-

Identification of Block-Oriented Nonlinear Systems Based on Input-Output Data

Li Guoqi

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirements for the degree of
Doctor of Philosophy

2011

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

....

Date Li Guoqi

Acknowledgement

First and foremost I would like to thank my supervisor, Dr. Wen Changyun, Professor of School of EEE, Nanyang Technological University, Singapore, who opened the door of scientific research for me. He gave me enormous guidance and help and offered me lots of chances and spaces to think independently. I appreciate the wisdom of his way of guidance, his attitude toward research as well as his admirable patience, from which I would benefit in all my life.

I would like to thank the school of EEE, Nanyang Technological University, Singapore, for the financial support during my Ph.D study. They also have kindly provided travel funds to attend overseas and local conferences.

I would like to thank Professor Zheng Wei Xing, School of Computing and Mathematics, University of Western Sydney, Australia, for his advices, suggestions and helps. I wish also to thank Professor Huang Guangbin and Professor Mao Kezhi, School of EEE, Nanyang Technological University, Singapore, for their suggestions and comments.

I wish to thank Dr. Li Zhengguo, Institute for Infocomn Research, Singapore. Thanks also go to Professor Zheng Yunfeng, a visiting professor from Dalian Maritime University, P. R. China. Many thanks to my former instructors Professor Zhang Aimin and Professor Zhao Guangshe form Xi'an Jiaotong University, P. R. China.

Singapore Singapore

I wish to express my sincere gratitude to my friends, Ms. Chen Yan, Mr. Yang Feng, Ms. Wang Wei, Mr. Han Shuchu, Mr. Wang Liang, Mr. Zhou Yong, Ms. Gao Tingting, Mr. Cui Song, Mr. Luo Dan, Mr. Huang Jiangshuai, Mr. Wang Junyan, Mr. Lu Jiwen and Mr. You Keyou for their kind friendships.

Special thanks must go to my parents, Mr. Li Shiqiu and Ms. Chen Shuie, for their unconditional support and encouragement throughout all these years.

Summary

This thesis deals with the identification of block-oriented nonlinear systems. Block-oriented nonlinear systems are composed of linear time-invariant dynamic systems and nonlinear static functions, which are interconnected in different ways. Hammerstein systems, Wiener systems and Hammerstein-Wiener systems as well as Wiener-Hammerstein systems are some well known examples of block-oriented systems. The main achievements are in the development of new models and new algorithms in identifying block-oriented nonlinear systems shown as follows.

Firstly, we propose a new class of block-oriented model and a new approach to identify the model based on kernel machine and space projection (KMSP). The well known Hammerstein-Wiener model is a subset of the proposed model. In the KMSP based approach, we use kernel machine to represent the functions and space projection to separate the represented functions. The asymptotic behavior of the proposed approach is analyzed.

Secondly, a long time open problem that an iterative identification algorithm in identifying block-oriented systems such as Hammerstein systems needs a proper initial condition to guarantee its convergence is solved in this thesis. To achieve this, we propose a new iterative algorithm by fixing the norm of the parameter estimates. The proofs of the method give a geometrical explanation on why the normalization guarantees the convergence.

Singapore Singapore

Thirdly, we introduce fixed point iteration to identifying both Hammerstein and Wiener systems. A unified iterative algorithm is proposed inspired from fixed point theory and the convergence is guaranteed. It is shown that the iteration is a contraction mapping on a metric space when the number of input-output data points approaches infinity. This implies the existence and uniqueness of a fixed point of the iterated function sequence and thus ensures the convergence of the iteration.

Fourthly, we consider identifying bilinear models which is more general than Hammerstein and Wiener systems based on fixed point iteration. As an application, a block-oriented system represented by a cascade of a dynamic linear (L), a static nonlinear (N) and a dynamic linear (L) subsystems is illustrated. This gives a solution to the long-standing convergence problem of iteratively identifying LNL Wiener-Hammerstein models. In addition, we extend the static nonlinear function (N) to a nonparametric model represented by using kernel machine.

Fifthly, we point out that the identification of block-oriented nonlinear systems can be formulated as a biconvex optimization problem. To achieve this, a common model is proposed to represent a class of block-oriented systems. Then it is shown that identifying the common model can be formulated as a biconvex optimization problem, where we only need to find the unique partial optimum point of a biconvex cost function in the formulated optimization problem to obtain its global minimum point. A normalized alterative convex search (NACS) algorithm is presented. Its convergence property is also established, which provides a unified framework for the iterative identification of block-oriented systems.

In addition to the above mentioned contributions, we also explore the area of identifying block-oriented systems such as Wiener systems with binary quantized observations. We propose a classification based support vector machine (SVM) and

formulate the identification problem as a convex optimization. The solution to the optimization problem converges to the true parameters of the linear system if it is a finite impulse response (FIR) system. In identifying a Wiener system with a stable infinite impulse response (IIR) system, an FIR system is used to approximate it and the problem is converted to identify the FIR system together with solving a set of nonlinear equations. This leads to biased estimates of parameters in the IIR system while the bias could be controlled by choosing the order of the approximated FIR system.

Contents

Acknowledgement **i**

Summary **iii**

List of Figures **xii**

1 Introduction **1**

1.1 Literature Review and Motivations 5

1.2 Organization of the Thesis and Major Contributions 7

2 Identification of Block-oriented Systems Using Kernel Machine and Space Projection **9**

2.1 Introduction 10

2.2 Motivation and Problem Formulation 13

2.2.1 Hammerstein-Wiener Models 13

2.2.2 A New Class of Nonlinear Systems 14

2.3 Kernel Machine and Space Projection in Nonlinear System Identification 17

2.3.1 Kernel Machine for Function Approximation 17

2.3.2 A Fundamental Model 23

2.3.3 Identification of Fundamental Model Based on Space Projection 25

2.3.4 Comparison Between Standard Least Square Estimation Method and the Proposed Space Projection Method 28

2.4 Identification of the New Model 30

2.4.1 Model Transformation 30

2.4.2 System Identification Based on the Transformed Models Using Space Projection 33

2.5 Ambiguity Analysis 38

2.5.1 Two Kinds of Ambiguities 38

2.5.2 Conditions for Avoiding Ambiguities 39

2.6 Results on Asymptotic Behavior 40

2.7 Simulation Results 42

2.8 Conclusion 52

3 Iterative Identification of Hammerstein Systems by Normalization 54

3.1 Introduction 54

3.2 Normalized Iterative Algorithm of a Hammerstein System 56

3.3 Convergence Analysis of the Iterative Algorithm 63

3.4 Illustrative Examples 70

3.5 Conclusion 75

4 Convergence of Fixed Point Iteration for the Identification of Hammerstein and Wiener Systems 76

4.1 Introduction 77

4.2 Fixed Point Iteration Algorithm for Hammerstein and Wiener Systems 78

4.2.1 Hammerstein and Wiener Systems 78

4.2.2 Fixed Point Iteration Algorithm 83

4.3 Convergence Analysis 87

4.4 Examples 92

4.5 Conclusion 96

5 Fixed Point Iteration for The Identification of Bilinear Models 97

5.1 Introduction 97

5.2 Bilinear Models and Fixed Point Theory 99

5.2.1 Bilinear Models 100

5.2.2 Iterative Identification Algorithm 103

5.2.3 Convergence Analysis 105

5.3 Identification of LNL Wiener-Hammerstein Models 115

5.3.1 Kernel Machine for Function Approximation 117

5.3.2 Model Transformation 118

5.3.3 Convergence Results 122

5.3.4 Extension to IIR Linear Systems 123

5.4 Example Illustration 125

5.5 Conclusion 129

6 Identification of Block-oriented Systems Based on Biconvex Optimization 130

6.1 Introduction 131

6.2 Biconvex Optimization Problem 132

6.2.1 Definition of Biconvex Optimization 132

6.2.2 Alterative Convex Search Algorithm (ACS) 134

6.3 Biconvex Optimization in Parameter Estimation 136

6.3.1 A Common Model 136

6.3.2 Identifying the Common Model 138

6.4 Convergence Analysis 140

6.5 Applications in the Identification of Block-oriented Systems 143

6.5.1 A New Class of Block-oriented Nonlinear Systems 143

6.5.2 Transformation to the Common model 144

6.6 Discussion of Model Generalization 149

6.7 Simulation Results 151

6.8 Conclusion 152

7 Identification of Wiener Systems with Clipped Observations 154

7.1 Introduction 155

7.2 Problem Formulation with Two-classes Classification SVM 157

7.3 Convergence Analysis 161

7.3.1 Convergence Analysis for Noise Free Case 162

7.3.2 Convergence Analysis in the Presence of Noise 164

7.4 Wiener Model with IIR Linear System Identification 169

7.4.1 Wiener Model with IIR Linear System 169

7.4.2 Solution of Nonlinear Equations by Using Trust Region Algorithm 170

7.5 Comparisons and Simulation Illustration 172

7.5.1 Comparisons Between Regression Based LS-SVM and Classification Based SVM in the Identification of the Wiener System with an FIR System 172

7.5.2 An Example of Wiener System with an IIR System 173

7.6 Summary 174

8 Conclusions and Future Works 175

Author's Publications 178

Bibliography **181**

List of Figures

1.1 Hammerstein systems 2

1.2 Wiener systems 3

1.3 Hammerstein-Wiener systems 4

1.4 Wiener-Hammerstein systems 5

2.1 Hammerstein-Wiener model 13

2.2 A new class of block-oriented nonlinear systems 15

2.3 Input sequence 43

2.4 Output sequence 44

2.5 Support vector sequence 45

2.6 True input nonlinear static function and estimated function 46

2.7 True output nonlinear function and estimated function 47

2.8 The change of error for the estimated function with respect to the data points N 48

2.9 The change of error for the estimated function with respect to the data points N 49

2.10 True input nonlinear static function f_1 and estimated function value for the generalized Hammerstein-Wiener model 50

2.11 The change of error for the estimated parameters with respect to the data points N 51

3.1 The geometrical illustration of the neighborhood of $\tilde{a}$ when $\tilde{b}$ is fixed 66

3.2 Estimation error with respect to the number of data points N . . . 72

3.3 The illustration that the iteration algorithm converges in a few iterations 73

3.4 Estimation error with respect to number of data points 74

4.1 The block diagram of Wiener systems 78

4.2 The illustration that fixed point iteration algorithm converges in a few iterations 93

4.3 Estimation error with respect to number of data points 94

4.4 The illustration that fixed point iteration algorithm converges in a few iterations 95

5.1 Estimates with respect to number of iterations k ($N = 20$) 126

5.2 Estimation error with respect to number of data points N 127

5.3 Illustration of convergence 128

5.4 True nonlinear function and estimated function 128

5.5 Estimation error respect to number of data points N 129

6.1 Examples of biconvex set which are non-convex 133

6.2 The change of error for the estimated parameters respect to N . . . 153

7.1 Wiener systems with quantized observations 157

7.2 Possible misclassification region when $h(\mathrm{x})$ is not close to $h^*(\mathrm{x})$ . . 162

Chapter 1

Introduction

System identification is actually a particular process of statistical inference based on two types of information. The first type of information is called a prior knowledge which is known before identification. The priori information concerns some general knowledge about the system, e.g., whether the system is continuous or discrete time, dynamic or static, the structure of the system is known or not and so on. The second type of information basically concerns the measured data, e.g., the designed inputs and observed outputs. In other words, system identification is to build mathematical models for identifying dynamic systems using statistical methods based on certain prior knowledge and measured input output data.

Mathematical models are often classified into black-box models, white-box models or gray-box models, according to how much a priori information is available from the system. A black-box model is a system without a priori information available. A white-box model (also called glass-box or clear-box) is a system with all necessary information available. Practically all systems are somewhere between the black-box and white-box models. Although the peculiarities of what is going on inside the system are not entirely known, a model based on both insight about

the system and experimental data can be constructed. However, this model still has a number of unknown free parameters which can be estimated using system identification. Such a model is called grey-box model.

A dynamical mathematical model in this thesis is a mathematical description of the dynamic behavior of a system or process in the time domain. We deal with system identification based on grey-box models. More particularly, we concern the identification of partially known block-oriented [1] nonlinear systems. They are composed of different blocks such as linear time-invariant dynamic systems and nonlinear static functions, which are interconnected in different ways. However, there are a number of unknown parameters in each block. Block-oriented systems can capture a large class of complex and nonlinear systems and have motivated a great deal of interest paid to them over the past twenty years. It is worth noting that the model (linear and nonlinear blocks) may not correspond to certain physical components. Consequently, the connection points between blocks are generally artificial. That is to say, they cannot be supposed to be accessible for measurements. The inaccessibility of such measurement points, together with the system nonlinearities, makes block-oriented system identification a quite complex problem. Therefore, most currently available solutions only concern relatively simple structures.

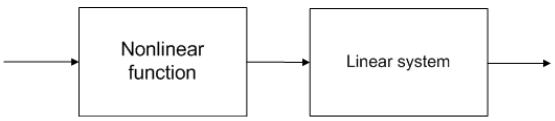

Figure 1.1: Hammerstein systems

In the identification of a block-oriented system, the very first step is to assume a

structure of the system, i.e., how many blocks are there in the system and what are their orders of connection. The second step is to design the input data and collect the output data. By employing different approaches, the final objective is to identify the block-oriented system with known structures and unknown parameters based on input output data.

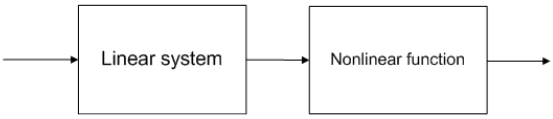

Figure 1.2: Wiener systems

The simplest and most well-known block-oriented nonlinear structures are composed of just two blocks connected in series as shown in Figures 1.1 and 1.2. The first one, the Hammerstein system, introduced in 1930 by the German mathematician A. Hammerstein [2], involves one input static nonlinear element in series with a dynamic linear subsystem. The nonlinear element may account for actuator nonlinearities and other nonlinear effects that can be brought to the system input. Despite their simplicity, Hammerstein models have been proved to be able to accurately describe a wide variety of nonlinear systems, for examples, chemical processes [3], electrically stimulated muscles [4], power amplifiers [5], electrical drives [6], thermal microsystems [7], physiological systems [8], sticky control valves [9], solid oxide fuel cells [10], and magneto-rheological dampers [11].

The permutation of the linear and nonlinear elements in the Hammerstein model leads to what is commonly referred to the Wiener model in Figure 1.2, as a model of this type was first studied by N. Wiener in 1958 [12]. In this model, the output nonlinear element may represent sensor nonlinearities as well as any nonlinear effects that can be brought to the system output. For instance, limit switch devices

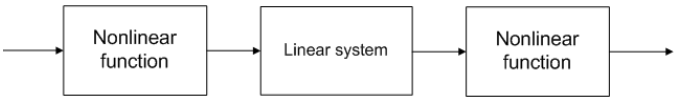

Figure 1.3: Hammerstein-Wiener systems

in mechanical systems and overflow valves can be modeled by output saturating nonlinearities. Moreover, the ability of Wiener models to capture complex nonlinear phenomena has been formally established. In this regard, it was shown that almost any nonlinear system can be approximated by a Wiener model with an arbitrarily high accuracy [13]. This theoretical fact has been experimentally verified through several practical applications, for examples, chemical processes [14] [15], biological systems [16] and others. A series combination of a Hammerstein and an Wiener model immediately yields a new model structure called the Hammersteir-Wiener system shown in Figure 1.3. The inverse combination leads to what is referred to as the Wiener-Hammerstein structure shown in Figure 1.4. These new structures offer higher modeling capabilities. Clearly, the Hammersteir-Wiener model is more convenient when both actuator and sensor nonlinearities are present. It has also been successfully applied to modeling several physical processes, for examples, polymerase reactors [17], ionospheric processes [18], PH processes [19], magnetospheric dynamics and so on. The Wiener-Hammerstein model (Figure 1.4) also finds applications. Since block-oriented systems have been widely applied in practice, different methodologies focusing on the identification of these systems have been generously and extensively researched. In this thesis, we focus on the identification of block-oriented systems such as Hammerstein systems, Wiener systems, Hammerstein-Wiener systems, Wiener-Hammestein systems and even more complicated systems.

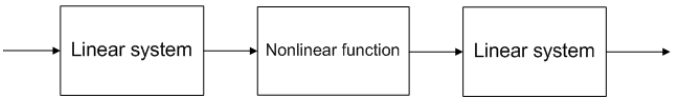

Figure 1.4: Wiener-Hammerstein systems

1.1 Literature Review and Motivations

As mentioned, the identification of block-oriented nonlinear systems has been extensively studied in recent two decades. Existing methods mainly include the over-parametrization method [24], the non-parametrization method [56] [57] [58], the stochastic method [53] [54] [50], the subspace and least-squares method [25] [26] and the iterative method [49] [51] [64] [70] [48].

Over-parametrization considers an over-parametric representation of block-oriented nonlinear systems. This method uses an optimal two stage identification method combining a least squares parameter estimation and a singular value decomposition of two matrices whose dimensions are fixed and do not increase as the number of the data point increases. The algorithm is shown to be convergent in the absence of noise and convergent with probability one in the presence of white noise.

The meaning of non-parametric methods covers techniques that do not assume that the structure of a model is fixed. Typically, the model grows in size to accommodate the complexity of the data. In [56], the nonparametric approach to block-oriented system identification was introduced by Greblicki and Pawlak. Kernel regression estimation [56] or the employing of the orthogonal series expansion is used to estimate the nonlinear static function of a Hammerstein or a Wiener system, which reduces the system to a linear system.

The stochastic method such as the maximum likelihood algorithm in [22] and [23] is introduced to identify Wiener systems. The noise is even allowed to be coloured making possible blind estimation of Wiener systems.

The subspace and least-squares method is introduced by Goethals et al [25] [26] where LS-SVM techniques are presented in the identification of Hammerstein, Wiener or Hammerstein-Wiener systems. The method is essentially based on the overparametrisation technique, and combines this with a regularisation framework and a suitable model description which fits nicely within the LS-SVM framework with primal and dual model representations.

Last but not least, the iterative method, which divides the unknown parameters into two sets, the linear part and the nonlinear part. At each iteration, one set of estimates is computed while the other set is fixed. Then the two sets alternate and their final parameter estimates are obtained iteratively. Such an iterative algorithm was first proposed to estimate Hammerstein systems by Narendra & Gallman in 1966 [51]. Since then, the idea has been used to identify Hammersteim systems in [49] [70] [48].

Though many methods have been proposed and analyzed, there are still a number of open problems in the identification of block-oriented nonlinear systems. Particularly in this thesis, we will address the following problems.

1) How to propose new identification algorithms which are different from the above existing schemes? The proposed new algorithms with different ideas should overcome certain disadvantages of the existing ones.

2) Generally speaking, block-oriented nonlinear system contain Hammerstein systems, Wiener systems, Hammerstein-Wiener systems and Wiener-Hammerstein systems. Is it still possible to consider a more general system which includes

Hammerstein-Wiener or Winner-Hammerstein systems as its special cases? On the other hand, the generalization should also relax the assumptions on both linear and nonlinear blocks. For example, for the identification of Wiener systems, how to allow that the linear systems are IIR systems and the static function is a non-invertible function.

3) If the block-oriented systems can be generalized to more general models, then how to identify such models.

4) Currently the convergence of some existing schemes are still unproven. For example, the convergence of the iterative identification is unknown even for a general Hammerstein system. Note that the convergence in developing new algorithms is essential. It is important to show how to guarantee the convergence of the proposed algorithms, i.e., how to ensure the estimates converges to their true values.

In this thesis, we shall find solutions to the above problems. The organization of the thesis and its contributions are summarized in the next subsection.

1.2 Organization of the Thesis and Major Contributions

The following are the main contributions of the thesis:

1) In Chapter 2, we propose a new class of block-oriented systems which includes Hammerstein-Wiener systems. A new algorithm called kernel machine and space projection method is proposed to identify the newly proposed model.

2) In Chapter 3, we propose a new iterative algorithm for a general Hammerstein system and prove its convergence. We also give a geometrical explanation of why the convergence property can be achieved.

3) In Chapter 4, we introduce fixed point iteration algorithm to identify both Hammerstein and Wiener systems. A unified iterative algorithm is proposed inspired from fixed point theory and the convergence is guaranteed. It is shown that the iteration process is a contraction mapping on a metric space when the number of input-output data points approaches infinity.

4) In Chapter 5, we formulate a new general bilinear model which actually represents a class of Wiener-Hammerstein systems. This new general bilinear model includes Hammerstein and Wiener systems as its special cases. The iterative algorithm is proposed based on the fixed point iteration which is shown to be convergent. This gives a new point of view in proving the convergence property in identifying block-oriented systems.

5) In Chapter 6, we extend the iterative algorithm to our newly proposed block-oriented systems in Chapter 2. A new common model is proposed which actually represents the newly proposed block-oriented systems. Biconvex optimization is introduced to such systems.

6) In Chapter 7, we also consider the identification of block-oriented nonlinear systems based on clipped (binary quantized) observations. For the first time, SVM for classification is introduced to identify block-oriented nonlinear systems such as Wiener systems with clipped observations.

7) Finally conclusions and suggestions for future research are given in Chapter 8.

Chapter 2

Identification of Block-oriented Systems Using Kernel Machine and Space Projection

In this chapter, we propose a new class of block-oriented systems which is more general than Hammerstein-Wiener systems and a new algorithm to identify the newly proposed models. The new algorithm is called kernel machine and space projection (KMSP), where kernel machine is used to represent the functions and space projection to separate the represented functions. We also discuss two possible ambiguities and give conditions to avoid such ambiguities. The asymptotic behavior of the proposed approach is analyzed. The performance of the proposed method is substantiated by simulation studies.

2.1 Introduction

Existing methods for Hammerstein or Wiener model identification can be broadly divided into three categories when refers to the degree of parametrization: the parametric method [20] [21] [24] [25] [26], the nonparametric method [27] [28] [29] [30] [31], and the semi-parametric method [32]. In order to approximate the nonlinear function, in the parametric method, the function is usually assumed to be a polynomial with a fixed order. Then the objective becomes to estimate the corresponding coefficients of the polynomial. If the nonlinear function is not a polynomial, the parametric method is unable to guarantee that the estimated function converges to the true function [33]. In [34], a method based on line segments is used to approximate the nonlinear function. In the nonparametric method, approaches based on Fourier series, polynomial series including Laguerre, Legendre or Hermite polynomials, the wavelet series and so on have been developed [35] [36] to approximate the nonlinear function. The function basis is fixed before the identification process starts. In [30] a nonparametric kernel regression estimator is proposed to approximate the nonlinear static function, without the requirement of knowing the basis functions in advance. In the semi-parametric approach, parametric methods are developed for the linear subsystem while nonparametric ideas are used to identify the nonlinear function [32], based on the theory of partial linear models [37] [38]. It is noted that all these methods are applicable to either the Hammerstein model or the Wiener model. As for the identification of Hammerstein-Wiener models, only several results have been reported (see, e.g., [24], [39], [40], [41]). In these schemes, the class of nonlinear static functions is limited.

In this chapter, we will consider the identification of a class of block-oriented nonlinear systems which belongs to nonlinear autoregressive models with exogenous

inputs (NARX). Generalized from Hammerstein-Wiener models, more than one input nonlinear functions may be allowed. Inspired by the idea of approximating nonlinear functions by support vector machines, we propose a new identification scheme, named the kernel machine and space projection (KMSP) method. Note that there have been reports on 'kernel' regression based approaches such as the methods proposed in [30]. However, their ideas are based on the method of interpolation, while ours is motivated by the kernel machine in support vector machine (SVM) in [42] and [43] . Thus the resulting approaches are different. In nonlinear system identification based on kernel machines, a kernel machine is a powerful tool to transform nonlinear relationships into linear relationships in a higher dimensional space. Then what remains is how to solve the transformed problem, which is important and challenging. There are also schemes proposed for identification of nonlinear dynamic systems based on support vector machines, see, for example, [44], [45]. For these schemes, it is stated that an implicit or explicit formulation of the auto-regressive and moving average model (ARMA) data structure was introduced in a reproducing kernel Hilbert space (RKHS) by using kernel machines. In the formulated structure, both the autocorrelation and cross correlation are taken into account. So this method turns out to work well in identifying Hammerstein, Wiener, and Hammerstein-Wiener systems. Both approaches in [44], [45] and in this chapter use kernel machines as a transformation tool. It is noted that for the identification method in [44] and [45], the main objective is to ensure the output of the identified model to track the output of the actual system. In our case, we employ kernel machines only to approximate the static nonlinear function instead of the whole dynamic system. Thus, our proposed method can estimate the parameters in the linear subsystems and also the nonlinear static functions. Therefore, more detailed structure of the system is explored with our approach. In addition, we can achieve the same output tracking objective shown in [44] and

[45]. It is also noted that in some cases, such as Example 2.7.2 to be given in the simulation section, certain input data required by the method in [44] and [45] may not be available.

The newly proposed KMSP method allows for the estimation of the static nonlinear function as well as the parameters in the linear system. One closely related approach is the schemes proposed by Goethals et. al [25] [26] [40]. In this approach, one kind of kernel machines, least squares support vector machine (LS-SVM), was proposed to identify Hammerstein, Wiener, Hammerstein-Wiener models. The differences between the KMSP and LS-SVM based identification methods are mainly in the following aspects. Firstly, in the LS-SVM method, every data point is a support vector while in the KMSP method only a subset of data points need to be support vectors, which leads to a different representation of nonlinear functions and less computational cost. Secondly, though both methods use kernel machines as a tool to transform the model to a solvable problem, the ways to solve the transformed problem are quite different. In the LS-SVM method, singular value decomposition (SVD) or kernel canonical correlation analysis (KCCA) was used, while a space projection approach is proposed in this chapter. Thirdly, our method can handle a more general class of models. For example, we can estimate multiple input functions (including saturation, deadzone, quantization, signum functions) and an output function. Note that such functions have never been addressed before. In this chapter, we also analyze possible ambiguities which may occur in the identification process, and propose some conditions to avoid such ambiguities. Asymptotic behaviors of the proposed KMSP method are further established.

The remaining part of this chapter is organized as follows. In Section 2.2, we present a new class of nonlinear systems to be identified. The idea of KMSP is introduced together with the derivation of a fundamental model in Section 2.3.

In Section 2.4, we present the new identification scheme of the nonlinear systems. Section 2.5 is devoted to the analysis of ambiguities possibly existing in parameter identification, while Section 2.6 is concerned with studying asymptotic behaviors of the proposed identification algorithm. Some simulation examples are given in Section 2.7 to show the performance of the proposed KMSP algorithm. Finally, this chapter is concluded in Section 2.8.

2.2 Motivation and Problem Formulation

2.2.1 Hammerstein-Wiener Models

Hammerstein and Wiener models belong to a specific class of nonlinear systems. As shown in Figure 2.1, if the linear subsystem is preceded by a nonlinear static function, it is a Hammerstein model; if followed by a nonlinear static function, it is a Wiener model; and if the linear subsystem is between the two nonlinear static functions, the system becomes a Hammerstein-Wiener model. Even though

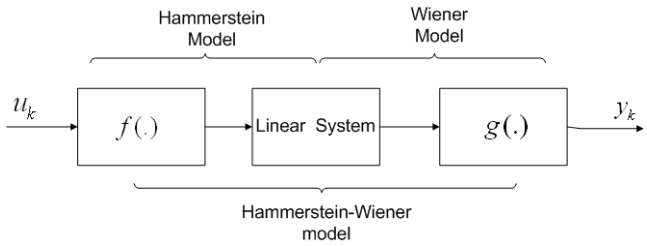

Figure 2.1: Hammerstein-Wiener model

Hammerstein-Wiener models represent a fairly large class of models in modeling practical nonlinear systems, we still feel that they are not sufficient to represent some more general nonlinear systems. In this chapter, we will consider a new class of nonlinear systems generalized from the Hammerstein-Wiener models.

2.2.2 A New Class of Nonlinear Systems

In this subsection, we present the new class of nonlinear systems to be identified, which is generalized from the Hammerstein-Wiener models. The system consists of three blocks as shown in Figure 2.2. The first block is a nonlinear subsystem which contains a number of paths. In each path, there is a nonlinear function followed by a linear gain. This block may be considered as a generalization of the Hammerstein model in which the linear subsystem is the moving average part. The second block is a linear subsystem described by an autoregressive model and the last block is an output nonlinear static function. The unknown nonlinear system can be represented by the following equations with nonlinear input functions:

$$\begin{aligned} z_k = \; & a_1 z_{k-1} + a_2 z_{k-2} + \ldots + a_n z_{k-n} \\ & + b_0 f_0(u_k) + b_1 f_1(u_{k-1}) + \ldots + b_m f_m(u_{k-m}) + v_k \end{aligned} \tag{2.1}$$

and output nonlinear function

$$y_k = g(z_k) \tag{2.2}$$

where $\{u_k\}$ and $\{y_k\}$ are the input and output sequences respectively, v_k denotes the random noise, n and m are the orders of the system, $r = \max(n, m) + 1$ and $k = r, r+1, \ldots, N$.

Our objective is to estimate the parameters in the linear subsystem and the nonlinear static functions in the NARX model described by (2.1) and (2.2). Note that when $f_0(.), \ldots, f_m(.)$ are all different from each other, there is no need to estimate $b_0, \ldots, b_m$. Here two functions are meant different if they are different in a region which has a nonzero measure. In this case, we estimate the parameters $a_1, \ldots, a_n$, nonlinear static functions $f_0(.), \ldots, f_m(.)$ and $g(.)$. Only when $f_0(.) = \ldots = f_m(.)$, estimating $b_0, \ldots, b_m$ becomes meaningful. To characterize the class of systems

and identify them, we make the following assumptions.

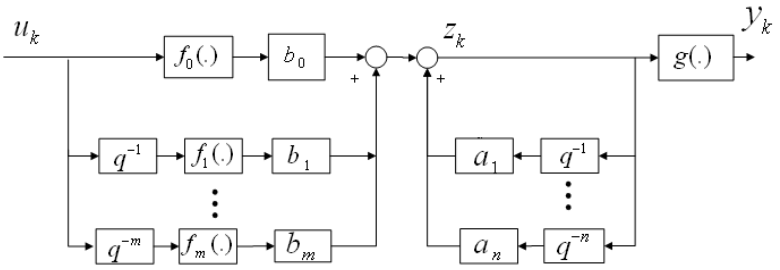

Figure 2.2: A new class of block-oriented nonlinear systems

Assumption 2.1. *Input $u_k \in [-C, C]$, where $C > 0$ is a constant and u_k is an i.i.d random variable with a probability density function $p_u(u)$. Noise v_k is i.i.d with zero mean and finite variance σ_v^2.*

Assumption 2.2. *All the nonlinear functions are static functions. The inverse of the output nonlinear function $g(.)$ exists, i.e., $z = g^{-1}(y)$.*

Since $z_k = g^{-1}(y_k)$, the model is represented as

$$\begin{aligned} g^{-1}(y_k) = a_1 g^{-1}(y_{k-1}) + \ldots + a_n g^{-1}(y_{k-n}) + b_0 f_0(u_k) \\ + b_1 f_1(u_{k-1}) + \ldots + b_m f_m(u_{k-m}) + v_k \end{aligned} \tag{2.3}$$

The available input-output data are $\{u_k, y_k\}_{k=r}^{N}$. The model in (2.3) is expressible as

$$[g^{-1}(y_r) \ \ldots \ g^{-1}(y_N)]' = \Phi\tau + v \tag{2.4}$$

where $\tau = [a_1 \ \ldots \ a_n \ b_0 \ \ldots \ b_m]'$, $v = [v_r \ \ldots \ v_N]'$ denotes the noise vector, the

superscript $'$ stands for transpose operation, and

$$\Phi = \begin{bmatrix} g^{-1}(y_{r-1}) & \dots & g^{-1}(y_{r-n}) & f_0(u_r) & \dots & f_m(u_{r-m}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ g^{-1}(y_{N-1}) & \dots & g^{-1}(y_{N-n}) & f_0(u_l) & \dots & f_m(u_{N-m}) \end{bmatrix}. \tag{2.5}$$

Assumption 2.3. *Input functions $f_i \in \mathfrak{F}$, $i = 0, \dots, m$, where $\mathfrak{F}$ denotes the set of any function whose derivative is absolutely integrable. In addition, the input and output functions ensure that Φ in (2.5) is of full column rank with inputs satisfying Assumption 2.1.*

Remark 2.1. *The given assumptions are more relaxed when compared with the existing schemes such as in Hammerstein-Wiener system identification [24][41]. The functions can be discontinuous and the input functions are extended to be measurable functions including saturation, deadzone, quantization, signum functions and so on. These functions may have zero measure discontinuities and they can be uniformly approximated by a sequence of continuous differentiable functions. For the discussion of the function set $\mathfrak{F}$, one could refer to [46]. Note that not all functions in $\mathfrak{F}$ can ensure the full column rank of Φ, for example, if $f_0, \dots, f_m$ are constant functions.*

Remark 2.2. *In the NARX model (2.1) and (2.2), if $f_0(u) = \dots = f_m(u)$, the model reduces to a Hammerstein-Wiener model. If $z = g(y) = y$, the model is a Hammerstein model. If $f_0(u) = \dots = f_m(u) = u$, the model becomes a Wiener model.*

2.3 Kernel Machine and Space Projection in Nonlinear System Identification

To solve the identification problem, appropriate expressions to represent functions $z = g^{-1}(y)$ and $f_0(u), \ldots, f_m(u)$ are important. In this section, we first use kernel machines to transform the system model, and then use space projection to identify the transformed model. This new identification approach is called the Kernel Machine and Space Projection (KMSP) method. To begin with, we introduce the theoretical basis of KMSP.

2.3.1 Kernel Machine for Function Approximation

A proper nonparametric representation of a nonlinear static function $y = f(x)$, $x \in R^{\mathcal{R}}$ where $\mathcal{R}$ denotes the dimension of x, with a kernel machine is employed here for estimating it. Such a representation has been widely used, see for example [59] [43]. Let $(x_i, y_i) \in R^{\mathcal{R}} \times R$ for $i = 1, ..., N$ be the inputs and the output of the nonlinear function $f(x)$ and $\{x_i\}_{i=1}^{N}$ be sampled i.i.d from its probability density function. Let the set $sv = (\tilde{x}_j)_{j=1}^{m_{sv}}$, where m_{sv} is the number of support vectors, be support vectors [59] [43] which are i.i.d sampled from $R^{\mathcal{R}}$. Note that usually $m_{sv} << N$, which means that the support vectors are sparse compared with the training data. Let $\{k(., \theta) : \theta \in \Theta\}$ be a family of bases on a compact set parameterized over the set $\Theta = R^{\mathcal{R}} \times R$. For example, we will consider a Gaussian kernel function $k(x, \theta) = k(x, \tilde{x}_i, \rho) = e^{-(x-\tilde{x}_i)^2/\rho^2}$ with $\theta = [\tilde{x}_i, \rho]$. Note that ρ is a user chosen parameter, so we write $k(x, \theta) = k(x, \tilde{x}_i) = e^{-(x-\tilde{x}_i)^2/\rho^2}$.

The observation noise at point x_i is v_i. With a regression based on kernel machine

approximation, static function $f(x)$ at x_i can be represented as

$$y_i = f(x_i) + v_i = \sum_{j=1}^{m_{sv}} a_j \bar{k}(x_i, \tilde{x}_j) + c_0 + \xi_i + v_i \tag{2.6}$$

where a_j, $j = 0, ..., m_{sv}$, is a weight to be determined from the training set, c_0 is the constant part, and ξ_i is the function approximation error at x_i. Note that $\bar{k}(x_i, \tilde{x}_j) = k(x_i, \tilde{x}_j) - \bar{k}$ where $\bar{k} = E(\bar{k}(., \tilde{x}_j))$. Clearly, $E(\bar{k}(x_i, \tilde{x}_j)) = 0$. In this thesis, we use $E(.)$ and $D(.)$ to denote the expectation and variance of a random variable, respectively.

Assumption 2.4. *ξ_i is a random variable with finite variance, i.e., $D(\xi_i) = \sigma_\xi^2$.*

Let $\varepsilon_i = \xi_i + v_i$. To determine the optimal weight vector $\{\gamma_i\}_{i=0}^{m_{sv}}$ which minimizes the least square error σ_ξ^2 in Assumption 2.4, we express (2.6) in the matrix equation form as

$$Y = K\gamma + \xi + v = K\gamma + \varepsilon \tag{2.7}$$

where $Y = [y_1, \ldots, y_N]'$, $\xi = [\xi_1, \ldots, \xi_N]'$, $v = [v_1, \ldots, v_N]'$, $\varepsilon = \xi + v = [\varepsilon_1, \ldots, \varepsilon_N]'$, $\gamma = [\gamma_0, \ldots, \gamma_{m_{sv}}]'$, and

$$\begin{aligned} K &= [e_1' \quad K_{sv}], \quad e_1 = [1 \quad \ldots \quad 1] \\ K_{sv} &= \begin{bmatrix} \bar{k}(x_1, \tilde{x}_1) & \ldots & \bar{k}(x_1, \tilde{x}_{m_{sv}}) \\ \vdots & \ldots & \vdots \\ \bar{k}(x_N, \tilde{x}_1) & \ldots & \bar{k}(x_N, \tilde{x}_{m_{sv}}) \end{bmatrix}. \end{aligned} \tag{2.8}$$

K is constructed to be a full column rank and zero mean matrix based on the input sequence $\{x_i\}_{i=1}^N$ and the set of support vectors $\{\tilde{x}_j\}_{j=1}^{m_{sv}}$ chosen randomly from the input sequence. This is ensured from the following analysis. We first cite Lemma 2.1 from [46]. Then, we analyze that K_{sv} can be a full rank matrix in

Lemma 2.3.

Lemma 2.1. *[46] Let μ be any probability measure on the definition domain of x, and define the norm $\|f\|_\mu = \int_D f^2(x)\mu(x)dx$. For any $f \in \mathfrak{F}$ and $0 < \delta < 1$, with the probability at least $1-\delta$ over $x_1, \ldots, x_{m_{sv}}$ drawn i.i.d from $p_x(x)$, there exist $\gamma_0, \gamma_1, \ldots, \gamma_{m_{sv}}$ such that the estimated function satisfies*

$$\|\hat{f} - f\|_\mu < \frac{\|f\|_p}{\sqrt{m_{sv}}}(1 + \sqrt{2\log\tfrac{1}{\delta}}) \tag{2.9}$$

where $\|.\|_p$ is a function norm defined in [46] and μ is the empirical measure over the data set in a finite size.

Assumption 2.5. *(Parameter Selection of Kernel Machine) Parameters N, m_{sv} and ρ are chosen such that $\rho \to 0$ and $m_{sv} \cdot \rho \to \infty$ as $N \to \infty$ and $m_{sv} \to \infty$.*

Note that the variance of the approximation error ξ_i is a function of m_{sv}, i.e., the following lemma shows $\lim_{m_{sv}\to\infty} \sigma_\xi^2 = \sigma_\xi^2(m_{sv}) = 0$ asymptotically almost surely.

Lemma 2.2. *For any $0 < \delta < 1$, we have $\lim_{m_{sv}\to\infty} \sigma_\xi^2 = 0$ asymptotically almost surely.*

Proof. Let $Y^* = \{f(x_i)\}_{i=1}^N$ and $\hat{Y} = \{\hat{f}(x_i)\}_{i=1}^N$ where $\hat{f}(x_i)$ is the estimate of $f(x_i)$. Note that $\sigma_\xi^2 = E(\|\hat{Y} - Y^*\|_2^2)$. As both $E(\|\hat{Y} - Y^*\|_2^2)$ and $\|\hat{f} - f\|_\mu^2$ are empirical norms of the difference between $\hat{f}$ and f, they are equivalent. Based on Lemma 2.1, for any $0 < \delta < 1$, we have the following satisfied with probability $1 - \delta$:

$$E(\|\hat{Y} - Y^*\|_2^2) = \kappa\|\hat{f} - f\|_\mu^2 < (\frac{\|f\|_p}{\sqrt{m_{sv}}}(1 + \sqrt{2\log\tfrac{1}{\delta}}))^2$$

which gives

$$\sigma_\xi^2 < \kappa(\frac{\|f\|_p}{\sqrt{m_{sv}}}(1+\sqrt{2\log\tfrac{1}{\delta}}))^2$$

where κ is the ratio between the $E(\|\hat{Y}-Y^*\|_2^2)$ and $\|\hat{f}-f\|_\mu^2$. Thus, we have $\lim_{m_{sv}\to\infty}\sigma_\xi^2 = 0$ asymptotically almost surely. □

Lemma 2.3. *Under Assumptions 2.1 and 2.5, the columns in K_{sv} in (2.8) can be constructed as a full column rank matrix.*

Proof. Note that the dimension of K_{sv} is $N\times m_{sv}$. Assume that the input sequence $\{x_i\}_{i=1}^N$ are all different and we have $\{\tilde{x}_j\}_{j=1}^{m_{sv}} \subset \{x_i\}_{i=1}^N$. Each element in K_{sv} is denoted as $k(x_i,\tilde{x}_j) = e^{-(x_i-\tilde{x}_j)^2/\rho^2}$. If $\rho\to 0$, then $k(x_i,\tilde{x}_j)=1$ for $x_i=\tilde{x}_j$ and $k(x_i,\tilde{x}_j)=0$ for $x_i\neq\tilde{x}_j$. So $\forall 1\le j',j\le m_{sv}$, the j-th and j'-th column vectors of K_{sv} are two different unit vectors, respectively. For example, in the j-th column vector, only the element corresponding the case that $x_i=\tilde{x}_j$ is 1 and all others are 0. As $\{x_i\}_{i=1}^N$ are all different, K_{sv} is of full column rank in this case. Based on the continuity of $k(x_i,\tilde{x}_j)$, there exists a sufficiently small $\rho_0>0$ such that K_{sv} is of full rank for all $0\le\rho\le\rho_0$. Thus, K_{sv} as well as K can be constructed to be full column rank matrices based on Assumptions 2.1 and 2.5 for sufficiently small ρ even when m_{sv} is sufficiently large. □

Now we proceed to find the solution of $\hat{\gamma}$ for equation (2.7). Once getting the estimated weights $\{\hat{\gamma}_i\}$ of $\{\gamma_i\}$, $\hat{f}$ is obtained. Let $P_K = KK^+ = K(K'K)^{-1}K'$ denote projection operators onto span$\{K\}$, where span$\{\cdot\}$ is the space spanned by the column vectors of a matrix. To solve γ in (2.7), we project Y to the span$\{K\}$ by operator P_K. Then we have

$$P_K Y = K\hat{\gamma} \tag{2.10}$$

which gives

$$\hat{\gamma} = (K^{'}K)^{-1}K^{'}Y = K^{+}Y \tag{2.11}$$

Note that, in this case, the solution of space projection is actually the same as the least square solution of (2.7).

Lemma 2.4. *Consider equation $Y = K\gamma + \varepsilon$ in (2.7) where γ denotes the m_{sv} dimensional optimal weight vector. For the solution in (2.11), we have $\lim_{N\to\infty}\hat{\gamma} = \gamma$ asymptotically almost surely, provided that K is of full column rank.*

Proof. From (2.11), we can see that $\hat{\gamma} = (K^{'}K)^{-1}K^{'}(K\gamma+\varepsilon) = \gamma+(K^{'}K)^{-1}K^{'}(\xi+v)$.. Note that $\sigma_{\varepsilon}^2 = (\sigma_{\xi}+\sigma_v)^2$. Then, we have $E(\hat{\gamma}) = \gamma + (K^{'}K)^{-1}K^{'}E(\xi)$ and $\sum_{i=0}^{m_{sv}} D(\hat{\gamma}_i) = \sigma_{\varepsilon}^2\mathrm{tr}((K^{'}K)^{-1}) \leq (\sigma_{\xi}+\sigma_v)^2\mathrm{tr}((K^{'}K)^{-1})$ where tr(.) denotes the trace of a matrix.

Now let K_N denote matrix K when its dimension is $N \times (m_{sv}+1)$. We also introduce matrix $K_N = K_N^{'}K_N$ and vector

$$g_{N+1} = [1 \; k(x_{N+1}, \tilde{x}_1) \quad \ldots \quad k(x_{N+1}, \tilde{x}_{m_{sv}})]$$

Then

$$K_{N+1} = [K_N' \quad g_{N+1}'] \begin{bmatrix} K_N \\ g_{N+1} \end{bmatrix} = K_N'K_N + g_{N+1}'g_{N+1} \tag{2.12}$$

Let $\lambda_i(N)$, $i = 0, \ldots, m_{sv}$, be the eigenvalues of K_N and assume that $\lambda_0(N) \geq \ldots \geq \lambda_{m_{sv}}(N) > 0$. There exists a nonsingular matrix P_N such that

$$P_N'K_NP_N = \mathrm{diag}[\lambda_0(N) \; \ldots \; \lambda_{m_{sv}}(N)]$$

Let $\alpha = [1\ 0\ \ldots\ 0]'$, then

$$\alpha' P_N' K_N P_N \alpha' = (P_N\alpha)' K_N (P_N\alpha) = \lambda_0(N)$$

We also have

$$\begin{aligned}\lambda_0(N+1) &= \alpha' P_N' K_{N+1}' P_N \alpha' \\ &= (P_N\alpha)' K_N' (P_N\alpha) + (P_N\alpha)' g_{N+1}' g_{N+1} (P_N\alpha) \\ &= \lambda_0(N) + (P_N\alpha)' g_{N+1}' g_{N+1} (P_N\alpha).\end{aligned}$$

Note that $\lambda_0(N+1) - \lambda_0(N) = (P_N\alpha)' g_{N+1}' g_{N+1}(P_N\alpha) = (g_{N+1}(P_N\alpha))^2$ for all N. So, there exists a constant $\lambda^* > 0$ such that the probability $p(\lambda_0(N+1) - \lambda_0(N) > \lambda^*)$ is nonzero for all N. Thus, we obtain $\lim_{N\to\infty} \lambda_0(N) \to \infty$ almost surely. Similarly, we have $\lim_{N\to\infty} \lambda_i(N) \to \infty$, for $i = 1, \ldots, m_{sv}$, and

$$\lim_{N\to\infty} \mathrm{tr}((K_N' K_N)^{-1}) = \textstyle\sum_{i=0}^{m_{sv}} \frac{1}{\lambda_i(N)} = 0$$

almost surely. This means the trace of $(K'K)^{-1}$ approaches zero almost surely as $N \to \infty$. Since σ_ξ and σ_v are bounded, σ_ε^2 is bounded. Therefore, we have $\lim_{N\to\infty} E(\hat{\gamma}) = \gamma$ and $\lim_{N\to\infty} \sum_{i=0}^{m_{sv}} D(\hat{\gamma}_i) = \sigma_\varepsilon^2 \mathrm{tr}((K'K)^{-1}) = 0$ almost surely and thus the lemma holds. □

Remark 2.3. *Lemma 2.1 basically shows how the error of approximating a nonlinear function reduces uniformly by increasing the number of randomly produced basis functions. This is applicable to our case, because in approximating a function by kernel machines, the basis functions are constructed based on the randomly distributed training set and their number increases with the increasing of the number of data points in the training set. As the number of data points increases, these constructed basis functions become dense in the interval $[-C, C]$ and thus the approximation error will approach zero almost surely. For a set of a finite number of*

basis functions, which corresponds to a finite number of data points in the training set, it is pointed out in [46] that there exist some weights γ satisfying the condition in (2.9). We use S_γ to denote the set containing all these weights. It is also known that, with the least square method, the vector having the observed function values as its components is projected onto the space spanned by the basis functions set and thus the least square estimate $\hat{\gamma}$ minimizes $\|\hat{f} - f\|$ on the space. So a least squares estimate belongs to S_γ and it actually gives the minimum approximation error in the set S_γ. Thus, we can have that $\lim_{m_{sv}\to\infty} \sigma_\xi^2 = 0$ in Lemmas 2.2 and 2.4.

Corollary 2.1. *$\forall x_i \in [-C, C]$ sampled from its probability distribution function $p_x(x)$, under Assumption 2.5, we have $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} |\hat{f}(x_i) - f(x_i)| = 0$ asymptotically almost surely, i.e., $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} \hat{f} \to f$ asymptotically almost surely.*

Proof. Note that ρ in the Gaussian kernel function approaches zero as m_{sv} and N tend to infinity under Assumption 2.5. This ensures the full column rank of K from Lemma 2.3. Thus, by combining Lemma 2.2 and Lemma 2.4. together, $\forall x_i \in [-C,\ C]$, we have $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} |\hat{f}(x_i) - f(x_i)| = 0$ with probability of $1 - \delta$, i.e., $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} \hat{f} \to f$ asymptotically almost surely. □

2.3.2 A Fundamental Model

In model (2.7), there is only one function approximated by $K\gamma$ with the least square estimate $\hat{\gamma}$ of γ given by (2.11). In this subsection, we consider the approximation of two independent functions and propose a fundamental, yet general model. To solve the identification problem formulated in Subsection 2.2.2, we will transform the class of considered systems to such a model and develop a suitable method to

identify it.

For convenience, we denote $\{.\}_{k=1}^{N}$ as a column vector. Suppose that sequence $Y = \{y_i\}_{i=1}^{N}$ is the sum of two sequences $\{h_1(t_i)\}_{i=1}^{N}$ and $\{h_2(s_i)\}_{i=1}^{N}$, namely,

$$\{y_i\}_{i=1}^{N} = \{h_1(t_i)\}_{i=1}^{N} + \{h_2(s_i)\}_{i=1}^{N} + \{v_i\}_{i=1}^{N} \tag{2.13}$$

where $\{h_1(t_i)\}_{i=1}^{N}$ and $\{h_2(s_i)\}_{i=1}^{N}$ are observation vectors of functions $h_1(.)$ and $h_2(.)$, respectively. If both t_i and s_i are i.i.d random variables, then $\{h_1(t_i)\}_{i=1}^{N}$ and $\{h_2(s_i)\}_{i=1}^{N}$ are two independent sequences. From the preceding subsection, we have

$$Y = K\gamma + G\beta + \varepsilon \tag{2.14}$$

where K and G are constructed based on their respective input sequence $\{t_i\}_{i=1}^{N}$, $\{s_i\}_{i=1}^{N}$ and the sequence of the support vector set $\{\tilde{t}_j\}_{j=1}^{m_{sv}}$, $\{\tilde{s}_j\}_{j=1}^{m_{sv}}$, i.e., $K = [e_1' \;\; K_{sv}]$ and $G = [e_1' \;\; G_{sv}]$ with G_{sv} being constructed similarly to K_{sv} given in (2.8). Basically, $K\gamma$ is used to approximate the sequence $\{h_1(t_i)\}$ generated by function $h_1(.)$ and $G\beta$ to approximate $\{h_2(s_i)\}$ generated by function $h_2(.)$. For convenience, the two constant vectors $e_1'\gamma_0$ and $e_1'\beta_0$ are combined together. In this way, we can assume that the constant part of $h_1(.)$ is zero, i.e., $\gamma_0 = 0$ and only need vector e_1' in matrix G. Then, we let

$$K = K_{sv} \quad \text{and} \quad G = [e_1' \;\; G_{sv}] \tag{2.15}$$

Due to the independence of $\{t_i\}_{i=1}^{N}$, $\{s_i\}_{i=1}^{N}$, $\{\tilde{t}_j\}_{j=1}^{m_{sv}}$ and $\{\tilde{s}_j\}_{j=1}^{m_{sv}}$, we have the following remark based on Lemma 2.3.

Remark 2.4. *Under Assumptions 2.1–2.5, matrix* $[K \;\; G]$ *is of full column rank. In addition, all the elements in* K_{sv} *and* G_{sv} *are i.i.d with zero mean and finite variance.*

Equations (2.13) and (2.14) are considered as a fundamental model, which is very important and useful in our subsequent discussions. However, what we know is only the sum of $\{h_1(.)\}$ and $\{h_2(.)\}$. Either $\{h_1(.)\}$ or $\{h_2(.)\}$ is not available. How to separate these two independent sequences $\{h_1(.)\}$ and $\{h_2(.)\}$ from the observation sequence Y is a critical yet challenging problem.

