Deterministic Variational Approach to System Modelling by Data Assimilation

Abhishek Pandey

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

February 2018

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering.

2018/02/21

To my parents, who have sacrificed a lot for me to be here.
To my professor, the kindest lady I know.

Abstract

A non-asymptotic approach to simultaneous model and state estimation is discussed in this thesis. Specifically, given a cloud of measurement points presumed to represent a dynamical system output trajectory, a linear model is constructed to fit it best. The approach is somewhat similar to variational data assimilation in that it employs a cost functional as a measure of model fitness and computes its gradient by solution of an adjoint sensitivity problem. A kernel system model is first constructed and viewed as a linear finite dimensional subspace of a reproducing kernel Hilbert space (RKHS). The subspace is linearly parametrized by the unknown system constants whose values determine the subspace "orientation" vis á vis the cloud of measurement points. The system constants are found simultaneously with reconstructing the system state. A geometrical interpretation of this process is particularly appealing as it can be visualized as subspace re-orientation in the RKHS by way of minimizing the residual of the modelling error.

The method does not employ numerical differentiation techniques. The joint model and state estimation does not require any information about boundary conditions of the presumed dynamical system nor the measurement noise characteristic. Additionally, as the optimization process searches for the best estimates of the system parameters the "empirical" statistical distribution of the modelling residual error can be tested for congruence with a priori knowledge about the measurement noise. The non-parametric Kolmogorov probability density test is used in this thesis to test such congruence. Examples confirm high accuracy of adaptive estimation, but the algorithm is computationally expensive as it employs optimization by iterative search.

The fifth chapter of the thesis presents a radical re-formulation of the problem that directly employs the basis functions of the RKHS. Such formulation bypasses the evaluation of the fundamental solutions of the system model and does not require variational techniques. Numerical results pertaining to the latest problem statement will be presented in a forthcoming publication.

Résumé

Cette thèse présente une approche non-asymptotique d'estimation simultanée du modèle du système et de son état. Plus précisément, un modèle linéaire est construit afin de s'ajuster à une trajectoire de sortie de système dynamique représentée par un nuage de points de mesure. L'approche est quelque peu similaire à l'assimilation variationnelle des données en ce sens qu'elle utilise un coût fonctionnel pour obtenir un ajustement de modèle et calcule son gradient par la résolution d'un problème de sensibilité adjoint. Un modèle de système à noyau est d'abord construit et considéré comme un sous-espace linéaire de dimension finie d'un espace de Hilbert à noyau de reproduction (RKHS). Le sous-espace est paramétré de manière linéaire par les constantes inconnues du système dont les valeurs déterminent l'orientation du sous-espace par rapport au nuage de points de mesure. Les paramètres du système sont trouvées simultanément avec la reconstruction de l'état du système. Une interprétation géométrique de ce processus est particulièrement attrayante car elle peut être visualisée comme une réorientation du sous-espace dans le RKHS en minimisant le résidu de l'erreur de modélisation.

La méthode n'utilise pas de techniques de différentiation numérique. Le modèle conjoint et l'estimation d'état ne nécessitent aucune information sur les conditions limites du présumé système dynamique ni sur la caractéristique de bruit de mesure. De plus, pendant que le processus d'optimisation recherche les meilleures estimations des paramètres du système, la distribution statistique empirique de l'erreur résiduelle de modélisation peut être testée pour la congruence, avec la connaissance a priori du bruit de mesure. Le test de densité de probabilité non paramétrique de Kolmogorov est utilisé dans cette thèse pour tester une telle congruence. Les exemples confirment la haute précision de l'estimation adaptative, mais l'algorithme est coûteux en calcul car il utilise un processus itératif d'optimisation.

Le cinquième chapitre présente une reformulation radicale du problème qui utilise directement les fonctions de base du RKHS. Une telle formulation contourne l'évaluation des solutions fondamentales du modèle de système et ne nécessite pas de techniques variationnelles. Les résultats numériques relatifs au dernier énoncé du problème seront présentés dans une prochaine publication.

Acknowledgments

I want to express my gratitude to several people who have helped me constantly throughout my academic endeavour to produce this thesis.

First and foremost, I thank my supervisor, Professor Hannah Michalska, who helped develop my capability so that I could engage in research. I would like to express appreciation for her patience, in helping me get past my intellectual limitations. I sincerely acknowledge her help in every aspect of my studies and regard her with great respect for the exceptional researcher and kind person that she is. I learnt from her the ability to keep good humor while being under pressure, and hope to retain this learning forever.

I would also take this opportunity to recognize the endless efforts of my parents, who currently reside quite far away, but whose support and encouragement gives me courage all the time. I hope this work repays for their sacrifice, at least partially.

I convey my gratefulness to Debarshi Ghoshal and Kumar Vishwanath Gopalakrishnan, friends and mentors, who have looked over my shoulder during my research, always available with their forethought and their improvisations.

This list would be incomplete without mentioning the fruitful discussions I had with my friends Aditya, Arbaaz and Shaunak. They have always been accommodating of all my requirements in completing this work and I have picked up good qualities from all of them.

I am in great debt to my friends, especially the ones I made in Montréal, Neelakshi, Priyanka, Ruchika, Alida, Sahil, Aditya, Stefania and Nachiket, who have been sources of strength, even if I don't convey this frequently.

At last, I would like to thank for the blessings of God, who always listens to my mother.

Preface

This is to declare that this thesis work was carried out by myself, but was part of a larger collaborative effort - that of the research team led by my supervisor Dr. Hannah Michalska. Although each student in the group was assigned a different research topic, present was a unifying theme of system modeling and estimation by concepts of machine learning and functional data analysis.

The kernel representation of the system in the RKHS employed here was first derived by Debarshi Patanjali Ghoshal, PhD scholar of this research group, but is properly corrected and verified only by me in this thesis.

The rigorous theoretical discussion leading to the problem re-formulation in the RKHS, presented in the last chapter of this thesis, is based on the research notes of my supervisor Dr. Michalska, as is duly acknowledged in the text.

Contents

List of Figures

List of Tables

List of Acronyms

RKHS	Reproducing Kernel Hilbert Space
KS	Kolmogorov Smirnov
LTI	Linear Time Invariant
SISO	Single Input Single Output

Chapter 1

Introduction

In an attempt to improve existing technologies, Control Systems has become one of the most important areas of research interest. Since technology has entered our lives rapidly, automatic controllers are all around us, now more than ever. They are at work in our homes, our cars, our factories, our transportation systems, our defense systems, everywhere we look [7]. A control system is a system by which any quantity of interest in a machine, mechanism or other equipment is maintained or altered in accordance with a desired manner [8]. For example, the automobile driving system, comprising of its accelerator, fuel injectors and the engine, constitute a control system because it works to maintain a desired speed.

A control law or controller is a set of rules that allows us to determine the commands to be sent to the governed plant (via an actuator) to achieve the desired evolution. These rules can be described as either *open-loop* control or *closed-loop* (feedback) control [6].

An *open-loop* control system is a system which cannot correct variations in the output because it does not employ feedback. A traffic light system with fixed intervals of time at which the lights change colors is an open loop system.

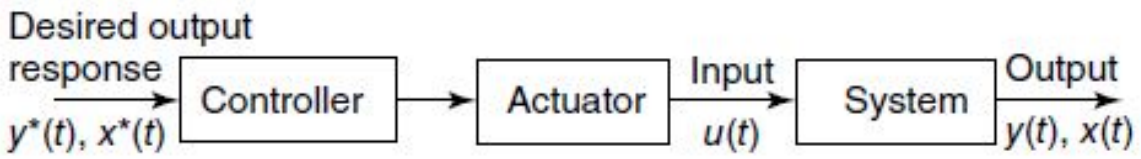

Figure 1.1 An open loop control system [6]

A *closed-loop* control , on the other hand, is a type of system which employs a feedback loop. An error detector compares a signal, a function of the output, obtained from a sensor, with a reference input. The difference between these signals is used by the controller to determine a control action and reduce the error.

Considering the previous example of the traffic system; if an arrangement was made to

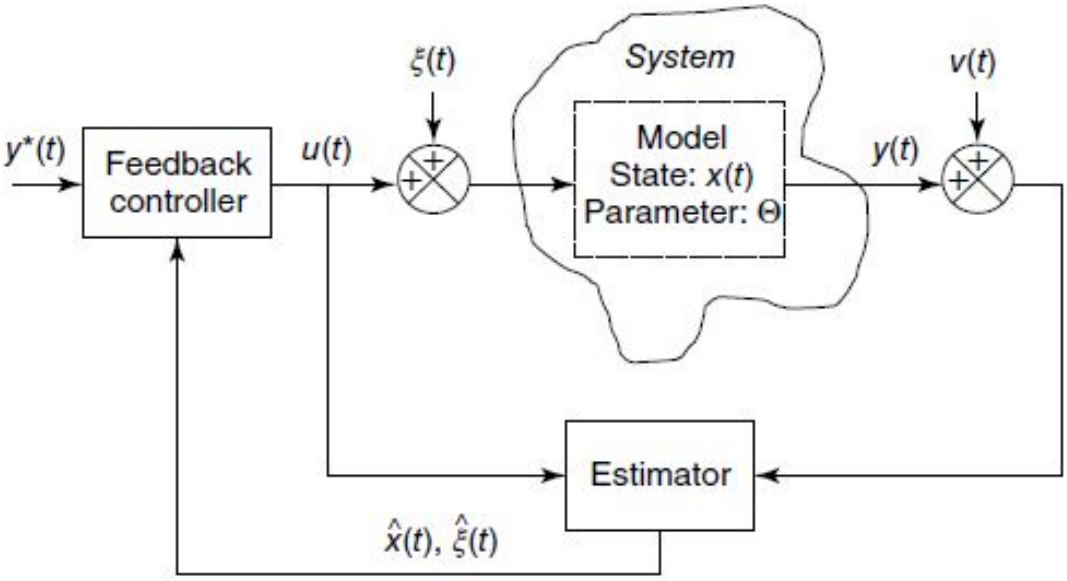

Figure 1.2 A closed loop control system [6]

take into account the traffic flow , it would be a closed loop system.

Feedback is a mechanism to command a system to evolve in a desired fashion so that the states, and outputs, exhibit a desired evolution (e.g., to track a reference trajectory) or stay at a prescribed equilibrium. Feedback can also be used to stabilize the state of a system, while also improving its performance [6].

1.1 Parameter Estimation in Linear Systems

A mathematical model of "the plant" is viewed as one of the most essential parts of a control system. Defining a model by a finite number of parameters leads to a parametric model, which is necessary to be known in order to implement a model based controller. This makes the problem of parameter estimation to be of equal, if not more important, than state estimation.

Parameter estimation is the experimental determination of values of parameters that govern the dynamic behavior of the system, assuming that the structure of the system is well known [6]. One of the most important early contributions in this area was the statistical view of the parameter identification problem, proposed by Gauss [9] and known as the "inverse problem of computing the response of a system with known characteristics", and the maximum likelihood procedure introduced by Fisher [10].

An example of a parameter estimation problem would be the following; consider a continuous-time (CT) linear time-invariant (LTI) system, whose order n is known, described by:

$$a_n \frac{d^n x}{dt_n} + a_{n-1} \frac{d^{n-1} x}{dt_{n-1}} + \cdots + a_0 x(t) = b_m \frac{d^m u}{dt_m} + b_{m-1} \frac{d^{m-1} u}{dt_{m-1}} + \cdots + u(t) \tag{1.1}$$

A noise corrupted version of x is measured, $z(t_k) = x(t_k) + \epsilon(t_k), k = 1, 2, ..., M$ where, $\epsilon(t_k)$ is Gaussian noise with mean 0 and variance σ^2, M is the number of samples taken. The goal of identification is to estimate the parameters $a_n, a_{n-1}, ..., a_1, a_0, b_m, b_{m-1}, ..., b_1$ from the input-output data i.e., $(u(t_k), z(t_k); k = 1, 2, ..., M$. [11]

The parameter estimation of a homogeneous system can be viewed as the identification of a differential invariant $\mathcal{I}$ ($\mathcal{I} \equiv 0$)under the action of the flow of some closed loop system (such as its characteristic equation) [12]:

$$\mathcal{I}(t, y(t), y^{(1)}(t), \cdots, y^{(n)}(t)) \equiv y^{(n)}(t) + a_{n-1} y^{(n-1)}(t) + \cdots + a_0 y(t); t \geq 0 \tag{1.2}$$

This thesis adopts the same method of parameter estimation using the Kernel function, as mentioned in [1]. But then it shows that these parameters aren't the BEST estimates for the system and develops the need for simultaneous denoisifcation and model shaping.

1.2 State Estimation in Linear Systems

The importance of state estimation is evident from figure 1.2. States $x(t)$ and parameters θ of the system model are required to produce the final output $y(t)$. The *state* of a system is a minimum set of variables (state variables) whose present values, together with the values of the input signals in the future, completely determine the future behavior of the system [6]. A feedback controller requires the states of the system to generate a control law in order to drive the final output to a desired output value $y^*(t)$. It can also be seen from the figure that the system is plagued by external input disturbances and measurement noise; hence, the controller needs to be robust to these "exogenous" signals. Hence, accurate state availability becomes even more important.

Unfortunately all the state variables are not measurable; hence, we need to build a state estimator as a first step to building a feedback controller. This problem has been tackled in depth by Kalman and Luenberger , which leads to an extensive theory development on *Observers* as mentioned in [13], [14], [15], [16], [17], [18], [19]. An observer for a linear system is readily feasible whenever the system is observable, that is, when its observability matrix has full rank. If the observer gains are optimized for the noise input to the system and to the sensor(s), the observer is called a Kalman filter. If the gains are not so optimized, and the setting is deterministic, the observer is called a Luenberger observer [6].

Time derivative estimators can also be used to estimate states of a dynamical system. If a system is observable, the state-space can be reconstructed from a set of measurements of the input, the output, and a finite number of their time derivatives. One way to reconstruct the state-space via time derivatives is by exploiting the property of differential flatness of systems [6].

1.2.1 Flat systems

A variable of a system is said to be *endogenous* if it can be expressed as a differential function of the input, the output and a finite number of their time derivatives. A system is said to be *flat*, if there exists an endogenous variable, called the flat output f, such that the input u and the output y can be expressed as linear combinations of the flat output and finite number of its time derivatives.

Consider a LTI SISO system described by:

$$y(s) = \frac{b_m s^m + b_{m-1} s^{m-1} + \cdots + b_0}{s^n + a_{n-1} s^{n-1} + \cdots + a_0} u(s), \qquad m < n \tag{1.3}$$

with the numerator and the denominator being co-prime. A flat output, f can be considered to be

$$f(s) = \frac{\kappa}{s^n + a_{n-1} s^{n-1} + \cdots + a_0} u(s), \qquad \kappa \neq 0 \tag{1.4}$$

Thus, the original system can now be represented as:

$$\begin{aligned} \frac{d^n f}{dt_n} + a_{n-1} \frac{d^{n-1} f}{dt_{n-1}} + \cdots + a_0 f = \kappa u \\ y = \frac{1}{\kappa} \left[b_m \frac{d^m f}{dt_m} + b_{m-1} \frac{d^{m-1} f}{dt_{m-1}} + \cdots + b_0 f \right] \end{aligned} \tag{1.5}$$

indicating that the system in (1.3) is a differentially flat system [20].

According to [21], a system variable is said to be *algebraically observable* if it can be represented in terms of the input, the output, and a finite number of their time derivatives, as expressed above. Thus, observer design is reduced to the design of a numerical differentiator [22], [23] . One problem which is encountered while estimation using time derivatives is that a typical differentiator is affected by noise. Thus, if we choose this estimation technique, an additional filtering algorithm has to be implemented in the estimation procedure to obtain reasonable results under noisy measurement conditions. Notable work in optimal filtering was done by Norbert Wiener [24] and also by Rudolf Kalman, in the form of Kalman-Bucy filter [14], [16].

1.2.2 Algebraic Methods in Control Theory

Algebraic Methods are an identification methodology used to estimate states, but not limited to it. They belong to the class of Time-derivative estimators and have the following advantages [6]:

1. Robust State estimators can be designed even in the presence of noise corrupting the

data.

2. No statistical knowledge of the noise is required for the estimation.
3. They are robust with respect to different initial conditions provided to the system.
4. Unlike classical observers as mentioned in this section, this methodology features faster convergence than an asymptotic one and does not focus on Lyapunov Stability theory for design.

This work focuses on using the kind of methods as contained by 1.2.2, specifically, based on the attenuation of noise achieved by the integration operation.

1.3 Thesis Objectives and Organization

The primary objective of the thesis is to develop a new method for system modelling and estimation which is non-asymptotic and can be applied to time windows of arbitrary length. It endeavours to extend the previous work on algebraic approaches to state and parameter estimation of linear systems as in [1]. Additionally, this work aims at:

- finding the best possible parameters to fit a linear model to a data cloud. [3]
- increasing the accuracy of the state estimation over the whole time window. [3]
- representing the system in an alternative form in RKHS.
- formulation of linear modelling problem in RKHS.

Summary of Thesis Content

Chapter 1 of the thesis introduces Control Systems while differentiating between Open-Loop and Closed-Loop control; stressing on the importance of feedback. The primary goal of the chapter is to introduce the concept of parameter and state estimation. A general problem is formulated to explicitly convey the meaning of parameter estimation. This chapter also contains a short section on *Flat Systems*, which is necessary for the understanding of algebraic estimation techniques mentioned in the later chapters.

Chapter 2 almost entirely focusses on the derivation of the double-sided kernel as developed in [1]. This chapter also states the "Joint State and Parameter Estimation" problem

which is being treated in this entire work. An example is also considered to demonstrate parameter estimation and the need for using Optimisation techniques in Simultaneous Denoisification and Model Shaping is established. In that, it serves as a critique to the work presented in [1].

Chapter 3 tackles the shortcomings of the state estimation method expressed before. A more radical, non-asymptotic approach to simultaneous model and state estimation is adopted in this chapter. This approach employs a cost functional as a measure of model fitness and computes its gradient by solution of an adjoint sensitivity problem. Broyden's method is also discussed as an alternative to the gradient calculation as described in the former approach.

The appendix provides some important concepts and results from Calculus of Variations and can be referred to for assistance.

Chapter 4 presents the results after application of the simultaneous denoisifcation and model shaping method as developed in Chapter 3. Most of the results consider white Gaussian noise with Signal to Noise Ratio(SNR) of 30dB but the method is also tested against a white Gaussian noise of SNR 10dB. The non-parametric Kolmogorov probability density test is used to assess congruence of the modelling residual errors with the (in this case assumed) statistical density of the measurement noise.

Chapter 5 aims at a theoretical construction of modelling of systems using RKHS. It delves into a detailed study of theory involving Reproducing Kernel Hilbert Spaces. It describes classical construction of a RKHS space and moves onto application of the theory to our specific problem of simultaneous denoisifcation and model shaping. This chapter formulates a Tikhonov Regularization problem pertaining to the original application of model shaping but does not contain results and is thus viewed as a potential area of future work.

Chapter 6 of the thesis provides a conclusion of the work presented in the previous chapters, reiterates the results and suggests future work.

Chapter 2

A Double Sided Kernel approach to Model and State Estimation

2.1 Introduction

The double sided kernel approach was developed and discussed in [1] but will be produced here.

2.2 An Overview of the Double Sided Kernel [1]

In this approach, the state equations are replaced with an output reproducing property on an arbitrary time $[t_a, t_b]$ which follows directly from the knowledge of the system characteristic equation. The behavioural model is derived from the differential invariance which is characteristic of the system and eliminates the need of initial conditions and is in the form of a homogeneous Fredholm integral equation of the second kind with a Hilbert-Schmidt kernel [2]. The mathematical interpretation as a Reproducing Kernel Hilbert Space (RKHS) of the behavioral model allows us to extract signal and its time derivatives that confirm the system invariance from output measurement subject to noise. The details concerning the RKHS interpretation of the problem are presented in the Chapter 5.

2.3 Development of the Double Sided Kernel [2]

Before proceeding with the discussion of the actual development of the method, let us first describe the system whose states we are estimating. We are using this extremely simple model only for simplicity and brevity of exposition. For higher order systems the procedure is identical and its general validity can be proved by mathematical induction.

2.3.1 System Description

Consider the following third order LTI system:

$$\begin{aligned} \dot{x}(t) &= Ax(t) \\ y(t) &= Cx(t) \end{aligned} \tag{2.1}$$

where A, C, x(t), $\dot{x}$(t) and y(t) are given as follows

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{2.2}$$

$$C = \begin{bmatrix} c_{11} & c_{12} & c_{13} \end{bmatrix} \tag{2.3}$$

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} \tag{2.4}$$

$$\dot{x}(t) = \begin{bmatrix} \dot{x_1}(t) \\ \dot{x_2}(t) \\ \dot{x_3}(t) \end{bmatrix} \tag{2.5}$$

$$y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{bmatrix} \tag{2.6}$$

$$\det(A - \lambda I) = \det \left\{ \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right\}$$

The characteristic equation for the above system is $\det(A - \lambda I) = 0$, i.e.

$$\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0 = 0 \tag{2.7}$$

Using Cayley Hamilton Theorem, we can rewrite (2.7) as follows.

$$A^3 + a_2A^2 + a_1A + a_0 = 0 \tag{2.8}$$

Now, if we multiply (2.8) by $x(t)$ from the right and by C from the left, we get,

$$CA^3x(t) + a_2CA^2x(t) + a_1CAx(t) + a_0Cx(t) = 0 \tag{2.9}$$

Differentiating (2.1) with respect to time yields,

$$y(t) = C\dot{x}(t) \tag{2.10}$$

Substituting the first part of equation (2.1) into (2.10)

$$y^{(1)}(t) = CAx(t) \tag{2.11}$$

Differentiating (2.11) and substituting again two more times we get,

$$y^{(2)}(t) = CA^2x(t) \tag{2.12}$$

and

$$y^{(3)}(t) = CA^3x(t) \tag{2.13}$$

Now using (2.11) , (2.12), (2.13) and substituting them in (2.9) we finally arrive at,

$$y^{(3)}(t) + a_2y^{(2)}(t) + a_1y^{(1)}(t) + a_0y(t) = 0 \tag{2.14}$$

2.3.2 Derivation of the Double Sided Kernel

We shall now derive the double-sided kernel for a general third order characteristic polynomial described in the previous subsection

$$y^{(3)}(t) + a_2 y^{(2)}(t) + a_1 y^{(1)}(t) + a_0 y(t) = 0 \tag{2.15}$$

on an interval $[a, b]$.

