xi_i} \right|
\]
\[
\leq L \cdot \|\zeta(\xi_i)\|_{t=k\Delta} - \|\zeta(\xi_i)\|_{t=\xi_i}
\]
\[
= L \cdot \|z(k\Delta) - z(\xi_i)\|
\]

where the existence of the Lipschitz constant \(L\) is guaranteed by the analyticity of \(f(\cdot), g(\cdot)\) and \(h(\cdot)\). In addition, analyticity guarantees the state trajectory \(z(t) = \left[\zeta^T(t) \quad \eta^T(t)\right]^T\) is bounded [11] Th. 112E in the following form
\[
\|z(k\Delta) - z(\xi_i)\| \leq C \cdot \frac{e^{L|k\Delta - \xi_i|} - 1}{L}.
\]

Therefore
\[
|F - F_i| \leq L \cdot C \cdot \frac{e^{L|k\Delta - \xi_i|} - 1}{L}
\]
\[
< C \cdot \left( e^{L\Delta} - 1 \right)
\]
\[
\in \mathcal{O}(\Delta)
\]
which establishes (6.12). The proof for \( e_\eta \) is analogous, on noting that

\[
\left| c(\zeta, \eta)|_{t=k\Delta} - c(\zeta, \eta)|_{t=k\Delta} \right| \leq L \cdot ||(\zeta, \eta)|_{t=k\Delta} - (\zeta, \eta)|_{t=k\Delta} ||
\]

from where (6.13) follows.

**Theorem 6.2.2** (Local Vector Fixed Step Truncation Error). Consider the same setup as in Theorem 6.2.1. Then, the approximate sampled data model given in Definition 5.3.2, with states \( (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_r, \hat{\eta}) \), has Local Vector Fixed Step Truncation Error of the order of \((\Delta^{r+1}, \Delta^r, \Delta^3, \Delta^2, \Delta)\).

**Proof.** Here we need to consider the cumulative effect of the errors when we iterate the approximate model over \( N \) steps where \( N \) is fixed.

Define \( e_i[p] = \hat{z}_i[p] - z_i[p], i = 1, \ldots, r \). Then, we can write

\[
\begin{bmatrix}
    e_1[k+1] \\
    e_2[k+1] \\
    e_3[k+1] \\
    \vdots \\
    e_{r-1}[k+1] \\
    e_r[k+1]
\end{bmatrix} =
\begin{bmatrix}
    1 & \Delta & \Delta^2 & \cdots & \Delta^{r-2} & \Delta^{r-1} \\
    0 & 1 & \Delta & \cdots & \Delta^{r-3} & \Delta^{r-2} \\
    0 & 0 & 1 & \cdots & \Delta^{r-4} & \Delta^{r-3} \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
    0 & 0 & 0 & \cdots & 1 & \Delta \\
    0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}
\begin{bmatrix}
    e_1[k] \\
    e_2[k] \\
    e_3[k] \\
    \vdots \\
    e_{r-1}[k] \\
    e_r[k]
\end{bmatrix} + \begin{bmatrix}
    \Delta^r \\
    \Delta^{r-1} \\
    \Delta^{r-2} \\
    \vdots \\
    \Delta^3 \\
    \Delta^2 \\
    \Delta
\end{bmatrix} \cdot E \quad (6.14)
\]

where, as shown in the proof of Theorem 6.2.1, \( E = \max_{t,k} |F - F_t|_{t=k\Delta} \in O(\Delta), \forall k \in \mathbb{N}, \forall i = 1, \ldots, r \). Since the system (6.14) has the form

\[
\theta[k+1] \leq A \cdot \theta[k] + \Phi \cdot E \quad (6.15)
\]

then, after \( N \) steps, we have

\[
\theta[k+N] \leq A^N \cdot \theta[k] + \sum_{i=0}^{N-1} A^i \cdot \Phi \cdot E \quad (6.16)
\]

It is easily seen that the matrices in (6.16) are as shown in (6.17) and (6.18)

\[
A^N =
\begin{bmatrix}
    1 & N\Delta & N^2\Delta^2 & \cdots & N^{r-2}\Delta^{r-2} & N^{r-1}\Delta^{r-1} \\
    0 & 1 & N\Delta & \cdots & N^{r-3}\Delta^{r-3} & N^{r-2}\Delta^{r-2} \\
    0 & 0 & 1 & \cdots & N^{r-4}\Delta^{r-4} & N^{r-3}\Delta^{r-3} \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
    0 & 0 & 0 & \cdots & 1 & N\Delta \\
    0 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix} \quad (6.17)
\]
where

\[
S_r^{N-1} = \sum_{i=0}^{N-1} i^r = \frac{N^{r+1}}{r+1} + \sum_{k=1}^{r} \frac{B_k}{r-k+1} \binom{r}{k} N^{r-k+1}
\]

and where \(B_k\) is the \(k\)-th Bernoulli number.

Finally, if \(N\) is a fixed integer, then both \(A^N \cdot \theta[k]\) and \(\sum_{i=0}^{N-1} A^i \cdot \Phi \cdot E\) preserve the order of \(\theta[k]\) element-wise.

The proof for the remaining states \(\eta\) is completely analogous to the proof for \(z_r\).

\[\text{Remark 6.2.1.} \quad \text{Note that the original analysis given in [124], uses a definition of local truncation error that refers to the output only and holds for one step of the model only. This analysis cannot be}\]

\[\text{Theorem 6.2.3 (Global Vector Truncation Error). Consider the same setup as in Theorem 6.2.1.}\]

Then, the approximate sampled data model given in Definition 5.3.2, with states \((\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{r-1}, \hat{z}_r, \hat{\eta})\), has Global Vector (Fixed Time) Truncation Error of the order of \((\Delta, \Delta, \ldots, \Delta, \Delta, \Delta)\).

\[\text{Proof.} \quad \text{Based on the proof of Theorem 6.2.2, we note that the sum } \sum_{i=0}^{N-1} i^r \text{ is a polynomial in } N \text{ of order } r + 1 \text{ (see [101]). Therefore, since } N = T/\Delta \text{ with } T \text{ fixed, every element in the last column of the matrix } \sum_{i=0}^{N-1} A^i \text{ is of order } O(\Delta^{-1}). \text{ Then, given that the last row of } \Phi \text{ is } \Delta, \text{ there will always be a term of order } O(1) \text{ multiplying } E \text{ for every state. Finally, since we already know } E \in O(\Delta), \text{ then it is straightforward to show that}\]

\[\sum_{i=0}^{N-1} A^i \cdot \Phi \cdot E \in O(\Delta)\]

On the other hand, note that all the elements of \(A^N\) are of order \(O(1)\), and since the initial errors are at least \(O(\Delta)\) in all states by assumption \((\bar{m}_i \geq m_i)\), we conclude that

\[e_i[k + N] \in O(\Delta), \quad i = 1, \ldots, r.\]

The proof for the remaining states \(\eta\) is completely analogous to the proof for \(z_r\). \]
extended to multiple steps since the latter inevitably requires information about all of the states. The definitions and analysis provided in this chapter, show that the model presented in [124] possesses appealing properties that hold for multiple iterations. Specifically we have shown how the truncation errors propagate over a finite number of steps and over a fixed time period.

Remark 6.2.2. The result shown in Theorem 6.2.3 has its roots in the fact that the lower order errors in the lower part of chain of integrators, e.g. the r-th state, start appearing (and accumulating) in the upper part of the chain, e.g. the first state. This behaviour is the one Butcher [11] calls Global Truncation Error, but because of the structure of the system, i.e. the interaction between the states and their errors, instead of dropping only one order from local to global truncation error, it drops all the way down to ∆ when a fixed time horizon is considered. Further insight as to why this happens can be found in the proof of Theorem 6.2.3 in the Appendix.

Remark 6.2.3. The results provided in Theorems 5.4.2 and 6.2.1 are closely related, i.e. the augmented order of accuracy over a finite number of iterations is precisely the property that accounts for the “high-frequency” mismatch between the continuous-time and sampled-data models, i.e. the sampling zero dynamics.

Remark 6.2.4. In the notation of Vector Truncation Errors, for a nonlinear system with states $(\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{r-1}, \hat{z}_r, \hat{\eta})$ with a Vector Truncation Error of the order of $(\Delta^{m_1}, \Delta^{m_2}, \ldots, \Delta^{m_{r-1}}, \Delta^{m_r}, \Delta^{m_j})$, the last term $\Delta^{m_j}$ has to be interpreted as a vector since it corresponds to the error of the state $\eta \in \mathbb{R}^{n-r}$.

6.3 Examples

In this section we present numerical examples which substantiate the previous results. We consider the following second order nonlinear system expressed in normal form:

\[
\begin{align*}
\dot{x}_1(t) &= x_2(t) \\
\dot{x}_2(t) &= -\alpha_1 x_2(t) - \alpha_0 x_1(t)(1 + \varepsilon_1 x_1(t)^2) + \beta_0 (1 + \varepsilon_2 x_1(t))u(t)
\end{align*}
\]

and the corresponding Truncated Taylor Series sampled-data model:

\[
\begin{align*}
x_1^+ &= x_1 + \Delta \cdot x_2 + \frac{\Delta^2}{2} \cdot F \\
x_2^+ &= x_2 + \Delta \cdot F
\end{align*}
\]

where

\[
F = -\alpha_1 x_2 - \alpha_0 x_1(1 + \varepsilon_1 x_1^2) + \beta_0 (1 + \varepsilon_2 x_1)u
\]
We choose the system parameters as $\alpha_0 = 2, \alpha_1 = 3, \beta_0 = 2, \varepsilon_1 = 0.5, \varepsilon_2 = 0.5$. According to Theorems 6.2.1, 6.2.2 and 6.2.3, the model (6.21)-(6.22) will have Local Vector Truncation Error, Local Vector Fixed Step Truncation Error and Global Vector (Fixed Time) Truncation Error of the order of $(\Delta^3, \Delta^2)$, $(\Delta^3, \Delta^2)$ and $(\Delta, \Delta)$ respectively. The exact discrete-time model is impossible to obtain. In order to accurately approximate the errors we simulated the continuous time system in Simulink with ode45 (Dormand-Prince) as a solver, which we recognise is not the true solution. We then sampled the state signals at the corresponding sampling instants.

We analyse the behaviour of the truncation errors as the sampling period $\Delta$ is varied, specifically as it tends to zero as $\Delta = T/2^n$, where $n = 3, \ldots, 10$ and $T = 1\text{[s]}$. We consider a step input and zero initial conditions.

Figure 6.1 and Table 6.1 show the truncation errors after one iteration of the approximate model,
for both states of the system, for different sampling times. The column termed ratio, in Table 6.1 denotes the slope obtained in the error plots for the i-th state, as the sampling period is reduced. It can be seen that, in the case of ratio1, every time the sampling period is decreased by a factor of 2, the error in state $x_1$ decreases by a factor of 8, from where we conclude that the error is indeed of the order of $\Delta^3$. Similarly, ratio2 shows that the error in state $x_2$ is of the order of $\Delta^2$. These results are as predicted by Theorem 6.2.1.

Figure 6.2 and Table 6.2 show the errors after a fixed number of steps. $N = 7$ was chosen for all the simulations. It can be seen that, for a sufficiently small sampling period, the accumulated errors, in both states, preserve the behaviour of the local truncation error. This result is as predicted by Theorem 6.2.2.

Figure 6.3 and Table 6.3 show the error for a fixed continuous time interval, i.e. after $N$ steps, where $N = T/\Delta$ is the last step in the given time horizon $T = 1$. It can be seen that the errors, in both states, are asymptotically of the order of $\Delta$. This result is as predicted by Theorem 6.2.3.

6.4 Conclusions

In this chapter we have provided three novel definitions of truncation error. These definitions lie at the intersection of numerical analysis and control theory. We have applied these definitions to a recently developed sampled-data model for nonlinear systems based on a normal form description. It has been shown that the model possesses previously unknown properties, regarding the propagation of truncation errors. These properties are of direct relevance to control applications.
Table 6.1: Local Vector Truncation Errors

| Δ  | $|e_1[1]|$   | $|e_2[1]|$   | ratio₁ | ratio₂ |
|----|-------------|-------------|---------|--------|
| $2^{-3}$ | 1.8005e-03 | 4.2067e-02 | -       | -      |
| $2^{-4}$ | 2.3429e-04 | 1.1094e-02 | 7.6848  | 3.7921 |
| $2^{-5}$ | 2.9892e-05 | 2.8499e-03 | 7.8380  | 3.8925 |
| $2^{-6}$ | 3.7753e-06 | 7.2235e-04 | 7.9178  | 3.9454 |
| $2^{-7}$ | 4.7436e-07 | 1.8184e-04 | 7.9586  | 3.9725 |
| $2^{-8}$ | 5.9450e-08 | 4.5618e-05 | 7.9792  | 3.9862 |
| $2^{-9}$ | 7.4409e-09 | 1.1424e-05 | 7.9896  | 3.9931 |
| $2^{-10}$ | 9.3072e-10 | 2.8585e-06 | 7.9948  | 3.9965 |

Table 6.2: Local Vector Fixed Step Truncation Error

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^{-3}$</td>
<td>3.2229e-02</td>
<td>1.9171e-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$2^{-4}$</td>
<td>8.2994e-03</td>
<td>2.4302e-02</td>
<td>3.8833</td>
<td>7.8888</td>
</tr>
<tr>
<td>$2^{-5}$</td>
<td>1.4894e-03</td>
<td>1.1474e-02</td>
<td>5.5723</td>
<td>2.1180</td>
</tr>
<tr>
<td>$2^{-6}$</td>
<td>2.2301e-04</td>
<td>3.8569e-03</td>
<td>6.6787</td>
<td>2.9749</td>
</tr>
<tr>
<td>$2^{-7}$</td>
<td>3.0505e-05</td>
<td>1.1132e-03</td>
<td>7.3105</td>
<td>3.4648</td>
</tr>
<tr>
<td>$2^{-8}$</td>
<td>3.9887e-06</td>
<td>2.9872e-04</td>
<td>7.6478</td>
<td>3.7265</td>
</tr>
<tr>
<td>$2^{-9}$</td>
<td>5.0994e-07</td>
<td>7.7352e-05</td>
<td>7.8220</td>
<td>3.8618</td>
</tr>
<tr>
<td>$2^{-10}$</td>
<td>6.4463e-08</td>
<td>1.9680e-05</td>
<td>7.9105</td>
<td>3.9305</td>
</tr>
</tbody>
</table>

This chapter is based on work previously published by the author in [25].
### Table 6.3: Global Vector (Fixed Time) Truncation Error

| $\Delta$ | $|e_1[N]|$  | $|e_2[N]|$  | $ratio_1$ | $ratio_2$ |
|-----------|-------------|-------------|-----------|-----------|
| $2^{-3}$  | 3.4015e-02 | 1.0889e-02 | -         | -         |
| $2^{-4}$  | 1.6650e-02 | 5.8648e-03 | 2.0430    | 1.8566    |
| $2^{-5}$  | 8.2397e-03 | 3.0103e-03 | 2.0207    | 1.9483    |
| $2^{-6}$  | 4.0990e-03 | 1.5218e-03 | 2.0102    | 1.9780    |
| $2^{-7}$  | 2.0443e-03 | 7.6479e-04 | 2.0050    | 1.9899    |
| $2^{-8}$  | 1.0209e-03 | 3.8332e-04 | 2.0025    | 1.9952    |
| $2^{-9}$  | 5.1013e-04 | 1.9189e-04 | 2.0013    | 1.9976    |
| $2^{-10}$ | 2.5498e-04 | 9.6001e-05 | 2.0006    | 1.9988    |
In the previous chapter we have shown that the TTS model has increased order of accuracy when compared to the usual Euler approximation, at least where local errors are concerned. In this chapter and the following two short chapters we will present three different cases where the TTS model, and its improved error properties, prove beneficial.

In this chapter we use the results presented in Chapter 6 regarding the truncation errors of the TTS model to develop an associated input-output sampled-data model. We will restrict ourselves to the class of nonlinear systems having no finite zeros. Important properties of the new model described are: (i) it has a local truncation error of order $\Delta^{n+1}$, where $n$ is the state dimension and $\Delta$ is the sample period, (ii) it depends on input and output only, and (iii) it makes explicit the extra sampling zero dynamics arising from the sampling process.

### 7.1 An input-output model

As stated above, we make the following additional restriction: the continuous system has no finite zero dynamics. Consequently we have $r = n$ in the model (5.14a)–(5.14e).

Based on the results of the previous chapter, we have the following result:

**Theorem 7.1.1.** For the class of nonlinear continuous time systems (5.6)–(5.9) having relative degree $n$, where $n$ is the state dimension, then an input-output sampled data model exists having
the following form:

\[(q - 1)^n \cdot \hat{y} = \Delta^n \cdot b(y, \ldots, \delta^{n-1}y) + \Delta^n \cdot a(y, \ldots, \delta^{n-1}y) \cdot p_n(q - 1) \cdot u \]

where \(q\) is the forward shift operator and its relationship to the delta operator is given by \(\delta = \frac{q - 1}{\Delta}\)

and where \(p_n(q - 1)\) is the Euler-Frobenius polynomial of order \(n\), as presented in Definition 5.4.2, defining the asymptotic sampling zeros.