2.3.3 Identification of Fundamental Model Based on Space Projection

We now proceed to solve equation (2.14) based on space projection. The normal equations of (2.13) and (2.14) are given by

$$G\beta = P_G(Y - K\gamma) \tag{2.16}$$

$$K\gamma = P_K(Y - G\beta) \tag{2.17}$$

where $P_K = KK^+$ and $P_G = GG^+$ denote projection operators onto span$\{K\}$ and span$\{G\}$, respectively. From (2.16) and (2.17), it is easy to obtain

$$G\beta = P_G(Y - P_K(Y - G\beta)) \tag{2.18}$$

which gives $(I - P_G P_K)G\beta = P_G(I - P_K)Y$. By using the same space projection as in (2.7), we obtain $\hat{\beta}$ as

$$\hat{\beta} \ = ((I - P_G P_K)G)^+ P_G(I - P_K)Y = H_G^+ P_G(I - P_K)Y \tag{2.19}$$

where $H_G = (I - P_G P_K)G$. Similarly, we have $(I - P_K P_G)K\gamma = P_K(I - P_G)Y$ and

$$\hat{\gamma} \ = ((I - P_K P_G)K)^+ P_K(I - P_G)Y = H_K^+ P_K(I - P_G)Y \tag{2.20}$$

where $H_K = (I - P_K P_G)K$. The estimated functions $h_1(t_i)$ and $h_2(s_i)$ are given by

$$\hat{h}_1(t_i) = \sum_{j=1}^{m_{sv}} \hat{\gamma}_j k(t_i, t_j), \quad \hat{h}_2(s_i) = \sum_{j=1}^{m'_{sv}} \hat{\beta}_j k(s_i, s_j) + \hat{\beta}_0 \tag{2.21}$$

As shown in Remark 2.6 later, the space projection method makes it possible for us to obtain the estimate of β even when there exist unknowing parts in $K\gamma$. Theorem 2.1 to be given later on shows that $\hat{\beta}$ and $\hat{\gamma}$ in (2.19) and (2.20) are the unbiased estimate of β and γ, respectively. To prove Theorem 2.1, we first present the following lemmas.

Lemma 2.5. *[47] If matrix A is of full column rank, then $A^+A = I$.*

Lemma 2.6. *Under Assumptions 2.1–2.5, $(I - P_G P_K)$ is of full rank.*

Proof. Note that it follows from Lemma 2.3 that $[K \;\; G]$ given in (2.15) is of full column rank under Assumptions 2.1–2.5. This means that K and G do not have a joint column space, i.e., $\text{span}\{K\} \cap \text{span}\{G\} = \{0\}$. Assume that $(I - P_G P_K)$ is not of full rank, then $\exists x \neq 0$ such that $(I - P_G P_K)x = 0$, i.e., $P_G P_K x = x$. As P_K and P_G are projection operators onto $\text{span}\{K\}$ and $\text{span}\{G\}$ respectively, we have $x \in \text{span}\{K\} \cap \text{span}\{G\} = \{0\}$, and this contradicts with that $x \neq 0$. So this lemma holds. □

Theorem 2.1. *For the estimates $\hat{\beta}$ and $\hat{\gamma}$ given in (2.19) and* (2.20)*, we have $\lim_{l\to\infty} E(\hat{\beta}) = \beta$, $\sum_i D(\hat{\beta}_i) = \sigma_\varepsilon^2 \text{tr}((G'G)^{-1})$ and $\lim_{N\to\infty} E(\hat{\gamma}) = \gamma$, $\sum_i D(\hat{\gamma}_i) = \sigma_\varepsilon^2 \text{tr}((K'K)^{-1})$, respectively, under Assumptions 2.1–2.5.*

Proof. Note that under Assumptions 2.1–2.3, matrix G is of full column rank.

Then from Lemma 2.5, $G^+G = (G^{'}G)^{-1}G^{'}G = I$. Thus

$$\begin{aligned}
\hat{\beta} &= ((I - P_G P_K)G)^+ P_G(I - P_K)Y \\
&= ((I - P_G P_K)G)^+ P_G(I - P_K)(G\beta + K\gamma + \varepsilon) \\
&= ((I - P_G P_K)G)^+ P_G(I - P_K)G\beta + ((I - P_G P_K)G)^+ P_G(I - P_K)(K\gamma + \varepsilon) \\
&= ((I - P_G P_K)G)^+ P_G(I - P_K)G\beta \\
&\quad + ((I - P_G P_K)G)^+ P_G((K - K(K^{'}K)^{-1}K^{'}K)\gamma + \varepsilon) \\
&= ((I - P_G P_K)G)^+ P_G(I - P_K)(G\beta + \varepsilon) \\
&= AG\beta + A\varepsilon
\end{aligned}$$

where $A = ((I - P_G P_K)G)^+ P_G(I - P_K)$. From Lemma 2.6, $(I - P_G P_K)$ is of full column rank. So

$$\begin{aligned}
AG &= ((I - P_G P_K)G)^+ P_G(I - P_K)G \\
&= ((I - P_G P_K)G)^+ (P_G G - P_G P_K G) \\
&= ((I - P_G P_K)G)^+ ((I - P_G P_K)G) \\
&= I
\end{aligned}$$

Then we have $A = G^+$. Thus, $\hat{\beta} = \beta + (G^{'}G)^{-1}G^{'}\varepsilon$. Similar with Lemma 2.4, $\lim_{l\to\infty} E(\hat{\beta}) = \beta$ and $\sum_i D(\hat{\beta}_i) = \sigma_\varepsilon^2 \mathrm{tr}((G^{'}G)^{-1})$. Similarly, we have $\hat{\gamma} = \gamma + (K^{'}K)^{-1}K^{'}\varepsilon$ and $\lim_{N\to\infty} E(\hat{\gamma}) = \gamma$ and $\sum_i D(\hat{\gamma}_i) = \sigma_\varepsilon^2 \mathrm{tr}((K^{'}K)^{-1})$. □

Remark 2.5. *Theorem 2.1 illustrates how good the fundamental model can be identified. That is, based on the two i.i.d input sequences* $\{t_i\}$, $\{s_i\}$ *and the sum* $\{h_1(t_i)\} + \{h_2(s_i)\}$ *, we can separate the sequences* $\{h_1(t_i)\}$ *and* $\{h_2(s_i)\}$ *satisfying the established performances.*

Corollary 2.2. *For the fundamental model in (2.13), we have the estimated functions in (2.21) satisfying that* $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} |\hat{h}_1(t_i) - h_1(t_i)| \to 0$ *and* $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} |\hat{h}_2(s_i) - h_2(s_i)| \to 0$ *almost surely.*

Proof. Note that we have $\lim_{N\to\infty} E(\hat{\gamma}) = \gamma$ and $\sum_i D(\hat{\gamma}_i) = \sigma_\varepsilon^2 \mathrm{tr}((K^{'}K)^{-1})$ from Theorem 2.1. Similar to Lemma 2.2 and Corollary 2.1, then it can be obtained that $\lim_{m_{sv}\to\infty} \sigma_\xi^2 \to 0$ and $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} |\hat{h}_1(x_i) - h_1(x_i)| \to 0$ which is satisfied with the probability $1-\delta$ for any $0 < \delta < 1$. Similarly, we have the same conclusion for h_2. We denote the above as $\hat{h}_1 \to h_1$ and $\hat{h}_2 \to h_2$. □

2.3.4 Comparison Between Standard Least Square Estimation Method and the Proposed Space Projection Method

It is mentioned that the space projection solution for (2.7) is the same as the standard least square solution. For the fundamental model (2.14), it can also be converted to the form of (2.7) as

$$Y = P\eta + \varepsilon \tag{2.22}$$

where

$$P = [K \quad G], \quad \eta = \begin{bmatrix} \gamma \\ \beta \end{bmatrix} \tag{2.23}$$

Then we obtain its least square solution based on (2.11) as

$$\hat{\eta}_{ls} = [K_l' \quad g_{l+1}'] \begin{bmatrix} \hat{\gamma}_{ls} \\ \hat{\beta}_{ls} \end{bmatrix} = (P'P)^{-1}P'Y = P^{+}Y \tag{2.24}$$

Now we compare the performance of the standard least square method and our proposed space projection method concerning the two solutions of (2.14), i.e.,

(2.24) and (2.19), (2.20). From (2.23),

$$P'P = \begin{bmatrix} K'K & K'G \\ G'K & G'G \end{bmatrix} \tag{2.25}$$

Then

$$(P'P)^{-1} = \begin{bmatrix} P_{11}^{-1} & * \\ * & P_{22}^{-1} \end{bmatrix} \tag{2.26}$$

where $P_{11} = [K'K - K'G(G'G)^{-1}G'K]$, $P_{22} = [G'G - G'K(K'K)^{-1}K'G]$. Obviously, we have $E(\hat{\eta}_{ls}) = \eta,\ D(\hat{\eta}_{ls}) = \sigma_\varepsilon^2(\mathrm{tr}(P_{11}^{-1}) + \mathrm{tr}(P_{22}^{-1}))$.

Thus, we have the following proposition.

Proposition 2.3.1. $\sum_i D(\hat{\gamma}_i) + \sum_i D(\hat{\beta}_i) \leq \sum_i D(\hat{\eta}_{ls_i})$.

Proof. Assume that the eigenvalues of the matrices $K'K$ and $K'G(G'G)^{-1}G'K$ are $\lambda_0, \ldots, \lambda_{m_{sv}}$ and $\tilde{\lambda}_0, \ldots, \tilde{\lambda}_{m_{sv}}$, respectively. Then, the eigenvalues of the matrix P_{11} are $\lambda_0 - \tilde{\lambda}_0, \ldots, \lambda_{m_{sv}} - \tilde{\lambda}_{m_{sv}}$. We have

$$\mathrm{tr}((K'K)^{-1}) = \textstyle\sum_{i=0}^{m_{sv}} \frac{1}{\lambda_i}, \quad \mathrm{tr}((P_{11})^{-1}) = \sum_{i=0}^{m_{sv}} \frac{1}{\lambda_i - \tilde{\lambda}_i}$$

As the matrices $K'K$ and $K'G(G'G)^{-1}G'K$ and $K'K - K'G(G'G)^{-1}G'K$ are all positive definite matrices, there holds $\lambda_i \geq \lambda_i - \tilde{\lambda}_i \geq 0$ and we have

$$\mathrm{tr}((K'K)^{-1}) \leq \mathrm{tr}((P_{11})^{-1}).$$

Then $\sum_i D(\hat{\gamma}_i) = \sigma_\varepsilon^2 \mathrm{tr}((K'K)^{-1}) \leq \sigma_\varepsilon^2 \mathrm{tr}(P_{11}^{-1})$. Similarly, we have $\sum_i D(\hat{\beta}_i) = \sigma_\varepsilon^2 \mathrm{tr}((G'G)^{-1}) \leq \sigma_\varepsilon^2 \mathrm{tr}(P_{22}^{-1})$, and thus the lemma is established. □

Proposition 2.3.1 shows that, for the fundamental model, the space projection method provides better estimates than the standard least square method.

2.4 Identification of the New Model

2.4.1 Model Transformation

In this subsection, we convert the new block-oriented model into the fundamental model in (2.14), that is, a set of linear equations based on kernel machines.

Define a function $\tilde{g}^{-1}(.)$ as $\tilde{g}^{-1}(y_k) = z_k - y_k$. Based on Assumption 2.2, the output function is invertible and we have

$$z_k = g^{-1}(y_k) = y_k + \tilde{g}^{-1}(y_k) \tag{2.27}$$

Also, let $F(u_k, \ldots, u_{k-m}) = b_0 f_0(u_k) + \ldots + b_m f_m(u_{k-m})$. Then (2.3) is expressed as

$$\begin{aligned} y_k = & -\tilde{g}^{-1}(y_k) + a_1[y_{k-1} + \tilde{g}^{-1}(y_{k-1})] + \ldots + a_n[y_{k-n} \\ & +\tilde{g}^{-1}(y_{k-n})] + b_0 f_0(u_k) + \ldots + b_m f_m(u_{k-m}) + v_k \\ = & -\tilde{g}^{-1}(y_k) + a_1\tilde{g}^{-1}(y_{k-1}) + \ldots + a_n\tilde{g}^{-1}(y_{k-n}) \\ & +a_1 y_{k-1} + \ldots + a_n y_{k-n} + F(u_k, \ldots, u_{k-m}) + v_k \end{aligned} \tag{2.28}$$

Based on the discussions in Subsection 2.3.2, the unknown functions will be represented by kernel machines. Let $U_i = [u_{i+r}\ u_{i+r-1}\ \ldots\ u_{i+r-m}]'$, Then

$$\begin{aligned} \{F(u_k, \ldots, u_{k-m})\}_{k=r+1}^{N} &= [F(U_1)\ \ \ldots\ F(U_{N-r})]' \\ &= \breve{K}\breve{\gamma} + \varepsilon_F \end{aligned} \tag{2.29}$$

where $\breve{K} = [k(U_i, U_j)]_{i=1,j=1}^{i=N-r,j=m_{sv}}$ and ε_F is the approximation error vector for

approximating $\{F(u_k, \ldots, u_{k-m})\}$. We also represent

$$\begin{aligned} \{\tilde{g}^{-1}(y_k)\}_{k=r+1}^{N} &= K_0\gamma_0 + \varepsilon_{g_0} \\ a_1\{\tilde{g}^{-1}(y_{k-1})\}_{k=r+1}^{N} &= K_1 a_1\gamma_1 + \varepsilon_{g_1} \\ &\vdots \\ a_n\{\tilde{g}^{-1}(y_{k-n})\}_{k=r+1}^{N} &= K_n a_n\gamma_n + \varepsilon_{g_n} \end{aligned} \tag{2.30}$$

where K_i is constructed based on input sequence $\{y_{k-i}\}_{k=r+1}^{N}$ and its support vector sequence, ε_{g_0}, ..., ε_{g_n} are the respective approximation error vectors for $\{\tilde{g}^{-1}(y_k)\}$, ..., $\{\tilde{g}^{-1}(y_{k-n})\}$. Thus, we can also represent (2.27) as follows

$$\{z_k\}_{k=r+1}^{k=N} = \{y_k\}_{k=r+1}^{k=N} + K_0\gamma_0 + \varepsilon_{g_0} \tag{2.31}$$

We can manipulate (2.28) to obtain the form of the fundamental model given in (2.14) for $k \geq r+1$.

$$\begin{aligned} Y &= -K_0\gamma_0 + K_1 a_1\gamma_1 + \ldots + K_n a_n\gamma_n + W\zeta + \breve{K}\breve{\gamma} + \varepsilon \\ &= \begin{bmatrix} -K_0 & K_1 & \ldots & K_n \end{bmatrix} \begin{bmatrix} \gamma_0 \\ a_1\gamma_1 \\ \vdots \\ a_n\gamma_n \end{bmatrix} + \begin{bmatrix} \breve{K} & W \end{bmatrix} \begin{bmatrix} \breve{\gamma} \\ \zeta \end{bmatrix} + \varepsilon \\ &\triangleq K\gamma + G\beta + \varepsilon \end{aligned} \tag{2.32}$$

where

$$
\begin{aligned}
Y &= [y_{r+1} \;\; y_{r+2} \;\; \dots \;\; y_N]' \\
W &= \begin{bmatrix} y_r & y_{r-1} & \dots & y_{r-n+1} \\ y_{r+1} & y_r & \dots & y_{r-n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{N-1} & y_{N-2} & \dots & y_{N-n+1} \end{bmatrix} \\
\varepsilon &= \varepsilon_F - \varepsilon_{g_0} + \textstyle\sum_{i=1}^{n} a_i \varepsilon_{g_i} \\
\zeta &= [\zeta_1 \; \zeta_2 \; \dots \; \zeta_n]' = [a_1 \; a_2 \; \dots \; a_n]'
\end{aligned}
\tag{2.33}
$$

Comparing (2.32) with (2.14), we have

$$
\begin{aligned}
K &= [-K_0 \; K_1 \; \dots \; K_n] \\
\gamma &= [\gamma_0 \; a_1\gamma_1 \dots \; a_n\gamma_n] \\
G &= \begin{bmatrix} \breve{K} & W \end{bmatrix} \\
\beta &= \begin{bmatrix} \breve{\gamma} \\ \zeta \end{bmatrix}
\end{aligned}
\tag{2.34}
$$

Remark 2.6. *It is important to note that, by exploiting the proposed space projection method in solving the fundamental model in (2.14), we could obtain the estimate $\hat{\beta}$ in (2.32) without knowing $a_1, ..., a_n$ in γ. Also note that K is a column full rank matrix, which implies that $K_0, \dots, K_n$ are linearly independent. It is not possible to find nonzero $\gamma_1, \dots, \gamma_n$ such that all columns in matrix $[K_1\gamma_0 \; \dots \; K_n\gamma_n]$ are constant vectors. For example, if $K_0\gamma_0 = K_n\gamma_n$, then we have $K_0\gamma_0 - K_n\gamma_n = 0$, which contradicts with that K_0, K_n are linear independent. In addition, if $[K_0\gamma_0 \; \dots \; K_n\gamma_n]$ is a constant matrix, parameter a is unidentifiable. This consists with Assumption 2.3 that the functions cannot be constant functions.*

2.4.2 System Identification Based on the Transformed Models Using Space Projection

In this subsection, we use space projection to identify the NARX system based on the converted model (2.32). We first identify the output static nonlinear function $g(.)$, then determine the estimates of the parameters $a_1, \ldots, a_n$, and finally estimate each input static nonlinear function and the parameters $b_0, b_1, \ldots, b_m$.

Estimation of the output static nonlinear function

The output static nonlinear function is $y = g(z)$. In order to obtain its estimate, we need to get the estimate of $\{z_k\}$ first. Based on (2.19) and (2.32), we obtain

$$\hat{\beta} = H_G^+ P_G (I - P_K) Y \tag{2.35}$$

where

$$G = \begin{bmatrix} \breve{K} & W \end{bmatrix}, \qquad \hat{\beta} = \begin{bmatrix} \hat{\breve{\gamma}} & \hat{\zeta} \end{bmatrix}' \tag{2.36}$$

Once obtaining the estimates of $\breve{\gamma}$ and ζ, we rewrite (2.32) in the form of the fundamental model (2.14) as

$$Y - W\hat{\zeta} - \breve{K}\hat{\breve{\gamma}} = -K_0\gamma_0 + K_1 a_1 \gamma_1 + \ldots + K_n a_n \gamma_n + \varepsilon \triangleq K\gamma + G\beta + \varepsilon \tag{2.37}$$

and in this case $K = -K_0$, $\gamma = \gamma_0$, $G = \begin{bmatrix} K_1 & \ldots & K_n \end{bmatrix}$, $\beta = \begin{bmatrix} a_1\gamma_1 & \ldots & a_n\gamma_n \end{bmatrix}'$. Then the estimate $\hat{\gamma}$ is given by

$$\hat{\gamma} = \hat{\gamma}_0 = H_K^+ P_K (I - P_G)(Y - W\hat{\zeta} - \breve{K}\hat{\breve{\gamma}}) \tag{2.38}$$

Remark 2.7. *Note that though we obtain an estimate of γ_0 when estimating γ*

based on the model (2.32), we re-estimate γ_0 based on the model (2.37) using the space projection method. We do this since from Proposition 2.3.1, space projection gives a better estimation than the standard least square method for the fundamental model. For example, consider a model $Y = K_0\nu_0 + K_1\nu_1 + K_2\nu_2$, where Y, K_0, K_1 and K_2 are given and ν_0, ν_1, ν_2 are the weights to be estimated. We first arrange this model in the standard fundamental form $Y = K\gamma + G\beta$ with $K = K_0$, $\gamma = \nu_0$ and $G = [K_1 \quad K_2]$, $\beta = \begin{bmatrix} \nu_1' & \nu_2' \end{bmatrix}'$. Note that the estimate of β is obtained by projecting Y onto $\text{span}\{G\}$*, i.e.,* $\text{span}\{K_1\} \cup \text{span}\{K_2\}$*, by using P_G. This is the same case as shown in (2.22) and thus the corresponding estimates $\hat{\nu}_1$ and $\hat{\nu}_2$ are the same as the least square estimates which are re-denoted as $\hat{\beta}_{ls} = \begin{bmatrix} \nu_{1ls}' & \nu_{2ls}' \end{bmatrix}'$. On the other hand, for the model $G\beta = K_1\nu_1 + K_2\nu_2$ with $G\beta$ replaced by $G\hat{\beta}$, we have the form of the fundamental model (2.14), i.e., $Y \triangleq K\gamma + G\beta$ with $Y = G\hat{\beta}, K = K_1, G = K_2$, $\gamma = \nu_1$ and $\beta = \nu_2$. Then ν_1 and ν_2 can be re-estimated based on space projection and obtain the new estimates $\hat{\nu}_1$, $\hat{\nu}_2$, which could have lower variances than their previous estimates $\hat{\nu}_{1ls}$ and $\hat{\nu}_{2ls}$ based on Proposition 2.3.1.*

Once obtaining the weights, we get function $\tilde{g}^{-1}(.)$ from equation (2.27). Then we can obtain the output static nonlinear function $g^{-1}(.)$. As $g^{-1}(.)$ is invertible, we have the output static nonlinear function $g(.)$ by using suitable fitting methods. From (2.31), we obtain the estimate of $\{z_k\}_{k=r+1}^N$ as follows:

$$\{\hat{z}_k\}_{k=r+1}^N = \{g^{-1}(y_k)\}_{k=r+1}^N = \{y_k\}_{k=r+1}^N + K_0\hat{\gamma}_0 \tag{2.39}$$

Estimation of the parameters $a_1, \ldots, a_n$

We first get the estimate of $\{F(u_k, u_{k-1}, \ldots, u_{k-m})\}_{k=r+1}^{N}$ as follows:

$$\{\hat{F}(u_k, u_{k-1}, \ldots u_{k-m})\}_{k=r+1}^{N} = \{b_0\hat{f}_0(u_k) + \ldots + b_m\hat{f}_m(u_{k-m})\}_{k=r+1}^{N} = \breve{K}\hat{\breve{\gamma}} \tag{2.40}$$

where $\hat{\breve{\gamma}}$ is given in (2.36). For $k \geq N + r + 1$, (2.28) and (2.1) are identical which may be written as the following linear equations

$$\begin{aligned}\{z_k\}_{k=n+r+1}^{N} - \{F(u_k, u_{k-1}, \ldots, u_{k-m})\}_{k=n+r+1}^{N} &= \begin{bmatrix} z_{n+r} & \ldots & z_{r+1} \\ \vdots & \vdots & \vdots \\ z_{N-1} & \ldots & z_{N-n} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \\ &= Xa \end{aligned} \tag{2.41}$$

To estimate a based on (2.41), we need to obtain $\{\hat{z}_k\}$ and $\{\hat{F}(u_k, u_{k-1}, \ldots, u_{k-m})\}$. Then the estimate $\hat{a} = \begin{bmatrix} \hat{a}_1 & \ldots & \hat{a}_n \end{bmatrix}'$ of a can be obtained by solving (2.41):

$$\hat{a} = \hat{X}^{+}\hat{\tilde{Y}} \tag{2.42}$$

where $\hat{\tilde{Y}} = \{\hat{z}_k - \hat{F}(u_k, u_{k-1}, \ldots, u_{k-m})\}_{k=n+r+1}^{N}$.

Estimation of the input static nonlinear functions and the parameters $b_0, \ldots, b_m$.

Two cases will be considered in the identification process.

Case 1: $f_0(.), f_1(.), \ldots, f_m(.)$ are all different. In this case, there is no need to estimate $b_0, b_1, \ldots, b_m$. Instead, we only need to identify $b_i f_i(.)$ $(i = 0, 1, \ldots, m)$. The main idea is to extract $b_i f_i(.)$ when identifying it.

After obtaining $\{F(u_k, u_{k-1}, \ldots, u_{k-m})\}_{k=r+1}^{N} = \{b_0 f_0(u_k) + b_1 f_1(u_{k-1}) + \ldots + b_m f_m(u_{k-m})\}_{k=r+1}^{N}$ in (2.40), we replace F by $\hat{F}$ and extract $b_i f_i(u_{k-i})$ from it one by one starting with $i = 0$. In this way, all the functions can be identified in $m+1$ steps given below:

Step 1: Estimation of $b_0 f_0(u_k)$

Let

$$\begin{aligned} \mathcal{U}_{ij} &= \begin{bmatrix} u_{j+r-1-i} & u_{j+r-i} & \ldots & u_{j+r-m} \end{bmatrix}' \\ \tilde{F}_i(u_{k-i}, \ldots, u_{k-m}) &= b_i f_i(u_{k-i}) + \ldots + b_m f_m(u_{k-m}) \\ i &= 0, 1, \ldots, m,\ 1 \le j \le N - r + 1 \end{aligned} \tag{2.43}$$

Note that the dimension of $\mathcal{U}_{ij}$ is $m + 1 - i$. We have $\tilde{F}_0(u_k, u_{k-1}, \ldots, u_{k-m}) = F(u_k, u_{k-1}, \ldots, u_{k-m})$. Then

$$\begin{aligned} &\{\hat{\tilde{F}}_0(u_k, u_{k-1}, \ldots, u_{k-m})\}_{k=r+1}^{N} \\ &= \{b_0 f_0(u_k) + b_1 f_1(u_{k-1}) + \ldots + b_m f_m(u_{k-m})\}_{k=r+1}^{N} \\ &= \{b_0 f_0(u_k)\}_{k=r+1}^{N} + \{\tilde{F}_1(u_{k-1}, \ldots, u_{k-m}\}_{k=r+1}^{N} \\ &\triangleq \mathcal{K}_0 \tilde{\gamma}_0 + \mathcal{G}_0 \tilde{\beta}_0 + \varepsilon \end{aligned} \tag{2.44}$$

We derive (2.44) with the same reason as stated in Remark 2.7. That is, $\tilde{\gamma}_0$ and $\tilde{\beta}_0$ are the newly defined weights to be estimated. The left-hand side of (2.44) is known since we have obtained $\hat{\tilde{F}}$ in (2.40). For the right-hand side , $\mathcal{K}_0$ and $\mathcal{G}_0$ are constructed based on their respective input sequence and selected support vectors. For the construction of K_0, the input set is $\{u_k\}_{k=r+1}^{N}$, while for the construction of $\mathcal{G}_0$, the input set is $\{\mathcal{U}_{0j}\}_{j=1}^{N-r-1}$. We use $\mathcal{K}_0 \hat{\tilde{\gamma}}_0$ to approximate $\{b_0 f_0(u_k)\}$ and $\mathcal{G}_0 \hat{\tilde{\beta}}_0$ to approximate $\{\tilde{F}_1(u_{k-1}, \ldots, u_{k-m}\}$ by solving the fundamental model.

Obviously (2.44) is in the form of (2.14), so $\tilde{\gamma}_0$ can be estimated by using the space

projection algorithm as

$$\hat{\tilde{\gamma}}_0 = H^+_{\mathcal{K}_0} P_{\mathcal{K}_0}(I - P_{\mathcal{G}_0})\hat{F} \tag{2.45}$$

where $H_{\mathcal{K}_0} = (I - P_{\mathcal{K}_0}P_{\mathcal{G}_0})\mathcal{K}_0$. Thus we have $\widehat{b_0 f_0(u_k)} = \mathcal{K}_0\hat{\tilde{\gamma}}_0$, which is an estimate of $b_0 f_0(u_k)$.

Step $i+1$: Estimation of $b_i f_i(u_{k-i})$, $i = 1, 2, \ldots, m$

Define

$$\begin{aligned} &\{\hat{\tilde{F}}_i(u_{k-i}, \ldots, u_{k-m})\}_{k=r+1}^N = \\ &\{\hat{F}_{i-1}(u_{k-i}, \ldots, u_{k-m})\}_{k=r+1}^N - \{\widehat{b_{i-1} f_{i-1}(u_{k-(i-1)})}\}_{k=r+1}^N \end{aligned} \tag{2.46}$$

where $\{\widehat{b_{i-1} f_{i-1}(u_{k-(i-1)})}\}_{k=r+1}^N = \mathcal{K}_{i-1}\hat{\gamma}_{i-1}$, which is obtained at step i. Then the same procedure as used for estimating $b_0 f_0(u_k)$ can be applied to extract $b_i f_i(u_{k-i})$ $(i = 1, 2, \ldots, m)$ from $\hat{\tilde{F}}$ and to rearrange (2.46) in the form of the fundamental model as

$$\{\hat{\tilde{F}}_i(u_{k-i}, \ldots, u_{k-m})\}_{k=r+1}^N = \mathcal{K}_i\tilde{\gamma}_i + \mathcal{G}_i\tilde{\beta}_i + \varepsilon. \tag{2.47}$$

This gives the estimate of $b_i f_i(u_{k-i})$ as

$$\begin{aligned} \hat{\tilde{\gamma}}_i &= H^+_{\mathcal{K}_i} P_{\mathcal{K}_i}(I - P_{\mathcal{G}_i})\hat{\tilde{F}}_i \\ \{\widehat{b_i f_i(u_{k-i})}\} &= \mathcal{K}_i\hat{\tilde{\gamma}}_i \end{aligned} \tag{2.48}$$

where $H_{\mathcal{K}_i} = (I - P_{\mathcal{K}_i}P_{\mathcal{G}_i})\mathcal{K}_i$, and $\mathcal{K}_i$ and $\mathcal{G}_i$ are constructed based on their respective input set and selected support vectors set. For the construction of $\mathcal{K}_i$, the input set is $\{u_{k-i}\}_{k=r}^N$, while for $\mathcal{G}_i$, the input set is $\{\mathcal{U}_{ij}\}_{j=1}^{N-r-1}$.

Case 2: $f = f_0(.) = f_1(.) = \ldots = f_m(.) = f(.)$. In this case, estimating $b_0, b_1, \ldots, b_m$ becomes meaningful. But before finding their estimates, we can use

the previous approach to get an estimate $\widehat{b_i f(.)}$ of $b_i f(.)$. Define

$$\overline{\widehat{b_i f(u_{k-i})}} = \frac{\sum_{k=r+1}^{N} \widehat{b_i f(u_{k-i})}}{N-r}. \tag{2.49}$$

Then we can obtain the following formula with its asymptotic analysis to be given in Theorem 2.2:

$$\frac{\hat{b}_i}{\hat{b}_0} = \sqrt{\frac{\left(\sum_{k=r+1}^{N}(\widehat{b_i f(u_{k-i})} - \overline{\widehat{b_i f(u_{k-i})}})\right)^2}{\left(\sum_{k=r+1}^{N}(\widehat{b_0 f(u_k)} - \overline{\widehat{b_0 f(u_k)}})\right)^2}} \tag{2.50}$$

where $i = 1, \ldots, m$. Hence, $b_0, \ldots, b_m$ could be identified when the norm of b is fixed.

2.5 Ambiguity Analysis

To the best of our knowledge, in all existing approaches of identifying Hammerstein-Wiener systems such as [24], there exist certain ambiguities in the identification process. Based on some analysis of our scheme, we give certain conditions to avoid such ambiguities in this section.

2.5.1 Two Kinds of Ambiguities

There are two types of ambiguities, constant deflection and scale deflection. In the case of constant deflection, the following two situations cannot be distinguished

through identification for any nonzero constant vector D.

$$\begin{aligned} Y &= K\gamma + G\beta + \varepsilon \\ Y &= (K\gamma + D) + (G\beta - D) + \varepsilon. \end{aligned} \tag{2.51}$$

In the case of scale deflection, the identification of the following two equations will not make any difference for any nonzero constant λ:

$$\begin{aligned} \{z_k\} &= \{a_1 z_{k-1} + \ldots + a_n z_{k-n}\} + F, \quad y_k = g(z_k) \\ \{\lambda z_k\} &= \{\lambda(a_1 z_{k-1} + \ldots + a_n z_{k-n})\} + \lambda F, \quad y_k = g(\lambda z_k) \end{aligned} \tag{2.52}$$

2.5.2 Conditions for Avoiding Ambiguities

Consider the fundamental model $Y = \{h_1(t_i)\}_{i=1}^{l} + \{h_2(s_i)\}_{i=1}^{l}$ in (2.13), where $t_i \in U(-C, C)$, $s_i \in U(-C, C)$ and $h_1(.)$ and $h_2(.)$ are measurable functions. As t_i and s_i are i.i.d, $\{h_1(t_i)\}$ and $\{h_2(s_i)\}$ are independent. Note that if $h_1(.)$ is an odd function, then in approximating $h_1(.)$, the constant part γ_0 is zero. Thus in (2.51), we may assume that e_1' is included in matrix G and the resulting constant vector D is only due to the function $h_2(.)$. Then, there will be no constant deflection for identifying $h_1(.)$ and $h_2(.)$. With this, we give some conditions to possibly avoid ambiguities as follows.

Constant deflection

a) If either $h_1(.)$ or $h_2(.)$ is an odd function, then there can be no constant deflection when estimating the functions. In this case, the constant part D is distributed to the non-odd function. b) If $h_1(.) = h_2(.)$, there is no need for $h_1(.)$ or $h_2(.)$ to be an odd function. In this case, there can be no constant deflection as the constant

is dividedly equally to each function.

Scale deflection

If $F(.) = \lambda F(.)$, then $g(z) = g(\lambda z)$. In order to avoid scale deflection, one can fix the definition domain of $g(.)$, i.e., the interval of z_k.

Remark 2.8.

1) As mentioned above, there exist constant deflection and scale deflection in all other existing nonlinear system identification schemes. In some approaches such as [24] and [48], both the norms of a and b are fixed and the nonlinear function is assumed to be an odd function to avoid ambiguities. Clearly, our conditions are more relaxed.

2) Based on the discussion of constant deflection, we also note that if both $h_1(.)$ and $h_2(.)$ are constant functions, then $h_1(.)$ and $h_2(.)$ can never be separated, since we do not know how to distribute the constant part to each function. This is why we have Assumption 2.2.

2.6 Results on Asymptotic Behavior

The space projection method extracts a function from the sum of functions in the fundamental model (2.13). In the KMSP algorithm, we first separate the input and output functions, then we extract input functions one by one. Now we analyze the asymptotic behavior of the proposed algorithm based on the results obtained in identifying the fundamental model, which shows that the input and output functions can be separated completely. By using Corollary 2.2, we have the multiple input functions $\widehat{b_i f_i} \to b_i f_i$, for $i = 0, 1, \ldots, m$, when $N \to \infty, m_{sv} \to \infty, \rho \to 0$.

Then we have the following theorems on the convergence of parameters estimates $\hat{b}$ and $\hat{a}$, respectively.

Theorem 2.2. *In (2.50), for $i = 1, 2, \ldots, m$, we have*

$$\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} \sqrt{\frac{(\sum_{k=r+1}^{N}(b_i\widehat{f_i(u_{k-i})} - \overline{b_i\widehat{f_i(u_{k-i})}))^2}{(\sum_{k=r+1}^{N}(b_0\widehat{f_0(u_k)} - \overline{b_0\widehat{f_0(u_k)}))^2}} = \frac{b_i}{b_0}$$

asymptotically almost surely.

Proof. Let $s = [s_1\ s_2\ \ldots\ s_{N-r}] = \{b_i\widehat{f(u_{k-i})}\}_{k=r+1}^{N}$ and $v = [v_1\ v_2\ \ldots\ v_{N-r}] = \{\widehat{b_0 f(u_k)}\}_{k=r+1}^{N}$. By using Corollary 2.2, we note that $s \to \{b_i f(u_{k-i})\}_{k=r+1}^{N}$ and $v \to \{b_0 f(u_k)\}_{k=r+1}^{N}$ asymptotically almost surely. By the law of large numbers, we have

$$\begin{aligned} &\lim_{N\to\infty} \frac{\sum_{j=1}^{N-r} s_j}{N-r} \sim N(E(s), D(s)) \\ &\lim_{N\to\infty} \frac{\sum_{j=1}^{N-r} v_j}{N-r} \sim N(E(v), D(v)) \end{aligned} \tag{2.53}$$

where $N(.,.)$ denotes the normal distribution. As we have that $\mathrm{D}(b_i f(u_{k-i})) = b_i^2 \mathrm{D}(f(u_{k-i}))$ and $\mathrm{D}(b_0 f(u_k)) = b_0^2 \mathrm{D}(f(u_k))$, then

$$\lim_{N\to\infty} \frac{\frac{\sum_{j=1}^{N-r} s_j}{N-r} - E(\frac{\sum_{j=1}^{N-r} s_j}{N-r})}{\sqrt{\frac{D(s)}{N-r}}} \sim N(0, b_i) \tag{2.54}$$

$$\lim_{N\to\infty} \frac{\frac{\sum_{j=1}^{N-r} v_j}{N-r} - E(\frac{\sum_{j=1}^{N-r} v_j}{N-r})}{\sqrt{\frac{D(v)}{N-r}}} \sim N(0, b_0) \tag{2.55}$$

Finally, it follows from (2.54) and (2.55) that Theorem 2.2 holds. □

Theorem 2.3. *For the estimate $\hat{a}$ given in (2.42), if $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} \hat{z}_k \to \lambda z_k$ and $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} \hat{F} \to \lambda F$ asymptotically almost surely, then we have $\lim_{N\to\infty, m_{sv}\to\infty, \rho\to 0} \hat{a} \to a$ asymptotically almost surely.*

Proof. Note that if $\lim_{N\to\infty,m_{sv}\to\infty,\rho\to 0}\hat{z}_k \to \lambda z_k$ and $\lim_{N\to\infty,m_{sv}\to\infty,\rho\to 0}\hat{F} \to \lambda F$ asymptotically almost surely, it can be obtained that $\hat{X} \to \lambda X$ and $\hat{\tilde{Y}} \to \lambda\tilde{Y} \to \lambda Xa$ asymptotically almost surely from (2.41). Then, $\lim_{N\to\infty,m_{sv}\to\infty,\rho\to 0}\hat{a} = \lim_{N\to\infty,m_{sv}\to\infty,\rho\to 0}\hat{X}^{+}\hat{\tilde{Y}} = ((\lambda X)'\lambda X)^{-1}(\lambda X)'\lambda Xa = a$ asymptotically almost surely. □

Theorem 2.3 implies that, though there may exist scale deflection for estimating z_k and F, there is no scale deflection for estimating a.

2.7 Simulation Results

In this section, we test and verify the performance of the newly proposed identification approach by using three examples.

Example 2.7.1. *We consider the Hammerstein-Wiener model given below:*

$$z_k = 0.5z_{k-1} + 0.2z_{k-2} + 0.6f(u_k) + 0.4f(u_{k-1})$$

$$y_k = g(z_k)$$

where

$$f(u) = \begin{cases} 1 & \text{if } u \le 0 \\ -1 & \text{if } u > 0 \end{cases}$$

$$g(z) = \begin{cases} \sqrt{z} & \text{if } z > 0 \\ -\sqrt{-z} & \text{if } z < 0 \end{cases}$$

We choose $\{u_k\}_{k=1}^{k=N} \in U(-2,2)$, $N = 400$ *and* $p = 0.5$. *The set of support vectors can be randomly chosen from the input data set. In this experiment, we set the number of support vectors as* $m_{sv} = \frac{N}{2} - r - 1$. *Note that* $[K\ G]$ *should be guaranteed*

to be of full column rank when constructing K and G based on the set of support vectors and the set of input data when using the fundamental model. The steps of identifying the above model are summarized as follows:

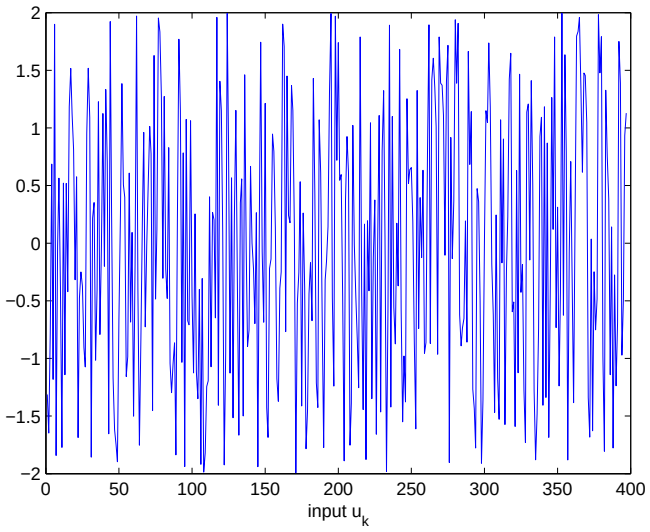

Figure 2.3: Input sequence

1) *Collect the output sequence $\{y_k\}_{k=1}^{N}$ based on the input sequence $\{u_k\}_{k=1}^{N}$.*

2) *Construct the support vector set SV as*

$$SV = \{\tilde{u}_j, \tilde{y}_j\}_{j=1}^{m_{sv}} = \{u_{2j}, y_{2j}\}_{j=r+1}^{N/2} \subset \{u_j, y_j\}_{j=r+1}^{N}$$

Then, construct matrices K and G using SV, $\{u_k\}_{k=1}^{N}$ and $\{y_k\}_{k=1}^{N}$ as well as the parameter ρ based on (2.8).

3) *Estimate the sequence $\{z_k\}_{k=r+1}^{N}$ using (2.39) based on Subsection 2.4.2.*

4) *Estimate a_1 and a_2 using (2.42).*

5) *Estimate input nonlinear functions and b_0 and b_1 based on Subsection 2.3.1.*

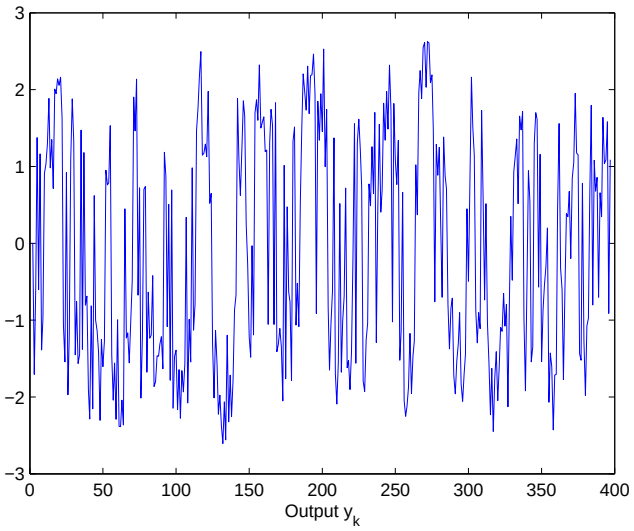

Figure 2.4: Output sequence

The input data, output data and the chosen support vector data are illustrated in Figures 2.3–2.5.

The estimated unknown input nonlinear function together with the true function is shown in Figure 2.6, while the estimated unknown output nonlinear function and its true value are displayed in Figure 2.7. The estimated parameters are $[\hat{a}\ \hat{b}] = [0.4999\ 0.2001\ 0.6001\ 0.3999]$. *Note that in order to get a unique solution, the norm* $||b||_1$ *is fixed as* $||b||_1 = 1$ *according to the condition for avoiding scale deflection.*

Now we investigate the performance of KMSP with respect to N and ρ. To do this, we define $\text{Error}_f = \frac{\|\hat{f}_0 - f\|_1 + \|\hat{g} - g\|_1}{2N}$. Firstly, we set $\rho = 0.5$ and change the number of data points N. Figure 2.8 shows how the error changes with the number of data points. We can see that when N becomes large, the error becomes small. But, the errors cannot approach zero even when N is continuously increasing. This due to the fixed value of parameter ρ. To further reduce the error, we need Assumption 2.5, i.e., $m_{sv}\rho \to \infty$ and $\rho \to 0$ when $N \to \infty$, $m_{sv} \to \infty$. To verify this, we set $\rho = 0.5e^{-m_{sv}/400}$ and do the experiment again with results shown in

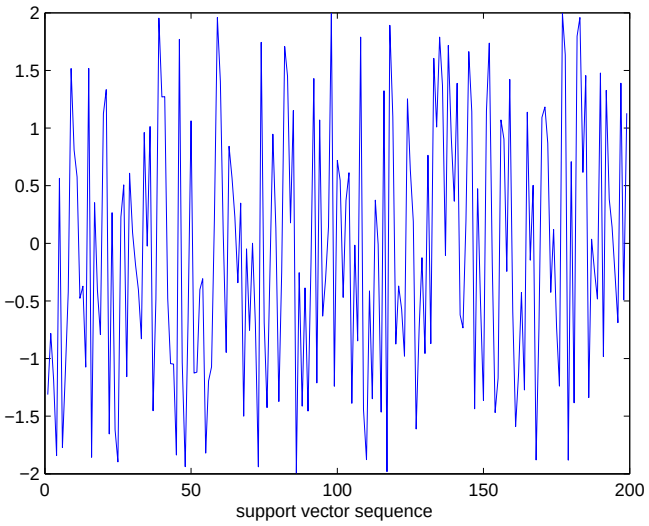

Figure 2.5: Support vector sequence

Figure 2.9. It is seen that the error can approach zero in this case.

Example 2.7.2. *In the case that* $f_0(.) \neq f_1(.)$*, the generalized Hammerstein-Wiener model (i.e., a new NARX model) is:*

$$z_k = 0.6z_{k-1} + 0.1z_{k-2} + 0.8f_0(u_k) + 0.2f_1(u_{k-1})$$

$$y_k = g(z_k)$$

where

$$f_0(u) = \begin{cases} 1 & \text{if } u \leq 0 \\ -1 & \text{if } u > 0 \end{cases}$$

$$f_1(u) = \sin(3u) + \sin(5u)$$

$$g(z) = \begin{cases} \sqrt{z} & \text{if } z > 0 \\ -\sqrt{-z} & \text{if } z < 0 \end{cases}$$

The identification steps are similar to those given in the previous example. We set

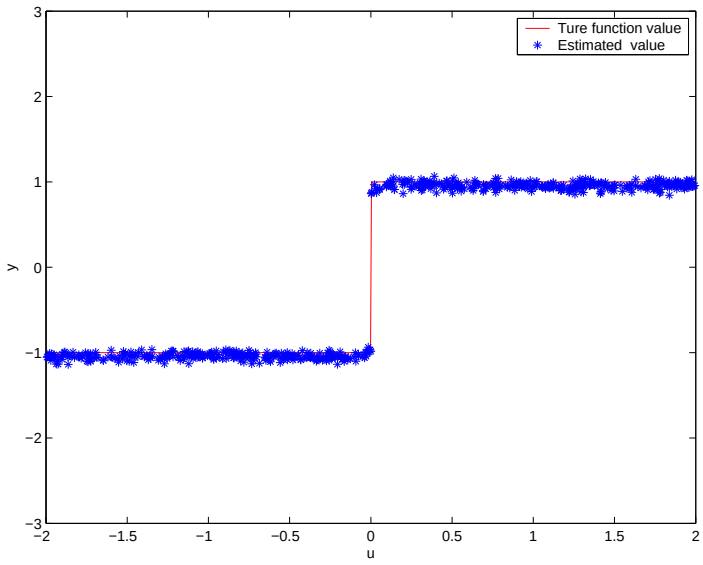

Figure 2.6: True input nonlinear static function and estimated function

$N = 400$, $\rho = 0.2$ *and* $m_{sv} = \frac{N}{2} - r - 1$. *As mentioned, we do not need to estimate parameters* b_0 *and* b_1 *in this case. The algorithm extracts* $b_0 f_0$ *and* $b_1 f_1$ *one by one. By applying the proposed identification algorithm, we obtain* $\hat{a} = [0.5992 \quad 0.1005]$. *The nonlinear functions at the input and output sides are also identified. The estimated* $f_0(u)$ *and* $g(z)$ *in this case are very similar to the estimated functions shown in Figures 2.6 and 2.7 of Example 2.7.1. The estimated* $f_1(u)$ *and its true value are shown in Figure 2.10.*

Now we investigate the performance in the presence of noise:

$$z_k = 0.6z_{k-1} + 0.1z_{k-2} + 0.8f_0(u_k) + 0.2f_1(u_{k-1}) + v_k$$

$$y_k = g(z_k)$$

where v_k *is white Gaussian noise with zero mean and standard deviation* 0.1. *In this case, we conduct the same experiment as in the noise-free case. Application of the KMSP method yields* $\hat{a} = [0.5968 \quad 0.1029]$. *Actually, as long as the noise is independent of system input and output, we can obtain the convergence property*

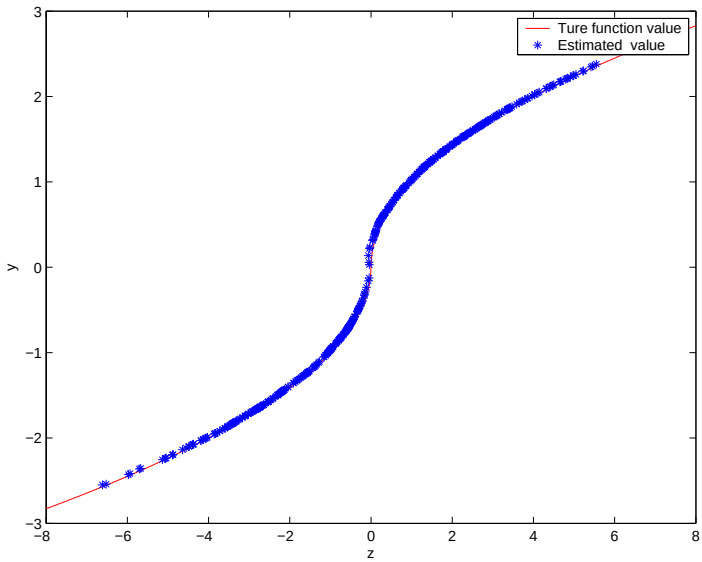

Figure 2.7: True output nonlinear function and estimated function

of our proposed method. When the variance of noise increases, we can still obtain a satisfactory result by increasing N and m_{sv} as well as decreasing ρ.

In this example, we also investigate the estimation error which are defined as $\text{Error}_p = \|\hat{a} - a\|_1 + \|\hat{b} - b\|_1$. Figure 2.11 shows how the estimation error changes with the number of data points. We can see that when N becomes large, the error converges to zero, thus giving a satisfactory result.

Clearly, the results of the above two examples illustrate the effectiveness of our proposed KMSP method.

Remark 2.9.

1) *For the Hammerstein-Wiener system given in Example 2.7.2, the whole dynamic system can be represented as $y_k = \mathcal{F}(z_{k-1}, z_{k-2}, u_k, u_{k-1}) = \mathcal{F}(\mathrm{x}_k)$. In order to use the method in [44] and [45], input x_k must be available. However, x_k is only partially known as signals z_{k-1}, z_{k-2} are unavailable, so that method may face difficulty in implementation. However, if the linear system*

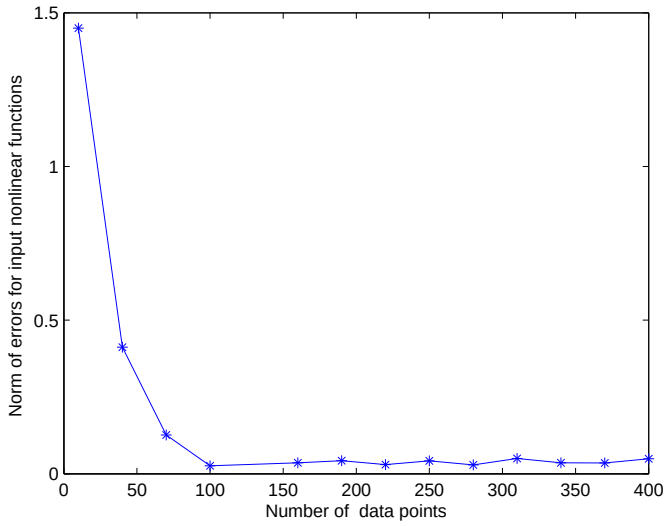

Figure 2.8: The change of error for the estimated function with respect to the data points N

is an FIR system, then the method in [44] and [45] can also obtain a good performance in tracking the output of the Hammerstein-Wiener system.