Let us consider the following two equations obtained from (2.15) by multiplying $(\xi - a)^3$ and $(b - \zeta)^3$

$$(\xi - a)^3 y^{(3)}(t) + a_2(\xi - a)^3 y^{(2)}(t) + a_1(\xi - a)^3 y^{(1)}(t) + a_0(\xi - a)^3 y(t) = 0 \tag{2.16}$$

$$(b - \zeta)^3 y^{(3)}(t) + a_2(b - \zeta)^3 y^{(2)}(t) + a_1(b - \zeta)^3 y^{(1)}(t) + a_0(b - \zeta)^3 y(t) = 0 \tag{2.17}$$

Now we integrate (2.16) and (2.17) thrice on the interval $[a, a + \tau]$ and $[b - \sigma, b]$. This effectively means that we would be integrating the (2.15) in the forward direction during the interval $[a, a + \tau]$ and in the backward direction on the interval $[b, b - \sigma]$.
Integrating the first term in (2.16) for the first time,

$$\begin{aligned}
\int_a^{a+\tau} (\xi - a)^3 y^{(3)}(\xi)\,\mathrm{d}\xi &= (\xi - a)^3 y^{(2)}(\xi)\,|_a^{a+\tau} - \int_a^{a+\tau} 3(\xi - a)^2 y^{(2)}(\xi)\,\mathrm{d}\xi \\
&= \tau^3 y^{(2)}(a + \tau) - \left[3(\xi - a)^2 y^{(1)}(\xi)\,|_a^{a+\tau} - \int_a^{a+\tau} 6(\xi - a) y^{(1)}(\xi)\mathrm{d}\xi\right] \\
&= \tau^3 y^{(2)}(a + \tau) - 3\tau^2 y^{(1)}(a + \tau) + 6(\xi - a) y^{(1)}(\xi)\,|_a^{a+\tau} - \int_a^{a+\tau} 6y(\xi)\,\mathrm{d}\xi \\
&= \tau^3 y^{(2)}(a + \tau) - 3\tau^2 y^{(1)}(a + \tau) + 6\tau y(a + \tau) - \int_a^{a+\tau} 6y(\xi)\,\mathrm{d}\xi
\end{aligned} \tag{2.18}$$

When we integrate again, the upper limit on the integral becomes a 'dummy variable', that is we set $\xi' = a + \tau$ then,
$\tau^3 y^{(2)}(a + \tau)$ is integrated as $(\xi' - a)^3 y^{(2)}(\xi')$

$3\tau^2 y^{(1)}(a+\tau)$ is integrated as $3(\xi'-a)^2 y^{(1)}(\xi')$
$6\tau y(a+\tau)$ is integrated as $6(\xi'-a)y(\xi')$
Integrating (2.18) again,

$$\begin{aligned}
\int_a^{a+\tau}\int_a^{\xi'}(\xi-a)^3 y^{(3)}(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' &= \int_a^{a+\tau}(\xi'-a)^3 y^{(2)}(\xi')\,\mathrm{d}\xi' - \int_a^{a+\tau} 3(\xi'-a)^2 y^{(1)}(\xi')\,\mathrm{d}\xi' \\
&\quad + \int_a^{a+\tau} 6(\xi'-a)y(\xi')\,\mathrm{d}\xi' - \int_a^{a+\tau}\int_a^{\xi'} 6y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' \\
&= (\xi'-a)^3 y^{(1)}(\xi')\,|_a^{a+\tau} - \int_a^{a+\tau} 3(\xi'-a)^2 y^{(1)}(\xi')\,\mathrm{d}\xi' \\
&\quad - \left[3(\xi'-a)^2 y(\xi')\,|_a^{a+\tau} - \int_a^{a+\tau} 6(\xi'-a)y(\xi')\,\mathrm{d}\xi'\right] \\
&\quad + \int_a^{a+\tau} 6(\xi'-a)y(\xi')\,\mathrm{d}\xi' - \int_a^{a+\tau}\int_a^{\xi'} 6y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' \\
&= \tau^3 y^{(1)}(a+\tau) - \left[3(\xi'-a)^2 y(\xi')\,|_a^{a+\tau} - \int_a^{a+\tau} 6(\xi'-a)y(\xi')\,\mathrm{d}\xi'\right] \\
&\quad - 3\tau^2 y(a+\tau) + \int_a^{a+\tau} 12(\xi'-a)y(\xi')\,\mathrm{d}\xi' - \int_a^{a+\tau}\int_a^{\xi'} 6y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' \\
&= \tau^3 y^{(1)}(a+\tau) - 6\tau^2 y(a+\tau) + \int_a^{a+\tau} 18(\xi'-a)y(\xi')\,\mathrm{d}\xi' - \int_a^{a+\tau}\int_a^{\xi'} 6y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'
\end{aligned} \tag{2.19}$$

As shown earlier, the upper limit again becomes a 'dummy variable' and now we set $\xi'' =$

$a+\tau$. Integrating again for the third time we get,

$$\begin{aligned}
\int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'}(\xi-a)^3y^{(3)}(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi'' &= \int_a^{a+\tau}(\xi''-a)^3y^{(1)}(\xi'')\,\mathrm{d}\xi'' - \int_a^{a+\tau}6(\xi''-a)^2y(\xi'')\,\mathrm{d}\xi'' \\
&\quad + \int_a^{a+\tau}\int_a^{\xi''}18(\xi'-a)y(\xi')\,\mathrm{d}\xi'\mathrm{d}\xi'' - \int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'}6y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi'' \\
&= \tau^3y(a+\tau) - \int_a^{a+\tau}3(\xi''-a)^2y(\xi'')\,\mathrm{d}\xi'' - \int_a^{a+\tau}6(\xi''-a)^2y(\xi'')\,\mathrm{d}\xi'' \\
&\quad + \int_a^{a+\tau}\int_a^{\xi''}18(\xi'-a)y(\xi')\,\mathrm{d}\xi'\mathrm{d}\xi'' - \int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'}6y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi'' \\
&= \tau^3y(a+\tau) - \int_a^{a+\tau}9(\xi''-a)^2y(\xi'')\,\mathrm{d}\xi'' \\
&\quad + \int_a^{a+\tau}\int_a^{\xi''}18(\xi'-a)y(\xi')\,\mathrm{d}\xi'\mathrm{d}\xi'' - \int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'}6y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi''
\end{aligned} \tag{2.20}$$

Integrating the second term in (2.16) first time,

$$\begin{aligned}
\int_a^{a+\tau}a_2(\xi-a)^3y^{(2)}(\xi)\,\mathrm{d}\xi &= a_2(\xi-a)^3y^{(1)}(\xi)\mid_a^{a+\tau} - \int_a^{a+\tau}3a_2(\xi-a)^2y^{(1)}(\xi)\,\mathrm{d}\xi \\
&= a_2\tau^3y^{(1)}(a+\tau) - \left[3a_2(\xi-a)^2y(\xi)\mid_a^{a+\tau} - \int_a^{a+\tau}6a_2(\xi-a)y(\xi)\,\mathrm{d}\xi\right] \\
&= a_2\tau^3y^{(1)}(a+\tau) - 3a_2\tau^2y(a+\tau) + \int_a^{a+\tau}6a_2(\xi-a)y(\xi)\,\mathrm{d}\xi
\end{aligned} \tag{2.21}$$

Using the guidelines as shown previously, we introduce a 'dummy variable' again and

integrating (2.21) again,

$$\begin{aligned}\int_a^{a+\tau}\int_a^{\xi'} a_2(\xi-a)^3y^{(2)}(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' &= \int_a^{a+\tau} a_2(\xi'-a)^3y^{(1)}(\xi')\,\mathrm{d}\xi' - \int_a^{a+\tau} 3a_2(\xi'-a)^2y(\xi')\,\mathrm{d}\xi' \\ &\quad + \int_a^{a+\tau}\int_a^{\xi'} 6a_2(\xi-a)y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' \\ &= a_2(\xi'-a)^3y(\xi')\,|_a^{a+\tau} - \int_a^{a+\tau} 3a_2(\xi'-a)^2y(\xi')\,\mathrm{d}\xi' \\ &\quad - \int_a^{a+\tau} 3a_2(\xi'-a)^2y(\xi')\mathrm{d}\xi' + \int_a^{a+\tau}\int_a^{\xi'} 6a_2(\xi-a)y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' \\ &= \tau^3a_2(a+\tau) - \int_a^{a+\tau} 6a_2(\xi'-a)^2y(\xi')\,\mathrm{d}\xi' \\ &\quad + \int_a^{a+\tau}\int_a^{\xi'} 6a_2(\xi-a)y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' \end{aligned} \tag{2.22}$$

Integrating the third time we have,

$$\begin{aligned}\int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'} a_2(\xi-a)^3y^{(2)}(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi'' &= \int_a^{a+\tau} a_2(\xi''-a)^3y(\xi'')\,\mathrm{d}\xi'' - \int_a^{a+\tau}\int_a^{\xi''} 6a_2(\xi'-a)^2y(\xi')\,\mathrm{d}\xi'\mathrm{d}\xi'' \\ &\quad + \int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'} 6a_2(\xi-a)y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi'' \end{aligned} \tag{2.23}$$

Integrating the third term in (2.16) first time,

$$\begin{aligned}\int_a^{a+\tau} a_1(\xi-a)^3y^{(1)}(\xi)\,\mathrm{d}\xi &= a_1(\xi-a)^3y(\xi)\,|_a^{a+\tau} - \int_a^{a+\tau} 3a_1(\xi-a)^2y(\xi)\,\mathrm{d}\xi \\ &= a_1\tau^3y(a+\tau) - \int_a^{a+\tau} 3a_1(\xi-a)^2y(\xi)\,\mathrm{d}\xi \end{aligned} \tag{2.24}$$

Integrating (2.24) second time yields,

$$\int_a^{a+\tau}\int_a^{\xi'} a_1(\xi-a)^3 y^{(1)}(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' = \int_a^{a+\tau} a_1(\xi'-a)^3 y(\xi')\,\mathrm{d}\xi' - \int_a^{a+\tau}\int_a^{\xi'} 3a_1(\xi-a)^2 y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi' \tag{2.25}$$

Integrating (2.25) the last time,

$$\begin{aligned}\int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'} a_1(\xi-a)^3 y^{(1)}(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi'' = &\int_a^{a+\tau}\int_a^{\xi''} a_1(\xi'-a)^3 y(\xi')\,\mathrm{d}\xi' \\ &- \int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'} 3a_1(\xi-a)^2 y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi''\end{aligned} \tag{2.26}$$

Finally, integrating the last term thrice, we get

$$\int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'} a_0(\xi-a)^3 y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi'' \tag{2.27}$$

Collecting the terms in (2.20) to (2.27) we have,

$$\begin{aligned}-\tau^3 y(a+\tau) = &\int_a^{a+\tau}\Big[-9(\xi''-a)^2 + a_2(\xi''-a)^3\Big]y(\xi'')\,\mathrm{d}\xi'' \\ &+ \int_a^{a+\tau}\int_a^{\xi''}\Big[+18(\xi'-a) - 6a_2(\xi'-a)^2 + a_1(\xi'-a)^3\Big]y(\xi')\,\mathrm{d}\xi'\mathrm{d}\xi'' \\ &+ \int_a^{a+\tau}\int_a^{\xi''}\int_a^{\xi'}\Big[-6 + 6a_2(\xi-a) - 3a_1(\xi-a)^2 + a_0(\xi-a)^3\Big]y(\xi)\,\mathrm{d}\xi\mathrm{d}\xi'\mathrm{d}\xi''\end{aligned} \tag{2.28}$$

This can be further simplified by recalling Cauchy formula for repeated integration. Let f

be a continuous function on the real line, then the nth repeated integral of f based at a.

$$f^{(-n)}(x) = \int_a^x \int_a^{\sigma_1} \cdots \int_a^{\sigma_{n-1}} f(\sigma_n) \mathrm{d}\sigma_n \cdots \mathrm{d}\sigma_2 \mathrm{d}\sigma_1 \tag{2.29}$$

is given by single integration

$$f^{(-n)}(x) = \frac{1}{(n-1)!} \int_a^x (x-t)^{n-1} f(t) \mathrm{d}t \tag{2.30}$$

Now we apply the Cauchy formula for repeated integration stated above on (2.28) while letting $a + \tau = t$, getting $\tau = t - a$. Now we get,

$$\begin{aligned} -(t-a)^3 y(t) = & \int_a^t \Big[-9(\tau-a)^2 + a_2(\tau-a)^3 \Big] y(\tau)\, \mathrm{d}\tau \\ & + \int_a^t (t-\tau) \Big[18(\tau-a) - 6a_2(\tau-a)^2 + a_1(\tau-a)^3 \Big] y(\tau)\, \mathrm{d}\tau \\ & + \frac{1}{2} \int_a^t (t-\tau)^2 \Big[-6 + 6a_2(\tau-a) - 3a_1(\tau-a)^2 + a_0(\tau-a)^3 \Big] y(\tau)\, \mathrm{d}\tau \\ \triangleq & \int_a^t K_F(t,\tau) y(\tau)\, \mathrm{d}\tau \end{aligned} \tag{2.31}$$

with $K_F(t,\tau)$ defined by,

$$\begin{aligned} K_F(t,\tau) \triangleq & \Big[-9(\tau-a)^2 + a_2(\tau-a)^3 \Big] + (t-\tau) \Big[18(\tau-a) - 6a_2(\tau-a)^2 + a_1(\tau-a)^3 \Big] \\ & + (t-\tau)^2 \Big[-6 + 6a_2(\tau-a) - 3a_1(\tau-a)^2 + a_0(\tau-a)^3 \Big] \end{aligned} \tag{2.32}$$

We now turn our attention to the equation (2.17). We start by integrating the first term

in (2.17) first time,

$$\begin{aligned}\int_{b-\sigma}^{b}(b-\zeta)^3y^{(3)}(\zeta)\,\mathrm{d}\zeta &= (b-\zeta)^3y^{(2)}(\zeta)\mid_{b-\sigma}^{b}+\int_{b-\sigma}^{b}3(b-\zeta)^2y^{(2)}(\zeta)\,\mathrm{d}\zeta\\ &= -\sigma^3y^{(2)}(b-\sigma)+\left[3(b-\zeta)^2y^{(1)}(\zeta)\mid_{b-\sigma}^{b}+\int_{b-\sigma}^{b}6(b-\zeta)y^{(1)}(\zeta),\mathrm{d}\zeta\right]\\ &= -\sigma^3y^{(2)}(b-\sigma)-3\sigma^2y^{(1)}(b-\sigma)+6(b-\zeta)y^{(1)}(\zeta)\mid_{b-\sigma}^{b}+\int_{b-\sigma}^{b}6y(\zeta)\,\mathrm{d}\zeta\\ &= -\sigma^3y^{(2)}(b-\sigma)-3\sigma^2y^{(1)}(b-\sigma)-6\sigma y(b-\sigma)+\int_{b-\sigma}^{b}6y(\zeta)\,\mathrm{d}\zeta\end{aligned} \tag{2.33}$$

When integrating again the upper limit on the integral becomes a 'dummy variable' as shown previously, i.e. we set $\zeta' = b-\sigma$ then,
$\sigma^3y^{(2)}(b-\sigma)$ is integrated as $(b-\zeta')^3y^{(2)}(\zeta')$
$3\sigma^2y^{(1)}(b-\sigma)$ is integrated as $3(b-\zeta')^2y^{(1)}(\zeta')$
$6\sigma y(b-\sigma)$ is integrated as $6(b-\zeta')y(\zeta')$.

We also flip the limits of the integration from $(\zeta' \longrightarrow b)$ to $-(b \longrightarrow \zeta')$ and hence a negative sign is introduced. And when we differentiate $(b-\zeta)^n$ with respect to ζ, we get $-n(b-\zeta)^{n-1}$, again a negative sign is introduced. We have to keep these two critical concepts in mind when evaluating the second and third integrals.

Integrating (2.33) again,

$$\begin{aligned}
\int_{b-\sigma}^{b}\int_{b}^{\zeta'}(b-\zeta)^3y^{(3)}(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' &= \int_{b-\sigma}^{b}(b-\zeta')^3y^{(2)}(\zeta')\,\mathrm{d}\zeta' + \int_{b-\sigma}^{b}3(b-\zeta')^2y^{(1)}(\zeta')\,\mathrm{d}\zeta' \\
&\quad + \int_{b-\sigma}^{b}6(b-\zeta')y(\zeta')\,\mathrm{d}\zeta' - \int_{b-\sigma}^{b}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \\
&= (b-\zeta')^3y^{(1)}(\zeta')\;|_{b-\sigma}^{b} + \int_{b-\sigma}^{b}3(b-\zeta')^2y^{(1)}(\zeta')\,\mathrm{d}\zeta' \\
&\quad + \left[3(b-\zeta')^2y(\zeta')\;|_{b-\sigma}^{b} + \int_{b-\sigma}^{b}6(b-\zeta')y(\zeta')\,\mathrm{d}\zeta'\right] \\
&\quad + \int_{b-\sigma}^{b}6(b-\zeta')y(\zeta')\,\mathrm{d}\zeta' + \int_{b}^{b-\sigma}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \\
&= -\sigma^3y^{(1)}(b-\sigma) + \left[3(b-\zeta')^2y(\zeta')\;|_{b-\sigma}^{b} + \int_{b-\sigma}^{b}6(b-\zeta')y(\zeta')\,\mathrm{d}\zeta'\right] \\
&\quad - 3\sigma^2y(b-\sigma) - \int_{b}^{b-\sigma}12(b-\zeta')y(\zeta')\,\mathrm{d}\zeta' + \int_{b}^{b-\sigma}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \\
&= -\sigma^3y^{(1)}(b-\sigma) + \left[-3\sigma^2y(b-\sigma) + \int_{b-\sigma}^{b}6(b-\zeta')y(\zeta')\,\mathrm{d}\zeta'\right] \\
&\quad - 3\sigma^2y(b-\sigma) - \int_{b}^{b-\sigma}12(b-\zeta')y(\zeta')\,\mathrm{d}\zeta' + \int_{b}^{b-\sigma}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \\
&= -\sigma^3y^{(1)}(b-\sigma) - 6\sigma^2y(b-\sigma) - \int_{b}^{b-\sigma}18(b-\zeta')y(\zeta')\,\mathrm{d}\zeta' \\
&\quad + \int_{b}^{b-\sigma}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \qquad (2.34)
\end{aligned}$$

As explained earlier, the upper limit becomes a 'dummy variable' and now we set $\zeta'' = b-\sigma$. Now integrating (2.34) third time,

$$\begin{aligned}
\int_{b-\sigma}^{b}\int_{b}^{\zeta''}\int_{b}^{\zeta'}(b-\zeta)^3 y^{(3)}(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' &= \int_{b-\sigma}^{b}(b-\zeta'')^3 y^{(1)}(\zeta'')\,\mathrm{d}\zeta'' + \int_{b-\sigma}^{b}6(b-\zeta'')^2 y(\zeta'')\,\mathrm{d}\zeta'' \\
&\quad - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}18(b-\zeta')y(\zeta')\,\mathrm{d}\zeta'\mathrm{d}\zeta'' - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' \\
&= \left[(b-\zeta'')^3 y(\zeta'')\mid_{b-\sigma}^{b} + \int_{b-\sigma}^{b}3(b-\zeta'')^2 y(\zeta'')\,\mathrm{d}\zeta''\right] \\
&\quad - \int_{b}^{b-\sigma}6(b-\zeta'')^2 y(\zeta'')\,\mathrm{d}\zeta'' - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}18(b-\zeta')y(\zeta')\,\mathrm{d}\zeta'\mathrm{d}\zeta'' \\
&\quad - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' \\
&= -\sigma^3 y(b-\sigma) - \int_{b}^{b-\sigma}9(b-\zeta'')^2 y(\zeta'')\,\mathrm{d}\zeta'' \\
&\quad - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}18(b-\zeta')y(\zeta')\,\mathrm{d}\zeta'\mathrm{d}\zeta'' - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}\int_{b}^{\zeta'}6y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta''
\end{aligned} \tag{2.35}$$

Integrating the second term in (2.17) first time,

$$\begin{aligned}\int_{b-\sigma}^{b} a_2(b-\zeta)^3 y^{(2)}(\zeta)\,\mathrm{d}\zeta &= a_2(b-\zeta)^3 y^{(1)}(\zeta)\mid_{b-\sigma}^{b} + \int_{b-\sigma}^{b} 3a_2(b-\zeta)^2 y^{(1)}(\zeta)\,\mathrm{d}\zeta \\ &= -a_2\sigma^3 y^{(1)}(b-\sigma) + \left[3a_2(b-\zeta)^2 y(\zeta)\mid_{b-\sigma}^{b} + \int_{b-\sigma}^{b} 6a_2(b-\zeta)y(\zeta)\,\mathrm{d}\zeta\right] \\ &= -a_2\sigma^3 y^{(1)}(b-\sigma) - 3a_2\sigma^2 y(b-\sigma) - \int_{b}^{b-\sigma} 6a_2(b-\zeta)y(\zeta)\,\mathrm{d}\zeta\end{aligned} \tag{2.36}$$

Using 'dummy variable' again and integrating (2.36) again,

$$\begin{aligned}\int_{b-\sigma}^{b}\int_{b}^{\zeta'} a_2(b-\zeta)^3 y^{(2)}(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' &= \int_{b-\sigma}^{b} a_2(b-\zeta')^3 y^{(1)}(\zeta')\,\mathrm{d}\zeta' + \int_{b-\sigma}^{b} 3a_2(b-\zeta')^2 y(\zeta')\,\mathrm{d}\zeta' \\ &\quad - \int_{b-\sigma}^{b}\int_{b}^{\zeta'} 6a_2(b-\sigma)y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \\ &= a_2(b-\zeta')^3 y(\zeta')\mid_{b-\sigma}^{b} + \int_{b-\sigma}^{b} 3a_2(b-\zeta')^2 y(\zeta')\,\mathrm{d}\zeta' \\ &\quad - \int_{b}^{b-\sigma} 3a_2(b-\zeta')^2 y(\zeta') - \int_{b-\sigma}^{b}\int_{b}^{\zeta'} 6a_2(b-\zeta)y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \\ &= -\sigma^3 a_2(b-\sigma) - \int_{b}^{b-\sigma} 6a_2(b-\zeta')^2 y(\zeta')\,\mathrm{d}\zeta' \\ &\quad - \int_{b-\sigma}^{b}\int_{b}^{\zeta'} 6a_2(b-\zeta)y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\end{aligned} \tag{2.37}$$

Integrating third time we have,

$$\begin{aligned}\int_{b-\sigma}^{b}\int_{b}^{\zeta''}\int_{b}^{\zeta'} a_2(b-\zeta-a)^3 y^{(2)}(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' &= \int_{b-\sigma}^{b} a_2(b-\zeta'')^3 y(\zeta'')\,\mathrm{d}\zeta'' \\ &\quad + \int_{b-\sigma}^{b}\int_{b}^{\zeta''} 6a_2(b-\zeta')^2 y(\zeta')\,\mathrm{d}\zeta'\mathrm{d}\zeta'' \\ &\quad + \int_{b-\sigma}^{b}\int_{b}^{\zeta''}\int_{b}^{\zeta'} 6a_2(b-\zeta) y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' \\ &= -\int_{b}^{b-\sigma} a_2(b-\zeta'')^3 y(\zeta'')\,\mathrm{d}\zeta'' \\ &\quad - \int_{b}^{b-\sigma}\int_{b}^{\zeta''} 6a_2(b-\zeta')^2 y(\zeta')\,\mathrm{d}\zeta'\mathrm{d}\zeta'' \\ &\quad - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}\int_{b}^{\zeta'} 6a_2(b-\zeta) y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta''\end{aligned} \tag{2.38}$$

Integrating the third term in (2.17) first time,

$$\begin{aligned}\int_{b-\sigma}^{b} a_1(b-\zeta)^3 y^{(1)}(\zeta)\,\mathrm{d}\zeta &= a_1(b-\zeta)^3 y(\zeta)\,|_{b-\sigma}^{b} + \int_{b-\sigma}^{b} 3a_1(b-\zeta)^2 y^{(2)}(\zeta)\,\mathrm{d}\zeta \\ &= -a_1\sigma^3 y(b-\sigma) - \int_{b}^{b-\sigma} 3a_1(b-\zeta)^2 y^{(2)}(\zeta)\,\mathrm{d}\zeta\end{aligned} \tag{2.39}$$