Key properties of the above model are that it depends only on input-output data and that the Local Fixed Step Truncation error between \(\hat{y}\) and \(y\) is of order \(\Delta^{n+1}\).

**Proof.** We first note that for \(r = n\), it is straightforward to rewrite the model in the following form by using the forward shift operator \(q\):

\[(q - 1)^n \cdot \hat{y} = p_n(q - 1) \cdot \Delta^n \cdot \hat{\bar{F}}\]

where \(\hat{\bar{F}} = \hat{F}(\hat{\bar{\zeta}})\) depends on the \(n\) states defined in the normal form given in Definition 5.3.2. Since we seek an input-output model we need to translate this dependance on the states to a dependance on the output only. Recall that the \(n\) states are a truncated approximation of the output and its first \(n - 1\) derivatives. Thus, it motivates us to replace the approximated states with a delta operator approximation of the derivatives of the output. With this in mind, we introduce the following Taylor series expansion of \(\hat{\bar{F}}\):

\[
\hat{F}(\hat{\bar{\zeta}}) = b(\hat{z}_1, \ldots, \hat{z}_r) + u \cdot a(\hat{z}_1, \ldots, \hat{z}_r)
= b(y, \ldots, \delta^{r-1}y) + u \cdot a(y, \ldots, \delta^{r-1}y) + \left(\frac{\partial b}{\partial \hat{z}_1} + u \cdot \frac{\partial a}{\partial \hat{z}_1}\right) \cdot (\hat{z}_1 - y)
+ \left(\frac{\partial b}{\partial \hat{z}_2} + u \cdot \frac{\partial a}{\partial \hat{z}_2}\right) \cdot (\hat{z}_2 - \delta y)
+ \cdots + \left(\frac{\partial b}{\partial \hat{z}_r} + u \cdot \frac{\partial a}{\partial \hat{z}_r}\right) \cdot (\hat{z}_r - \delta^{r-1}y)
\]

We also note that, for \(i = 1, \ldots, n\), we can write

\[\hat{z}_i - \delta^{i-1}y = (\hat{z}_i - z_i) + (z_i - \delta^{i-1}y)\]
Next, using the known Local Truncation Error properties of the model (5.14a)–(5.14e) shown in Chapter 6, we have

\[ \hat{z}_i - z_i \in O(\Delta^{n+2-i}), \quad i = 1, \ldots, n \]
\[ z_i - \delta^{i-1} y \in O(\Delta), \quad i = 2, \ldots, n \]
\[ z_1 - y = 0 \]

We can immediately conclude that

\[ \hat{F}(\hat{\zeta}) = b(y, \ldots, \delta^{n-1} y) + u \cdot a(y, \ldots, \delta^{n-1} y) + O(\Delta) \]

and therefore

\[ \Delta^n \cdot \hat{F}(\hat{\zeta}) = \Delta^n \cdot \{ b(y, \ldots, \delta^{n-1} y) + u \cdot a(y, \ldots, \delta^{n-1} y) \} + O(\Delta^{n+1}) \]

On the other hand, since \( p_n(q - 1) \) is a monic polynomial in powers of \( (q - 1) \) we have that

\[ (q - 1)(b + u \cdot a) = (b^+ - b) + (u^+ a^+ - ua) \]

Using an Euler expansion of \( b \) we have that

\[ (b^+ - b) \in O(\Delta) \]

and also that

\[ u^+ a^+ - ua = (u^+ a^+ - u^+ a) + (u^+ a - ua) \]
\[ = u^+ (a^+ - a) + a (u^+ - u) \]
\[ = O(\Delta) + a (u^+ - u) \]
\[ = O(\Delta) + a \{(q - 1) \cdot u\} \]

since, using an Euler expansion of \( a \) we know that \( (a^+ - a) \in O(\Delta) \). Therefore, we can write

\[ (q - 1)(b + u \cdot a) = \{(q - 1) \cdot u\} \cdot a + O(\Delta) \]

Following the same argument, it is easily shown that

\[ (q - 1)^n (b + u \cdot a) = \{(q - 1)^n \cdot u\} \cdot a + O(\Delta) \]
Hence, we can write the model as

\[(q - 1)^n \hat{y} = \Delta^n \cdot b(y, \ldots, \delta^{n-1} y) + \Delta^n \cdot a(y, \ldots, \delta^{n-1} y) \cdot p_n(q - 1) \cdot u + O(\Delta^{n+1})\]

from where it is straightforward to note that the Local Fixed Step Truncation error of the output is of order $\Delta^{n+1}$. This completes the proof.

\[\square\]

**Remark 7.1.1.** We note that $p_n(q - 1) \cdot u$ coincides with the extra sampling zero polynomial that arises in the linear case [124, 5].

\[\square\]

### 7.2 Conclusions

This chapter has developed a sampled data output predictor model for continuous time systems having no finite zeros. Key properties of the predictor are that it depends on input-output data only and has Local Fixed Step Truncation Error of order $\Delta^{n+1}$. The model makes explicit the additional sampling zero dynamics that arise due to the sampling process. A potential application of the ideas described here include nonlinear system identification based on input-output data.

This chapter is based on work previously published by the author in [21].
Bias and Variance Issues in the Identification of Continuous Time Nonlinear Systems from Sampled Data

Many (and arguably most) physical systems are described by continuous time differential equation models. Moreover, the parameters within these equations are often of direct physical interest and hence of importance in any estimation study. On the other hand, experimental data collected from a system is almost always in sampled form. This raises the issue as to how one can best estimate the parameters in a continuous time model from sampled data.

Even in the case of linear systems, the sampled response of a continuous time system is a highly nonlinear function of the continuous parameters. For example, the sampled data state transition matrix takes the form $e^{A\Delta}$ where $A$, $\Delta$ are respectively the continuous system matrix and the sampling period. The nonlinear case is even more difficult since no closed form exists for the sampled data model.

In this chapter we argue that the use of the TTS sampled-data model allows one to obtain a much improved bias-variance tradeoff in equation error based system identification than would be achieved if an Euler integration based model were to be used.

The remainder of this chapter is organised as follows: In Section 8.1 we give notation and basic definitions. In Section 8.2 we briefly review the approximate sampled data models to be used.
In Section 8.3 we show the main result regarding the parameter estimation error. In Section 8.4 provide an interpretation of the result for finite data length. Finally, in Section 8.5 we draw conclusions.

8.1 Preliminaries

Throughout this chapter we limit attention to the single-input single-output case. We consider a nonlinear system of order \( n \) described as follows:

\[
\dot{x}(\tau, \theta_o) = f(x(\tau, \theta_o)) + g(x(\tau, \theta_o))u(\tau) \tag{8.1}
\]

\[
y(\tau, \theta_o) = h(x(\tau, \theta_o)) \tag{8.2}
\]

where \( x(\tau) \) is the state evolving in an open subset \( M \subset \mathbb{R}^n \), and where the vector fields \( f(\cdot), g(\cdot) \), and the output function \( h(\cdot) \) are analytic. The system is assumed to have a vector of true parameters \( \theta_o \) and uniform relative degree \( r \) when \( x(\tau) \in M \).

We will be interested in the problem of estimating \( \theta_o \) from a finite sample of observations where the sample period is \( \Delta \). We assume that the samples satisfy

\[
y_t = y(\tau, \theta_o) + v_t, \quad t \in \mathbb{N}^+ \tag{8.3}
\]

where \( y_t \) denotes \( y(t\Delta) \) and \( \{v_t\} \) is a stationary stochastic process.

Remark 8.1.1. In the sequel, for simplicity of exposition, we will take \( \{v_t\} \) to be an i.i.d gaussian sequence of mean zero and variance \( \sigma^2 \).

We introduce an (approximate) sampled model having output \( \hat{y}_t(\theta) \) parameterised by \( \theta \). We assume that the approximate model has local (fixed step) truncation error of order \( \Delta^m \). This is expressed as

\[
\hat{y}_t(\theta_o) - y_t(\theta_o) = E(\Delta, \theta_o) \in O(\Delta^m), \quad \forall t \in \mathbb{N} \tag{8.4}
\]

Our goal is to quantify the impact that the truncation error \( E(\Delta, \theta_o) \) has on the estimation of \( \theta_o \). Before proceeding, we describe two possible choices for \( \hat{y}_t(\theta_o) \).

8.2 Approximate Sampled Data Models

Our analysis will cover any model for which an error quantification of the form (8.4) is available. Two possible choices are discussed below. We assume throughout that the input is generated by a
8.3 Maximum Likelihood Estimation

zero-order hold, i.e. \( u(\tau) = u_t, \tau \in [t\Delta, t\Delta + \Delta) \).

8.2.1 Euler Integration

The simplest possible (approximate) model is obtained from (8.1)–(8.2) by Euler integration, which leads to:

\[
\hat{x}_{t+1} = \hat{x}_t + \Delta \cdot (f(\hat{x}_t, \theta) + g(\hat{x}_t, \theta) \cdot u_t) \quad (8.5)
\]

\[
\hat{y}_t(\theta) = h(\hat{x}_t, \theta), \quad t \in \mathbb{N} \quad (8.6)
\]

It is well known [11] that the Local and Local Fixed Step truncation error in the output for Euler integration is of order \( \Delta^2 \), i.e. \( E(\Delta, \theta_o) \in \mathcal{O}(\Delta^2) \).

8.2.2 Truncated Taylor Series

Here we consider the TTS sampled-data model presented in Definition 5.3.2.

It has been shown in Chapter 6 that the Local and Local Fixed Step truncation errors in the output for this model are of order \( \Delta^{r+1} \) and \( \Delta^{r+1} \) respectively, where \( r \) is the relative degree of the system. This implies that \( E(\Delta, \theta_o) \in \mathcal{O}(\Delta^{r+1}) \) for this specific model.

Remark 8.2.1. Note that for \( r > 1 \), the model in Section 8.2.2 has smaller output truncation errors than the model in Section 8.2.1. We will discuss the implications of this observation in Section 8.4. □

8.3 Maximum Likelihood Estimation

We now return to the problem of estimating of \( \theta_o \). We assume we are given observations \( \{y_0, \ldots, y_N\} \). We also assume that the initial state is known, say the origin. Since, \( \{v_t\} \) is an i.i.d. gaussian sequence, the (negative) likelihood cost function can be written as

\[
J_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} [y_t - \hat{y}_t(\theta)]^2 \quad (8.7)
\]

We then define

\[
\hat{\theta}_N = \arg \min_{\theta} J_N(\theta) \quad (8.8)
\]
We will also be interested in the asymptotic estimate $\theta^*$, defined by

$$\theta^* = \arg \min_{\theta} \lim_{N \to \infty} J_N(\theta) = \lim_{N \to \infty} \hat{\theta}_N$$

(8.9)

The existence of $\theta^*$ requires the assumption that there exist well defined limits:

$$J_N(\theta) \to J_\infty(\theta)$$

(8.10)

$$\hat{\theta}_N \to \theta^*$$

(8.11)

Note also that $\theta_o$ satisfies

$$\theta_o = \arg \min_{\theta} \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^N [y_t - \tilde{y}_t(\theta)]^2$$

(8.12)

when the model structure corresponds to the true system, i.e. $\tilde{y}_t(\theta) = \hat{y}_t(\theta)$.

In addition we assume that

$$J''_\infty(\theta) \geq cI, \quad c > 0, \quad \forall \theta \in B$$

(8.13)

where $B$ is a neighbourhood that includes $\theta_0$ and $\theta^*$.

Our intention is to quantify the estimation error $\theta_o - \hat{\theta}_N$. For this purpose we will establish the following key result:

**Theorem 8.3.1.** Subject to the above assumptions and assuming the output truncation error satisfies $E(\Delta, \theta) \in O(\Delta^m)$, then the estimation error $\theta_o - \hat{\theta}_N$ can be written as

$$\theta_o - \hat{\theta}_N = C(N) \left\{ \frac{1}{\sqrt{N}} \cdot A(N) + B(N) \right\}$$

(8.14)

where $C(N) \in O(1)$ is such that

$$\lim_{N \to \infty} C(N) \leq \sup_{\theta \in B} \|J''_\infty(\theta)^{-1}\| \leq I \cdot 1/c$$

(8.15)

and where $A(N)$ converges to a normal random variable having zero mean and covariance

$$\Omega = \lim_{N \to \infty} \frac{1}{N} \left( \sum_{t=1}^N \hat{\Phi}_t \hat{\Phi}_t^T \right) \sigma^2$$

(8.16)

where

$$\hat{\Phi}_t = \frac{\partial \hat{y}_t}{\partial \theta} \bigg|_{\theta_o}$$

(8.17)

and $B(N)$ is of order $\Delta^m$. 

Proof. By definition we have

$$\hat{\theta}_N = \arg\min J_N(\theta)$$

$$= \arg\min \frac{1}{N} \sum_{t=1}^{N} [y_t - \hat{y}_t(\theta)]^2 \quad (8.18)$$

It follows that

$$J'_N(\hat{\theta}_N) = 0 \quad (8.19)$$

Then we can write

$$J'_N(\theta_0) = J'_N(\theta_0) - J'_N(\hat{\theta}_N)$$

$$= J''_N(\xi)[\theta_0 - \hat{\theta}_N] \quad (8.20)$$

where we have used the mean value theorem. Note that

$$J''_N(\theta) = \frac{1}{N} \sum_{t=1}^{N} 2v_t \left( -\frac{\partial^2 \hat{y}_t}{\partial \theta^2} \right) \bigg|_{\theta}$$

$$+ \frac{2}{N} \sum_{t=1}^{N} \left[ \{\hat{y}_t - \hat{y}_t(\theta)\} \left( -\frac{\partial^2 \hat{y}_t}{\partial \theta^2} \right) \bigg|_{\theta} + \left( \frac{\partial \hat{y}_t}{\partial \theta} \right)^2 \bigg|_{\theta} \right] \quad (8.21)$$

where

$$J''_\infty(\theta) = \lim_{N \to \infty} J''_N(\theta)$$

$$= \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} 2 \left[ \{\hat{y}_t - \hat{y}_t(\theta)\} \left( -\frac{\partial^2 \hat{y}_t}{\partial \theta^2} \right) \bigg|_{\theta} + \left( \frac{\partial \hat{y}_t}{\partial \theta} \right)^2 \bigg|_{\theta} \right] \in O(1) \quad (8.22)$$

For convenience, we define

$$C(N) = J''_N(\xi)^{-1} \quad (8.23)$$

and (8.15) follows from the assumptions. From (8.20)

$$(\theta_0 - \hat{\theta}_N) = C(N) \left[ \frac{1}{N} \sum_{t=1}^{N} (y_t - \hat{y}_t(\theta_0)) \right] \hat{\Phi}_t(\theta_0)$$

$$= C(N) \frac{1}{N} \sum_{t=1}^{N} [v_t + \hat{y}_t(\theta_0) - \hat{y}_t(\theta_0)] \hat{\Phi}_t(\theta_0) \quad (8.24)$$

$$= C(N) \left\{ \frac{1}{\sqrt{N}} \left[ \frac{1}{\sqrt{N}} \sum_{t=1}^{N} v_t \hat{\Phi}_t(\theta_0) \right] + \frac{1}{N} \sum_{t=1}^{N} [\hat{y}_t(\theta_0) - \hat{y}_t(\theta_0)] \hat{\Phi}_t(\theta_0) \right\}$$

$$= C(N) \left\{ \frac{1}{\sqrt{N}} \sum_{t=1}^{N} v_t \hat{\Phi}_t(\theta_0) \right\}$$
We now define

\[ A(N) = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} v_t \hat{\Phi}_t(\theta_0) \]  

\[ B(N) = \frac{1}{N} \sum_{t=1}^{N} [\hat{y}_t(\theta_0) - \hat{y}_t(\theta_0)] \hat{\Phi}_t(\theta_0) \]

A standard result, e.g. [81, 121], shows that \( A(N) \) converges to a normal random variable as stated in the theorem. On the other hand, it is easy to see that

\[ |B(N)| = \left| \frac{1}{N} \sum_{t=1}^{N} (\hat{y}_t(\theta_0) - \hat{y}_t(\theta_0)) \hat{\Phi}_t(\theta_0) \right| \]

\[ \leq \sqrt{\frac{1}{N} \sum_{t=1}^{N} |\hat{y}_t(\theta_0) - \hat{y}_t(\theta_0)|^2} \sqrt{\frac{1}{N} \sum_{t=1}^{N} |\Phi_t|^2} \]

\[ \leq K E(\Delta, \theta_0) \]

which completes the proof.

\[ \square \]

### 8.4 Finite Data Length Interpretation

The quantification given in Section 8.3 is asymptotic in \( N \). In practice one can use this result to motivate a quantification of the error for finite \( N \) as:

\[ MSE \approx \left| \lim_{N \to \infty} C(N) \right|^2 \cdot |B(N)|^2 + \frac{1}{N} \cdot \text{trace}\left\{ \lim_{N \to \infty} C(N) \left( \frac{1}{N} \sum_{t=1}^{N} \Phi_t^T \Phi_t \right) \right\} \]

\[ (8.31) \]

where \( |B(N)|^2 \in \mathcal{O}(\Delta^{2m}) \).

Moreover, we have seen that, for fixed \( N \), \( E(\Delta, \theta_\ast) \in \mathcal{O}(\Delta^2) \) for the model of Section 8.2.1 and \( E(\Delta, \theta_\ast) \in \mathcal{O}(\Delta^{r+1}) \) for the model of Section 8.2.2.