2*) Note that the other existing schemes, for example, the LS-SVM scheme presented in [24] [39] [40], cannot be employed to identify the models in the above two examples, due to the nonlinear static functions and system model considered.*

Example 2.7.3. *To compare KMSP with the LS-SVM based identification method in [25], we consider the following Hammerstein model*

$$y_k = 0.6y_{k-1} + 0.1y_{k-2} + 0.6f(u_k) + 0.4f(u_{k-1}) + v_k$$

where v_k *is white Gaussian noise with zero mean and standard deviation* 0.1,

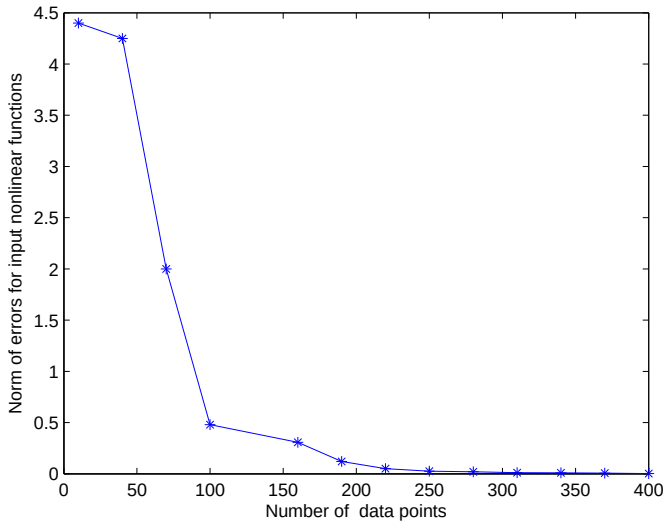

Figure 2.9: The change of error for the estimated function with respect to the data points N

$f(u) = \sin c(u)u^2$. *In LS-SVM,* $f(u) = \sum_{j=1}^{N} \gamma_j k(u, u_j) + \gamma_0$. *Then,*

$$\begin{aligned} y_k &= a_1 y_{k-1} + a_2 y_{k-2} + b_0(\textstyle\sum_{j=1}^{N} \gamma_j k(u_k, u_j)) \\ &\quad + b_1(\textstyle\sum_{j=1}^{N} \gamma_j k(u_{k-1}, u_j)) + (b_0 + b_1)\gamma_0 \\ &= a_1 y_{k-1} + a_2 y_{k-2} + d + \textstyle\sum_{i=0}^{1} \sum_{\gamma=1}^{N} b_i \gamma_j k(u_{k-i}, u_j)) \\ &= a_1 y_{k-1} + a_2 y_{k-2} + d + \textstyle\sum_{i=0}^{1} \sum_{\gamma=1}^{N} \theta_{ij} k(u_{k-i}, u_j)) \end{aligned}$$

where $\theta_{ij} = b_k \gamma_j$, $d = (b_0 + b_1)\gamma_0$ *are solved by using the standard least square algorithm. Estimates of* b_i *and* f *are obtained from singular value decomposition (SVD) of matrix*

$$A = \begin{bmatrix} \hat{\theta}_{01} & \dots & \hat{\theta}_{0N} \\ \hat{\theta}_{11} & \dots & \hat{\theta}_{1N} \end{bmatrix} \begin{bmatrix} k(u_1, u_1)) & \dots & k(u_1, u_N)) \\ \vdots & \vdots & \vdots \\ k(u_N, u_1)) & \dots & k(u_N, u_N)) \end{bmatrix}.$$

By subtracting the mean of every row in A *and making SVD of the resultant matrix, the estimate of* b_i, $i = 0, \dots, m$ *is extracted and then* γ_j, $j = 0, \dots, N$ *can be found.*

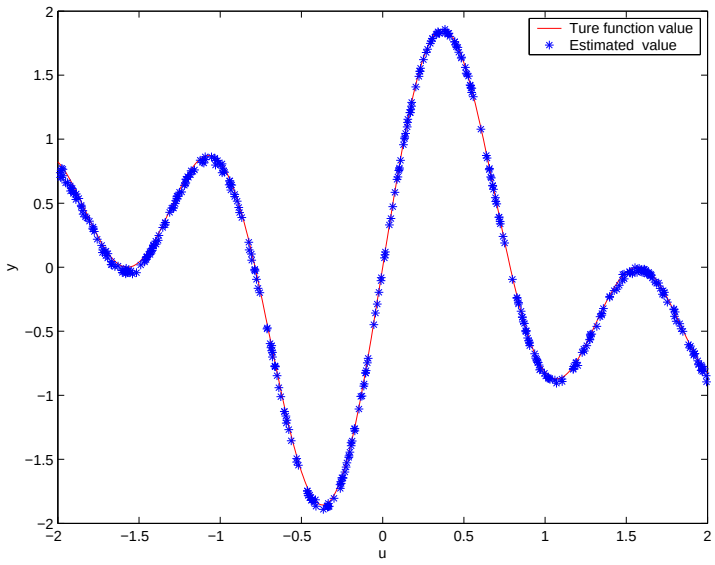

Figure 2.10: True input nonlinear static function f_1 and estimated function value for the generalized Hammerstein-Wiener model

The estimated $[\hat{a} \;\; \hat{b}]$ *using KMSP and LS-SVM are* $[0.5988 \;\; 0.1013 \;\; 0.6015 \;\; 0.3985]$ *and* $[0.5978 \;\; 0.1013 \;\; 0.6020 \;\; 0.3980]$*, respectively. Thus, both KMSP and LS-SVM can obtain satisfactory estimates of* f *and parameters. However, the differences between the KMSP and LS-SVM based identification methods are obvious in the identification process as mentioned in the Introduction section and the above two examples as well as the discussions on the computational cost and numerical problems to be made in the following remark.*

Remark 2.10. *It is also worthwhile to discuss the computational cost and numerical problems in the identification. Theoretically, the estimation error tends to zero when* $N \to \infty, m_{sv} \to \infty$*,* $\rho \to 0$ *and* $m_{sv} \cdot \rho \to \infty$*. However, numerical problems exist and become obvious when the matrix dimension is more than thousands or even larger. Usually, we can obtain a satisfactory estimation result when* N *is less than one thousand. For example, it is observed form Figure 2.11 that when* N *is larger than* 250*, the estimation* Error_p *is smaller than* 0.02*. Thus, numerical problems are negligible with our scheme. Note that the sparse representation of a*

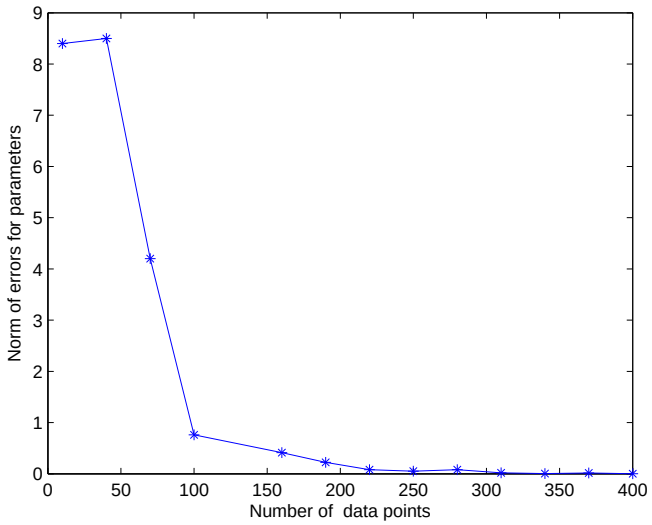

Figure 2.11: The change of error for the estimated parameters with respect to the data points N

nonlinear function f using a subset of input-output data points as a support vector set in this thesis can also deal with numerical problems well. For example, if we try to solve $Y = K\gamma$, we get $\hat{\gamma} = (K'K)^{-1}K'Y$. In this case, the dimension of K is $N \times m_{sv}$. Thus, we only need to determine the inverse of a matrix of dimension $m_{sv} \times m_{sv}$ instead of $N \times N$. This can reduce the computational complexity as well as numerical errors. On the other hand, just as we have mentioned, all data points are support vectors in the LS-SVM based identification method, so the concerned dimension for that method is $N \times N$. Clearly in these aspects our method is advantageous over the LS-SVM based identification method.

Remark 2.11. *Note that for Example 2.7.3, the methods given in [44] and [45] can also be applied to identify the whole dynamic system by representing $y_k = \mathcal{F}(y_{k-1}, y_{k-2}, u_k, u_{k-1}) = F(x_k)$, where $x_k = [y_{k-1}\ y_{k-2}\ u_k\ u_{k-1}]'$ is treated as the input of the nonlinear system $\mathcal{F}(.)$. Then the identification is seen as an implicit nonlinear ARMA model in an RKHS. As a comparison, the method called SVM-ARMA$_{2k}$ in [44] and KMSP proposed herein are applied to track the output of*

the nonlinear system. One can refer to [44] for setting the parameters of SVM-ARMA$_{2k}$. We choose $N = 400$ and $\rho = 0.2$ for KMSP. Let the output of the model be $\hat{y}_k = \hat{\mathcal{F}}(x_k)$ and define a criterion for the tracking error as Error $= \sqrt{\frac{1}{N}\sum_{k=r}^{N}(\hat{y}_k - y_k)^2}$. *From our simulation results,* Error $= 0.1024$ *with SVM-ARMA$_{2k}$ in [44], while* Error $= 0.1025$ *with KMSP. Clearly, both are very close to the level of the standard deviation of the noise $v = 0.1$. So both methods perform well for Example 2.7.3.*

2.8 Conclusion

In this chapter, we have considered the identification of new block-oriented nonlinear systems based on kernel machine and space projection. The major contributions of the chapter are summarized as follows:

1) We have proposed a new class of block-oriented nonlinear system. The proposed model includes the Hammerstein model, the Wiener model, and the Hammerstein-Wiener model as special cases. Input nonlinearities include saturation nonlinearity, deadzone nonlinearity, quantization nonlinearity, signum nonlinearity and so on.

2) We have derived a new method to identify nonlinear systems based on kernel machine and space projection.

3) We have analyzed two ambiguities that may occur in the identification process. Based on our analysis, we have proposed some conditions to avoid such ambiguities.

4) The convergence results of the proposed identification algorithm have also been established in this chapter.

5) The performance of the proposed method has been exemplified by simulation results.

Chapter 3

Iterative Identification of Hammerstein Systems by Normalization

Among the existing schemes in identifying block-oriented systems, an iterative scheme may be the simplest and easiest method to be complemented [70]. In this chapter, we will propose a new iterative algorithm for a general class of Hammerstein systems and prove its convergence. By doing this, we also give a geometrical explanation of why the convergence property can be achieved.

3.1 Introduction

One class of block-oriented nonlinear systems with a static nonlinear function followed by a linear dynamic system is called Hammerstein systems. The identification of Hammerstein systems has been extensively studied in recent two decades. Existing methods mainly include the over-parametrization method [24], the non-

parametric method [56] [57] [58], the stochastic method [53] [54] [50], the kernel machine and space projection method [59] in Chapter 2, and the iterative method ([51] [49] [52] [70] [48]). All the methods have their own advantages and of course weak points. For example, with the non-parametric approach, it is possible to identify more general systems with less assumptions. But usually its convergence speed is relatively slow.

Among the above methods, the iterative method seems to yield the simplest and the most efficient algorithms [70]. Basically, the iterative identification approach divides the unknown parameters into two sets, the linear part and the nonlinear part. At each iteration, one set of estimates is computed while the other set is fixed. Then the two sets alternate and their final parameters estimates are obtained iteratively. Such an iterative approach was first proposed to estimate Hammerstein systems in [51]. However, its convergence needs proper initialization as pointed out by Stoica [49] and Bai and Li [70]. Bai and Li proved the convergence provided that the linear system is FIR. Later, Liu & Bai [48] established the convergence under a given initial condition for the Hammerstein system with an IIR linear system and a static function represented by odd basis functions. It was also pointed out by them that whether the convergence can be extended to a general Hammerstein system was not clear and in fact appeared to be questionable. The convergence for the case of non odd functions or even more general functions under arbitrary nonzero initial conditions is still an open issue.

In this chapter, we propose a normalized iterative algorithm to address this issue. The nonlinear static function is allowed to be any square-integrable functions. It will be shown that the unknown true parameters of the Hammerstein system correspond to the unique partial optimum point of a cost function under the constraint that the norm of the estimates is fixed. With this result, the proposed normalized

algorithm ensures that the estimates converge to the true parameters.

3.2 Normalized Iterative Algorithm of a Hammerstein System

Consider the Hammerstein system consisting of an ARMA linear system and a nonlinear static function as follows:

$$\begin{aligned} x_t &= f(u_t) \\ &= a_0 k_0(u_t) + a_1 k_1(u_t) + ... + a_l k_l(u_t) + e_t \\ y_t &= d_1 y_{t-1} + ... + d_n y_{t-n} + b_0 x_t + ... + b_m x_{t-m} + v_t \end{aligned} \tag{3.1}$$

where u_t is the input signal, $f(.)$ is a nonlinear function represented by the combination of known basis functions and unknown coefficients $a_0, ..., a_l$, x_t and y_t are the input and output of the linear sub-system with known structure but unknown parameters $d_1, ..., d_n$ (AR part of the ARMA system) and $b_0, ..., b_m$ (MA part), ξ_t denotes the approximation error corresponding to u_t, and v_t denotes the noise.

Note that (3.1) can be rewritten as

$$\begin{aligned} y_t = \; & d_0 + d_1 y_{t-1} + ... + d_n y_{t-n} + b_0(a_1 k_1(u_t) + ... + a_l k_l(u_t)) \\ & + ... + b_m(a_1 k_1(u_{t-m}) + ... + a_l k_l(u_{t-m})) + v_t + \xi_t \end{aligned} \tag{3.2}$$

where v_t is the noise term, d_0 is the constant term and ξ_t is the approximation error term:

$$\begin{aligned} d_0 &= a_0 \textstyle\sum_{i=0}^{m} b_i \\ \xi_t &= \textstyle\sum_{i=1}^{m} b_i e_{t-i} \end{aligned} \tag{3.3}$$

The identification objective is to estimate the unknown parameters $d = [d_0 \; ... \; d_n]'$, $b = [b_0 \; ... \; b_m]'$ and $a = [a_0 \; ... \; a_l]'$ in model (3.1) and (3.2) based on the observed

input and output data $\{u_t, y_t\}$, $t = -r, ..., 0, 1, ..., N$ where $r = max(m, n)$ for sufficiently large N.

Denote $Y = [y_1 \ ... \ y_N]'$, $v = [v_1 \ ... \ v_N]'$ and $\xi = [\xi_1 \ ... \ \xi_N]'$. The Hammerstein system in (3.2) can be rewritten as the matrix form:

$$\begin{aligned} Y &= \mathcal{G}d + b_0K_0a + ... + b_mK_ma + v + \xi \\ &= \mathcal{G}d + (\mathcal{K} \otimes a)b + v + \xi \\ &= \mathcal{G}d + (b \cdot \mathcal{K})a + v + \xi \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} &\mathcal{K} = [K_0 \ ... \ K_m] \in R^{N \times (m+1)l} \\ &b \cdot \mathcal{K} \triangleq b_0K_0 + ... + b_mK_m \\ &\mathcal{K} \otimes a \triangleq [K_0a \ \ ... \ \ K_ma] \\ &\mathcal{G} = \begin{bmatrix} 1 & y_0 & ... & y_{1-n} \\ 1 & \vdots & \vdots & \vdots \\ 1 & y_{N-1} & ... & y_{N-n} \end{bmatrix}, \ K_i = \begin{bmatrix} k_1(u_{1-i}) & ... & k_l(u_{1-i}) \\ \vdots & \vdots & \vdots \\ k_1(u_{N-i}) & ... & k_l(u_{N-i}) \end{bmatrix}, i = 0, ..., m \end{aligned} \tag{3.5}$$

and $(b \cdot \mathcal{K})a = (\mathcal{K} \otimes a)b = b_0K_0a + ... + b_mK_ma$. To achieve the identification objective, we have the following assumptions.

Assumption 3.1. *Assume that $f(u)$ is a square-integrable function such that $\int_{-U_0}^{U_0} f(u)^2du < \infty$ and $k_i(u), i = 0, ..., l$ are orthonormal basis functions on a given interval $[-U_0, U_0]$ with $k_0(u)$ being a constant basis function.*

Assumption 3.2. *Input u_t and noise v_t are i.i.d random variables. In addition, $u_t \sim U(-U_0, U_0)$ where $U(-U_0, U_0)$ denotes the uniform distribution on the interval $[-U_0, U_0]$, $E(v_t) = 0$ and $D(v_t) = \sigma_v^2 < \infty$.*

Assumption 3.3. *$[\mathcal{G} \ \ \mathcal{K}]$ is full column rank.*

Assumption 3.4. *Either $\|b\|_2$ or $\|a\|_2$ is known and the first nonzero entry of b or a is positive.*

Remark 3.1. *Assumption 3.3 actually refers to the requirement of the input signals. It is noted that when the input signals are i.i.d, it is not hard to guarantee linear independence of $\mathcal{G}$ and $\mathcal{K}$. Assumption 3.3 also implies that, for any $b \neq 0$, $[\mathcal{G} \quad b \cdot \mathcal{K}]$ is full column rank, and for any $a \neq 0$, $[\mathcal{G} \quad \mathcal{K} \otimes a]$ is full column rank. Assumption 3.4 is to guarantee a unique expression of the Hammerstein system, as any pair λa and b/λ for some non-zero λ provides the same input-output data. Later it will be seen that the cost function in (3.8) is actually a bilinear combination of a and b. So either $\|b\|_2$ or $\|a\|_2$ should be known. In addition, if the true parameter (a, b) globally minimizes cost function (3.8), $(-a, -b)$ also makes the cost function attain its minimum point. To avoid such a case, we need to assume that the first entry of b or a is positive. In addition, considering the case that $\tilde{b}$ is fixed in Figure 3.1, we can always find a unique true parameter a in the upper semisphere in the proof of Lemma 3.3 (The upper semisphere can be defined as a semisphere in which the first entry of a is positive). So both Assumptions 3.3 and 3.4 are related to the identifiability of the nonlinear system.*

Remark 3.2. *Consider the constant term $d_0 = a_0 \sum_{i=0}^{m} b_i$ in (3.3). If $\sum_{i=0}^{m} b_i = 0$, a_0 which is the coefficient of the constant basis function $k_0(.)$ becomes unidentifiable. In this case, we only identify d_0 and $a = [a_1 \ \ldots \ a_l]'$. If $\sum_{i=0}^{m} b_i \neq 0$, all the parameters in (3.2) are identifiable under Assumptions 3.1-3.4. Note that the identifiability of constant term a_0 is related to the constant deflection as discussed in Chapter 2. We also note that most existing schemes including [58] [57] [48] assume that $f(0) = 0$ which implies $a_0 = 0$.*

Lemma 3.1. *In (3.1), the variance of the approximation error $D(e_t) = \sigma_e^2 \to 0$ almost surely as $l \to \infty$.*

Proof. The proof is obvious and omitted. More details of Lemma 3.1 can be seen in Lemma 2.2 in Chapter 2. When $D(e_t) \to 0$, we have $D(\xi_t) \to 0$ in (3.3) and (3.4). □

Remark 3.3. *For the nonlinear static function $f(.)$, if there exists a known upper bound of order l such that $D(e_t) = 0$, i.e, $\xi = 0$, the proposed iterative algorithm belongs to a parametric approach. For more general square-integrable function $f(.)$ like a discontinuous function, we still have $D(e_t) \to 0$ almost surely as $l \to \infty$ in Lemma 3.1. In other words, l is allowed to increase with the increase of learning data. Later it will be seen that our method becomes a semi-parametric approach in this case.*

Lemma 3.2. *Under Assumptions 3.1-3.2, for any $\mathcal{K} \in R^{N\times(m+1)l}, (m+1)l < N$, we have $\lim_{N\to\infty} \frac{\mathcal{K}'\mathcal{K}}{N} = I$ almost surely where I is an identity matrix with dimension $(m+1)l \times (m+1)l$.*

Proof. As the basis functions $k_0(.), k_1(.), ..., k_l(.)$ are orthonormal with $k_0(.)$ being a constant basis function, $\int_{-U_0}^{U_0} k_i(u_t)k_j(u_t)du_t = \delta_{ij}$ where δ_{ij} is 1 if and only if $i = j$, otherwise, $\delta_{ij} = 0$. When $i = 0$ and $j \neq 0$, $\int_{-U_0}^{U_0} k_j(u_t)du_t = 0$, and then $k_i(u_t)$ is a zero mean variable under Assumptions 3.1-3.2. It is known that if u_t and $u_{\tilde{t}}$ $(-r \leq t \neq \tilde{t} \leq N)$ are i.i.d, then $k_j(u_t)$ and $k_j(u_{\tilde{t}})$ for $1 \leq j \leq l$ are all zero mean i.i.d variables. Thus all elements in $\mathcal{K}$ are random variables with zero mean and variance 1. So we have $\lim_{N\to\infty} \frac{\mathcal{K}'\mathcal{K}}{N} = I$ almost surely. □

Remark 3.4. *Note that almost surely means that an event occurs with probability 1. In Lemma 3.2, it is possible that $\mathcal{K}$ is a singular matrix in one realization for a particular sequence $\{u_t\}_{i=-r}^{N} \in [-U_0\ U_0] \otimes [-U_0\ U_0] \otimes ... \otimes [-U_0\ U_0] \subset R^{N+r+1}$ but the measure of such sequences is 0. So such an event occurs with probability 0.*

Remark 3.5. *Legendre polynomials $p_0(u), .., p_j(u), ..., p_l(u)$ are well known orthogonal basis functions in the interval $[-1, 1]$ for $0 \leq j \leq l$ with j denoting the order*

of each basis function. Legendre polynomials can be produced by using Rodrigues' formula: $p_j(u) = \frac{1}{2^k k!}\frac{d^k}{du^k}(u^2-1)^j$. *Note that* $\int_{-1}^{1} p_i(u)p_j(u)du = \frac{2}{2j+1}\delta_{ij}$. *Based on this, it is easy to construct orthonomal basis functions in the interval* $[-C, C]$ *by the substitution* $k_j(u) = \frac{2j+1}{2}p_j(\frac{u}{C})$ *for* $j = 0, ..., l$. *Obviously,* $k_0(u)$ *is a constant function and* $E(k_j(u)) = 0$.

Define a cost function $J_{N,l}(.)$ as

$$\begin{aligned} J_{N,l}(\bar{a}, \bar{b}, \bar{d}) &= \tfrac{1}{N}(\bar{Y} - Y)'(\bar{Y} - Y) \\ &= \tfrac{1}{N}(\bar{Y} - (Y^* + v + \xi))'(\bar{Y} - (Y^* + v + \xi)) \end{aligned} \tag{3.6}$$

where $\bar{Y} = \mathcal{G}\bar{d} + \bar{b}_0 K_0 \bar{a} + ... + \bar{b}_m K_m \bar{a}$, $Y^* = \mathcal{G}d + b_0 K_0 a + ... + b_m K_m a$. Let

$$\begin{aligned} J_N(\bar{a}, \bar{b}, \bar{d}) &= \lim_{l\to\infty} J_{N,l}(\bar{a}, \bar{b}, \bar{d}) \\ J(\bar{a}, \bar{b}, \bar{d}) &= \lim_{N\to\infty, l\to\infty} J_{N,l}(\bar{a}, \bar{b}, \bar{d}) \end{aligned} \tag{3.7}$$

From Lemma 3.1, we have $\lim_{l\to\infty} \|\xi\|_2 = 0$ almost surely. Also, from Assumption 3.2, it can be obtained that

$$\begin{aligned} J(\bar{a}, \bar{b}, \bar{d}) &= \lim_{N\to\infty} \tfrac{1}{N}\|\bar{Y} - Y^*\|_2^2 + \sigma_v^2 \\ &= \lim_{N\to\infty} \tfrac{1}{N}\|\mathcal{G}(\bar{d} - d) + (\bar{b}\cdot\mathcal{K})\bar{a} - (b\cdot\mathcal{K})a\|_2^2 + \sigma_v^2 \end{aligned} \tag{3.8}$$

The estimates $\hat{a}$, $\hat{d}$ and $\hat{b}$ are determined by minimizing $J(\bar{a}, \bar{b}, \bar{d})$, i.e,

$$\{\hat{a}, \hat{b}, \hat{d}\} = \text{argmin } J(\bar{a}, \bar{b}, \bar{d}) \tag{3.9}$$

We first obtain $\hat{d}$ without knowing a and b. We have

$$\begin{aligned} Y &= \mathcal{G}d + b_0K_0a + ... + b_mK_ma + v + \xi \\ &= \mathcal{G}d + [K_0 \ ... \ K_m]\begin{bmatrix} b_0a \\ \vdots \\ b_ma \end{bmatrix} + v + \xi \\ &= \mathcal{G}d + \mathcal{K}\gamma + v + \xi \end{aligned} \tag{3.10}$$

where $\gamma' = \begin{bmatrix} (b_0a)' \ ... \ (b_ma)' \end{bmatrix}$. Let $P_{\mathcal{K}} = \mathcal{K}\mathcal{K}^+$ and $P_{\mathcal{G}} = \mathcal{G}\mathcal{G}^+$ denote projection operators onto span$\{\mathcal{K}\}$ and span$\{\mathcal{G}\}$, respectively, where span$\{.\}$ is the space spanned by the column vectors of a matrix and $\mathcal{K}^+ = (\mathcal{K}'\mathcal{K})^{-1}\mathcal{K}'$ is the generalized matrix inverse. We first ignore the approximation error term in (3.10). Then we obtain $P_{\mathcal{G}}\mathcal{G}d = P_{\mathcal{G}}(Y - \mathcal{K}\gamma) - P_{\mathcal{G}}v$ and $P_{\mathcal{K}}\mathcal{K}\gamma = P_{\mathcal{K}}(Y - \mathcal{G}d) - P_{\mathcal{K}}v$. Note that the noise space is independent of space $span\{\mathcal{K}\} \cup span\{\mathcal{G}\}$ with $E(v) = \{0\}$ based on Assumption 3.2. Thus the noise space is orthogonal to $span\{\mathcal{K}\} \cup span\{\mathcal{G}\}$. And as $P_{\mathcal{G}}v$ and $P_{\mathcal{K}}v$ are operators projecting the noise to the space $span\{\mathcal{G}\}$ and $span\{\mathcal{K}\}$, we have $P_{\mathcal{G}}v = 0$ and $P_{\mathcal{K}}v = 0$. As $P_{\mathcal{G}}\mathcal{G} = \mathcal{G}\mathcal{G}^+\mathcal{G} = \mathcal{G}$ and $P_{\mathcal{K}}\mathcal{K} = \mathcal{K}\mathcal{K}^+\mathcal{K} = \mathcal{K}$, it can be obtained that

$$\mathcal{G}d = P_{\mathcal{G}}(Y - \mathcal{K}\gamma) \tag{3.11}$$

$$\mathcal{K}\gamma = P_{\mathcal{K}}(Y - \mathcal{G}d) \tag{3.12}$$

which is the same as the transformed fundamental model in Chapter 2. Then the KMSP method proposed in Chapter 2 can be used to solve the above model. It is easy to establish that $\mathcal{G}d = P_{\mathcal{G}}(Y - P_{\mathcal{K}}(Y - \mathcal{G}d))$ from (4.11) and (4.12), which gives $(I - P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}d = P_{\mathcal{G}}(I - P_{\mathcal{K}})Y$. Let $\hat{d}$ be the estimate of d. Then

$$\hat{d} = H_{\mathcal{G}}^+P_{\mathcal{G}}(I - P_{\mathcal{K}})Y \tag{3.13}$$

where $H_{\mathcal{G}} = (I - P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}$. After $\hat{d}$ is obtained, the cost function $J_N(\bar{a}, \bar{b}, \bar{d})$ in (6.3) becomes $J_N(\bar{a}, \bar{b}, \hat{d})$.

Definition 3.1. *A point $(a^\star, b^\star)$ is a partial optimum point of the cost function $J_N(\bar{a}, \bar{b}, d)$ if $J_N(a^\star, b^\star, d) \leq J_N(a, b^\star, d)$ when $b^\star$ is fixed and $J_N(a^\star, b^\star, d) \leq J_N(a^\star, b, d)$ when $a^\star$ is fixed.*

Now we are at the position to present our normalized iterative algorithm as follows with k denoting the $k-$th iteration. Based on Assumption 3.4, we fix the norm $\|\hat{b}\|_2 = \|b\|_2$ in the iteration.

Step 1**:** Obtain estimates $\hat{d}$ by using (3.13) and choose an arbitrary nonzero initial value $\hat{b}(0)$.

Step 2**:** Solve the problem $\hat{a}(k) = \text{argmin}_{\bar{a}} J_{N,l}(\bar{a}, \hat{b}(k-1), \hat{d})$.

Step 3**:** Find $\hat{b}_{op}(k) = \text{argmin}_{\bar{b}} J_{N,l}(\hat{a}(k), \bar{b}, \hat{d})$. Let $\hat{b}(k) = sgn(\kappa) \cdot \|b\|_2 \cdot \frac{\hat{b}_{op}(k)}{\|\hat{b}_{op}(k)\|_2}$ where κ is the first nonzero entry of $\hat{b}_{op}(k)$. Clearly, $\|\hat{b}(k)\|_2 = \|b\|_2$ and the first nonzero entry of $\hat{b}(k)$ is positive.

Step 4**:** If a stopping criterion is satisfied, end. Otherwise, replace k by $k+1$ and go back to Step 2.

Remark 3.6. *1). The order of Step 2 and Step 3 can be permuted. With permutation, we start the estimation with nonzero initial value $\hat{a}(0)$. 2). There are several ways to define the stopping criterion in Step 4 of the algorithm. For example, one can consider the difference of cost function values $J_{N,l}(k)$ and $J_{N,l}(k-1)$ since $J_{N,l}(k)$ is a decreasing sequence. Or the absolute value of the difference between $(\hat{a}(k), \hat{b}(k))$ and $(\hat{a}(k-1), \hat{b}(k-1))$. In handling the case that l increases with the increase of learning data, the algorithm can be generalized to a semi-parametric format.*

Remark 3.7. *With the proposed approach in Steps 1-4 in the end of this section, parametric representation is developed for the AR part of the ARMA linear subsystem while its MA part and the nonparametric represented nonlinearity are identified by using the iterative algorithm. The resulted semi-parametric method links the parametric and nonparametric methods.*

Remark 3.8. *The dimensions in Steps 2 and 3 are l and $m+1$, respectively. This avoids the high dimension problem in the well known over parametrization approach in Bai (1998), where the dimension is $l(m+1)$.*

Remark 3.9. *With the non-parametric method proposed by Greblicki and Pawlak [56], the nonlinearity is identified first without knowing the parameters in the linear block, followed by the estimation of the parameters in the linear system. With the proposed method here, the Hammerstein model is identified in a reverse order. Namely the estimate $\hat{d}$ in the AR part of the ARMA linear system is obtained without knowing the nonlinearity, followed by iteratively estimating the nonlinearity and the parameters in the MA part of the linear system.*

3.3 Convergence Analysis of the Iterative Algorithm

We first prove that $\lim_{N\to\infty, l\to\infty} \hat{d} = d$. Then, it is shown that the true parameters correspond to the unique partial optimum point of the cost function and we obtain the convergence of the normalized iterative algorithm.

Lemma 3.3. *Under Assumption 3.3, matrix $(I - P_{\mathcal{G}}P_{\mathcal{K}})$ is full rank.*

Proof. This can be easily proved following the same procedure in Lemma 2.6. □

Theorem 3.1. *Under Assumptions 3.1-3.3, the estimate $\hat{d}$ given in (3.13) satisfies that $\lim_{N\to\infty,l\to\infty}\hat{d}=d$ almost surely.*

Proof. Note that under Assumptions 3.1-3.3, matrix $\mathcal{G}$ is full column rank. As $\mathcal{G}^{+}\mathcal{G}=(\mathcal{G}'\mathcal{G})^{-1}\mathcal{G}'\mathcal{G}=I$, we have

$$\begin{aligned}\hat{d} &= [(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}P_{\mathcal{G}}(I-P_{\mathcal{K}})Y\\ &= [(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}P_{\mathcal{G}}[I-\mathcal{K}(\mathcal{K}'\mathcal{K})^{-1}\mathcal{K}'](\mathcal{G}d+\mathcal{K}\gamma+v+\xi)\\ &= [(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}P_{\mathcal{G}}(I-P_{\mathcal{K}})[\mathcal{G}d+(\mathcal{K}-\mathcal{K}(\mathcal{K}'\mathcal{K})^{-1}\mathcal{K}'\mathcal{K})\gamma+v+\xi]\\ &= [(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}P_{\mathcal{G}}(I-P_{\mathcal{K}})(\mathcal{G}d+v+\xi)\\ &= A\mathcal{G}d+A(v+\xi)\end{aligned} \tag{3.14}$$

where $A=[(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}P_{\mathcal{G}}(I-P_{\mathcal{K}})$. From Lemma 3.3, $(I-P_{\mathcal{G}}P_{\mathcal{K}})$ is full column rank, and then

$$\begin{aligned}A\mathcal{G} &= [(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}[\mathcal{G}(\mathcal{G}'\mathcal{G})^{-1}\mathcal{G}'(I-\mathcal{K}(\mathcal{K}'\mathcal{K})^{-1}\mathcal{K}')\mathcal{G}]\\ &= [(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}[\mathcal{G}-\mathcal{G}(\mathcal{G}'\mathcal{G})^{-1}\mathcal{G}'\mathcal{K}(\mathcal{K}'\mathcal{K})^{-1}\mathcal{K}'\mathcal{G}]\\ &= [(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]^{+}[(I-P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}]=I\end{aligned} \tag{3.15}$$

So we have $A=\mathcal{G}^{+}$. When $l\to\infty$, we have $\lim_{l\to\infty}\|\xi\|_2=0$ almost surely. Based on Assumption 3.2, we have

$$\begin{aligned}&\lim_{l\to\infty}\hat{d}=A\mathcal{G}d+Av=d+\mathcal{G}^{+}(v+\xi)=d+(\mathcal{G}'\mathcal{G})^{-1}\mathcal{G}'v\\ &\lim_{l\to\infty}E((\hat{d}-d)'(\hat{d}-d))=(\mathcal{G}'\mathcal{G})^{-1}\sigma_v^2\end{aligned} \tag{3.16}$$

which gives

$$\lim_{l\to\infty}\textstyle\sum_i E((\hat{d}_i-d_i)'(\hat{d}_i-d_i))=tr((\mathcal{G}'\mathcal{G})^{-1})\sigma_v^2 \tag{3.17}$$

where $tr(.)$ is the trace of a matrix. Now we show that $\lim_{N\to\infty,l\to\infty}\hat{d}=d$ almost

surely. Let $\mathcal{G}_N$ denote a matrix $\mathcal{G}$ with dimension $N \times (n+1)$. Note that $\mathcal{G}_N'\mathcal{G}_N$ is a symmetrical positive matrix as $\mathcal{G}_N$ is full column rank. By following the same procedure in Lemma 2.4 in Chapter 2, where K is the matrix $\mathcal{G}$ here, we have

$$\lim_{N\to\infty} tr((\mathcal{G}_N'\mathcal{G}_N)^{-1}) = \sum_{i=1} \frac{1}{\lambda_i(N)} = 0 \tag{3.18}$$

almost surely. Then, in (3.17), it is easy to obtain that

$$\lim_{N\to\infty,l\to\infty} \sum_i E((\hat{d}_i - d_i)'(\hat{d}_i - d_i)) = 0 \tag{3.19}$$

almost surely. Thus, $\lim_{N\to\infty,l\to\infty} \hat{d} = d$ almost surely and this theorem holds. □

As $Y = \mathcal{G}d + (b\cdot\mathcal{K})a + v + \xi$, we have $lim_{N\to\infty,l\to\infty}\frac{1}{N}a'(b\cdot\mathcal{K})'(b\cdot\mathcal{K})a = \|b\|_2^2\|a\|_2^2 \leq lim_{N\to\infty}\frac{1}{N}\|Y-\mathcal{G}d\|_2^2+\sigma_v^2 = C$ where C is a constant. So if $\|\bar{b}\|_2$ is fixed in $J(\bar{a},\bar{b},\bar{d})$, then $\|\bar{a}\|_2$ is bounded. Define the domain $D = \{(\bar{a},\bar{b})|\ \|\bar{b}\|_2 = \|b\|_2, \|\bar{a}\|_2 \leq M\}$ where M is a constant denoting the bound of $\|\bar{a}\|_2$.

Lemma 3.4. *Under Assumptions 3.1-3.4, the cost function $J(\bar{a},\bar{b},d)$ has a unique partial minimum point (a,b,d) in the domain $D = \{(\bar{a},\bar{b})|\ \|\bar{b}\|_2 = \|b\|_2, \|\bar{a}\|_2 \leq M\}$.*

Proof. We first prove that (a,b,d) is a partial optimum point of $J(\bar{a},\bar{b},d)$. From (3.8), we get

$$J(a+\Delta a, b+\Delta b, d) = \lim_{N\to\infty}(\sigma_v^2 + \tfrac{1}{N}\|\Delta Y\|_2) \geq \lim_{N\to\infty} J_N(a,b,d) = \sigma_v^2 \tag{3.20}$$

for any Δa and Δb. This shows that (a,b,d) is a global minimum point and of course a partial optimum point.

Now we prove the uniqueness of the partial optimum point (a,b,d) by contradiction. Assume that $(\tilde{a},\tilde{b},d)$ is a partial optimum point in D with $\tilde{a} \neq a$ or $\tilde{b} \neq b$.

Let $N_{\tilde{a}}$ and $N_{\tilde{b}}$ denote the neighborhoods of $(\tilde{a}, \tilde{b})$ when either $\tilde{b}$ or $\tilde{a}$ is fixed, respectively, and ρ_1, ρ_2 be the respective radii of $N_{\tilde{a}}$ and $N_{\tilde{b}}$. Figure 3.1 gives a geometrical illustration of the neighborhood $N_{\tilde{a}}$ when $\tilde{b}$ is fixed. Without loss of generality, a locates on the upper semi-sphere based on Assumption 3.4. We have

$$N_{\tilde{a}} = \{(\tilde{a} + \Delta a, \tilde{b}) \mid \|\Delta a\|_2 \leq \rho_1\} \tag{3.21}$$

where $\Delta a = \bar{\tilde{a}} - a = \rho_1 \frac{\bar{\tilde{a}} - \tilde{a}}{\|\bar{\tilde{a}} - \tilde{a}\|_2}$ and $\bar{\tilde{a}}$ locates at the margin of the neighborhood seen in Figure 3.1. Similarly,

$$N_{\tilde{b}} = \{(\tilde{a}, \tilde{b} + \Delta b) \mid \|\Delta b\|_2 \leq \rho_2\} \tag{3.22}$$

where $\Delta b = \bar{\tilde{b}} - b = \rho_2 \frac{\bar{\tilde{b}} - \tilde{b}}{\|\bar{\tilde{b}} - \tilde{b}\|_2}$ when $\tilde{a}$ is fixed. Assume that θ_1, θ_2 are the angles between Δa and $a - \tilde{a}$, Δa and $\tilde{a}$; θ_3, θ_4 are the angles between Δb and $b - \tilde{b}$, Δb and $\tilde{b}$, respectively. We choose the clockwise as the positive direction of the angles. From Figure 3.1, it is easy to establish that

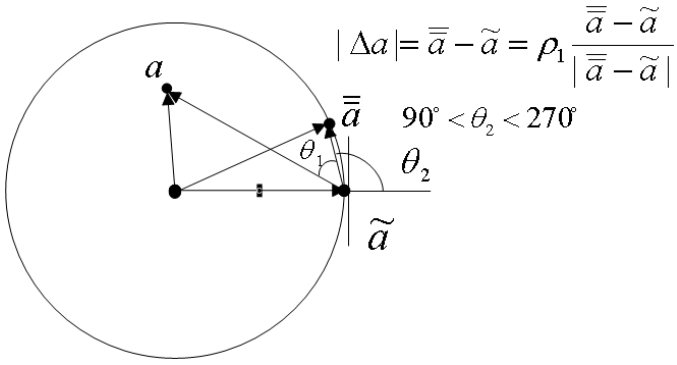

Figure 3.1: The geometrical illustration of the neighborhood of $\tilde{a}$ when $\tilde{b}$ is fixed

$$\begin{aligned} \Delta a \cdot (a - \tilde{a}) &= \rho_1 \|a - \tilde{a}\|_2 \cos\theta_1 \\ \Delta a \cdot \tilde{a} &= \tilde{a}'(\bar{\tilde{a}} - \tilde{a}) = \rho_1 \|\tilde{a}\|_2 \cos\theta_2 \end{aligned} \tag{3.23}$$

Similarly when $\tilde{a}$ is fixed, we have

$$\begin{aligned} \Delta b \cdot (b - \tilde{b}) &= \rho_2 \|b - \tilde{b}\|_2 \cos\theta_3 \\ \Delta b \cdot \tilde{b} &= \tilde{b}'(\bar{\tilde{b}} - \tilde{b}) = \rho_2 \|\tilde{b}\|_2 \cos\theta_4 \end{aligned} \tag{3.24}$$

Note that

$$\begin{aligned} J(\tilde{a} + \Delta a, \tilde{b} + \Delta b, d) &= \lim_{N\to\infty} \tfrac{1}{N} \|\tilde{Y} + \Delta Y - Y^*\|_2^2 + \sigma_v^2 \\ &= \lim_{N\to\infty} (J(\tilde{a}, \tilde{b}, d) + \tfrac{1}{N}\|\Delta Y\|_2^2 \\ &\quad - \tfrac{2}{N}(Y^* - \tilde{Y})'\Delta Y) + \sigma_v^2 \end{aligned} \tag{3.25}$$

where

$$\begin{aligned} \tilde{Y} &= \mathcal{G}d + (\tilde{b} \cdot \mathcal{K})\tilde{a} \\ Y^* &= \mathcal{G}d + (b \cdot \mathcal{K})a \\ \Delta Y &= (\tilde{b} + \Delta b \cdot \mathcal{K})(\tilde{a} + \Delta a) - (\tilde{b} \cdot \mathcal{K})\tilde{a} \\ &= (\tilde{b} \cdot \mathcal{K})\Delta a + (\Delta b \cdot \mathcal{K})\tilde{a} + (\Delta b \cdot \mathcal{K})\Delta a \\ Y^* - \tilde{Y} &= (b \cdot \mathcal{K})a - (\tilde{b} \cdot \mathcal{K})\tilde{a} \\ &= ((b - \tilde{b}) \cdot \mathcal{K})(a - \tilde{a}) + (\tilde{b} \cdot \mathcal{K})(a - \tilde{a}) + ((b - \tilde{b}) \cdot \mathcal{K})\tilde{a} \end{aligned} \tag{3.26}$$

From Lemma 3.2, we have

$$\lim_{N\to\infty} x' \tfrac{(b\cdot\mathcal{K})'(\tilde{b}\cdot\mathcal{K})}{N} y = b'\tilde{b}x'y \tag{3.27}$$

Based on (3.27), we get

$$\begin{aligned} \lim_{N\to\infty} \tfrac{1}{N}\|\Delta Y\|_2^2 &= \lim_{N\to\infty} \tfrac{1}{N}(\Delta Y)'(\Delta Y) \\ &= s_1 + s_2 + s_3 + s_4 \end{aligned} \tag{3.28}$$

where

$$\begin{aligned}
s_1 &= \lim_{N\to\infty} \tfrac{1}{N}((\tilde{b}\cdot\mathcal{K})\Delta a + (\Delta b\cdot\mathcal{K})\tilde{a})'((\tilde{b}\cdot\mathcal{K})\Delta a + (\Delta b\cdot\mathcal{K})\tilde{a}) \\
&= \rho_1^2\tilde{b}'\tilde{b} + \rho_2^2\tilde{a}'\tilde{a} + 2\rho_1\rho_2\|\tilde{a}\|_2\|\tilde{b}\|_2\cos\theta_2\cos\theta_4 \\
&= \rho_1^2\|\tilde{b}\|_2^2 + \rho_2^2\|\tilde{a}\|_2^2 + 2\rho_1\rho_2\|\tilde{a}\|_2\|\tilde{b}\|_2\cos\theta_2\cos\theta_4 \\
s_2 &= \lim_{N\to\infty} \tfrac{1}{N}((\Delta b\cdot\mathcal{K})\Delta a)'((\Delta b\cdot\mathcal{K})\Delta a) = \rho_1^2\rho_2^2 \\
s_3 &= \lim_{N\to\infty} \tfrac{1}{N}((\tilde{b}\cdot\mathcal{K})\Delta a)'((\Delta b\cdot\mathcal{K})\Delta a) = \rho_1^2\rho_2\cos\theta_4\|\tilde{b}\|_2 \\
s_4 &= \lim_{N\to\infty} \tfrac{1}{N}((\Delta b\cdot\mathcal{K})\tilde{a})'((\Delta b\cdot\mathcal{K})\Delta a) = \rho_1\rho_2^2\cos\theta_2\|\tilde{a}\|_2
\end{aligned} \tag{3.29}$$

and

$$\lim_{N\to\infty} \tfrac{2}{N}(Y^* - \tilde{Y})'\Delta Y = s_5 + s_6 + s_7 + s_8 \tag{3.30}$$

where

$$\begin{aligned}
s_5 &= \lim_{N\to\infty} \tfrac{2}{N}(((b-\tilde{b})\cdot\mathcal{K})(a-\tilde{a}))'((\Delta b\cdot\mathcal{K})\Delta a)) \\
&= 2\rho_1\rho_2\cos\theta_1\cos\theta_3\|a-\tilde{a}\|_2\|b-\tilde{b}\|_2 \\
s_6 &= \lim_{N\to\infty} \tfrac{2}{N}((\tilde{b}\cdot\mathcal{K})(a-\tilde{a}) + ((b-\tilde{b})\cdot\mathcal{K})\tilde{a})'((\tilde{b}\cdot\mathcal{K})\Delta a + (\Delta b\cdot\mathcal{K})\tilde{a}) \\
&= 2(\rho_1\cos\theta_1\|a-\tilde{a}\|_2\|\tilde{b}\|_2^2 + \rho_2\cos\theta_3\|\tilde{a}\|_2^2\|b-\tilde{b}\|_2 \\
&\quad + \rho_1\cos\theta_2\tilde{b}'(b-\tilde{b})\|\tilde{a}\|_2 + \rho_2\cos\theta_4\|\tilde{b}\|_2\tilde{a}'(a-\tilde{a})) \\
s_7 &= \lim_{N\to\infty} \tfrac{2}{N}((\tilde{b}\cdot\mathcal{K})(a-\tilde{a}))'((\Delta b\cdot\mathcal{K})\Delta a) \\
&= 2\rho_1\rho_2\cos\theta_1\cos\theta_4\|\tilde{b}\|_2\|a-\tilde{a}\|_2 \\
s_8 &= \lim_{N\to\infty} \tfrac{2}{N}(((b-\tilde{b})\cdot\mathcal{K})\tilde{a})'((\Delta b\cdot\mathcal{K})\Delta a) \\
&= 2\rho_1\rho_2\cos\theta_2\cos\theta_3\|b-\tilde{b}\|_2\|\tilde{a}\|_2
\end{aligned} \tag{3.31}$$

Since only s_6 includes the first order terms of ρ_1 and ρ_2, when $\rho_1 \to 0, \rho_2 \to 0$, we have $[(s_1+s_2+s_3+s_4)-(s_5+s_6+s_7+s_8)] \to -s_6$ and thus (3.25) becomes

$$\begin{aligned}
J(\tilde{a}+\Delta a, \tilde{b}+\Delta b, d) &= J(\tilde{a},\tilde{b},d) + (s_1+s_2+s_3+s_4) - (s_5+s_6+s_7+s_8) \\
&\to J(\tilde{a},\tilde{b},d) - s_6
\end{aligned} \tag{3.32}$$

When $\tilde{b}$ is fixed, $\rho_2 = 0$ and

$$s_6 = 2\rho_1(\cos\theta_1\|a-\tilde{a}\|_2\|\tilde{b}\|_2^2 + \cos\theta_2\tilde{b}'(b-\tilde{b})\|\tilde{a}\|_2) \tag{3.33}$$

Note that

$$\tilde{b}'(b-\tilde{b}) = \tilde{b}'b - \tilde{b}'\tilde{b} \leq 0 \tag{3.34}$$

under $\|b\|_2 = \|\tilde{b}\|_2$. Seen from Figure 3.1, there always exists an $\bar{a}$ such that $-90^\circ < \theta_1 < 90^\circ$ ($\cos\theta_1 > 0$) and $90^\circ < \theta_2 < 270^\circ$ ($\cos\theta_2 < 0$) in $N_{\tilde{a}}$. Thus we have $s_6 > 0$ and $J(\tilde{a}+\Delta a, \tilde{b}, d) < J(\tilde{a}, \tilde{b}, d)$ which means $(\tilde{a}, \tilde{b}, d)$ cannot be a partial optimum point in this case. Similar conclusion can be obtained when $\tilde{a}$ is fixed. So as long as the point $(\tilde{a}, \tilde{b}, d)$ is different from (a, b, d), we can always find certain points in its corresponding neighborhood $N_{\tilde{a}}$ or $N_{\tilde{b}}$ with smaller cost function values. This contradicts with the assumption that $(\tilde{a}, \tilde{b}, d)$ is a partial optimum point. Therefore the conclusion of this lemma holds. □

Remark 3.10. *As mentioned in Remark 3.6, estimating a and b can be permuted if there exists a finite l such that $\xi = 0$. Define $\tilde{D} = \{(\bar{a}, \bar{b})|\ \|\bar{a}\|_2 = \|a\|_2, \|\bar{b}\|_2 \leq \tilde{M}\}$ where $\tilde{M}$ is a constant denoting the bound of $\|\bar{b}\|_2$. Then under Assumptions 3.1-3.4, the cost function $J(\bar{a}, \bar{b}, d)$ has a unique partial minimum point (a, b, d) in the domain $\tilde{D}$.*

Theorem 3.2. *Under Assumptions 3.1-3.4,* $\lim_{N\to\infty,l\to\infty,k\to\infty}\hat{a}(k) = a$ *and* $\lim_{N\to\infty,l\to\infty,k\to\infty}\hat{b}(k) = b$ *almost surely.*

Proof. Let $\hat{z}(k) = (\hat{a}(k),\ \hat{b}(k),\ \hat{d}(k))$. Employing the proposed normalized iterative algorithm, we obtain $\lim_{N\to\infty,l\to\infty}\hat{d} = d$ almost surely in Theorem 3.1 in one iteration step. Since sequence $\{J(k)\}$ is positive and decreasing, it is convergent. Then the generated sequence $\{\hat{z}(k)\}$ has at least one accumulation point denoted

as $z^* = (a^*,\ b^*, d)$. By Lemma 3.4, z^* will be the unique partial optimum point if either $\|\hat{b}(k)\|_2$ or $\|\hat{a}(k)\|_2$ is fixed. Thus, $\lim_{N\to\infty,l\to\infty,k\to\infty} \hat{a}(k) = a^* = a$ and $\lim_{N\to\infty,l\to\infty,k\to\infty} \hat{b}(k) = b^* = b$ almost surely for arbitrary nonzero initial conditions. □

Remark 3.11. *In employing the iterative algorithm, if the norm $\|\hat{a}\|_2$ or $\|\hat{b}\|_2$ is not fixed to be $\|a\|_2$ or $\|b\|_2$, the iteration sequence may diverge as explained below. Let $\tilde{a}$ and $\tilde{b}$ denote the current estimates of a and b at the k-th iteration. Assume that $\tilde{b}'(b-\tilde{b}) > 0$, which is possible when $\|\tilde{b}\|_2 < \|b\|_2$. Also seen in Figure 3.1, we could choose an $\bar{\tilde{a}}$ locating outside the sphere such that $90° < \theta_1 < 270°$ $(\cos\theta_1 < 0)$ and $-90° < \theta_2 < 90°$ $(\cos\theta_2 > 0)$. Then $\cos\theta_2\tilde{b}'(b-\tilde{b})\|\tilde{a}\|_2 > -\cos\theta_1\|a-\tilde{a}\|_2^2\|\tilde{b}\|_2^2$, i.e, $s_6 > 0$. The case that $\cos\theta_1 > 0$ means the iteration point moves toward a, while $\cos\theta_1 < 0$ corresponds to that the iteration point moves away from a. This implies that there also exist certain directions along which the iteration point moves away from true parameter a while the cost function decreases. Thus, if the norm of $\hat{a}$ is not fixed, the sequence $\{(\hat{a}, \hat{b}, d)\}$ will move to $(\infty, 0, d)$. Similarly, $(0, \infty, d)$ could also be an accumulation point. This explains why a counterexample could be provided by Stoica (1981) and also explains why we need to fix the norm of $\hat{a}$ or $\hat{b}$.*

3.4 Illustrative Examples

Example 3.4.1. *Consider a Hammerstein system with an FIR linear system given by Stocia (1981): $y_t = b_0(a_1u_t + a_2u_t^2) + b_1(a_1u_{t-1} + a_2u_{t-1}^2) = 1*(0*u_t + 2*u_t^2) - 2*(0*u_{t-1} + 2*u_{t-1}^2)$ where b_0 is fixed to be 1.*

In Stocia (1981), the following situation is considered when the iterative method is used to do estimation. If $\hat{b}_1(0) > -1/3$ and $\hat{b}_1(k-1) < -\frac{3+\hat{b}_1(0)}{1+3\hat{b}_1(0)}$, then $\hat{b}_1(k) < \hat{b}_1(k-1) < \hat{b}_1(0)$. If $\hat{b}_1(0) < -1/3$ and $\hat{b}_1(k-1) > -\frac{3+\hat{b}_1(0)}{1+3\hat{b}_1(0)}$, then $\hat{b}_1(k) >$

$\hat{b}_1(i-1) > \hat{b}_1(0)$. Thus, as long as $\hat{b}_1(0) \neq -1/3$ one can always initialize $\hat{b}_1(0)$ such that the estimates diverge. This implies that it is impossible to ensure the convergence of the method under a general initial condition. Recently in [48], for the case that the nonlinear static function is odd and initial estimate is chosen as $\hat{b}(0) = [1\ 0\ ...\ 0]'$, the convergence of iterative algorithm is ensured. While in our proposed algorithm, by orthonormalizing the function basis u and u^2 to orthonormal polynomials, the convergence of the iterative identification for any nonzero initial conditions can be guaranteed. This is achieved by excluding the divergence cases with normalization. As $\|b\|_2 = \sqrt{5}$, we fix the norm $\|\hat{b}\|_2 = \sqrt{5}$ in the iterative estimation. The above divergence conditions in Stocia (1981) will not be satisfied after the normalization, since $\|\hat{b}(k)\|_2$ cannot become large.