Integrating (2.39) second time,

$$\begin{aligned}\int_{b-\sigma}^{b}\int_{b}^{\zeta'} a_1(b-\zeta)^3 y^{(1)}(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' &= \int_{b-\sigma}^{b} a_1(b-\zeta)^3 y(\zeta')\,\mathrm{d}\zeta' - \int_{b-\sigma}^{b}\int_{b}^{\zeta'} 3a_1(b-\zeta)^2 y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta' \\ &= -\int_{b}^{b-\sigma} a_1(b-\zeta)^3 y(\zeta')\,\mathrm{d}\zeta' + \int_{b}^{b-\sigma}\int_{b}^{\zeta'} 3a_1(b-\zeta)^2 y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\end{aligned} \tag{2.40}$$

Integrating (2.40) the final time,

$$\begin{aligned}\int_{b-\sigma}^{b}\int_{b}^{\zeta''}\int_{b}^{\zeta'} a_1(b-\zeta)^3 y^{(1)}(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' &= \int_{b-\sigma}^{b}\int_{b}^{\zeta''} a_1(b-\zeta')^3 y(\zeta')\,\mathrm{d}\zeta' \\ &\quad + \int_{b-\sigma}^{b}\int_{b}^{\zeta''}\int_{b}^{\zeta'} 3a_1(b-\zeta)^2 y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' \\ &= -\int_{b}^{b-\sigma}\int_{b}^{\zeta''} a_1(b-\zeta')^3 y(\zeta')\,\mathrm{d}\zeta' \\ &\quad - \int_{b}^{b-\sigma}\int_{b}^{\zeta''}\int_{b}^{\zeta'} 3a_1(b-\zeta)^2 y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta''\end{aligned} \tag{2.41}$$

Integrating the last term in (2.17) thrice we have,

$$\int_{b-\sigma}^{b}\int_{b}^{\zeta''}\int_{b}^{\zeta'} a_0(b-\zeta)^3 y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' = -\int_{b}^{b-\sigma}\int_{b}^{\zeta''}\int_{b}^{\zeta'} a_0(b-\zeta)^3 y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta'' \tag{2.42}$$

Collecting the thrice integrated terms from (2.34) to (2.42):

$$\begin{aligned}\sigma^3 y(b-\sigma) = &\int_b^{b-\sigma} \Big[-9(b-\zeta'')^2 - a_2(b-\zeta'')^3 \Big] y(\zeta'')\,\mathrm{d}\zeta'' \\ &+ \int_b^{b-\sigma}\int_b^{\zeta''} \Big[-18(b-\zeta') - 6a_2(b-\zeta')^2 - a_1(b-\zeta')^3 \Big] y(\zeta')\,\mathrm{d}\zeta'\mathrm{d}\zeta'' \\ &+ \int_b^{b-\sigma}\int_b^{\zeta''}\int_b^{\zeta'} \Big[-6 - 6a_2(b-\zeta) - 3a_1(b-\zeta)^2 - a_0(b-\zeta)^3 \Big] y(\zeta)\,\mathrm{d}\zeta\mathrm{d}\zeta'\mathrm{d}\zeta''\end{aligned} \tag{2.43}$$

Now, applying the Cauchy formula for repeated integration stated earlier and simultaneously, while letting $b-\sigma = t$ and $\sigma = b-t$. We get:

$$\begin{aligned}(b-t)^3 y(t) = &\int_b^{t} \Big[-9(b-\sigma)^2 - a_2(b-\sigma)^3 \Big] y(\sigma)\,\mathrm{d}\sigma \\ &+ \int_b^{t} (t-\sigma) \Big[-18(b-\sigma) - 6a_2(b-\sigma)^2 - a_1(b-\sigma)^3 \Big] y(\sigma)\,\mathrm{d}\sigma \\ &+ \frac{1}{2}\int_b^{t} (t-\sigma)^2 \Big[-6 - 6a_2(b-\sigma) - 3a_1(b-\sigma)^2 - a_0(b-\sigma)^3 \Big] y(\sigma)\,\mathrm{d}\sigma\end{aligned} \tag{2.44}$$

Rewriting (2.44) after flipping the limits on the integrals,

$$\begin{aligned}(b-t)^3 y(t) = &-\int_t^b \Big[-9(b-\tau)^2 - a_2(b-\tau)^3\Big] y(\tau)\,\mathrm{d}\tau \\ &-\int_t^b (t-\tau)\Big[-18(b-\tau) - 6a_2(b-\tau)^2 - a_1(b-\tau)^3\Big] y(\tau)\,\mathrm{d}\tau \\ &-\frac{1}{2}\int_t^b (t-\tau)^2\Big[-6 - 6a_2(b-\tau) - 3a_1(b-\tau)^2 - a_0(b-\tau)^3\Big] y(\tau)\,\mathrm{d}\tau \\ \triangleq &\int_b^t K_B(t,\tau) y(\tau)\mathrm{d}\tau \end{aligned} \tag{2.45}$$

with $K_B(t,\tau)$ defined by,

$$\begin{aligned}K_B(t,\tau) \triangleq &\Big[9(b-\tau)^2 + a_2(b-\tau)^3\Big] \\ &+ (t-\tau)\Big[18(b-\tau) + 6a_2(b-\tau)^2 + a_1(b-\tau)^3\Big] \\ &+ \frac{1}{2}(t-\tau)^2\Big[6 + 6a_2(b-\tau) + 3a_1(b-\tau)^2 + a_0(b-\tau)^3\Big]\end{aligned} \tag{2.46}$$

Redefining the partial kernels as: '*forward*' & '*backward*'

$$\begin{aligned}K_F(t,\tau) \triangleq &\,\mu(\tau-a)\Big[9(\tau-a)^2 - a_2(\tau-a)^3\Big] \\ &+ (t-\tau)\Big[-18(\tau-a) + 6a_2(\tau-a)^2 - a_1(\tau-a)^3\Big] \\ &+ \frac{1}{2}(t-\tau)^2\Big[6 - 6a_2(\tau-a) + 3a_1(\tau-a)^2 - a_0(\tau-a)^3\Big]\end{aligned} \tag{2.47}$$

$$\begin{aligned} K_B(t,\tau) \triangleq \mu(b-\tau)&\Big[9(b-\tau)^2 + a_2(b-\tau)^3\Big] \\ &+ (t-\tau)\Big[18(b-\tau) + 6a_2(b-\tau)^2 + a_1(b-\tau)^3\Big] \\ &+ \frac{1}{2}\int_a^t (t-\tau)^2\Big[6 + 6a_2(b-\tau) + 3a_1(b-\tau)^2 + a_0(b-\tau)^3\Big] \end{aligned} \tag{2.48}$$

With,

$$\mu(\tau - a) = \begin{cases} 1 & : \tau \geq a \\ 0 & : \tau < a \end{cases}$$

and

$$\mu(b - \tau) = \begin{cases} 1 & : \tau \leq b \\ 0 & : \tau > b \end{cases}$$

Now (2.47) and (2.48) can be compactly written as :

$$(t-a)^3 y(t) = \int_a^t K_F(t,\tau) y(\tau)\,\mathrm{d}\tau \tag{2.49}$$

$$(b-t)^3 y(t) = \int_t^b K_B(t,\tau) y(\tau)\,\mathrm{d}\tau \tag{2.50}$$

Now we define:

$$K_{DS}(t,\tau) \triangleq \begin{cases} K_F(t,\tau) & : \tau \leq t \\ K_B(t,\tau) & : \tau > t \end{cases} \tag{2.51}$$

Combining (2.50) and (2.51) side by side while dividing both sides by $[(t-a)^3 + (b-t)^3]$ which is always greater than zero. This results in the following expression:

$$y(t) = \frac{1}{[(t-a)^3 + (b-t)^3]} \int_a^b K_{DS}(t,\tau) y(\tau)\,\mathrm{d}\tau \tag{2.52}$$

The recursive expressions of the derivatives can be derived by proceeding similarly as we used to derive the K_{DS}. To obtain the expression for $y^{(1)}(t)$ the equations (2.16) and

(2.17) are integrated twice. We now get,

$$\begin{aligned}(t-a)^3y^{(1)}(t) = 6(t-a)^2y(t) - a_2(t-a)^3y(t) \\ + \int_t^a \Big[-18(\tau-a) + 6a_2(\tau-a)^2 - a_1(\tau-a)^3\Big]y(\tau)\mathrm{d}\tau \\ + \int_t^a (t-\tau)\Big[6 - 6a_2(\tau-a)^2 - 3a_1(\tau-a)^2 - a_0(\tau-a)^3\Big]y(\tau)\mathrm{d}\tau\end{aligned} \tag{2.53}$$

and

$$\begin{aligned}(b-t)^3y^{(1)}(t) = -6(b-t)^2y(t) - a_2(b-t)^3y(t) \\ + \int_t^b \Big[18(b-\tau) + 6a_2(b-\tau)^2 - a_1(b-\tau)^3\Big]y(\tau)\mathrm{d}\tau \\ + \int_t^b (t-\tau)\Big[6 - 6a_2(b-\tau)^2 - 3a_1(b-\tau)^2 - a_0(b-\tau)^3\Big]y(\tau)\mathrm{d}\tau\end{aligned} \tag{2.54}$$

The final expression for $y^{(1)}(t)$ is obtained by adding the results of (2.53) and (2.54) and dividing by $[(t-a)^3 + (b-t)^3]$.

To get the expression for $y^{(2)}(t)$, (2.16) and (2.17) are integrated once. We get the following expression:

$$\begin{aligned}(t-a)^3y^{(2)}(t) = 3(t-a)^2y^{(1)}(t) - a_2(t-a)^3y^{(1)}(t) \\ - 6(t-a)y(t) + 3a_2(t-a)^2y(t) - a_1(t-a)^3y(t) \\ + \int_t^a \Big[6 - 6a_2(\tau-a) + 3a_1(\tau-a)^2 - a_0(\tau-a)^3\Big]y(\tau)\mathrm{d}\tau\end{aligned} \tag{2.55}$$

and

$$\begin{aligned}(b-t)^3y^{(2)}(t) = -3(b-t)^2y^{(1)}(t) - a_2(b-t)^3y^{(1)}(t) \\ - 6(b-t)y(t) - 3a_2(b-t)^2y(t) - a_1(b-t)^3y(t) \\ + \int_t^b \Big[6 + 6a_2(b-\tau) + 3a_1(b-\tau)^2 + a_0(b-\tau)^3\Big]y(\tau)\mathrm{d}\tau\end{aligned} \tag{2.56}$$

the expression for $y^{(2)}(t)$ is obtained by adding (2.55) and (2.56) while dividing by a factor of $[(t-a)^3 + (b-t)^3]$.

2.4 Model and State Estimation [3]

2.4.1 Joint Estimation Problem

Given a data cloud of time-tagged measurements, a linear model can be sought to best explain the local output behavior of a presumed dynamical system. For simplicity of exposition, consider single output, zero-input, LTI systems as candidates of such dynamical system models. The order of the model being fitted is not of critical importance as it can also be determined during the estimation process; see [25].

The goal of simultaneous model parameter and state estimation is then stated as follows:

Linear Simultaneous State and Parameter Estimation
Assuming an LTI SISO system model :$\dot{x} = Ax; y = Cx$ of order n, identify the values of the parameters $a_i, i = 0, .., n-1$, in its characteristic equation

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1 y^{(1)}(t) + a_0 y(t) = 0 \tag{2.57}$$

from a noisy system output measurement $\tilde{y}(t)$ over a finite, but arbitrary, interval of time $t \in [0, T]$, $T > 0$. No assumption is made about the noise characteristic. Also, provide the estimates of the noise-free output $y(t)$ and all output derivatives : $y^{(i)}(t), i = 1, \cdots, n-1$, for $t \in [0, T]$.

What makes this estimation problem different is the lack of any assumptions about the initial conditions of the system as well as its non-asymptotic nature - the observation interval is not only finite but, in principle, can be arbitrarily short. The left hand side of (2.57) is clearly a differential invariant for the system as it results from the validity of the Cayley-Hamilton's theorem, a quantity that remains invariant under the action of the flow of the system.

2.4.2 A Kernel Model Representation

The following kernel representation of the system model of Section 2.4.1 was introduced and derived in [12]:

Theorem 1. *There exist Hilbert-Schmidt kernels K_{DS}, K^i_{DS}, $i = 1, \cdots n-1$, such that the*

output function y in (2.57) *is reproduced on the interval* $[a, b]$ *in accordance with the action of the evaluation functional*

$$y(t) = \int_a^b K_{DS}(t, \tau) y(\tau) \, \mathrm{d}\tau \; ; \; \forall t \in [a, b] \tag{2.58}$$

and the derivatives of the output $y^{(1)}, \cdots y^{(n-1)}$ *can be computed recursively by way of output integration, so that for* $i = 1, \cdots n-1$ *and for all* $t \in [a, b]$*:*

$$y^{(i)}(t) = \sum_{k=0}^{i-1} b_k(t) y^{(k)}(t) + \int_a^b K_{DS}^i(t, \tau) y(\tau) \, \mathrm{d}\tau \tag{2.59}$$

where $y^{(0)} \equiv y$ *and* $b_k(\cdot)$ *are rational functions of* t*. Hilbert-Schmidt kernels are square integrable functions on* $L^2[a, b] \times L^2[a, b]$*.*

The kernel in (2.58) induces a reproducing kernel Hilbert space (RKHS) uniquely corresponding to the symmetric, positive-type kernel function
$K(t_1, t_2) \triangleq \langle K_{DS}(t_1, \cdot) | K_{DS}(t_2, \cdot) \rangle_2$
for all $t_1, t_2 \in [a, b]$ where $\langle \cdot | \cdot \rangle_2$ denotes the scalar product in $L^2[a, b]$. The RKHS, here denoted by $\mathcal{H}_K$, is then simply defined as the image of the space $L^2[a, b]$ under the integral transform defined by the double-sided kernel K_{DS} of (2.58) with a K-weighted norm as defined in [26]. The reproducing equality (2.58) has then yet another useful interpretation - that of a linear subspace of $\mathcal{H}_K$:

$$\mathcal{S}_I \triangleq \{ y \in \mathcal{H}_K \mid y \text{ satisfies (2.58) } \} \tag{2.60}$$

More importantly, by construction of the kernel, the behavioral system model of Theorem 1 is equivalent to the differential model as described by the invariance equation (2.57). This fact is stated as follows:

Corollary 1. *An output function* $y : [a, b] \to \mathbb{R}$ *satisfies the invariance equation (2.57) on the interval* $[a, b]$ *if an only if it is reproduced by the evaluation functional in (2.58).*

The proof of this fact is redundant as the multiple iterated integration in the derivation of the reproducing kernel can be reversed by multiple differentiation to retrieve the original invariance equation; see [2]. Clearly, at this point, the initial conditions of the

original system play no role as the behavior of the system is fully characterized in terms of its "trajectory behavior" over $[a, b]$. The filtering problem for the system output and its derivatives of any order then amounts to a "output trajectory reconstruction" that best fits the noisy measurement while preserving the system invariant.

2.4.3 Parameter Estimation and Practical Identifiability

Model parameter estimation, as described briefly below; see also [12], is obviously a component of the joint estimation approach as proposed here.

The proof of Theorem 1 ; see [12], shows that the kernel K_{DS} is in fact linear with respect to the system parameters $a_i, i = 0, \cdots, n-1$, i.e. one can write

$$y(t) = \int_a^b K_{DS}(t,\tau)y(\tau)\ \mathrm{d}\tau = \sum_{i=1}^{n} \tilde{a}_i g_i(t,y) \tag{2.61}$$

$$\text{where } g_i(t,y) \stackrel{def}{=} \int_a^b K_{DS}^{(i)}(t,\tau)y(\tau)\ \mathrm{d}\tau;\ t \in [a,b] \tag{2.62}$$

where the $K_{DS}^{(i)}; i = 1, \ldots, n$ are "component kernels" of K_{DS} and $\tilde{a}_i = a_{i-1}$ for notational convenience.

Given distinct time instants $t_1, \cdots, t_m \in (a, b]$, $m \geq n$, equation (2.61) is now re-written point-wise in the form of a linear algebraic system of equations

$$q(y) = P(y)a\ ;\ \text{mapping trajectories } y : [0,t] \to \mathbb{R} \tag{2.63}$$

$$\tag{2.64}$$

$$q(y) \stackrel{def}{=} \begin{bmatrix} y(t_1) \\ \vdots \\ y(t_m) \end{bmatrix}; a \stackrel{def}{=} \begin{bmatrix} \tilde{a}_1 \\ \vdots \\ \tilde{a}_n \end{bmatrix};$$

$$P(y) \stackrel{def}{=} \begin{bmatrix} g_1(t_1,y) \cdots g_n(t_1,y) \\ \ddots \\ g_1(t_m,y) \cdots g_n(t_m,y) \end{bmatrix} \tag{2.65}$$

As no assumptions are made about the noise which may determine the existence of solutions to the linear system (2.63), we give the following practical definition of linear identifiability.

2.4.4 Practical Linear Identifiability

Definition 1. The homogeneous system (2.57) is practically linearly identifiable on $[a, b]$ with respect to a particular realization of the output measurement, $y(t), t \in [a, b]$, if and only if there exist distinct time instants $t_1, \cdots, t_m \in (a, b]$ such that $\text{rank}P(y) = n$. By analogy with the nomenclature used in [27] output trajectories which render $\text{rank}P(y) = n$ will be called *persistent*.

In practical applications the m distinct time instants needed to generate (2.65) can be placed equidistantly over the interval $(a, b]$ or else generated randomly. Since no assumptions are made about system perturbations or measurement noise, the estimation equation (2.65) is best solved in terms of a pseudo-inverse $P^{\dagger}$ of P:

$$a = P^{\dagger}(y)q(y) \tag{2.66}$$

Finally, note that the output reproducing property, as written in (2.61), implies that the model representing subspace (2.60) has explicit linear expression in terms of the system model parameters. This fact is important in further development of the joint estimation approach.

2.5 Example

The third order LTI system as mentioned below is considered:

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -10 & 0 \end{bmatrix} x \; ; y = x_1 \; ; \; x(0) = [1, 1, 0] \tag{2.67}$$

with its corresponding characteristic equation

$$y^{(3)}(t) + a_2 y^{(2)}(t) + a_1 y^{(1)} + a_0 y(t) = 0 \tag{2.68}$$

All the parameters a_0, a_1, a_2 were assumed to be unknown, but the “nominal" reference values used for comparison are however given in Table 1 as True values.

The system output was first quantized using Lloyd’s algorithm. Gaussian noise of 30dB

SNR was added to the quantized signal and the so deformed "raw measurement data" was then used to compute initial estimates of the parameters a_0, a_1, a_2 as in section 2.4.3 . The initial estimation result is shown in Table 2.1. The quantized, noisy "raw measurement" is shown in Figure 2.1. Note that quantization is not necessary for parameter estimation.

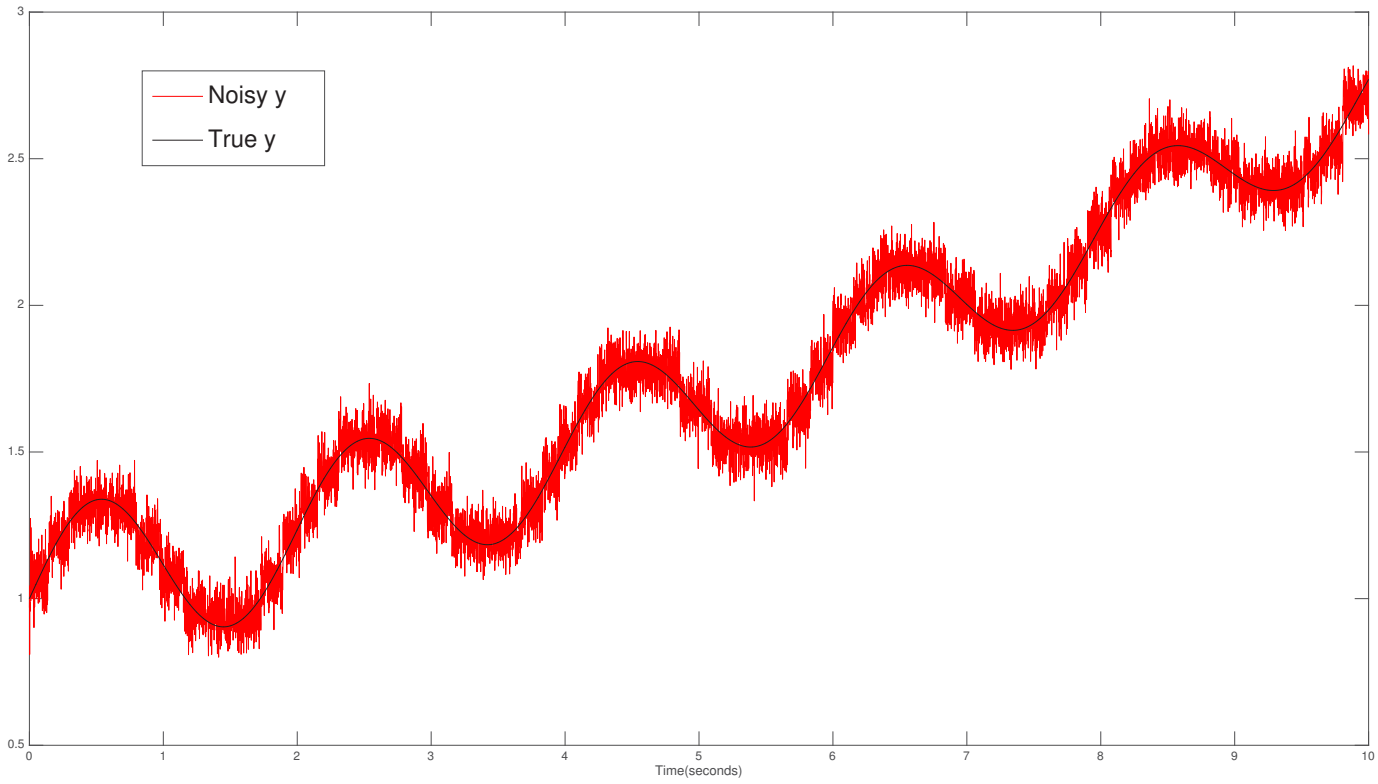

Figure 2.1 Quantized, noisy measurement vs. the "true" output signal

Comparing the estimated parameter values to their true counterparts could be very misleading (the absolute differences seem small as seen in Table 2.1).