The expression \( (8.31) \) shows that \( |B(N)|^2 \) represents a lower bound on the achievable MSE no matter how much data we collect. In particular, there is no point choosing an \( N \) such that the second term (variance) is much smaller than the unavoidable bias term.

Thus, if one can collect as much data as one likes, then it is preferable to use the model presented in Section 8.2.2 since the bias error is smaller than if the Euler model were to be used for a given sampling period.
8.5 Conclusions

In this chapter we have examined the question of bias-variance tradeoffs when estimating the parameters of continuous time models using sampled data. We have shown that when approximate sampled data models are used then bias errors result which place an unavoidable lower bound on the estimation accuracy no matter how much data is collected. We have also shown that this lower bound can be reduced by using models other than those based on simple Euler integration.

This chapter is based on work previously published by the author in [27].
High Gain Control with Feedback Linearisation under Sampling

In this chapter we study the possible implications of the TTS model, introduced in Chapter 5, in the problem of high-gain sampled-data feedback control.

Most real world systems exhibit nonlinear behaviour. The study of nonlinear systems is of considerable practical importance and has raised significant research interest \[112, 64, 108, 67, 3, 3\]. The problems involved are highly nontrivial and lead to significant theoretical challenges. Nonetheless, there is a significant, and growing, body of work, covering many different aspects of nonlinear control under different headings, e.g. optimal control \[117\], nonlinear model predictive control \[58\], feedback linearisation \[64\], flatness \[39\], dissipativity \[10\], etc. Much of this work can be characterised by two observations: (i) the majority of the work deals with continuous time systems or (ii) treats inherently discrete time systems. There remains a significant gap relating to the digital control of underlying continuous time systems \[77\].

Sampled-data control of nonlinear systems represents a major challenge because it raises difficulties associated with the impact of sampling in a nonlinear environment. Indeed, it is well known that algorithms that work perfectly well for continuous time systems can fail catastrophically when implemented digitally \[98\]. The controller design problem for sampled-data systems has been explored from many different angles, e.g. \[92, 94, 77, 93, 76, 93, 95\]. The framework that has been developed in \[99, 98\] has served as a starting point for many important results. Specifically, results are available regarding stability conditions and tools to aid the controller design based on
approximate discrete models.

In this framework, it may seem that, provided one samples quickly enough, then any controller design based on a simple discrete approximate model, e.g. using Euler integration, will stabilise a continuous time nonlinear system. Indeed, the same hypothesis seems reasonable for linear systems. Unfortunately, the hypothesis is false. There is substantial literature on this topic and many counter examples exist, see for example [57, 99, 98].

Indeed, a clear example of what happens when the sampling zeros are not included in the sampled-data model for the controller design is given in the triple integrator motivational example discussed in [98], where a deadbeat controller, designed based on the Euler approximation of the states, fails to stabilise the real plant regardless of the sampling period. Designing the same controller, but based on a model that includes the sampling zeros (which for this simple example actually leads to an exact description), results in a stable control loop.

We argue that there is a major deficiency in some of the commonly used approximate models, namely they do not capture the, so-called, “sampling zeros”. We argue that the sampled-data model has to account, in some way, for the sampling zero dynamics if high-bandwidth control of a continuous nonlinear system is required.

The above line of reasoning leads us to suggest that one should include the sampling zero dynamics when designing sampled data controllers. This is straightforward in the linear case, where an exact discrete model can be computed. In addition, it seems reasonable that, at fast sampling rates, one should be able to approximate the “true” sampling zeros by their asymptotic locations. More importantly, as explained above, this idea extends to the nonlinear case. The key observation being that, no matter how fast one samples, the sampling zero dynamics cannot be ignored.

The key contribution in this chapter is to show the importance of including the sampling zeros in the model for controller/observer design. We will exploit the fact that the TTS model accounts for the asymptotic location of the sampling zero dynamics to gain insight into one failure mechanism for sampled-data controllers for continuous time nonlinear systems.

The remainder of this chapter is organised as follows: In Section 9.1 we discuss high-gain feedback control using the the TTS model via feedback linearisation. In Section 9.2 we illustrate the ideas for both linear and nonlinear systems. Finally, in Section 9.3 we draw conclusions.
9.1 Feedback Linearisation

A possible approach to test the above ideas, is to use feedback linearisation [64] to transform the nonlinear system to a linear one. This would be a simple way to inject the idea of sampling zeros. However, several difficulties arise.

The first problem that arises is that feedback and sampling do not commute, so it is much harder to linearise a system under sampling [57] and even if the continuous-time system is feedback linearisable the sampled counterpart may not be. Stronger conditions are needed for a system to be feedback linearisable under sampling [2]. Different solutions have been given in the literature to overcome these restrictions and to recover the property of feedback linearisability under sampling. For example, in [57], a multirate sampling strategy is employed, in which the control move is implemented each sampling period while the state is measured every $N \geq 2$ sampling periods. In [94] it is shown how to compute an extended feedback control law that provides an approximate solution to the linearisation problem.

The second problem that arises is that the exact nonlinear sampled data model is usually unknown and therefore one must use approximate models to design the linearising controller. The problem of stabilising a nonlinear system based only on an approximate model is addressed in [99, 98]. However, there are no guidelines as to how to choose, or obtain, a better sampled-data model.

The idea of feedback linearisation applied to the TTS model is straightforward. In the sequel we will define $v_k$ such that

$$b(\hat{\zeta}, \hat{\eta}) + a(\hat{\zeta}, \hat{\eta}) \cdot u_k = v_k$$

(9.1)

gives a linear system when substituted in (5.14).

9.2 Illustrative Examples

9.2.1 A simple linear example

To illustrate the path we intend to follow, it is instructive to first review the linear case, where all models can be quantified explicitly. In this case, the sampling zero dynamics are perfectly known and the associated control design is well documented. Of course linear systems are a special case of nonlinear systems, so it is instructive to look at the linear case to gain insight into what might happen in the nonlinear case.
Consider the following simple unstable linear system
\[ G(s) = \frac{1}{(s - 1)^3} \]

We notice that this system has relative degree three and therefore linear theory tells us that there will, generically, be two additional sampling zeros. The exact sampled-data model with sampling period \( \Delta = 0.01 \) is easily calculated
\[ G_q(z) = \frac{1.6792e - 07(z + 3.76)(z + 0.27)}{(z - 1.01)^3} \]  \hspace{1cm} (9.2)

Of course, in the nonlinear case, we cannot expect an exact model and so we need to have approximate solutions. The simplest of those, and the model which reflects the centre of gravity of nonlinear control theory, is to use Euler integration, which leads, for the above example, to the following model:
\[ G_q^E = \frac{1e - 06}{(z - 1.01)^3} \]  \hspace{1cm} (9.3)

Alternatively, if one follows the ideas in Section 5.2 of Chapter 5 specialised to linear systems, then one obtains a new model which, inter alia, includes the sampling zero dynamics. We call the associated transfer function \( G_q^{TTS} \):
\[ G_q^{TTS} = \frac{1.6667e - 07(z + 3.732)(z + 0.2679)}{(z - 1.009)(z^2 - 2.021z + 1.021)} \]  \hspace{1cm} (9.4)

We see that this model has two additional zeros which are not present in the continuous model. Their locations are at \( z = \{-3.732, -0.2679\} \) which is almost exactly the location of the extra zeros in the true model (9.2). If we plot the bode diagram of the three transfer functions: the true one, the Euler model and \( G_q^{TTS} \) we obtain the plots in Fig. 9.1, where \( \omega \in [0, \pi/\Delta] \). We notice that, at high frequencies, the TTS model and the true system correspond very well both in magnitude and phase, but the Euler model has a major deficiency with respect to phase. Indeed there is a 180 degree phase error at the Nyquist frequency. Of course, this is a major issue for wide-bandwidth control system design.

In the context of this simple linear problem, we next follow the approach that we plan to adopt in the nonlinear case. We will carry out a linear design for the three models, assigning the closed loop poles to the origin. We apply the resultant controller to the true system. Of course the design based on the true model works perfectly. The design based on the Euler model, even though the sampling rate is extremely high (\( \Delta = 0.01 \)), and even if we measure the states of the system, leads to an unstable solution! Perhaps this is not surprising because we have already revealed the 180 degree phase error in the vicinity of the folding frequency. Output feedback
design is no better and, if anything, a little worse. The closed loop poles, when applied to the true system, are as follows: (i) state feedback design \( z = \{-2.7079, 0.4897 \pm j0.1157\} \), (ii) output feedback design \( z = \{-2.5167, 0.1002 \pm j2.5682, 0.7982 \pm j0.3139, 0.7201\} \), which are clearly outside the unit circle. On the other hand, the design based on the TTS model leads to the following closed loop poles: (i) state feedback design \( z = \{-0.2902, 0.1257 \pm j0.1497\} \), (ii) output feedback design \( z = \{-0.4849 \pm j0.0353, 0.0402 \pm j0.6392, 0.4451 \pm j0.1950\} \), which are clearly stable.

### 9.2.2 A nonlinear example

The conjecture in this case is that the applicability of the sampled-data feedback linearisation strategy will be extended by improving the model via the addition of the zero dynamics. For all cases described below the sampling period is \( \Delta = 0.1 \), the initial conditions are \( x_1(0) = 1, x_2(0) = 0 \) and the discrete closed loop poles are assigned to the origin. To illustrate the ideas we begin with
the system:

\[
\dot{x}_1 = x_2 \\
\dot{x}_2 = (1 + |x_1|) \cdot u
\]

(9.5a)

(9.5b)

The first step is to obtain the sampled-data model based on the ideas in Section 5.2. This leads to the following approximate sampled-data model

\[
x_1^+ = x_1 + \Delta x_2 + \frac{\Delta^2}{2} (1 + |x_1|)u \\
x_2^+ = x_2 + \Delta (1 + |x_1|)u
\]

We then make the following substitution

\[
v = (1 + |x_1|)u
\]

We see that \(v\) does not appear in a simple way in the model but appears in the two discrete state equations. Moreover there are marginally stable zero dynamics. As we know from linear theory there will be certain output trajectories which cannot be achieved with finite input, in particular, \(\sin(\pi t/\Delta)\). But with that caveat in mind, we now have a linear system. We then perform pole assignment on this system.

Fig. 9.2 shows the output response of the true system when the controllers, designed based on the Euler and TTS models, are implemented. Clearly, the design based on the TTS model is perfectly satisfactory, whereas the design based on an Euler approximation is unstable.

Similar behaviour is observed with other systems. Consider the following

\[
\dot{x}_1 = x_2 \\
\dot{x}_2 = (1 + 0.2 \cdot |x_1|) \cdot u
\]

(9.6a)

(9.6b)

Fig 9.3 shows the output response for the different controllers. It can be seen that there is a perfectly acceptable result for the TTS model, and whilst the Euler model is now stable, the response has deteriorated.

**Remark 9.2.1.** We note that for this particular example, the strategies we are adopting are related to some observations made in past literature, in particular, [57] uses a multirate controller. From our perspective, the multirate controller exposes the unstable zero dynamics arising from sampling. The rest of the design in [57] deals with this appropriately. It is not explicitly stated in any of the
previous work, but our perspective on that solution is that the sampling zero dynamics are crucial and the multirate controller captures the issue. Of course, having recognised the zero dynamics, there are many other ways to proceed.

Remark 9.2.2. There exists an implicit trade-off between the sampling period and the achievable closed-loop bandwidth. In other words, the faster one samples, there exists a larger feedback gain that stabilises the sampled-data system. The reason for this is that the real sampling zero dynamics approach their asymptotic locations as the sampling period goes to zero and therefore, the “high frequency” mismatch between the approximate and the real sampled-data model vanishes. On the other hand, if a high sampling rate is not feasible because of physical limitations, then there will be an upper bound on the closed-loop bandwidth that can be achieved, i.e. deadbeat control may not be possible.
The ideas previously presented have lead us to formulate the following conjecture:

**Conjecture 9.2.1.** To improve the performance of a high-bandwidth controller in the sampled-data control of nonlinear systems, e.g. enlarge the region of attraction or, in some cases, achieve stability, the following statements are proposed:

- **It is necessary that the sampled-data model accounts for the sampling zero dynamics.**
- **Under fast sampling, it is sufficient for the sampled-data model to contain the asymptotic sampling zero dynamics.**

\[ \square \]

This conjecture goes beyond the scope of the thesis, however its detailed study is already the aim of subsequent research.

**9.3 Conclusions**

In this chapter, we have discussed the role of asymptotic sampling zero dynamics in the sampled-data control of continuous nonlinear systems. The result builds on the properties of the TTS model, in particular, the inclusion of the sampling zero dynamics. Several examples have been presented showing that including the asymptotic zero dynamics can lead to significantly improved high-bandwidth performance of sampled-data controllers for a class of continuous nonlinear dynamic systems.

This chapter is based on work previously published by the author in [23].
So far in this thesis we have only addressed deterministic systems. In this chapter we extend the analysis to stochastic systems. In particular, we demonstrate that the concept of relative degree plays a key role in obtaining higher order of accuracy for integration procedures compared to the stochastic Euler-Maruyama approximation.

We first extend the definitions of truncation errors proposed in Chapter 6 in a meaningful sense to the stochastic framework. Furthermore, we will apply the new definitions to a specific stochastic sampled-data model, a stochastic counterpart of the TTS model, aptly named Stochastic TTS model or STTS, and will show that this new model exhibits improved order of accuracy over the Euler-Maruyama method at no significant computational cost.

The associated challenge is non-trivial. Firstly, the rules of classic calculus do not apply, i.e. one needs to turn to new definitions, e.g. Ito or Stratonovich integrals. Secondly, as is well known, one cannot just apply the standard numerical algorithms developed for deterministic differential equations (DDE) to stochastic differential equations (SDE) and preserve the same order of accuracy.

Also, there are different types of definitions for the errors. For example, in [72] only global errors with fixed (continuous) time interval are considered. Similarly, Milstein [90] has considered global errors with fixed continuous time interval based on the knowledge of the local errors of the approximate SDE, for both the mean and mean squared errors.

In addition, we will show that, given the improved error properties, the model includes stochastic sampling zero dynamics. These additional dynamics are a consequence of the sampling process and have no counterpart in the continuous-time system. In fact, the additional dynamics are linked to
10. A Sampled-Data Model for Stochastic Nonlinear Systems

The sampling zero dynamics of deterministic systems \cite{5,124} and linear stochastic systems \cite{119}.

The presence of these additional zeros in stochastic models has proven to be a key issue, for example, in parameter estimation of continuous-time auto-regressive (CAR) systems. In \cite{114} and \cite{36} bias compensation was proposed to obtain accurate parameter estimates when using Euler approximation of derivatives. A related discussion is provided in \cite{33} and the references therein. In fact, this problem can be avoided by taking into account the stochastic sampling zeros of the output spectrum \cite{79} or by restricting the estimation bandwidth to reduce the impact of the approximation error \cite{126, Chapter 19}. Although the results above are restricted to linear systems, similar issues are expected to arise for nonlinear stochastic systems when using sampled-data models. This idea will be explored in this chapter.

The remainder of this chapter is organised as follows: In Section 10.1 we motivate the problem. In Section 10.2 we define the different error measures for stochastic systems. In Section 10.3 we review the Stochastic Truncated (Ito)-Taylor Series (STTS) sampled-data model that will be used in this chapter. In Section 10.4 we define the truncation errors of the STTS model. In Section 10.5 we establish the order of the different truncation errors of the STTS model. In Section 10.6 we illustrate, via simulation, the truncation errors of the STTS model. In Section 10.7 we prove that the STTS model includes specific sampling zero dynamics that are tightly related to the improved order of accuracy of the model. In Section 10.8 we illustrate the importance of the inclusion of the sampling zero dynamics via a parameter estimation problem. Finally, in Section 10.9 we draw conclusions.

10.1 Motivation

In the deterministic case, it is known from Chapter \ref{5} that a nonlinear system

\begin{align}
\dot{x}(t) &= f(x, t) + g(x, t) \cdot u(t) \\
y(t) &= h(x, t) \cdot x(t)
\end{align}

(10.1)
with relative degree \( r < n \), can be expressed in a special form, called the normal form \([64]\), via a transformation \( z = \Phi(x) \), where

\[
\begin{align*}
\dot{z}_1 &= z_2 \\
\vdots \\
\dot{z}_{r-1} &= z_r \\
\dot{z}_r &= b(\zeta, \eta) + a(\zeta, \eta) \cdot u(t) \\
\dot{\eta} &= c(\zeta, \eta)
\end{align*}
\]

where the output is \( y = z_1 \) and the states have been partitioned as follows:

\[
\begin{align*}
\zeta &= \begin{bmatrix} z_1 & \ldots & z_r \end{bmatrix}^T \\
\eta &= \begin{bmatrix} z_{r+1} & \ldots & z_n \end{bmatrix}^T
\end{align*}
\]

and the functions \( a(\cdot), b(\cdot), c(\cdot) \) are well defined.

We have seen in Chapters 5 and 6 that a sampled-data deterministic model, i.e. the TTS model, can be obtained from the system expressed in normal form \((10.3)\) by using a Taylor series expansion of different order on each state. The resulting sampled-data model has two main properties,

(i) it has improved order of accuracy, compared to an Euler approximation, at no extra cost (see Section 6.2) and,

(ii) it accounts for the asymptotic location of sampling zero dynamics (see Section 5.4.3).