Example 3.4.2. *In this example, we consider the following Hammerstein system given in [48] for the purpose of comparison with existing techniques in the area.*

$$\begin{aligned} y_t &= 0.3y_{t-1} + 0.2y_{t-2} + 0.1y_{t-3} + 3x_t - 2x_{t-1} + v_t \\ x_t &= 0.9454u_t + 0.2983u_t^3 \end{aligned}$$

We first orthogonalize the odd function basis u and u^3 with coefficients 0.9545 and 0.2983 to orthogonal polynomials (Legendre polynomials) given as 1 , u, $\frac{1}{2}(3u^2-1))$ and $\frac{1}{2}(5u^3-3u))$. Then the orthonomal basis is constructed as $k_0(u) = \frac{1}{2}, k_1(u) = \frac{3}{2}u, k_2(u) = \frac{5}{2}\cdot\frac{1}{2}(3u^2-1), k_3(u) = \frac{7}{2}\cdot\frac{1}{2}(5u^3-3u)$ with coefficients 0, 0.4534, 0 and 0.0341 as discussed in Remark 3.5. Compared with (3.1), when $l = 3$, we have the approximation error can be zero, i.e, we have $e_t = 0$ in (3.1). The true parameters of the Hammerstein system in the form of (4.2) are $a = [a_0\ a_1\ a_2\ a_3] = [0\ 0.4534\ 0\ 0.0341]'$, $b = [b_0\ b_1] = [\ 3\ -2]'$ and $d = [d_0\ d_1\ d_2\ d_3] = [0\ 0.3\ 0.2\ 0.1]'$. The input u_t is i.i.d in the interval $[-1, 1]$ and the noise v_t is zero mean with standard derivation 0.1. Let $\|b\|_2 = \sqrt{13}$ and consider the l_2 norm of the estimation error

$\|e\|_2 = \sqrt{e'e}$ where $e = (\hat{a}, \hat{b}, \hat{d}) - (a, b, d)$. The comparison results are shown in

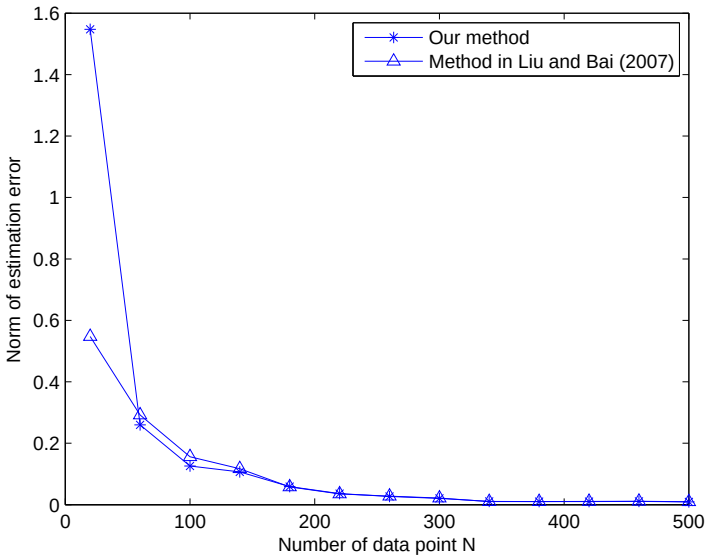

Figure 3.2: Estimation error with respect to the number of data points N

Figure 3.2 in which $\|e\|_2$ with respect to the number of data points N is plotted. It can be observed that both methods perform very well. Note that the initial value $\hat{b}(0)$ is randomly given in our method and it is $\hat{b}(0) = [1\ 0\ 0\ ...]'$ in [48].

Example 3.4.3. *In our proposed algorithm, the nonlinear function is not necessary to be odd even when the linear system is IIR . To illustrate this, we consider the Hammerstein system:*

$$\begin{aligned} y_t &= 0.4y_{t-1} + 0.1y_{t-2} + 0.5x_t + 0.3x_{t-1} + 0.2x_{t-2} + v_t \\ x_t &= 0.1 + 0.6u_t + 0.3u_t^2 \end{aligned}$$

where $u_t \sim U[-1, 1]$ *and* $v_t \sim U[-0.1, 0.1]$.

By orthonormalizing the non-odd function basis 1, u and u^2 to orthonomal polynomials given as $k_0(u) = \frac{1}{2}$, $k_1(u) = \frac{3}{2}u$ and $k_2(u) = \frac{5}{2} \cdot \frac{1}{2}(3u^2 - 1)$ on the interval $[-1, 1]$, we get $x_t = 0.4k_0(u_t) + 0.4k_1(u_t) + 0.08k_2(u_t)$. Then this Hammerstein

system can be rewritten as

$$\begin{aligned} y_t &= 0.4y_{t-1} + 0.1y_{t-2} + 0.5x_t + 0.3x_{t-1} + 0.2x_{t-2} + v_t \\ x_t &= 0.4k_0(u_t) + 0.4k_1(u_t) + 0.08k_2(u_t) \end{aligned}$$

and true parameters to be estimated are $a = [0.4\ 0.4\ 0.08]$, $d = [0.4\ 0.4\ 0.1]$ and $b = [0.5\ 0.3\ 0.2]$. We choose $N = 500$ and fix $\|b\|_1 = 1$ with $b_1 > 0$. Note that

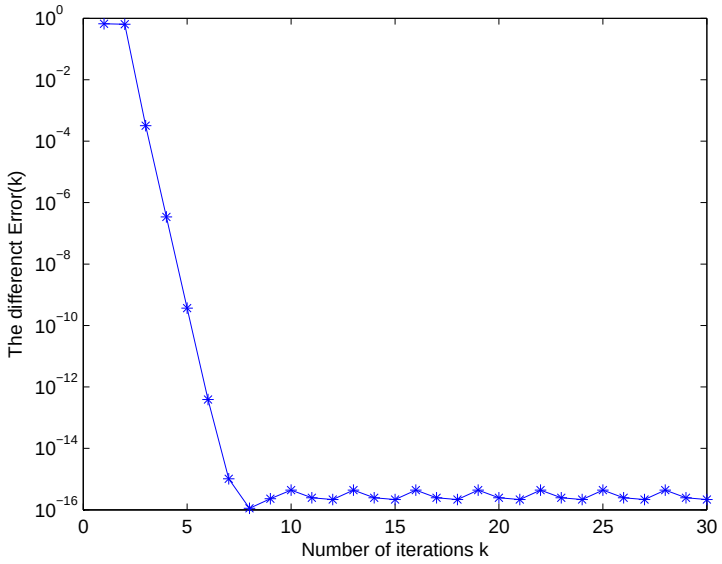

Figure 3.3: The illustration that the iteration algorithm converges in a few iterations

both $\|.\|_1$ and $\|.\|_2$ can be used to fix the norm of the parameters vector. By implementing the proposed iterative algorithm, we obtain $\hat{a} = [0.3990\ 0.3980\ 0.0813]$, $d = [0.4037\ 0.4110\ 0.0922]$ and $\hat{b} = [0.5015\ 0.2988\ 0.1997]$. Obviously, the estimates are very close to the true values. To show how the estimates converges to the fixed point with respect to the number of iterations, we calculate the difference $Error(k) = \sum_{i=0}^{m} |\hat{b}_i(k+1) - \hat{b}_i(k)|$ at each iteration and use it as a stop criterion. Figure 3.4 shows that the algorithm converges in only a few iterations. To show how the estimates behave with the number of data points N, we plot the square

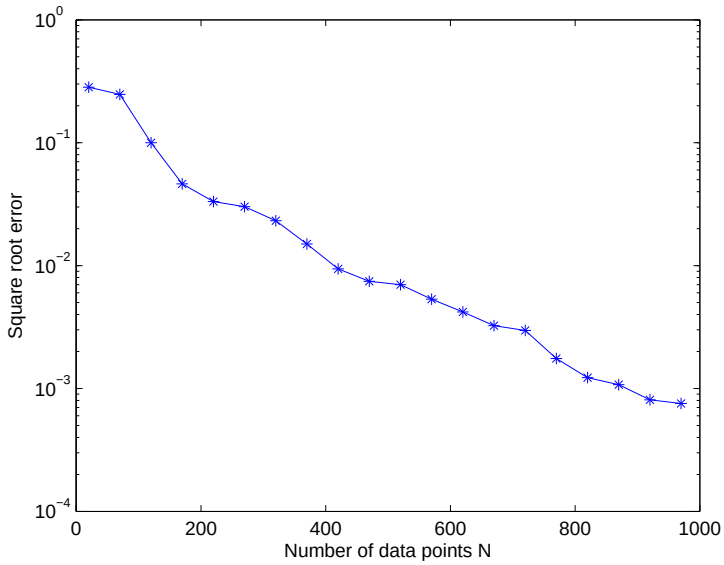

Figure 3.4: Estimation error with respect to number of data points

root error $\|\hat{b} - b\|_2$ with respect to N in Figure 3.4. This figure shows that the square error can be made small enough by choosing N appropriately.

Remark 3.12. *The main advantages of the iterative method are its simplicity, computational efficiency with fast convergence speed and so on. It is also noticed that usually the iterative algorithm can be easily understood and implemented. To illustrate these, we make a comparison based on the observations from the above simulation results and those in the example given by [57] (Hasiewicz and Mzyk, 2009, section 5.1, pp. 449).*

1) *In Hasiewicz and Mzyk (2009), the nonparametric instrumental variables method is illustrated for the identification of a Hammerstein system similar to Example 3. After the nonparametric regression method is employed to estimate the static nonlinear function, the instrumental variables method together with Levenberg-Marquardt method is used to identify the parameters in the nonlinear function and the linear system. The estimation errors* $\Delta_b(N) = \|\hat{b}-b\|_2/\|b\|_2 \cdot 100\%$ *is de-*

fined to show its performance. It is reported that when $N > 800$, $\Delta_b(N) < 1\%$.

2*) As seen in Figure 3.4 and Figure 3.4, estimates are ensured to converge to the true parameters rapidly, usually in few iteration steps even for a relatively small number of data points* N. *Also observed in Figure 3.4,* $\Delta_b(N)$ *can be much smaller than* 1% *when* $N > 800$ *under the noise level* $v_t \sim U(-0.1, 0.1)$. *However, compared with the methods in Hasiewicz and Mzyk (2009), coloured noise cannot be considered for the iterative method in its present form. How to employ the instrumental variables approach using iterative algorithm in the identification of Hammerstein systems may be an interesting research direction in the future.*

In summary, nonparametric and parametric approaches have their own advantages and weak points. These two approaches complement each other in the identification of both linear and nonlinear systems.

3.5 Conclusion

We revisit the iterative identification method and propose a normalized algorithm for Hammerstein systems. Convergence property is established under arbitrary nonzero initial conditions. As pointed out by [70], if the iterative algorithm converges, it converges fast and is very efficient. The static function is extended to square-integrable functions. Examples are also used to illustrate the effectiveness of the proposed scheme. Note that we obtain the convergence property under the condition that the noise is white. It is also pointed out that how to employ the instrumental variables approach using iterative algorithm in the identification of Hammerstein systems may be an interesting research direction.

Chapter 4

Convergence of Fixed Point Iteration for the Identification of Hammerstein and Wiener Systems

In this chapter, fixed point iteration is introduced to identifying both Hammerstein and Wiener systems with a unified algorithm. It is shown that the iteration is a contraction mapping on a metric space when the number of input-output data points approaches infinity. This implies the existence and uniqueness of a fixed point of the iterated function sequence and thus ensures the convergence of the iteration.

4.1 Introduction

A Hammerstein or Wiener system is a cascade system with a static nonlinear function followed or preceded by a linear dynamic system. The identification of such systems has been extensively studied, see for examples, [70] [50] [48] [66] [52] [64] [65] [56] [54] [59].

One kind of identification method is to use iterative methods [70]. There are two types of iterative identification algorithms. One approach uses the dynamic inverse of the linear system to iteratively estimate an intermediate signal. The other approach divides the unknown parameters into two sets, the linear part and the nonlinear part. At each iteration, estimates of parameters for one set is computed while the other is fixed. Then the two sets alternate and their final estimates are obtained iteratively. The first type initially appeared in [63]. As pointed out by [52] [64], the convergence is still unknown for both Hammerstein and Wiener systems. The second type was first presented for Hammerstein systems in [51]. However, proper initialization was required for its convergence as pointed out in [48], [70] and [49].

Only until recently convergence under arbitrary nonzero initial conditions was established for Hammerstein systems with a nonlinear function represented by general basis functions in [67] (Chapter 3), which was achieved by showing that true parameters correspond to the unique partial optimum point of a proposed cost function.

In fact, convergence property of the iterative algorithm is generally difficult to establish as a Hammerstein or Wiener system contains certain unmeasurable internal variables. In this chapter, we propose a fixed point iteration for identifying both Hammerstein and Wiener systems with a unified algorithm, which belongs

to the second type of iterative algorithms. We show that parameter estimation can be reduced to a fixed point iteration of a nonlinear equation and the proposed algorithm ensures the estimates converging to the true parameters under arbitrary nonzero initial conditions.

The remaining part of this chapter is organized as follows: In Section 4.2 the iterative algorithm of Hammerstein and Wiener systems is formulated. The convergence property for arbitrary nonzero initializations is established in Section 4.3. Examples are studied in Section 4.4 and this chapter is concluded in Section 4.5.

4.2 Fixed Point Iteration Algorithm for Hammerstein and Wiener Systems

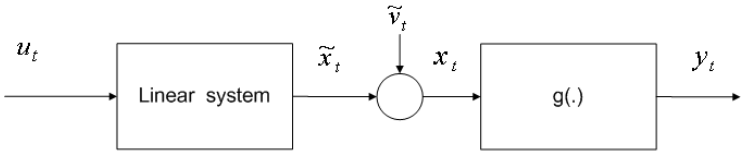

Figure 4.1: The block diagram of Wiener systems

4.2.1 Hammerstein and Wiener Systems

Consider the Hammerstein system below:

$$\begin{aligned} y_t &= d_1 y_{t-1} + ... + d_n y_{t-n} + b_1 x_{t-1} + ... + b_m x_{t-m} + v_t \\ x_t &= f(u_t) = a_0 k_0(u_t) + a_1 k_1(u_t) + ... + a_l k_l(u_t) \end{aligned} \tag{4.1}$$

where u_t is the input signal, $f(.)$ is a nonlinear function represented by the combination of known basis functions and unknown coefficients $a_0, ..., a_l$, and x_t and y_t are the input and output of a sub linear system with known structure and unknown parameters $d_1, ..., d_n$ and $b_1, ..., b_n$, and v_t denotes the noise.

Assumption 4.1. *Basis functions $k_i(u), i = 0, ..., l$ are orthonormal basis functions on a given interval $[-U_0, U_0]$ and $k_0(u)$ is a constant function. In addition, the parameters of $b_1, ..., b_m$ satisfy that $\sum_{i=1}^{m} b_i \neq 0$.*

Based on (4.1) and Assumption 4.1, we get

$$\begin{aligned} y_t = \quad & d_0 + d_1 y_{t-1} + ... + d_n y_{t-n} + b_1(a_1 k_1(u_{t-1}) + ... + a_l k_l(u_t)) \\ & + a_l k_l(u_{t-1})) + ... + b_m(a_1 k_1(u_{t-m}) + ... + a_l k_l(u_{t-m})) + v_t \end{aligned} \tag{4.2}$$

where $d_0 = k_0(u) \cdot a_0 \cdot \sum_{i=1}^{m} b_i$. The identification purpose is to estimate the unknown parameters $d = [d_0 \ ... \ d_n]'$, $b = [b_1 \ ... \ b_m]'$ and $a = [a_0 \ ... \ a_l]'$ in model (4.1) and (4.2) based on the observed input and output data $\{u_t, y_t\}$, $t = -r, ...0, 1, ..., N$ where $r = max(m, n)$ for sufficiently large N.

Let $Y = [y_1 \ ... \ y_N]'$. The Hammerstein system can be rewritten as the following matrix form:

$$\begin{aligned} Y &= \mathcal{G}d + b_1 K_1 a + ... + b_m K_m a + v \\ &= \mathcal{G}d + (b \cdot \mathcal{K})a + v \\ &= \mathcal{G}d + (\mathcal{K} \otimes a)b + v \\ &= [\mathcal{G} \quad b \cdot \mathcal{K}] \begin{bmatrix} d \\ a \end{bmatrix} + v \end{aligned} \tag{4.3}$$

where $b = [b_1 \ ... \ b_m]'$, $a = [a_1 \ ... \ a_l]'$, $d = [d_0 \ ... \ d_n]'$, $v = [v_1 \ ... \ v_N]'$ and for

$i = 1, ..., m$

$$\mathcal{G} = \begin{bmatrix} 1 & y_0 & ... & y_{1-n} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & y_{N-1} & ... & y_{N-n} \end{bmatrix}, \; K_i = \begin{bmatrix} k_1(u_{1-i}) & ... & k_l(u_{1-i}) \\ \vdots & \vdots & \vdots \\ k_1(u_{N-i}) & ... & k_l(u_{N-i}) \end{bmatrix}$$

$$\mathcal{K} = [K_1 \; ... \; K_m], \; \mathcal{K} \otimes a \triangleq [K_1 a \;\; ... \;\; K_m a]$$

$$b \cdot \mathcal{K} \triangleq b_1 K_1 + ... + b_m K_m, \; (b \cdot \mathcal{K})a = (\mathcal{K} \otimes a)b \triangleq b_1 K_1 a + ... + b_m K_m a$$

Assumption 4.2. *Input $u_t \in [-U_0, U_0]$ and noise v_t are i.i.d random variables. In addition, $E(v_t) = 0$ and $E(v_t^2) = D(v_t) = \sigma_v^2 < \infty$.*

Assumption 4.3. *$[\mathcal{G} \;\; \mathcal{K}]$ is full column rank.*

Assumption 4.4. *Either $\|b\|_2$ or $\|a\|_2$ is known and the first nonzero entry of b or a is positive.*

Remark 4.1. *We do not need the assumption that $k_i(u)$, $i = 0, ..., l$ are odd symmetrical functions assumed in [48]. Note that every set of function basis in a given interval can always be orthonormalized. For example, polynomial basis $1, x, x^2, x^3...$ including the odd basis $x, x^3, ...$ as a subset basis is easy to be orthonormalized. The assumption on $\sum_{i=1}^m b_i \neq 0$ is to guarantee parameter a_0 identifiable as $d_0 = k_0(u) \cdot a_0 \cdot \sum_{i=1}^m b_i$.*

Remark 4.2. *Legendre polynomials $p_0(u), .., p_j(u), ..., p_l(u)$ are well known orthogonal basis functions in the interval $[-1, 1]$ for $0 \leq j \leq l$ with j denoting the order of each basis function. Legendre polynomials can be produced by using Rodrigues' formula: $p_j(u) = \frac{1}{2^k k!} \frac{d^k}{du^k}(u^2 - 1)^j$. Note that $\int_{-1}^1 p_i(u) p_j(u) du = \frac{2}{2j+1} \delta_{ij}$. Based on this, it is easy to construct orthonomal basis functions in the interval $[-C, C]$ by the substitution $k_j(u) = \frac{2j+1}{2} p_j(\frac{u}{C})$ for $j = 0, ..., l$. Obviously, $k_0(u)$ is a constant function and $E(k_j(u)) = 0$ for $j = 1, ..., l$.*

Remark 4.3. *Assumption 4.3 actually refers to the requirement of the input signals. From Lemma 4.1, it is noted that when the input signals are i.i.d, it is not hard to guarantee the linear independence of* $\mathcal{G}$ *and* $\mathcal{K}$ *as they are constructed from input and output signals, respectively. Assumption 4.3 also implies that, for any* $b \neq 0$, $[\mathcal{G} \quad b \cdot \mathcal{K}]$ *is full column rank, and for any* $a \neq 0$, $[\mathcal{G} \quad \mathcal{K} \otimes a]$ *is full column rank. Assumption 4.4 is to guarantee a unique expression of the Hammerstein system model, as any pair* λa *and* b/λ *for some non-zero* λ *provides the same input-output data. Both Assumptions 4.3 and 4.4 are related to the identifiability of the nonlinear system.*

Lemma 4.1. *Under Assumptions 4.1-4.3, for any* $\mathcal{K} \in R^{N \times ml}, ml < N$, *we have* $\lim_{N\to\infty} \frac{\mathcal{K}'\mathcal{K}}{N} = I$ *almost surely where* I *is an identity matrix with dimension* $ml \times ml$.

Proof. The proof is the same as the proof of Lemma 3.2 in Chapter 3. □

Remark 4.4. *Note that almost surely means that an event occurs with probability* 1. *In Lemma 3.2, it is possible that* $\mathcal{K}$ *is a singular matrix in one realization for a particular sequence* $\{u_t\}_{i=-r}^{N} \in [-U_0 \; U_0] \otimes [-U_0 \; U_0] \otimes ... \otimes [-U_0 \; U_0] \subset R^{N+r+1}$ *but the measure of such sequences is* 0. *So such an event occurs with probability* 0.

Similarly, a Wiener system shown in Figure 4.1 is modeled by

$$\begin{aligned} \tilde{x}_t &= \tilde{b}_1\tilde{x}_{t-1} + ... + \tilde{b}_m\tilde{x}_{t-m} + \tilde{d}_1 u_{t-1} + ... + \tilde{d}_n u_{t-n} \\ x_t &= \tilde{x}_t + \tilde{v}_t \\ y_t &= g(x_t) \end{aligned} \tag{4.4}$$

Assumption 4.5. *The nonlinear output function of the Wiener system* $g(.)$ *is invertible and can be represented as* $x_t = g^{-1}(y_t) = a_0 y_t + a_1 k_1(y_t) + ... + a_l k_l(y_t)$ *where* $a_0 \neq 0$, $k_i(.), i = 1, ..., l$ *are known orthonormal basis functions in the symmetric*

interval $[-Y_0, Y_0]$ *determined by the range of* y_t.

Under Assumption 4.5, we can introduce a new intermediate variable $z_t = a_1 k_1(y_t) + ... + a_l k_l(y_t)$. Then system (4.4) is re-written as:

$$\begin{aligned} a_0 y_t + z_t - \tilde{v}_t &= \tilde{b}_1(a_0 y_{t-1} + z_{t-1} - \tilde{v}_{t-1}) + ... + \\ &\quad \tilde{b}_m(a_0 y_{t-m} + z_{t-m} - \tilde{v}_{t-m}) + \tilde{d}_1 u_{t-1} + ... + \tilde{d}_n u_{t-n} \\ z_t &= a_1 k_1(y_t) + ... + a_l k_l(y_t). \end{aligned} \tag{4.5}$$

Dividing a_0 on both sides of the first equation of (4.5) yields

$$\begin{aligned} y_t &= -\tfrac{1}{a_0} z_t + \tfrac{\tilde{b}_1}{a_0} z_{t-1} + ... + \tfrac{\tilde{b}_m}{a_0} z_{t-m} + \tilde{b}_1 y_{t-1} \\ &\quad + ... + \tilde{b}_m y_{t-m} + \tfrac{\tilde{d}_1}{a_0} u_{t-1} + ... + \tfrac{\tilde{d}_n}{a_0} u_{t-n} + v_t \end{aligned} \tag{4.6}$$

where $v_t = \tilde{v}_t - \sum_{i=1}^{m} \tilde{b}_i \tilde{v}_{t-i}$. Thus, we obtain

$$\begin{aligned} y_t &= b_0 z_t + b_1 z_{t-1} + ... + b_m z_{t-m} + c_1 y_{t-1} \\ &\quad + ... + c_m y_{t-m} + d_1 u_{t-1} + ... + d_n u_{t-n} + v_t \\ z_t &= a_1 k_1(y_t) + ... + a_l k_l(y_t) \end{aligned} \tag{4.7}$$

where $b_0 = -\frac{1}{a_0}$, $b_1 = \frac{\tilde{b}_1}{a_0}$, ..., $b_m = \frac{\tilde{b}_m}{a_0}$, and $c_1 = \tilde{b}_1$, ... , $c_m = \tilde{b}_m$, and $d_1 = \frac{\tilde{d}_1}{a_0}$, ..., $d_n = \frac{\tilde{d}_n}{a_0}$. In this way, we obtain a similar matrix form as that of the Hammerstein systems (4.3):

$$Y = \mathcal{G}d + b_0 K_0 a + b_1 K_1 a + ... + b_m K_m a + v \tag{4.8}$$

where $Y = [y_1 \ ... \ y_N]'$, $b = [b_0 \ b_1 \ ... \ b_m]'$, $a = [a_1 \ ... \ a_l]'$ and $d = [c_1 \ ... \ c_m \ d_1 \ ... \ d_n]'$

and

$$K_i = \begin{bmatrix} k_1(y_{1-i}) & ... & k_l(y_{1-i}) \\ \vdots & \vdots & \vdots \\ k_1(y_{N-i}) & ... & k_l(y_{N-i}) \end{bmatrix}$$

$$\mathcal{G} = \begin{bmatrix} y_0 & ... & y_{1-m} & u_0 & ... & u_{1-n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{N-1} & ... & y_{N-m} & u_{N-1} & ... & u_{N-n} \end{bmatrix}$$

As both Hammerstein and Wiener systems can be formulated to the matrix from given in (4.3), we investigate how to identify (4.3) based on the observed input and output data $\{u_t, y_t\}$ sequences where $t = -r, ..., 0, 1, ..., N$ and $r = max(m, n)$ for sufficiently large N.

Remark 4.5. *In order to achieve the above objective with a unified algorithm based on fixed point iteration, we introduce new intermediate variable z_t in transforming Wiener systems (4.7) into the form of Hammerstein systems (4.3).*

4.2.2 Fixed Point Iteration Algorithm

The ideas of estimating a, b and d in (4.3) are outlined as follows. Before employing the fixed point iteration algorithm, we obtain $\hat{d}$ first without knowing a and b. Then the estimate $\hat{b}$ of the unknown parameter vector b can be represented as

$$\hat{b} = \mathcal{F}(\hat{b}) \tag{4.9}$$

The function $\mathcal{F}(.)$ will be abstracted later from an iterative algorithm derived for estimating $\hat{b}$ and we will show that (4.9) has a unique fixed point which corresponds

to the true parameter b when N tends to infinity. We will also prove that the sequence $\hat{b}(0), \hat{b}(1), ...\hat{b}(k)...$ generated by the iterative algorithm converges to the fixed point as $k \to \infty$.

Simultaneously, at each iteration step, determining the estimate $\hat{a}$ of a is to solve a linear equation by substituting $\hat{b}$ and $\hat{d}$ into (4.3).

Now we show how to obtain $\hat{d}$. From (4.3),

$$\begin{aligned} Y &= \mathcal{G}d + b_1K_1a + ... + b_mK_ma + v \\ &= \mathcal{G}d + [K_1 \ ... \ K_m] \begin{bmatrix} b_1a \\ \vdots \\ b_ma \end{bmatrix} + v \\ &= \mathcal{G}d + \mathcal{K}\gamma + v \end{aligned} \tag{4.10}$$

where $\gamma' = \begin{bmatrix} (b_1a)' & ... & (b_ma)' \end{bmatrix}$. This is the same model as in (3.4) where the approximation error term $\xi = 0$ shown in Chapter 3. For convenience, we also present its solving procedure as follows. Let $P_{\mathcal{K}} = \mathcal{K}\mathcal{K}^+$ and $P_{\mathcal{G}} = \mathcal{G}\mathcal{G}^+$ denote projection operators onto span$\{\mathcal{K}\}$ and span$\{\mathcal{G}\}$, respectively, where span$\{.\}$ is the space spanned by the column vectors of a matrix and $\mathcal{K}^+ = (\mathcal{K}'\mathcal{K})^{-1}\mathcal{K}'$ is the generalized matrix inverse. From (4.10), we obtain $P_{\mathcal{G}}\mathcal{G}d = P_{\mathcal{G}}(Y - \mathcal{K}\gamma) - P_{\mathcal{G}}v$ and $P_{\mathcal{K}}\mathcal{K}\gamma = P_{\mathcal{K}}(Y - \mathcal{G}d) - P_{\mathcal{K}}v$. Note that the noise space is independent of space $span\{\mathcal{K}\} \cup span\{\mathcal{G}\}$ with $E(v) = \{0\}$ based on Assumption 4.2. Thus the noise space is orthogonal to $span\{\mathcal{K}\} \cup span\{\mathcal{G}\}$. And as $P_{\mathcal{G}}v$ and $P_{\mathcal{K}}v$ are operators projecting the noise to the space $span\{\mathcal{G}\}$ and $span\{\mathcal{K}\}$, we have $P_{\mathcal{G}}v = 0$ and $P_{\mathcal{K}}v = 0$. Based on Lemma 4.2, we also have $P_{\mathcal{G}}\mathcal{G} = \mathcal{G}\mathcal{G}^+\mathcal{G} = \mathcal{G}$ and $P_{\mathcal{K}}\mathcal{K} =$

$\mathcal{K}\mathcal{K}^{+}\mathcal{K} = \mathcal{K}$. Thus, we obtain

$$\mathcal{G}d = P_{\mathcal{G}}(Y - \mathcal{K}\gamma) \tag{4.11}$$

$$\mathcal{K}\gamma = P_{\mathcal{K}}(Y - \mathcal{G}d) \tag{4.12}$$

From (4.11) and (4.12), it is easy to establish that $\mathcal{G}d = P_{\mathcal{G}}(Y - P_{\mathcal{K}}(Y - \mathcal{G}d))$, which gives $(I - P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G}d = P_{\mathcal{G}}(I - P_{\mathcal{K}})Y$. Then $\hat{d}$ is obtained as

$$\hat{d} = ((I - P_{\mathcal{G}}P_{\mathcal{K}})\mathcal{G})^{+}P_{\mathcal{G}}(I - P_{\mathcal{K}})Y. \tag{4.13}$$

Note that we do not need any information of b or a to obtain $\hat{d}$. For a currently obtained $\hat{b}(k)$ and $\hat{d}$, a least square solution $\hat{a}_{op}(k)$ for a is given as

$$\hat{a}_{op}(k) = \mathcal{F}_1(\hat{b}(k)) = ((\hat{b}(k) \cdot \mathcal{K})'(\hat{b}(k) \cdot \mathcal{K}))^{-1}(\hat{b}(k) \cdot \mathcal{K})(Y - \mathcal{G}\hat{d}) \tag{4.14}$$

Based on Assumption 4.4, the estimates $\hat{a}$ with normalization is given by

$$\hat{a}(k) = \mathcal{F}_2(\hat{a}_{op}(k)) = \frac{\hat{a}_{op}(k)\|a\|_2}{\|\hat{a}_{op}(k)\|_2} \tag{4.15}$$

Clearly, $\|\hat{a}(k)\| = \|a\|_2$. After replacing a and b with $\hat{a}(k)$ and $\hat{d}$, respectively, equation (4.3) becomes

$$Y - \mathcal{G}\hat{d} = (\mathcal{K} \otimes \hat{a}(k))b + v \tag{4.16}$$

From (4.16), we can have the following iterative algorithm which gives the least square estimate of b at step $k + 1$

$$\hat{b}(k+1) = \mathcal{F}_3(\hat{a}(k)) = ((\mathcal{K} \otimes \hat{a}(k))'(\mathcal{K} \otimes \hat{a}(k)))^{-1}(\mathcal{K} \otimes \hat{a}(k))'(Y - \mathcal{G}\hat{d}) \tag{4.17}$$

Based on (4.15) and (4.14), it can be obtained that

$$\hat{b}(k+1) = \mathcal{F}_3(\mathcal{F}_2(\hat{a}_{op}(k))) = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(\hat{b}(k)))) \tag{4.18}$$

To study the convergence property of $\hat{b}(k)$ in (4.18), we represent $\hat{b}$ in the form of (4.9) as follows

$$\hat{b} = \mathcal{F}(\hat{b}) \triangleq ((\mathcal{K} \otimes \mathcal{F}_2(\mathcal{F}_1(\hat{b}))'(\mathcal{K} \otimes \mathcal{F}_2(\mathcal{F}_1(\hat{b})))^{-1}(\mathcal{K} \otimes \mathcal{F}_2(\mathcal{F}_1(\hat{b}))'(Y - \mathcal{G}\hat{d}) \tag{4.19}$$

where $\mathcal{F}(.) = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(.)))$.

Now we summarize the iterative algorithm which starts with an arbitrary nonzero initial value $\hat{b}(0)$ as follows.

Step 1: Estimate $\hat{d}$ by using (4.13) and let $k = 0$.

Step 2: Obtain estimates $\hat{a}(k)$ by using (4.14) and (4.15) for given $\hat{b}(k)$.

Step 3: Obtain $\hat{b}(k+1)$ from (4.17) for given $\hat{a}(k)$ and $\hat{d}$.

Step 4: If a stopping criterion is satisfied, then let $\hat{a} = \hat{a}(k) \cdot sgn(\hat{a}_1(k))$ and $\hat{b} = \hat{b}(k+1) \cdot sgn(\hat{a}_1(k))$, and end. Otherwise, go to Step 2. Finally, the obtained estimates are denoted as $\hat{a}$, $\hat{b}$ and $\hat{d}$.

Note that $\hat{a}(k)$ and $\hat{b}(k)$ can be permuted at Steps 1-4 and $\hat{a} = \hat{a}(k) \cdot sgn(\hat{a}_1(k))$ is to guarantee that the first element of $\hat{a}$ remains positive. There are several ways to define the stopping criterion in Step 3 of the algorithm. For example, one can consider the difference of absolute value of the difference between $\hat{a}(k)$ and $\hat{a}(k-1)$ or ($\hat{b}(k)$ and $\hat{b}(k-1)$).

4.3 Convergence Analysis

In this section, we first prove that $\lim_{N\to\infty} \hat{d} = d$ without knowing a and b. Then we show that the fixed point of equation (4.19) is unique and obtain the convergence property of the algorithm.

Lemma 4.2. *[47] If A is a full column rank, then $A^+A = I$.*

Lemma 4.3. *Under Assumption 4.3, matrix $(I - P_{\mathcal{G}}P_{\mathcal{K}})$ is full column rank.*

Proof. The proof is the same as the proof in in Lemma 2.6 in Chapter 2. □

Theorem 4.1. *Under Assumptions 4.1-4.3, for the estimates $\hat{d}$ given in (4.13), we have $\lim_{N\to\infty} \hat{d} = d$ almost surely.*

Proof. The proof follows the same procedure as Theorem 3.1 in Chapter 3. □

Lemma 4.4. Contraction Mapping Theorem. *[71]*
Let (X, D) be a non-empty complete metric space where $D(.,.)$ is a metric on X. Let $\mathcal{F} : X \to X$ be a contraction mapping on X, i.e., there is a nonnegative real number $Q < 1$ such that $D(\mathcal{F}(x), \mathcal{F}(y)) \leq Q \cdot D(x, y)$, for all $x, y \in X$. Then the map $\mathcal{F}$ admits one and only one fixed point $x^* \in X$ which means $x^* - \mathcal{F}(x^*) = 0$. Furthermore, this fixed point can be found as follows: start with an arbitrary element $x(0)$ in X and define an iterative sequence by $x(k+1) = \mathcal{F}(x(k))$ for $k = 1, 2, 3, ...$ This sequence converges to x^* .

From Assumption 4.5, define $X_a = \{\hat{a} \mid \|\hat{a}\|_2 = \|a\|_2, \hat{a}_0 > 0\}$, $X_b = \{\hat{b} \mid \|\hat{b}\|_2 \leq \|b\|_2\}$, $D(x, y) = \|x - y\|_2$.

Theorem 4.2. *Under Assumptions 4.1- 4.4, when $N \to \infty$, $\mathcal{F}(\hat{b}) : X_b \to X_b$ defined in (4.19) is a contraction mapping on X_b, thus equation (4.19) has a unique fixed point of $\hat{b} = \mathcal{F}(\hat{b})$ on X_b which is the true parameter b.*

Proof. Firstly, we prove that when N approaches infinity, $\mathcal{F}(\hat{b})$ maps $b \in X_b$ into X_b, i.e, $\mathcal{F}(\hat{b}) : X_b \to X_b$. Secondly, we show that $\mathcal{F}(\hat{b})$ is a contraction mapping on X_b and finally the true parameters is the unique fixed point of $\hat{b} = \mathcal{F}(\hat{b})$.

From Assumption 4.3, for any $\hat{a} \neq 0 \in X_a$, $(\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes \hat{a})$ has an inverse. Then

$$\hat{b} = \mathcal{F}_3(\hat{a}) = ((\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes \hat{a}))^{-1}(\mathcal{K} \otimes \hat{a})'(Y - \mathcal{G}\hat{d}) \tag{4.20}$$

Note that $Y - \mathcal{G}\hat{d} = (b \cdot \mathcal{K})a + \mathcal{G}(d - \hat{d}) + v = (\mathcal{K} \otimes a)b + \mathcal{G}(d - \hat{d}) + v$. Based on Theorem 4.1,

$$\lim_{N\to\infty} \hat{d} = d \tag{4.21}$$

almost surely. Then (4.20) becomes

$$\hat{b} = \mathcal{F}(\hat{b}) = ((\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes \hat{a}))^{-1}(\mathcal{K} \otimes \hat{a})'((\mathcal{K} \otimes a)b + v) \tag{4.22}$$

By following similar analysis in deriving Theorem 4.1, we have

$$\lim_{N\to\infty}((\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes \hat{a}))^{-1}(\mathcal{K} \otimes \hat{a})'v = 0 \tag{4.23}$$

almost surely. Then

$$\hat{b} = ((\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes \hat{a}))^{-1}(\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes a)b \tag{4.24}$$

From Lemma 4.1, $k_i(.)$ for $i = 0, 1, ...l$ are orthonormal functions, which gives

$$\lim_{N\to\infty} \frac{\mathcal{K}'\mathcal{K}}{N} = I \tag{4.25}$$

where I is an identity matrix. Therefore,

$$\|\mathcal{F}(\hat{b})\|_2 = \frac{\|(\mathcal{K}\otimes\hat{a})'(\mathcal{K}\otimes a)\|_2}{\|(\mathcal{K}\otimes\hat{a})'(\mathcal{K}\otimes\hat{a})\|_2} \cdot \|b\|_2 = \frac{\|\hat{a}'a\|_2}{\|\hat{a}'\hat{a}\|_2}\|b\|_2 \tag{4.26}$$

Based on the definition of X_a from Assumption 4.4, if $\hat{a}$ belongs to X_a, then $-\hat{a}$ does not belong to X_a. Thus,

$$\frac{\|\hat{a}'a\|_2}{\|\hat{a}'\hat{a}\|_2} = \begin{cases} 1 & \text{if and only if } \hat{a} = a \\ \mathcal{F}_b \ , \ 0 < \mathcal{F}_b < 1 & \text{otherwise} \end{cases} \tag{4.27}$$

This gives $\|\mathcal{F}(\hat{b})\|_2 \leq \|b\|_2$. Therefore, $\mathcal{F}(\hat{b}) \in X_b$ and $\mathcal{F}(\hat{b}) : X_b \to X_b$.

Now we prove that $\mathcal{F}(\hat{b}) : X_b \to X_b$ is a contraction mapping on X_b as N approaches infinity. Note that the contour planes of $\|\mathcal{F}(\hat{b})\|_2$ are concentric spheres $\|\hat{b}\|_2 = B$ with different radius B as seen in (4.26). This means that the gradient of $\mathcal{F}(\hat{b})$ is in the radial direction. It is known that the maximum magnitude of the directional derivative $D_u\mathcal{F}(\hat{b})$ along direction u attains its maximum when the vector u is aligned with the gradient of $\mathcal{F}(\hat{b})$, which is the radial direction $\hat{b}$. The maximum value of the directional derivative is the magnitude of its gradient $\|\nabla\mathcal{F}\|_2$. Define

$$Q = \|\nabla\mathcal{F}\|_2 = \|\frac{\partial\mathcal{F}(\hat{b})}{\partial\hat{b}}\|_2 = max\{D_u\mathcal{F}(\hat{b})\} = D_{\hat{b}}\mathcal{F}(\hat{b}) \tag{4.28}$$

where $\|\frac{\partial\mathcal{F}(\hat{b})}{\partial\hat{b}}\|_2$ is the directional derivative of $\mathcal{F}(\hat{b})$ with respect to $\hat{b}$ along the

direction $\hat{b}$. From (4.19), $\hat{b} = \mathcal{F}(b) = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(\hat{b})))$, then

$$\begin{aligned} Q &= \|\frac{\partial \mathcal{F}(\hat{b})}{d\hat{b}}\|_2 \\ &= \|\frac{\partial \mathcal{F}_3}{\partial \hat{a}} \cdot \frac{\partial \mathcal{F}_2}{\partial \hat{a}_{op}} \cdot \frac{\partial \mathcal{F}_1}{\partial \hat{b}}\|_2 \\ &= \|\frac{\partial \mathcal{F}_3}{d\hat{a}}\|_2 \cdot \|\frac{\partial \mathcal{F}_2}{\partial \hat{a}_{op}}\|_2 \cdot \|\frac{\partial \mathcal{F}_1}{\partial \hat{b}}\|_2 \\ &= \|\frac{\partial \mathcal{F}_3}{\partial \hat{a}}\|_2 \cdot \|\frac{\partial \hat{a}}{\partial \hat{a}_{op}}\|_2 \cdot \|\frac{\partial \hat{a}_{op}}{\partial \hat{b}}\|_2 \end{aligned} \tag{4.29}$$

For a nonzero $\hat{a} \neq 0$, we have

$$\begin{aligned} \|\frac{\partial \mathcal{F}_3}{\partial \hat{a}}\|_2 &= max\{D_u \mathcal{F}_3(\hat{a})\} = D_{\hat{a}} \mathcal{F}_3(\hat{a}) \\ &= \lim_{\|\Delta a\|_2 \to 0} \frac{\|\mathcal{F}_3(\hat{a}+\Delta a) - \mathcal{F}_3(\hat{a})\|_2}{\|\Delta a\|_2} \\ &= \lim_{\|\Delta a\|_2 \to 0} \frac{\|(\mathcal{K} \otimes (\Delta a))'(\mathcal{K} \otimes a)\|_2}{\|(\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes \hat{a})\|_2 \cdot \|\Delta a\|_2} \cdot \|b\|_2 \\ &= \frac{\|(\mathcal{K} \otimes (\vec{a}))'(\mathcal{K} \otimes a)\|_2}{\|(\mathcal{K} \otimes \hat{a})'(\mathcal{K} \otimes \hat{a})\|_2} \cdot \|b\|_2 \\ &= \frac{\|\vec{a}' a\|_2}{\|\hat{a}' \hat{a}\|_2} \cdot \|b\|_2 \end{aligned} \tag{4.30}$$

where $\vec{a} = \frac{\Delta a}{\|\Delta a\|_2}$ is a unit vector along the direction of $\hat{a}$. It is easy to obtain that

$$\|\frac{\partial \hat{a}}{\partial \hat{a}_{op}}\|_2 = \|\frac{\partial \mathcal{F}_2}{\partial \hat{a}_{op}}\|_2 = \frac{\|a\|_2}{\|a_{op}\|_2} \tag{4.31}$$

From (4.25), we have

$$\|\frac{\partial \hat{a}_{op}}{\partial \hat{b}}\|_2 = max\{D_u \mathcal{F}_1(\hat{b})\} = D_{\hat{b}} \mathcal{F}_1(\hat{b}) = \frac{\|(\vec{b} \cdot \mathcal{K})'(b \cdot \mathcal{K})\|_2}{\|(\hat{b} \cdot \mathcal{K})'(\hat{b} \cdot \mathcal{K})\|_2} \cdot \|a\|_2 = \frac{\|\vec{b}' b\|_2}{\|\hat{b}' \hat{b}\|_2} \cdot \|a\|_2 \tag{4.32}$$

where $\vec{b}$ is a unit vector along the direction of $\hat{b}$. Combining (4.29)-(4.32) yields

$$\begin{aligned} Q &= \|\frac{\partial \mathcal{F}_3}{\partial \hat{a}}\|_2 \cdot \|\frac{\partial \hat{a}}{\partial \hat{a}_{op}}\|_2 \cdot \|\frac{\partial \hat{a}_{op}}{\partial \hat{b}}\|_2 \\ &= \frac{\|\vec{a}' a\|_2}{\|\hat{a}' \hat{a}\|_2} \cdot \|b\|_2 \cdot \frac{\|a\|_2}{\|a_{op}\|_2} \cdot \frac{\|\vec{b}' b\|_2}{\|\hat{b}' \hat{b}\|_2} \cdot \|a\|_2 \\ &= \frac{\|\hat{a}' a\|_2}{\|\hat{a}' \hat{a}\|_2} \cdot \frac{\|a\|_2}{\|a_{op}\|_2} \cdot \frac{\|\hat{b}' b\|_2}{\|\hat{b}' \hat{b}\|_2} \end{aligned} \tag{4.33}$$

where $\hat{a} = \|a\|_2\vec{a}$ and $\hat{b} = \|b\|_2\vec{b}$. Also, (4.14), (4.21) and (4.23) give

$$\begin{aligned} \hat{a}_{op} &= ((\hat{b}\cdot\mathcal{K})'(\hat{b}\cdot\mathcal{K}))^{-1}(\hat{b}\cdot\mathcal{K})(Y - G\hat{d}) \\ &= ((\hat{b}\cdot\mathcal{K})'(\hat{b}\cdot\mathcal{K}))^{-1}(\hat{b}\cdot\mathcal{K})((b\cdot\mathcal{K})a + \mathcal{G}(\hat{d} - d) + v) \\ &= ((\hat{b}\cdot\mathcal{K})'(\hat{b}\cdot\mathcal{K}))^{-1}(\hat{b}\cdot\mathcal{K})(b\cdot\mathcal{K})a \end{aligned} \tag{4.34}$$

Multiplying $\mathcal{K}\otimes\hat{a}$ and $\hat{b}\cdot\mathcal{K}$ on both sides of (4.24) and (4.34), respectively, we get

$$\begin{aligned} (\mathcal{K}\otimes\hat{a})\hat{b} &= (\mathcal{K}\otimes\hat{a})((\mathcal{K}\otimes\hat{a})'(\mathcal{K}\otimes\hat{a}))^{-1}(\mathcal{K}\otimes\hat{a})'(\mathcal{K}\otimes a)b \\ (\hat{b}\cdot\mathcal{K})\hat{a}_{op} &= (\hat{b}\cdot\mathcal{K})((\hat{b}\cdot\mathcal{K})'(\hat{b}\cdot\mathcal{K}))^{-1}(\hat{b}\cdot\mathcal{K})(b\cdot\mathcal{K})a \end{aligned} \tag{4.35}$$

i.e,

$$\begin{aligned} \hat{b}'(\mathcal{K}\otimes\hat{a})'(\mathcal{K}\otimes\hat{a})\hat{b} &= b'(\mathcal{K}\otimes a)'(\mathcal{K}\otimes\hat{a})((\mathcal{K}\otimes\hat{a})'(\mathcal{K}\otimes\hat{a}))^{-1}(\mathcal{K}\otimes\hat{a})'(\mathcal{K}\otimes a)b \\ \hat{a}_{op}(\hat{b}\cdot\mathcal{K})'(\hat{b}\cdot\mathcal{K})\hat{a}_{op} &= a'(b\cdot\mathcal{K})'(\hat{b}\cdot\mathcal{K})((\hat{b}\cdot\mathcal{K})'(\hat{b}\cdot\mathcal{K}))^{-1}(\hat{b}\cdot\mathcal{K})(b\cdot\mathcal{K})a \end{aligned} \tag{4.36}$$

By combining (4.36) and (4.25), it can be obtained that

$$\begin{aligned} \|\hat{b}\|_2\|\hat{a}\|_2 &= \|b\|_2\|a\|_2 \\ \|\hat{b}\|_2\|\hat{a}_{op}\|_2 &= \|b\|_2\|a\|_2 \end{aligned} \tag{4.37}$$

which gives

$$\|\hat{a}_{op}\|_2 = \|\hat{a}\|_2. \tag{4.38}$$

Then (4.33) becomes

$$Q = \|\frac{\partial\mathcal{F}}{\partial\hat{b}}\|_2 = \frac{\|\hat{a}'a\|_2}{\|\hat{a}'\hat{a}\|_2}\cdot\frac{\|\hat{b}'b\|_2}{\|\hat{b}'\hat{b}\|_2} \tag{4.39}$$

As $\|\hat{a}\|_2 = \|a\|_2$ and $\|\hat{b}\|_2 = \|b\|_2$, we obtain $Q = \|\frac{\partial\mathcal{F}(\hat{b})}{\partial\hat{b}}\|_2 < 1$ as long as $\hat{a} \neq a$ or $\hat{b} \neq b$. So we have $\forall x, y \in X_b$,

$$D(\mathcal{F}(x), \mathcal{F}(y)) \leq Q \cdot D(x, y)$$

Finally from Lemma 3.4 and (4.39), when $k \to \infty$, $\hat{b}(k)$ converges to the unique fixed point b of $\hat{b} = \mathcal{F}(\hat{b})$, which implies Q converges to 1 and $\hat{a}(k)$ converges to a. □

Remark 4.6. *Note that parameters a and b can be permuted. If we fix the norm $\hat{b}$ in (4.17), we obtain an equation for iteration with respect to a, i.e, $\hat{a} = \tilde{\mathcal{F}}(\hat{a})$. Define $\tilde{X}_a = \{\hat{a} \mid \|\hat{a}\|_2 \leq \|a\|_2\}$, $\tilde{X}_b = \{\hat{b} \mid \|\hat{b}\|_2 = \|b\|_2, b_0 > 0\}$. We have that $\hat{a} = \tilde{\mathcal{F}}(\hat{a}) : \tilde{X}_a \to \tilde{X}_a$ which is a contraction mapping on $\tilde{X}_a$. When $N \to \infty$ and $k \to \infty$, the unique fixed point a^* of equation $\hat{a} = \tilde{\mathcal{F}}(\hat{a}) : \tilde{X}_a \to \tilde{X}_a$ corresponds to the true parameter a almost surely.*

4.4 Examples

We now use two examples to illustrate the proposed fixed point iteration algorithm and verify the convergence results for both Hammerstein and Wiener systems.