When these estimates were subsequently used in the following kernels corresponding to

	a_0	a_1	a_2
True values	-1	10	0
Initial estimates	-1.0601	10.5965	0.0541

Table 2.1 Initial estimates for parameter values (prior to denoisification and model shaping).

the example system (2.67) via Theorem 1 (as derived before):

$$y(t) = \int_a^b K_{DS}(t,\tau)y(\tau)\,\mathrm{d}\tau \tag{2.69}$$

with

$$K_{DS}(t,\tau) \triangleq \frac{1}{[(t-a)^3+(b-t)^3]} \begin{cases} K_F(t,\tau) & : \tau \le t \\ K_B(t,\tau) & : \tau > t \end{cases} \tag{2.70}$$

$$\begin{aligned} K_F(t,\tau) \triangleq & \left[9(\tau-a)^2 - a_2(\tau-a)^3\right] \\ & + (t-\tau)\left[-18(\tau-a) + 6a_2(\tau-a)^2 - a_1(\tau-a)^3\right] \\ & + \frac{(t-\tau)^2}{2}\left[6 - 6a_2(\tau-a) + 3a_1(\tau-a)^2 - a_0(\tau-a)^3\right] \end{aligned} \tag{2.71}$$

$$\begin{aligned} K_B(t,\tau) \triangleq & \left[9(b-\tau)^2 + a_2(b-\tau)^3\right] \\ & + (t-\tau)\left[18(b-\tau) + 6a_2(b-\tau)^2 + a_1(b-\tau)^3\right] \\ & + \frac{(t-\tau)^2}{2}\left[6 + 6a_2(b-\tau) + 3a_1(b-\tau)^2 + a_0(b-\tau)^3\right] \end{aligned} \tag{2.72}$$

they failed to reproduce the system output correctly as proved by Figure 2.2. The need for further refinement of the initial estimates was then transparent.

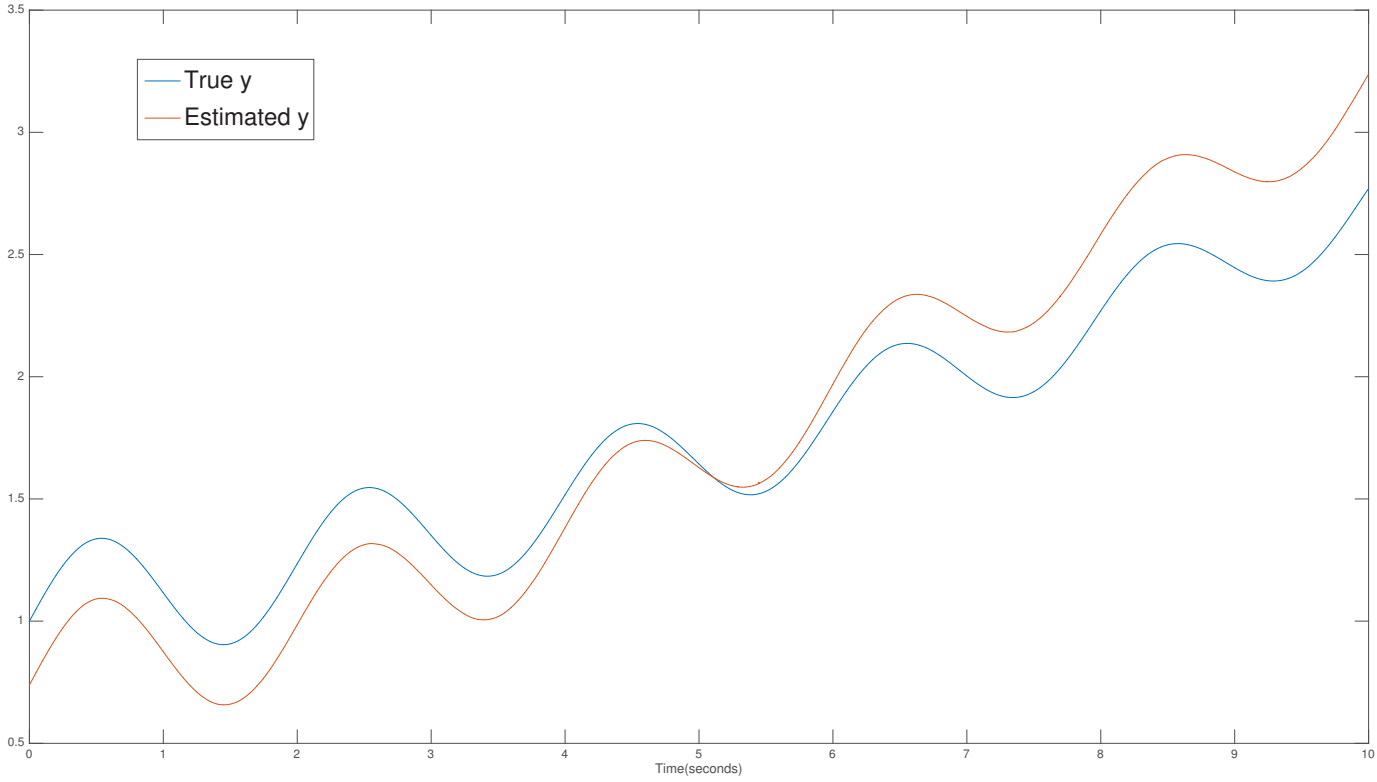

Figure 2.2 Comparison of the true and estimated outputs prior to denoisification and model shaping.

Chapter 3

Simultaneous Denoisification and Model Shaping

3.1 Formulation [3]

With essentially no information about the system model nor measurement noise the task of output differentiation requires a special treatment.

While the first step of the approach will use raw data to deliver the initial estimate of the system parameters as outlined in 2.4.3, the latter estimate must be further refined during an exhaustive search for an optimal orientation of the model representing subspace $\mathcal{H}$ of (2.60) to best fit the data cloud. Such a refinement process will be referred to as "simultaneous denoisification and model shaping". To pose this problem in a mathematical form, it is convenient to re-write (2.57) in state space form:

$$\dot{x} = A_K x \ ; \ y = Cx; \quad x \in \mathbb{R}^n \tag{3.1}$$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \stackrel{def}{=} \begin{bmatrix} y \\ \vdots \\ y^{(n-1)} \end{bmatrix} \ ; \ A_K \stackrel{def}{=} \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ 0 & 0 & 0 & \ddots \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{bmatrix}$$

$$K \stackrel{def}{=} \begin{bmatrix} k_1 & k_2 & \cdots & k_n \end{bmatrix} \stackrel{def}{=} \begin{bmatrix} -a_0 & -a_1 & \cdots & -a_{n-1} \end{bmatrix}$$

$$C \stackrel{def}{=} \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix};$$

with an unknown initial condition $x(0) = x_0$. The simultaneous denoisification and parameter estimation problem is now cast as an optimization problem

$$\min\{Q(K, x_0) \mid \dot{x} = A_K x \; ; x(0) = x_0; y = Cx\} \tag{3.2}$$

$$Q(K, x_0) \stackrel{def}{=} \frac{1}{2} \int_0^T (y - \tilde{y})^2 dt \tag{3.3}$$

where the cost Q represents the model residual error. As the initial condition is unknown the problem is equivalently re-formulated in terms of fundamental solutions of (3.1) as follows. Let, $x_i^e, i = 1, \cdots, n$ denote the solutions of the system equation $\dot{x} = A_K x$ corresponding to the initial conditions $x(0) = e_i$, respectively, where e_i are unit canonical basis vectors in $\mathbb{R}^n$. All solutions of $\dot{x} = A_K x$ then form an n-dimensional subspace $\mathcal{S}_x \stackrel{def}{=} \text{span}\{x_i^e; i = 1, \cdots, n\}$ i.e. any solution of the system equation (3.1) is a linear combination

$$x = \sum_{i=1}^{n} c_i x_i^e; \;\; c \stackrel{def}{=} \begin{bmatrix} c_1, \cdots, c_n \end{bmatrix} \tag{3.4}$$

and (3.2) - (3.3) takes a form of a finite dimensional optimization problem with equality constraints which is, however, free of any initial conditions

$$\min\{J(K, c) \mid \dot{x}_i^e = A_K x_i^e; \;\; i = 1, \cdots, n\} \tag{3.5}$$

$$J(K, c) \stackrel{def}{=} \frac{1}{2} \int_0^T (C \sum_{i=1}^{n} c_i x_i^e - \tilde{y})^2 dt \tag{3.6}$$

$$J : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

It is now easy to see that *for any fixed value* of the vector K the minimum of J is attained for a vector $\hat{c} \in \mathbb{R}^n$ which renders the orthogonal projection

$$\hat{y} = \sum_{i=1}^{n} \hat{c}_i y_i^e; \quad y_i^e \stackrel{def}{=} C x_i^e \tag{3.7}$$

of the output measurement function $\tilde{y}(\cdot)$ onto the n-dimensional subspace

$$\mathcal{S}_y \stackrel{def}{=} \text{span}\{y_i^e; i = 1, \cdots, n\} \subset L^2[0, T]; \tag{3.8}$$

so that

$$(\hat{y} - \tilde{y}) \perp \mathcal{S}_y \tag{3.9}$$

$$\text{i.e. } (\sum_{i=1}^{n} \hat{c}_i y_i^e - \tilde{y}) \perp y_k^e \text{ for all } k = 1, \cdots, n \tag{3.10}$$

$$\implies \hat{c} = G(y_i^e)^{-1} h(\tilde{y}, y_i^e); \tag{3.11}$$

$$G(y_i^e) \stackrel{def}{=} \begin{bmatrix} (y_1^e|y_1^e) & \cdots & (y_n^e|y_1^e) \\ (y_1^e|y_2^e) & \cdots & \cdots \\ \cdots & \ddots & \cdots \\ (y_1^e|y_n^e) & \cdots & (y_n^e|y_n^e) \end{bmatrix}; h(\tilde{y}, y_i^e) \stackrel{def}{=} \begin{bmatrix} (\tilde{y}|y_1^e) \\ \vdots \\ (\tilde{y}|y_n^e) \end{bmatrix}$$

where the Gram matrix $G(y_i^e)$ is invertible as the fundamental set spanning $\mathcal{S}_y$ is linearly independent.

Remark 1. *Note that if the system parameters in vector K are known then equations (3.7) and (3.11) deliver a denoisified measurement corresponding to $\tilde{y}$ in the form of $y(t) = \hat{y}(t), t \in [0, T]$.*

Importantly, the optimization problem (3.5) - (3.6) has infinitely many solutions if the noisy measurement is identically zero (i.e. when $\tilde{y} \equiv 0$). This is because then the orthogonal projection (3.7) is also identically zero, i.e. $\hat{y} \equiv 0$. Consequently, $\hat{y}^{(i)} \equiv 0$ for $i = 1, \cdots, n-1$, so the system homogeneous equation (2.57) is satisfied for any set of parameters $a_i; i = 0, \cdots, n-1$. The system (2.57) is then non-identifiable. As such a situation is unlikely to occur the minima of (3.5) - (3.6) can be sought by iterative search direction optimization methods. Computation of such search directions requires evaluation of the gradient of the cost index J at any given parameter pair (K, c). This is a complex task as the computation of $\frac{\partial J}{\partial K}(\bar{K}, \bar{c})$ at any given point $(\bar{K}, \bar{c})$ requires calculation of the sensitivities of the fundamental solutions of $\dot{x}_i^e = A_{\bar{K}} x_i^e$ with respect to variation of the parameters in $\bar{K}$. The solutions x_i^e depend on the parameters K non-linearly and a direct forward calculation of sensitivities (via the solution of the respective variational equations; [28]) is not helpful especially that

the sensitivities enter not only the cost index J but also the constraints. The adjoint method is used instead as explained next.

3.1.1 Necessary Conditions of Optimality and Gradient Computation

The adjoint gradient method for dynamical systems was provided rigorously in [29] and less formally in [30]. It will be explained here with strict reference to our particular problem.

The gradient of the cost index in the optimization problem (3.5) - (3.6) is efficiently computed in a dual approach by incorporating a Lagrangian function:

$$\mathcal{L}(K,c) \stackrel{def}{=} \frac{1}{2}\int_0^T (C\sum_{i=1}^n c_i x_i^e - \tilde{y})^2 dt + \int_0^T \sum_{i=1}^n c_i \lambda_i^T (\dot{x}_i^e - A_K x_i^e) dt \tag{3.12}$$

where the Lagrangian multipliers are *linearly independent vector functions* $\lambda^T : [0,T] \to \mathbb{R}^n$. The Lagrange multipliers constrain the dynamical system to variations around the path of $\dot{x}_i^e = A_K x_i^e$. Clearly, once the constraint equation $\dot{x}_i^e = A_K x_i^e$ is satisfied the Lagrange multiplier term will disappear leaving

$$\frac{\partial \mathcal{L}(K,c)}{\partial k_j} = \frac{\partial J(K,c)}{\partial k_j};\ j = 1, \cdots, n \tag{3.13}$$

Therefore, in the dual representation of the constrained problem the contribution of the derivative $\frac{dx_i^e}{dK}$ to the gradient $\frac{d\mathcal{L}}{dK}$ can be "annihilated" by choosing $\lambda^T(\cdot)$ in a specific way (in effect circumventing the need for explicit forward calculation of $\frac{dx_i^e}{dK}$). Such adjoint procedure is explained as follows. Calculating the derivatives with respect to the components of K yields

$$\frac{\partial \mathcal{L}(K,c)}{\partial k_j} = \int_0^T (C\sum_{l=1}^n c_l x_l^e - \tilde{y})^T C \sum_{i=1}^n c_i \frac{dx_i^e}{dk_j} dt + \int_0^T \sum_{i=1}^n c_i \lambda_i^T (\frac{d\dot{x}_i^e}{dk_j} - A_K \frac{dx_i^e}{dk_j} - \frac{dA_K}{dk_j} x_i^e) dt \tag{3.14}$$

for all $k_j, j = 1, \cdots, n$, where explicit evaluation gives the sensitivity of the matrix A_K with respect to parameter k_j

$$\frac{dA_K}{dk_j} = I_K;\ I_K \stackrel{def}{=} \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{3.15}$$

where $I_K \in \mathbb{R}^n \times \mathbb{R}^n$ is a matrix with a single non-zero element in the last row and k-th column : $I_K^{(k,n)} = 1$. Using integration by parts allows to write

$$\int_0^T \lambda_i^T \frac{d\dot{x}_i^e}{dk_j} dt = \lambda_i^T \frac{dx_i^e}{dk_j}\Big|_0^T - \int_0^T \frac{d\lambda_i^T}{dt} \frac{dx_i^e}{dk_j} dt \tag{3.16}$$

Since the Lagrangian multiplier functions have no effect on the value of the Lagrangian when the constraint is satisfied; see (3.13), and the initial conditions needed to compute the fundamental solutions $x_i^e; i = 1, \cdots n$ do not depend on the parameters k_j then we can set, without the loss of generality

$$\lambda_i^T(T) = 0; \quad \frac{dx_i^e}{dk_j}(0) = 0; \quad i, j = 1, \cdots, n \tag{3.17}$$

for an immediate simplification

$$\int_0^T \lambda_i^T \frac{d\dot{x}_i^e}{dk_j} dt = -\int_0^T \frac{d\lambda_i^T}{dt} \frac{dx_i^e}{dk_j} dt \tag{3.18}$$

Substituting (3.15) and (3.18) into (3.14) yields

$$\frac{\partial \mathcal{L}(K, c)}{\partial k_j} = \int_0^T \sum_{i=1}^n c_i \left[(C \sum_{l=1}^n c_l x_l^e - \tilde{y})^T C \frac{dx_i^e}{dk_j} \right. \tag{3.19}$$

$$\left. - \dot{\lambda}_i^T \frac{dx_i^e}{dk_j} - \lambda_i^T A_K \frac{dx_i^e}{dk_j} \right] dt - \int_0^T \sum_{i=1}^n c_i \lambda_i^T I_j x_i^e dt \tag{3.20}$$

It is now time to select a “trajectory" for the Lagrange multipliers that best serves the goal of removing excessive computational burden by imposing that

$$\dot{\lambda}_i^T = -\lambda_i^T A_K + (C\sum_{l=1}^{n} c_l x_l^e - \tilde{y})^T C \tag{3.21}$$
$$\lambda_i^T(T) = 0; \quad i = 1, \cdots, n$$

The remaining term of (3.20) then constitutes the essential part of the gradient with respect to the parameters in K:

$$\frac{\partial \mathcal{L}(K,c)}{\partial k_j} = -\int_0^T \sum_{i=1}^{n} c_i \lambda_i^T I_j x_i^e dt; j = 1, \cdots, n \tag{3.22}$$

It is also easy to see that the gradient of J with respect to the parameters c_i that replace the influence of the initial conditions is given by

$$\frac{\partial J(K,c)}{\partial c_i} = \int_0^T (C\sum_{l=1}^{n} c_l x_l^e - \tilde{y})^T C x_i^e dt; \ i = 1, \cdots, n \tag{3.23}$$

It is now possible to state:

Theorem 2. *The necessary conditions for optimality in the simultaneous state and parameter estimation optimization problem (3.5) - (3.6) are given by*

$$\frac{\partial J(K,c)}{\partial k_j} = -\int_0^T \sum_{i=1}^{n} c_i \lambda_i^T I_j x_i^e dt = 0; \tag{3.24}$$
$$\frac{\partial J(K,c)}{\partial c_i} = \int_0^T (C\sum_{l=1}^{n} c_l x_l^e - \tilde{y})^T C x_i^e dt = 0; \tag{3.25}$$
$$\dot{\lambda}_i^T = -\lambda_i^T A_K + (C\sum_{l=1}^{n} c_l x_l^e - \tilde{y})^T C; \tag{3.26}$$
$$\dot{x}_i^e = A_K x_i^e; \tag{3.27}$$
$$\lambda_i^T(T) = 0; \quad x_i^e(0) = e_i; \tag{3.28}$$
$$\textit{for all} \ \ i = 1, \cdots, n; \quad j = 1, \cdots, n$$

The above conditions can readily be used in a gradient-search direction based iterative

optimization procedure. For clarity, the gradient of the cost function with respect to parameters K at a nominal value $\bar{K}$ (say, a current value in iteration k), $\frac{\partial J(\bar{K},c)}{\partial K}$, is calculated via the following procedure:

Gradient Calculation

(1) For the current value of the parameter $\bar{K}$ solve the system equations (3.27) obtaining the corresponding set $\{x_i^e; i = 1, \cdots, n\}$ of fundamental solutions;

(2) Calculate the optimal coefficients $\hat{c}_i; i = 1, \cdots, n$ from equation (3.25) or, equivalently use formula (3.11);

(3) solve the adjoint equations (3.26) backwards in time as indicated by (3.28);

(4) Calculate the components of the gradient from the formula

$$\frac{\partial J(\bar{K}, \hat{c})}{\partial k_j} = -\int_0^T \sum_{i=1}^{n} \hat{c}_i \lambda_i^T I_j x_i^e dt; \ j = 1, \cdots, n$$

A steepest descent Armijo type of optimization algorithm can be applied directly emloying a gradient calculation outlined above, but will converge too slowly. We therefore suggest employing a quasi-Newton algorithm known under the name of Broyden's method that converges much faster (super-linearly). The Broyden algorithm is a version of a secant algorithm and is given here for completeness of exposition.

3.1.2 Broyden's Method for Simultaneous State and Parameter Estimation

Broyden's method is a type of secant algorithm which can be applied to our problem by solving the gradient equation $\frac{\partial J(K,\hat{c})}{\partial K} = 0$; see [31] for an improved version of the algorithm.

We adopt the following notation for the algorithm to be stated succinctly.

$$\begin{aligned} F(K) &\stackrel{def}{=} \left[\frac{\partial J(K,\hat{c})}{\partial k_1} \quad \cdots \quad \frac{\partial J(K,\hat{c})}{\partial k_n}\right]^T \\ F &: \mathbb{R}^n \to \mathbb{R}^n; \ F : K \mapsto F(K) \end{aligned} \tag{3.29}$$

Let K_0 denote an initial parameter for the solution of the vector equation $F(K) = 0$ and let B_0 denote an initial finite difference approximation of the Jacobian of F at K_0, also

denote

$$\begin{aligned} s_j &\stackrel{def}{=} K_{j+1} - K_j \\ d_j &\stackrel{def}{=} F(K_{j+1}) - F(K_j); \quad j = 1, 2, \cdot \end{aligned} \tag{3.30}$$

Broyden's Algorithm

For $j = 0, 1, 2, \cdots$ until convergence criterion is met

$$\begin{aligned} &\text{Solve } B_j s_j = -F(K_j) \text{ for } s_j \\ &\text{Set } K_{j+1} = K_j + s_j \\ &\text{Set } d_j = F(K_{j+1}) - F(K_j) \\ &\text{Update } B_{j+1} = B_j + (d_j - B_j s_j)\frac{s_j^T}{s_j^T s_j} \end{aligned}$$

It should be noted that the Broyden iteration need to be performed only with respect to the parameter K_j because the update/calculation of the corresponding optimal parameter $\hat{c}_j$ is done as part of the Gradient Calculation (i.e. the calculation of $F(K_j)$). No step size needs to be calculated, but invertibility of the Jacobian approximation B_j is implied. If B_j is singular the method should be restarted. If the initial guess for the parameter value K_0 is not distant from the minimum the Jacobian will be non-singular by local convexity of the cost index.

As already pointed out, convergence of any iterative optimization algorithm benefits from an good initial choice of the optimization parameter. As already suggested, initial values for the system parameters K can be obtained from raw data by application of the parameter estimation procedure of 2.4.3.

The latter can be used in terms of a starting parameter in Broyden's algorithm. Once satisfactory precision is achieved during the iterative search for the minimum in (3.5) - (3.6) the algorithm is exited. The latest values of the iterated parameters $\hat{c}$ and $\hat{K}$ are then used in the next stage of the estimation process in which flatness-based estimates of the system states requires prior estimation of all the output derivatives. Their calculation is described next.

3.1.3 Estimation of Output Derivatives and System States

Estimation of the output $y(t)$ is immediately obtained in terms of the optimized values $\hat{c}$, $\hat{K}$, as found at the exit from the optimization algorithm of the previous section ,i.e.

$$\hat{y}(t) = \sum_{i-1}^{n} \hat{c}_i \hat{x}_i^e(t);\ t \in [0, T] \tag{3.31}$$

where $\hat{x}_i^e; i = 1, \cdot, n$ are is the fundamental set of solutions corresponding to the optimal parameter $\hat{K}$. The estimate of the state of the system is then calculated in terms of all the derivatives of the output $y^{(1)}, \cdots y^{(n-1)}$ explicitly using the integral transform formula as delivered by Theorem 1 of 2.4.2; see (2.59) .

The advantage of using integral transforms to compute derivatives in place of conventional numerical differentiation methods should be clear as such an approach provides additional robustness with respect to computational noise.

Chapter 4

Results

Simultaneous denoisification and model shaping was subsequently applied to the example in 2.5 in an attempt to eliminate the drawbacks as depicted before. Two cases are presented below: in both the cases the output of the dynamical system in 2.5 is distorted by White Gaussian noise of known SNR.

Case 1: Additive White Gaussian Noise with 30dB SNR

Figure 4.1 shows the refinements of the output trajectory estimates in three iterations of the model shaping process as it converges through the use of a simple gradient descent method.

Figure 4.2 shows the estimated output trajectories obtained in five iterations of the model shaping process as it was forced to start from arbitrarily chosen parameter values $a_0 = 1, a_1 = 1, a_2 = 1$ while omitting the initial parameter estimation step of section 2.4.3. The parameters are clearly far away from both the true and the initial estimates of Table 2.1. Still, the model shaping process converges very well to the true output trajectory of the system. This test condition is important to assess the robustness of the method developed in the previous chapter.

Figures 4.3, 4.4 and 4.5 show the y, $y^{(1)}$, $y^{(2)}$ estimates after convergence was achieved in denoisification and model shaping. These estimates are the same in both cases: with and without prior parameter estimation from raw data.