More specifically, sampling zeros (sampling zero dynamics) arise in the discretisation of continuous time systems and have no continuous time counterpart. The above result is well known for linear systems, both deterministic \([5]\) and stochastic \([119]\). The extension to nonlinear deterministic systems has been presented in \([91, 124]\). To the best of the current authors’ knowledge the question has previously remained open for the nonlinear stochastic case.

In this context, this chapter investigates whether or not sampling zero dynamics arise in the sampling of nonlinear stochastic systems. To this end, we will study the following set of Ito stochastic differential equations, which are analogous to the deterministic normal form described
previously in \([10.3]\),

\[
\begin{align*}
\frac{dx_1}{dt} &= x_2 \quad \text{(10.6a)} \\
\vdots \\
\frac{dx_{r-1}}{dt} &= x_r \quad \text{(10.6b)} \\
\frac{dx_r}{dt} &= a(X_t)dt + b(X_t)dv_t \quad \text{(10.6c)}
\end{align*}
\]

where \(v(t)\) is a Wiener process of unitary incremental variance, i.e. \(E\{dv_t^2\} = dt\).

A sampled-data model based on \([10.6]\) has been proposed in \([125]\) following ideas analogous to those presented in \([124]\) for deterministic systems. However, neither of the two aforementioned properties derived in the deterministic case, i.e. the improved order of accuracy or the inclusion of the asymptotic sampling zeros, have been proven in the stochastic framework.

### 10.2 Truncation Errors

Similar to the deterministic case, in the stochastic framework we define local and global errors, corresponding to a fixed number of steps or fixed continuous-time interval, respectively. The following notation will be used:

We define \(T = N \cdot \Delta\) as the continuous time horizon over which the approximate model will be used, where \(\Delta\) is the length of the time discretisation and \(N\) is the number of steps. In addition, we will use \(x_i[k]\) or \(x_{i,k}\) to denote \(x_i(k\Delta)\).

We introduce the following definitions of truncation error:

**Definition 10.2.1 (Local Vector Truncation Error).** Consider a dynamical system with states \((x_1, \ldots, x_n)\) and an associated approximate model with states \((\hat{x}_1, \ldots, \hat{x}_n)\). Let \(M(x - \hat{x})\) be a measure of the error. Then, the local vector truncation error of the approximate model is said to be of the order of \((\Delta^{m_1}, \ldots, \Delta^{m_n})\) if and only if

\[
\begin{align*}
M(x_1[k] - \hat{x}_1[k]) &= 0 \\
\vdots \\
M(x_n[k] - \hat{x}_n[k]) &= 0 \\
M(x_1[k+1] - \hat{x}_1[k+1]) &= O(\Delta^{m_1}) \\
\vdots \\
M(x_n[k+1] - \hat{x}_n[k+1]) &= O(\Delta^{m_n})
\end{align*}
\]

where \(m_1, \ldots, m_n\) are integers.

**Definition 10.2.2 (Local Vector Fixed Step Truncation Error).** The local vector fixed step truncation error of an approximate model is said to be of the order of \((\Delta^{m_1}, \ldots, \Delta^{m_n})\) if and only if,
for initial state errors satisfying

\[ M(x_1[k] - \hat{x}_1[k]) \in O(\Delta^{\bar{m}_1}) \] (10.8)

\[ \vdots \]

\[ M(x_n[k] - \hat{x}_n[k]) \in O(\Delta^{\bar{m}_n}) \] (10.9)

for any \( \bar{m}_i \geq \bar{m}_i, i = 1, \ldots, n \) then after \( \bar{N} \) steps, where \( \bar{N} \) is a finite fixed number, i.e. does not depend on \( \Delta \), we have that

\[ M(x_1[k + \bar{N}] - \hat{x}_1[k + \bar{N}]) \in O(\Delta^{\bar{m}_1}) \] (10.10)

\[ \vdots \]

\[ M(x_n[k + \bar{N}] - \hat{x}_n[k + \bar{N}]) \in O(\Delta^{\bar{m}_n}) \] (10.11)

\[ \square \square \square \]

**Definition 10.2.3** (Global Vector Truncation Error). The global vector (fixed time) truncation error of an approximate model is said to be of the order of \( (\Delta^{\bar{m}_1}, \ldots, \Delta^{\bar{m}_n}) \) if and only if, for initial state errors satisfying

\[ M(x_1[k] - \hat{x}_1[k]) \in O(\Delta^{\bar{m}_1}) \] (10.12)

\[ \vdots \]

\[ M(x_n[k] - \hat{x}_n[k]) \in O(\Delta^{\bar{m}_n}) \] (10.13)

for any \( \bar{m}_i \geq \bar{m}_i, i = 1, \ldots, n \) then after a fixed (continuous) time \( T \), i.e. after \( N = \lfloor T/\Delta \rfloor \) steps, where \( \lfloor \cdot \rfloor \) denotes the floor function, we have that

\[ M(x_1[k + N] - \hat{x}_1[k + N]) \in O(\Delta^{\bar{m}_1}) \] (10.14)

\[ \vdots \]

\[ M(x_n[k + N] - \hat{x}_n[k + N]) \in O(\Delta^{\bar{m}_n}) \] (10.15)

\[ \square \square \square \]

The previous definitions only deal with the evolution of the errors, e.g. when the error is measured, but say nothing on how the error is measured. We define two error cost functions \( M(\cdot) \) \[72, 90\], namely:

(i) Weak Convergence, i.e. the error of the mean:

\[ M(\cdot) = |E\{\cdot\}| \]
(ii) Strong Convergence, i.e. the mean of the error:

\[ M(\cdot) = E\{\cdot\} \]

where \( E\{\cdot\} \) is the expectation operator. Note that, through the Lyapunov inequality, strong convergence can also be characterised and bounded in the mean square sense,

\[ E\{|\cdot|\} \leq E\{|\cdot|^2\}^{1/2}. \]  \( (10.16) \)

**Remark 10.2.1.** The main difference between strong and weak convergence is that the former implies sample-path convergence, i.e. the true and approximate trajectories are “close” to one another, whereas the latter is concerned with convergence of a function of the process, e.g. first or second order moments.

**Remark 10.2.2.** It is important to note that reference [72] deals with Global Truncation Errors (Definition 10.2.3). In addition, [72] goes one step further and establishes an uniform error bound over the whole interval \([0, T]\) instead of just an error bound at the final instant \(T\). It is because of this worst case error bound that the results in [72] are more conservative than those which apply to Local errors, which explains the results presented in [125].

### 10.3 A Stochastic Sampled Data Model

In this section we review the sampled data model described in [125] for a class of stochastic nonlinear systems.

We consider the set of Ito stochastic differential equations

\[ dZ(t) = A(t, Z)dt + B(t, Z)dv \] \( (10.17) \)

where the functions \( A : \mathbb{R} \times \mathbb{R}^{r \times 1} \to \mathbb{R}^{r \times 1} \) and \( B : \mathbb{R} \times \mathbb{R}^{r \times 1} \to \mathbb{R}^{r \times 1} \) are nonlinear functions, and are assumed analytic and \( v(t) \) is a Wiener process of unitary incremental variance.

By analogy with the deterministic case [64], we consider a class of stochastic nonlinear systems in normal form, i.e. we assume there exists a transformation \( X = \Phi(Z) \) such that the system \( (10.17) \)
can be rewritten as

\begin{align}
      dx_1 &= x_2 \, dt \\
            &\vdots \\
      dx_{r-1} &= x_r \, dt \\
      dx_r &= a(X_t) \, dt + b(X_t) \, dv_t 
\end{align}

(10.18a)

(10.18b)

(10.18c)

We then consider the following approximate sampled data model, based on a truncated Ito-Taylor series expansion of the states:

**Definition 10.3.1 (STTS Model [125]).** Consider the continuous-time stochastic system (10.18), with output \( y = x_1 \). Then the STTS approximate model is

\[ \hat{X}_{k+1} = A_q \hat{X}_k + B_q a(\hat{X}_k) + b(\hat{X}_k) \bar{V}_k \]

(10.19)

where \( \hat{X}_k = \hat{X}(k\Delta) \), and

\[ A_q = \begin{bmatrix} 1 & \Delta & \cdots & \Delta^{n-1} \atop 0 & 1 & \cdots & \Delta^{n-2} \atop \vdots & \ddots & \ddots & \vdots \atop 0 & \cdots & 0 & 1 \end{bmatrix} \]

(10.20)

\[ B_q = \begin{bmatrix} \Delta^n \atop \Delta^{n-1} \atop \vdots \atop 0 \end{bmatrix}^T \]

(10.21)

\[ \bar{V}_k = \begin{bmatrix} \bar{v}_1 & \cdots & \bar{v}_n \end{bmatrix}^T \]

(10.22)

\[ \bar{v}_\ell = \int_{k\Delta}^{(k+1)\Delta} (\Delta - \tau)^{(n-\ell)} \, dv_\tau, \quad \ell = 1, \ldots, n \]

(10.23)

with \( a(\cdot) \) and \( b(\cdot) \) defined as in (10.18) and where

\[ E\{\bar{v}_i \bar{v}_j\} = \frac{\Delta^{2n-i-j+1}}{(n-1)!(n-j)!(2n-i-j+1)!} \]

(10.24)

are the elements of the covariance matrix \( E\{\bar{V}_k \bar{V}_k^T\} \).

We will refer to this model as the Stochastic Truncated (Ito)-Taylor Series (STTS) model.
10.4 Defining the Errors of the STTS Model

In order to compute the errors associated with the STTS model, we return to the original set of stochastic differential equations (10.18) and express them in integral form, i.e.,

\[ x_1(\Delta) = x_1(0) + \int_0^{\Delta} x_2(\tau_1)d\tau_1 \]  
\[ \vdots \]  
\[ x_{r-1}(\Delta) = x_{r-1}(0) + \int_0^{\Delta} x_r(\tau_{r-1})d\tau_{r-1} \]  
\[ x_r(\Delta) = x_r(0) + \int_0^{\Delta} a(X_r)d\tau + \int_0^{\Delta} b(X_r)dv \tau \]  

If we substitute the \( r \)-th state into the \((r-1)\)-th state and proceed similarly until we reach the first state, then we can rewrite the above system of equations as:

\[ x_1(\Delta) = x_1(0) + x_2(0) \int_0^{\Delta} d\tau_1 + \ldots \]  
\[ + x_{r-1}(0) \int_0^{\Delta} \int_0^{\tau_1} \ldots \int_0^{\tau_{r-3}} d\tau_{r-2} \ldots d\tau_1 \]  
\[ + x_r(0) \int_0^{\Delta} \int_0^{\tau_1} \ldots \int_0^{\tau_{r-2}} d\tau_{r-1} \ldots d\tau_1 \]  
\[ + \int_0^{\Delta} \int_0^{\tau_1} \ldots \int_0^{\tau_{r-1}} a(X_r)d\tau \ldots d\tau_1 \]  
\[ + \int_0^{\Delta} \int_0^{\tau_1} \ldots \int_0^{\tau_{r-1}} b(X_r)dv \tau \ldots d\tau_1 \]  
\[ \vdots \]  
\[ x_{r-1}(\Delta) = x_{r-1}(0) + x_r(0) \int_0^{\Delta} d\tau_{r-1} \]  
\[ + \int_0^{\Delta} \int_0^{\tau_{r-1}} a(X_r)d\tau d\tau_{r-1} \]  
\[ + \int_0^{\Delta} \int_0^{\tau_{r-1}} b(X_r)dv \tau d\tau_{r-1} \]  
\[ x_r(\Delta) = x_r(0) + \int_0^{\Delta} a(X_r)d\tau + \int_0^{\Delta} b(X_r)dv \tau \]
10.4 Defining the Errors of the STTS Model

By solving the multiple time integrals with constant integrand, we obtain

\[x_1(\Delta) = x_1(0) + \ldots + x_{r-1}(0)\frac{\Delta^{r-2}}{(r-2)!} + x_r(0)\frac{\Delta^{r-1}}{(r-1)!}\]

\[+ \int_0^\Delta \int_0^{\tau_1} \int_0^{\tau_{r-1}} a(X_\tau)d\tau_1 \ldots d\tau_1 \]

\[+ \int_0^\Delta \int_0^{\tau_1} \int_0^{\tau_{r-1}} b(X_\tau)d\tau_1 \ldots d\tau_1 \]

\[(10.29a)\]

\[x_r(\Delta) = x_r(0) + \int_0^\Delta a(X_\tau)d\tau + \int_0^\Delta b(X_\tau)dv_\tau \]

\[(10.29c)\]

Note that this representation is still exact. The approximate STTS model presented in Definition 10.3.1 is obtained by replacing \(X_\tau\) by \(X_0\) in the above equations, leading to

\[\hat{x}_1(\Delta) = \hat{x}_1(0) + \ldots + \hat{x}_{r-1}(0)\frac{\Delta^{r-2}}{(r-2)!} + \hat{x}_r(0)\frac{\Delta^{r-1}}{(r-1)!}\]

\[+ a(X_0) \int_0^\Delta \int_0^{\tau_1} \int_0^{\tau_{r-1}} d\tau_1 \ldots d\tau_1 \]

\[+ b(X_0) \int_0^\Delta \int_0^{\tau_1} \int_0^{\tau_{r-1}} dv_\tau \ldots d\tau_1 \]

\[(10.30a)\]

\[\hat{x}_{r-1}(\Delta) = \hat{x}_{r-1}(0) + \hat{x}_r(0)\Delta
\]

\[+ a(X_0) \int_0^\Delta \int_0^{\tau_1} \int_0^{\tau_{r-1}} d\tau_1 \ldots d\tau_1 \]

\[+ b(X_0) \int_0^\Delta \int_0^{\tau_1} \int_0^{\tau_{r-1}} dv_\tau \ldots d\tau_1 \]

\[(10.30b)\]

\[\hat{x}_r(\Delta) = \hat{x}_r(0) + a(X_0) \int_0^\Delta d\tau + b(X_0) \int_0^\Delta dv_\tau \]

\[(10.30c)\]

The expansion of the last state \(\hat{x}_r\) corresponds to the so called Euler-Maruyama approximation [72] of \(x_r\). As an illustration of the error introduced by the approximation, e.g. in \(x_r\), consider the following:

\[x_r(\Delta) = \hat{x}_r(\Delta) + e_r(\Delta) \]

\[(10.31)\]

\[\hat{x}_r(\Delta) = \hat{x}_r(0) + a(X_0) \int_0^\Delta d\tau + b(X_0) \int_0^\Delta dv_\tau \]

\[(10.32)\]
where, assuming \( \hat{x}_r(0) = x_r(0) \), we have

\[
e_r(\Delta) = x_r(\Delta) - \hat{x}_r(\Delta) = \int_0^\Delta \{a(X_\tau) - a(X_0)\} d\tau + \int_0^\Delta \{b(X_\tau) - b(X_0)\} d\tau
\]

which is the one step truncation error incurred by the Euler-Maruyama expansion.

Lemma 10.4.1. The Euler-Maruyama approximation has the following local truncation error properties

\[
|E\{e_r(\Delta)\}| \in \mathcal{O}(\Delta^2) \quad (10.34a)
\]

\[
E\left\{|e_r(\Delta)|^2\right\}^{1/2} \in \mathcal{O}(\Delta) \quad (10.34b)
\]

Proof. See [90, Section 1.6].

Corollary 10.4.1 ([90, Section 1.6]). In the additive noise case the local truncation errors of the Euler-Maruyama approximation are

\[
|E\{e_r(\Delta)\}| \in \mathcal{O}(\Delta^2) \quad (10.35a)
\]

\[
E\left\{|e_r(\Delta)|^2\right\}^{1/2} \in \mathcal{O}(\Delta^{3/2}) \quad (10.35b)
\]

In a similar fashion we can obtain the truncation errors of the full state vector \( x_1, \ldots, x_{r-1} \) as \( e_i = x_i - \hat{x}_i \). It is easily seen that the errors \( e_i \) satisfy

\[
e_i = \begin{cases} 
\int_0^\Delta \int_0^{\tau_{r-2}} \cdots \int_0^{\tau_2} e_r(\tau_{r-1}) d\tau_{r-1} \cdots d\tau_i & i = 1, \ldots, r - 2 \\
\int_0^\Delta e_r(\tau_{r-1}) d\tau_{r-1} & i = r - 1 
\end{cases}
\]

(10.36)

Our aim in the next section is to quantify the errors \( e_i \) for the whole state vector and to analyse their propagation through time.

Remark 10.4.1. The Ito-taylor expansions presented in [72] assume a system with relative degree one and that the derivatives of the functions \( a(\cdot), b(\cdot) \) exist. The STTS model is based on the relative degree concept and, moreover, no derivatives of the functions are needed.

10.5 Truncation Errors of the STTS Model

In this section we present a general result on the Local Vector Truncation Error and the Local Vector Fixed Step Truncation Error for the STTS model:
Theorem 10.5.1. The Local Vector Truncation Errors of the STTS model presented in Definition 10.3.1 are of the order of

(i) \((\Delta^r, \ldots, \Delta)\) in the strong convergence sense,

(ii) \((\Delta^{r+1}, \ldots, \Delta^2)\) in the weak convergence sense.