Example 4.4.1. *Consider a Hammerstein system*

$$\begin{aligned} y_t &= 0.4y_{t-1} + 0.1y_{t-1} + 0.5x_t + 0.3x_{t-1} + 0.2x_{t-2} + v_t \\ x_t &= 0.1 + 0.6u_t + 0.3u_t^2 \end{aligned}$$

where $u_t \in [-1, 1]$ and v_t is white noise with zero mean and standard derivation 0.3.

By orthonormalizing the function basis 1, u and u^2 to orthonormal polynomials given as $k_1(u) = \frac{1}{2}$, $k_2(u) = \frac{3}{2}u$ and $k_3(u) = \frac{5}{2}\frac{1}{2}(3u^2 - 1)$ on the interval $[-1, 1]$, we get $x_t = 0.1k_1(u_t) + 0.4k_2(u_t) + 0.08k_3(u_t)$. Then this Hammerstein system can be

rewritten as

$$\begin{aligned} y_t &= 0.4y_{t-1} + 0.1y_{t-1} + 0.5x_t + 0.3x_{t-1} + 0.2x_{t-2} + v_t \\ x_t &= 0.1k_1(u_t) + 0.4k_2(u_t) + 0.08k_3(u_t) \end{aligned}$$

and true parameters to be estimated are $a = [0.1\ 0.4\ 0.08]$, $d = [0.4\ 0.1]$ and $b = [0.5\ 0.3\ 0.2]$. We choose $N = 500$ and fix $\|b\|_1 = |b_1| + |b_2| + |b_3| = 1$ with

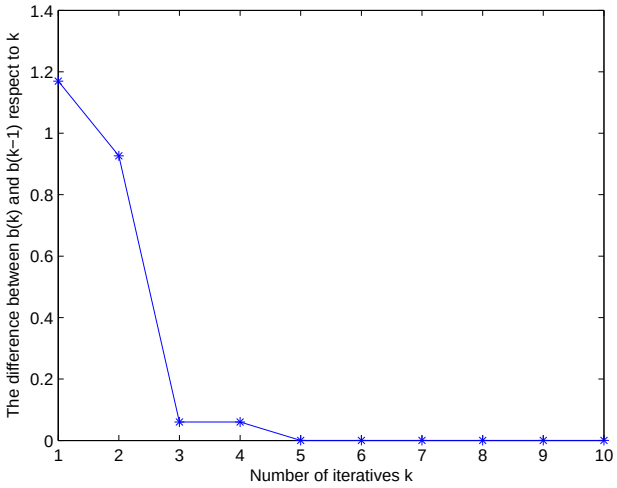

Figure 4.2: The illustration that fixed point iteration algorithm converges in a few iterations

$b_1 > 0$. Note that both $\|.\|_1$ and $\|.\|_2$ can be used to fix the norm of the parameters vector. The ratio of the standard derivation of the noise v_t to that of the output y_t is $\frac{\sigma_v}{\sigma_y} = 66\%$, which shows a high noise level. By implementing the proposed iterative algorithm, we obtain $\hat{a} = [0.0993\ 0.3974\ 0.0837]$, $d = [0.4110\ 0.0922]$ and $\hat{b} = [0.5015\ 0.2988\ 0.1997]$. Obviously, the estimates are very close to the true values. To show how the estimates converges to the fixed point with respect to the number of iterations, we calculate the difference $\sum_{i=1}^{m} |\hat{b}_i(k+1) - \hat{b}_i(k)|$ at each iteration and use it as a stop criterion. Figure 4.2 shows that the algorithm converges in only a few iterations. To show how the estimates behave with the number of data points N, we plot the square error of $\|\hat{b} - b\|_2^2$ with respect to N

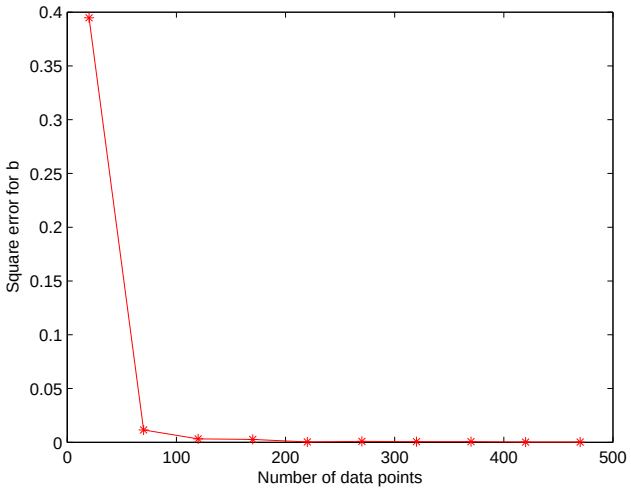

Figure 4.3: Estimation error with respect to number of data points

in Figure 4.3. This figure shows that the square error can be made small enough by choosing N appropriately.

Example 4.4.2. *Consider a Wiener system*

$$\begin{aligned} \tilde{x}_t &= 0.4\tilde{x}_{t-1} + 0.3\tilde{x}_{t-2} - 0.2u_t - 0.1u_{t-1} + 0.1u_{t-2} \\ x_t &= \tilde{x}_t + \tilde{v}_t \\ y_t &= g(x_t) \end{aligned}$$

where $\tilde{v}_t$ *is white noise with zero mean and standard derivation* 0.1 *and the inverse function of the nonlinear function* $g(.)$ *is* $x_t = g^{-1}(y_t) = y_t + 0.9y_t^2 + 0.1y_t^3$.

In this example, one can orthonormalize the basis function y_t y_t^2 and y_t^3 to a set of orthonormal basis functions and do the identification in the same way as in Example 4.4.1. The identification process is similar and the algorithm converges to the true parameters in a few iteration steps.

We now explore the possibility of identifying also the system without orthonormalizing the basis functions to Legendre polynomials, namely without the orthormal

requirement of Assumption 4.5. In this case, we identify the model as follows

$$\begin{aligned} y_t &= 0.4y_{t-1} + 0.3y_{t-2} - 0.2u_t - 0.1u_{t-1} - 0.5u_{t-2} \\ &\quad - z_t + 0.4z_{t-1} + 0.3z_{t-2} + v_t \\ z_t &= 0.9k_1(y_t) + 0.1k_2(y_t) \end{aligned}$$

where the basis functions $k_1(y_t) = y_t^2$ and $k_2(y_t) = y_t^3$ are not orthonormal basis functions. By comparing with (4.8), the true parameters to be estimated are

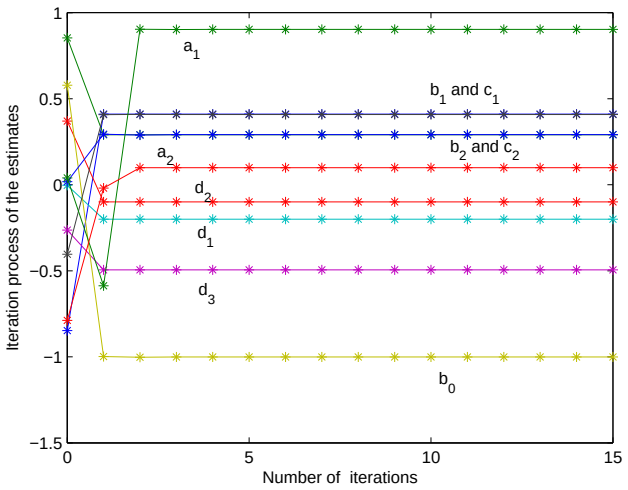

Figure 4.4: The illustration that fixed point iteration algorithm converges in a few iterations

$a = [0.9\ 0.1]$, $b = [b_0\ b_1\ b_2] = [-1\ 0.4\ 0.3]$ and $d = [c_1\ c_2\ d_1\ d_2\ d_3] = [0.4\ 0.3\ -0.2\ -0.1\ -0.5]$. The estimates of the parameters in the sub linear system are obtained iteratively as shown in Figure 4.4. Note that the initial values of the estimates are given arbitrarily. The obtained estimates are $\hat{a} = [0.8992\ 0.0983]$, $\hat{b} = [b_0\ b_1\ b_2] = [-1.007\ 0.4000\ 0.2933]$ and $\hat{d} = [c_1\ c_2\ d_1\ d_2\ d_3] = [0.3963\ 0.2898\ -0.1993\ -0.0996\ 0.4957]$ when $N = 500$. As observed in Figure 4.4, we obtain $\hat{d}(k) = [\hat{c}_1(k), \hat{c}_2(k)\ \hat{d}_1(k)\ \hat{d}_2(k)\ \hat{d}_3(k)]$ in one iteration step ($k = 1$) and all the parameters converge to their respective true values.

From the simulation results, we conjecture that our proposed algorithm may be also applicable to identifying Hammerstein and Wiener systems in which the nonlinear function is represented by a linear combination of a set of general basis functions.

4.5 Conclusion

In this chapter, fixed point iteration algorithm is introduced to identifying both Hammerstein and Wiener systems with a unified algorithm. This newly proposed estimation algorithm gives consistent estimates of the parameters under any arbitrary nonzero initial conditions. The performance of the proposed method is also verified by simulation examples. We feel that the idea of using fixed point iteration for convergence analysis in this chapter can be generalized to study the convergence property of many other nonlinear parameter estimation algorithms. In next Chapter, we extend the Hammerstein and Wiener models to more general bilinear models.

Chapter 5

Fixed Point Iteration for The Identification of Bilinear Models

In this chapter, we consider identifying bilinear models which is more general than the model in Chapter 4 based on fixed point iteration. As an application, a block-oriented system represented by a cascade of a dynamic linear (L), a static nonlinear (N) and a dynamic linear (L) subsystems is illustrated. This gives a solution to the long-standing convergence problem of iteratively identifying LNL Wiener-Hammerstein models. In addition, we extend the static nonlinear function (N) to a nonparametric model represented by using kernel machine.

5.1 Introduction

One common model that arises in science and engineering is the bilinear model [60], especially in nonlinear system identification [61], signal processing and classification [62], and many other areas of socioeconomics and biology. For example, one class of block-oriented systems consisting of a dynamic linear (L), a static nonlinear

(N) and a dynamic linear (L) subsystems in series can be conveniently formulated as bilinear models. Such an LNL cascade system is called an Wiener-Hammerstein system [63] and its identification has been widely studied in [63]-[70].

Due to the wide range applications of bilinear models, there is a strong motivation to develop identification algorithms for such models. Among the existing schemes, an iterative algorithm originated in [69] has been extensively used. As pointed out in [70], if the iterative algorithm converges, it converges rapidly and is simple to be implemented. However, the convergence is generally hard to achieve and unknown in identifying bilinear models. In fact, it was pointed out in [52] and [64] that the convergence even for LNL systems with a parametric model [70] representing the N part is still unknown. The main difficulty in obtaining the convergence property is that a block-oriented system contains internal variables, which are generally unmeasurable. It is noted that using nonparametric models [27] to represent nonlinear static functions, which is more efficient and general than the parametric model especially when the nonlinear functions are non-smooth or discontinuous, makes the convergence property even more difficult to obtain.

In this chapter, we propose an algorithm for the identification of bilinear models inspired by fixed point theory [71]. Fixed point of a function is a point that is mapped to itself by the function. In many fields, equilibrium is a fundamental concept that can be described in terms of fixed points and the convergence of a sequence can be analyzed. By exploiting the fixed point theory, it can be proven that the iteration produced by the proposed iterative algorithm is a contraction mapping on a metric space when the number of data points approaches infinity. This implies the existence and uniqueness of a fixed point of the iterated function sequence. Therefore the convergence of the iteration can be established. For application, we show that LNL Wiener-Hammerstein models with a nonparametric

model representing the N part belong to bilinear models. With this, the long-standing convergence problem of iteratively identifying LNL Wiener-Hammerstein models is solved by applying the proposed algorithm in this chapter.

The remaining part of this chapter is organized as follows. In Section 5.2, we introduce bilinear models and fixed point theory as well as the iterative identification method in identifying bilinear models. The representation of LNL Wiener-Hammerstein systems as bilinear models and the analysis are shown in Section 5.3. Some simulation examples are given in Section 5.4 to show the performance of the proposed algorithm. Finally, this chapter is concluded in Section 5.5.

5.2 Bilinear Models and Fixed Point Theory

In this section, we first present a common model of bilinear systems. Then an iterative algorithm is proposed to achieve the identification objective with available input-output data points. We mainly show that the estimate $\hat{b}$ of an unknown parameter vector b can be represented as $\hat{b} = \mathcal{F}(\hat{b})$, where function $\mathcal{F}(.)$ is obtained from the iterative algorithm. It will be shown that $\hat{b} = \mathcal{F}(\hat{b})$ has a unique fixed point which corresponds to the true parameter vector b when the number of data points tends to infinity. We will also prove that the sequence $\{\hat{b}(0), \hat{b}(1),\ \hat{b}(2),\ ...,\ \hat{b}(k)\ ...\}$ generated by the iterated function sequence $\{\hat{b}(0),\ \mathcal{F}(\hat{b}(0)),\ \mathcal{F}(\mathcal{F}(\hat{b}(0))),\ ...\}$ converges to the fixed point as $k \to \infty$.

5.2.1 Bilinear Models

Usually a linear system described in (5.1) is considered for its simplicity,

$$y_i = \phi_i a + v_i,\ i = 1, ..., N \tag{5.1}$$

where $\phi_i \in R^{1\times M}$ is a known system matrix, $a \in R^{M\times 1}$ is an unknown parameter vector, and y_i denotes an observation of the system output with unknown noise v_i. Another common yet more general model in science and engineering is the following bilinear model [74].

$$\begin{aligned} y_i &= b'\Psi^i a + v_i \\ &= b' \begin{bmatrix} \Psi^i_{11} & ... & \Psi^i_{1L} \\ \vdots & \cdots & \vdots \\ \Psi^i_{M1} & ... & \Psi^i_{ML} \end{bmatrix} a + v_i, \quad i = 1, ..., N \end{aligned} \tag{5.2}$$

where $b = [b_1\ ...\ b_M]' \in R^{M\times 1}$ and $a = [a_1\ ...\ a_L]' \in R^{L\times 1}$ are two vectors of unknown parameters with superscript $'$ denoting the transpose, and $\Psi^i \in R^{M\times L}$, for $i = 1, ..., N$, is a sequence of $M \times L$ dimensional matrices which describes a bilinear map from the parameter space to the observation space. The model is called 'bilinear model' because when either a or b is fixed, the relationship between y_i and b or a is linear. Note that Ψ^i_{jt}, for $j = 1, ..., M$, and $t = 1, ..., L$, denoting a component of matrix Ψ^i is usually related to the available input output data points. Let

$$\begin{aligned} Y &= [y_1\ ...\ y_N]' \\ v &= [v_1\ ...\ v_N]' \end{aligned} \tag{5.3}$$

We express the bilinear model in a matrix form $Y = F(a, b) + v$ where $F(.,.)$ denotes the nonlinearity of the bilinear model, and it can be divided into the

following two sub-linear models:

$$Y = A_a b + v = (A \otimes a)b + v, \ if \ a \ is \ known \tag{5.4}$$

$$Y = B_b a + v = (b \cdot A)a + v, \ \ if \ b \ is \ known \tag{5.5}$$

where

$$\begin{aligned}
&A_a \triangleq A \otimes a = [A_1 a \ ... \ A_J a \ ... \ A_M a] \\
&A = [A_1 \ ... \ A_J \ ... \ A_M] \in R^{N \times ML} \\
&B_b \triangleq b \cdot A = b_1 A_1 + ... + b_J A_J + ... + b_M A_M \\
&A_J = \begin{bmatrix} \Psi^1_{J1} & ... & \Psi^1_{JL} \\ \vdots & \cdots & \vdots \\ \Psi^N_{J1} & ... & \Psi^N_{JL} \end{bmatrix} \\
&J = 1, ..., M
\end{aligned} \tag{5.6}$$

Note that $A_a \in R^{N \times M}$ or $B_b \in R^{N \times L}$ is a known matrix after either a or b is given. In some cases such as an LNL system, which is consisting of a dynamic linear (L), a static nonlinear (N) and a dynamic linear (L) subsystems in series, we may have other unknown parameters that are independent of the bilinear pairs a and b. LNL systems are also called Wiener-Hammerstein systems and their identification has been an active research field in these years. To include these cases, model (5.2) can be generalized to $Y = F(a, b, d) + v$ by adding another parameter vector $d \in R^{r \times 1}$ which is independent of a and b. With this, models (5.4) and (5.5) are generalized to

$$Y = \mathcal{G}d + A_a b + v, \ if \ a \ is \ known \tag{5.7}$$

$$Y = \mathcal{G}d + B_b a + v, \ if \ d \ and \ b \ \ are \ known \tag{5.8}$$

where $\mathcal{G} \in R^{N\times r}$ is a known full column rank matrix. It will be seen how an LNL nonlinear system is formulated to the forms of (5.7) and (5.8) in Section 3. Our identification objective is to propose an algorithm to iteratively estimate a, b and d in the bilinear model of (5.7) and (5.8), based on sufficiently large number of input output data pairs.

Assumption 5.1. *Each component of v is i.i.d with zero mean and finite variance $D(v_i) = \sigma_v^2$.*

Assumption 5.2. *A_{jt} is a random i.i.d variables sampled from the probability density function $p_\Phi(.)$, where A_{jt} denotes the component of the matrices A. More particularly, $E(A_{jt}) = 0$ and $D(A_{jt}) = \sigma_\Phi^2$.*

Assumption 5.3. *Matrix $[\mathcal{G}\ A] = [\mathcal{G}\ A_1\ ...\ A_M]$ is full column rank.*

Assumption 5.4. *Either $\|b\|_2$ or $\|a\|_2$ is known and the first nonzero entry of b or a is positive.*

Remark 5.1. *Note that Assumption 5.1 requires the noises to be white and Assumptions 5.2 and 5.3 are compatible. Assumption 5.2 guarantees that matrix A can be constructed as a full column rank matrix if the row numbers of A is not less than its column number. In addition, as matrices $\mathcal{G}$, A are constructed based on random input and output signals, Assumption 5.3 should be satisfied under Assumption 5.2 provided that the row number of $[\mathcal{G}\ A]$ is not less than its column number. Assumptions 5.2 and 5.3 imply that $\rho_1 I \leq \frac{1}{N}[\mathcal{G}\ A][\mathcal{G}\ A]' \leq \rho_2 I$ where ρ_1 and ρ_2 are positive numbers. Clearly, this has the same implication as that of the input/output signals being persistently exciting (PE) [72]. The verification of these assumptions is illustrated in the simulation section (Section 4) later. One can also refer to Remark 5.5 when applying the proposed algorithm to LNL systems. Assumption 5.4 is to guarantee a unique representation of the LNL nonlinear system, as any pair of κa and b/κ for some non-zero κ will give the same input-output data.*

5.2.2 Iterative Identification Algorithm

Denote the estimates of a, b and d as $\hat{a}$, $\hat{b}$ and $\hat{d}$, respectively. We first obtain $\hat{d}$ without knowing a and b in the first iteration step. Then we estimate $\hat{a}(k)$ and $\hat{b}(k)$ iteratively.

Note that (5.7) can be rewritten as

$$\begin{aligned} Y &= \mathcal{G}d + (b \cdot A)a + v \\ &= \mathcal{G}d + (A \otimes a)b + v \\ &= \mathcal{G}d + b_1 A_1 a + ... + b_M A_M a + v \\ &= \mathcal{G}d + [A_1 \ ... \ A_M] \begin{bmatrix} a_1 b \\ \vdots \\ a_M b \end{bmatrix} + v \\ &= \mathcal{G}d + A\gamma + v \end{aligned} \tag{5.9}$$

where $\gamma = \begin{bmatrix} b_1 a \\ \vdots \\ b_M a \end{bmatrix}$. We first give the estimate $\hat{d}$ as follows

$$\hat{d} = (\mathcal{G}'(I_N - A(A'A)^{-1}A')\mathcal{G})^{-1}(\mathcal{G}'(I_N - A(A'A)^{-1}A'))Y \tag{5.10}$$

where I_N is an identity matrix of dimension N. Later we will show how to derive (5.10) and obtain the consistency of $\hat{d}$ in Theorem 5.1 in the convergence analysis subsection. After $\hat{d}$ is obtained, let $\hat{b}(k)$ be the current estimates of b at the $k-$th iteration step. When $\hat{d}$ and $\hat{b}(k)$ become available, determining the estimate $\hat{a}(k)$ of a is to solve a linear equation by substituting them into (5.8). Thus, with currently obtained $\hat{b}(k)$ and $\hat{d}$, a least square optimal solution $\hat{a}_{op}(k)$ of (5.8) for

estimating a is obtained as

$$\hat{a}_{op}(k) = \mathcal{F}_1(\hat{b}(k)) = (B'_{\hat{b}(k)}B_{\hat{b}(k)})^{-1}B'_{\hat{b}(k)}(Y - \mathcal{G}\hat{d}). \tag{5.11}$$

Based on Assumption 5.4, the estimates $\hat{a}(k)$ can be obtained after normalization:

$$\hat{a}(k) = \mathcal{F}_2(\hat{a}_{op}(k)) = \frac{\hat{a}_{op}(k)\|a\|_2}{\|\hat{a}_{op}(k)\|_2} \tag{5.12}$$

Clearly, $\|\hat{a}(k)\|_2 = \|a\|_2$. After replacing a with $\hat{a}(k)$ and d with $\hat{d}$, (5.7) becomes

$$Y - \mathcal{G}\hat{d} = A_{\hat{a}(k)}b + v \tag{5.13}$$

which gives the least square estimate of b at step $k+1$

$$\hat{b}(k+1) = \mathcal{F}_3(\hat{a}(k)) = (A'_{\hat{a}(k)}A_{\hat{a}(k)})^{-1}A'_{\hat{a}(k)}(Y - \mathcal{G}\hat{d}) \tag{5.14}$$

Combining (5.11), (5.12) and (5.14), it can be obtained that

$$\hat{b}(k+1) = \mathcal{F}_3(\mathcal{F}_2(\hat{a}_{op}(k))) = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(\hat{b}(k)))) = \mathcal{F}(\hat{b}(k)) \tag{5.15}$$

Thus we could obtain the iterated function sequence $\{\hat{b}(0), \mathcal{F}(\hat{b}(0)), \mathcal{F}(\mathcal{F}(\hat{b}(0))), ...\}$. To study the convergence property of the sequence, we represent $\hat{b}$ in the form of $\hat{b} = \mathcal{F}(\hat{b})$ as follows

$$\hat{b} = \mathcal{F}(\hat{b}) = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(\hat{b}))) \tag{5.16}$$

where $\mathcal{F}(.) = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(.)))$. It will be shown that (5.16) has a unique fixed point and thus the algorithm (5.15) is regarded as finding this fixed point iteratively.

We now summarize the iterative algorithm starting with an arbitrary nonzero initial value $\hat{b}(0)$ as follows.

Step 1: Obtain $\hat{d}$ from (5.10).

Step 2: Obtain estimates $\hat{a}(k)$ by using (5.11) and (5.12) with current estimate $\hat{b}(k)$.

Step 3: Replace k by $k+1$ and obtain $\hat{b}(k+1)$ from (5.14) with estimate $\hat{a}(k)$.

Step 4: If a stopping criterion is satisfied, then let $\hat{a} = \hat{a}(k) \cdot sgn(\hat{a}_1(k))$ and $\hat{b} = \hat{b}(k+1) \cdot sgn(\hat{a}_1(k))$, and end. Otherwise, go to Step 2. Finally, the obtained estimates are denoted as $\hat{a}$, $\hat{b}$ and $\hat{d}$.

Remark 5.2. *Actually $\hat{a}(k)$ and $\hat{b}(k)$ can be permuted at Steps 1-4. Also $\hat{a} = \hat{a}(k) \cdot sgn(\hat{a}_1(k))$ is to guarantee that the first element of $\hat{a}$ remains positive. As long as the initial value $\hat{b}(0)$ is nonzero, $B'_{\hat{b}(k)} B_{\hat{b}(k)}$ will be a full column rank matrix and then $\hat{a}(k)$ is ensured nonzero for all $k \geq 0$ under Assumptions 5.1-5.4. This is explained below. It is noted that $\hat{a}_{op}(k)$ is a zero vector if and only if $Y - \mathcal{G}\hat{d}$ is a zero vector, due to the full column rank of $B'_{\hat{b}(k)} B_{\hat{b}(k)}$. However, $Y - \mathcal{G}\hat{d}$ is a zero vector implies that the output of the bilinear system can be represented by $\mathcal{G}\hat{d}$ at any sample point of time. This is not possible for bilinear systems with random inputs. So the iteration only requires a nonzero initial condition of $\hat{a}(0)$ and the convergence property will be analyzed in the next convergence analysis subsection. In practice, a stopping criterion can be set as $|\hat{b}(k+1) - \hat{b}(k)| < \epsilon$ ($\epsilon = 10^{-5}$).*

5.2.3 Convergence Analysis

We now establish the convergence properties of the proposed iterative algorithm. To achieve this, some required preliminaries are presented first.

Lemma 5.1. *For matrix $\mathcal{K}_N \in R^{N \times M}$ with its component $\mathcal{K}_{jt}$ denoting a random i.i.d variable sampled from a probability density function $p_{\mathcal{K}}(.)$, we have $\lim_{N \to \infty} tr((\mathcal{K}'_N \mathcal{K}_N)^{-1}) = 0$ almost surely where $tr(.)$ denotes the matrix trace.*

Proof. Note that $\mathcal{K}_N \in R^{N\times M}$ is full column rank matrix as long as $N \geq M$. Note that $\mathcal{K}_N^{'}\mathcal{K}_N$ is a symmetrical positive definite matrix as $\mathcal{K}_N$ is full column rank. Let g_N' be the $N-$th row vector of $\mathcal{K}_N$, then

$$\mathcal{K}_{N+1}'\mathcal{K}_{N+1} = \begin{bmatrix} \mathcal{K}_N' & g_{N+1} \end{bmatrix} \begin{bmatrix} \mathcal{K}_N \\ g_{N+1}' \end{bmatrix} = \mathcal{K}_N'\mathcal{K}_N + g_{N+1}g_{N+1}' \tag{5.17}$$

Let $\lambda_i(N)$, $i = 1, ..., M$, denote the eigenvalues of $\mathcal{K}_N$ and assume that $\lambda_1(N) \geq ... \geq \lambda_M(N) > 0$. There exists a matrix P_N such that

$$P_N'\mathcal{K}_N P_N = diag[\lambda_1(N) \ ... \ \lambda_M(N)] \tag{5.18}$$

Let $\alpha = [1 \ 0 \ ... \ 0]'$, then

$$\alpha' P_N'\mathcal{K}_N'\mathcal{K}_N P_N\alpha' = (P_N\alpha)'\mathcal{K}_N'\mathcal{K}_N(P_N\alpha) = \lambda_1(N) \tag{5.19}$$

Following the similar procedure of Lemma 2.4, we have

$$\lambda_1(N+1) = \lambda_1(N) + ((P_N\alpha)'g_{N+1})^2 \tag{5.20}$$

Let $\lambda_1^*(N) = \lambda_1(N+1) - \lambda_1(N) = ((P_N\alpha)'g_{N+1})^2$. As $\mathcal{K}_{kj}$ is i.i.d sampled from a probability density function $p_{\mathcal{K}}(.)$ and thus components in the column vector g_{N+1} are i.i.d. Therefore, $\lambda_1^*(N) = ((P_N\alpha)'g_{N+1})^2$ is positive almost surely. There exists a constant c such that the probability $p(\lambda_1^*(N) > c)$ is nonzero for all N. Thus, we obtain $\lim_{N\to\infty}\lambda_1(N) \to \infty$ almost surely. Similarly, $\lim_{N\to\infty}\lambda_i(N) \to \infty$, for $i = 2, ..., M$ almost surely. Therefore,

$$\lim_{N\to\infty} tr((\mathcal{K}_N'\mathcal{K}_N)^{-1}) = \sum_{i=1}^{M}\frac{1}{\lambda_i(N)} = 0 \tag{5.21}$$

almost surely. Finally, we have $tr((\mathcal{K}_N^{'}\mathcal{K}_N)^{-1})$ approaches zero almost surely when its dimension tends to infinity. □

Denote $f(\hat{a}) = \lim_{N\to\infty} \|(A_{\hat{a}}'A_{\hat{a}})^{-1}A_{\hat{a}}\|_2$ where $\hat{a} \neq 0$ and $A_a \in R^{N\times M}$. Also $\|A_{\hat{a}}\|_2$ is defined as the square root of the maximum eigenvalue of $A_{\hat{a}}'A_{\hat{a}}$, i.e, $\|A_{\hat{a}}\|_2 = \sqrt{\lambda_{max}(A_{\hat{a}}'A_{\hat{a}})}$.

Lemma 5.2. *Under Assumption 5.2, the magnitude of the directional derivative of $f(\hat{a})$ along a direction vector* **u** *attains its maximum when* **u** *is in the same direction as $\hat{a}$.*

Proof. Under Assumption 5.2 and the strong law of large numbers, we have

$$\lim_{N\to\infty} \tfrac{1}{N}A_{\hat{a}}'A_{\hat{a}} = \hat{a}'\hat{a}\sigma_\Phi^2 I \tag{5.22}$$

almost surely where I is an identity matrix and σ_Φ^2 denotes the variance of A_{jt} as shown in Assumption 5.2. Then we obtain

$$\begin{aligned}\sqrt{N}f(\hat{a}) &= \sqrt{N}\cdot\|(A_{\hat{a}}'A_{\hat{a}})^{-1}A_{\hat{a}}'\|_2 \\ &= \sqrt{N}\cdot\tfrac{1}{N\hat{a}'\hat{a}\sigma_\Phi^2}\cdot\sqrt{N}\cdot\sqrt{\lambda_{max}(\tfrac{1}{N}A_{\hat{a}}'A_{\hat{a}})} \\ &= \tfrac{1}{\hat{a}'\hat{a}\sigma_\Phi^2}\sqrt{\hat{a}'\hat{a}\sigma_\Phi^2} \\ &= \tfrac{1}{\|\hat{a}\|_2}\tfrac{1}{\sqrt{\sigma_\Phi^2}}\end{aligned} \tag{5.23}$$

Note that N is a constant and then the contour planes of $f(\hat{a})$ are concentric spheres. This means that the gradient of $f(\hat{a})$ is in the radial direction. It is known that the maximum magnitude of the directional derivative $D_{\mathbf{u}}f(\hat{a})$, i.e. $\|\nabla f\|_2$, occurs when the vector **u** is aligned with its gradient $\nabla f(\hat{a})$, which is the radial direction $\hat{a}$. □

Lemma 5.3. *For conformable matrices A, B and C, $tr(ABC) = tr(BCA) = tr(CAB)$. Also, if matrices A and B are addable, $tr(A+B) = tr(A) + tr(B)$.*

Proof. To prove $tr(ABC) = tr(BCA) = tr(CAB)$, actually we only need to prove that $tr(AB) = tr(BA)$. As $tr(AB) = \sum_i \sum_j A_{ij} B_{ji} = tr(BA)$ and $tr(A + B) = \sum_i (A_{ii} + B_{ii}) = \sum_i (A_{ii}) + \sum (B_{ii}) = tr(A) + tr(B)$, then this Lemma holds. □

Theorem 5.1. *Under Assumptions 5.1-5.3, we have $\hat{d}$ given in (5.10) satisfies that $\lim_{N\to\infty} \hat{d} = d$ almost surely.*

Proof. From Assumption 5.3, $[\mathcal{G}\ A]$ is full column rank. Then

$$\begin{bmatrix} \mathcal{G}' \\ A' \end{bmatrix} [\mathcal{G}\ A] = \begin{bmatrix} \mathcal{G}'\mathcal{G} & \mathcal{G}'A \\ A'\mathcal{G} & A'A \end{bmatrix} \tag{5.24}$$

is positive definite and its inverse exists. We obtain the least square estimates of $\begin{bmatrix} \hat{d} \\ \hat{\gamma} \end{bmatrix}$ in (5.9) as follows

$$\begin{bmatrix} \hat{d} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \mathcal{G}'\mathcal{G} & \mathcal{G}'A \\ A'\mathcal{G} & A'A \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{G}' \\ A' \end{bmatrix} Y \tag{5.25}$$

Let

$$\begin{bmatrix} \mathcal{G}'\mathcal{G} & \mathcal{G}'A \\ A'\mathcal{G} & A'A \end{bmatrix}^{-1} = \begin{bmatrix} X_1 & X_2 \\ X_2' & X_3 \end{bmatrix} \tag{5.26}$$

then

$$\begin{aligned} \mathcal{G}'\mathcal{G}X_1 + \mathcal{G}'AX_2' &= I_{\mathcal{G}} \\ \mathcal{G}'\mathcal{G}X_2 + \mathcal{G}'AX_3' &= 0 \\ A'\mathcal{G}X_1 + A'AX_2' &= 0 \\ A'\mathcal{G}X_2 + A'AX_3 &= I_A \end{aligned} \tag{5.27}$$

where $I_{\mathcal{G}}$ and I_A denote the identity matrices with the same dimension of $\mathcal{G}'\mathcal{G}$ and

$A'A$, respectively. We get

$$\begin{aligned} X_1 &= (\mathcal{G}^{'}(I_N - A(A'A)^{-1}A')\mathcal{G})^{-1} \\ X_2 &= -(\mathcal{G}^{'}(I_N - A(A'A)^{-1}A')\mathcal{G})^{-1}\mathcal{G}'A(A'A)^{-1} \\ X_3 &= (A^{'}(I_N - \mathcal{G}(\mathcal{G}'\mathcal{G})^{-1}\mathcal{G}')A)^{-1} \end{aligned} \tag{5.28}$$

where I_N is an identity matrix with dimension $N \times N$. Let

$$\mathcal{A}_N = I_N - A(A'A)^{-1}A' \tag{5.29}$$

Note that $\mathcal{A}_N^2 = \mathcal{A}_N$ as $\mathcal{A}_N$ is a projection matrix. From (5.25), (5.26) and (5.28), we have

$$\begin{aligned} \hat{d} &= [X_1 \ X_2] \begin{bmatrix} \mathcal{G}' \\ A' \end{bmatrix} Y \\ &= (\mathcal{G}'(I_N - A(A'A)^{-1}A')\mathcal{G})^{-1}\mathcal{G}'(I_N - A(A'A)^{-1}A')Y \\ &= (\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}\mathcal{G}'\mathcal{A}_N Y \end{aligned} \tag{5.30}$$

which is same as (5.10). Now we prove the consistency of $\hat{d}$. Substituting $Y = \mathcal{G}d + A\gamma + v$ in (5.9) to (5.30) gives

$$\hat{d} = (\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}\mathcal{G}'\mathcal{A}_N(\mathcal{G}d + A\gamma + v) \tag{5.31}$$

By observing (5.31) carefully, it can be noted that

$$\begin{aligned} &(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N)\mathcal{G} = I_{\mathcal{G}} \\ &\mathcal{G}'\mathcal{A}_N A = \mathcal{G}'(I_N - A(A'A)^{-1}A')A = \mathcal{G}'(A - A) = 0 \end{aligned} \tag{5.32}$$

Then,

$$\hat{d} = d + (\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}\mathcal{G}'\mathcal{A}_N v \tag{5.33}$$

and

$$\begin{aligned} & E((\hat{d}-d)'(\hat{d}-d)) \\ & = \mathcal{A}_N'\mathcal{G}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}\mathcal{G}'\mathcal{A}_N\sigma_v^2 \end{aligned} \tag{5.34}$$

So we have $\sum_i E((\hat{d}_i-d_i)'(\hat{d}_i-d_i)) = tr((\mathcal{G}'\mathcal{A}_N)'(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N))\sigma_v^2$ where d_i denotes the element of vector d. Based on Lemma 5.3 and using the property $\mathcal{A}_N^2 = \mathcal{A}_N$ of the projection matrix $\mathcal{A}_N$, we obtain

$$\begin{aligned} & \textstyle\sum_i E((\hat{d}_i-d_i)'(\hat{d}_i-d_i)) \\ & = tr(\mathcal{A}_N'\mathcal{G}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}\mathcal{G}'\mathcal{A}_N)\sigma_v^2 \\ & = tr(\mathcal{G}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}\mathcal{G}'\mathcal{A}_N^2)\sigma_v^2 \\ & = tr(\mathcal{G}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}\mathcal{G}'\mathcal{A}_N)\sigma_v^2 \\ & = tr(\mathcal{G}'\mathcal{A}_N\mathcal{G}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}(\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1})\sigma_v^2 \\ & = tr((\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1})\sigma_v^2 \\ & = tr((\mathcal{G}'(I_N - A(A'A)^{-1}A')\mathcal{G})^{-1})\sigma_v^2 \end{aligned} \tag{5.35}$$

Also, as $\mathcal{A}_N = I_N - A(A'A)^{-1}A'$ is a projection matrix, there exists a positive number $\kappa_1 \geq 1$ such that $0 \leq \mathcal{A}_N \leq \kappa_1 I_N$. This implies that $\mathcal{A}_N$ and $\kappa I_N - \mathcal{A}_N$ are semi-positive definite matrices. Then we have

$$\tfrac{1}{\kappa_1} \cdot tr((\mathcal{G}'\mathcal{G})^{-1}) \leq tr((\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}) \leq tr((\mathcal{G}'\mathcal{G})^{-1}) \tag{5.36}$$

So there exists a κ_2 satisfying $\frac{1}{\kappa_1} < \kappa_2 < 1$ such that $tr((\mathcal{G}'\mathcal{A}_N\mathcal{G})^{-1}) = \kappa_2 \cdot tr((\mathcal{G}'\mathcal{G})^{-1})$. We then obtain

$$\lim_{N\to\infty} \textstyle\sum_i E((\hat{d}_i-d_i)'(\hat{d}_i-d_i)) = \lim_{N\to\infty} \kappa_2 \cdot tr((\mathcal{G}'\mathcal{G})^{-1})\sigma_v^2 = 0 \tag{5.37}$$

almost surely based on Lemma 5.1. Therefore, we have $\lim_{N\to\infty} \hat{d} = d$ almost surely and the Theorem holds. □

Lemma 5.4. The Contraction Mapping Theorem *[71] [73]. Let (X, D) be a nonempty complete metric space where D is a metric on X. Let $\mathcal{F} : X \to X$ be a contraction mapping on X, i.e., there is a nonnegative real number $Q < 1$ such that $D(\mathcal{F}(x), \mathcal{F}(y)) \leq Q \cdot D(x, y)$, for all $x, y \in X$. Then the map $\mathcal{F}$ admits one and only one fixed point $x^* \in X$ which means $x^* - \mathcal{F}(x^*) = 0$. Furthermore, this fixed point can be found as follows: start with an arbitrary element $x(0)$ in X and define an iterative sequence by $x(k+1) = \mathcal{F}(x(k))$ for $k = 1, 2, ...$ This sequence converges to x^*.*

This Lemma can be seen on page 267 in [71] and Definition 1.1 in [73].

Now we define $X_a = \{\hat{a} \mid \|\hat{a}\|_2 = \|a\|_2, \hat{a}_1 > 0\}$, $X_b = \{\hat{b} \mid \|\hat{b}\|_2 \leq \|b\|_2\}$, $D(x, y) = \|x - y\|_2$.

Theorem 5.2. *Under Assumptions 5.1-5.4, $\mathcal{F}(\hat{b}) : X_b \to X_b$ defined in (5.16) is a contraction mapping on X_b when $N \to \infty$. Thus equation (5.16) has a unique fixed point on X_b which corresponds to the true parameter b.*

Proof. Firstly, we prove that $\mathcal{F}(\hat{b}) : X_b \to X_b$ as $N \to \infty$. Secondly, we show that $\mathcal{F}(\hat{b})$ is a contraction mapping on X_b and finally the true parameters corresponds to the unique fixed point of $\hat{b} = \mathcal{F}(\hat{b})$.

From Assumption 5.3, for any nonzero $\hat{a} \in X_a$, $A'_{\hat{a}} A_{\hat{a}}$ has an inverse. Then $\hat{b} = \mathcal{F}_3(\hat{a}) = (A'_{\hat{a}} A_{\hat{a}})^{-1} A'_{\hat{a}} (Y - \mathcal{G}\hat{d})$. Note that $Y = \mathcal{G}d + A_a b + v$. From Theorem 5.1, we get $\lim_{N\to\infty} \hat{d} = d$ almost surely under Assumptions 5.1-5.3 and thus

$$\begin{aligned} \lim_{N\to\infty} \hat{b} &= \lim_{N\to\infty} \mathcal{F}_3(\hat{a}) \\ &= \lim_{N\to\infty} (A'_{\hat{a}} A_{\hat{a}})^{-1} A'_{\hat{a}} (A_a b + \mathcal{G}(d - \hat{d}) + v) \\ &= \lim_{N\to\infty} (A'_{\hat{a}} A_{\hat{a}})^{-1} A'_{\hat{a}} (A_a b + v) \\ &= \lim_{N\to\infty} (A'_{\hat{a}} A_{\hat{a}})^{-1} A'_{\hat{a}} A_a b + (A'_{\hat{a}} A_{\hat{a}})^{-1} A'_{\hat{a}} v \end{aligned} \tag{5.38}$$

Based on Assumption 5.1 and Lemma 5.1, we have

$$\lim_{N\to\infty} \frac{1}{N}\|(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}v\|_2 \leq \lim_{N\to\infty} \frac{1}{\sqrt{N}}\frac{1}{\|\hat{a}\|_2}\frac{1}{\sqrt{\sigma_\Phi^2}}\|v\|_2 = 0 \tag{5.39}$$

almost surely. This yields

$$\begin{aligned}\lim_{N\to\infty} \|\hat{b}\|_2 &= \lim_{N\to\infty} \|\mathcal{F}_3(\hat{a})\|_2 \\ &= \lim_{N\to\infty} \|(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_a b\|_2 \\ &\leq \lim_{N\to\infty} \frac{\|A_{\hat{a}}A_a\|_2}{\|A_{\hat{a}}A_{\hat{a}}\|_2}\|b\|_2\end{aligned} \tag{5.40}$$

From Assumption 5.4, if $\hat{a}$ belongs to X_a, then $-\hat{a}$ does not belong to X_a based on the definition of X_a. And from (5.22), we have

$$\begin{aligned}\lim_{N\to\infty} \frac{\|A_{\hat{a}}A_a\|_2}{\|A_{\hat{a}}A_{\hat{a}}\|_2} &= \lim_{N\to\infty} \frac{\|\frac{1}{N}A_{\hat{a}}A_a\|_2}{\|\frac{1}{N}A_{\hat{a}}A_{\hat{a}}\|_2} = \frac{\|\hat{a}'a\|_2\sigma_\Phi^2}{\|\hat{a}'\hat{a}\|_2\sigma_\Phi^2} = \frac{\|\hat{a}'a\|_2}{\|\hat{a}'\hat{a}\|_2} \\ &= \begin{cases} 1 & \text{if and only if } \hat{a} = a \\ \mathcal{F}_b\ ,\ 0 < \mathcal{F}_b < 1 & \text{otherwise} \end{cases}\end{aligned} \tag{5.41}$$

This gives $\|\hat{b}\|_2 = \|\mathcal{F}(\hat{b})\|_2 = \|\mathcal{F}_3(\hat{a})\|_2 \leq \|b\|_2$. Therefore $\mathcal{F}(\hat{b}) \in X_b$ and $\mathcal{F}(\hat{b}) : X_b \to X_b$.

Now we prove that $\mathcal{F}(\hat{b}) : X_b \to X_b$ is a contraction mapping on X_b as $N \to \infty$. It can be seen that (5.38) and (5.39) give

$$\hat{b} = (A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}(Y - \mathcal{G}\hat{d}) = (A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_a b \tag{5.42}$$

as $N \to \infty$. Multiplying $A_{\hat{a}}$ on both sides of (5.42), we get

$$A_{\hat{a}}\hat{b} = A_{\hat{a}}(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_a b \tag{5.43}$$

That is

$$\begin{aligned}\hat{b}'A'_{\hat{a}}A_{\hat{a}}\hat{b} &= b'A'_aA_{\hat{a}}(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_{\hat{a}}(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_ab \\ &= b'A'_aA_{\hat{a}}(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_ab\end{aligned} \tag{5.44}$$

Note that $A_{\hat{a}}(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}$ is a projection matrix which projects a vector onto the space spanned by $A_1, ..., A_m$. Then we have $A_{\hat{a}}(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_ab = A_ab$. So we have

$$\begin{aligned}&\hat{b}'A'_{\hat{a}}A_{\hat{a}}\hat{b} \\ &= b'A'_aA_{\hat{a}}(A'_{\hat{a}}A_{\hat{a}})^{-1}A'_{\hat{a}}A_ab \\ &= b'A'_aA_ab\end{aligned} \tag{5.45}$$

As $\lim_{N\to\infty}\frac{1}{N}A'_{\hat{a}}A_{\hat{a}} = \hat{a}'\hat{a}\sigma^2_\Phi I$ and $\lim_{N\to\infty}\frac{1}{N}A'_aA_a = a'a\sigma^2_\Phi I$, we obtain

$$\|\hat{b}\|_2\|\hat{a}\|_2\sigma^2_\Phi = \|b\|_2\|a\|_2\sigma^2_\Phi \tag{5.46}$$

Similarly, (5.11) and (5.39) also give

$$\hat{a}_{op} = (B'_{\hat{b}}B_{\hat{b}})^{-1}B'_{\hat{b}}B_ba \tag{5.47}$$

Multiplying $B_{\hat{b}}$ on both sides of (5.47) gives

$$\|\hat{b}\|_2\|\hat{a}_{op}\|_2\sigma^2_\Phi = \|b\|_2\|a\|_2\sigma^2_\Phi \tag{5.48}$$

Thus under Assumption 5.4, we have

$$\|\hat{a}_{op}\|_2 = \|a\|_2,\ \|\hat{b}\|_2 = \|b\|_2 \tag{5.49}$$

when combining (5.46) and (5.48). Let $Q = \|\frac{d\mathcal{F}(\hat{b})}{d\hat{b}}\|_2$ be the magnitude of the

derivative of $\mathcal{F}(\hat{b})$ with respect to $\hat{b}$. Then, (5.15) becomes

$$\hat{b}(k+1) \;\; = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1((\hat{b}(k))))) \tag{5.50}$$

and then

$$\begin{aligned} Q \;\; &= \| \tfrac{d\mathcal{F}}{d\hat{a}} \cdot \tfrac{d\hat{a}}{d\hat{a}_{op}} \cdot \tfrac{d\hat{a}_{op}}{d\hat{b}} \|_2 \\ &= \| \tfrac{d\mathcal{F}_3}{d\hat{a}} \|_2 \cdot \| \tfrac{d\mathcal{F}_2}{d\hat{a}_{op}} \|_2 \cdot \| \tfrac{d\mathcal{F}_1}{d\hat{b}} \|_2 \end{aligned} \tag{5.51}$$

For a nonzero $\hat{a}$, based on Lemma 5.2, we know that the magnitude of the directional derivative $D_u\mathcal{F}_3(.)$ with respect to $\hat{a}$ attains its maximum when u is in the same direction as $\hat{a}$, i.e.,

$$\begin{aligned} \| \triangledown \mathcal{F}_3(.) \|_2 \;\; &= \| \tfrac{d\mathcal{F}_3}{d\hat{a}} \|_2 \\ &= \lim_{\|\Delta a\|_2 \to 0} \tfrac{\|\mathcal{F}_3(\hat{a}+\Delta a) - \mathcal{F}_3(\hat{a})\|_2}{\|\Delta a\|_2} \\ &= \tfrac{\|A'_a A_a\|_2}{\|A'_{\hat{a}} \;\; A_{\hat{a}}\|_2} \cdot \|b\|_2 \\ &= \tfrac{\|\vec{a}' a\|_2}{\|\hat{a}'\hat{a}\|_2} \|b\|_2 \end{aligned} \tag{5.52}$$

where $\vec{a} = \frac{\Delta a}{\|\Delta a\|_2}$ is a unit vector along the direction of $\hat{a}$. We also have

$$\| \tfrac{d\hat{a}}{d\hat{a}_{op}} \|_2 = \| \tfrac{d\mathcal{F}_2}{d\hat{a}_{op}} \|_2 = \tfrac{\|a\|_2}{\|\hat{a}_{op}\|_2} = 1 \tag{5.53}$$

as $\|a\|_2 = \|a_{op}\|_2$ in (5.49). Similarly to $\| \frac{d\mathcal{F}_3}{d\hat{a}} \|_2$, we obtain

$$\| \tfrac{d\hat{a}_{op}}{d\hat{b}} \|_2 = \| \tfrac{d\mathcal{F}_1}{d\hat{b}} \|_2 = \tfrac{B'_{\hat{b}} B_b}{B'_b B_{\hat{b}}} \|a\|_2 = \tfrac{\|\vec{b}' b\|_2}{\|\hat{b}'\hat{b}\|_2} \|a\|_2 \tag{5.54}$$

where $\vec{b}$ is a unit vector along the direction of $\hat{b}$. Combining (5.51)-(5.54), and using $\|\vec{a}' a\|_2 \cdot \|a\|_2 = \|\hat{a}' a\|_2$ and $\|\vec{b}' b\|_2 \cdot \|b\|_2 = \|\hat{b}' b\|_2$, we get

$$Q = \tfrac{\|\vec{a}' a\|_2}{\|\hat{a}'\hat{a}\|_2} \|b\|_2 \cdot \tfrac{\|a\|_2}{\|\hat{a}_{op}\|_2} \cdot \tfrac{\|\vec{b}' b\|_2}{\|\hat{b}'\hat{b}\|_2} \|a\|_2 = \tfrac{\|\hat{a}' a\|_2}{\|\hat{a}'\hat{a}\|_2} \cdot \tfrac{\|\hat{b}' b\|_2}{\|\hat{b}'\hat{b}\|_2} < 1 \tag{5.55}$$

as long as $\hat{a} \neq a$, or $\hat{b} \neq b$. So we have $\forall x, y \in X_b$, $d(\mathcal{F}(x), \mathcal{F}(y) \leq Q \cdot D(x, y)$. Finally based on Lemma 5.4, $\hat{b} = \mathcal{F}(\hat{b})$ has a unique fixed point which is the true parameter $b \in X_b$. □

Corollary 5.1. *If noise $v = 0$ and matrix A in (5.6) is a matrix satisfying that $\frac{1}{N}A'A = \sigma_\Phi^2 I$, Theorem 5.2 holds for a finite N.*

Proof. Firstly, from (5.33) in the proof Theorem 5.1, we could have $\hat{d} = d$ for a finite N since $v = 0$. Secondly, as $\frac{1}{N}A'A = \sigma_\Phi^2 I$ which means $\sqrt{\frac{1}{N\sigma_\Phi^2}}A$ is an orthonormal matrix, it can be obtained that $\frac{\|A'_{\hat{a}}A_a\|_2}{\|A'_{\hat{a}}A_{\hat{a}}\|_2} = \frac{\|\hat{a}'a\|_2}{\|\hat{a}'\hat{a}\|_2}$ for a finite N instead of (5.41) where $N \to \infty$ is required. Thus, we have this Corollary holds. □

Remark 5.3. *In real applications like the identification of LNL systems considered in the next section, the effects of noise v can be made small enough and matrix $\sqrt{\frac{1}{N\sigma_\mathcal{K}^2}}\mathcal{K}$ ($\mathcal{K}$ is A in the general bilinear model) can be constructed as an orthonormal matrix approximately based on the input output data and the chosen kernel for a finite N. Then, with a properly chosen finite N, good performances can be achieved, as illustrated in the application examples later.*

5.3 Identification of LNL Wiener-Hammerstein Models

In this section, it is shown that an LNL Wiener-Hammerstein cascade system described below can be represented by a bilinear model

$$x_i = \mu_0 u_i + ... + \mu_m u_{i-m} \tag{5.56}$$

$$z_i = f(x_i) \tag{5.57}$$

$$y_i = \omega_1 y_{i-1} + ... + \omega_n y_{i-n} + b_0 z_i + ... + b_s z_{i-s} + v_i \tag{5.58}$$

where $\{u_i\}$ and $\{y_i\}$ are the input and output sequences, respectively, and v_i is the observation noise, and m, n and s are the orders of the system. The identification objective is to estimate the parameters μ_0 ... μ_m, ω_1 ... ω_n and b_0 ... b_s as well as the nonlinear static function $f(.)$ based on the available input-output data points $\{u_i, y_i\}_{i=r}^N$ where $r = \max(n, m, s) + 1$ and $i = r, ..., N$. Let

$$\begin{aligned} Y^r &= [y_r \ ... \ y_N]' \\ v^r &= [v_r \ ... \ v_N]' \end{aligned} \tag{5.59}$$

where v_r is the noise vector. Then (5.58) can be expressed as $Y^r = \Gamma\theta + v^r$ where

$$\begin{aligned} \theta &= [\omega' \ b'] = [\omega_1 \ ... \ \omega_n \ b_0 \ ... \ b_s]' \\ \omega &= [\omega_1 \ ... \ \omega_n]' \\ b &= [b_0 \ ... \ b_s]' \\ \Gamma &= \begin{bmatrix} y_{r-1} & \cdots & y_{r-n} & f(x_r) & \cdots & f(x_{r-s}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{N-1} & \cdots & y_{N-n} & f(x_N) & \cdots & f(x_{N-s}) \end{bmatrix}. \end{aligned} \tag{5.60}$$

To guarantee that the input output signals are persistently exciting (PE), we have the following assumptions.