Figure 4.6 shows that the system parameters estimated after initial parameter estimation, as

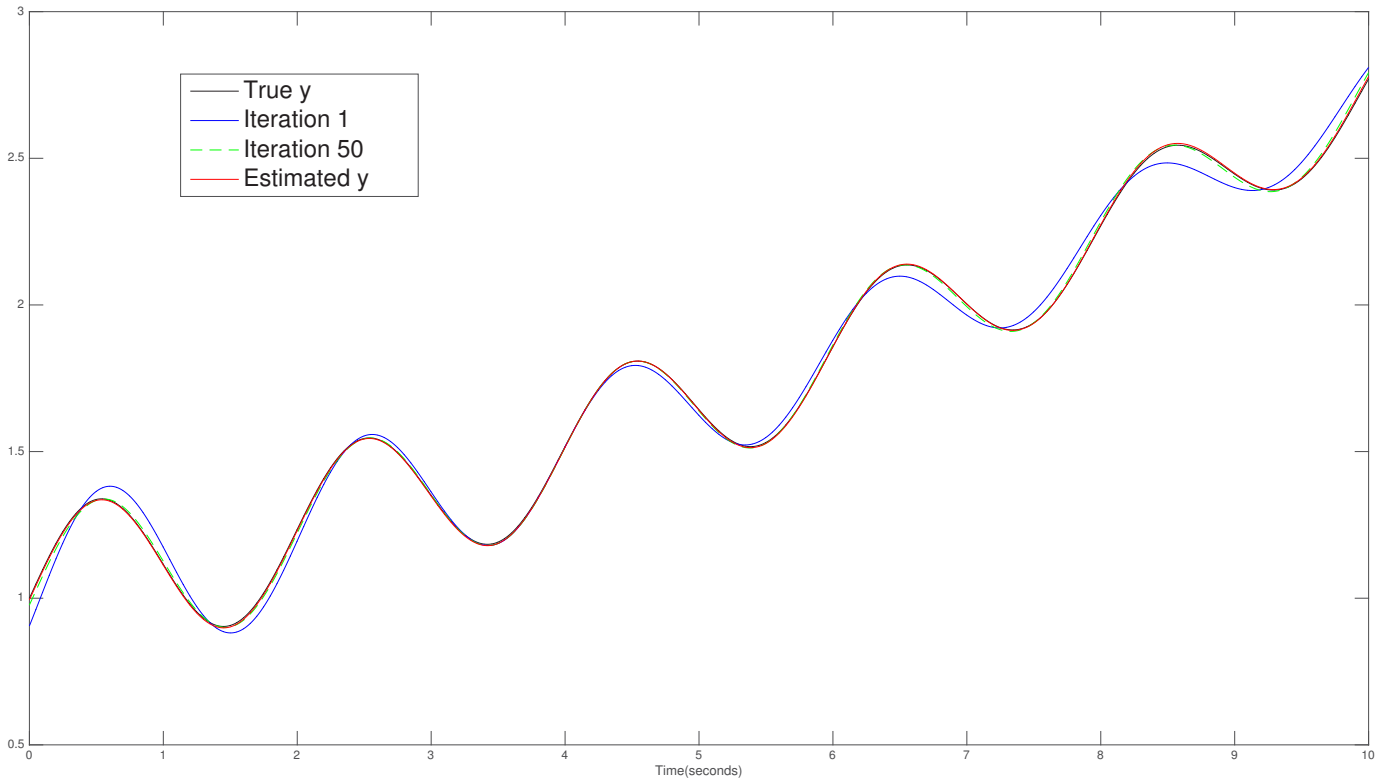

Figure 4.1 Model shaping process gradually converging as preceded by initial parameter estimation. Convergence is shown in terms of output trajectories.

mentioned in 2.4.3, undergo variation when subjected to denoisification and model shaping. A geometrical interpretation of this process is particularly appealing as it can be visualized as re-orientation of subspace containing estimated y by way of minimizing the residual of the modelling error (i.e. the cost functional in this case). The final estimated parameters after application of this method can be seen in Table 4.1 and the quantitative as well as qualitative difference between the values from that of Table 2.1 is thus ascertained.

It is not hard to notice that convergence of state trajectories is more accurate with the application of Simultaneous denoisification and model shaping. Comparisons can be made with relevant graphs in [1].

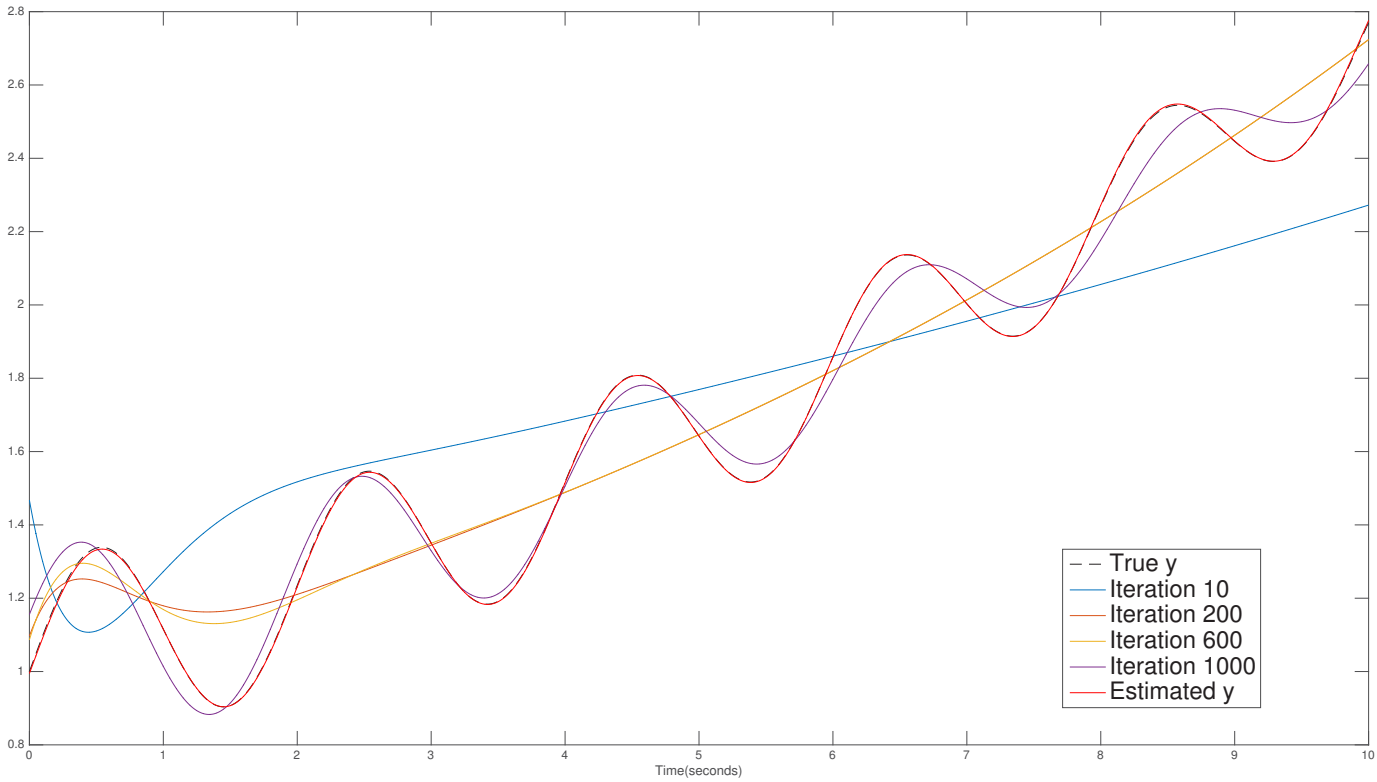

Figure 4.2 Model shaping process gradually converging while skipping initial parameter estimation. Convergence is shown in terms of output trajectories.

Case 2: Additive White Gaussian Noise with 10dB SNR

An additional case of 10dB SNR is considered in this work to highlight the high noise attenuation property of this algorithm, see Figure 4.7. Only the convergence of y estimate to the actual output of the dynamical system is presented here.

	a_0	a_1	a_2
True values	-1	10	0
Final estimates	-1.0009	9.9978	-0.0106

Table 4.1 Final estimates for parameter values (after denoisification and model shaping).

4.0.1 Kolmogorov Smirnov test

The Kolmogorov Smirnov test can be used to test whether two samples come from the same distributions. The KS statistic in this case is given by :

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

where $F_n(x)$ and $G_n(x)$ are empirical distribution functions of the two samples of size n and m respectively.
The null hypothesis, that data in the two samples comes from the same distribution, is rejected at significance level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$$

A built-in function for K-S test offered by a standard numerical computing software was used to perform the two-sample test employing the empirical samples from $\tilde{y} - y$ and $\tilde{y} - \hat{y}$, where $\tilde{y}$ is the noise signal, y is the true system trajectory, $\hat{y}$ is the estimated trajectory.

The null hypothesis was not rejected at a 5% significance level, indicating a good fit of the residual, confirming successful model shaping.

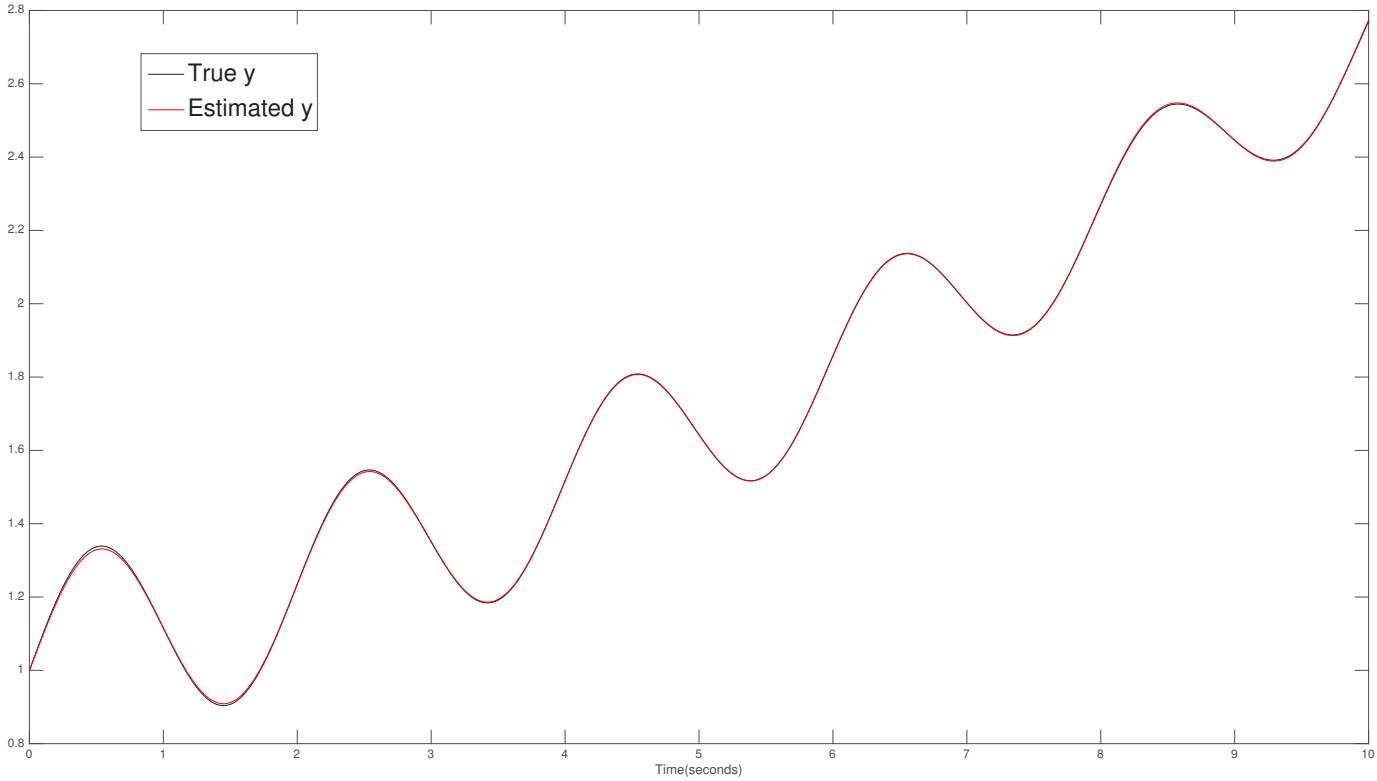

Figure 4.3 Estimated system output y after convergence is achieved.(30dB SNR)

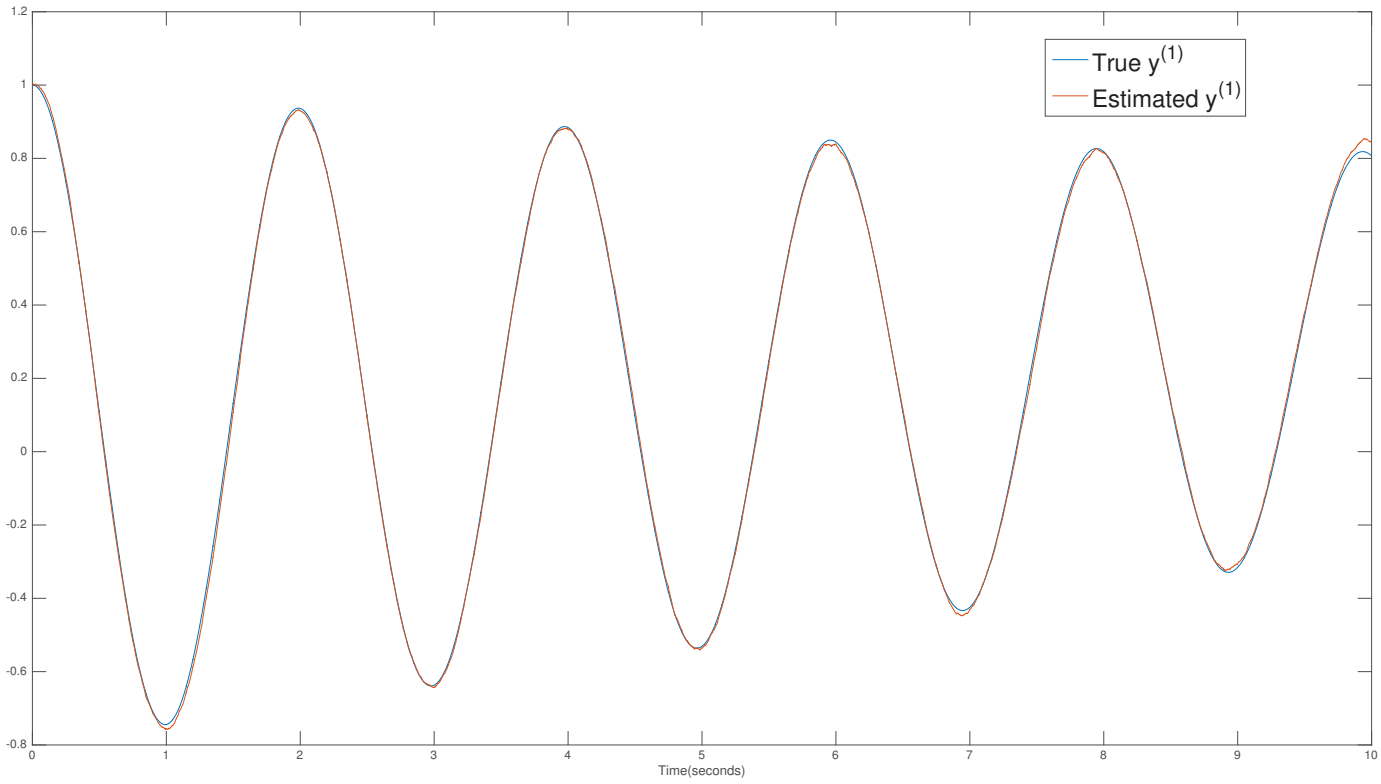

Figure 4.4 Estimated first derivative of the output $y^{(1)}$ after convergence is achieved.

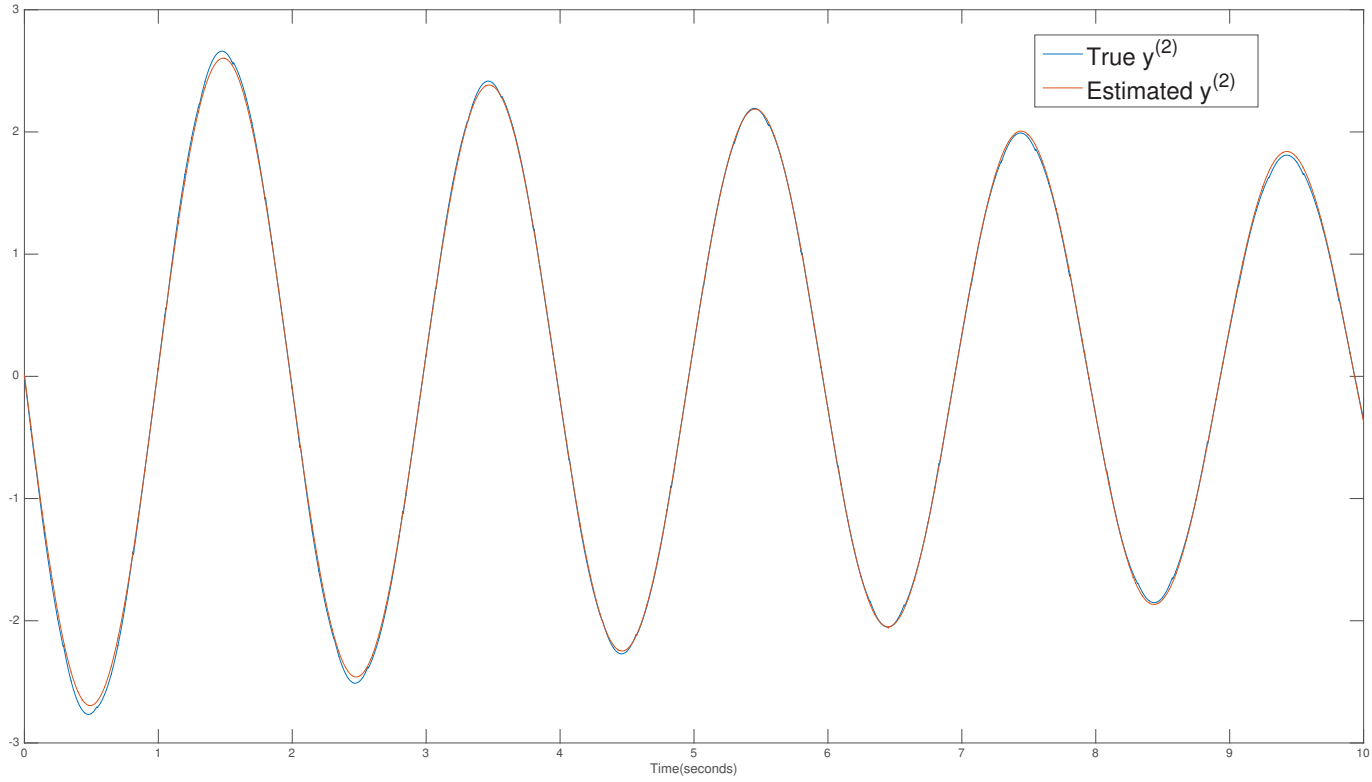

Figure 4.5 Estimated second derivative of the output $y^{(2)}$ after convergence is achieved.

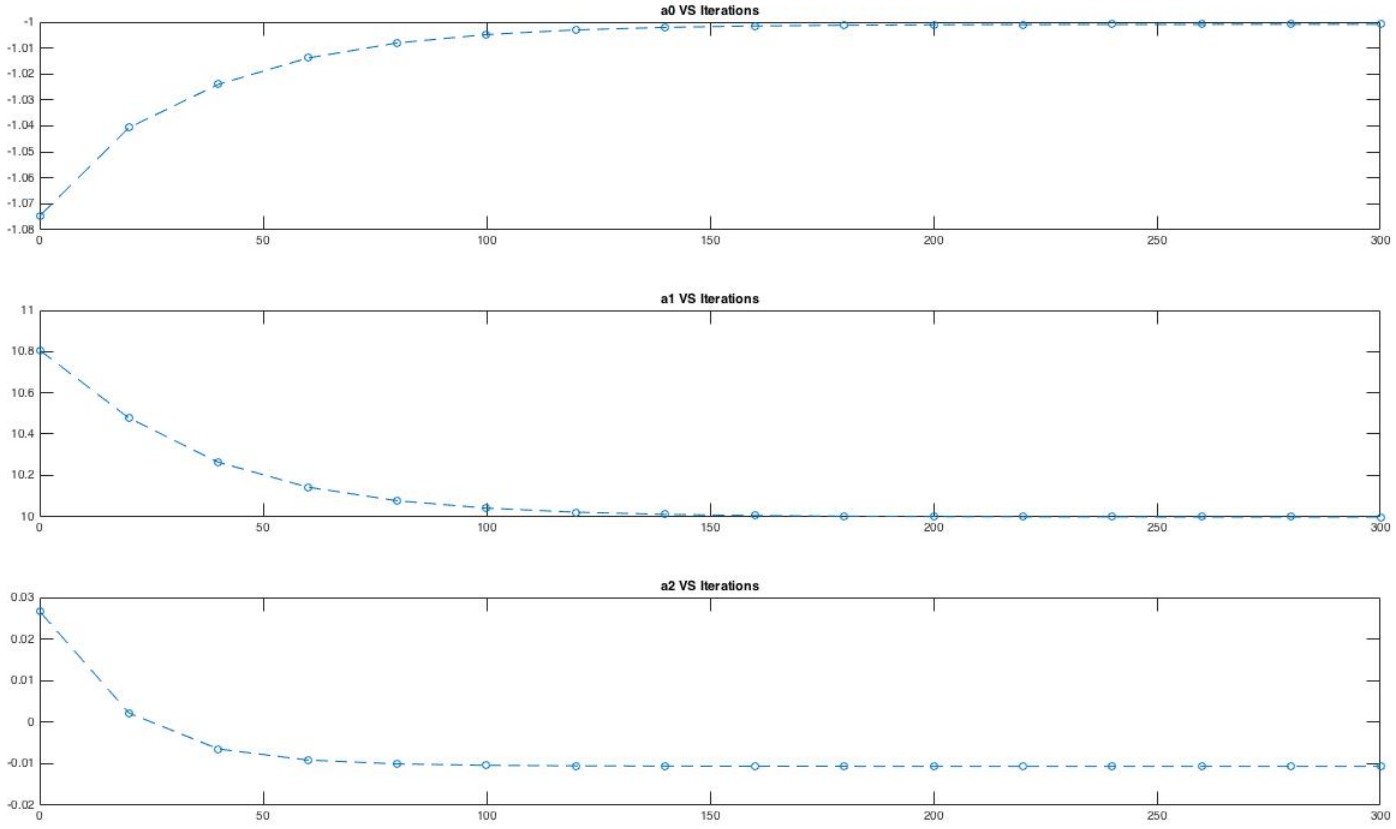

Figure 4.6 Variation of the system parameters

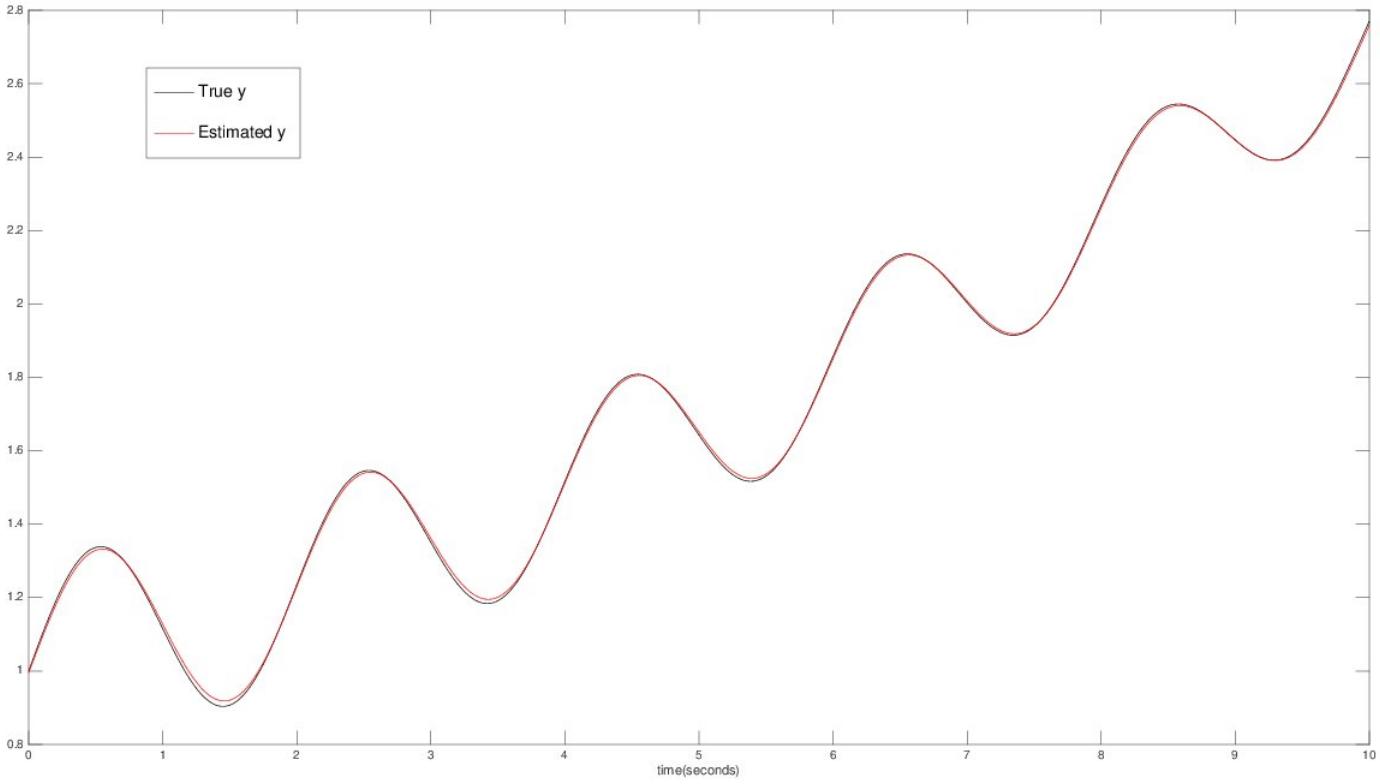

Figure 4.7 Estimated system output y after convergence is achieved.(10dB SNR)

Chapter 5

Reproducing Kernel Hilbert Spaces

5.1 Introduction

Given a set X, if we equip the set of all functions from X to $\mathbb{F}$,$\mathcal{F}(X, \mathbb{F})$ with the usual operations of addition, $(f + g)(x) = f(x) + g(x)$, and scalar multiplication, $(\lambda\Delta f)(x) = \lambda\Delta(f(x))$, then $\mathcal{F}(X, \mathbb{F})$ is a vector space over $\mathbb{F}$ [32].