Proof. Note that Lemma 10.4.1 holds for \(e_r\). The order of the Local Vector Truncation Errors can be translated into proving the following inequalities:

\[ |E\{e_i(\Delta)\}| \in \mathcal{O}(\Delta^{r-i+2}) \quad (10.37) \]

\[ E\left\{ |e_i(\Delta)|^{2} \right\}^{1/2} \in \mathcal{O}(\Delta^{r-i+1}) \quad (10.38) \]

for \(i = 1, \ldots, r - 1\). First, we will prove the truncation errors in the weak sense. It is easily seen that:

\[ |E\{e_i(\Delta)\}| = \left| E\left\{ \int_{0}^{\Delta} \int_{0}^{\tau_i} \cdots \int_{0}^{\tau_{r-2}} e_r(\tau_{r-1}) d\tau_{r-1} \cdots d\tau_{i} \right\} \right| \]

\[ = \int_{0}^{\Delta} \int_{0}^{\tau_i} \cdots \int_{0}^{\tau_{r-2}} E\{e_r(\tau_{r-1})\} d\tau_{r-1} \cdots d\tau_{i} \]

\[ \leq \int_{0}^{\Delta} \int_{0}^{\tau_i} \cdots \int_{0}^{\tau_{r-2}} |E\{e_r(\tau_{r-1})\}| d\tau_{r-1} \cdots d\tau_{i} \]

(10.39)

Considering (10.34a) and the definition of Big-O, there exists a constant \(C_1\) such that

\[ |E\{e_i(\Delta)\}| \leq \int_{0}^{\Delta} \int_{0}^{\tau_i} \cdots \int_{0}^{\tau_{r-2}} C_1 \cdot \tau_{i}^{2+(r-2)-i+1} d\tau_{i} \]

\[ \leq C_1 \cdot \Delta^{r-i+2} \quad (10.40) \]

On the other hand, for the truncation errors in the strong convergence sense, by repeatedly applying the H"older inequality [72, eq. 1.4.40, 4.5.15], we have that

\[ E\left\{ |e_i(\Delta)|^{2} \right\} = E\left\{ \int_{0}^{\Delta} \int_{0}^{\tau_i} \cdots \int_{0}^{\tau_{r-2}} e_r(\tau_{r-1}) d\tau_{r-1} \cdots d\tau_{i} \right\}^{2} \]

\[ \leq \Delta \cdot \int_{0}^{\Delta} E\left\{ \int_{0}^{\tau_i} \cdots \int_{0}^{\tau_{r-2}} e_r(\tau_{r-1}) d\tau_{r-1} \cdots d\tau_{i-1} \right\}^{2} d\tau_{i} \]

\[ \leq \Delta \cdot \int_{0}^{\Delta} \tau_{i} \cdot \int_{0}^{\tau_i} E\left\{ \int_{0}^{\tau_{r-1}} \cdots \int_{0}^{\tau_{r-2}} e_r(\tau_{r-1}) d\tau_{r-1} \cdots d\tau_{2} \right\}^{2} d\tau_{r-1} d\tau_{i} \]

\[ \leq \Delta \cdot \int_{0}^{\Delta} \tau_{i} \cdot \int_{0}^{\tau_{i}} \cdots \int_{0}^{\tau_{r-2}} E\{e_r(\tau_{r-1})\} \left\{ \int_{0}^{\tau_{r-1}} \cdots \int_{0}^{\tau_{r-2}} d\tau_{r-1} \cdots d\tau_{2} \right\}^{2} d\tau_{r-1} \cdots d\tau_{i} \]

(10.41)
Considering (10.34b) and the definition of Big-$\mathcal{O}$, there exists a constant $C_2$ such that

$$E \left\{ |e_i(\Delta)|^2 \right\} \leq \Delta \cdot \int_0^\Delta C_2 \cdot \tau_i^{2+2(r-i+1)} d\tau_i$$

$$\leq C_2 \cdot \Delta^{2+2(r-i+1)+2}$$

$$\leq C_2 \cdot \Delta^{2(r-i+1)}$$

(10.42)

The result is then obtained by taking the square root.

**Corollary 10.5.1.** In the special case of additive noise, i.e. $b(X) = c$, then the Local Vector Truncation Errors of the STTS model are of the order of

(i) $(\Delta^{(2r+1)/2}, \ldots, \Delta^{3/2})$ in the strong convergence sense,

(ii) $(\Delta^{r+1}, \ldots, \Delta^2)$ in the weak convergence sense.

**Proof.** The result follows by using inequalities (10.35) instead of (10.34) in the proof of Theorem 10.5.1.

**Theorem 10.5.2.** The Local Vector Fixed Step Truncation Errors of the STTS model presented in Definition 10.3.1 are of the order of

(i) $(\Delta^r, \ldots, \Delta)$ in the strong convergence sense,

(ii) $(\Delta^{r+1}, \ldots, \Delta^2)$ in the weak convergence sense.

**Proof.** Recall Lemma 10.4.1. We adopt the alternative view proposed in [12] pp. 167-168, where the local errors are referred to the exact solution rather than to the approximate solution. The alternative view proposed in [12] allows us to consider the propagation of errors through a finite number of steps. In particular, it allows us to define

$$F^s_k = \int_{(k-1)\Delta}^{k\Delta} \{a(X_s) - a(X_k)\} ds + \int_{(k-1)\Delta}^{k\Delta} \{b(X_s) - b(X_k)\} dv_s$$  \hspace{1cm} (10.43)

$$F^r_i = \int_{(k-1)\Delta}^{k\Delta} \cdots \int_{(k-1)\Delta}^{(k-1)\Delta} \int_{(k-1)\Delta}^{(k-1)\Delta} F^r_{\tau_i} d\tau_1 \cdots d\tau_{r-1}, \quad i = 1, \ldots, r-1$$  \hspace{1cm} (10.44)
where $F_{\tau_1}'$ in (10.44) is $F_k'$ as defined in (10.43) but with the upper limit of the integrals replaced by $\tau_1$. Considering (10.34), we notice that

$$E\{|F_{i[k]}|\} \in O(\Delta^{r-i+1}), \quad i = 1, \ldots, r \quad (10.45)$$

$$|E\{F_{i[k]}\}| \in O(\Delta^{r-i+2}), \quad i = 1, \ldots, r \quad (10.46)$$

In addition, define

$$e_i[k] = x_i(k\Delta) - \hat{x}_i(k\Delta), \quad i = 1, \ldots, r \quad (10.47)$$

Then, from (10.29) and (10.30), we can write

$$E\{|\bar{e}_k^{N}|\} \leq A \cdot |E\{e[k]\}| + E\{|\bar{F}\|$$

$$|E\{e[k+1]\}| \leq A \cdot |E\{e[k]\}| + |E\{\bar{F}\}|$$

where

$$\bar{e}[k] = \begin{bmatrix} e_1[k] & e_2[k] & \cdots & e_r[k] \end{bmatrix}^T$$

$$A = \begin{bmatrix} 1 & \Delta & \cdots & \frac{\Delta^{r-1}}{(r-1)!} \\ 0 & 1 & \cdots & \frac{\Delta^{r-2}}{(r-2)!} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\bar{F} = \begin{bmatrix} F_1' \ F_2' \ \cdots \ F_r' \end{bmatrix}^T$$

Following a similar argument to that used in [25, Theorem 2], and since both (10.48) and (10.49) have the same structure, we have that

$$E\{|e[k+N]|\} \leq A^N \cdot |E\{e[k]\}| + \sum_{j=0}^{N-1} A^j \cdot E\{|\bar{F}\|$$

$$|E\{e[k+N]\}| \leq A^N \cdot |E\{e[k]\}| + \sum_{j=0}^{N-1} A^j \cdot |E\{\bar{F}\}|$$

where

$$A^N = \begin{bmatrix} 1 & N\Delta & \frac{N^2\Delta^2}{2} & \cdots & \frac{N^{r-2}\Delta^{r-2}}{(r-2)!} & \frac{N^{r-1}\Delta^{r-1}}{(r-1)!} \\ 0 & 1 & N\Delta & \cdots & \frac{N^{r-3}\Delta^{r-3}}{(r-3)!} & \frac{N^{r-2}\Delta^{r-2}}{(r-2)!} \\ 0 & 0 & 1 & \cdots & \frac{N^{r-4}\Delta^{r-4}}{(r-4)!} & \frac{N^{r-3}\Delta^{r-3}}{(r-3)!} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & N\Delta \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

(10.55)
\[ \sum_{j=0}^{N-1} A^j = \begin{bmatrix} N & \frac{N(N-1)}{2} & \cdots & \frac{S_{r-4} \Delta^{r-4}}{(r-4)!} & \frac{S_{r-3} \Delta^{r-3}}{(r-3)!} \\ 0 & N & \cdots & \frac{S_{r-2} \Delta^{r-2}}{(r-2)!} & \frac{S_{r-1} \Delta^{r-1}}{(r-1)!} \\ & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & N & \frac{N(N-1)}{2} \Delta \\ 0 & 0 & \cdots & 0 & N \end{bmatrix} \] (10.56)

where

\[ S_{r}^{N-1} = \frac{N^{r+1}}{r+1} + \sum_{k=1}^{r} \frac{B_k}{r-k+1} \binom{r}{k} N^{r-k+1} \] (10.57)

and where \( B_k \) is the \( k \)-th Bernoulli number \([29, \text{Section 1.14}]\).

Therefore, assuming that

\[ E[|e_i[k]|] \in O(\Delta^{r-i+1}), \quad i = 1, \ldots, r \] (10.58)

\[ |E[e_i[k]]| \in O(\Delta^{r-i+2}), \quad i = 1, \ldots, r \] (10.59)

and if \( N \) is a fixed number (not depending on \( \Delta \)), then

\[ [A^N]_i \cdot E[|\bar{e}[k]|] \in O(\Delta^{r-i+1}) \] (10.60)

\[ [A^N]_i \cdot |E[\bar{e}[k]]| \in O(\Delta^{r-i+2}) \] (10.61)

where \([A^N]_i\) denotes the \( i \)-th row of \( A^N \). The same argument applies to

\[ \sum_{j=0}^{N-1} A^j \cdot E[|\bar{e}[k]|] \in O(\Delta^{r-i+1}) \] (10.62)

\[ \sum_{j=0}^{N-1} A^j \cdot |E[\bar{e}[k]]| \in O(\Delta^{r-i+2}) \] (10.63)

which, in turn, implies

\[ E[|e_i[k+N]|] \in O(\Delta^{r-i+1}), \quad i = 1, \ldots, r \] (10.64)

\[ |E[e_i[k+N]]| \in O(\Delta^{r-i+2}), \quad i = 1, \ldots, r \] (10.65)

which completes the proof.

\[ \blacksquare \]

**Corollary 10.5.2.** In the special case of additive noise, i.e. \( b(X) = c \), then the Local Vector Fixed Step Truncation Errors of the STTS model are of the order of
(i) \((\Delta^{(2r+1)/2}, \ldots, \Delta^{3/2})\) in the strong convergence sense,

(ii) \((\Delta^{r+1}, \ldots, \Delta^2)\) in the weak convergence sense.

Proof. The result follows by considering inequalities (10.35) instead of (10.34) in the proof of Theorem 10.5.2. 

\[\text{\[113x672\] 10.6 Example I: Truncation Errors} \]

In this section we present a numerical example that substantiates the results presented in the previous section. Consider the following second order nonlinear system

\[\begin{align*}
   dx_1 &= x_2 dt \\
   dx_2 &= (-2x_2 - \sqrt{|x_1|}) dt + (1 + \text{sat}(x_1)) \cdot dv
\end{align*}\]  

(10.66a) (10.66b)

The corresponding STTS sampled-data model is

\[\begin{align*}
   \hat{x}_1^+ &= \hat{x}_1 + \Delta \cdot \hat{x}_2 + \frac{\Delta^2}{2} \cdot (-2\hat{x}_2 - \sqrt{\hat{x}_1}) + (1 + \text{sat}(\hat{x}_1)) \cdot \tilde{v}_1 \\
   \hat{x}_2^+ &= \hat{x}_2 + \Delta \cdot (-2\hat{x}_2 - \sqrt{\hat{x}_1}) + (1 + \text{sat}(\hat{x}_1)) \cdot \tilde{v}_2
\end{align*}\]  

(10.67a) (10.67b)

with initial states \(x_1(0) = \hat{x}_1(0) = 1\) and \(x_2(0) = \hat{x}_2(0) = 1\). An exact sampled-data model for system (10.66) is not directly available. In order to accurately approximate the errors we simulated an upsampled version of the Euler-Maruyama approximation:

\[\begin{align*}
   \bar{x}_1^+ &= \bar{x}_1 + \Delta_u \cdot \bar{x}_2 \\
   \bar{x}_2^+ &= \bar{x}_2 + \Delta_u \cdot (-2\bar{x}_2 - \sqrt{\bar{x}_1}) + (1 + \text{sat}(\bar{x}_1)) \cdot \tilde{v}_2^u
\end{align*}\]  

(10.68a) (10.68b)

where \(\Delta_u = \Delta/N_u, \ N_u = 10000, \ \bar{x}_1(0) = \bar{x}_2(0) = 1\) and \(\tilde{v}_2^u \sim N(0, \sqrt{\Delta_u})\). The output of the upsampled system is then sampled every \(N_u\) samples.

We analyse the behaviour of the truncation errors as the sampling period \(\Delta\) is varied as \(\Delta = T/2^n\), with \(n = 3, \ldots, 10\) and \(T = 1[s]\). The errors are averaged over \(N_{sim} = 1000\) runs for each sampling period.

Figure 10.1 and Table 10.1 show the Local Vector Truncation Errors in the strong sense, for both states of the system, for different sampling times. Since, for every new simulation batch, the sampling period is halved, we compute the ratio of the averaged errors for two subsequent sampling periods. The columns \(\log_2(\text{ratio})\) in Tables 10.1 and 10.2 represent the power of two by which the errors decrease, i.e. the order of the errors for the state \(x_i\).
Figure 10.1: Local Vector Truncation Errors for $x_1$ and $x_2$

Table 10.1: Local Vector Truncation Errors (Strong)

| $\Delta$ | $E[|e_1(1)|]$ | $E[|e_2(1)|]$ | $\log_2(ratio_1)$ | $\log_2(ratio_2)$ |
|----------|----------------|----------------|-------------------|-------------------|
| $2^{-3}$ | 5.2198e-03     | 1.1317e-01     | -                 | -                 |
| $2^{-4}$ | 9.1957e-04     | 3.8937e-02     | 2.5049            | 1.5392            |
| $2^{-5}$ | 1.5637e-04     | 1.3075e-02     | 2.5559            | 1.5742            |
| $2^{-6}$ | 2.8366e-05     | 4.6921e-03     | 2.4627            | 1.4785            |
| $2^{-7}$ | 5.0228e-06     | 1.6367e-03     | 2.4975            | 1.5194            |
| $2^{-8}$ | 8.5098e-07     | 5.4741e-04     | 2.5613            | 1.5801            |
| $2^{-9}$ | 1.5422e-07     | 1.9512e-04     | 2.4641            | 1.4882            |

Figure 10.2 and Table 10.2 show the Local Vector Fixed Step Truncation Errors. We chose $N = 7$ for all simulations. It can be seen that, for a sufficiently small sampling period, the accumulated errors, in both states, preserve the behaviour of the local truncation error. This result is as predicted by Theorem 10.5.2.

**Remark 10.6.1.** Note that, since $X_0 = \hat{X}_0 = 1$ is not a random variable but a known constant, the model effectively behaves like an additive noise system for the purposes of Local and Local Fixed Step errors.
10.7 Stochastic Sampling Zero Dynamics

In this section we will show that a slightly modified form of the STTS model presented in Definition 10.3.1 accounts for the asymptotic location of the sampling zeros. First, we confirm that the STTS model has relative degree 1, and hence, contains extra zero dynamics that have no continuous-time counterpart. We first recall the following definition of relative degree:

Definition 10.7.1 ([6]). A discrete-time system has relative degree \( r \) if

(i) \[ \frac{\partial y_k}{\partial u_k} \bigg|_{(x_k, u_k)} = 0, \quad \forall i = 0, \ldots, r - 1 \]

(ii) \[ \frac{\partial y_k}{\partial u_k} \bigg|_{(x_k, u_k)} \neq 0 \]

where \( y_k \) is the output and \( u_k \) is the sampled exogenous input at time \( k\Delta \).
Theorem 10.7.1. The STTS sampled-data model described in Definition 10.3.1 generically has relative degree 1.

Proof. Recall that the output of the STTS model is given by
\[ y_k = \hat{x}_{1,k}, \]
where
\[ \hat{x}_{1,k+1} = x_{1,k} + x_{2,k} \cdot \Delta + \ldots + x_{r,k} \cdot \Delta^{r-1} \left( \frac{\Delta}{r-1}! \right) + a(X_k) \cdot \frac{\Delta^r}{r!} + b(X_k) \cdot \tilde{v}_{1,k} \] (10.69)

Although the inputs \( \tilde{v}_i \) are all different, in order to define the relative degree of the discrete-time system it suffices to show that all states (particularly \( \hat{x}_1 \)) share a common input. This is done by considering
\[ \tilde{v}_{\ell,k} = \int_{k\Delta}^{k\Delta + \Delta} \frac{(\Delta - \tau)^{(n-\ell)}}{(n-\ell)!} \, dv_{\ell}, \ell = 1, \ldots, n \]
(10.70)

Hence, we can consider \( u_k = \tilde{v}_{n,k} \) in the partial derivatives used in Definition 10.7.1. It is then noticed that:
\[ \frac{\partial y_{k+0}}{\partial \tilde{v}_{n,k}} = 0 \] (10.71)
\[ \frac{\partial y_{k+1}}{\partial \tilde{v}_{n,k}} \neq 0 \] (10.72)

In addition, following the same argument, it can be shown that if the output is chosen as any other state, it would also have relative degree 1, with respect to the input \( u_k = \tilde{v}_{n,k} \).