Assumption 5.5. *Input $u_i \in U(-1, 1)$ where $U(-1, 1)$ denotes a uniform distribution in the interval $[-1, 1]$, and $\sum_{j=0}^m |\mu_j| \leq C$ with $\mu_0 = 1$ where $C > 1$ is a constant.*

Assumption 5.6. *The nonlinear assumption $f(.)$ is an invertible function which satisfies that Γ is a full column rank under the condition that input and output signals are persistently exciting (PE). The inverse function is denoted as $f^{-1}(.)$ on $[-C, C]$.*

Remark 5.4. *Note that Assumptions 5.5 guarantees that the definition domain of x of $z = f(x)$ is on the interval $[-C, C]$. Assumption 5.6 refers to the identifiability of the whole LNL system.*

In [46], it is noted that a nonlinear function can be uniformly approximated by increasing the number of randomly produced basis functions. So in next subsection, we will discuss the kernel machine for function approximation, which plays an important role in transforming an LNL Wiener-Hammerstein system.

5.3.1 Kernel Machine for Function Approximation

Here we use the kernel machine introduced in Subsection 2.3.1 in Chapter 2 to represent the nonlinear function $z_i = f(x_i)$ in (5.57). With a regression based on kernel machine approximation, static function $f(x)$ at x_i can be represented as

$$z_i = f(x_i) + v_i = \sum_{j=1}^{m_{sv}} a_j \bar{k}(x_i, \tilde{x}_j) + c_0 + \xi_i \tag{5.61}$$

where a_j, $j = 0, ..., m_{sv}$, is a weight to be determined from the training set, c_0 is the constant part, and ξ_i is the function approximation error at x_i. Note that $\bar{k}(x_i, \tilde{x}_j) = k(x_i, \tilde{x}_j) - \bar{k}$ where $\bar{k} = E(\bar{k}(., \tilde{x}_j))$. Clearly, $E(\bar{k}(x_i, \tilde{x}_j)) = 0$. To determine the weights $\{a_i\}_{i=0}^{m_{sv}}$, we express (5.61) in the matrix equation form as

$$\{z_i\}_{i=1}^{N} = Ka + c_0 + \xi \tag{5.62}$$

where $\xi = [\xi_1, ..., \xi_N]'$, $a = [a_1, ..., a_{m_{sv}}]'$ and

$$K = \begin{bmatrix} \bar{k}(x_1, \tilde{x}_1) & ... & \bar{k}(x_1, \tilde{x}_{m_{sv}}) \\ \vdots & ... & \vdots \\ \bar{k}(x_N, \tilde{x}_1) & ... & \bar{k}(x_N, \tilde{u}_{m_{sv}}) \end{bmatrix} = [\bar{k}(x_i, \tilde{x}_j)]_{i=1,j=1}^{i=N,j=m_{sv}}. \tag{5.63}$$

Note that as long as $i \neq i'$ and $j \neq j'$, x_i, $x_{i'}$, $\tilde{x}_j$ and $\tilde{x}_{j'}$ are i.i.d and we have the following Lemma.

Lemma 5.5. *If $i \neq i'$ or $j \neq j'$, $\bar{k}(x_i, \tilde{x}_j)$ and $\bar{k}(x_{i'}, \tilde{x}_{j'})$ are i.i.d.*

Proof. Note that when the random variables are joint Gaussian distribution, factorisation of expectations is necessary and sufficient for independence of random variables.

We first prove the case that $i \neq i'$ but $j = j'$. Since x_i and $\tilde{x}_j$ are uniformly sampled from $[-C, C]$, x_i and $x_{i'}$ are i.i.d. Then, $p_{xx}(x_i x_{i'}) = p_x(x_i)p_x(x_{i'})$. Note that $\bar{k}(x_i, \tilde{x}_j)$ and $\bar{k}(x_{i'}, \tilde{x}_j)$ are joint Gaussian distribution. In addition, we have

$$\begin{aligned} & E(\bar{k}(x_i, \tilde{x}_j)\bar{k}(x_{i'}, \tilde{x}_j)) \\ &= \int\int p_{xx}(x_i, x_{i'})\bar{k}(x_i, \tilde{x}_j)\bar{k}(x_{i'}, \tilde{x}_j)dx_i dx_{i'} \\ &= \int\int p_x(x_i)p_x(x_{i'})\bar{k}(x_i, \tilde{x}_j)\bar{k}(x_{i'}, \tilde{x}_j)dx_i dx_{i'} \\ &= \int p_x(x_i)\bar{k}(x_i, \tilde{x}_j)dx_i \int p_x(x_{i'})\bar{k}(x_{i'}, \tilde{x}_j)dx_{i'} \\ &= E(\bar{k}(x_i, \tilde{x}_j))E(\bar{k}(x_{i'}, \tilde{x}_j)) \end{aligned} \tag{5.64}$$

Thus, $\bar{k}(x_i, \tilde{x}_j)$ and $\bar{k}(x_{i'}, \tilde{x}_j)$ are i.i.d. Similarly, when $i = i'$ but $j \neq j'$, the same result can be obtained. So as long as $i \neq i'$ or $j \neq j'$, $\bar{k}(x_i, \tilde{x}_j)$ and $\bar{k}(x_{i'}, \tilde{x}_{j'})$ are i.i.d. □

5.3.2 Model Transformation

In this subsection, it will be shown that the NL subsystem,which is the model in (5.57) and (5.58), and the LN subsystem, which is the model in (5.56) and (5.57), of the LNL system (5.56)-(5.58) can be transformed to a bilinear model and a linear model, respectively.

Transforming NL subsystem to a bilinear model

Let $U_i = [u_i, ..., u_{i-m}]'$. From Assumption 5.6, $z_i = f(x_i) = f(u_i, ..., u_{i-m}) = f(U_i)$. Substituting $z_i = f(U_i)$ into (5.58), one could have

$$y_i = \omega_1 y_{i-1} + ... + \omega_n y_{i-n} + b_0 f(U_i) + ... + b_s f(U_{i-s}) + v_i \tag{5.65}$$

Now we use the kernel machine to approximate function $f(U_i)$ based on Subsection 2.3.1. Similar to (5.62), we have the matrix expression

$$\{f(U_i)\}_{i=1}^{N} = K_U a + c_0 + \xi \tag{5.66}$$

where

$$K_U = [\bar{k}(U_i, \tilde{U}_j)]_{i=1,j=1}^{i=N,j=m_{sv}}. \tag{5.67}$$

Note that the definitions of U_i and $\tilde{U}_j$ are the same as that of u_i and $\tilde{u}_j$. Obviously, $\{f(U_i)\}_{i=r}^{N} = K_0 a + c_0 + \{\xi_i\}_{i=r}^{N}$, ... , and $\{f(U_{i-s})\}_{i=r}^{N} = K_s a + d_0 + \{\xi_i\}_{i=r-s}^{N-s}$ where $K_J = \{K_U\}_{(r-J \longrightarrow N-J)}, J = 0, 1, \ldots, s$. The symbol $r - J \rightarrow N - J$ means that K_J is a sub-matrix of K_U with its rows from $r - J$ to $N - J$. For example, K_0 is a sub-matrix of K_U including its r-th row until its N-th row. The dimension of each K_J is $(N - r + 1) \times m_{sv}$. Thus, we obtain the matrix form of equations (5.56)-(5.58) as follows:

$$\begin{aligned} Y^r &= \mathcal{G}d + b_0 K_0 a + ... + b_s K_s a + \xi^r + v^r \\ &= \mathcal{G}d + (b \cdot \mathcal{K})a + \xi^r + v^r \\ &= \mathcal{G}d + (\mathcal{K} \otimes a)b + \xi^r + v^r \end{aligned} \tag{5.68}$$

where

$$\begin{aligned}
\mathcal{G} &= \begin{bmatrix} 1 & y_{r-1} & \dots & y_{r-n} \\ \vdots & \vdots & \dots & \vdots \\ 1 & y_{N-1} & \dots & y_{N-n} \end{bmatrix}, \quad d = \begin{bmatrix} c_0 \\ \omega \end{bmatrix} \\
\mathcal{K} &= [K_0 \ \dots \ K_s], \quad b \cdot \mathcal{K} = b_0 K_0 + \dots + b_m K_s \\
\mathcal{K} \otimes a &= [K_0 a \ \dots \ K_s a] \\
\xi^r &= \{\xi_i\}_{i=r}^N = [\xi_r \ \dots \ \xi_N]'
\end{aligned}$$

Note that the vector $[1 \ \dots \ 1]'$ in $\mathcal{G}$ corresponds to the constant part c_0. Comparing with the bilinear model in (5.7) and (5.8), we have $A_a = \mathcal{K} \otimes a$ and $B_b = b \cdot \mathcal{K}$.

Transforming LN subsystem to a linear model

By solving the bilinear model in (5.68), one can get that $\{\hat{z}_i\}_{i=r}^N = \{\hat{f}(U_i)\}_{i=r}^N = K_0\hat{a} + \hat{c}_0$. Also, based on Assumption 5.6, (5.57) is written as

$$x_i = f^{-1}(\hat{z}_i) + e_i \tag{5.69}$$

where the term e_i is due to the existence of error $\hat{z}_i - z_i$. Combining (5.69) with Assumption 5.5, (5.56) and (5.57) can be rewritten as

$$u_k = f^{-1}(\hat{z}_i) - \mu_1 u_{i-1} + \dots - \mu_m u_{k-m} + e_i \tag{5.70}$$

Similar to (5.62), we have the matrix expression $\{f^{-1}(\hat{z}_i)\}_{i=1}^N = \bar{K}_z\bar{a} + \bar{c}_0 + \bar{\xi}$ where $\bar{\xi} = \{\bar{\xi}_i\}_{i=1}^N$ and $\bar{K}_z = [\bar{k}(\hat{z}_i, \tilde{\hat{z}}_j)]_{i=1,j=1}^{i=N,j=m_{sv}}$. Finally, the LN subsystem (5.56) and (5.57) can be transformed to a linear model, i.e,

$$\bar{Y}^r = \bar{\mathcal{G}}\bar{d} + \bar{K}_0\bar{a} + \bar{\xi}^r \tag{5.71}$$

where

$$\begin{aligned}
\bar{Y}^r &= \{u_i\}_{i=r}^N = [u_r \ ... \ u_N]' \\
\bar{\mathcal{G}} &= \begin{bmatrix} 1 & -u_{r-1} & ... & -u_{r-n} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & -u_{N-1} & ... & -u_{N-n} \end{bmatrix} \\
\bar{d} &= \begin{bmatrix} \bar{c}_0 \\ \mu \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix} \\
\bar{K}_0 &= \{\bar{K}_z\}_{r\to N} = [\bar{k}(\hat{z}_i, \tilde{\hat{z}}_j)]_{i=r,j=1}^{i=N,j=m_{sv}} \\
\bar{a} &= [\bar{a}_1 \ ... \ \bar{a}_{m_{sv}}]' \\
\bar{\xi}^r &= \{\bar{\xi}_k\}_{i=r}^N + \{e_i\}_{i=r}^N
\end{aligned} \tag{5.72}$$

Before giving the convergence results, we present the following two assumptions in employing the techniques of kernel machine.

Assumption 5.7. *[59]. (Parameter Selection of Kernel Machine) Parameters N, m_{sv} and ρ are chosen such that $\rho \to 0$ and $m_{sv} \cdot \rho \to \infty$ as $N \to \infty$, $m_{sv} \to \infty$.*

Assumption 5.8. *$[\mathcal{G} \ \ \mathcal{K}] = [\mathcal{G} \ \ K_0 \ ... \ K_s]$ in (5.68) is full column rank. Also, $[\bar{\mathcal{G}} \ \bar{K}_0]$ in (5.71) is full column rank.*

Remark 5.5. *Note that Assumptions 5.2 and 5.3 can be guaranteed for LNL nonlinear systems by Assumptions 5.5-5.8. Lemma 5.1 basically implies that all the components in $\mathcal{K}$ sampled from K in (5.67) can be considered as random i.i.d variables sampled from a probability density function. Thus Assumption 5.3 can be satisfied by LNL nonlinear systems provided that the input output signals are i.i.d as stated in Assumption 5.5, which is to guarantee that the input output signals are persistently exciting (PE). Assumption 5.6 refers to the identifiability of the system when an LNL system is considered. Assumption 5.7 shows how to choose the parameters when using a nonparametric model to represent a nonlinear function by kernel machine. One can refer to [59] for more details about this assumption.*

Assumption 5.8 corresponds to Assumption 5.2 and it implies that, for any $b \neq 0$, $[\mathcal{G} \quad b \cdot \mathcal{K}]$ is full column rank, and for any $a \neq 0$, $[\mathcal{G} \quad \mathcal{K} \otimes a]$ is full column rank. Actually Assumption 5.8 requires that input and output signals are persistently exciting (PE) and the parameters in the kernel machine are properly chosen.

For LNL systems, one needs to estimate the parameters in models (5.68) and (5.71). In next subsection, we will analyze the convergence results in LNL systems.

5.3.3 Convergence Results

Comparing the model (5.68) with the model in (5.7) and (5.8), one could notice that the only difference is the existence of the approximation error ξ^r in (5.68). Here by noting that the variance of the approximation error ξ_i is a function of N, m_{sv}, ρ, we introduce the following Lemma from [59].

Lemma 5.6. *[59]. Under Assumptions 5.7 and 5.8, for any function $f(.)$ assumed in Assumption 5.6, we have $\lim_{m_{sv}\to\infty} \sigma_\xi^2 = \sigma_\xi^2(m_{sv}) = 0$ asymptotically almost surely.*

Lemma 5.6 implies that the approximation error will decrease by increasing the dimension of $\mathcal{K}$ in (5.68), which means that the basis functions can be dense on a closed set.

Theorem 5.3. *Consider the identification of the NL subsystem (5.56) and (5.58) of the LNL Wiener-Hammerstein system in (5.57)-(5.58) under Assumptions 5.1-5.8. By using the proposed iterative algorithm to solve (5.68), it can be obtained that $\hat{a} \to a$, $\hat{b} \to b$ and $\hat{d} \to d$ asymptotically almost surely as $N \to \infty, m_{sv} \to \infty, \rho \to 0$.*

Proof. Firstly, we obtain $\lim_{N\to\infty} \hat{d} = d$ almost surely by Theorem 5.1. From

Lemma 5.6, we have $\lim_{m_{sv}\to\infty}\sigma_{\bar{\xi}}^2 = 0$ asymptotically almost surely under Assumptions 5.1 and 5.5, 5.7 and 5.8. Then the model in (5.68) and the model in (5.7) and (5.8) become equivalent. When using the proposed fixed point iteration, we obtain $\hat{a} \to a$, $\hat{b} \to b$ and $\hat{d} \to d$ asymptotically almost surely as $N \to \infty, m_{sv} \to \infty, \rho \to 0$ by Theorem 5.2. □

Theorem 5.4. *Consider the identification of the LN subsystem (5.56) and (5.57) of the LNL Wiener-Hammerstein system in (5.56)-(5.58) under Assumptions 5.1-5.8. Substituting $\{\hat{z}_k\}$ obtained from (5.68), it can be obtained that $\hat{\mu} \to \mu$ and $\hat{f} \to f$ as $N \to \infty, m_{sv} \to \infty, \rho \to 0$ through solving (5.71).*

Proof. Note that we have $\{\hat{z}_k\} \to \{z_k\}$ asymptotically almost surely as the consistency of $\hat{a}$, $\hat{b}$ and $\hat{d}$ in Theorem 5.3. Also, in the linear model (5.71), we have $\sigma_{\bar{\xi}}^2 \to 0$ asymptotically almost surely as $N \to \infty, m_{sv} \to \infty, \rho \to 0$ by Lemma 5.6. Thus, it can be obtained that $\hat{\mu} \to \mu$ by Theorem 5.1. If all the consistency of the parameters in the LNL system have been achieved, one can directly obtain the estimated $\hat{f}$ based on (5.61) and $\lim_{N\to\infty,m_{sv}\to\infty,\rho\to 0}\|\hat{f} - f\|_2 = \lim_{N\to\infty,m_{sv}\to\infty,\rho\to 0}\sigma_{\bar{\xi}}^2 = 0$ asymptotically almost surely. □

5.3.4 Extension to IIR Linear Systems

We now consider the case that the first linear system is a stable IIR system, namely

$$\begin{aligned} x_i &= \alpha_1 x_{i-1} + ... + \alpha_{n_0} x_{i-n_0} + \beta_0 u_i + ... + \beta_{m_0} u_{i-m_0} \\ z_i &= f(x_i) \\ y_i &= \omega_1 y_{i-1} + ... + \omega_n y_{i-n} + b_0 z_i + ... + b_s z_{i-s} + v_i \end{aligned} \tag{5.73}$$

where $\alpha = [\alpha_1 \ ... \ \alpha_{n_0}]'$ and $\beta = [\beta_0 \ ... \ \beta_{m_0}]'$ are unknown parameter vectors. Assume that the impulse response of the IIR system is $\{\mu_j\}_{j=0}^{\infty}$, i.e,

$$\begin{aligned} x_i = \ & \mu_0 u_i + ... + \mu_m u_{i-m} \\ & + \mu_{m+1} u_{i-m-1} + ... + \mu_{m+\infty} u_{i-\infty}. \end{aligned} \tag{5.74}$$

Let $e_i = \mu_{m+1} u_{i-m-1} + ... + \mu_{m+\infty} u_{i-\infty} = \bar{\mu}_m' \bar{\mathrm{u}}_m$ where $\bar{\mu}_m = [\mu_{m+1} \ ... \ \mu_{m+\infty}]'$ and $\bar{\mathrm{u}}_m = [u_{i-m-1} \ ... \ u_{i-\infty}]'$. An FIR model can be used to approximate the IIR model, which gives

$$x_i \ = \mu_0 u_i + ... + \mu_m u_{i-m} + e_i \tag{5.75}$$

where e_k denotes the approximation error in representing the IIR linear system. Note that $\|e_i\|_1 \leq \|\bar{\mu}_m\|_1 \|\bar{\mathrm{u}}_m\|_\infty$ where $\|.\|_1$ and $\|.\|_\infty$ denote the one norm and infinity norm, respectively. As the IIR system is stable, $\lim_{m\to\infty} \|\bar{\mu}_m\|_1 \to 0$. Also, from Assumption 5.5, $\|\bar{\mathrm{u}}_m\|_\infty \leq u_{max}$. Therefore $\|e_i\|_1 \leq \|\bar{\mu}_m\|_1 \|\bar{\mathrm{u}}_m\|_\infty \leq \|\bar{\mu}_m\|_1 u_{max}$. Thus, we have $\lim_{m\to\infty} \|e_i\|_1 = \lim_{m\to\infty} \|\bar{\mu}_m\|_1 u_{max} = 0$ asymptotically almost surely, which means the approximation error e_k can be reduced by increasing the order m in the FIR model. In this case, the matrix form of system (5.68) needs to be revised to

$$Y^r = \mathcal{G} d + (\mathcal{K} \otimes a) b + \xi^r + \zeta^r + v^r \tag{5.76}$$

where ζ^r is an error vector with its element depending on e_k. Since the IIR system is stable, the variance $D(\zeta) = \sigma_\zeta^2$ can be reduced by increasing the order m in the FIR model and $\lim_{m\to\infty} \sigma_\zeta^2 \to 0$ asymptotically almost surely from the above analysis. Then we have the following corollary whose proof is the same as that of Theorem 5.3 and 5.4.

Corollary 5.2. *Consider the identification of the LNL Wiener-Hammerstein models in (5.75), (5.57) and (5.58) satisfying Assumptions 5.1 and 5.5-5.8. By us-*

ing the proposed fixed point iteration in solving (5.76), it can be obtained that $\hat{a} \to a$, $\hat{b} \to b$ *and* $\hat{d} \to d$ *as well as* $\hat{f} \to f$ *asymptotically almost surely as* $N \to \infty, m_{sv} \to \infty, \rho \to 0$ *and* $m \to \infty$.

5.4 Example Illustration

Example 5.4.1. *(System of bilinear equations) In this example, we consider the bilinear system of equations:* $\sum_{j=1}^{m}\sum_{k=1}^{n} T_{jt}^{i} a_j b_t = y_i$, *for* $i = 1, ..., N$ *where* $T_{jt}^{i} \in U(-1, 1)$ *is a known i.i.d random number and* y_i *is a known observation. This system consists of* N *equations with* $m + n$ *unknown parameters which are set as* $a = [0.1826\ 0.3651\ 0.5477\ 0.7303]'$ *and* $b = [-0.6521\ -0.3477\ -0.1651\ 0.2826]'$. *The system can be rewritten in the forms of (5.4) and (5.5) with* $v = 0$ *and*

$$A_a = \sum_{j=1}^{m} a_j A_j, A_j = \begin{bmatrix} T_{j1}^{1} & ... & T_{jn}^{1} \\ \vdots & \ddots & \vdots \\ T_{j1}^{N} & ... & T_{jn}^{N} \end{bmatrix} \in R^{N \times n}$$

$$B_b = \sum_{t=1}^{n} b_t B_t,\ B_t = \begin{bmatrix} T_{1t}^{1} & ... & T_{mt}^{1} \\ \vdots & \ddots & \vdots \\ T_{1t}^{N} & ... & T_{mt}^{N} \end{bmatrix} \in R^{N \times n}.$$

We identify the system with the proposed iterative algorithm by fixing $\|a\|_2 = 1$ *and* $a_0 > 0$ *and choosing* $N = 20$. *Figure 5.1 shows the estimates with respect to number of iterations* k. *It is observed that the estimates converge to the true parameters* $\hat{a} = [0.1826\ 0.3651\ 0.5477\ 0.7303]'$ *and* $b = [-0.6521\ -0.3477\ -0.1651\ 0.2826]'$ *in only a few iterations. In addition, to show how the estimates behave with different number of data points* N, *we calculate the error* $E_a(N) = \|\hat{a} - a\|_2$ *with respect*

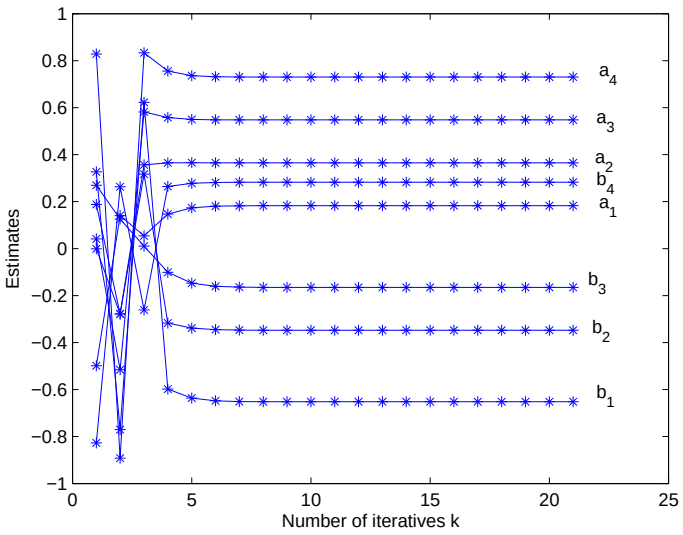

Figure 5.1: Estimates with respect to number of iterations k ($N = 20$)

to different N and the results are shown in Figure 5.2. It is observed that when $N > 15$, the error $E_a(N)$ can be made zero for the bilinear system shown in Example 5.4.1 for noise free case.

Example 5.4.2. *(LNL block-oriented models) Consider the following system*

$$\begin{aligned} x_i = &\quad 1u_i + 0.3u_{i-1} + 0.1u_{i-2} + 0.1u_{i-3} \\ z_i = &\quad \arctan(x_i) \\ y_i = &\quad 0.4y_{i-1} + 0.1y_{i-2} \\ &\quad +0.8111z_i + 0.4867z_{i-1} + 0.3244z_{i-2} + v_i \end{aligned}$$

where v_i is white noise with zero mean and standard derivation 0.1. *We use the proposed iterative algorithm to identify the system by fixing $\|b\|_2 = 1$ and $b_0 > 0$. The input signal $\{u_t\} \in [-1, 1]$ which is i.i.d and $m_{sv} = \frac{N}{4} - r$, $\rho = e^{-m_{sv}/1000}$. Note that all the Assumptions 5.1 -5.8 can be satisfied in this example. It is obtained that $\hat{\mu} = [1.0000\ 0.3008\ 0.1009\ 0.0991]'$, $\hat{\omega} = [0.4020\ 0.0910]'$ and $\hat{b} = [0.8122\ 0.4853\ 0.3237]'$. Clearly, the estimates are very close to the true values. To show how the estimates converge to the fixed point with respect to the number*

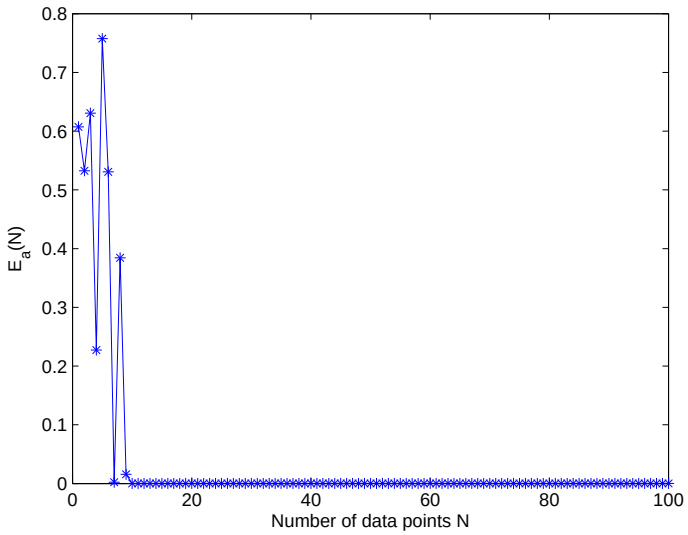

Figure 5.2: Estimation error with respect to number of data points N

of iterations, let $\hat{c} = [\hat{\mu}'\ \hat{\omega}'\ \hat{b}']$ *and we plot the difference* $d_c(k) = \|\hat{c}(k+1) - \hat{c}(k)\|_2$ *at each iteration. Figure 5.3 shows that the iteration converges in only a few iterations. Also, as seen in Figure 5.3,* Q *in (5.55) can be strictly smaller than* 1 *even for a finite* $N = 800$ *in this example. The estimated unknown nonlinear function together with the true function is shown in Figure 5.4. It is observed that the estimated function is hardly distinguishable from the true function. To show how the estimates behave with different data points* N*, we calculate the error* $E_c(N) = \|\hat{c} - c\|_2$ *with respect to different* N *and the results are shown in Figure 5.5. This illustrates that errors can be sufficiently small when* N *becomes large enough.*

In Theorem 5.2, it is shown that the consistency of the estimates is obtained when $N \to \infty$. But in real applications as shown in the above two Examples, N is not necessarily large enough. For instance, good performances are obtained with finite N in both Example 5.4.1 and Example 5.4.2. In case $v = 0$ in Example 5.4.1, N can be as small as 15. In Example 5.4.2, when $N > 800$, the performance of the estimates becomes quite good.

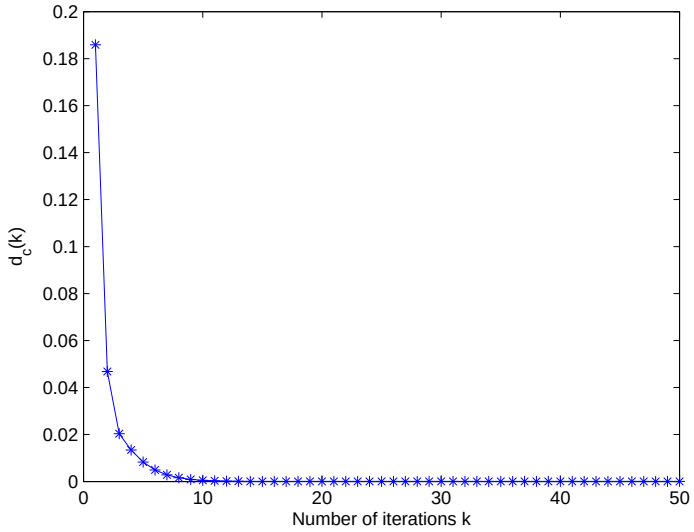

Figure 5.3: Illustration of convergence

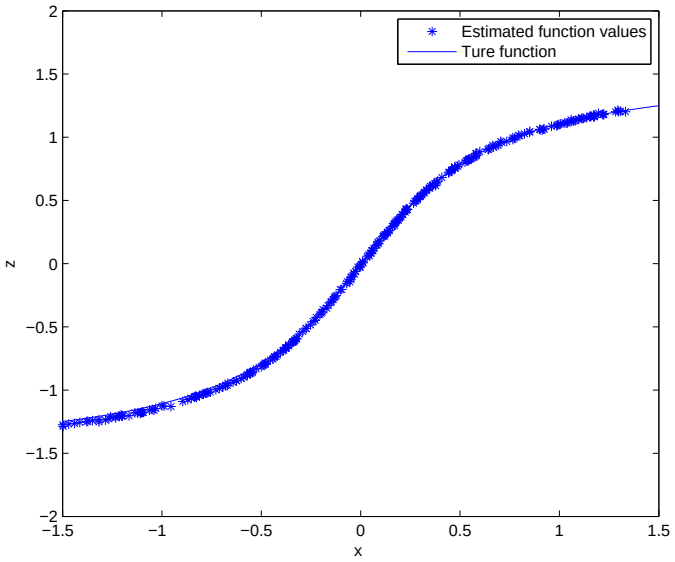

Figure 5.4: True nonlinear function and estimated function

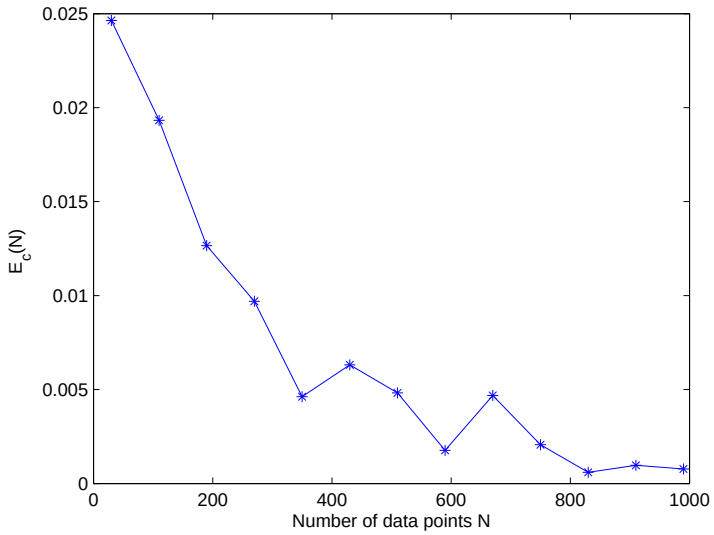

Figure 5.5: Estimation error respect to number of data points N

5.5 Conclusion

In this chapter, we propose a fixed point iteration approach for the identification of bilinear models with convergence properties established. For applications, parameter estimation of a system of bilinear equations and the identification of LNL systems are illustrated. With this, the long-standing convergence problem of iteratively identifying LNL Wiener-Hammerstein models has been solved. In addition, we extend the static nonlinear function (N) to a nonparametric model represented by using kernel machine.

Chapter 6

Identification of Block-oriented Systems Based on Biconvex Optimization

In this chapter, we investigate biconvex optimization in the identification of the class of block-oriented nonlinear systems newly proposed in Chapter 2. A common model is proposed to represent such block-oriented systems. It is shown that identifying the common model can be formulated as a biconvex optimization problem, where we only need to find the unique partial optimum point of a biconvex cost function in the formulated optimization problem to obtain its global minimum point. The normalized alterative convex search (NACS) algorithm is proposed and its convergence property is achieved. Thus, we provide a unified framework for the identification of block-oriented systems.

6.1 Introduction

Nonlinear programming plays a major role in mathematical modeling and programming. In nonlinear programming, many algorithms lead to a local minimum point instead of the global minimum point. If the programming is convex, a local minimum point and a global minimum point become equivalent. So in parameter estimation [75], convexity naturally implies the convergence of the estimates. This is why convex optimization has been widely used. Different from convex optimization, biconvex optimization [76] belongs to general global optimization problems which may have a large number of local minimum points. However, the convex substructures indeed can be exploited in solving biconvex optimization problems. In this chapter, biconvex optimization is investigated in the identification of nonlinear systems such as the well known and widely implemented block-oriented nonlinear systems [56].

Block-oriented systems are composed of linear dynamic systems and nonlinear static functions [56]. For example, Hammerstein-Wiener [77] systems, which include Hammerstein systems [26] and Wiener systems [78] as its special cases, are one of the most well known member of block-oriented systems with the linear dynamic block between two nonlinear blocks. One popular identification scheme for block-oriented systems is the iterative method mentioned in our previous chapters. The original iterative algorithm may be divergent as seen in [49]. But with normalization, we obtained the globally asymptotical convergence property for Hammerstein systems in Chapter 3. Also, the bilinear models are considered in Chapter 4. However, the convergence of iterative algorithm for a general block-oriented system, for example, the newly proposed model shown in Chapter 2, is still unknown. In this chapter, we will show how the identification of such systems can be formulated as a biconvex optimization problem, which is to minimize a

proposed biconvex cost function on a convex set. It is proved that the formulated biconvex optimization only needs to find the unique partial optimum point of the cost function to obtain its global minimum point under arbitrary non-zero initial conditions. Based on this, a normalized alterative convex search (NACS) algorithm is presented. Its convergence property is also established. This provides a unified framework for the iterative identification of block-oriented systems.

The remaining part of this chapter is organized as follows. We introduce biconvex optimization problem and investigate some theoretical results in Section 6.2. Identification of the proposed common model is formulated as a biconvex optimization problem in Section 6.3. The convergence property is analyzed in Section 6.4. In Section 6.5, NACS algorithm is presented for the identification of Hammerstein-Wiener system. Model generalization is discussed in Section 6.6 and simulation examples are given in Section 6.7. Finally, this chapter is concluded in Section 6.8.

6.2 Biconvex Optimization Problem

In practice, biconvex optimization problems frequently occur in industrial applications. Here we review certain theoretical results for biconvex sets and biconvex functions for general biconvex optimization problems.

6.2.1 Definition of Biconvex Optimization

Recall that a set $S \subseteq R^N$ is said to be convex if for any two points $s_1, s_2 \in S$, the line segment joining s_1 and s_2 is completely contained in S. Function $F : S \to R$ is convex if $F(\lambda s_1 + (1-\lambda)s_2) \leq \lambda F(s_1) + (1-\lambda)F(s_2)$ where $\lambda \in [0,1]$ and $s_1, s_2 \in S$.

Now we give some definitions on biconvex set and biconvex functions. Let $F : X \times Y \rightarrow R$ where $X \subseteq R^N$ and $Y \subseteq R^M$ are non-empty convex sets. Let $S \subseteq X \times Y$. The subsets S_x and S_y of S are defined as $S_x = \{y \in Y | (x, y) \in S)\}$ and $S_y = \{x \in X | (x, y) \in S)\}$, respectively. In other words, S_x denotes the subset for a fixed value of x, and S_y denotes the subset for a fixed value of y.

Definition 6.1. *The set $S \subseteq X \times Y$ is called a biconvex set on $X \times Y$, if S_x is convex for every $x \in X$ and S_y is convex for every $y \in Y$.*

Note from [76] that set $S \subseteq X \times Y$ is biconvex if and only if for all quadruples (x_1, y_1), (x_1, y_2), (x_2, y_1), $(x_2, y_2) \in S$, it holds that for every $(\lambda, \mu) \in [0, 1] \times [0, 1]$, $(x_\lambda, y_\mu) := ((1 - \lambda)x_1 + \lambda x_2, (1 - \mu)y_1 + \mu y_2) \in S$. Obviously, a convex set must be a biconvex set while the converse is not true. For example, the sets in Figure 6.1 are biconvex but not convex.

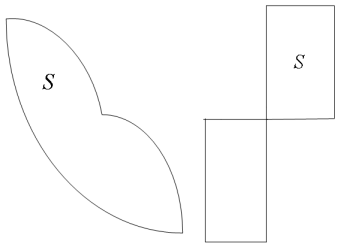

Figure 6.1: Examples of biconvex set which are non-convex

Definition 6.2. *A function $F : S \rightarrow R$ on a biconvex set $S \subseteq X \times Y$ is called biconvex function if $F_x(x, .) : S_x \rightarrow R$, $F_y(., y) : S_y \rightarrow R$ are convex functions on B_x and B_y, respectively.*

Definition 6.3. *An optimization problem of the form $min(F(x, y) : (x, y) \in S)$ is said to be a biconvex optimization problem or biconvex in short, if the feasible set S is biconvex on $X \times Y$, and the objective function is biconvex on S.*

Let $X \subseteq R^N$ and $Y \subseteq R^M$ be two non-empty convex sets, and let F be a real valued function on $X \times Y$. As shown in [81], F is biconvex if and only if for all quadruples (x_1, y_1), (x_1, y_2), (x_2, y_1), $(x_2, y_2) \in X \times Y$, it holds that for every $(\lambda, \mu) \in [0, 1] \times [0, 1]$: $F(x_\lambda, y_\mu) \leq (1-\lambda)(1-\mu)F(x_1, y_1) + (1-\lambda)\mu F(x_1, y_2) + \lambda(1-\mu)F(x_2, y_1) + \lambda\mu F(x_2, y_2)$ where $(x_\lambda, y_\mu) := ((1-\lambda)x_1 + \lambda x_2, (1-\mu)y_1 + \mu y_2)$.

Definition 6.4. *Biconvex optimization with a differentiable biconvex function and separable constraints is defined as* $min(F(x, y) : x \in X \subseteq R^M, y \in Y \subseteq R^N)$ *where* $F(x, y)$ *is a differentiable biconvex function from* $X \times Y \rightarrow R$.

Definition 6.5. *Let* $F : S \rightarrow R$ *be a given biconvex function and* $(x^\star, y^\star) \in S$. *Then,* $(x^\star, y^\star)$ *is a partial optimum of* F *on* S *if* $F(x^\star, y^\star) \leq F(x, y^\star) \ \ \forall x \in S_{y^\star}$ *and* $F(x^\star, y^\star) \leq F(x^\star, y) \ \ \forall y \in S_{x^\star}$.

Remark 6.1. *In this chapter, we consider biconvex optimization defined in Definition 6.4. Different from convex optimization problems, biconvex optimization problems are in general global optimization problems which may have a large number of local minima. However, the convex substructures in Definition 6.4 indeed can be exploited when solving the biconvex problems. By this way, some meaningful results in biconvex optimization problems may be achieved as that in convex optimization problems. The definitions of biconcave and bilinear functions are obtained by replacing the corresponding properties of biconvex of being concave and bilinear, respectively.*

6.2.2 Alterative Convex Search Algorithm (ACS)

After representing the biconvex problem defined in Definition 6.4, the remaining work is how to solve it. Note that biconvex optimization has become a hot research topic in recent years and there are many schemes focusing the methodology of

solving it and applications [83] [84] [76] and so on. Among existing schemes, the most well known algorithm is the alterative convex search (ACS) algorithm [76]. Let k be the kth iteration. The alterative convex search (ACS) algorithm is given as:

Step 1: Set $k = 0$ and choose an arbitrary starting point $z(0) = (x(0), y(0)) \in S$.

Step 2: Solve the convex optimization problem $min\ \{F(x, y(k)), x \in S_{y(k)}\}$ for fixed $y(k)$. If there exists an optimal solution $x^\star \in S_{y(k)}$ to this problem, set $x(k+1) = x^\star$, otherwise stop.

Step 3: Solve the convex optimization problem $min\ \{F(x(k+1), y), y \in S_{x(k+1)}\}$ for fixed $x(k+1)$. If there exists an optimal solution $y^\star \in S_{x(k+1)}$ to this problem, set $y(k+1) = y^\star$, otherwise stop.

Step 4: Set $z(k+1) = (x(k+1), y(k+1))$. If a stopping criterion is satisfied, then stop, otherwise replace k by $k+1$ and go back to Step 2.

Remark 6.2.

1) *The order of the optimization problems in Step 2 and Step 3 can be permuted, i.e., it is possible first to minimize $F(x, y)$ with respect to y-variable, followed by an optimization with respect to x-variable.*

2) *There are several ways to define the stopping criterion in Step 4. For example, one can consider the absolute value of the difference between $z(k-1)$ and $z(k)$ (or the difference in their function values) or the relative change in the z-variable compared to the previous iteration.*

Lemma 6.1. *[76] For the biconvex optimization problem in Definition 6.4, let $z^\star \in X \times Y$ be the limit of the sequence $\{z(k)\}$ generated by ACS. Then $z^\star$ is a partial optimum of $F(x, y)$.*

6.3 Biconvex Optimization in Parameter Estimation

6.3.1 A Common Model

In science end engineering, convex optimization has been investigated and applied extensively. However in many cases, we obtain a biconvex problem instead of a convex one. For example, estimating the parameters in the common model (6.1) and (6.2) can be formulated as a biconvex problem. This model is called "common" model because it actually represents many parameter estimation problems. Later in Section 6.5, it will be shown that the new class of block-oriented nonlinear systems in Chapter 2 can be formulated as the common model.

$$Y = \mathcal{G}d + \mathcal{K}\gamma + v \tag{6.1}$$

$$Y - L(d) = b_1 K_1 a_1 + ... + b_m K_m a_m + v \tag{6.2}$$

where $\mathcal{G} \in R^{N\times n}$, $\mathcal{K} \in R^{N\times l}$ and $K_i \in R^{N\times l}$, for $i = 1, ..., m$ are known matrices constructed based on learning data points, $d = [d_1 \ ... \ d_n]' \in R^{n\times 1}$ and $\gamma = [\gamma_1 \ ... \ \gamma_l]' \in R^{l\times 1}$, $b = [b_1 \ ... \ b_m]' \in R^{m\times 1}$ and $a = [a_1' \ ... \ a_i' \ ... \ a_m'] \in R^{lm\times 1}$ with $a_i = [a_{i1} \ ... \ a_{il}]' \in R^{l\times 1}$ are unknown parameters, and $L(d)$ is a known function with respect to parameter d in (6.1), $Y = [y_1 \ ... \ y_N]'$ is the observation vector and $v = [v_1 \ ... \ v_N]'$ denotes the noise vector. Our objective is to obtain the estimates $\hat{a}, \hat{b}$ and $\hat{d}$ of a, b and d, respectively, in (6.1) and (6.2).

To characterize the above common model and estimate its parameters, we make the following assumptions.

Assumption 6.1. *The noise v_t is an i.i.d random variable with zero mean and*

$E(v_t^2) = D(v_t) = \sigma_v^2 < \infty$.

Assumption 6.2. *$\|b\|_2$ is known and the first nonzero entry of b is positive.*

Assumption 6.3. *In (6.1), $[\mathcal{G}\ \mathcal{K}] \in R^{N\times(l+n)}$ is a full column rank matrix. In addition, $\lim_{N\to\infty} tr((\mathcal{G}'\mathcal{G})^{-1}) = 0$ where $tr(.)$ is the trace of a matrix.*

Assumption 6.4. *In (6.2), $K = [K_1\ ...\ K_m] \in R^{N\times lm}$ is a matrix such that $lim_{N\to\infty}\frac{K'K}{N} = I$ where I is an identity matrix.*

Remark 6.3. *Note that in Assumption 6.3, $\lim_{N\to\infty} tr((\mathcal{G}'\mathcal{G})^{-1}) = 0$ is easy to be satisfied as we analyzed in Chapter 5, where it is shown that if $\mathcal{G}$ is a random matrix, then the condition is definitely satisfied. Also, Assumption 6.2 is to guarantee a unique expression of the common model.*

In next subsection, it will be shown how to estimate d in (6.1), a and b in (6.2), respectively. In Section 6.4, we will prove that $\hat{d}$, $\hat{a}$ and $\hat{b}$ converge to their true values under Assumptions 6.1-6.4.

After $\hat{d}$ is obtained in (6.1), one needs to minimize the following cost function $J_N^{\hat{d}}(.)$ to obtain the estimates of a and b.

$$J_N^{\hat{d}}(\bar{a},\bar{b}) = \tfrac{1}{N}(\bar{Y} - Y^* - v)'(\bar{Y} - Y^* - v) \tag{6.3}$$

where $\bar{a} = [\bar{a}_1'\ ...\ \bar{a}_m']$, $\bar{b} = [\bar{b}_1\ ...\ \bar{b}_m]'$, $\bar{Y} = L(\hat{d}) + \bar{b}_1K_1\bar{a}_1 + ... + \bar{b}_mK_m\bar{a}_m$ and $Y^* = L(d) + b_1K_1a_1 + ... + b_mK_ma_m$. Let

$$\begin{aligned} K_b &= [b_1K_1\ ...\ b_mK_m] \\ K_a &= [K_1a_1\ \ ...\ \ K_ma_m] \\ J^{\hat{d}}(\bar{a},\bar{b}) &= \lim_{N\to\infty} J_N^{\hat{d}}(\bar{a},\bar{b}) \end{aligned} \tag{6.4}$$

It is necessary to mention that $\lim_{N\to\infty} J_N^{\hat{d}}(\bar{a},\bar{b})$ exists since we have $lim_{N\to\infty}\frac{K'K}{N} = I$ in Assumption 6.4.

As the noise is independent of Y^* and $\bar{Y}$, we obtain

$$\begin{aligned} J^{\hat{d}}(\bar{a},\bar{b}) &= lim_{N\to\infty}\tfrac{1}{N}\|L(\hat{d}-d)+\mathcal{K}_{\bar{b}}\bar{a}-\mathcal{K}_b a\|_2^2+\sigma_v^2 \\ &= \lim_{N\to\infty}\tfrac{1}{N}\|L(\hat{d}-d)+\mathcal{K}_{\bar{a}}\bar{b}-\mathcal{K}_a b\|_2^2+\sigma_v^2 \end{aligned} \tag{6.5}$$

Lemma 6.2. *Minimizing $J^{\hat{d}}(\bar{a},\bar{b})$ in a bounded domain is a biconvex optimization defined in Definition 6.4.*

Proof. Firstly, it is easy to see that the bounded definition domain is a convex set and of course a biconvex set from Definition 1. Then, it is easy to see that $J^{\hat{d}}(\bar{a},\bar{b})$ is a biconvex function with respect to separable $\bar{a}$ and $\bar{b}$ based on Definitions 6.2-6.4. Thus, this lemma holds. □

Note that if $\hat{d}=d$ in (6.5), the cost function becomes

$$\begin{aligned} J(\bar{a},\bar{b}) &= lim_{N\to\infty}\tfrac{1}{N}\|\mathcal{K}_{\bar{b}}\bar{a}-\mathcal{K}_b a\|_2^2+\sigma_v^2 \\ &= \lim_{N\to\infty}\tfrac{1}{N}\|\mathcal{K}_{\bar{a}}\bar{b}-\mathcal{K}_a b\|_2^2+\sigma_v^2 \end{aligned} \tag{6.6}$$

6.3.2 Identifying the Common Model

Firstly, we use the space projection method [59] to obtain $\hat{d}$ from (6.1), which is given as

$$\hat{d}=((I-\mathcal{G}\mathcal{G}^+\mathcal{K}\mathcal{K}^+)\mathcal{G})^+\mathcal{G}\mathcal{G}^+(I-\mathcal{K}\mathcal{K}^+)Y \tag{6.7}$$

where I is an identity matrix and the superscript $^+$ denotes a generalized inverse of a matrix, for example, $\mathcal{G}^+=(\mathcal{G}'\mathcal{G})^{-1}\mathcal{G}'$. Then we focus on the identification of (6.2) based on the newly estimated $\hat{d}$. This is to obtain $\hat{a}$ and $\hat{b}$ in (6.2) by minimizing the cost function in (6.5). Unlike convex optimization, minimizing (6.5) has multiple local minimums.

Based on the general ACS algorithm discussed in Section 6.2.2, here we present a normalized ACS (NACS) for the identification of the common model (6.1) and (6.2) in this subsection.

Step 1: Obtain $\hat{d}$ based on (6.7).

Step 2: Give $\hat{b}$ an arbitrary nonzero initial value $\hat{b}(0)$.

Step 3: Solve the problem $\hat{a}(k) = argmin_{\bar{a}} J_N^{\hat{d}}(\bar{a}, \hat{b}(k-1))$, which gives the least square estimates

$$\hat{a}(k) = (K'_{\hat{b}(k-1)} K_{\hat{b}(k-1)})^{-1} K'_{\hat{b}(k-1)} (Y - L(\hat{d})) \tag{6.8}$$

when $\hat{b}(k-1)$ and $\hat{d}$ become known.

Step 4: Solve the problem $\hat{b}_{op}(k) = argmin_{\bar{b}} J_N^{\hat{d}}(\hat{a}(k), \bar{b})$. This gives

$$\hat{b}_{op}(k) = (K'_{\hat{a}(k)} K_{\hat{a}(k)})^{-1} K'_{\hat{a}(k)} (Y - L(\hat{d})) \tag{6.9}$$

Let κ be the first nonzero entry of $\hat{b}_{op}(k)$. Then estimate $\hat{b}(k)$ is given as

$$\hat{b}(k) = sgn(\kappa) \cdot \|b\|_2 \cdot \hat{b}_{op}(k) / \|\hat{b}_{op}(k)\|_2 \tag{6.10}$$

such that the norm $\|\hat{b}(k)\|_2 = \|b\|_2$ and the first nonzero entry of $\hat{b}(k)$ is positive.

Step 4: If a stopping criterion is satisfied, end iteration. Otherwise, $k := k + 1$ and go back to Step 2.