Definition 1. A Hilbert Space is an inner product space that is complete and separable with respect to the norm defined by the inner product.

Examples of Hilbert spaces include:

1. The vector space $\mathbb{R}^n$ with $\langle a, b\rangle = a'b$, the vector dot product of a and b.
2. The space l_2 of square summable sequences, with inner product $\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i$
3. The space L^2 of square integrable functions (i.e., $\int_s f(x)^2 \mathrm{d}x < \infty$), with inner product $\langle f, g\rangle = \int_s f(x)g(x)\mathrm{d}x < \infty$) [33]

Definition 2. Given a set X, we will say that $\mathcal{H}$ is a reproducing kernel Hilbert space(RKHS) on X over $\mathbb{F}$, provided that:

1. $\mathcal{H}$ is a vector subspace of $\mathcal{F}(X, \mathbb{F})$
2. $\mathcal{H}$ is endowed with an inner product, $\langle ., .\rangle$ making it into a Hilbert space,
3. or every $y \in X$, the linear evaluation functional, $E_y : \mathcal{H} \mapsto \mathbb{F}$, defined by $E_y(f) = f(y)$, is bounded [32].

For instance, the L^2 space is a Hilbert space, but not an RKHS because the delta function which has the reproducing property:

$$f(x) = \int_s \delta(x-u)f(u)\mathrm{d}u \tag{5.1}$$

does not satisfy the square integrable condition, that is,

$$\int_s \delta(u)^2 \mathrm{d}u \not< \infty \tag{5.2}$$

thus the delta function is not in L^2 [33].

5.2 Reproducing Kernels for Modelling of LTI Systems [4]

The class of systems considered here comprises linear time invariant systems described in terms of their characteristic equations :

$$D^m y(t) = a_{m-1}D^{m-1}y(t) + a_{m-2}D^{m-2}y(t) + ... + a_1 D^1 y(t) + a_0 D^0 y(t), \tag{5.3}$$

with the usual definition of the differential operators :

$$D^0 y(t) = y(t); \quad D^1 y(t) = \frac{d}{dt}y(t); \quad D^k y(t) = \frac{d^k}{dt^k}y(t); k = 1, ..., m \tag{5.4}$$

that are satisfied by the system "output functions" y on some interval of time $t \in [a, b] \subset \mathbb{R}$, where $a_i \in \mathbb{R}, i = 1, .., m-1$, as real coefficients.

As mentioned in 2.3.2 and with a change of notation from K_{DS} to K^{DS}, it will be assumed that a double-sided kernel $K^{DS} : (t, \xi) \mapsto K^{DS}(t, \xi),\ (t, \xi) \in [a, b] \times [a, b]$, can be constructed such that every solution of (5.3) on an interval $[a, b]$ satisfies

$$y(p) = \int_a^b K^{DS}(p, \xi)y(\xi)d\xi\ ; \quad p \in [a, b] \tag{5.5}$$

Since the output variable y is assumed to be measured, and hence is likely to be corrupted by noise, it is convenient to regard it as a member of the Hilbert space $L^2[a, b]$. It follows from the construction the double-sided kernel that K^{DS} is also an L^2 function, so $K^{DS} \in L^2[a, b] \times L^2[a, b]$. For brevity it will be convenient to adopt the following shorthand

notation:

$$K_p^{DS}(\xi) := K^{DS}(p,\xi); \quad K_p^{DS} \in L^2[a,b]; \text{ i.e. } K^{DS}(p,\cdot) \in L^2[a,b]; \text{ for all } p \in [a,b] \quad (5.6)$$

allowing to re-write (5.5) in terms of the scalar product on $L^2[a,b]$:

$$y(p) = \langle y, K_p^{DS} \rangle_2, \quad p \in [a,b] \quad (5.7)$$

$$\text{i.e. } y = \langle y, K^{DS} \rangle_2 \; ; \quad \text{where } K^{DS} : p \mapsto K_p^{DS} \in L^2[a,b] \; ; \; p \in [a,b] \quad (5.8)$$

where $\langle \cdot, \cdot \rangle_2$ is the scalar product in $L^2[a,b]$. It should be noted that with these definitions, the functions K_p^{DS}, although considered as members of $L^2[a,b]$ are point wise defined. The reproducing property (5.7) holds for any functions y satisfying (5.3) .
Since the reproducing property is linear in the reproduced signal y, the condition (5.44) already holds on a linear vector subspace of $L^2[a,b]$ comprising all solutions of the characteristic equation on the interval $[a,b]$ (regardless of their initial conditions).

5.2.1 Classical Construction of RKHS Spaces

In classical theory of reproducing kernel Hilbert spaces there are two approaches of constructing such spaces: (a) "from the inside of an RKHS", and (b) "from the outside of an RKHS". In the "inside" approach one starts with a given Hilbert space in which there already are defined: (i) a scalar product, (ii) a symmetric positive definite "kernel" that is "generating it". In the inside approach no reference is made to any outer space in which the given RKHS may be embedded as a subspace.

On the other hand, the "outside" approach starts from a given linear operator in a larger Hilbert space and proceeds to define the RKHS as a range space of this operator - a closed subspace of the larger space. The outside approach seems particularly useful when projections of functions from the larger space onto the RKHS as its subspace are required. Such is the situation when one wishes to "filter" a noisy signal y. Below, we shall review the "outside" approach to see how it can serve in achieving our objectives.

To characterize the $\mathcal{H}_K$ that satisfies all the "canons" of the classical theory of the RKHS

spaces while adopting the "outside approach", one starts with a given linear operator, acting in the space of noisy signals (say in $L^2[a,b]$). At such a starting point, the construction of this operator is nebulous if one already has a class of functions that one requires to lie in the constructed RKHS. Ignoring this difficulty for now, let us review how any given linear and bounded operator in $L^2[a,b]$ can generate an RKHS subspace. The operator, referred here by the symbol L, is defined point-wise, in terms of a "kernel" function $\mathbf{h}$:

$$\mathbf{h} : [a,b] \to L^2[a,b] \; ; \quad \text{with} \;\; \mathbf{h}(p,\cdot) := \mathbf{h}_p \in L^2[a,b] \quad \text{for any} \;\; p \in [a,b] \tag{5.9}$$

where L is explicitly given by

$$L : L^2[a,b] \to Ran(L) \subset L^2[a,b]; \tag{5.10}$$

$$(L\mathbf{f})(p) := \langle \mathbf{f}, \mathbf{h}_p \rangle_2 := \int_a^b \mathbf{h}(p,\xi)\mathbf{f}(\xi)d\xi \; ; \quad \mathbf{f} \in L^2[a,b], \;\; p \in [a,b] \tag{5.11}$$

$$\text{the value of } L \;\; \text{is thus a function} \quad f := L\mathbf{f} \tag{5.12}$$

where $Ran(L)$ denotes the range of the operator L while its domain is clearly $D(L) := L^2[a,b]$. The norm on the domain vector space is considered to be induced by the scalar product $\langle \cdot, \cdot \rangle_2$, so for all $\mathbf{f} \in D(L)$, $\| \mathbf{f} \|_2^2 = \langle \mathbf{f}, \mathbf{f} \rangle_2$. However, let the norm on the range space $R(L)$ be introduced differently by the expression

$$\| f \|_R := \inf\{ \| \mathbf{f} \|_2 \quad \text{s.t.} \quad f = L\mathbf{f} \} \tag{5.13}$$

The details of this setting will be clear later.
Induced by the norm (5.13) is the scalar product on $Ran(L)$:

$$\langle f_1, f_2 \rangle_R := \langle \mathbf{f}_1, \mathbf{f}_2 \rangle_2, \quad \text{for any } f_1, f_2 \in Ran(L) \tag{5.14}$$

$$\text{where} \;\; f_i, i = 1,2 \;\; \text{are such that} \tag{5.15}$$

$$\| f_i \|_R := \inf\{ \| \mathbf{f}_i \|_2 \quad \text{s.t.} \quad f_i = L\mathbf{f}_i \}, \; i = 1,2 \tag{5.16}$$

Further, define a function called the kernel for the RKHS associated with the reproducing property (5.7) by the product

$$K(p,q) := \langle \mathbf{h}_q, \mathbf{h}_p \rangle_2 = \int_a^b \mathbf{h}(p,\xi)\mathbf{h}(q,\xi)d\xi \quad \text{for all} \quad p,\ q \in [a,b] \tag{5.17}$$

we shall also write

$$K_q := \int_a^b \mathbf{h}(\cdot,\xi)\mathbf{h}(q,\xi)d\xi := K(\cdot,q) \in L^2[a,b] \ \text{ as a function of } p \tag{5.18}$$

$$\text{and } K_q(p) := K(p,q) \in \mathbb{R} \tag{5.19}$$

The properties of the kernel K and its link with the operator in (5.11) are discussed in the following.

Proposition 1.
(1) The kernel as a function of two variables is symmetric; i.e. for all $p, q \in [a,b]$ we have $K(p,q) = K(q,p)$;
(2) The function K is positive definite in the sense of Moore; i.e. for any finite sets of numbers $c_p, c_q \in \mathbb{R}$

$$\sum_{p,q} c_p c_q K(p,q) \geq 0 \ \text{ for all } p,q \in [a,b] \tag{5.20}$$

(3) The functions K_q, $q \in [a,b]$, are all members of the range of the operator L; i.e. $K_q \in Ran(L)$. As $Ran(L)$ is a linear vector space, it follows that

$$span\{K_q, q \in [a,b]\} \subset Ran(L) \tag{5.21}$$

where *span* denotes the usual linear span of vectors in a vector space.

Proof:

Property (1) follows from symmetric property of the scalar product in (5.17).
Property (2) is shown as follows. For any finite sets of numbers (of equal cardinality)

$c_p, c_q \in \mathbb{R}$

$$\sum_{p,q} c_p c_q K(p,q) = \sum_{p,q} c_p c_q \langle \mathbf{h}_q, \mathbf{h}_p \rangle_2 = \sum_{p,q} \langle c_q \mathbf{h}_q, c_p \mathbf{h}_p \rangle_2$$
$$\langle \sum_q c_q \mathbf{h}_q, \sum_p c_p \mathbf{h}_p \rangle_2 = \| \sum_p c_p \mathbf{h}_p \|_2^2 \geq 0 \quad \text{for all } p, q \in [a,b] \tag{5.22}$$

Property (3) also follows from (5.17) because, according to this definition the function K_q is the result of $\mathbf{h}_p$ acting on $\mathbf{h}_q$ i.e.

$$L\mathbf{h}_q = \int_a^b \mathbf{h}(\cdot, \xi)\mathbf{h}(q, \xi) d\xi = K_q \ , \quad \text{for any} \quad q \in [a,b] \ , \quad \text{so} \quad K_q \in Ran(L) \tag{5.23}$$

This completes the proof.

□

Proposition 2.

(1) The domain of the operator L can be represented as a direct sum of subspaces :
$D(L) = L^2[a,b] = Ker(L) \oplus Ker(L)^\perp$
where $Ker(L)$ denotes the null space of L and $Ker(L)^\perp$ is its orthogonal complement in $L^2[a,b]$.

(2) The null space of L and its orthogonal complement are given by

$$Ker(L) = span\{\mathbf{h}_p, p \in [a,b]\}^\perp \quad \text{and} \tag{5.24}$$
$$Ker(L)^\perp = [span\{\mathbf{h}_p, p \in [a,b]\}^\perp]^\perp = \overline{span\{\mathbf{h}_p, p \in [a,b]\}} \tag{5.25}$$

where $\overline{\mathcal{S}}$ denotes the closure of a set $\mathcal{S}$ in the topology of the space containing it.

(3) The range space of L is given by

$$Ran(L) = \overline{span\{K_q, q \in [a,b]\}} \tag{5.26}$$

(4) The restricted mapping $L_K : Ker(L)^\perp \to Ran(L)$ is an isometry in that

$$\| L(P_{K^\perp}\mathbf{f}) \|_R = \| P_{K^\perp}\mathbf{f} \|_2 \ ; \quad \text{for all } \mathbf{f} \in L^2[a,b] \tag{5.27}$$

where $P_{K^\perp}$ denotes the projection operator on $Ker(L)^\perp$ and $\| \cdot \|_R$ denotes the

norm introduced in $Ran(L)$. In terms of scalar products the latter means

$$\langle L(P_{K^\perp}\mathbf{f}), L(P_{K^\perp}\mathbf{g})\rangle_R = \langle P_{K^\perp}\mathbf{f}, P_{K^\perp}\mathbf{g}\rangle_2 \ ; \quad \text{for all } \mathbf{f},\mathbf{g} \in L^2[a,b] \tag{5.28}$$

Proof: [4]
(1) The operator L is linear, hence its null space is closed so the result is implied by the Orthogonal Projection Theorem.
(2) The null space of L, $Ker(L)$, is characterized as follows:

$$\text{if } \mathbf{f} \in Ker(L) \ \text{ then } L\mathbf{f} = 0 \ \text{ i.e. } \ (L\mathbf{f})(p) = \langle \mathbf{f}, \mathbf{h}_p\rangle_2 = 0 \ \text{ for all } p \in [a,b] \tag{5.29}$$

As L is linear the above implies that

$$\text{if } \mathbf{f} \in Ker(L) \ \text{ then } \ \mathbf{f} \perp span\{\mathbf{h}_p, p \in [a,b]\} \ \text{ i.e. } \mathbf{f} \in span\{\mathbf{h}_p, p \in [a,b]\}^\perp \tag{5.30}$$

Conversely,

$$\text{if } \mathbf{f} \in span\{\mathbf{h}_p, p \in [a,b]\}^\perp = \{\mathbf{h}_p, p \in [a,b]\}^\perp \tag{5.31}$$

$$\text{then } \ (L\mathbf{f})(p) = \langle \mathbf{f}, \mathbf{h}_p\rangle_2 = 0 \ \text{ for all } p \in [a,b] \ \text{ and so } \mathbf{f} \in Ker(L) \tag{5.32}$$

which proves (5.24) and (5.25) follows as double orthogonal complement induces closure of the subspace involved.
(3) By definition the operator L is linear, mapping $L\mathbf{h}_q = K_q$, for all $q \in [a,b]$, thus the image under L satisfies

$$L(span\{\mathbf{h}_q, q \in [a,b]\}) = span\{K_q, q \in [a,b]\} \ ; \ \text{ i.e. } \ span\{K_q, q \in [a,b]\} \subset Ran(L) \tag{5.33}$$

The operator L is obviously bounded in $L^2[a,b]$ because, by Cauchy-Schwartz inequality

$$|L\mathbf{f}(p)| = |\langle \mathbf{f}, \mathbf{h}_p\rangle_2| \leq \| \mathbf{h}_p \|_2 \cdot \| \mathbf{f} \|_2 \ ; \ \text{ for all } \ p \in [a,b], \ \mathbf{f} \in L^2[a,b] \tag{5.34}$$

By continuity of L, we get

$$L(\overline{span\{\mathbf{h}_q, q \in [a,b]\}}) = \overline{span\{K_q, q \in [a,b]\}} \; ; \quad \text{i.e.} \quad \overline{span\{K_q, q \in [a,b]\}} = Ran(L) \tag{5.35}$$

because any $\mathbf{f} \notin \overline{span\{\mathbf{h}_q, q \in [a,b]\}}$ satisfies $L\mathbf{f} = 0$ and hence contributes trivially to the range $Ran(L)$.
(4) The restricted mapping $L_K : Ker(L)^{\perp} \rightarrow Ran(L)$ is linear and bounded because the unrestricted L is linear and bounded. It is one-to-one because $Ker(L_K) = \{0\}$ by virtue of part (1). The mapping L_K is also onto by virtue of part (3), i.e. (5.35). Hence L_K is a bijective and linear mapping between vector spaces. To show that it is an isometry we need to demonstrate that it preserves scalar products. To this end, let

$$\mathbf{f}, \; \mathbf{g} \in span\{\mathbf{h}_q, q \in [a,b]\} \; ; \quad \text{i.e.} \quad \mathbf{f} = \sum_q \alpha_q \mathbf{h}_q \; ; \; \mathbf{g} = \sum_r \beta_r \mathbf{h}_r \tag{5.36}$$

Then

$$\langle L(P_{K^{\perp}}\mathbf{f}), L(P_{K^{\perp}}\mathbf{g})\rangle_R = \langle L\mathbf{f}, L\mathbf{g}\rangle_R = \langle \sum_q \alpha_q \langle \mathbf{h}_q, \mathbf{h}_p\rangle_2, \sum_r \beta_r \langle \mathbf{h}_r, \mathbf{h}_p\rangle_2 \rangle_R \tag{5.37}$$

$$= \langle \sum_q \alpha_q K_q, \sum_r \beta_r K_r \rangle_R = \langle \sum_q \alpha_q L\mathbf{h}_q, \sum_r \beta_r L\mathbf{h}_r \rangle_R \tag{5.38}$$

However, by definition of the scalar product in $Ran(L)$

$$\langle L\mathbf{h}_q, L\mathbf{h}_r\rangle_R = \langle \mathbf{h}_q, \mathbf{h}_r\rangle_2 \quad \text{for all} \quad q, \; r \tag{5.39}$$

Hence

$$\langle L(P_{K^{\perp}}\mathbf{f}), L(P_{K^{\perp}}\mathbf{g})\rangle_R = \langle \sum_q \alpha_q L\mathbf{h}_q, \sum_r \beta_r L\mathbf{h}_r \rangle_R \tag{5.40}$$

$$= \sum_{q,r} \alpha_q \beta_r \langle \mathbf{h}_q, \mathbf{h}_r\rangle_2 = \langle \sum_q \alpha_q \mathbf{h}_q, \sum_r \beta_r \mathbf{h}_r \rangle_2 = \langle \mathbf{f}, \mathbf{g}\rangle_2 = \langle P_{K^{\perp}}\mathbf{f}, P_{K^{\perp}}\mathbf{g}\rangle_2 \tag{5.41}$$

because for $\mathbf{f}, \mathbf{g}$ satisfying (5.36), $\mathbf{f} = P_{K^{\perp}}\mathbf{f}$ and $\mathbf{g} = P_{K^{\perp}}\mathbf{g}$; see (5.24). Also, note that in the above, we used the definition of the scalar product on $Ran(L)$ as given by (5.16) and

the fact that any $\mathbf{h}_q$ is already a function of minimal norm satisfying $L\mathbf{h}_q = K_q$ because $\mathbf{h}_q \in Ker(L)^\perp$ and

$$\inf\{\|\mathbf{h_q}\|_2 \text{ s.t. } L\mathbf{h}_q = K_q\} \tag{5.42}$$

$$= \inf\{\|\mathbf{h}_q + \mathbf{g}\|_2 \text{ s.t. } L\mathbf{h}_q = K_q \text{ and } \mathbf{g} \in Ker(L)\} \tag{5.43}$$

The invariance of the scalar product extends to the closed subspaces $Ker(L)^\perp$ and $Ran(L)$ by continuity of L_K. The isometric equality (5.27) follows readily by setting $\mathbf{g} = \mathbf{f}$ in the above. The proof is complete.

□

Let us now investigate if f is reproductive on $Ran(L)$. The following result explains how $f \in Ran(L)$ is reproduced *inside* $Ran(L)$ in terms of the scalar product $\langle \cdot, \cdot \rangle_R$ on $Ran(L)$.