The following two lemmas are preliminary results that will be used later to reveal the sampling zero dynamics in the STTS model.

Lemma 10.7.1. Define
\[ F_i(q) = (q - 1)^{n-i} \frac{\Delta^{i-1}}{(i-1)!} B_{i-1}(q) \] (10.73)
where \( B_i(q) \) is the \( i \)-th Euler-Frobenius polynomial, with the convention that \( B_0(q) = 1, B_1(q) = 1 \), then the following equality holds
\[ \sum_{i=1}^{n} \sum_{j=1}^{n} F_i(z) F_j(z^{-1}) \sigma_{ij} = \frac{\Delta^{2n-1}}{(2n-1)!} \cdot z^{1-n} \cdot B_{2n-1}(z) \] (10.74)
where
\[
\sigma_{ij} = E\{\tilde{v}_i\tilde{v}_j\} = \frac{\Delta^{2n-i-j+1}}{(n-1)!(n-j)!(2n-i-j+1)!} \tag{10.75}
\]
are the elements of the covariance matrix of \(\tilde{V}_k\) in Definition \textit{10.3.1}.

**Proof.** Consider a \(n\)-th order integrator

\[
\begin{align*}
dx_1 &= x_2 dt \quad (10.76a) \\
\vdots \\
dx_{n-1} &= x_n dt \quad (10.76b) \\
dx_n &= dv \quad (10.76c)
\end{align*}
\]

where \(v(t)\) is a Wiener process of unitary incremental variance. In [119] it is shown that the output spectrum of a sampled \(n\)-th order stochastic integrator, for finite sampling period \(\Delta\), is given by

\[
\Phi_T = \frac{\Delta}{2\pi} \cdot \frac{\Delta^{2n-1}}{(2n-1)!} \cdot \frac{z^{1-n} B_{2n-1}(z)}{(z-1)^n(z^{-1}-1)^{n}} \tag{10.77}
\]

On the other hand, from [125], an exact discretisation of an \(n\)-th order stochastic integrator is given by

\[
X_{k+1} = A_q \cdot X_k + \tilde{V}_k \tag{10.78}
\]

where \(y = x_1\) and

\[
X_k = \left[ x_{1,k} \quad \cdots \quad x_{n,k} \right]^T \tag{10.79}
\]

\[
A_q = \begin{bmatrix}
1 & \Delta & \cdots & \Delta^{n-1} \\
0 & 1 & \cdots & \Delta^{n-2} \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & 1
\end{bmatrix} \tag{10.80}
\]

\[
\tilde{V}_k = \left[ \tilde{v}_{1,k} \quad \cdots \quad \tilde{v}_{n,k} \right]^T \tag{10.81}
\]

\[
\tilde{v}_\ell = \int_{k\Delta}^{(k+1)\Delta} \frac{dx_\ell}{(n-\ell)!} , \ell = 1, \ldots, n \tag{10.82}
\]
10. A Sampled-Data Model for Stochastic Nonlinear Systems

Hence the output is given path-wise exactly at the sampling instants by

\[ y_k = [1 \ldots 0] \cdot (qI - Aq)^{-1} \cdot \tilde{V}_k \]

\[ = [1 \ldots 0] \cdot \frac{1}{(q - 1)^n} \times \begin{bmatrix}
(q - 1)^{n-1} & \Delta(q - 1)^{n-2} & \frac{\Delta^2}{2} (q - 1)^{n-3} (q + 1) & \ldots \\
0 & (q - 1)^{n-1} & \Delta(q - 1)^{n-2} & \frac{\Delta^2}{2} (q - 1)^{n-3} (q + 1) & \ldots \\
0 & 0 & (q - 1)^{n-1} & \Delta(q - 1)^{n-2} & \ldots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\end{bmatrix} \times \begin{bmatrix}
\tilde{v}_{1,k} \\
\tilde{v}_{2,k} \\
\tilde{v}_{3,k} \\
\tilde{v}_{4,k} \\
\vdots \\
\end{bmatrix}^T \\
= \frac{1}{(q - 1)^n} \sum_{i=1}^{n} F_i(q) \cdot \tilde{v}_{i,k} \quad (10.83) \]

Then, the output spectrum is given by

\[ \Phi_y = \frac{\Delta}{2\pi} \cdot \frac{1}{(z - 1)^n(z^{-1} - 1)^n} \sum_{i=1}^{n} \sum_{j=1}^{n} F_i(z) F_j(z^{-1}) \sigma_{ij} \quad (10.84) \]

where the factor \( \Delta/2\pi \) comes from the definition of discrete-time power spectrum given in [119, Eq. 15].

Since (10.77) and (10.84) are both exact representations of the sampled spectrum of a \( n \)-th order stochastic integrator, it follows that \( \Phi_T = \Phi_y \) and therefore

\[ \sum_{i=1}^{n} \sum_{j=1}^{n} F_i(z) F_j(z^{-1}) \sigma_{ij} = \frac{\Delta^{2n-1}}{(2n - 1)!} \cdot z^{1-n} \cdot B_{2n-1}(z) \quad (10.85) \]

This completes the proof.

\[ \Box \]

Lemma 10.7.2. The polynomial \( z^{1-n} B_{2n-1}(z) \) admits a spectral factorisation, i.e.

\[ z^{1-n} B_{2n-1}(z) = K \cdot N_e(z) \cdot N_e(z^{-1}) \quad (10.86) \]

where \( N_e(z) = \prod_{i=1}^{n-1} (z + a_i) \) has all its roots \( z = -a_i \) inside the unit circle, and \( K = \prod_{i=1}^{n-1} (1/a_i) \).

Proof. The Euler-Frobenius polynomial \( B_{2n-1} \) is, by definition, a polynomial of order \( 2n - 2 \). In addition, \( z^{1-n} = z^{-(2n-2)/2} \) is always a power of \( z \) that is half of the highest power of \( z \) in \( B_{2n-1} \). Also, since \( B_n(z) \) is a symmetrical polynomial for all \( n \), it follows that

\[ z^{1-n} B_{2n-1}(z) = z^{-(1-n)} B_{2n-1}(z^{-1}) \quad (10.87) \]

Also, it is known that \( B_{2n-1}(1) = (2n - 1)! > 0 \), [126, Appendix A]. Therefore a spectral factorisation is always possible.
On the other hand, it is known that the Euler-Frobenius polynomials $B_n$ have simple real roots \[113\], and if $z = z_0$ is a root, with $|z_0| < 1$, then $1/z_0$ is also a root. Therefore, it follows that

$$z^{1-n}B_{2n-1}(z) = z^{1-n}\prod_{i=1}^{n-1}(z + a_i)(z + 1/a_i)$$

(10.88)

$$= z^{1-n}\prod_{i=1}^{n-1}(z + a_i)(a_i z + 1)/a_i$$

(10.89)

$$= \prod_{i=1}^{n-1}(z + a_i)(a_i + z^{-1})/a_i$$

(10.90)

$$= \prod_{i=1}^{n-1}(1/a_i)\prod_{i=1}^{n-1}(z + a_i)(z^{-1} + a_i)$$

(10.91)

with $|a_i| < 1$, which completes the proof. See \[126\] Appendix A for a summary of the properties of Euler-Frobenius polynomials.

The next step is to write the STTS model, presented in Definition \[10.3.1\] as an input-output model. It can be shown that the STTS model is equivalent to the following model:

$$(q - 1)^n\cdot \hat{y}_k = \frac{\Delta^n}{n!}\cdot G_0(q)\cdot a(\hat{X}_k) + \sum_{i=1}^{n} F_i(q)\cdot \tilde{v}_{i,k}$$

(10.92)

where $G_i(q)$ is a polynomial of order $i - 1$ and $F_i(q)$ is as defined in Lemma \[10.7.1\].

We then have the following property of the collapsed model.

**Lemma 10.7.3.** The model \[10.92\] has the following truncation error property:

$$E|\,(q - 1)^n\cdot y_k - (q - 1)^n\cdot \hat{y}_k| \in \mathcal{O}(\Delta^n)$$

(10.93)

where $n$ is the relative degree of the system.

**Proof.** Define $e_{1,k} = x_{1,k} - \hat{x}_{1,k} = y_k - \hat{y}_k$. Lemma \[10.7.3\] can then be translated to

$$E|e_{1,k+i}| \in \mathcal{O}(\Delta^n), \quad i = 1, \ldots, n$$

(10.94)

which is ensured by the order of the Local Fixed Step Truncation Error for the state $x_1$. ■

It is now possible to perform a further approximation which preserves the property presented in Lemma \[10.7.3\].

**Definition 10.7.2.** The Modified STTS model (MSTTS) for the output is:

$$(q - 1)^n\cdot \bar{y}_k = \frac{\Delta^n}{n!}\cdot G_0(q)\cdot a(\hat{X}_k) + b(\hat{X}_k)\cdot \sum_{i=1}^{n} F_i(q)\cdot \tilde{v}_{i,k}$$

(10.95)
Remark 10.7.1. Note that $G_n(q)$ and $F_i(q)$ are operators on $a(\hat{X}_k)$ and $b(\hat{X}_k)\tilde{v}_{i,k}$ respectively, hence

$$F_i(q)b(\hat{X}_k)\tilde{v}_{i,k} \neq b(\hat{X}_k)F_i(q)\tilde{v}_{i,k}$$  \hspace{1cm} (10.96)

However, the MSTTS model \[10.95\] is an approximation of \[10.92\] that preserves the order of the truncation errors, as we will see below in Lemma 10.7.4. It is this approximation which allows us to combine the noise vector $\tilde{v}_1, \ldots, \tilde{v}_n$ into a single noise. Indeed, it is precisely this step which has previously precluded the elucidation of the stochastic sampling zero dynamics. □

Lemma 10.7.4. The MSTTS model has the following local truncation error property:

$$E \left| (q - 1)^n \cdot y_k - (q - 1)^n \cdot \tilde{y}_k \right| \in O(\Delta^n)$$  \hspace{1cm} (10.97)

Proof. By the triangle inequality, we have

$$E \left| (q - 1)^n \cdot y_k - (q - 1)^n \cdot \tilde{y}_k \right| \leq E \left| (q - 1)^n \cdot y_k - (q - 1)^n \cdot \hat{y}_k \right| + E \left| (q - 1)^n \cdot \hat{y}_k - (q - 1)^n \cdot \tilde{y}_k \right|$$  \hspace{1cm} (10.98)

Using Lemma 10.7.3 we are only left to prove

$$E \left| (q - 1)^n \cdot \hat{y}_k - (q - 1)^n \cdot \tilde{y}_k \right| \in O(\Delta^m)$$  \hspace{1cm} (10.99)

where $m > n$. Define $e_{\hat{y}, n} = (q - 1)^n \cdot \hat{y}_k - (q - 1)^n \cdot \tilde{y}_k$, then

$$e_{\hat{y}, n} = \sum_{i=1}^{n} F_i(q) \cdot b(\hat{X}_k) \cdot \tilde{v}_{i,k} - b(\hat{X}_k) \cdot \tilde{v}_{i,k} \hspace{1cm} (10.100)$$

By determining the expressions for $E |e_{\hat{y}, n}|^2$, $n = 1, 2, 3, \ldots$, and since

$$E |q^n\tilde{v}_{i,k}|^2 = E |q\tilde{v}_{i,k}|^2$$  \hspace{1cm} (10.101)

$$E \left| q^n b(\hat{X}_k) - b(\hat{X}_k) \right|^2 \in O(E \left| q b(\hat{X}_k) - b(\hat{X}_k) \right|^2)$$  \hspace{1cm} (10.102)

it can be seen that it suffices to prove that

$$E |e_{\hat{y}, n}|^2 \leq \sum_{i=1}^{n} \Delta^{2(i-1)} \cdot E \left| q b(\hat{X}_k) - b(\hat{X}_k) \right|^2 \cdot E |q\tilde{v}_{i,k}|^2 \in O(\Delta^{2m})$$  \hspace{1cm} (10.103)

It is known \[125\] Eq. 41 that

$$E |q\tilde{v}_{i,k}|^2 = E |\tilde{v}_{i,k}|^2 \in O(\Delta^{2(n-i)+1})$$  \hspace{1cm} (10.104)

and that \[50\] Eq. 1.41

$$\left| q^n b(\hat{X}_k) - b(\hat{X}_k) \right| \in O(\Delta), \hspace{1cm} n = 1, 2, 3 \ldots$$  \hspace{1cm} (10.105)
Hence
\[ E |q^n b(\hat{X}_k) - b(\hat{X}_k)|^2 \in O(\Delta^2) \] (10.106)
and therefore
\[ E |\bar{e}_{y,n}|^2 \in O(\Delta^{2(i-1)+2(n-i)+1}) \] (10.107)
which corresponds to
\[ E |\bar{e}_{y,n}|^2 \in O(\Delta^{2n+1}) \] (10.108)
and therefore \( m = n + 1/2 > n \), which proves the result.

\[ \blacksquare \]

**Remark 10.7.2.** For the special case \( n = 3 \), \( E |\bar{e}_{y,3}|^2 \) turns out to be
\[ E |\bar{e}_{y,3}|^2 \leq E |q^n b(\hat{X}_k) - b(\hat{X}_k)|^2 \]
\[ + \Delta^2 E |q^n b(\hat{X}_k) - b(\hat{X}_k)|^2 \]
\[ + \Delta^4 E |q^n b(\hat{X}_k) - b(\hat{X}_k)|^2 \]
\[ + \frac{\Delta^4}{4} E |q^n b(\hat{X}_k) - b(\hat{X}_k)|^2 E |\tilde{v}_{3,k}|^2 \] (10.109)
which complies with all the assumptions of the previous proof.
\[ \blacksquare \]

The above results provide the basis of the following key result:

**Theorem 10.7.2.** The MSTTS model (10.95) is path-wise equivalent to the following model
\[ (q - 1)^n \cdot \bar{y}_k = \frac{\Delta^n}{n!} \cdot G_n(q) \cdot a(\hat{X}_k) + b(\hat{X}_k) \cdot N_e(q) \cdot w_k \] (10.110)
where \( N_e(q) \) is defined as in Lemma 10.7.2, i.e. the stable spectral factor of \( q^{1-n} B_{2n-1}(q) \), and \( w_k \) is a discrete-time white noise, with \( E |w_k|^2 = L \), with \( L = K \cdot \Delta^n/(2n - 1)! \) and \( K \) defined as in Lemma 10.7.2.

The process \( w_k \) is given pathwise by the innovations model [65, Section 8.3.5]:
\[ z_{k+1} = (A - K_p C) z_k - K_p \phi_k \] (10.111)
\[ w_k = C z_k - \phi_k \] (10.112)
where
\[ \phi_k = q^{-n} \cdot \sum_{i=1}^{n} F^n_i(q) \cdot \tilde{v}_{i,k} \] (10.113)
\[ K_p = (APCT + BS) \cdot R^{-1} \] (10.114)
\[ R_e = R + CPC^T \] (10.115)
and $P$ is the unique positive semi-definite solution of the DARE

$$P = APA^T + BQB^T - K_P R_P K_P.$$  \hspace{1cm} (10.116)

**Proof.** We have already shown, through Lemmas [10.7.1] and [10.7.2] that the spectral factorisation problem has a solution, and hence we can write

$$\phi_{k+n} = \sum_{i=1}^{n} F^n_i(q) \cdot \tilde{v}_{i,k} = N_c(q) \cdot w_k$$  \hspace{1cm} (10.117)

where $E\{|w_k^2|\} = L$. It is only left to prove that $w_k$ can be defined pathwise through inverting a specific innovations model. In general

$$\phi_{k+n} = \sum_{i=1}^{n} F^n_i(q) \cdot \tilde{v}_{i,k}$$  \hspace{1cm} (10.118)

can be written in state space form as

$$x_{k+1} = A x_k + B u_k$$  \hspace{1cm} (10.119a)

$$\phi_k = C x_k + \tilde{v}_{1,k}$$  \hspace{1cm} (10.119b)

where $A \in \mathbb{R}^{(n-1)\times(n-1)}$, $B \in \mathbb{R}^{(n-1)\times n}$ and $C \in \mathbb{R}^{1 \times (n-1)}$ are given by

$$A = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \ddots & \ddots \\ & & & 1 & 0 \end{bmatrix}$$  \hspace{1cm} (10.120)

$$B = \begin{bmatrix} [F^n_1]_1 & [F^n_2]_1 & \cdots & [F^n_n]_1 \\ [F^n_1]_2 & [F^n_2]_2 & \cdots & [F^n_n]_2 \\ \vdots & \vdots & \ddots & \vdots \\ [F^n_1]_{n-1} & [F^n_2]_{n-1} & \cdots & [F^n_n]_{n-1} \end{bmatrix}$$  \hspace{1cm} (10.121)

$$C = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}$$  \hspace{1cm} (10.122)

$$u_k = \begin{bmatrix} \tilde{v}_{1,k} & \tilde{v}_{2,k} & \cdots & \tilde{v}_{n,k} \end{bmatrix}^T$$  \hspace{1cm} (10.123)

where $[F^n_i]_j$ denotes the coefficient of $F^n_i(q)$ that multiplies $q^j$; and where

$$E\{[u_k \, \tilde{v}_{1,k}] [u_k \, \tilde{v}_{1,k}]^T\} = \begin{bmatrix} [Q] & [S] \\ [S]^T & R \end{bmatrix}$$  \hspace{1cm} (10.124)
is given element-wise by
\[ E\{\tilde{v}_{i,k}\tilde{v}_{j,k}\} = \frac{\Delta^{2n-i-j+1}}{(n-1)!(n-j)!(2n-i-j+1)!} \] (10.125)

Then, the innovations \( w_k \) of the process \( \phi_k \), associated with model (10.119) are given by [65, Section 8.3.5]

\[ z_{k+1} = (A - K_pC)z_k - K_p\phi_k \] (10.126)
\[ w_k = Cz_k - \phi_k \] (10.127)

with \( E|w_k|^2 = R_e \), and which requires to compute the following matrices

\[ K_p = (APC^T + BS) \cdot R_e^{-1} \] (10.128)
\[ R_e = R + CPC^T \] (10.129)
\[ P = APA^T + BQB^T - K_pR_eK_p \] (10.130)

which turn out to be the steady state equations of the Kalman filter associated with model (10.119) that solve the spectral factorisation problem. It is clear then that \( R_e = L \).