Note that $K_{\hat{b}(k-1)}$ and $K_{\hat{a}(k)}$ are constructed by replacing b and a with $\hat{b}(k-1)$ and $\hat{a}(k)$, respectively, based on (6.4). As mentioned in Remark 6.2, estimating $\hat{a}(k)$ and $\hat{b}(k)$ can be permuted. We will analyze the convergence of our proposed NACS in next section Under Assumptions 6.1-6.4.

6.4 Convergence Analysis

In this section, we consider the convergence of the estimates $\hat{d}$, $\hat{b}$ and $\hat{a}$ in the common model (6.1) and (6.2). We first prove $\lim_{N\to\infty}\hat{d} = d$ almost surely. Then, based on this, we prove that $\lim_{N\to\infty}\hat{b} = b$ and $\lim_{N\to\infty}\hat{a} = a$ almost surely.

Theorem 6.1. *For the estimates $\hat{d}$ in (6.7), we have $\lim_{N\to\infty}\hat{d} = d$ under Assumptions 6.1 and 6.3.*

Proof. As shown in Theorem 3.1, we have $\hat{d} = d+(\mathcal{G}'\mathcal{G})^{-1}\mathcal{G}'v$ and $\lim_{N\to\infty} tr((\mathcal{G}'\mathcal{G})^{-1}) = 0$. Thus, we have this Theorem. □

Theorem 6.2. *Under Assumptions 6.1-6.4, when $N \to \infty$, using NACS algorithm to minimize biconvex cost function $J(\bar{a},\bar{b})$ in (6.6) leads to a unique partial minimum point (a,b), i.e, $\lim_{N\to\infty}\hat{b} = b$ and $\lim_{N\to\infty}\hat{a} = a$ almost surely.*

Proof. From Theorem 6.1, we have $\hat{d} = d$ almost surely as $N \to \infty$ under Assumptions 6.1 and 6.3. Then the cost function in (6.5) becomes $J(\bar{a},\bar{b})$ in (6.6). We first prove that the point (a,b) corresponding to true parameters a and b is a partial optimum point of $J(\bar{a},\bar{b})$. From (6.5), we get

$$\begin{aligned} J(a+\Delta a, b+\Delta b) &= \sigma_v^2 + \lim_{N\to\infty}\tfrac{1}{N}\|\Delta Y\|_2 \\ &\geq \lim_{N\to\infty} J_N(a,b) \\ &= \sigma_v^2 \end{aligned} \tag{6.11}$$

for any Δa and Δb. This shows that (a,b) is a global minimum point and of course a partial optimum point. Now we prove the uniqueness of the partial optimum point (a,b) by contradiction. Assume that $(\hat{a},\hat{b})$ is a partial optimum point with $\hat{a} \neq a$ or $\hat{b} \neq b$. We consider the following two cases:

Case 1: $\hat{a} \neq a$

In this case, let

$$J_{\Delta a}(\hat{a},\hat{b}) = \lim_{\|\Delta a\|_2 \to 0}(J(\hat{a}+\Delta a,\hat{b}) - J(\hat{a},\hat{b})) \tag{6.12}$$

We have

$$\begin{aligned} J_{\Delta a}(\hat{a},\hat{b}) &= \lim_{N\to\infty} \tfrac{1}{N}((\hat{a}+\Delta a)' K_{\hat{b}}' K_{\hat{b}} (\hat{a}+\Delta a) \\ &\quad - \hat{a}' K_{\hat{b}}' K_{\hat{b}} \hat{a} - 2(K_b a)'(K_{\hat{b}} \Delta a)) \\ &= \lim_{N\to\infty} \tfrac{2}{N}(\Delta a' K_{\hat{b}}' K_{\hat{b}} \hat{a} - (K_b a)'(K_{\hat{b}} \Delta a)) \end{aligned} \tag{6.13}$$

From Assumption 6.4, we have

$$\begin{aligned} \lim_{N\to\infty} \tfrac{1}{N} K_{\hat{b}}' K_{\hat{b}} &= \hat{b}'\hat{b} I \\ \lim_{N\to\infty} \tfrac{1}{N} K_b' K_{\hat{b}} &= b'\hat{b} I \end{aligned} \tag{6.14}$$

Then,

$$J_{\Delta a}(\hat{a},\hat{b}) = \lim_{\|\Delta a\|_2 \to 0} 2((\Delta a)'\hat{a}\hat{b}'\hat{b} - (\Delta a)' a b'\hat{b}) \tag{6.15}$$

Let $\Delta a = \rho(a - \hat{a})$ where $\rho > 0$. As $\|\hat{b}\|_2 = \|b\|_2$, i.e, $\hat{b}'\hat{b} = b'b$. If $\hat{b} \neq b$,

$$\hat{b}' b - \hat{b}'\hat{b} \leq 0 \tag{6.16}$$

Then, we have

$$\begin{aligned} J_{\Delta a}(\hat{a},\hat{b}) &= \lim_{\rho\to 0} 2(\rho(a-\hat{a})'\hat{a}\hat{b}'\hat{b} - \rho(a-\hat{a})' a b'\hat{b}) \\ &\leq \lim_{\rho\to 0} 2(\rho(a-\hat{a})'\hat{a} b'\hat{b} - \rho(a-\hat{a})' a b'\hat{b}) \\ &= \lim_{\rho\to 0} 2(\rho(a-\hat{a})'(\hat{a}-a) b'\hat{b}) \\ &= -\lim_{\rho\to 0} 2\rho \|\Delta a\|_2^2 b'\hat{b} \end{aligned} \tag{6.17}$$

Let θ be the angle between b and $\hat{b}$. Note that both the first nonzero entry of $\hat{b}$

and b are positive under Assumption 6.2. Then we have $|\theta| < 90°$, which gives $b'\hat{b} = \|b\|_2\|\hat{b}\|_2 cos(\theta) > 0$. Thus, we have

$$J_{\Delta a}(\hat{a}, \hat{b}) < 0 \tag{6.18}$$

as long as $\hat{b} \neq b$ and $\hat{a} \neq a$, which means $(\hat{a}, \hat{b})$ cannot be a partial optimum point in this case.

Case 2: $\hat{a} = a$ while $\hat{b} \neq b$.

In this case, let $\Delta b = \rho(b - \hat{b})$ where $\rho > 0$ and $J_{\Delta b}(\hat{a}, \hat{b}) = \lim_{\|\Delta b\|_2 \to 0}(J(\hat{a}, \hat{b} + \Delta b) - J(\hat{a}, \hat{b}))$. Note that we still have $\|\hat{b}\| = \|b\|$. Then we obtain

$$\begin{aligned} J_{\Delta b}(\hat{a}, \hat{b}) &= \lim_{\rho \to 0} 2(\rho(b - \hat{b})'\hat{b}\hat{a}'\hat{a} - \rho(b - \hat{b})'ba'\hat{a}) \\ &= \lim_{\rho \to 0} 2(\rho(b - \hat{b})'\hat{b}a'a - \rho(b - \hat{b})'ba'a) \\ &= \lim_{\rho \to 0} 2\rho(\hat{b}'b - \hat{b}'\hat{b} - b'b + \hat{b}'b)a'a \\ &= \lim_{\rho \to 0} 4\rho(\hat{b}'b - b'b)a'a \\ &< 0 \end{aligned} \tag{6.19}$$

as long as $\hat{b} \neq b$.

So as long as the point $(\hat{a}, \hat{b})$ is different from (a, b) in both Case 1 and Case 2, we can always find certain points in its neighborhood with smaller cost function values. This contradicts with the assumption that $(\hat{a}, \hat{b})$ is a partial optimum point. Therefore, the true parameter (a, b) is the only partial optimum point of $J(\bar{a}, \bar{b})$.

Finally, based on Lemma 6.1, it is known that ACS converges to a partial optimum point. Thus the partial optimum is unique and corresponds to the true parameters (a, b) under normalization. Therefore we have $\lim_{N\to\infty} \hat{b} = b$ and $\lim_{N\to\infty} \hat{a} = a$ almost surely and the conclusion of this Theorem holds. □

Remark 6.4. *Note that $\hat{a}$ and $\hat{b}$ can be permuted as discussed in Remark 6.2. In*

employing the iterative algorithm, if the norm $\|\hat{b}\|_2$ is not fixed to $\|b\|_2$, the iteration sequence may diverge as explained below. Let $\hat{a}$ and $\hat{b}$ denote the current estimates of a and b at the k-th iteration. Without normalization, we cannot guarantee the condition in (6.16). The reason can be explained as follows. Since there may exist $\|\hat{b}\|_2 < \|b\|_2$ such that $\hat{b}'b > \hat{b}'\hat{b}$, which implies that there may exist certain directions along which the iteration point moves away from true parameter a while the cost function decreases. This explains why a counterexample could be provided in [49] and the reason why we fix the norm of $\hat{b}$.

6.5 Applications in the Identification of Block-oriented Systems

6.5.1 A New Class of Block-oriented Nonlinear Systems

As mentioned in Chapter 2, a Hammerstein-Wiener system can be represented by the following equations with nonlinear input and output functions:

$$\begin{aligned} z_k &= c_1 z_{k-1} + ... + c_n z_{k-n} + b_0 f(u_k) + ... + b_m f(u_{k-m}) + v_k \\ y_k &= g(z_k) \end{aligned} \tag{6.20}$$

where $c = [c_1 \ ... \ c_n]'$ and $b = [b_0 \ ... b_m]'$ are the unknown parameters, $\{u_k\}$ and $\{y_k\}$ are the input and output sequences respectively, $r = \max(n, m)$. The output nonlinear function $g(.)$ is an invertible function with its inverse function being $g^{-1}(.)$. Let the available input-output data be $\{u_k, y_k\}_{k=1-r}^{N}$. Even though Hammerstein-Wiener systems represent a fairly large class of systems in modeling practical nonlinear systems, we still feel that they are not sufficient to represent some more general nonlinear systems. This is why we consider the new class of

nonlinear systems generalized from the Hammerstein-Wiener systems in Chapter 2 as follows

$$g^{-1}(y_k) = c_1 g^{-1}(y_{k-1}) + ... + c_n g^{-1}(y_{k-n}) + b_0 f_0(u_k) + ... + b_m f_m(u_{k-m}) + v_k \tag{6.21}$$

In (6.21), if $f_0(.) = ... = f_m(.)$, the model reduces to a Hammerstein-Wiener system in (6.21). In addition, if $z = g(.) = y$ and $f_0(u) = ... = f_m(u) \neq u$, the model is a Hammerstein system; if $f_0(u) = ... = f_m(u) = u$ and $z = g(.) \neq y$, it becomes a Wiener system. If $f_0(u) = ... = f_m(u) = u$ and $z = g(.) = y$, the nonlinear system becomes a linear system.

To have a unique representation of system (6.21), we have the following assumptions.

Assumption 6.5. *The noise v_t is an i.i.d random variable with zero mean and finite variance σ_v^2. The input is also an i.i.d random variable such that $u_k \in U(-C, C)$.*

Assumption 6.6. *$f_i(u_k) = a_{i0}k_0(u_k) + ... + a_{il}k_l(u_k)$ where $k_0(.), ..., k_l(.)$ are orthornomal basis functions with $k_0(.) = 1$ and $E(k_j(.)) = 0$ ($1 \leq j \leq l$) in the interval $[-C, C]$. The inverse of the output nonlinear function $g(.)$ exists and can be represented by $z_k = g^{-1}(y_k) = h_0k_0(y_k) + h_1k_1(y_k) + ... + h_lk_l(y_k)$.*

Assumption 6.7. *Assume that $\|b\|_1 = 1$ where the first nonzero entry of b is positive, $h_1 = 1$ and $a_{i0} = 0$ for $i = 0, ..., m$.*

6.5.2 Transformation to the Common model

To apply the proposed NACS in Section 6.2, we need to transform (6.21) to the form of the common model in (6.1) and (6.2). Based on Assumptions 6.6 and 6.7

we have

$$\begin{aligned} k_1(y_k) &= -(h_2k_2(y_k) + ... + h_lk_l(y_k)) + c_0 + c_1k_1(y_{k-1}) + c_nk_1(y_{k-n}) + \\ &\quad + c_1(h_2k_2(y_{k-1}) + ... + h_lk_l(y_{k-1})) + ... + c_n(h_2k_2(y_{k-n}) + ... + h_lk_l(y_{k-n})) \\ &\quad + b_0(a_{01}k_1(u_k) + ... + a_{0l}k_l(u_k)) \\ &\quad + ... + b_m(a_{m1}k_1(u_{k-m}) + ... + a_{ml}k_l(u_{k-m})) + v_k \end{aligned} \tag{6.22}$$

where

$$c_0 = (\textstyle\sum_{i=1}^{n} c_i - 1)h_0 + \sum_{i=0}^{m} b_i a_{i0} \tag{6.23}$$

is the constant part.

Remark 6.5. *Note that the Legendre polynomials $k_0(u), .., k_j(u), ..., k_l(u)$ are well known orthonomal basis functions in the interval $[-1, 1]$ with $k_0(u) = 1$ and $E(k_j(u)) = 0$ where $0 \leq j \leq l$ denotes the order of each basis function. Legendre polynomial $k_j(u)$ can be produced using Rodrigues' formula: $k_j(u) = \frac{1}{2^k k!}\frac{d^k}{du^k}(u^2 - 1)^j$. Based on Legendre polynomials, it is easy to construct orthonomal basis functions in the interval $[-C, C]$ by the substitution $k_j(u) := k_j(\frac{u}{C})$ for $j = 0, ..., l$.*

Remark 6.6. *Note that the constant term c_0 is $(\sum_{i=1}^{n} c_i - 1)h_0 + \sum_{i=0}^{m} b_i a_{i0}$ in (6.23). Then it is impossible to estimate the coefficients of constant basis function $k_0(.)$ (h_0 and a_{i0} for $i = 0, ..., m$) from c_0. This is due to the existence of a constant deflection in identifying block-oriented systems such as Hammerstein-Wiener systems [59]. Note that scalar deflection also exists as seen in [59]. To avoid the constant deflection, we assume that $f_i(u)$ satisfies that $E(f_i(u)) = 0$ on $[-C, C]$, i.e, $a_{i0} = 0$ for $i = 0, ..., m$ in Assumption 6.7. Then we have $c_0 = (\sum_{i=1}^{n} c_i - 1)h_0$ and then $h_0 = c_0/(\sum_{i=1}^{n} c_i - 1)$, which implies the estimate $\hat{h}_0 = \hat{c}_0/(\sum_{i=1}^{n} \hat{c}_i - 1)$. To avoid scalar deflection, it is assumed that $\|b\|_1 = 1$ where the first nonzero entry of b is positive and $h_1 = 1$. Thus, a unique representation of the nonlinear system can be obtained. However, even when $E(f_i(u)) \neq 0$, c_0*

is still identifiable, although h_0 and a_{i0} cannot be identified. This means that the system (6.22) can still be identified although the constant terms cannot be separated between each static functions.

Denote $\{z_k\}_{k=k_0}^N$ as $\{z_k\}_{k=k_0}^N = [z_{k_0} \ \dots z_N]'$ where $k_0 \leq N$. In other words, $\{.\}_{k=k_0}^N$ is treated as a column vector. Let

$$\begin{aligned}
&Y = \{k_0(y_k)\}_{k=1}^N \\
&v = \{v_k\}_{k=1}^N \\
&Y_1 = \{-(h_2k_2(y_k) + ... + h_lk_l(y_k)) + c_0 + c_1k_1(y_{k-1}) + c_nk_1(y_{k-n})\}_{k=1}^N \\
&Y_2 = \{c_1(h_2k_2(y_{k-1}) + ... + h_lk_l(y_{k-1})) + ... + c_n(h_2k_2(y_{k-n}) + ... + h_lk_l(y_{k-n}))\}_{k=1}^N \\
&Y_3 = \{b_0(a_{01}k_1(u_k) + ... + a_{0l}k_l(u_k)) + ... + b_m(a_{m1}k_1(u_{k-m}) + ... + a_{ml}k_l(u_{k-m}))\}_{k=1}^N
\end{aligned} \tag{6.24}$$

We have

$$\begin{aligned}
Y &= Y_1 + Y_2 + Y_3 + v \\
&= \mathcal{G}d + \mathcal{K}_1\gamma_1 + \mathcal{K}_2\gamma_2 + v
\end{aligned} \tag{6.25}$$

where

$$\begin{aligned}
Y_1 &= \mathcal{G}d \\
Y_2 &= \mathcal{K}_1\gamma_1 \\
Y_3 &= \mathcal{K}_2\gamma_2 = b_0K_1\mathrm{a}_0 + ... + b_mK_m\mathrm{a}_m
\end{aligned} \tag{6.26}$$

and

$$
\begin{aligned}
\mathcal{G} &= \begin{bmatrix} -k_2(y_1) & ... & k_l(y_1) & 1 & k_1(y_0) & ... & k_1(y_{1-n}) \\ \vdots & ... & \vdots & \vdots & \vdots & ... & \vdots \\ -k_2(y_N) & ... & k_l(y_N) & 1 & k_1(y_{N-1}) & ... & k_1(y_{N-n}) \end{bmatrix} \in R^{N\times(n+l)} \\
d &= [h_2 \ ... \ h_l \ c_0 \ c_1 \ ... \ c_n]' \\
\mathcal{K}_1 &= \begin{bmatrix} k_2(y_0) & ... & k_l(y_0) & ... & k_2(y_{1-n}) & ... & k_l(y_{1-n}) \\ \vdots & ... & \vdots & ... & \vdots & ... & \vdots \\ k_2(y_{N-1}) & ... & k_l(y_{N-1}) & ... & k_2(y_{N-n}) & ... & k_n(y_{N-n}) \end{bmatrix} \in R^{N\times nl} \\
\gamma_1 &= [c_1h_2 \ ... \ c_1h_l \ ... \ c_nh_2 \ ... \ c_nh_l] \\
\mathcal{K}_2 &= \begin{bmatrix} k_1(u_1) & ... & k_l(u_1) & ... & k_1(u_{1-m}) & ... & k_l(u_{1-m}) \\ \vdots & ... & \vdots & ... & \vdots & ... & \vdots \\ k_1(u_N) & ... & k_l(u_N) & ... & k_1(u_{N-m}) & ... & k_l(u_{N-m}) \end{bmatrix} \in R^{N\times(l+1)m} \\
\gamma_2 &= [b_0\mathrm{a}_0 \ ... \ b_m\mathrm{a}_m] = [b_0a_{01} \ ... \ b_0a_{0l} \ ... \ b_ma_{m1} \ ... \ b_ma_{ml}]' \\
\mathrm{a}_i &= [a_{i1} \ ... \ a_{il}]', \ i = 0, ..., m
\end{aligned}
\tag{6.27}
$$

Note that when the elements in d are known, γ_1 is also obtained as in (6.27), namely $\gamma_1 = r(d)$. We have

$$L(d) = \mathcal{G}d + \mathcal{K}_1\gamma = \mathcal{G}d + \mathcal{K}_1 r(d) \tag{6.28}$$

Then the common model can be obtained as

$$Y = \mathcal{G}d + \mathcal{K}\gamma + v \tag{6.29}$$

$$Y - L(d) = b_0K_1a_0 + ... + b_mK_ma_m + v \tag{6.30}$$

where

$$K_i = \begin{bmatrix} k_0(u_{1-i}) & ... & k_l(u_{1-i}) \\ \vdots & ... & \vdots \\ k_0(u_{N-i}) & ... & k_l(u_{N-i}) \end{bmatrix} \in R^{N\times(l+1)}, i = 0, 1, ..., m \tag{6.31}$$

Assumption 6.8. *$[\mathcal{G}\ \mathcal{K}]$ in (6.29) is the same as in Assumption 6.3.*

Remark 6.7. *The model in (6.29) and (6.30) are identical to the common model (6.1) and (6.2). So the convergence property of the estimates of (6.29) and (6.30) can be achieved provided that Assumptions 6.1-6.4 are all satisfied. Note that Assumptions 6.1-6.3 can be directly achieved under Assumptions 6.5 - 6.8. We now focus on how the condition in Assumption 6.4 is satisfied.*

Let $K = [K_1\ ...\ K_m]$ where K_i for $i = 0, ..., m$ is defined in (6.31) and we have the following Lemma.

Lemma 6.3. *Under Assumption 6.5, K can be constructed such that $\lim_{N\to\infty} \frac{K'K}{N} = I$ almost surely.*

Proof. As the basis functions $k_0(.), k_1(.), ..., k_l(.)$ are orthonormal on the interval $[-C, C]$, we have $\int_{-C}^{C} k_i(u_t)k_j(u_t)du_t = \delta(i-j)$ where $\delta(.)$ is 1 if and only if $i = j$, otherwise, $\delta(i-j) = 0$. Note that $k_0(.) = 1$ and then $\int_{-C}^{C} k_j(u_t)du_t = \delta(j)$. This means $k_i(u_t)$ and $k_j(u_t)$ $(i, j > 0$ and $i \neq j)$ are independent variables with zero mean and variance 1 on the interval $[-C, C]$ under Assumptions 6.5. It is known that as long as u_t and $u_{\tilde{t}}$ $(t \neq \tilde{t})$ are i.i.d, $k_j(u_t)$ and $k_j(u_{\tilde{t}})$ for $1 \leq j \leq l$ are also i.i.d. Then $k_j(u_t)$ and $k_j(u_{\tilde{t}})$ are independent variables with zero mean and variance 1 on $[-C, C]$. Thus all elements in K are random variables with zero mean and variance 1. So we have $\lim_{N\to\infty} \frac{K'K}{N} = I$ almost surely. □

Remark 6.8. *Note that almost surely means that an event occurs with probability* 1. *In Lemma 6.3, it is possible that* K *is a singular matrix in one realization for a particular sequence* $\{u_t\}$ *but the measure of such a sequence is zero, namely, such an event occurs with probability* 0.

6.6 Discussion of Model Generalization

As seen in system model (6.21), f_i is assumed as $f_i(u_k) = a_{i0}k_0(u_k) + ... + a_{il}k_l(u_k)$. So f_i can be different functions for different i while there is only one output function $g(.)$. In this section, we discuss whether we can generalize the system model in (6.21) to the following two cases.

Case 1: Is it possible that there are multiple static functions on the output block?

In this case, we investigate whether the system model in (6.21) can be generalized to

$$g_0^{-1}(y_k) = c_1 g_1^{-1}(y_{k-1}) + ... + c_n g_n^{-1}(y_{k-n}) + b_0 f_0(u_k) + ... + b_m f_m(u_{k-m}) + v_k \tag{6.32}$$

i.e, there are $n+1$ output functions to be estimated.

Remark 6.9. *To guarantee the convergence of the estimates in (6.21), the i.i.d assumption of* $\{u_k\}$ *is required as shown in Assumption 6.5 to make sure that* $\lim_{N\to\infty} \frac{K'K}{N} = I$ *in Lemma 6.3. Note that the i.i.d of* u_k *and* u_{k-1} *can be satisfied in designing the input sequence. However, we notice that the outputs* y_k *and* y_{k-1} *are dependent due to the dependence of* z_k *and* z_{k-1}. *So we cannot guarantee the convergence of the estimates in (6.32) in this case. Then the system in model (6.21) cannot be generalized to Case 1 by using our proposed method.*

However, Remark 6.9 actually points out a potential future research topic, i.e, how to construct a matrix K such that $\lim_{N\to\infty} \frac{K'K}{N} = I$ based on the output observation sequence $\{y_k\}$ that is not an i.i.d sequence. If one can achieve this, then the system model can be generalized to Case 1.

Case 2: Is it possible that $f_i(.)$ is allowed a general nonlinear function in the input block?

For a general function $f_i(u_k)$ in the interval $[-C, C]$, we assume that $f_i(u_k) = a_{i0}k_0(u_k) + ... + a_{il}k_l(u_k) + \varepsilon_k$ where ε_k denotes the approximation error at u_k. To avoid constant deflection, assume that $a_{i0} = 0$. In this case, (6.29) and (6.30) become

$$Y = \mathcal{G}d + \mathcal{K}\gamma + \varepsilon + v \tag{6.33}$$

$$Y - L(d) = b_0K_1a_0 + ... + b_mK_ma_m + \tilde{\varepsilon} + v \tag{6.34}$$

where $\varepsilon = [\varepsilon_1 \ ... \ \varepsilon_N]'$ and $\tilde{\varepsilon} = [\tilde{\varepsilon}_1 \ ... \ \tilde{\varepsilon}_N]'$ are the approximation error vectors. We know that the variance of ε_k and $\tilde{\varepsilon}_k$ are dependent on N denoted as $D(\varepsilon_k) = \sigma^2_\varepsilon(N)$ and $D(\tilde{\varepsilon}_k) = \sigma^2_{\tilde{\varepsilon}}(N)$, respectively. Note that $\sigma^2_\varepsilon(N)$ and $D(\tilde{\varepsilon}_k)$ approach zero almost surely as the number of basis functions approaches infinity. Then (6.33) and (6.34) will tend to (6.29) and (6.30) as $N \to \infty$ and the proposed algorithm can be applied. Even if $a_{i0} \neq 0$, we can still identify the system without separating the constant terms in all the static functions as discussed in Remark 6.6. Then it is possible to generalize $f_i(.)$ to a general nonlinear function.

6.7 Simulation Results

For the illustration of our proposed method, we consider the following block-oriented system which is more general than a Hammerstern-Wiener system.

$$\begin{aligned} g^{-1}(y_k) &= 0.4g^{-1}(y_{k-1}) + 0.1g^{-1}(y_{k-2}) \\ &\quad + 0.6f_0(u_k) + 0.3f_1(u_{k-1}) + 0.1f_1(u_{k-2}) + v_k \\ f_0(u) &= 0.9k_1(u_k) + 0.8k_2(u_k) + 0.7k_3(u_k) \\ f_1(u) &= 0.6k_1(u_{k-1}) + 0.5k_2(u_{k-1}) + 0.4k_3(u_{k-1}) \\ f_2(u) &= 0.3k_1(u_{k-2}) + 0.2k_2(u_{k-2}) + 0.1k_3(u_{k-2}) \\ g^{-1}(y) &= 0.2k_0(y_k) + k_1(y_k) + 0.2k_2(y_k) + 0.3k_3(y_k) \end{aligned}$$

where v_k is zero mean with $\sigma_v^2 = 0.1$ and $k_i(.)$ for $i = 0, 1, ..., 3$, are the Legendre polynomial functions (as seen in Remark 6.5) which are $k_0(u) = 1$, $k_1(u) = u$, $k_2(u) = \frac{1}{2}(3u^2 - 1)$ and $k_3(u) = \frac{1}{2}(5u^3 - 3u)$. By transforming the above model into the common model in (6.29) and (6.30), the true parameters in the common model are given as $d = [h_2\ h_3\ c_0\ c_1\ c_2]' = [0.2\ 0.3\ -0.1\ 0.4\ 0.1]'$, $b = [b_0\ b_1\ b_2]' = [0.6\ 0.3\ 0.1]'$ and $\mathrm{a} = [\mathrm{a}_0\ \ \mathrm{a}_1\ \mathrm{a}_2] = [a_{01}\ a_{02}\ a_{03}\ a_{11}\ a_{12}\ a_{13}\ a_{21}\ a_{22}\ a_{23}] = [0.9\ 0.8\ 0.7\ 0.6\ 0.5\ 0.4\ 0.3\ 0.2\ 0.1]'$ (Note that $c_0 = (\sum_{i=1}^{2} c_i - 1)h_0$ and $h_1 = 1$). To consist with the assumption, we have $\|b\|_1 = 1$ and input sequence is designed such that it is uniformly distributed in the interval $[-1, 1]$ ($C = 1$). In order to estimate both the parameters and nonlinear functions, we choose $N = 1000$ and the steps of identifying the above model are summarized as follows:

1) Collect the output sequence $\{y_k\}_{k=-1}^{N}$ based on the white noise input sequence $\{u_k\}_{k=-1}^{N} \in [-1, 1]$.

2) Construct observation vector $Y = [k_1(y_1)\ ...\ k_1(y_N)]'$, $\mathcal{G}$ and $\mathcal{K}$ based on (6.27), $K = [K_0\ K_1\ ...\ K_2]$ based on (6.31) by using the input and output

sequences. We also have $L(\hat{d})$ given in (6.28). Then, we obtain the common model in (6.29) and (6.30), which are identical to (6.1) and (6.2).

3) Initialize and employ the NACS algorithm in Section 5.3.2 to obtain estimates of parameters in the common model (6.1) and (6.2).

The initial values of the estimates are arbitrarily given. By employing the above identification procedure, we obtain the estimates as

$$\begin{aligned}\hat{d} &= [0.2010\ 0.3008\ -0.9890\ 0.3983\ 0.1007]'\\ \hat{b} &= [0.6021\ 0.2986\ 0.9993]'\\ \hat{\mathrm{a}} &= [0.8990\ 0.8015\ 0.7010\ 0.6012\ 0.4995\ 0.4007\ 0.3006\ 0.1896\ 0.1004]'\end{aligned}$$

and $\hat{h}_0 = \hat{c}_0/(\sum_{i=1}^{2}\hat{c}_i - 1) = 0.2011$. Actually, estimation error of NACS depends on the number of data points N. Let the l_1 norm of error be $\|e\|_1 = \|\hat{a} - a\|_1 + \|\hat{b} - b\|_1 + \|\hat{d} - d\|_1$. Figure 6.2 shows how the l_1 norm of the error changes with the number of data points N. We can see that the error converges to zero and thus gives us a satisfactory result.

6.8 Conclusion

In this chapter, we focus on the biconvex optimization in the identification of block-oriented systems. We propose a common model on the basis of the new class of block-oriented nonlinear systems introduced in Chapter 2. It is shown that identifying the common model can be formulated as a biconvex optimization problem with a suitable biconvex cost function. Such a biconvex optimization problem only needs to find the unique partial optimum point of its biconvex cost function on a convex set. The normalized alterative convex search (NACS) algorithm is proposed with the guaranteed convergence property. This provides a

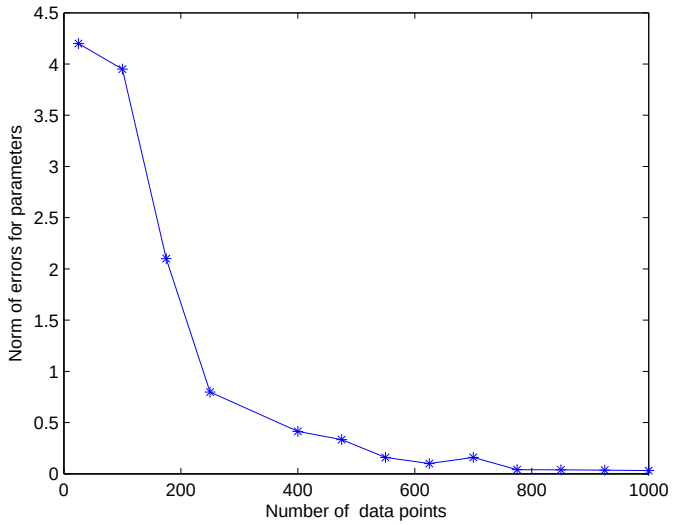

Figure 6.2: The change of error for the estimated parameters respect to N

unified framework for the identification of block-oriented systems. On the other hand, we also solve the problem that Hammerstein or Winener systems as well as Hammerstein-Winner systems need a proper initialization as pointed out in [49] [48] [80].

Chapter 7

Identification of Wiener Systems with Clipped Observations

In previous chapters, we consider identification of block-oriented nonlinear systems based on designed input data and directly measured output data. While in some cases, the system outputs need to be transmitted by binary sensors. This leads to clipped outputs. In this chapter, we focus on the parametric version of Wiener systems where both the linear and nonlinear parts are identified with clipped observations in the presence of internal and external noises. Also the static functions are allowed non-invertible. We propose a classification based SVM and formulate the identification problem as a convex optimization. The solution to the optimization problem converges to the true parameters of the linear system if it is an FIR system, even though clipping reduces a great deal of information about the system characteristics. In identifying a Wiener system with a stable IIR system, an FIR system is used to approximate it and the problem is converted to identify the FIR system together with solving a set of nonlinear equations. This leads to biased estimates of parameters in the IIR system while the bias could be controlled by

choosing the order of the approximated FIR system.

7.1 Introduction

A Wiener system consists of a linear dynamic system (FIR or IIR) followed by a nonlinear static function and its identification has been well studied in [27] [85] [30] [56] and [86]. In engineering, binary-valued sensor is commonly used for its convenience such as ease for transmitting the measured signals in communication systems compared with conventional sensors [86]. Binary-valued sensor produces output clipping. When the output of a Wiener system is observed through a binary-valued sensor, it can be represented as a Wiener system with clipped output observations [87]. In this chapter, we concern the parametric version of such Wiener systems where both the linear and nonlinear parts are identified in the presence of internal and external noises. The new identification approach is achieved by using support vector machines (SVM) [88]-[82] from classification point of view.

It is noted that SVM and LS-SVM in identifying Wiener systems is not new, especially LS-SVM has become a very powerful tool in both Wiener systems and Hammerstern-Wiener systems identification [90] [91] [40] [26]. However, these methods are based on SVM for regression instead of classification. Though SVM for classification and regression are similar, they play very different roles in Wiener systems identification. For example, as seen in the LS-SVM for regression based methods [91] [40], in order to identify the parameters in the linear system, the static function is required to be invertible, as this enables the output of the linear system to be represented as a function of the output of the static function. LS-SVM for regression based method becomes unapplicable if the static function in a Wiener system is non-invertible. Note that a clipped observer actually denotes

a non-invertible function. Thus, LS-SVM for regression based method cannot be applied here to identify the parameters in the linear part of the Wiener system with clipped observations. In fact, there are only few papers addressing non-invertible functions in Wiener systems by using nonparametric approaches, see, for example, [27][78]. As pointed out in [86] [92], clipped (binary-valued) observations provide very limited information on the system output and hence introduce difficulties in system modeling, identification, and control.

The main contribution of this chapter is to propose a classification based SVM approach so that the parameters in the linear system can be estimated even when the static function is non-invertible, the output observations are clipped and in the presence of both internal and external noises. The consistency of the estimated parameters is established when the linear system is FIR. When the linear system is a stable IIR system, an FIR system can be used to approximate it and the problem is converted to identifying the FIR system together with solving a set of nonlinear equations. Based on trust region algorithm [93], a scheme is developed to solve the nonlinear equations. The approximation leads to biased estimates of parameters in the IIR system while the bias could be controlled by choosing the approximation order of the FIR system. It is also worth pointing out that only sign information of system outputs, which is equivalent to clipped observations, is considered in schemes [86][94] [95] [54] [97]. However, our ideas and approaches are different from these schemes. For example, the idea in [86] is to decompose the identification problem into a finite number of core identification and only FIR linear models are considered.

7.2 Problem Formulation with Two-classes Classification SVM

Wiener systems which belong to the class of block-oriented [56] nonlinear systems can be represented as

$$\begin{aligned} z_k &= a_1 z_{k-1} + ... + a_n z_{k-n} + b_0 u_k + ... + b_m u_{k-m} + \eta_k, \\ y_k &= g(z_k) + e_k = c_1 k_1(z_k) + ... + c_h k_h(z_k) + e_k \end{aligned} \tag{7.1}$$

where u_k and y_k are the system input and output, η_k and e_k denote the internal and external noises, integers n and m denote the known system order, $g(.)$ is a static function with unknown parameters $c = [c_1...c_h]^T$ but known function basis $k_1(.), ..., k_h(.)$, $a = [a_1...a_n]^T$ and $b = [b_0...b_m]^T$ are vectors with unknown parameters in the linear system, and $z_1, ..., z_{r-1} = 0$ where $r = min(m, n) + 1$. A Wiener system with clipped observations is shown in Figure 7.1. We have

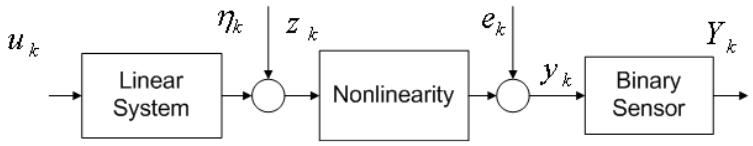

Figure 7.1: Wiener systems with quantized observations

$Y_k = sgn[y_k - C]$ with C being a known threshold and is assumed to be 0 in this chapter.

Remark 7.1. *As $Y_k = sgn(y_k - C)$, all moments of Y_k depend merely on the distribution of y_k. C is a threshold such that $E(y_k) = 1 - 2P_y(C) = 0$ and $var(y_k) = 1 - (1 - 2P_y(C))^2$, i.e., $P_y(C) = \frac{1}{2}$ where $P_y(C)$ is the probability of the event $y_k \leq C$.*

Assumption 7.1. *$g(z)$ satisfies that $z(g(z) - C) \geq 0$.*

Assumption 7.2. *Input u_k is a symmetric i.i.d process and the noises $\eta_k \in [-d, d]$, $e_k \in [-e, e]$ are bounded white noises.*

Assumption 7.3. *The linear system is stable. $b = [b_0, ..., b_m]^T \in B$ where B is a bounded set and $b_0 = 1$.*

Note that $b_0 = 1$ is to obtain a unique representation of the system. Our objective is to estimate a, b and c only using the input and the signs of the output. We first formulate the identification problem for an FIR system ($a = [0 \ ... \ 0]^T$), which will be extended to IIR systems in Section 7.4. Under Assumption 7.1 , we have

$$Y_k = sgn(g(b_0u_k + ... + b_mu_{k-m})) = sgn(b_0u_k + ... + b_mu_{k-m}) \tag{7.2}$$

for the noise free case. Note that $b_0u_k + ... + b_mu_{k-m} = 0$ denotes a hyperplane in R^{m+1}. Let $\mathrm{x}_k = (u_k, ..., u_{k-m})^T \in \mathcal{X} = R^{m+1}$ and $Y_k = sgn(y_k) \in \mathcal{Y} = \{-1, 1\}$. We identify the system using methodologies in classification based on N pairs of training data $T_N = \{(\mathrm{x}_1, Y_1), ...(\mathrm{x}_N, Y_N)\}$. As $\{\mathrm{x}_k\}$ is an overlapping sequence, z_k and $z_{k+m'}$ are dependent if $m' \leq m$. But as long as $m' > m$, z_k and $z_{k+m'}$ become independent, i.e, y_k and $y_{k+m'}$ become independent. Let $m' = m + 1$ and $N = lm'$. Through sampling, we can obtain i.i.d l pairs of training data $T_l = \{(\mathrm{x}_1, Y_1), (\mathrm{x}_{2m'-m}, Y_{2m'-m}), ..., (\mathrm{x}_{lm'-m}, Y_{lm'-m})\} \in \{\mathcal{X} \times \mathcal{Y}\}$. For simplicity of expression, the i.i.d T_l is written as $T_l = \{(\mathrm{x}_1, Y_1), ..., (\mathrm{x}_l, Y_l)\}$. For a given parameter b one can define a hyperplane $h(\mathrm{x}) = \mathrm{x}^T b = 0$. Then the classification rule is based on the following decision function

$$H(\mathrm{x}) = sgn(h(\mathrm{x})) = sgn(\mathrm{x}^T b) \tag{7.3}$$

i.e, if $H(\mathrm{x}) = 1$, x is determined to belong to class +1; otherwise it is determined

to belong to class -1.

Remark 7.2. *As original training set T_N is transformed into T_l, we actually reduce the size of data from N to l being of length N/m'. The remaining of data is wasted. Based on Lemma 12.5 in [56], T_N can be divided into m' blocks of training data set $T_l^1, ..., T_l^{m'}$ such that the data points are i.i.d in each block. So the data belonging to each block can all be used to estimate b. This process will give m' estimates of b. The final estimation of b can be obtained by taking the average of these estimates. Possibly we can obtain a better estimate with lower variance.*

Definition 7.1. *0-1 loss function is given by $c(Y, H(\mathrm{x})) = 0$ when $Y = H(\mathrm{x})$ and $c(Y, H(\mathrm{x})) = 1$ when $y \neq H(\mathrm{x})$.*

Definition 7.2. *Empiric risk: The empiric risk of a decision function $H(\mathrm{x}) = sgn(\mathrm{x}^T b)$ is defined as $R_l^{emp}(b) = \frac{1}{l}\sum_{k=1}^{l} c(\mathrm{x}_k, Y_k, H(\mathrm{x}_k)) = \frac{1}{l}\sum_{k=1}^{l} c(\mathrm{x}_k, Y_k, sgn(\mathrm{x}_k^T b))$ based on the l pairs of training data in T_l.*

Definition 7.3. *Expected risk: Suppose $P(\mathrm{x}, y)$ is a probability distribution on $\mathcal{X} \times \mathcal{Y}$. The expected risk of $H(\mathrm{x}) = sgn(\mathrm{x}^T b)$ is defined as $R(b) = E(c(y, H(\mathrm{x}))) = E(c(y, sgn(\mathrm{x}^T b))) = \int_{\mathcal{X}\times\mathcal{Y}} c(y, sgn(\mathrm{x}^T b)))dP(\mathrm{x}, y)$.*

The identification problem is formulated as a classification problem of finding a decision function $H(\mathrm{x})$ such that a suitably defined risk is minimized. In the presence of noise, we cannot correctly classify all the data points such that $\forall k, Y_k \mathrm{x}_k^T b \geq 1$. Then slack variables $\xi = (\xi_1 ... \xi_k ... \xi_l)$ are introduced such that $\forall k,\ Y_k \mathrm{x}_k^T b \geq (1 - \xi_k)$. Here we apply classification SVM based approach to minimize a penalized risk as follows

$$min_{\{b,\xi\}}\ \tfrac{1}{2} b^T b + \gamma \sum_{k=1}^{l} \xi_k \quad s.t \quad Y_k \mathrm{x}_k^T b \geq (1 - \xi_k), \quad \xi_k \geq 0 \tag{7.4}$$

where γ is a penalized factor that plays the role of regularization. By minimizing

the empirical risk through choosing a proper penalized factor γ, (7.4) is to determine a decision function which is restricted to the class of linear planes. With (7.4), a quadratic function is minimized under linear inequality constraints. So it is a convex optimization problem. Based on SVM, the estimate denoted as $\hat{b}_l^{svm}$ which minimizes the cost function in (7.4) is

$$\hat{b}_l^{svm} = \sum_{k=1}^{l} \alpha_k^* Y_k \mathrm{x}_k \tag{7.5}$$

where α_k^*, $k = 1, ...l$ can be easily and uniquely obtained by solving the dual problem of (7.4) as in [88]-[82]. We know that very few α_k^* are non-zero. The corresponding data points with nonzero α_k^* are called support vectors.

Definition 7.4. *Define* $R_l^{svm}(b) = \frac{1}{l}\sum_{k=1}^{l}\xi_k = \frac{1}{l}\sum_{k=1}^{l} max(0, 1 - Y_k b^T \mathrm{x}_k)$ *which is the risk term in (7.4).*

If the VapnikChervonenkis (VC) dimension of the hyperplane $h(\mathrm{x})$ is fixed, SVM tends to minimize $R_l^{svm}(b)$ as l becomes large. Let $\hat{b}_l^{svm}$ and $\hat{b}_l^{emp}$ be the minimum points of $R_l^{svm}(b)$ and $R_l^{emp}(b)$, respectively. Then, based on [98], we have the following Lemma.

Lemma 7.1. $lim_{l\to\infty}\hat{b}_l^{svm} = lim_{l\to\infty}\hat{b}_l^{emp}$.

Proof. This result is the conclusion of [98]. In [98], it is concluded that by minimizing $R_l^{svm}(b)$, one also indirectly minimizes the empirical risk $R_l^{emp}(b)$ (classification error) as $l \to \infty$. They both converge to the optimal Bayes error. □

Remark 7.3. *Let* b^* *be the true parameter of* b. *Our aim is to show that* $\lim_{l\to\infty}\hat{b}_l^{svm} = b^*$. *Based on Lemma 7.1, this can be ensured if we can show that* $\lim_{l\to\infty}\hat{b}_l^{emp} = b^*$, *In next section, we will analyze how* $\hat{b}_l^{emp}$ *converges to* b^*.

7.3 Convergence Analysis

Lemma 7.2. *([88]) Assume that the data set T_l is i.i.d. For an arbitrary $b \in B$, $R_l^{emp}(b)$ and $R(b)$ can be defined. Let $Q(l) = sup_{b\in B}|R(b) - R_l^{emp}(b)|$. We have $Q(l) \to 0$ almost surely (a.s) when $l \to \infty$.*

Proof. The proof follows directly from Theorem 5.3 in [88]. Note that $R_l^{emp}(b)$ and $R(b)$ in this chapter are considered as $R_l^{emp}(f)$ and $R(f)$ in [88] , respectively. In [88], it is shown that $P(Q(l) > \varepsilon) \leq 4V_{ch}(l)e^{(-\frac{l\varepsilon^2}{8})}$ where $V_{ch}(l)$ denotes the VC dimension of the hyperplane $h(\mathrm{x})$ in the decision function. In this chapter, $h(\mathrm{x})$ is a linear plane with dimension $m+1$, so $V_{ch}(l)$ is fixed as $V_{ch}(l) = m+2$. Then $P(Q(l) > \varepsilon) \leq 4(m+2)e^{-\frac{l\varepsilon^2}{8}}$. Obviously, $\sum_{l=1}^{\infty} P(Q(l) > \varepsilon) = 4(m+2)\frac{e^{-\varepsilon^2/8}}{1-e^{-\varepsilon^2/8}} < \infty$. Thus we can conclude that $Q(l) \to 0$ almost surely (a.s) when $l \to \infty$ by Bore-Cantelli Lemma. □

Remark 7.4. *The necessary and sufficient condition of $Q(l) \to 0$ as $l \to \infty$ almost surely is presented in the Theorem of [99] and [100]. The condition is that the VC dimension of the decision function $H(\mathrm{x}) = sgn(h(\mathrm{x}))$ denoted as $V_{ch}(l)$ satisfies that $lim_{l\to\infty}\frac{V_{ch}(l)}{l} = 0$ almost surely [99] [100]. On the other hand, in [82], it is also shown that upper bound $Q(l)$ is a decreasing function of $\|b\|$. This explains why we minimize b^Tb in (7.4). Lemma 7.2 basically shows that $R_l^{emp}(b)$ converges to $R(b)$ as $l \to \infty$ (a.s). To show $\hat{b}_l^{emp}$ converging to b^*, we only need to prove the minimizer of $R(b)$ corresponds to the true parameter b^*. Since the expressions of $R(b)$ are different for the noise free case and noise cases. We analyze these cases separately.*

7.3.1 Convergence Analysis for Noise Free Case

In this case, the two-class data are linearly classifiable. As shown in Figure 7.2(a), $h(\mathrm{x}) = x^T b = 0$ is a decision hyperplane and $h^*(\mathrm{x}) = \mathrm{x}^T b^* = 0$ is the true hyperplane. Let S be the total space and S_1, S_2 be the regions between $h(\mathrm{x}) = 0$ and $h^*(\mathrm{x}) = 0$, and A, A_1 and A_2 be the volume of S, S_1 and S_2, respectively. Obviously S_1 and S_2 denote the sets which cannot be correctly classified by $H(\mathrm{x})$. $S_1 \cup S_3$ is the region above $h^*(\mathrm{x}) = 0$ denoting the region where the first class data points locate, while $S_2 \cup S_4$ below $h^*(\mathrm{x}) = 0$ denotes the region where the second class data points locate. In space $\mathcal{X} \times \mathcal{Y}$, assume that the probability density

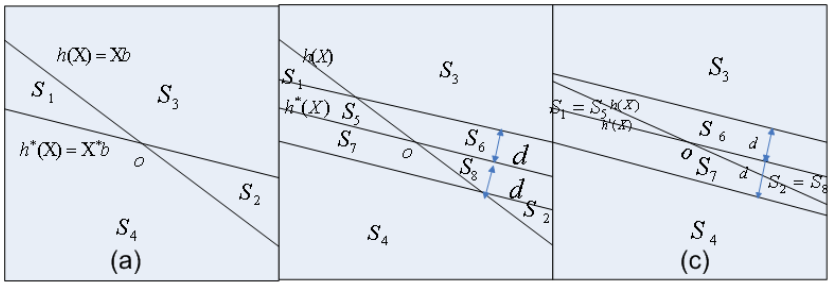

Figure 7.2: Possible misclassification region when $h(\mathrm{x})$ is not close to $h^*(\mathrm{x})$

function $p(\mathrm{x}, y)$ is given as

$$
\begin{aligned}
p(\mathrm{x},1) \quad &= \begin{cases} \frac{\mathbf{p}(\mathrm{x})}{A} & h^*(\mathrm{x}) \geq 0,\ i.e, x \in S_1 \cup S_3 \\ 0 & h^*(\mathrm{x}) \leq 0,\ i.e, x \in S_2 \cup S_4 \end{cases} \\
p(\mathrm{x},-1) \quad &= \begin{cases} 0 & h^*(\mathrm{x}) \geq 0,\ i.e, \mathrm{x} \in S_1 \cup S_3 \\ \frac{\mathbf{p}(\mathrm{x})}{A} & h^*(\mathrm{x}) \leq 0\ i.e, \mathrm{x} \in S_2 \cup S_4 \end{cases}
\end{aligned}
$$

where $\mathbf{p}(\mathrm{x})$ is a probability density function such that $\int_{\mathrm{x}\in\mathcal{X}}(p(\mathrm{x},1)+p(\mathrm{x},-1))d\mathrm{x} = 1$, $p(Y=1) = \int_{h^*(\mathrm{x})\geq 0} \frac{\mathbf{p}(\mathrm{x})}{A} dx = \frac{1}{2}$, $p(Y=-1) = \int_{h^*(\mathrm{x})\leq 0} \frac{\mathbf{p}(\mathrm{x})}{A} dx = \frac{1}{2}$ and $p(y = 1|h^*(\mathrm{x}) \geq 0) = p(y=-1|h^*(\mathrm{x}) < 0) = 1$.