Theorem 1
Any $f \in Ran(L)$ is reproduced *inside* $Ran(L)$ by the action of K. More precisely, the reproducing property is stated in terms of the scalar product $\langle \cdot, \cdot \rangle_R$ on $Ran(L)$ as

$$\langle f, K_q \rangle_R = f(q) \ ; \text{ for all } q \in [a, b] \tag{5.44}$$

Equivalently, for any $\mathbf{f} \in Ker(L)^\perp$ such that $L\mathbf{f} = f$

$$\langle \mathbf{f}, \mathbf{h}_q \rangle_2 = (L\mathbf{f})(q) = f(q) \ ; \quad \text{for all } \ q \in [a, b] \tag{5.45}$$

Proof:
If $f \in Ran(L)$ then there exists an $\mathbf{f} \in Ker(L)^\perp$ such that $L\mathbf{f} = f$. Note that $\mathbf{f} \in Ker(L)^\perp$ is already a function of minimal norm satisfying $L\mathbf{f} = f$ because

$$\inf\{\|\mathbf{f}\|_2 \text{ s.t. } L\mathbf{f} = f\} = \inf\{\|\mathbf{f} + \mathbf{g}\|_2 \text{ s.t. } L\mathbf{f} = f \text{ and } \mathbf{g} \in Ker(L)\} \tag{5.46}$$

Since for any $q \in [a, b]$, $\mathbf{h}_q \in Ker(L)^\perp$ and $L\mathbf{h}_q = K_q$ in a one-to-one way (the restricted operator L_K is bijective) then $\mathbf{h}_q$ is also a function of minimal norm. From the isometry

property (5.41) it follows that the scalar product of f and K_q satisfies:

$$\langle f, K_q\rangle_R = \langle L\mathbf{f}, L\mathbf{h}_q\rangle_R = \langle LP_{K^\perp}\mathbf{f}, LP_{K^\perp}\mathbf{h}_q\rangle_R = \langle P_{K^\perp}\mathbf{f}, P_{K^\perp}\mathbf{h}_q\rangle_2 \tag{5.47}$$

$$= \langle \mathbf{f}, \mathbf{h}_q\rangle_2 = (L\mathbf{f})(q) = f(q)\ ;\ \text{for all } q \in [a,b] \tag{5.48}$$

as claimed. Finally, it is worth noting that (5.45) holds (by the very definition of L) on the *entire* space $L^2[a,b]$.

□

Definition (RKHS generated by L)
The reproducing kernel Hilbert space induced by the reproductive mapping L is defined as

$$\mathcal{H}_K := Ran(L)\ ; \quad \text{with scalar product } \ \langle\cdot,\cdot\rangle_R \tag{5.49}$$

This definition is correct as $Ran(L)$ is a closed subspace of a Hilbert space $L^2[a,b]$ and hence is also complete. The space $\mathcal{H}_K$ is separable for the same reason.

5.2.2 From Theory to Practice

It is now the aim to investigate how the above theoretical construction can be put to work with reference to our objectives.

To this end, assume that $\mathbf{h}_p := K_p^{DS}$, for all $p \in [a,b]$, so that

$$L : L^2[a,b] \to Ran(L) \subset L^2[a,b]; \tag{5.50}$$

$$(L\mathbf{f})(p) := \langle \mathbf{f}, K_p^{DS}\rangle_2\ ; \quad \mathbf{f} \in L^2[a,b], \quad p \in [a,b] \tag{5.51}$$

$$\text{the value of } L \ \text{ is thus a function } \quad f := L\mathbf{f} \tag{5.52}$$

Explicitly

$$(L\mathbf{f})(p) := \int_a^b K_p^{DS}(\xi)\mathbf{f}(\xi)d\xi = \int_a^b K^{DS}(p,\xi)\mathbf{f}(\xi)d\xi\ ; \quad \mathbf{f} \in L^2[a,b], \quad p \in [a,b] \tag{5.53}$$

By definition, our RKHS space is then $Ran(L) = \mathcal{H}_K$ and is endowed with a scalar

product defined by (5.16), re-stated here

$$\langle f_1, f_2\rangle_R := \langle \mathbf{f}_1, \mathbf{f}_2\rangle_2, \quad \text{for any } f_1, f_2 \in Ran(L) \tag{5.54}$$

$$\text{where } \ f_i, i = 1, 2 \ \text{ are such that} \tag{5.55}$$

$$\| f_i \|_R := \inf\{\| \mathbf{f}_i \|_2 \quad \text{s.t.} \quad f_i = L\mathbf{f}_i \}, \ i = 1, 2 \tag{5.56}$$

where the RKHS kernel is computed as

$$K(p, q) := \langle K_q^{DS}, K_p^{DS}\rangle_2 = \int_a^b K^{DS}(p, \xi)K^{DS}(q, \xi)d\xi \ ; \quad p, q \in [a, b] \tag{5.57}$$

$$\text{i.e.} \quad K_q := K(\cdot, q) = \int_a^b K^{DS}(\cdot, \xi)K^{DS}(q, \xi)d\xi \ ; \quad q \in [a, b] \tag{5.58}$$

Continuing to draw on the results stated by Proposition 2, we have

$$Ker(L) = span\{K_p^{DS}, p \in [a, b]\}^\perp \quad \text{and} \tag{5.59}$$

$$Ker(L)^\perp = [span\{K_p^{DS}, p \in [a, b]\}^\perp]^\perp = \overline{span\{K_p^{DS}, p \in [a, b]\}} \tag{5.60}$$

$$Ran(L) = \overline{span\{K_q, q \in [a, b]\}} = \overline{span\{\int_a^b K^{DS}(\cdot, \xi)K^{DS}(q, \xi)d\xi, q \in [a, b]\}} \tag{5.61}$$

It is now apparent that characterizing the range of an RKHS is one of the biggest challenges of the RKHS theory. The set of spanning functions in (5.61) is amazingly rich. It is, however, quite clear that any solution of the characteristic equation (5.3) belongs to that range, despite that an explicit expression of it in terms of the spanning vectors in (5.61) is not readily available. To see that $y \in Ran(L)$ simply recall that y satisfies (by our construction)

$$y(p) = \int_a^b K^{DS}(p, \xi)y(\xi)d\xi = \langle y, K_p^{DS}\rangle_2 = (Ly)(p) \ ; \quad p \in [a, b] \tag{5.62}$$

which directly yields

$$y = Ly \in Ran(L) \tag{5.63}$$

Re-interpreting the isometry property (5.41) is necessary to understand in detail how K_q reproduce functions $f \in Ran(L)$ and thus also y. To this end, consider arbitrary $f, g \in Ran(L)$. There exist vectors $\mathbf{f}, \mathbf{g} \in Ker(L)^{\perp}$ of minimal norm that generate them which can be expressed as a point-wise limits of functions in the span (5.60)

$$\mathbf{f}_i = \sum_q \alpha_q^i K_q^{DS} \; ; \quad \mathbf{g}_i = \sum_r \beta_r^i K_r^{DS} \; ; \quad i \in \mathbb{I} \tag{5.64}$$

$$\text{with} \quad L\mathbf{f}_i = f_i \; ; \quad L\mathbf{g}_i = g_i \; ; \quad i \in \mathbb{I} \tag{5.65}$$

$$\text{and with} \quad f_i(p) \to f(p) \; ; \quad g_i(p) \to g(p) \; ; \quad \text{for all} \quad p \in [a,b] \quad \text{as } i \to \infty \tag{5.66}$$

$$\text{where there exist} \quad \alpha_q \; , \; \beta_r \quad \text{s.t.} \quad \mathbf{f} = \sum_q \alpha_q K_q^{DS} \; ; \quad \mathbf{g} = \sum_r \beta_r K_r^{DS} \tag{5.67}$$

Note that since $\mathbf{f}_i$ and $\mathbf{g}_i$ are already taken to be of minimal norm and hence satisy $\mathbf{f}_i, \mathbf{g}_i \in Ker(L)^{\perp}$, this also means that $P_{K^{\perp}}\mathbf{f}_i = \mathbf{f}_i$ and $P_{K^{\perp}}\mathbf{g}_i = \mathbf{g}_i$, for all $i \in \mathbb{I}$. Then, by virtue of the definition of the scalar product in $Ran(L)$; see (5.56) ,

$$\langle f_i, g_i \rangle_R = \langle \mathbf{f}_i, \mathbf{g}_i \rangle_2 = \langle P_{K^{\perp}}\mathbf{f}_i, P_{K^{\perp}}\mathbf{g}_i \rangle_2 = \int_a^b \sum_q \alpha_q^i K^{DS}(q,\xi) \cdot \sum_r \beta_r^i K^{DS}(r,\xi) \; d\xi \tag{5.68}$$

$$= \sum_{q,r} \alpha_q^i \beta_r^i \int_a^b K^{DS}(q,\xi) K^{DS}(r,\xi) \; d\xi = \sum_{q,r} \alpha_q^i \beta_r^i K(q,r) \tag{5.69}$$

By continuity we obtain in the limit as $i \to \infty$

$$\langle f, g \rangle_R = \langle \mathbf{f}, \mathbf{g} \rangle_2 = \langle P_{K^{\perp}}\mathbf{f}, P_{K^{\perp}}\mathbf{g} \rangle_2 \tag{5.70}$$

$$= \sum_{q,r} \alpha_q \beta_r \int_a^b K^{DS}(q,\xi) K^{DS}(r,\xi) \; d\xi = \sum_{q,r} \alpha_q \beta_r K(q,r) \tag{5.71}$$

The above also finally delivers a concrete formula for computation of scalar products in our RKHS. Clearly, to compute $\langle f, g \rangle_R$ one has to know their expansions in (5.67) which are not usually available. The scalar product itself is given in terms of the RKHS kernel function; a fact that is touted by the standard theory of RKHS.

To understand the reproducing property on our $\mathcal{H}_K$, assume that $f \in Ran(L)$, i.e.

$$\int_a^b K^{DS}(q,\xi)\mathbf{f}(\xi)\ d\xi = f(q) \quad \text{for all} \quad q \in [a,b] \tag{5.72}$$

where $\mathbf{f}$ has the minimum norm expansion (5.67). Remembering that the reproducing property is originally defined from "within" $Ran(L)$ and via "scaling " by the RKHS kernel (see (5.69)), one writes

$$\langle f, K_q\rangle_R = \langle L\mathbf{f}, L\mathbf{h}_q\rangle_R = \langle LP_{K^\perp}\mathbf{f}, LP_{K^\perp}\mathbf{h}_q\rangle_R = \langle P_{K^\perp}\mathbf{f}, P_{K^\perp}\mathbf{h}_q\rangle_2 \tag{5.73}$$

$$= \langle \mathbf{f}, \mathbf{h}_q\rangle_2 = \int_a^b K^{DS}(q,\xi)\mathbf{f}(\xi)\ d\xi = f(q)\ ;\ \text{for all } q \in [a,b] \tag{5.74}$$

Slight confusion arises when the function f is actually a solution of our characteristic equation, i.e. $f = y$. In this case the reproducing property should really be written as

$$\int_a^b K^{DS}(q,\xi)\mathbf{y}(\xi)\ d\xi = y(q)\ ;\ \text{for all } q \in [a,b] \tag{5.75}$$

where the boldface $\mathbf{y}$ signifies the "expansion" of y in terms of the K^{DS} functions, but does not differ from it, otherwise. It should also be clear how the functions K_r, for $r \in [a,b]$, reproduce themselves within $Ran(L)$:

$$\langle K_r, K_q\rangle_R = \langle L\mathbf{h}_r, L\mathbf{h}_q\rangle_R = \langle LP_{K^\perp}\mathbf{h}_r, LP_{K^\perp}\mathbf{h}_q\rangle_R = \langle P_{K^\perp}\mathbf{h}_r, P_{K^\perp}\mathbf{h}_q\rangle_2 \tag{5.76}$$

$$= \langle \mathbf{h}_r, \mathbf{h}_q\rangle_2 = \int_a^b K^{DS}(q,\xi)K^{DS}(r,\xi)\ d\xi = K(q,r) = K(r,q) = K_r(q)\ ;\ \text{for all } q \in [a,b] \tag{5.77}$$

while it is seen that symmetricity of the kernel function is essential for this to occur.

Finally, we need to determine what happens if one starts from an arbitrary $\mathbf{g} \in L^2[a,b]$ and wishes to "isolate" its best approximation within $\mathcal{H}_K$. Denoting by $P_{\mathcal{H}}\mathbf{g} \in \mathcal{H}_K$ the orthogonal projection of $\mathbf{g}$ onto $\mathcal{H}_K$, (the de-noisified signal closest to the observed $\mathbf{g}$), and using (5.60), the latter must satisfy the orthogonality condition

$$(\mathbf{g} - P_{\mathcal{H}}\mathbf{g}) \perp \mathcal{H}_K \quad \text{i.e.} \quad \Longleftrightarrow \quad (\mathbf{g} - P_{\mathcal{H}}\mathbf{g}) \perp K_r \quad \text{for all } r \in [a,b] \tag{5.78}$$

where orthogonality is understood in the sense of $L^2[a,b]$ as $\mathcal{H}_K$ was constructed here as a subspace of the latter. If $P_{\mathcal{H}}\mathbf{g} = \sum_q \eta_q K_q$ is the expansion of $P_{\mathcal{H}}\mathbf{g}$ in $\mathcal{H}_K$ then it follows that the above orthogonality condition can be rewritten as

$$\int_a^b K(r,\xi)\mathbf{g}(\xi)d\xi - \int_a^b \sum_q \eta_q K(q,\xi)K(r,\xi)d\xi = 0 \quad \text{for all } r \in [a,b] \tag{5.79}$$

so that

$$\int_a^b K(r,\xi)\mathbf{g}(\xi)d\xi = \int_a^b \sum_q \eta_q K(q,\xi)K(r,\xi)d\xi \tag{5.80}$$

$$= \sum_q \eta_q \langle K_r, K_q\rangle_R = \sum_q \eta_q \langle \mathbf{h}_q, \mathbf{h}_r\rangle_2 = \sum_q \eta_q K_q(r) = P_{\mathcal{H}}\mathbf{g}(r)\ ; \quad \text{for all } r \in [a,b] \tag{5.81}$$

by virtue of the self-reproducing property (5.76) - (5.77). The above directly states that the desired projection onto $\mathcal{H}_K$ is given point-wise by

$$P_{\mathcal{H}}\mathbf{g}(r) = \int_a^b K(r,\xi)\mathbf{g}(\xi)d\xi\ ; \quad \text{for all } r \in [a,b] \tag{5.82}$$

To re-state this expression in more detail,

$$P_{\mathcal{H}}\mathbf{g}(r) = \int_a^b K(r,\xi)\mathbf{g}(\xi)d\xi = \int_a^b \langle \mathbf{h}_r, \mathbf{h}_\xi\rangle_2\ \mathbf{g}(\xi)d\xi = \int_a^b \int_a^b \mathbf{h}(r,\zeta)\mathbf{h}(\xi,\zeta)d\zeta\ \mathbf{g}(\xi)d\xi \tag{5.83}$$

$$= \int_a^b \mathbf{h}(r,\zeta) \int_a^b \mathbf{h}(\xi,\zeta)\mathbf{g}(\xi)d\xi\ d\zeta \quad \text{for all } r \in [a,b] \tag{5.84}$$

5.2.3 Projection by Tikhonov Regularization

Calculating projections at every point in the interval $[a,b]$ is not practical, and it is in fact not necessary if the characteristic equation is known as the exact model that produces the measured data. The process of function reconstruction from noisy data is known as Tikhonov regularization. In the case at hand the regularization problem is formulated as follows.

Define a finite set of discrete points $p_i,\ i = 1,...,n,\ p_i \in [a,b]$, calculating corresponding

values $y_i,\ i = 1, ..., n$, along a "measured trajectory" $\mathbf{f}(\xi),\ \xi \in [a, b]$, by evaluating:

$$y_i := \int_a^b K^{DS}(p_i; \xi)\mathbf{f}(\xi)d\xi\ ,\quad i = 1, ..., n \tag{5.85}$$

Consider an optimization problem of finding a function $f \in \mathcal{H}_K$ that best approximates the noisy point-wise data given in (5.85) :

$$\inf\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i - f(p_i))^2 + \frac{\lambda}{2} \parallel f \parallel_R^2\ ;\ \text{w.r.t.}\quad f \in \mathcal{H}_K\right\} \tag{5.86}$$

where $\lambda \in \mathbb{R}, \lambda \geq 0$ is a regularization parameter to be fixed a priori and $\mathcal{H}_K$ is an RKHS containing all solutions of our characteristic equation.

First, since the RKHS $\mathcal{H}_K$ is closed, and the cost function is convex, the minimum exists in (5.86) so we denote it by $\hat{f} \in \mathcal{H}_K$. Next, Representer Theorem states that the optimal solution of (5.86) can be sought in the form

$$\hat{f}(\cdot) = \sum_{j=1}^{n} c_j K(\cdot, p_j) = \sum_{j=1}^{n} c_j \int_a^b K^{DS}(\cdot, \xi)K^{DS}(p_j, \xi)d\xi \tag{5.87}$$

$$\hat{f}(p_i) = \sum_{j=1}^{n} c_j K(p_i, p_j)\ , i = 1, ..., n \tag{5.88}$$

for some coefficients $c_j, j = 1, ..., n$. The square of the norm in $\mathcal{H}_K$ is then re-written as

$$\parallel f \parallel_R^2 = \langle f, f\rangle_R = \langle\sum_{i=1}^{n} c_i K_{p_i}, \sum_{j=1}^{n} c_j K_{p_j}\rangle_R \tag{5.89}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle K_{p_i} K_{p_j}\rangle_R = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(p_i, p_j) := c^T \mathbf{K} c \tag{5.90}$$

where we define for brevity:

$$K_{p_i} := K(\cdot, p_i)\ ,\ c^T := [c_1, ..., c_n]^T \in \mathbb{R}^n\ ;\quad \mathbf{K} := [K(p_i, p_j)]_{i,j=1,...,n} \in \mathbb{R}^n \times \mathbb{R}^n \tag{5.91}$$

$$\text{and also}\quad y^T := [y_1, ..., y_n]^T \in \mathbb{R}^n \tag{5.92}$$

Substituting (5.87) and (5.90) into (5.86) yields

$$\min \left\{ \frac{1}{2} \parallel y - \mathbf{K}c \parallel_2^2 + \frac{\lambda}{2} c^T \mathbf{K} c \; ; \; \text{w.r.t.} \quad c \in \mathbb{R}^n \right\} \tag{5.93}$$

The unique stationary point for (5.93) is found from the equation

$$\begin{aligned} &\frac{\partial}{\partial c} \left\{ \frac{1}{2} \parallel y - \mathbf{K}c \parallel_2^2 + \frac{\lambda}{2} c^T \mathbf{K} c \right\} = 0 \quad \Longrightarrow \\ &- \mathbf{K}(y - \mathbf{K}c) + \lambda \mathbf{K}c = 0 \iff (\mathbf{K} + \lambda \mathbf{I})c = y \iff c = (\mathbf{K} + \lambda \mathbf{I})^{-1} y \end{aligned} \tag{5.94}$$

Note that the matrix in (5.94) is invertible for $\lambda > 0$ and it can be found by solving a set of linear equations using e.g. Cholesky factorization. If $\hat{c}$ denotes the stationary point in (5.94) then the optimal $\hat{f} \in \mathcal{H}_K$ is given by

$$\hat{f}(q) = \sum_{j=1}^{n} \hat{c}_j K(q, p_j) = \mathbf{K}_q^T c \; , \quad \text{for all} \quad q \in [a, b] \tag{5.95}$$

$$\text{with} \quad \mathbf{K}_q := [K(q, p_j)]_{j=1,\ldots,n} \tag{5.96}$$

The solution of the regularization problem $\hat{f}(q), \; q \in [a, b]$, can next be used as "data curve" to calculate the corresponding initial conditions of our characteristic equation $y(a)$, $\frac{d}{dt} y(a)$ and $\frac{d^2}{dt^2} y(a)$ while using the kernel K^{DS}. Finally, the so retrieved initial conditions can be employed to explicitly integrate the characteristic equation to deliver an exact trajectory $y(q), q \in [a, b]$ that satisfies this equation. The latter is in fact a de-noisified reconstruction of the system trajectory that fits it exactly and hence satisfies the reproducing property employing K^{DS}.

Chapter 6

Conclusion and Future Work

Model estimation, adaptive filtering, and joint parameter and state estimation are topics of paramount importance to applications in almost all disciplines of science. The literature is overwhelmingly vast and diverse, ranging from general estimation methods, [34], [35], [36], through recursive stochastic filtering approaches [14], [37] to system modeling via data assimilation [38], [39]. Recursive stochastic approaches prevail due to their elegance, simplicity, and capability of efficient noise attenuation. Most filtering techniques, maybe with the exception of the Wiener filters [40], assume at least partial information about the initial state of the system and measurement noise characteristics. Convergence of such classical methods is conditioned by the validity of the assumptions made.

New trends in system estimation and modelling are rapidly developing and include: algebraic dead-beat observers based on the concept of differential flatness, algebraic parameter estimation [27], [41], [6], [42], and invariant observers [43], [44].The algebraic dead-beat methods are increasingly important as many applications, such as target tracking, call for finite time reliable estimation.

This work addresses the problem of simultaneous state and parameter estimation in linear systems from the measurement of its output over a finite interval of time, focusing on algebraic dead-beat methods. Additionally, the output measurement is subject to noise of an unknown characteristics. Chapter 2 studies the development of the double-sided kernel needed for the integral transform which reconstructs the state. But the inaccuracy of this reconstruction is shown at the conclusion of this chapter. Chapter 3 then proposes a new

method,by adopting a variational approach, of refining the system parameters and increasing the accuracy of estimation. Results of the application of this method can be seen in Chapter 4. The fitness of the estimated curves to the true curve is unarguably improved in all the three states : y,$y^{(1)}$ and $y^{(2)}$ as can be seen from the results and the KS test.

As noise is habitually represented by stochastic processes that also posses kernel representations, it is not hard to see the importance of the above approach. The two kernel representations: that of the hypothetical model and that of the stochastic process noise give rise to two objective functions: the value of the model residual error and the distance measure between two statistical distributions, that of the presumed noise and that of the observed residual. The need of a trade-off between model fitness to data and fitness of residual to noise statistics is apparent. Such and other probabilistic approaches to the problem will be explored elsewhere. Finally, the method is shown to be remarkably accurate and raises hopes for application in nonlinear flatness-based non-asymptotic estimation that relies on accurate computation of system output derivatives [3].

In Chapter 5, a kernel system model to be constructed is viewed as a linear finite dimensional subspace of a reproducing Hilbert space. Since, the subspace is linearly parametrized by the unknown system constants, their values determine the subspace "orientation" with respect to the cloud of measurement points. A "good orientation" is then aimed to be achieved by posing a problem stated in this chapter in the form of Tikhonov Regularization, constructing the best estimate of *true* y from the available noisy data. But the results of this will be discussed elsewhere.

Appendices

Appendix A

Calculus of Variations

A.1 Introduction [5]

In optimal control problems, the objective is to determine a function that minimizes a specified *functional*-the performance measure. The analogous problem in calculus is to determine a point that yields the minimum value of a function. This section introduces the concept of Variational Calculus.

Functionals
A functional **J** is a rule of correspondence that assigns to each function **x** in a certain class Ω a unique real number. Ω is called the domain of the functional, and the set of real numbers associated with the functions in Ω is called the range of the functional.
For example,

$$J(x) = \int_{t_0}^{t_f} x(t)dt \tag{A.1}$$

is a functional J, assigning to each function x, defined on $[t_0, t_f]$ a real number given by the area under the $x(t)$ curve.

Linear Functionals
J is a linear functional of x if and only if it satisfies the principle of homogeneity

$$J(\alpha x) = \alpha J(x) \tag{A.2}$$

for all $x \in \Omega$ and for all real numbers α such that $\alpha x \in \Omega$, and the principle of additivity

$$J(x^1 + x^2) = J(x^1) + J(x^2) \tag{A.3}$$

for all x^1,x^2 and $x^1 + x^2$ in Ω.