**Corollary 10.7.1.** In the special case of additive noise, i.e. \( b(X) = c \in \mathbb{R} \), the result presented in Theorem 10.7.2 follows immediately from the STTS model and no extra approximation is needed.

**Remark 10.7.3.** Theorem 10.7.2 proves the existence of extra zero dynamics that have no continuous-time counterpart. Furthermore, just as in the linear and nonlinear deterministic case and as in the linear stochastic case, they also depend on the Euler-Frobenius polynomials.

### 10.8 Example II: Application to Nonlinear Parameter Estimation

In this section we present a specific application of the MSTTS model in the identification of the parameters in a continuous-time stochastic nonlinear system. We show that the existence of the sampling zero polynomials plays a crucial role in obtaining unbiased estimates of the continuous time model parameters. Consider the system

\[ dx_1 = x_2dt \] (10.131a)
\[ dx_2 = a(x_1, x_2)dt + b(x_1, x_2) \cdot dv \] (10.131b)
where \( y = x_1, \) \( v(t) \) is a Wiener process of unitary incremental variance and the functions \( a(\cdot), b(\cdot) \) are given by

\[
a(x_1, x_2) = -\theta_1 \sqrt{|x_1|} - \theta_2 x_2 \tag{10.132}
\]

\[
b(x_1, x_2) = 1 + \text{sat}(x_1) \tag{10.133}
\]

The function \( \text{sat}(y) \) is given by

\[
\text{sat}(y) = \begin{cases} 
0.3 & y > 0.3 \\
\frac{y}{-0.3} & -0.3 < y < 0.3 \\
-0.3 & y < -0.3 
\end{cases} \tag{10.134}
\]

The parameter vector, is chosen to be

\[
\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \tag{10.135}
\]

Note that the continuous-time system (10.131) has relative degree two. Hence, the MSTTS model has additional zero dynamics which correspond to the root inside the unit circle of the polynomial

\[
B_3(z) = z^2 + 4z + 1. \tag{10.136}
\]

Since the system (10.131) does not have an exact discretisation we have used an upsampled Euler-Maruyama approximation to simulate the continuous-time system:

\[
x_1^{+} = x_1 + \Delta_u \cdot x_2 \tag{10.137}
\]

\[
x_2^{+} = x_2 + \Delta_u \cdot a(x_1, x_2) + b(x_1, x_2) \cdot \tilde{v}_2 \tag{10.138}
\]

where \( \tilde{v}_2 \sim N(0, \sqrt{\Delta_u}) \) are the increments of the driving Wiener process, \( \Delta_u = \Delta/N_u \) is the sampling period of the upsampling process, and \( N_u = 100 \). The obtained data is then sampled every \( N_u \) samples.

We perform \( N_{\text{sim}} = 250 \) simulations, using \( N = 10000 \) data points in each run. We first use least squares to fit the parameters in a Euler-Maruyama approximate model of the form:

\[
\frac{(q-1)^2}{\Delta^2} y_k + \theta_2 \frac{(q-1)}{\Delta} y_k + \theta_1 \sqrt{|y_k|} = (1 + \text{sat}(y_k)) \cdot \tilde{v}_k \tag{10.139}
\]

where \( \tilde{v}_k \sim N(0, \sqrt{\Delta}) \) are the increments of the driving Wiener process, i.e. discrete-time white noise.

We form the equation error model

\[
\frac{(q-1)^2}{1 + \text{sat}(y_k)} y_k + \theta_2 \frac{(q-1)}{1 + \text{sat}(y_k)} y_k + \theta_1 \frac{\sqrt{|y_k|}}{1 + \text{sat}(y_k)} = \epsilon_{k+2} \tag{10.140}
\]
This leads to estimates having the following mean values, averaged over the $N_{\text{sim}}$ trials:

\[
\begin{bmatrix}
\hat{\theta}_1 \\
\hat{\theta}_2
\end{bmatrix} = \begin{bmatrix}
0.6493 \\
1.9455
\end{bmatrix}
\]  

(10.141)

Clearly, the obtained estimates are far from the true continuous-time parameters $[1 \ 3]^T$.

We next consider the MSTTS model which here takes the form:

\[
\frac{(q - 1)^2}{\Delta^2} y_k + \theta_2 \frac{(q - 1)}{\Delta} y_k + \theta_1 \sqrt{|y_k|} = (1 + \text{sat}(y_k)) [q + (2 - \sqrt{3})] \cdot \tilde{v}_k
\]  

(10.142)

We again perform least squares estimation of the parameters, but first filter the regressors

\[
\frac{(q-1)^2 y_k}{1 + \text{sat}(y_k)} \quad \frac{(q-1) y_k}{1 + \text{sat}(y_k)} \quad \frac{\sqrt{|y_k|}}{1 + \text{sat}(y_k)}
\]  

(10.143)

by the asymptotic sampling zero polynomial, i.e. by the prefilter:

\[
F(q) = \frac{1}{q + (2 - \sqrt{3})}
\]  

(10.144)

leading to the following Equation Error model:

\[
F(q) \left\{ \frac{(q-1)^2 y_k}{1 + \text{sat}(y_k)} \right\} + \theta_2 F(q) \left\{ \frac{(q-1) y_k}{1 + \text{sat}(y_k)} \right\} + \theta_1 F(q) \left\{ \frac{\sqrt{|y_k|}}{1 + \text{sat}(y_k)} \right\} = e_{k+2}
\]  

(10.145)

The following mean estimates were obtained:

\[
\begin{bmatrix}
\hat{\theta}_1 \\
\hat{\theta}_2
\end{bmatrix} = \begin{bmatrix}
0.9688 \\
2.9058
\end{bmatrix}
\]  

(10.146)

The parameter estimates obtained using knowledge of the nonlinear stochastic sampling zero dynamics are clearly a better representation of the true continuous-time system parameters. Therefore, the inclusion of the stochastic sampling zero dynamics in (10.142) is crucial in obtaining meaningful estimates of the true continuous-time parameters. On the other hand, the parameter estimates obtained in (10.141) from the Euler-Maruyama model (10.139) are clearly biased.

10.9 Conclusions

In this chapter we have presented several results regarding sampled-data models for stochastic nonlinear systems. We first defined novel notions of truncation errors for stochastic systems. We then proved that a specific sampled-data model (STTS model) has improved order of accuracy when compared to the usual Euler-Maruyama scheme. We then introduced a closely related model
(MSTTS) which preserves the same accuracy property. In addition, we showed that, because of the improved order of accuracy, the MSTTS model has extra zero dynamics that have no continuous-time counterpart. Furthermore, the stochastic sampling zero dynamics are closely related to the Euler-Frobenius polynomials, which arise in the sampling of linear and nonlinear deterministic and linear stochastic systems. We have provided numerical simulations to illustrate the order of the error dynamics and the advantageous use of the sampling zero dynamics in the parameter estimation problem for nonlinear continuous-time stochastic systems.

This chapter is based on work previously described by the author in [26].
Conclusions and Suggestions for Future Research

In this thesis we have studied different robustness issues that arise in linear and nonlinear systems. Throughout the chapters we have reviewed existing results and presented novel contributions. In this chapter we will summarise the main results and discuss future research directions.

Linear Systems

In the first part of the thesis we have seen that, in the presence of uncertainty, closed-loop performance is usually compromised. We have presented two different strategies that allow improvements in closed-loop performance. They key idea behind both strategies is to augment the structure of the controller or the observer. In the FFMPC case, we use a separate optimisation problem that deals only with the performance variable. In the observer case, we showed that there are extra degrees of freedom in the design of a unbiased observer and that these can be used to account for certain types of mode errors or uncertainty.

Nonlinear Deterministic Systems

In the second part of the thesis, we turned our attention to nonlinear systems. The first important results is that we extended the definitions of truncation errors that are used in the numerical analysis literature to better account for control theory ideas. In particular, we introduced Vector Truncation Errors which provide an objective comparison of approximate sampled-data models, allowing the possibility to assign a different error bound to each element of the state vector. The novel definitions of Fixed Step truncation errors arise from the necessity in control engineering of
running a model over a finite number of steps, for example, in model predictive control or system identification. We applied these definitions to the existing TTS model, which we know includes the asymptotic sampling zero dynamics, and showed how the improved order of accuracy allowed us to (i) obtain better bias-variance tradeoffs in the parameter estimation of continuous-time systems from sampled-data, (ii) obtain a sampled-data model that depends only on input-output data that retains the Local Fixed Step order of the errors, and (iii) obtain better performance in the high-gain sampled-data feedback control of nonlinear systems.

Nonlinear Stochastic Systems

The idea of sampling zero dynamics is a well known one. Their existence has been proven for linear systems, both deterministic and stochastic, and for linear stochastic systems, but the question has been previously open for nonlinear stochastic systems. In linear stochastic systems, the sampling zeros appear in the sampled output spectrum, and a sampled-data model that includes the (asymptotic) sampling zeros can be obtained by spectral factorisation. None of the above is feasible in the nonlinear framework. To overcome this problem, we first showed that a specific stochastic sampled-data model, the STTS model, has improved order of accuracy, in the same way the TTS model has for the deterministic case. We then showed, based on the improved order of accuracy of the STTS model, that an equivalent model preserves the order of the errors in the output and explicitly contains an extra polynomial that corresponds to the sampling zero dynamics. It is interesting to note that, when compared to the deterministic case, the order of the results is reversed.

11.1 Summary of Contributions by Chapter

In Chapter 2 we introduced the ideas behind the use of feedforward in model predictive control (FFMPC). We presented the core concepts that underlie the potential advantages of using feedforward. The strategy relies on the use of a separate MPC optimisation problem embedded on top of a feedback MPC strategy. We showed that it is advantageous to use FFMPC when feedback control is severely compromised by the level of uncertainty. FFMPC is then built around a nominal model. Depending on the characteristics of the uncertainty, whether or not the real model is close to the nominal one, performance can be vastly improved. The downside is that if the opposite happens then performance will be deteriorated. Any probability distribution of the uncertainty that is accumulated around a nominal model will then clearly benefit from the scheme.
11.1 Summary of Contributions by Chapter

In Chapter 3 we have shown an application of FFMPC to the problem of feedback control of blood glucose. The key conclusion is that, for realistic variations in the model parameters of the actuator, plant and sensor, the achievable feedback closed-loop bandwidth is severely restricted. On the other hand, feedforward control does not suffer from this limitation. Thus feedforward provides the potential for major performance improvements.

In Chapter 4 we presented a relationship between output feedback sensitivity functions, filtering sensitivity functions and auxiliary variable feedback sensitivity functions. The main result is that the output feedback complementary sensitivity function can be decomposed into the product of the other two complementary sensitivity functions, i.e. the filtering and the auxiliary variable feedback complementary sensitivity functions. We showed that robust stability can be achieved, without severely compromising performance, by redesigning the observer beyond the Luenberger structure to the more general class of unbiased observers.

In Chapter 5 we defined three novel definitions of truncation error. These definitions extend the usual error definitions used in numerical analysis, where a uniform error bound is given across the elements of the state vector. In particular, we have considered the case where each element of the state vector may have different truncation errors. We have applied these definitions to a sampled-data model, named the TTS model, for nonlinear systems based on a normal form description. We showed that the TTS model possesses previously unknown properties, regarding the propagation of truncation errors. These properties are of direct relevance to control applications.

In Chapter 6 we used the properties of the TTS model to develop a sampled-data output predictor model for continuous time systems having no finite zeros. The model has two key properties, namely that it depends on input-output data only and that it has truncation error (fixed step) of order $\Delta^{n+1}$. The model makes explicit the additional sampling zero dynamics that arise due to the sampling process.

In Chapter 7 we examined the question of bias-variance tradeoffs when estimating the parameters of continuous time models using sampled data. We showed that, when approximate sampled-data models are used, then bias errors result which place an inescapable lower bound on the estimation accuracy no matter how much data is collected. We also showed that this lower bound can be reduced by using models with increased order or accuracy, e.g. the TTS model.

In Chapter 8 we discussed the role of asymptotic sampling zero dynamics in the sampled-data control of continuous nonlinear systems. The ideas build on the properties of the TTS model, namely the inclusion of the asymptotic sampling zero dynamics. We presented several examples
that illustrate the advantages of using a model that includes sampling zero dynamics. In particular, we showed that it can lead to significant improvement in the closed-loop performance of high-gain sampled-data controllers, at least for a certain class of continuous nonlinear dynamic systems.

In Chapter 10 presented and extended several results regarding sampled-data models for stochastic nonlinear systems. We first defined novel notions of truncation errors for stochastic systems. We then proved that a specific sampled-data model (STTS model) has improved order of accuracy when compared to the usual Euler-Maruyama scheme. In addition, we showed that, because of this improved order of accuracy, the STTS model has extra zero dynamics that have no continuous-time counterpart. Furthermore, the stochastic sampling zero dynamics are closely related to the Euler-Frobenius polynomials, which arise in the sampling of linear and nonlinear deterministic and linear stochastic systems. We illustrated the advantages of including the sampling zero dynamics in the parameter estimation problem for nonlinear stochastic systems.

11.2 Future Research

The work reported in this thesis has the potential to be extended in several directions:

FFMPC Design

The design of the FFMPC strategy is based on the knowledge of a nominal model. However, the choice of a “nominal” model within the uncertainty set is not unique. It would therefore be interesting to investigate a method to quantify, a priori, the ratio of systems that would get better/worse performance with a given choice of a “nominal” model. A further step would be to improve the FFMPC design by considering the probability distribution of the uncertainty, i.e. deciding which subset of the uncertainty would worsen their performance and which subset would improve their performance.

Controller design for sampled-data nonlinear systems

The controller design problem for sampled-data nonlinear systems is widely recognised as being difficult and has been explored from many different angles. In the work of Nešić et al. [99, 98], important results are presented on stability conditions when a digital controller, based on an approximate discrete-time model, is applied to a continuous time system. However, as stated
in [96], the results based on [99, 98] available today are prescriptive rather than constructive. The results state how a controller has to be designed and what properties the model has to possess, but the results do not tell one how to design the controller or how to obtain the model. The work on this topic presented in this thesis was of preliminary nature. The aim was to show that there are advantages to using a model with improved accuracy. It would be interesting to investigate the repercussions of using sampled-data models with improved accuracy in the above framework.

**Sampled-data Models**

The lack of good sampled-data models for nonlinear systems is still recognised as an important issue for control design [97]. Many results in the thesis are based on the TTS sampled-data model. It would be interesting to study other types of approximations, instead of the taylor series expansion used for the TTS model, and analyse the properties of the model. e.g. relative degree, sampling zero dynamics, truncation errors. Several options are available, e.g. Tchebyshev polynomials, Bernstein polynomials.
Euler-Frobenius Polynomials

This chapter presents a historical account of what are generally called Euler-Frobenius polynomials. Definitions and properties found in the literature are summarised. The intention is to clarify the difference between Euler, Eulerian and Euler-Frobenius numbers, fractions and polynomials found in the mathematics literature. A good introduction to this topic is given in [34].

Searching for Euler-Frobenius polynomials in the literature will not provide many references. The reason being, the polynomials referred to here as Euler-Frobenius, have been rediscovered and redefined many times in history. For instance, Sobolev [113] calls them Euler polynomials, Frobenius [43], Carlitz [17, 18] and Comtet [29, p. 244] call them Eulerian polynomials, and Reimer [103], Dubeau et al. [32] and Weller et al. [120] call them Euler-Frobenius polynomials. Adding to the confusion is the fact that there exist other polynomials, different to the three referred above, also called Euler polynomials [29, p. 48]. Furthermore, several generalisations of these polynomials are also called Eulerian, Euler-Frobenius or Frobenius-Euler, dropping the generalised prefix.

A first version of this chapter was published in [126] as a contributed chapter.