Lemma 7.3. *$R(b)$ is an increasing function of A_1 and A_2.*

Proof. Note that

$$\begin{aligned} R(b) &= \int_{\mathcal{X}\times\mathcal{Y}} c(Y, H(\mathrm{x}))d(P(\mathrm{x}, y)) \\ &= \int_{\mathcal{X}} c(-1, H(\mathrm{x}))p(\mathrm{x}, -1)d\mathrm{x} + \int_{\mathcal{X}} c(-1, H(\mathrm{x}))p(\mathrm{x}, 1)d\mathrm{x} \end{aligned}$$

As $H(\mathrm{x}) = 1$ if $x \in S_2 \cup S_3$ and $H(\mathrm{x}) = -1$ if $\mathrm{x} \in S_1 \cup S_4$. Then

$$\begin{aligned} R(b) &= \int_{\mathrm{x}\in S_1} c(-1,-1)0dx + \int_{\mathrm{x}\in S_2} c(-1,1)\tfrac{\mathbf{p}(\mathrm{x})}{A}dx + \int_{\mathrm{x}\in S_3} c(-1,1)0dx \\ &\quad + \int_{x\in S_4} c(\mathrm{x},-1,-1)\tfrac{\mathbf{p}(\mathrm{x})}{A}dx + \int_{\mathrm{x}\in S_1} c(1,-1)\tfrac{\mathbf{p}(\mathrm{x})}{A}dx \\ &\quad + \int_{\mathrm{x}\in S_2} c(1,1)0dx + \int_{\mathrm{x}\in S_3} c(1,1)\tfrac{\mathbf{p}(\mathrm{x})}{A}dx + \int_{\mathrm{x}\in S_4} c(1,-1)0dx \\ &= \int_{x\in S_2} \tfrac{\mathbf{p}(\mathrm{x})}{A}dx + \int_{x\in S_1} \tfrac{\mathbf{p}(\mathrm{x})}{A}dx \end{aligned}$$

Thus, $R(b)$ is an increasing function of A_1 and A_2. □

Theorem 7.1. *The estimate $\hat{b}_l^{svm}$ in (7.5) satisfies that $\lim_{l\to\infty} \hat{b}_l^{svm} = \lim_{l\to\infty} \hat{b}_l^{emp} = b^*$ (a.s) for the noise free case.*

Proof. From Lemma 7.2 , we have $\lim_{l\to\infty} R^{emp}(b) = R(b)$ (a.s). From Lemma 7.3, $R(b)$ obtains its minimum point if and only if $A_1, A_2 \to 0$ (a.s), which gives $\lim_{l\to\infty} \hat{b}_l^{emp} = b^*$ (a.s). Based on Lemma 7.1, we have $\lim_{l\to\infty} \hat{b}_l^{svm} = b^*$ (a.s). □

7.3.2 Convergence Analysis in the Presence of Noise

Only e_k is considered

In this case, $y_k = g(z_k) + e_k = \tilde{g}(\mathrm{x}_k^T b) + e_k$, the probability density function $p(\mathrm{x}, y)$ is given as:

$$
\begin{aligned}
p(\mathrm{x},1) \quad &= \begin{cases} \frac{\mathbf{p}(\mathrm{x})(1-P_1(\mathrm{x}))}{A} & \mathrm{x} \in S_1 \cup S_3 \\ \frac{\mathbf{p}(\mathrm{x})P_2(\mathrm{x})}{A} & \mathrm{x} \in S_2 \cup S_4 \end{cases} \\
p(\mathrm{x},-1) \quad &= \begin{cases} \frac{\mathbf{p}(\mathrm{x})(1-P_2(\mathrm{x}))}{A} & \mathrm{x} \in S_2 \cup S_4 \\ \frac{\mathbf{p}(\mathrm{x})P_1(\mathrm{x})}{A} & \mathrm{x} \in S_1 \cup S_3 \end{cases}
\end{aligned}
$$

where

$$
\begin{aligned}
P_1(\mathrm{x}) = \textstyle\int_{e_k+\tilde{g}(x)<0} de_k &= \begin{cases} \frac{e-\tilde{g}(\mathrm{x})}{2e} & 0 \le \tilde{g}(\mathrm{x}) < e \\ 0 & \tilde{g}(\mathrm{x}) \ge e \\ 0 & \tilde{g}(x) < 0 \end{cases} \\
P_2(\mathrm{x}) = \textstyle\int_{e_k+\tilde{g}(\mathrm{x})>0} de_k &= \begin{cases} \frac{e+\tilde{g}(\mathrm{x})}{2e} & -e < \tilde{g}(\mathrm{x}) \le 0 \\ 0 & \tilde{g}(\mathrm{x}) \le -e \\ 0 & \tilde{g}(\mathrm{x}) > 0 \end{cases}
\end{aligned}
$$

Here $P_1(\mathrm{x})$ is the probability that x locates in the region below $h^*(\mathrm{x}) = 0$ but $y = 1$, $P_2(\mathrm{x})$ is the probability that x locates in the region above $h^*(\mathrm{x}) = 0$ but $y = -1$. Actually, $P_1(\mathrm{x})$ and $P_2(\mathrm{x})$ are dependent on $\tilde{g}(.)$ and satisfy that

$\int_{x\in\mathcal{X}}(p(\mathrm{x},1)+p(\mathrm{x},-1))d\mathrm{x}=1$. As shown in Lemma 7.3, we have

$$\begin{aligned}R(\hat{b}) &= \int_{x\in S_2}\frac{\mathbf{p}(\mathrm{x})(1-p_2(\mathrm{x}))}{A}d\mathrm{x}+\int_{\mathrm{x}\in S_3}\frac{\mathbf{p}(\mathrm{x})p_1(\mathrm{x})}{A}d\mathrm{x}\\ &\quad+\int_{\mathrm{x}\in S_1}\frac{\mathbf{p}(\mathrm{x})(1-p_1(\mathrm{x}))}{A}d\mathrm{x}+\int_{\mathrm{x}\in S_4}\frac{\mathbf{p}(\mathrm{x})p_2(x)}{A}d\mathrm{x}\\ &= \int_{\mathrm{x}\in S_2}\frac{\mathbf{p}(\mathrm{x})(1-2P_2(x))}{A}d\mathrm{x}+\int_{\mathrm{x}\in S_1}\frac{\mathbf{p}(\mathrm{x})(1-2P_1(\mathrm{x}))}{S}d\mathrm{x}\\ &\quad+\int_{\mathrm{x}\in S}\frac{\mathbf{p}(\mathrm{x})P_1(\mathrm{x})}{A}d\mathrm{x}+\int_{\mathrm{x}\in S}\frac{\mathbf{p}(\mathrm{x})P_2(\mathrm{x})}{A}d\mathrm{x}\end{aligned}$$

From the definitions of $p_1(\mathrm{x}), p_2(\mathrm{x})$ above, $P_1(\mathrm{x})\leq\frac{1}{2}$ and $P_2(\mathrm{x})\leq\frac{1}{2}$. Then $1-2P_1(\mathrm{x})\geq 0$ and $1-2P_2(\mathrm{x})\geq 0$. Note that $\int_{\mathrm{x}\in S}\frac{P_1(\mathrm{x})}{A}d\mathrm{x}+\int_{\mathrm{x}\in S}\frac{P_2(\mathrm{x})}{A}d\mathrm{x}$ is a constant independent of A_1 and A_2. Thus, $R(b)$ is also an increasing function of A_1, A_2. Also, if $min(\|g(\mathrm{x})\|)>e$, $p_1(\mathrm{x})=p_2(\mathrm{x})=0$. This means that the noise has no influence on classification.

Only η_k is considered

Assume that $\eta_k\in[-d,d]$. The probability density function $p(\mathrm{x},y)$ is as below

$$p(\mathrm{x},1) = \begin{cases}\frac{\mathbf{p}(\mathrm{x})}{A} & x\in(S_1\cup S_3)-(S_5\cup S_6)\\ \frac{\mathbf{p}(\mathrm{x})(1-P_3(\mathrm{x}))}{A} & \mathrm{x}\in(S_5\cup S_6)\\ \frac{\mathbf{p}(\mathrm{x})P_4(\mathrm{x})}{A} & \mathrm{x}\in(S_7\cup S_8)\\ 0 & \mathrm{x}\in(S_2\cup S_4)-(S_7\cup S_8)\end{cases}$$

$$p(\mathrm{x},-1) = \begin{cases}0 & \mathrm{x}\in(S_1\cup S_3)-(S_5\cup S_6)\\ \frac{\mathbf{p}(\mathrm{x})P_3(\mathrm{x})}{A} & \mathrm{x}\in(S_5\cup S_6)\\ \frac{\mathbf{p}(\mathrm{x})(1-P_4(\mathrm{x}))}{A} & \mathrm{x}\in(S_7\cup S_8)\\ \frac{\mathbf{p}(\mathrm{x})}{A} & \mathrm{x}\in(S_2\cup S_4)-(S_7\cup S_8)\end{cases}$$

where $P_3(\mathrm{x})$ and $P_4(\mathrm{x})$ are defined similarly to $P_1(\mathrm{x})$ and $P_2(\mathrm{x})$. We also have $P_3(\mathrm{x})\leq\frac{1}{2}$ and $P_4(\mathrm{x})\leq\frac{1}{2}$. As illustrated in Figures 7.2(b) and 7.2(c), let S_5, S_6, S_7

and S_8 be the sets in which the distance to $h^*(\mathrm{x}) = 0$ is less than d with the corresponding volumes A_5, A_6, A_7 and A_8, respectively. The boundaries are two parallel planes $h^*(\mathrm{x}) = \pm d$. In the region between the two planes, η_k may lead to misclassification. When $h(\mathrm{x})$ is not close to $h^*(\mathrm{x})$, $S_5 \subset S_1$ and $S_8 \subset S_2$ as seen in Figure 7.2(b). When $h(\mathrm{x})$ approaches to $h^*(\mathrm{x})$, $S_5 = S_1$ and $S_8 = S_2$, which can be seen in Figure 7.2(c). The expected risk $R(b)$ is given as

$$R(b) = \begin{cases} \int_{\mathrm{x}\in S_1 - S_5} \frac{\mathbf{p}(\mathrm{x})}{A} d\mathrm{x} + \int_{\mathrm{x}\in S_5} \frac{\mathbf{p}(\mathrm{x})(1-2p_3(\mathrm{x}))}{A} d\mathrm{x} + \int_{\mathrm{x}\in S_2 - S_8} \frac{\mathbf{p}(\mathrm{x})}{A} d\mathrm{x} \\ + \int_{\mathrm{x}\in S_8} \frac{\mathbf{p}(\mathrm{x})(1-2p_4(\mathrm{x}))}{A} d\mathrm{x} + \int_{\mathrm{x}\in S} \frac{\mathbf{p}(\mathrm{x})p_4(\mathrm{x})}{A} d\mathrm{x} + \int_{\mathrm{x}\in S} \frac{\mathbf{p}(\mathrm{x})p_3(\mathrm{x})}{A} d\mathrm{x} \\ if \quad S_5 \subseteq S_1, S_8 \subseteq S_2 \\ \\ \int_{\mathrm{x}\in S_1} \frac{\mathbf{p}(\mathrm{x})(1-p_3(\mathrm{x}))}{A} dx + \int_{\mathrm{x}\in S_2} \frac{\mathbf{p}(\mathrm{x})(1-p_4(\mathrm{x}))}{A} + \int_{\mathrm{x}\in S_1} \frac{\mathbf{p}(\mathrm{x})(1-2p_3(\mathrm{x}))}{A} dx \\ + \int_{\mathrm{x}\in S_2} \frac{\mathbf{p}(\mathrm{x})(1-2p_4(\mathrm{x}))}{A} d\mathrm{x} + \int_{\mathrm{x}\in S} \frac{\mathbf{p}(\mathrm{x})p_4(\mathrm{x})}{A} d\mathrm{x} + \int_{\mathrm{x}\in S} \frac{\mathbf{p}(\mathrm{x})p_3(\mathrm{x})}{A} d\mathrm{x} \\ if \quad S_1 = S_5, S_2 = S_8 \end{cases} \tag{7.6}$$

As $\int_{\mathrm{x}\in S} \frac{p(x)p_4(\mathrm{x})}{A} d\mathrm{x} + \int_{\mathrm{x}\in S} \frac{p(x)p_3(\mathrm{x})}{A} d\mathrm{x}$ is a constant independent of S_1, S_2, $R(\hat{b})$ is an increasing function of A_1,A_2.

Both e_k and η_k are considered

In this case, the probability density function $p(\mathrm{x}, y)$ becomes

$$p(\mathrm{x},1) = \begin{cases} \frac{\mathbf{p}(\mathrm{x})(1-P_1(\mathrm{x}))}{A} & \mathrm{x} \in (S_1 \cup S_3) - (S_5 \cup S_6) \\ \frac{\mathbf{p}(\mathrm{x})(1-P_5(\mathrm{x}))}{A} & \mathrm{x} \in (S_5 \cup S_6) \\ \frac{\mathbf{p}(\mathrm{x})P_6(\mathrm{x})}{A} & \mathrm{x} \in (S_7 \cup S_8) \\ \frac{\mathbf{p}(\mathrm{x})P_2(\mathrm{x})}{A} & \mathrm{x} \in (S_2 \cup S_4) - (S_7 \cup S_8) \end{cases}$$

$$p(\mathrm{x},-1) = \begin{cases} \frac{\mathbf{p}(\mathrm{x})P_1(\mathrm{x})}{A} & \mathrm{x} \in (S_1 \cup S_3) - (S_5 \cup S_6) \\ \frac{\mathbf{p}(\mathrm{x})P_5(\mathrm{x})}{A} & \mathrm{x} \in (S_5 \cup S_6) \\ \frac{\mathbf{p}(\mathrm{x})(1-P_6(\mathrm{x}))}{A} & \mathrm{x} \in (S_7 \cup S_8) \\ \frac{\mathbf{p}(\mathrm{x})(1-P_2(\mathrm{x}))}{A} & \mathrm{x} \in (S_2 \cup S_4) - (S_7 \cup S_8) \end{cases}$$

where $P_5(\mathrm{x})$ is the probability that $\mathrm{x} \in S_5 \cup S_6$ but $y = 1$ and $P_6(\mathrm{x})$ is the probability that $x \in S_7 \cup S_8$ but $y = -1$. $P_5(\mathrm{x})$ and $P_6(\mathrm{x})$ are related to $P_1(\mathrm{x})$, $P_2(\mathrm{x})$, $p_3(\mathrm{x})$, $P_4(\mathrm{x})$, $g(.)$, e_k and η_k. Note that we also have $P_5(\mathrm{x}) \leq \frac{1}{2}$ and $P_6(\mathrm{x}) \leq \frac{1}{2}$. When $h(\mathrm{x})$ approaches to $h^*(\mathrm{x})$, $S_5 = S_1$ and $S_8 = S_2$, $R(b)$ can be obtained as:

$$\begin{aligned} R(b) &= \int_{\mathrm{x}\in S_1} \frac{\mathbf{p}(\mathrm{x})(1-2P_5(\mathrm{x}))}{A} d\mathrm{x} + \int_{\mathrm{x}\in S_1} \frac{\mathbf{p}(\mathrm{x})(1-P_5(\mathrm{x}))}{A} d\mathrm{x} \\ &+ \int_{\mathrm{x}\in S} \frac{\mathbf{p}(\mathrm{x})P_6(\mathrm{x})}{A} d\mathrm{x} + \int_{\mathrm{x}\in S_4-S_7} \frac{\mathbf{p}(\mathrm{x})(P_2(\mathrm{x}))}{A} d\mathrm{x} \\ &+ \int_{\mathrm{x}\in S_2} \frac{\mathbf{p}(\mathrm{x})(1-2P_6(\mathrm{x}))}{A} d\mathrm{x} + \int_{\mathrm{x}\in S_3-S_6} \frac{\mathbf{p}(\mathrm{x})P_1(\mathrm{x})}{A} d\mathrm{x} \\ &+ \int_{\mathrm{x}\in S} \frac{\mathbf{p}(\mathrm{x})P_5(\mathrm{x})}{A} d\mathrm{x} + \int_{\mathrm{x}\in S_2} \frac{\mathbf{p}(\mathrm{x})(1-P_6(\mathrm{x}))}{A} d\mathrm{x} \end{aligned}$$

Thus, $R(b)$ is an increasing function of A_1,A_2..

Theorem 7.2. *Under Assumptions 7.1-7.3, the estimate $\hat{b}_l^{svm}$ in (7.5) satisfies that $\lim_{l\to\infty} \hat{b}_l^{svm} = \lim_{l\to\infty} \hat{b}_l^{emp} = b^*$ (a.s).*

Proof. The proof is similar to the proof of Theorem 7.1. For all the three cases

considered above, namely noise free case, the presence of either e_k or η_k, and the presence of both e_k and η_k, $R(b)$ is an increasing function of A_1, A_2. $R(b)$ attains its minimum if and only if $A_1, A_2 \to 0$ (a.s), i.e, $\lim_{l\to\infty} \hat{b}_l^{svm} = \lim_{l\to\infty} \hat{b}_l^{emp} = b^*$ (a.s). □

Remark 7.5. *By using our method, a consistent estimates of b can be obtained. This is due to the following reasons. Firstly, the constraints in classification are inequalities (see ξ in the inequalities constraints) rather than equalities. Secondly, we have the conclusion that minimizing the SVM loss and minimizing the classification error are equivalent from [98]. Finally it can be proven that the expected classification error is an increasing function of the norm of the bias between the estimates and the true parameters.*

Remark 7.6.

1) *If $\hat{b}_l^{svm} = b^*$, we redesign another i.i.d input sequence and reconstruct the training data set denoted as $T_l' = \{(\hat{z}_1, Y_1), ..., (\hat{z}_l, Y_l)\}$ so that T_l' still satisfies the i.i.d condition. Let $k(\hat{z}_k) = [k_1(\hat{z}_k)...k_h(\hat{z}_k)]'$. Then estimating the coefficients in the static function can be formulated as the same optimization problem as in (6.4): $min_{\{c,\xi\}}\ \frac{1}{2}c^Tc + \gamma\sum_{k=1}^{l}\xi_k$ s.t. $Y_k k(\hat{z}_k)^T c \geq (1-\xi_k),\ \xi_k \geq 0$. The estimates $\hat{c}_l^{svm}$ can be obtained as $\hat{c}_l^{svm} = \sum_{k=1}^{l}\alpha_k^* Y_k k(\hat{z}_k)$. Let c^* be the true parameter of c. Since the model is identical to (6.4), we can also obtain $\hat{c}_l^{svm} \to c^*$ (a.s) as $l \to \infty$.*

2) *Note that $\hat{b}_l^{svm} \to b^*$ (a.s) as $l \to \infty$. On the other hand, $\hat{c}_l^{svm} \to c^*$ depends on whether $\hat{b}_l^{svm} = b^*$. In practical application, we cannot obtain $\hat{b}_l^{svm} = b^*$ since l cannot be infinity. So the issue of recovering the nonlinear static function remains open in this case.*

7.4 Wiener Model with IIR Linear System Identification

7.4.1 Wiener Model with IIR Linear System

In this section, we extend the results established in Section 7.3 to a stable IIR system qualitatively. A stable IIR linear system can be approximated by an FIR system as follows: $z_k = \mu_0 u_k + \mu_1 u_{k-1} + ... + \mu_{nq} u_{k-nq} + \epsilon_k$ where $(\mu_0 \ ... \ \mu_{nq})$ are the coefficients of the newly transformed FIR system. Obviously, $\mu_0 = b_0$. Denoting $\mu = (\mu_1 \ ... \ \mu_{nq})^T$ and expanding equation of the IIR system, we get $\mu = Ab$ where $A \in R^{nq \times (m+1)}$ and

$$A = \begin{bmatrix} a_1 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_n & ... & a_1 & 1 & ... & \vdots \\ a_1^2 & a_n & ... & a_1 & 1 & ... \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ a_n^2 & a_{n-1}^2 & ... & a_n & a_{n-1} & ... \end{bmatrix}$$

Remark 7.7. *When the linear system is IIR , its output gives a dependent sequence. Thus we cannot guarantee the i.i.d property of T_l. As an IIR system is approximated by an FIR system with an approximation error ϵ_k depending on the order nq, we can use the same identification method as for the Wiener system with an FIR system. Then Lemma 3.1 should be revised to* $\lim_{l\to\infty} P(Q(l) > \varepsilon + \epsilon^*) = 0$ *where $\epsilon^* \to 0$ as $nq \to \infty$. Employing the proposed SVM classification based method leads to biased estimates of μ_k, for $k = 0, 1, ..., nq$, because one cannot guarantee that A_1 and A_2 converge to zero (a.a.s) due to the existence of ϵ^* with a finite nq. Nevertheless, as a larger nq is chosen, more accurate estimates can be*

achieved.

By employing the proposed SVM classification method, μ can be estimated and denoted as $\hat{\mu}$. Now we investigate how to estimate the parameters a and b in the IIR system based on $\hat{\mu}$. This needs to solve the following multi-variable high order nonlinear equations: $f_i = 0: \ A(i,:)b - \hat{\mu}_i = 0, i = 1, 2, ..., nq$ where $A(i,:)$ is the $i-th$ row of matrix A. Let $f = (f_1, ..., f_{nq})^T$ and $x = (a, b)$ with the dimension $m + n + 1$. We need to solve $f(x) = 0$.

7.4.2 Solution of Nonlinear Equations by Using Trust Region Algorithm

Denote that $f(x) = 0$, where $f = (f_1, ..., f_N)^T$, $R^{m+n+1} \rightarrow R^{nq}$, $m + n + 1$ is the dimension of variable x, nq is the number of nonlinear equations. Usually $N > M$, which is called over-determined equations. Solving equation $f(x) = 0$ is equivalent to minimizing the following unconstrained optimization problem

$$min \ F(x) = f(x)^T f(x) = \frac{1}{2} \sum_{i=1}^{nq} f_i^2(x) \tag{7.7}$$

Gauss-Newton iteration method is widely used to search for its solution. The problem of Gauss-Newton method is that we cannot guarantee its convergence [96]. In order to guarantee the convergence of the iteration, we use the trust region algorithm which has the property of global convergence [93]. As $F(x)$ is twice differentiable, then for sufficiently small s, the quadratic model of $F(x)$ given as $q^{(x)}(s) = F(x) + J(x)^T f(x)s + \frac{1}{2}s^T G_k s$ can be used to approximate $F(x + s)$ where $J(x)$ is the Jacobi matrix of $f(x)$ and G_k is the Hessian matrix of F at point x. Trust region algorithm can always find a feasible descent direction even if G_k is non-positive or x_k is a saddle point. This method searches for the descent

direction, denoted as s_k, under a step length with upper bound h_k^2. Define $\Omega_k = \{x | \|x - x_k\| \leq h_k^2\}$. Then we have $x_{k+1} = x_k + s_k$. The model of trust region algorithm is given by

$$min\ q^k(s) = F(x_k) + J(x)^T f(x)s + \tfrac{1}{2}s^T G_k s \quad s.t. \quad s^T s \leq h_k^2 \tag{7.8}$$

with its Lagrange function being $L(s, \lambda) = q^{(k)}(s) + \lambda(s^T s - h_k^2)$. For the $k-$th iteration, $\lambda = \lambda_k$. As $L(s, \lambda_k)$ is a quadric function of s, let $\nabla L(s, \lambda_k) = 0$ and we have

$$\bigtriangledown_s L(s, \lambda_k) = q^{(k)}(s_k) + \tfrac{1}{2}(s - s_k)^T (G_k + \lambda_k I)(s - s_k) = 0 \tag{7.9}$$

where λ_k is chosen such that $G_k + \lambda_k I$ is positive definite. Then the iterative algorithm is given as:

$$x_{k+1} = x_k - s_k = x_k - (G_k + \lambda_k I)^{-1}(J(x_k)^T f(x_k)) \tag{7.10}$$

Note that the convergence of trust region methods can be found in [96].

Let the notation 'hat' $\hat{}$ be the estimate of a parameter. For the identification of Wiener systems based on clipped observations, the procedure of our algorithm is summarized in the following steps

Step 1: If the linear system is FIR, obtain $\hat{b}$ which is $\hat{b}_l^{svm}$ in (7.5).

Step 2: Else, transform the IIR stable linear system to an FIR model with finite order. After obtaining $\hat{\mu}$ using step 1), obtain $\hat{a}$ and $\hat{b}$ by solving the multi-variable high order nonlinear equations using trust region algorithm.

Step 3: Obtain $\hat{a}$ and $\hat{b}$ and reconstruct training set T_l', estimate the coefficients c of the nonlinear function using the same method as in step 1).

7.5 Comparisons and Simulation Illustration

7.5.1 Comparisons Between Regression Based LS-SVM and Classification Based SVM in the Identification of the Wiener System with an FIR System

Note that LS-SVM for regression is a well known method and has become a very powerful tool in Wiener and Hammerstein-Wiener systems identification. In this subsection, we focus on the comparison of LS-SVM for regression based method and our proposed SVM for classification based method in the identification of Wiener systems. For comparison, we consider a Wiener system modeled in (7.1) with $a^* = [0 \ ... \ 0]^T$, $b^* = [1 \ 0.6 \ 0.5 \ 0.4 \ 0.3 \ 0.2 \ 0.1]^T$ and have the following three cases.

Case 1. The inverse of the static function $g^{-1}(.)$ exists.

This requires $e_k = 0$. Since otherwise the existence of e_k may cause that z_k cannot be represented as $z_k = g^{-1}(y_k)$. Then (1) becomes $g^{-1}(y_k) = b^T \mathrm{x}_k + \eta_k$. With both regression based LS-SVM and our proposed classification based SVM, parameter b can be estimated and this Wiener system can be well identified.

Case 2. The static function $g(z_k)$ is a non-invertible function.

For this case, LS-SVM for regression based method is unable to estimate b as seen in [26] [91]. However, in this case, one can treat the Wiener system as a nonlinear mapping $y_k = \mathcal{F}(\mathrm{x}_k)$ where $\mathrm{x}_k = [u_k \ ... \ u_{k-6}]$. Then the identification problem can be converted to a function approximation problem, i.e, finding $\hat{\mathcal{F}} \in S(\mathcal{F})$ to approximate $\mathcal{F}$ with $S(\mathcal{F})$ denoting the function space that $\hat{\mathcal{F}}$ belongs to. Since LS-SVM leads to a convex optimization, the global optimal of the cost function in

LS-SVM based identification method can always be obtained. With the increasing of the complexity of $\hat{\mathcal{F}}$ and learning data points, the approximation error $\|\hat{\mathcal{F}} - F\|$ can be made arbitrarily small if the internal noise $\eta_k = 0$. This means that the output of $\hat{\mathcal{F}}(.)$ can track the output of the Wiener system $\mathcal{F}(.)$ very well. However, a function approximation problem cannot provide more detailed information of the system, for example, no knowledge about b. In certain applications like controller design such details are required. By using our proposed method, estimating b becomes possible for this case even when the observations are clipped as shown in Case 3.

Case 3. System output is clipped in the presence of both internal and external noises.

In this case, the nonlinear part of the Wiener system is $y_k = g(z_k) + e_k$ and we have $Y_k = sgn(y_k)$ after y_k. We take $g(z) = z - \frac{2}{3}z^2$ as an example. Then $c^* = [1 \ -\frac{2}{3}]$. As both $g(z)$ and the binary-valued function are non-invertible, LS-SVM becomes unapplicable to estimate parameters b but it may be used for approximation as in Case 2. Using our proposed method, one can obtain satisfactory estimates of b even when both internal and external noises are presented. For illustration, let $e = 0.05$ and $d = 0.05$ be the upper bound of η_k and e_k and use $l = 300$ and $N = lm' = l(m+1) = 2400$ to identify the system. We obtain $\hat{b} = [1\ 0.5930\ 0.5026\ 0.4011\ 0.3031\ 0.1990\ 0.0961]$ and $\hat{c} = [1 \ -0.6638]$.

7.5.2 An Example of Wiener System with an IIR System

In this example, the system to be identified is $z_k = 0.2z_{k-1} + 0.1z_{k-2} + 1u_k + 0.5u_{k-1} + 0.5u_{k-1} + \eta_k$, $y_k = g(z_k) + e_k$ and we also choose $g(z) = z - \frac{2}{3}z^2$, $Y_k = sgn(y_k)$. We use white noise input and $e = 0.05$ and $d = 0.05$ as the upper bound of η_k and e_k. We choose an FIR model with $m = 8$ to approximate the

IIR model and use $l = 400$ ($N = lm' = 3600$) to identify the system. We get $\hat{\mu} = [0.6941\ 0.7272\ 0.2020\ 0.0959\ 0.0358\ 0.0132\ 0.0066\ 0.0019]$. By using the trust region method, we obtain the estimates of a^* and b^* as $\hat{a} = [0.2005\ 0.1112]$, $\hat{b} = [1.0000\ 0.4941\ \ 0.5062]$. For the polynomial model of the nonlinear function, we obtain $\hat{c} = [1\ \ -0.6531]$. Clearly, when we choose $q \geq 4$, i.e, $nq \geq 8$, the estimates becomes very close to their true values. Thus the proposed method also performs well in identifying Wiener system with an IIR system.

7.6 Summary

In this chapter, the idea from support vector machine based on classification approach is used to identify Wiener systems with binary quantized observations. The model structure consists of an FIR or a stable IIR linear system followed by a static function. It is shown that identification of the FIR system can be formulated as a convex problem and the estimates converge to the true parameters of the FIR system. For a stable IIR model, an FIR model is used to approximate IIR model and the identification problem is converted to identifying an FIR model together with solving a set of nonlinear equations. Examples show the effectiveness of our proposed schemes.

Chapter 8

Conclusions and Future Works

The conclusions and contributions of this thesis are given as follows:

1) In Chapter 2, we propose a new class of block-oriented systems which includes Hammerstein-Wiener systems. A new algorithm called kernel machine and space projection method is proposed to identify the newly proposed model.

2) In Chapter 3, we propose a new iterative algorithm for a general Hammerstein systems and prove its convergence. We also give a geometrical explanation of why the convergence property can be achieved.

3) In Chapter 4, we introduce fixed point iteration to identifying both Hammerstein and Wiener systems. A unified iterative algorithm is proposed inspired from fixed point theory and the convergence is guaranteed. It is shown that the iteration is a contraction mapping on a metric space when the number of input-output data points approaches infinity.

4) In Chapter 5, we formulate a new general bilinear model which actually represents a class of Wiener-Hammerstein systems. This new general bilinear model includes Hammerstein and Wiener systems as its special cases. The iterative

algorithm is proposed based on the fixed point iteration which is shown to be convergent. This gives a new point of view in proving the convergence property in identifying block-oriented systems.

5) In Chapter 6, we extend the iterative algorithm to our newly proposed block-oriented systems in Chapter 2. A new common model is proposed which actually represents the newly proposed block-oriented systems. Biconvex optimization is introduced to such systems.

6) In Chapter 7, we also consider the identification of block-oriented nonlinear systems based on clipped (binary quantized) observations. For the first time, SVM for classification is introduced to identify block-oriented nonlinear systems such as Wiener systems with clipped observations.

Based on our achievement, we feel that the following research directions can be further considered.

1) In this thesis we extend the Hammerstein-Hammerstein model to the new model shown in Figure 2.2. In that model, the input functions can be different in different paths, i.e, there are multiple input functions in the input block. One future research direction is to investigate whether the output block can contain multiple output functions or not. Obviously, the existing schemes cannot guarantee the convergence property for this case.

2) Currently the noise is assumed to be white noise in this thesis, which means the noise is ergodic. Considering coloured noise definitely generalize the applications of identification methods. When the coloured noise is present, the methods proposed by us need to be improved to adapt to this case. This also gives a potential future research direction.

3) All these methods proposed in this thesis are based on batched input output data points, which means that we have to collect all the data before identification. It will be more interesting if our methods allows that the input output data points can be processed one by one. How to derive such methods with guaranteed convergence property will also be a possible topic in future.

4) Last but not the least, the inputs are assumed to be i.i.d random points in order to guarantee the identifiability in this thesis. So the methods may not be applicable to some practical systems in which the inputs are not appropriate to be designed i.i.d. Thus, to investigate whether the i.i.d condition can be relaxed is an attractive direction for future research in system identification of block-oriented nonlinear systems.

Author's Publications

Journal Papers:

1) G. Li, C. Wen, and W. X. Zheng, "A new iterative identification scheme for Hammerstein systems with support vector machine based on biconvex optimization," *Australian Journal of Intelligent Information Processing Systems*, vol. 11, pp. 29-34, 2010.

2) G. Li, C. Wen, W. X. Zheng, and Y. Chen, "Identification of a class of nonlinear autoregressive models with exogenous inputs based on kernel machines," *IEEE Transactions on Signal Processing*, Vol. 59, pp. 2146-2159, 2011.

3) G. Li, C. Wen, G. B. Huang, and Y. Chen, "Error tolerance based support vector machine for regression," *Neurocomputing*, vol. 74, pp. 771-782, 2011.

4) G. Li and C. Wen, "Convergence of normalized iterative identification of Hammerstein systems," *Systems and Control Letters*, vol. 60, pp. 929-935, 2011.

5) G. Li and C. Wen, "Identification of wiener systems with clipped observations," *IEEE Transactions on Signal Processing*, Accepted, 2012.

6) G. Li and C. Wen, "Convergence of fixed point iteration for the identification of Hammerstein and Wiener Systems, " *International Journal of Robust and Nonlinear Control*, Accepted, 2012.

7) Y. Chen, C. Wen, G. Tao, M. Bi, and G. Li, "Continuous and noninvasive blood pressure measurement: A novel modeling methodology of the relationship between blood pressure and pulse wave velocity," *Annals of Biomedical Engineering*, vol. 37, pp. 2222-2233, 2009.

8) G. Li, C. Wen, Z. G. Li, Y. Feng, and K. Z. Mao, "Model based online learning with kernels," *Submitted to IEEE Transactions on Neural Networks*, Accepted with minor revision, 2011.

9) G. Li and C. Wen, "Convergence of fixed point iteration in identifying bilinear models," *IEEE Transactions on Signal Processing*, Revised and Resubmitted (under second round review), 2010.

10) G. Li, C. Wen, W. X. Zheng and G. Zhao, "Identification of block-oriented systems based on biconvex optimization," *Automatica*, In revision (under second round review), 2011.

11) K. Ramanathan, N. Ning, G. Li and L. Shi, "A model for the design of element cells in artificial cognitive memory using a habituating synapse and a persistent firing neuron," *International Journal of Neural Systems*, In revision (under second round review), 2011.

12) G. Li, K. Ramanathan, N. Ning, L. Shi and C. Wen, "A New Energy Function Based Memory Dynamics in Attractor Networks, " *Submitted to Neurocomputing*, 2012.

13) N. Ning, L. Pan, G. Li, K. Ramanathan, L. Shi and H. Tan "A Simple Model of Persistent Firing Neurons," *Submitted to Neural Computation*, 2012.

Conference Papers:

1) G. Li and C. Wen, "Legendre polynomials in signal reconstruction and compression," *5th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Taiwan, 2010.

2) G. Li and C. Wen, "Identification of Wiener systems based on fixed point theory," *11th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, Singapore, 2010.

3) G. Li, C. Wen, and Z. G. Li, "A new online learning with kernels method in novelty detection," *The 37th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Australia, 2011.

4) G. Zhao, G. Li, C. Wen and F. Yang, "Convergence of normalized iterative identification of Hammerstein systems," *ICAR2011*, Dubai, 2012.

5) G. Li, C. Wen, W. X. Zheng and G. Zhao, "On the iterative identification of block-oriented systems based on biconvex optimization," *16th IFAC Symposium on System Identification, SYSID 2012*, Brussels, Belgium, 2012.

6) G. Li, C. Wen, D. Cui and F. Yang, "A New Method of Online Learning with Kernels for Regression," *7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Singapore, 2012.

Bibliography

[1] F. Giri and E. W. Bai, "Block-oriented nonlinear system identification," *Springer*, 2010.

[2] A. Hammerstein, "Nichtlineare integralgleichung nebst anwendungen," *Acta Mathematica*, vol. 54, pp. 117-176, 1930.

[3] E. Eskinat, S. Johnson, W. L. Luyben, "Use of Hammerstein models in identification of nonlinear systems," *AICHE Journal*, vol. 37, pp. 255-268, 1991.

[4] K. J. Hunt, M. Munih, N. D. Donaldson, and F. M. D. Barr, "Investigation of the Hammerstein hypothesis in the modeling of electrically stimulated muscle," *IEEE Transactions on Biomedical Engineering* , vol. 45, pp. 998-1009, 1998.

[5] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *IEE Electronics Letters*, vol. 37, pp.1417-1418, 2001

[6] A. Balestrino, A. Landi, M. Ould-Zmirli, and L. Sani, "Automatic nonlinear auto-tuning method for Hammerstein modeling of electrical drives," *IEEE Transactions on Industrial Electronics*, vol. 48, pp 645-655, 2001.

[7] S. Sung, "System identification method for Hammerstein processes," *Industrial and Engineering Chemistry Research*, vol. 41, pp.4295-4302, 2002.

[8] E. Dempsey and D. Westwick, "Identification of Hammerstein models with cubic spline nonlinearities," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 237-245, 2004.

[9] R. Srinivasan, R. Rengaswamy, S. Narasimhan, and R. Miller, "Control loop performance assessment-Hammerstein model approach for stiction diagnosis," *Industrial and Engineering Chemistry Research* vol. 44, pp. 6719-6728, 2005.

[10] F. Jurado, "A method for the identification of solid oxide fuel cells using a Hammerstein model," *Journal of Power Sources* , vol. 154, pp. 145-152, 2006.

[11] J. Wang, A. Sano, T. Chen, and B. Huang, "Identification of Hammerstein systems without explicit parameterization of nonlinearity," *International Journal of Control*, vol. 82, pp.937-952, 2009.

[12] N. Wiener, "Nonlinear problems in random theory," *Wiley*, New York, 1958.

[13] S. Boyd and L. O. Chu, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Transactions on Circuits and Systems* , vol. 32, pp. 1150-1161, 1985.

[14] Y. Zhu, "Distillation column identification for control using Wiener model," *American Control Conference*, USA, vol. 5, pp. 3462-3466, 1999.

[15] A. Kalafatis, L. Wang, and W. R. Cluett, "Identification of time-varying pH processes using sinusoidal signals," *Automatica*, vol. 41, pp. 685-691, 2005.

[16] I.W. Hunter, and M. J. Korenberg, " The identification of nonlinear biological systems: Wiener and Hammerstein cascade models," *Biological Cybernetics*, vol. 55, pp. 135-144, 1986.

[17] Y. J. Lee, S. W. Sung, and S. Park, "Input test signal design and parameter estimation method for the Hammerstein-Wiener processes," *Industrial and Engineering Chemistry Research* , vol. 43, pp 7521-7530, 2004.

[18] H. J. Palanthandalam-Madapusi, A. J. Ridley, and D. S. Bernstein, "Identification and prediction of ionospheric dynamics using a Hammerstein-Wiener model with radial basis functions," *In American Control Conference*, USA, pp. 5052-5057, 2005.

[19] H. C. Park, S. W. Sung, and J. Lee, "Modeling of Hammerstein-Wiener processes with special input test signals," *Industrial and Engineering Chemistry Research*, vol. 45, pp 1029-1038, 2006.

[20] N. Kalouptsidis and P. Koukoulas, "Blind identification of volterra-Hammerstein systems," *IEEE Transactions on Signal Processing*, vol. 53, pp. 2777-2787, 2005.

[21] L. Vanbeylen, R. Pintelon, and J. Schoukens, "Blind maximum-likelihood identification of Wiener systems," *IEEE Transactions on Signal Processing*, vol. 57, pp. 3017-3029, 2009.

[22] A. Hagenblad and L. Ljung, "Maximum likelihood estimation of wiener models," *in Proceeding, 39:th IEEE Conference on Decision and Control*, Australia, pp. 2417- 2418, 2000.

[23] A. Hagenblad, L. Ljung, and A. Wills, "Maximum likelihood identification of Wiener models," *Automatica*, vol. 44 , pp 2697-2705, 2008.

[24] E. W. Bai, "An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear systems," *Automatica*, vol. 34, pp. 333-338, 1998.

[25] I. Goethals, K. Pelckmans, J. A. K. Suykens, and B. De Moor, "Identification of MIMO Hammerstein models using least squares support vector machines," *Automatica*, vol. 41, pp. 1263-1272, 2005.

[26] I. Goethals, K. Pelckmans, J. A. K. Suykens, and B. De Moor, "Subspace identification of Hmamerstein systems using least squares support vectors machines," *IEEE Transactions on Automatic Control*, vol. 50, pp. 1509-1519. 2005.

[27] M. Pawlak, Z. Hasiewicz, and P. Wachel, "On nonparametric identification of Wiener systems, " *IEEE Transactions on Signal Processing*, vol. 55, pp. 482-492, 2007.

[28] P. Sliwinski and Z. Hasiewicz, "Computational algorithms for multiscale identification of nonlinearities in Hammerstein systems with random inputs," *IEEE Transactions on Signal Processing*, vol. 53, pp. 360-364, 2005.

[29] P. Sliwinski, and Z. Hasiewicz, "Computational algorithms for wavelet identification of nonlinearities in Hammerstein systems with random inputs," *IEEE Transactions on Signal Processing*, vol. 56, pp. 846-851, 2008.

[30] W. Greblicki, "Nonparametric approach to Wiener system identification," *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Appllications*, vol. 44, pp. 538-545, 1997.

[31] Z. Q. Lang, "A nonparametric polynomial identification algorithm for the Hammerstein system," *IEEE Transactions on Automatic Control*, vol. 42, pp. 1435-1441, 1997.

[32] M. Espinoza, J. A. K. Suykens, and B. De Moor, "Kernel based partically nonlinear identification," *IEEE Transactions on Automatic Control*, vol. 50, pp. 1602-1606, 2005.

[33] P. G. Gallman, "A comparsion of two Hammerstein model identification algorithms," *IEEE Transactions on Automatic Control*, vol. 20, pp.771-775, 1975.

[34] G. Dolanc and S. Strmcnik, "Identification of nonlinear systems using a piecewise-linear Hammerstein model," *Systems and Control Letters*, vol. 54, pp. 145-158, 2005.

[35] W. Greblicki and M. Pawlak, "Nonparametric recovering nonlinearities in block-oriented systems with the help of Laguerre-polynomials," *Control-Theory and Advanced Technology*, vol. 10, pp. 771-791, 1994.

[36] J. Zheng, W. J. Yan, and J. Zhu, "Identification of Hammerstein model based on spline approximation and wavelet decomposition," *Journal of System Simulation*, vol. 17, pp. 1063-1067, 2005.

[37] P. Green, "Linear models for field trials, smoothing and cross-validation," *Biometrika* , vol. 72, pp. 527-537, 1985.

[38] P. Speckman, "Kernel smoothing in patrical linear model," *Journal of the Royal Statistical Society: Series B*, vol. 50, pp. 413-436, 1988.

[39] E. W. Bai, "A blind approach to the Hammerstein-Wiener model identification," *Automatica*, vol. 38, pp. 967-979, 2005.

[40] I. Goethals, K. Pelckmans, L. Hoegaerts, J. A. K. Suykens, and B. De Moor, "Subspace intersection identification of Hammerstein-Wiener systems," in *Proc. Joint 44th IEEE Conference Decision and Control and 2005 European Control Conference*, Spain, 2005, pp. 7108-7113.

[41] R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic systems-A survey of input/output approaches," *Automatica*, vol. 26, pp. 651-677, 1990.

[42] V. Vapnik, "Statistical learning theroy," *Wiley*, New York, 1998.

[43] G. Li, C. Wen, B. G. Huang, and Y. Chen, "Error tolerance based support vector machine for regression," *Neurocompting*, vol. 74, pp. 771-782, 2011.

[44] M. Martinez-Ramon, J. L. Rojo-Alvarez, G. Camps-Valls, J. Munoz-Mari, A. Navia-Vazquez, E. Soria-Olivas, and A. R. Figueiras-Vidal, "Support vector machines for nonlinear kernel ARMA system identification," *IEEE Transactions on Neural Network*, vol. 17, pp. 1617-1622, 2006.

[45] H. R. Zhang, X. D. Wang, C. J. Zhang, and X. S. Cai, "Robust identification of non-linear dynamic systems using support vector machine," *IEE Proceedings. Part A, Science, Measurement and Technology*, vol. 153, pp. 125-129, 2006.

[46] A. Rahimi and B. Recht, "Uniform approximation of functions with random bases," *The 46th Annual Allerton Conference on Communication, Control and Computing*, Monticello, 2008.

[47] B. Israel and T. N. E. Greville, "Generalized inverse: theory and applications," *Wiley*, New York, 1974.

[48] Y. Liu and E. W. Bai, "Iterative identification of Hammerstein systems," *Automatica*, vol. 43, pp. 346-354, 2007.

[49] P. Stoica, "On the convergence of an iterative algorithm used for Hammerstein system identification," *IEEE Transactions on Automatic Control*, vol. 26, pp. 967-969, 1981.

[50] H. F. Chen, "Pathwise convergence of recursive identification algorithms for Hammerstein systems ," *IEEE Transactions on Automatic Control* , vol. 49, pp. 1641-1649.

[51] K. S. Narendra and P. G. Gallman. "An Iterative Method for the Identification of Nonlinear Systems Using a Hammerstein Model," *IEEE Transactions on Automatic Control*, vol. 11, pp. 546-550, 1966.

[52] J. Vörös, "Recursive identification of Hammerstein systems with discontinuous nonlinearites containing dead-zones," *IEEE Transactions on Automatic Control*, vol. 48, pp. 2203-2206, 2003.

[53] W. X. Zhao and H. F. Chen, "Adaptive tracking and recursive identification for Hammerstein systems," *Automatica*, vol. 45, pp. 2773-2783, 2009.

[54] Y. L. Zhao, J. F. Zhang, L. Y. Wang, and G. G. Yin, "Identification of Hammerstein systems with quantized observations," *SIAM Journal on Control and Optimization*, vol. 48, pp. 4352-4376.

[55] E. W. Bai and K. Li, "Convergence of the iterative algorithm for a general hammerstein system Identification," *Automatica*, vol. 46, pp.1891-1896, 2010.

[56] W. Greblicki and M. Pawlak, "Nonparametric system identification," *Cambridge University Press*, 2008.

[57] Z. Hasiewicz and G. Mzyk, "Hammerstein system identification by nonparametric instrumental variables," *International Journal of Control*, vol. 82, pp. 440-455, 2009.

[58] Z. Hasiewicz and G. Mzyk "Combined parametric-nonparametric identification of Hammerstein systems," *IEEE Transactions on Automatic Control*, vol. 49, pp. 1370-1375, 2004.

[59] G. Li, C. Wen, W. X. Zheng, and Y. Chen, "Identification of a class of nonlinear autoregressive with exogenous inputs models based on kernel machines," *IEEE Transactions on Signal Processing*, vol. 59, pp. 2146-2158, 2011.

[60] L. A. Johnston and V. Krishnamurthy, "Finite dimensional smoothers for MAP state estimation of bilinear systems," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2444-2459, 1999.

[61] E. W. Bai and Y. Liu, "Least squares solutions of bilinear equations," *System and Control Letters*, vol. 55, pp. 466-472, 2006.

[62] N. Hazarika, A. Tsoi, and A. Sergejew, "Nonlinear considerations in EEG signal classification," *IEEE Transactions on Signal Processing*, vol. 45, pp. 829-836, 1997.

[63] M. J. Korenberg and I. W. Hunter, "The identification of nonlinear biological systems: LNL cascade models," *Biological Cybernetics*, vol. 55, pp. 125-134, 1986.

[64] J. Vörös, "An iterative method for Wiener-Hammerstein systems parameters identification," *Journal of Electrical Engineering*, vol. 58, pp. 114-117, 2007.

[65] J. Vörös, "Parameter identification of Wiener systems with multisegment piecewise-linear nonlinearities," *Systems and Control Letters*, vol. 56, pp. 99-105, 2007.

[66] W. Greblicki, "Continous time Hammerstein system identification," *IEEE Transactions on Automatic Control*, vol.45, pp. 1232-1236, 2000.

[67] G. Li and C. Wen, "Convergence of normalized iterative identification of Hammerstein systems," *Systems and Control Letters*, vol. 60, pp. 929-935, 2011.

[68] M. Boutayeb and M. Darouach "Recursive identification method for MISO Wiener-Hammerstein model," *IEEE Transactions on Automatic Control*, vol. 40, pp. 287-291, 1995.

[69] K. S. Narendra and P. G. Gallman, "Continuous time Hammerstein system identification," *IEEE Transactions on Automatic Control*, vol. 11, pp. 546-550, 1966.

[70] E. W. Bai and D. Li, "Convergence of the iterative Hammerstein system identification algorithm," *IEEE Transactions on Automatic Control*, vol. 49, pp. 1929-1940, 2004.

[71] A. Granas, and J. Dugundji, "Fixed point theory," *Springer-Verlag*, New York, 2001.

[72] N. Shimkin and A. Feuer, "On the necessity of "Block Invariance" for the convergence of adaptive pole-placement algorithm with persistently exciting input," *IEEE Transactions of Automatic Control*, vol. 33, pp. 775-780,1988.

[73] T. Butsan, S. Dhompongsaa, and W. Takahashi, "A fixed point theorem for pointwise eventually nonexpansive mappings in nearly uniformly convex banach spaces," *Nonlinear Analysis*, vol. 74, pp. 1694-1701, 2011.

[74] J. B. Tenenbaum, and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, pp. 1247-1283, 2000.

[75] J. Roll, A. Nazin, and L. Ljung, "Nonlinear system identification via direct weight optimization," *Automatica*, vol. 41, pp. 475-490, 2005.

[76] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions-a survey and extensions," *Mathematical Methods of Operations Research*, vol. 66. pp. 373-407, 2007.

[77] Y. Zhu "Estimation of an N-L-N Hammerstein-Wiener model," *Automatica*, vol 38, pp. 1607-1614, 2002.

[78] W. Greblicki, " Nonparametric input density-free estimation of the nonlinearity in Wiener systems, " *IEEE Transactions on Information Theory*, vol. 56, pp. 3575-3580, 2010.

[79] S. Rangan, G. Wolodkin and K. Poolla, "Identification method for Hammerstein systems," *Proceeding CDC*, New Orleans, pp. 697-702, 1995.

[80] A. E. Nordsjo and L. H. Zetterberg, "Identification of certain time-varying nonlinear Wiener and Hammerstein systems," *IEEE Transactions on Signal Processing*, vol. 49, pp. 577-592, 2001.

[81] K. Goh, L. Turan, M. Safonov, G. Papavassilopoulos, and J. Ly, "Biaffine matrix inequality properties and computational methods," *In Proceedings of the American Control Conference Baltimore*, Maryland, pp. 850-855, 1994.

[82] R. Herbrich, "Learning kernel classifiers theory and algorithms," *MIT Press*, 2002.

[83] H. Tuyen and L. Muu, "Biconvex programming approach to optimization over the weakly efficient set of a multiple objective affine fractional problem," *Operations Research Letters*, vol. 28, pp. 81-92, 2001.

[84] H. Sherali, A. Alameddine and T. Glickman, "Biconvex models and algorithms for risk management problems," *Operations Research Management Science*, vol. 35, pp. 405-408, 1995.

[85] W. Greblicki, "Nonparametric identification of Wiener systems," *IEEE Transactions Information Theory* , vol. 38, pp. 1487-1493, 1992.

[86] Y. Zhao, L. Wang, G. Yin, and J. Zhang, "Identification of Wiener systems with binary-valued output observations," *Automatica* , vol. 43, pp. 1752-1765, 2007.

[87] L. Y. Wang, Y. Kim, and J. Sun, "Prediction of oxygen storage capicity and stored NOx using HEGO sensor model for improved LNT control stategies," *ASME International Mechanical Engineering Congress and Exposition*, New Orleans, 2002.

[88] B.Scholkopf and A. Smola, "Learning with kernels," *MIT Press*, 2001.

[89] I. Steinwart and A. Christman, "Support vector machines," *Springer*, 2008.

[90] J. R. Alvarez, M. M. Ramon, M. P. Cumplido, A. A. Rodriguez and A. F. Vidal, "Support vector method for robust ARMA system identification," *IEEE Transactions on Signal Processing*, Vol. 52, pp. 155-164, 2004.

[91] T. Falck, K. Pelckmans, J. Suykens, and B. De Moor, "Identification of Wiener-Hammerstein systems using LS-SVMs," *15th IFAC Symposium on System Identification Saint-Malo*, France, pp. 820-825, 2009.

[92] L. Y. Wang, J. F. Zhang, and G. Yin, "System identification using binary sensors," *IEEE Transactions on Automatic Control*, vol. 48, pp. 1892-1907, 2003.

[93] R. H. Byrd, R. B. Schnabel, and G. A. Schultz, "A trust region algorithm for nonlinearly constrained optimization," *SIAM Journal on Numerical Analysis*, vol. 24, pp. 1152-1170, 1987.

[94] M. C. Hughes, and D. T. Westwick, "Identification of IIR Wiener systems with spline nonlinearities that have variable knots," *IEEE Transactions on Automatic Control*, pp. 1617-1622, vol. 50, 2005.

[95] V. Krishnamarty, "Estimation of quantized linear errors-in-variables models," *Automatica*, vol. 31, pp. 1459-1464, 1995.

[96] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM Journal on Applied Mathematics*, vol. 11, pp. 431-441, 1963.

[97] E. Rafaljowicz, "Linear systems identification from random threshold binary data," *IEEE Transactions on Signal Processing*, vol. 44, pp. 2064-2070, 1996.

[98] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *The Annals of Statistics*, vol. 32, pp. 56-134, 2004.

[99] V. Vapnik and A. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of their expectations," *Theory of Probability and Its Applications*, vol. 26, pp. 532-553, 1981.

[100] V. Vapnik and A. Chervonenkis, "The necessary and sufficient conditions for consitency in the empirical risk minimizations method," *Pattern Recognition and Image Analysis*, vol. 1, pp. 283-305, 1991.