Function Norms

The norm of a function is a rule of correspondence that assigns to each function $x \in \Omega$, defined for $t \in [t_0, t_f]$, a real number. The norm of x, denoted by $\| \mathbf{x} \|$ satisfies the following properties:

1.

$$\| \mathbf{x} \| \geq 0 \; and \; \| \mathbf{x} \| = 0 \tag{A.4}$$

if and only if $x(t) = 0 \; \forall t \in [t_0, t_f]$

2.

$$\| \alpha \mathbf{x} \| = |\alpha| . \| \mathbf{x} \| \; \forall \alpha \in \mathbb{R} \tag{A.5}$$

3.

$$\| \mathbf{x^1} + \mathbf{x^2} \| \leq \| \mathbf{x^1} \| + \| \mathbf{x^2} \| \tag{A.6}$$

Intuitively speaking, the norm of the difference of two functions should be zero if the functions are identical, small if the functions are "close", and large if the functions are "far apart".

Increment

If x and $x + \delta x$ are functions for which the functional J is defined, then the increment of J, denoted by ΔJ, is

$$\Delta J(x, \delta x) \triangleq J(x + \delta x) - J(x) \tag{A.7}$$

δx is called the variation of the function x.

The Variation of a Functional

The variation of a functional plays the same role in determining extreme values of functionals as the differential does in finding maxima and minima of functions.
The increment of a functional can be written as

$$\Delta J(x, \delta x) \ = \delta J(x, \delta x) + g(x, \delta x). \parallel \delta x \parallel \tag{A.8}$$

where δJ is linear in δx. If

$$\lim_{\|\delta x\| \to 0} g(x, \delta x) \ = 0$$

then J is said to be *differentiable* on x and δJ is the *variation* of J evaluated for the function x.

Maxima and Minima of Functionals

A functional J with a domain Ω has a relative extremum at x^* if there is an $\epsilon > 0$ such that for all functions x in Ω that satisfy $\parallel x - x^* \parallel < \epsilon$ the increment of J has the same sign. If

$$\Delta J = J(x) - J(x^*) \geq 0 \tag{A.9}$$

$J(x^*)$ is a relative minimum;if

$$\Delta J = J(x) - J(x^*) \leq 0 \tag{A.10}$$

$J(x^*)$ is a relative maximum.
If (A.9) is satisfied for arbitrarily large ϵ, then $J(x^*)$ is a global minimum. Similarly, if (A.10) is satisfied for arbitrarily large ϵ, then $J(x^*)$ is a global maximum. x^* is called an *extremal* and $J(x^*)$ is called an *extremum*.

The fundamental theorem of the Calculus of Variations

The fundamental theorem used in finding extreme values of functions is the necessary condition that the differential vanish at an extreme point (except extrema at the boundaries of closed regions). In variational problems, the analogous theorem is that the variation must be zero on an extremal curve, provided that there are no bounds imposed

on the curves.

Let x be a vector function of t in the class Ω, and $J(x)$ be a differentiable functional of x. Assume that the functions in Ω are not constrained by any boundaries. The fundamental theorem of the calculus of variations is:

If x^* is an extremal, the variation of J must vanish on x^*;that is

$$\delta J(x^*, \delta x) = 0 \tag{A.11}$$

for all admissible δx. By admissible δx we mean that $x + \delta x$ must be a member of the class Ω; thus, if Ω is the class of continuous functions, x and δx are required to be continuous.

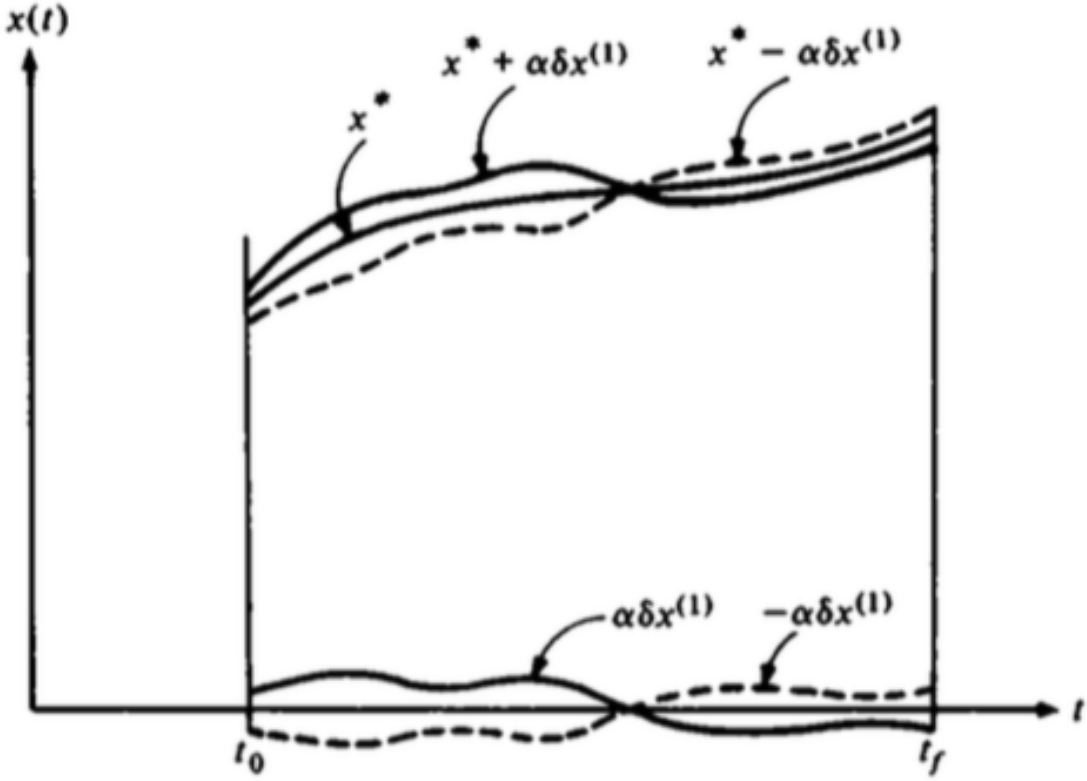

Figure A.1 An extremal and two neighboring curves

A.2 Variational Problems [5]

A Simple Variational Problem

Let x be a scalar function in the class of functions with continuous first derivatives. It is desired to find the function x* for which the functional

$$J(x) = \int_{t_0}^{t_f} g(x(t), \dot{x}(t), t)dt \tag{A.12}$$

has a relative extremum. It is assumed that the integrand g has continuous first and second partial derivatives with respect to all of its arguments; t_0 and t_f, are fixed,and the end points of the curve are specified as x_0 and x_f. Curves in the class Ω which also satisfy the end conditions are called admissible. Several admissible curves are shown in A.2. We

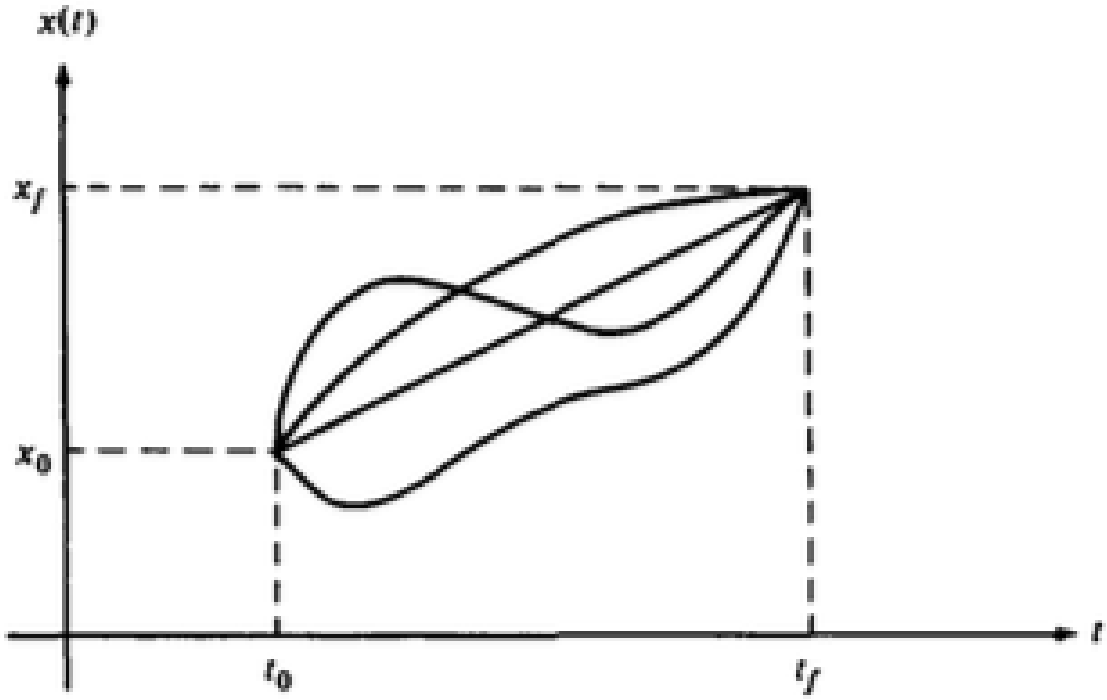

Figure A.2 Admissible curves

wish to find the curves (if any exist) that extremize $J(x)$. The search begins by finding the curves that satisfy the fundamental theorem. Let x be any curve in Ω, and determine

the variation $\delta J(x, \delta x)$ from the increment:

$$\begin{aligned} \Delta J(x, \delta x) &= J(x + \delta x) - J(x) \\ &= \int_{t_0}^{t_f} g(x(t) + \delta x(t), \dot{x(t)} + \delta \dot{x(t)}, t) dt \\ &\quad - \int_{t_0}^{t_f} g(x(t), \dot{x(t)}, t) dt \end{aligned} \tag{A.13}$$

Combining the intergrals gives

$$\Delta J(x, \delta x) = \int_{t_0}^{t_f} [g(x(t) + \delta x(t), \dot{x}(t) + \delta \dot{x}(t), t) \ - g(x(t), \dot{x}(t), t)] dt \tag{A.14}$$

Expanding the integrand of A.14 in a Taylor series about the point $x(t), \dot{x}(t)$ gives

$$\begin{aligned} \Delta J = \int_{t_0}^{t_f} [g(x(t), \dot{x}(t), t) \ + \ & \frac{\partial g}{\partial x}(x(t), \dot{x}(t), t) \delta x(t) \\ & + \frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t) \delta \dot{x}(t) \ - g(x(t), \dot{x}(t), t)] dt \end{aligned} \tag{A.15}$$

where the expansion is limited to the first order.
$\delta x(t)$ and $\delta \dot{x}(t)$ are related by

$$\delta x(t) = \int_{t_0}^{t_f} \delta \dot{x}(t) dt + \delta x(t_0)$$

thus, selecting δx uniquely determines $\delta \dot{x}$. Integrating by parts to express (A.15) entirely in δx:

$$\begin{aligned} \delta J(x, \delta x) = [\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)] \delta x(t) \Big|_{t_0}^{t_f} + \int_{t_0}^{t_f} \Bigg[& \frac{\partial g}{\partial x}(x(t), \dot{x}(t), t) \delta x(t) \\ & - \frac{d}{dt} [\frac{\partial g}{\partial \dot{x}}(x(t), \dot{x}(t), t)] \delta x(t) \Bigg] dt \end{aligned} \tag{A.16}$$

Since $x(t_0)$ and $x(t_f)$ are specified, all admissible curves must pass through these points; therefore, $\delta x(t_0) = 0$, $\delta x(t_f) = 0$ and the terms outside the integral vanish. If we now

consider an extremal curve, applying the fundamental theorem yields:

$$\delta J(x^*, \delta x) = 0 = \int_{t_0}^{t_f} \left[\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)] \right] \delta x(t) dt \tag{A.17}$$

Using the *fundamental lemma of calculus of variations* a [45], the integrand must be zero as well. Hence:

$$\frac{\partial g}{\partial x}(x^*(t), \dot{x}^*(t), t) - \frac{d}{dt}[\frac{\partial g}{\partial \dot{x}}(x^*(t), \dot{x}^*(t), t)] = 0 \quad \forall t \in [t_0, t_f] \tag{A.18}$$

which is called the *Euler equation*. Thus, we see that to obtain the optimal trajectory x^*, a *nonlinear, two-point boundary-value problem* (because we know $x(t_0)$ and $x(t_f)$) must be solved. The problem is difficult because of the combination of split boundary values and the non-linearity of the differential equation.

Constrained Minimization of Functionals using Lagrange Multipliers

In control problems the state trajectory is determined by the control $\mathbf{u}$; thus, we wish to consider functionals of $n + m$ functions, $\mathbf{x}$ and $\mathbf{u}$, but only m of the functions are independent-the controls. So, let us consider the presence of *point* constraints in variational problems. Let us determine a set of necessary conditions, as before, for a function w^* to be an extremal for a functional of the form:

$$J(w) = \int_{t_0}^{t_f} g(w(t), \dot{w}(t), t) dt \tag{A.19}$$

where w is a dimension $n + m$ vector of functions that satisfies $f_i(w(t), t) = 0 \;\; i = 1, 2, ..n$ which are called *point* constraints. The first step is to form the *augmented functional* by adjoining the constraining relations to J, which yields:

$$J_a(w) = \int_{t_0}^{t_f} \left[g(w(t), \dot{w}(t), t) + \mathbf{p}(t)^T [\mathbf{f}(w(t), t) \right] dt \tag{A.20}$$

where $\mathbf{p}$ is a vector of p_i's which are functions of time because the constraints need to be satisfied $\forall t \in [t_0, t_f]$. Similarly, $\mathbf{f}$ is a vector of f_i's. Notice that if the constraints are

satisfied, $J_a = J$ for any function $\mathbf{p}$. The variation of the functional J_a,

$$\delta J_a(w, \delta w, \mathbf{p}, \delta \mathbf{p}) = \int_{t_0}^{t_f} \left(\left[\frac{\partial g^T}{\partial w}(w(t), \dot{w}(t), t) \ + \mathbf{p}(t)^T [\frac{\partial \mathbf{f}}{\partial w}(w(t), t)] \right] \delta w(t) \right. \\ \left. + \frac{\partial g^T}{\partial \dot{w}}(w(t), \dot{w}(t), t) \delta \dot{w}(t) \ + \mathbf{f}^T(w(t), t) \delta \mathbf{p} \right) dt \tag{A.21}$$

where $\dfrac{\partial \mathbf{f}}{\partial w}$ denotes a $n \times (n+m)$ matrix:

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial w_1} & \cdots & \dfrac{\partial f_1}{\partial w_{n+m}} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial f_n}{\partial w_1} & \cdots & \dfrac{\partial f_n}{\partial w_{n+m}} \end{bmatrix}$$

Integrating by parts the term containing $\delta \dot{w}(t)$ and retaining only the terms inside the integral, we obtain:

$$\delta J_a(w, \delta w, \mathbf{p}, \delta \mathbf{p}) = \int_{t_0}^{t_f} \left(\left[\frac{\partial g^T}{\partial w}(w(t), \dot{w}(t), t) \ + \mathbf{p}(t)^T [\frac{\partial \mathbf{f}}{\partial w}(w(t), t)] \right. \right. \\ \left. \left. - \frac{d}{dt} [\frac{\partial g^T}{\partial \dot{w}}(w(t), \dot{w}(t), t)] \right] \delta w(t) \ + \mathbf{f}^T(w(t), t) \delta \mathbf{p} \right) dt \tag{A.22}$$

On an extremal, the variation must be zero; that is, $\delta J_a(w^*, p) = 0$. In addition, the point constraints must also be satisfied by an extremal; therefore, $\mathbf{f}(w^*(t), t) = 0, \forall t \in [t_0, t_f]$. Also, the coefficient of $\delta \mathbf{p}$ in (A.22) is zero. Since the constraints are satisfied, we can select the n Lagrange multipliers arbitrarily-let us choose the p's so that the coefficients of n of the components of $\delta w(t)$ are zero throughout the interval $[t_0, t_f]$. The remaining $(n+m) - n = m$ components of δw are then independent; hence, the coefficients of these components of $\delta w(t)$ must be zero. The final result is that, **in addition to the constraints**, the equations:

$$\frac{\partial g}{\partial w}(w^*(t), \dot{w}^*(t), t) \ + \mathbf{p}^*(t)^T [\frac{\partial \mathbf{f}}{\partial w}(w^*(t), t)] - \frac{d}{dt} [\frac{\partial g}{\partial \dot{w}}(w^*(t), \dot{w}^*(t), t) \ = 0 \tag{A.23}$$

must be satisfied.

Bibliography

[1] "A new method of non-asymptotic estimation for linear systems," Master's thesis.

[2] D. Ghoshal, K. Gopalakrishnan, and H. Michalska, "Using invariance to extract signal from noise," in *American Control Conference, 2017. ACC 2017*, 2017.

[3] D. Ghoshal, A. Pandey, and H. Michalska, "Variational approach to joint linear model and state estimation - submitted," in *American Control Conference, 2018. ACC 2018*, 2018 forthcoming.

[4] H. Michalska, " Instructor Notes: Reproducing Kernels for Modelling of Systems," 2016.

[5] D. E. Kirk. Dover Publications, 1998. [Online]. Available: http://app.knovel.com/hotlink/toc/id:kpOCTAI00A/optimal-control-theory/optimal-control-theory

[6] H. Sira-Ramírez, C. G. Rodríguez, J. C. Romero, and A. L. Juárez, *Algebraic identification and estimation methods in feedback control systems*. John Wiley & Sons, 2014.

[7] G. Stein, "Respect the unstable," *IEEE Control Systems*, vol. 23, no. 4, pp. 12–25, 2003.

[8] I. J. Nagrath and M. Gopal, *Control Systems Engineering*, 2nd ed. New York, NY, USA: Halsted Press, 1982.

[9] t. b. C. H. D. Gauss, KF., *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections*. Dover Publications, 1809.

[10] J. Aldrich *et al.*, "Ra fisher and the making of maximum likelihood 1912-1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176, 1997.

[11] S. K. Varanasi and P. Jampana, "Parameter estimation and model order identification of lti systems," *IFAC-PapersOnLine*, vol. 49, no. 7, pp. 1002–1007, 2016.

[12] D. Ghoshal, K. Gopalakrishnan, and H. Michalska, "Algebraic parameter estimation using kernel representation of linear systems - submitted," in *IFAC World Congress, 2017. 20th World Congress*, 2017.

[13] R. Kalman, "On the general theory of control systems," *IRE Transactions on Automatic Control*, vol. 4, no. 3, pp. 110–110, 1959.

[14] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[15] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Journal of basic engineering*, vol. 83, no. 1, pp. 95–108, 1961.

[16] R. E. Kalman, "Mathematical description of linear dynamical systems," *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, vol. 1, no. 2, pp. 152–192, 1963.

[17] D. G. Luenberger, "Observing the state of a linear system," *IEEE transactions on military electronics*, vol. 8, no. 2, pp. 74–80, 1964.

[18] D. Luenberger, "Observers for multivariable systems," *IEEE Transactions on Automatic Control*, vol. 11, no. 2, pp. 190–197, 1966.

[19] G. David, "An introduction to observers," *IEEE Transactions on automatic control*, vol. 16, no. 6, pp. 596–602, 1971.

[20] H. Sira-Ramirez and S. K. Agrawal, *Differentially flat systems*. Crc Press, 2004.

[21] S. Diop and M. Fliess, "Nonlinear observability, identifiability, and persistent trajectories," in *Decision and Control, 1991., Proceedings of the 30th IEEE Conference on*. IEEE, 1991, pp. 714–719.

[22] S. Diop, J. Grizzle, P. Moraal, and A. Stefanopoulou, "Interpolation and numerical differentiation for observer design," in *Proceedings of the American Control Conference*, vol. 2. AMERICAN AUTOMATIC CONTROL COUNCIL, 1994, pp. 1329–1329.

[23] S. Diop, J. Grizzle, and F. Chaplais, "On numerical differentiation algorithms for nonlinear estimation," in *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*, vol. 2. IEEE, 2000, pp. 1133–1138.

[24] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*. MIT press Cambridge, 1949, vol. 2.

[25] D. P. Ghoshal, K. Gopalakrishnan, and H. Michalska, "Kernel-based adaptive multiple model target tracking," in *IEEE Conference on Control Technology and Applications*, 2017.

[26] S. Saitoh, *Theory of reproducing kernels and its applications*. Longman, 1988, vol. 189.

[27] M. Fliess and H. Sira-Ramírez, "An algebraic framework for linear identification," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 9, pp. 151–168, 2003.

[28] M. Eslami, *Theory of sensitivity in dynamic systems: an introduction*. Springer Science & Business Media, 2013.

[29] Y. Cao, S. Li, L. Petzold, and R. Serban, "Adjoint sensitivity analysis for differential-algebraic equations: The adjoint dae system and its numerical solution," *SIAM Journal on Scientific Computing*, vol. 24, no. 3, pp. 1076–1089, 2003.

[30] A. C. Hindmarsh and R. Serban, "User documentation for cvodes, an ode solver with sensitivity analysis capabilities," *Livermore (California): Lawrence Livermore National Laboratory*, vol. 189, 2002.

[31] M. Ziani and F. Guyomarc'h, "An autoadaptative limited memory broyden's method to solve systems of nonlinear equations," *Applied Mathematics and Computation*, vol. 205, no. 1, pp. 202–211, 2008.

[32] V. I. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.* Cambridge University Press, 2016.

[33] P. Bartlett, " Instructor Notes: Reproducing Kernel Hilbert Spaces," 2008. [Online]. Available: https://people.eecs.berkeley.edu/ bartlett/courses/281b-sp08/7.pdf

[34] L. Ljung, "System identification: Theory for the user, ptr prentice hall information and system sciences series," *ed: Prentice Hall, New Jersey*, 1999.

[35] K. J. Åström and P. Eykhoff, "System identification—a survey," *Automatica*, vol. 7, no. 2, pp. 123–162, 1971.

[36] G. Goodwin and K. Sin, "Adaptive filtering prediction and control. 1984," *Englewood Clifs: Prentice Ha ll*.

[37] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation.* Prentice Hall Upper Saddle River, NJ, 2000, vol. 1.

[38] E. Kalnay, *Atmospheric modeling, data assimilation and predictability.* Cambridge university press, 2003.

[39] J. L. Anderson, "An ensemble adjustment kalman filter for data assimilation," *Monthly weather review*, vol. 129, no. 12, pp. 2884–2903, 2001.

[40] T. Kailath, "Lectures on wiener and kalman filtering," in *Lectures on Wiener and Kalman Filtering.* Springer, 1981, pp. 1–143.

[41] M. Fliess and H. Sira-Ramírez, "Closed-loop parametric identification for continuous-time linear systems via new algebraic techniques," in *Identification of Continuous-time Models from sampled Data.* Springer, 2008, pp. 363–391.

[42] M. Mboup, C. Join, and M. Fliess, "Numerical differentiation with annihilators in noisy environment," *Numerical algorithms*, vol. 50, no. 4, pp. 439–467, 2009.

[43] S. Bhattacharyya, "Parameter invariant observers," *International Journal of Control*, vol. 32, no. 6, pp. 1127–1132, 1980.

[44] P. Martin and E. Salaun, "Invariant observers for attitude and heading estimation from low-cost inertial and magnetic sensors," in *Decision and Control, 2007 46th IEEE Conference on*. IEEE, 2007, pp. 1039–1045.

[45] L. Elsgloc, *Calculus of Variations*. Addison-Wesley Publishing Company, 1962.