A.1 Euler-Frobenius relationship

Leonhard Euler [47, 118] first introduced the polynomials in 1749 in his work “Remarques sur un beau rapport entre les series des puissances tant direct que reciproques”, published in [35], while describing a method of computing values of the zeta function at negative integer points. Euler introduced them in the following way:

\[
\sum_{k=0}^{\infty} (k + 1)^n \cdot t^k = \frac{A_n(t)}{(1 - t)^{n+1}}.
\]  

(A.1)
For \( n = 0, \ldots, 4 \), the following expressions were obtained:

\[
1 + 2^0 t + 3^0 t^2 + 4^0 t^3 + \ldots = \frac{1}{1 - t} \tag{A.2}
\]
\[
1 + 2^1 t + 3^1 t^2 + 4^1 t^3 + \ldots = \frac{1}{(1 - t)^2} \tag{A.3}
\]
\[
1 + 2^2 t + 3^2 t^2 + 4^2 t^3 + \ldots = \frac{t + 1}{(1 - t)^3} \tag{A.4}
\]
\[
1 + 2^3 t + 3^3 t^2 + 4^3 t^3 + \ldots = \frac{t^2 + 4t + 1}{(1 - t)^4} \tag{A.5}
\]
\[
1 + 2^4 t + 3^4 t^2 + 4^4 t^3 + \ldots = \frac{t^3 + 11t^2 + 11t + 1}{(1 - t)^5} \tag{A.6}
\]

Ferdinand Georg Frobenius [61, 30] named the polynomials \( A_n(t) \), the Eulerian polynomials, and proved that \( A_n(t) \) has only simple negative real roots in the year 1910 [43]. Frobenius also showed that these polynomials could be written in terms of the Stirling numbers of the second kind, \( S(n,k) \) [29, Ch. V], by the following expression:

\[
A_n(t) = \sum_{k=1}^{n} k! \cdot S(n,k) \cdot (t - 1)^{n-k} \tag{A.7}
\]

We will now define the Euler-Frobenius polynomials, numbers and fractions. Note that some of the definitions may differ from the ones found in the literature by the polynomial \( A_0(t) \), which is not always defined. In this chapter we have modified the definitions so they are consistent. In this sense, the easiest way compare and distinguish between all of the polynomials and their different definitions is in terms of their generating function [54, 104, 122].

### A.2 Euler-Frobenius polynomials

#### A.2.1 Definition

The Euler-Frobenius polynomials are defined by the following exponential generating function

\[
\sum_{n=0}^{\infty} A_n(t) \cdot \frac{x^n}{n!} = \frac{t - 1}{t - e^{(t-1)x}} \tag{A.8}
\]
The first few polynomials are given by:

\[ A_0(t) = 1 \] (A.9)
\[ A_1(t) = 1 \] (A.10)
\[ A_2(t) = t + 1 \] (A.11)
\[ A_3(t) = t^2 + 4t + 1 \] (A.12)
\[ A_4(t) = t^3 + 11t^2 + 11t + 1 \] (A.13)

The Euler-Frobenius polynomials can be computed by the recurrence:

\[ A_0(t) = 1 \] (A.14)
\[ A_n(t) = t(1 - t) \cdot A'_n(t) + (1 + (n - 1)t) \cdot A_{n-1}(t) \] (A.15)

or they can be defined inductively by

\[ A_0(t) = 1 \] (A.16)
\[ A_n(t) = \sum_{k=0}^{n-1} \binom{n}{k} A_k(t) \cdot (t - 1)^{n-k-1} \] (A.17)

or, considering the notation \( A^n = A_n(t) \), by

\[ (A + (t - 1))^n - tA_n(t) = \begin{cases} (1 - t), & n = 0 \\ 0, & n > 0 \end{cases} \] (A.18)

They can also be computed explicitly by

\[ A_0(t) = 1 \] (A.19)
\[ A_n(t) = \frac{(1 - t)^{n+1}}{t} \left( \frac{d}{dt} \right)^{n-1} \frac{t}{(1 - t)^2}, \quad n > 0 \] (A.20)

Note that, to be consistent with the numbering of the polynomials, the last definition is shifted when compared to the original expression given in [113].

### A.2.2 Properties

The Euler-Frobenius polynomials have the following properties:

(i) \( A_n(t) \) is a palindromic polynomial, i.e. the sequence of its coefficients form a palindrome.
(ii) \( A_n(t) \) is a self-reciprocal polynomial, i.e. \( A_n(t) = t^{n-1} \cdot A_n(t^{-1}) \)

(iii) \( A_n(t) \) is a unimodal polynomial, i.e. the sequence of its coefficients is unimodal.

(iv) \( A_n(t) \) is a polynomial of degree \( n - 1 \).

(v) The roots of \( A_n(t) \) are simple, negative, real and distinct for all \( n \).

(vi) If \( t = t_0 \) is a root of \( A_n(t) \), i.e. \( A_n(t_0) = 0 \), then \( t = t_0^{-1} \) is also a root, i.e. \( A_n(t_0^{-1}) = 0 \).

(vii) In the case the degree of \( A_n(t) \) is odd, then \( A_n(t) \) always has a root at \( t = -1 \).

(viii) The roots of \( A_n(t) \) are separated by the roots of \( A_{n-1}(t) \), i.e. every root of \( A_{n-1}(t) \) lies in between two subsequent roots of \( A_n(t) \).

(ix) \( A_n(1) = n! \).

A.3 Euler-Frobenius numbers

A.3.1 Definition

The polynomials \( A_n(t) \) can also be written as

\[
A_n(t) = \sum_{k=0}^{n} A_{n,k} \cdot t^k
\]

(A.21)

where \( A_{n,k} \) are the Euler-Frobenius numbers, most commonly known as Eulerian numbers.

These numbers are defined by (see [40] [50])

\[
A_{n,k} = \sum_{i=0}^{k} \binom{n+1}{i} (k-i+1)^n (-1)^i
\]

(A.22)

The Eulerian numbers can also be obtained by the following recurrence relation

\[
A_{n,k} = (k+1) \cdot A_{n-1,k} + (n-k) \cdot A_{n-1,k-1}
\]

(A.23)

\[
A_{n,0} = 1
\]

(A.24)

\[
A_{n,k} = 0
\]

(A.25)

Table A.1 illustrates the first few numbers.
A.3 Euler-Frobenius numbers

<table>
<thead>
<tr>
<th></th>
<th>k=0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>26</td>
<td>66</td>
<td>26</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table A.1: Eulerian Numbers

A.3.2 Combinatorial interpretation of Eulerian numbers

The Eulerian numbers $A_{n,k}$ have a special property relating to Combinatorial Analysis (see [29 [50] and [100] Section 26.14].

A permutation of a set $\mathcal{N}$ is a bijective map of $\mathcal{N}$ onto itself. The set of all permutations of $\mathcal{N}$ is denoted by $\mathfrak{S}(\mathcal{N})$. If $\mathcal{N} = \{1, \ldots, n\}$, a permutation $\sigma = \sigma_1 \sigma_2 \cdots \sigma_n \in \mathfrak{S}(\mathcal{N})$ can be thought of as an ordered selection or arrangement of these integers where the integer in position $j$ is $\sigma_j$, e.g. if $\mathcal{N} = \{1, 2, 3\}$ then 231 is the permutation $\sigma_1 = 2, \sigma_2 = 3, \sigma_3 = 1$.

The following are a subset of properties defined for any permutation:

- An ascent of a permutation is a pair of adjacent elements for which $\sigma_j < \sigma_{j+1}$.
- A descent of a permutation is a pair of adjacent elements for which $\sigma_j > \sigma_{j+1}$.
- An excedance in a permutation is a position $j$ for which $\sigma_j > j$.
- A weak excedance in a permutation is a position $j$ for which $\sigma_j \geq j$.

The Eulerian number $A_{n,k}$ then represents:

- the number of permutations $\sigma \in \mathfrak{S}(\mathcal{N})$ that have $k$ ascents (descents).
- the number of permutations $\sigma \in \mathfrak{S}(\mathcal{N})$ with exactly $k$ excedances.
- the number of permutations $\sigma \in \mathfrak{S}(\mathcal{N})$ with exactly $k+1$ weak excedances.

The total number of permutations of a set with $n$ elements is well known to be $n!$. In the same way, the sum of all the permutations that have up to $n$ ascents is the total number of permutations.
Therefore, the *Euler-Frobenius* polynomials and numbers satisfy the following identity:

\[ A_n(1) = \sum_{k=0}^{n} A_{n,k} = n! \quad (A.26) \]

### A.4 Euler-Frobenius fractions

#### A.4.1 Definition

The name *Euler-Frobenius fractions* is less common in the literature, but is kept here for clarity of exposition. They can be found under the name of *Eulerian number* in the work of Carlitz [17][18] and *Euler-Frobenius numbers* in the work of Kim [69][68]. They first appeared in Euler’s treaty *Institutiones Calculi Differentialis, Section 173, Chapter 7, Part II* [34].

The *Euler-Frobenius fractions* \( H_n(t) \) are defined by the following exponential generating function:

\[
\sum_{n=0}^{\infty} H_n(t) \cdot \frac{x^n}{n!} = \frac{1 - t}{e^x - t} \quad (A.27)
\]

The first few fractions are given by

\[
\begin{align*}
H_0(t) &= 1 \quad (A.28) \\
H_1(t) &= \frac{1}{t - 1} \quad (A.29) \\
H_2(t) &= \frac{t + 1}{(t - 1)^2} \quad (A.30) \\
H_3(t) &= \frac{t^2 + 4t + 1}{(t - 1)^3} \quad (A.31) \\
H_4(t) &= \frac{t^3 + 11t^2 + 11t + 1}{(t - 1)^4} \quad (A.32)
\end{align*}
\]

The *Euler-Frobenius fractions* can be generated inductively, considering the notation \( H^n = H_n(t) \), by

\[
(H + 1)^n - t \cdot H_n(t) = 0, \quad H_0(t) = 1 \quad (A.33)
\]

#### A.4.2 Properties

The *Euler-Frobenius fractions* possess the property that

\[
\frac{dH_n(t)}{dt} = n \cdot H_{n-1}(t) \quad (A.34)
\]
A.5 Generalised Eulerian polynomials

A.5.1 Definition

Carlitz introduced a generalisation of the Euler-Frobenius fractions. He referred to them as Eulerian polynomials [17, 18]. The generalisation is defined by the following exponential generating function

\[
\sum_{k=0}^{\infty} H_n(u,t) \cdot \frac{x^k}{k!} = (1 - t) e^{ux} e^{x - t} \quad (A.37)
\]

These polynomials can be computed inductively by means of the following identity

\[
H_n(u,t) = \sum_{k=0}^{n} \binom{n}{k} H_k(t) \cdot u^{n-k} \quad (A.38)
\]

where \(H_k(t)\) are the Euler-Frobenius fractions previously defined. The first few polynomials are given by

\[
H_0(u,t) = 1 \quad (A.39)
\]
\[
H_1(u,t) = u + \frac{1}{t-1} \cdot u^0 \quad (A.40)
\]
\[
H_2(u,t) = u^2 + \frac{1}{t-1} \cdot 2u + \frac{t+1}{(t-1)^2} \cdot u^0 \quad (A.41)
\]
\[
H_3(u,t) = u^3 + \frac{1}{t-1} \cdot 3u^2 + \frac{t+1}{(t-1)^2} \cdot 3u + \frac{t^2 + 4t + 1}{(t-1)^3} \cdot u^0 \quad (A.42)
\]

It is then straightforward to see that

\[
H_n(0,t) = H_n(t) \quad (A.43)
\]
\[
(t - 1)^n \cdot H_n(0,t) = A_n(t) \quad (A.44)
\]
On the other hand, by simply comparing their respective generating function, the polynomials defined by Carlitz have the following property

\[ H_n(u, -1) = E_n(u) \]  \hspace{1cm} (A.45)

where \( E_n(u) \) are the Euler polynomials (defined later). The Eulerian polynomials also satisfy:

\[ \frac{\partial H_n(u, t)}{\partial u} = n \cdot H_{n-1}(u, t) \]  \hspace{1cm} (A.46)

### A.5.2 Properties

Carlitz proved several multiplication and summation properties of the Eulerian polynomials, e.g. the following summation identity holds:

\[ \sum_{k=0}^{n} \binom{n}{k} H_k(u, t) \cdot H_{n-k}(v, -t) = H_n(u + v, t^2) \]  \hspace{1cm} (A.47)

In addition, the following multiplication formula holds provided \( t \neq 1, z \neq 1, tz \neq 1 \):

\[ H_m(u, t)H_n(u, z) = H_{m+n}(u, tz) \]  \hspace{1cm} (A.48)

\[ + \frac{t(1-z)}{1-tz} \sum_{r=1}^{m} \binom{m}{r} H_r(t)H_{m+n-r}(u, tz) \]

\[ + \frac{z(1-t)}{1-tz} \sum_{s=1}^{n} \binom{n}{s} H_s(z)H_{m+n-s}(u, tz) \]

On the other hand, when \( tz = 1, t \neq 1 \), the following relationship holds:

\[ H_m(u, t) \cdot H_n(u, t^{-1}) = (t-1) \cdot \sum_{r=1}^{m} \binom{m}{r} H_r(t) \frac{B_{m+n-r+1}(t)}{m+n-r+1} \]  \hspace{1cm} (A.49)

\[ + (t^{-1}-1) \cdot \sum_{s=1}^{n} \binom{n}{s} H_s(t^{-1}) \frac{B_{m+n-s+1}(t)}{m+n-s+1} \]

\[ + (-1)^{n+1} \frac{m!}{(m+n+1)!} (1-t)H_{m+n+1}(t) \]

Several other identities are provided in [18].

### A.6 Euler polynomials and numbers

The following polynomials are defined here for the sake of completeness.
The Euler polynomials are defined by the following exponential generating function:

\[
\sum_{k=0}^{\infty} E_n(t) \cdot \frac{x^k}{k!} = \frac{2e^{tx}}{e^x + 1}
\]  

(A.50)

where the first few polynomials are given by

\[
E_0(t) = 1
\]

(A.51)

\[
E_1(t) = t - \frac{1}{2}
\]

(A.52)

\[
E_2(t) = t^2 - t
\]

(A.53)

\[
E_3(t) = t^3 - \frac{3}{2}t^2 + \frac{1}{4}
\]

(A.54)

\[
E_4(t) = t^4 - 2t^3 + t
\]

(A.55)

The Euler polynomials have only one root in the interval \( t \in (0, 1) \).

The Euler numbers \( E_n \) are defined in terms of the Euler polynomials \( E_n(t) \) by evaluating \( t = 1/2 \) and normalising by \( 2^n \), i.e.

\[
E_n = 2^n \cdot E_n \left( \frac{1}{2} \right)
\]

(A.56)

The first few Euler numbers are

\[
E_0 = 1
\]

(A.57)

\[
E_1 = 0
\]

(A.58)

\[
E_2 = -1
\]

(A.59)

\[
E_3 = 0
\]

(A.60)

\[
E_4 = 5
\]

(A.61)

These are also the coefficients of the Taylor-Maclaurin series

\[
\frac{1}{\cosh(t)} = \frac{2e^x}{e^{2x} + 1} = \sum_{k=0}^{\infty} E_n \cdot \frac{x^k}{k!}
\]

(A.62)

### A.7 Bernoulli polynomials and numbers

On the other hand, the Bernoulli polynomials are defined by the following exponential generating function:

\[
\sum_{k=0}^{\infty} B_n(t) \cdot \frac{x^k}{k!} = \frac{x \cdot e^{tx}}{e^x - 1}
\]

(A.63)
The first few *Bernoulli* polynomials are
\[ B_0(t) = 1 \] (A.64)
\[ B_1(t) = t - \frac{1}{2} \] (A.65)
\[ B_2(t) = t^2 - t + \frac{1}{6} \] (A.66)
\[ B_3(t) = t^3 - \frac{3}{2} t^2 + \frac{1}{2} t \] (A.67)
\[ B_4(t) = t^4 - 3 t^3 + t^2 - \frac{1}{30} \] (A.68)

The well-known *Bernoulli* numbers are given by
\[ B_n = B_n(0) \] (A.69)

Finally, it is important to note that the *Euler* and *Bernoulli* numbers satisfy
\[ E_{2n+1} = 0, \quad B_{2n+1} = 0, \quad n = 1, 2, \ldots \] (A.70)

In addition, similar to the *Euler-Frobenius* fractions, both *Euler* and *Bernoulli* polynomials form an Appell sequence, i.e., they satisfy the identity
\[ \frac{dE_n(t)}{dt} = n \cdot E_{n-1}(t) \] (A.71)
\[ \frac{dB_n(t)}{dt} = n \cdot B_{n-1}(t) \] (A.72)

There exists a vast literature on many other properties and relations that apply between *Euler* and *Bernoulli* polynomials and numbers. However, only basic definitions and properties will suffice for the purpose of this chapter.

### A.8 Summary

This chapter has given a thorough presentation of the *Euler-Frobenius* polynomials. The definitions and properties presented here should allow the reader to easily identify the different types of numbers, polynomials and fractions present in the literature.
Bibliography


[34] L. Euler. *Institutiones calculi differentialis cum eius usu in analysi finitorum ac doctrina serierum*. Academiae Imperialis Scientiarum Petropolitanae, St.Petersbourg, 1755.


