BIRKHÄUSER

Systems & Control: Foundations & Applications

Le Yi Wang
G. George Yin
Ji-Feng Zhang
Yanlong Zhao

System Identification with Quantized Observations

BIRKHÄUSER

Systems & Control: Foundations & Applications

Le Yi Wang
G. George Yin
Ji-Feng Zhang
Yanlong Zhao

System Identification with Quantized Observations

Birkhäuser
Boston • Basel • Berlin

Le Yi Wang
Department of Electrical
and Computer Engineering
Wayne State University
Detroit, MI 48202
USA
lywang@ece.eng.wayne.edu

G. George Yin
Department of Mathematics
Wayne State University
Detroit, MI 48202
USA
gyin@math.wayne.edu

Ji-Feng Zhang
Key Laboratory of Systems and Control
Academy of Mathematics
and Systems Science
Chinese Academy of Sciences
Beijing 100190
China
jif@iss.ac.cn

Yanlong Zhao
Key Laboratory of Systems and Control
Academy of Mathematics
and Systems Science
Chinese Academy of Sciences
Beijing 100190
China
ylzhao@amss.ac.cn

ISBN 978-0-8176-4955-5 e-ISBN 978-0-8176-4956-2
DOI 10.1007/978-0-8176-4956-2

Library of Congress Control Number: 2010927909

Mathematics Subject Classification (2010): Primary: 93B30, 93E12, 93C55; Secondary: 60B10, 62L20

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my wife, Hong, for her ability to find joy in everything and share it with me, from French cooking to pilates, propofol controlling, or just doing nothing; and hopefully by a long stretch of that ability, in reading this book. And to my daughters, Jackie and Abby, for their support, encouragement, and sympathy after failing to find interesting topics in this book

Le Yi Wang

For Uncle Peter & Aunt Nancy and Uncle Richard & Aunt Helen, who helped me come to the U.S. and launch my academic career, with eternal gratitude

George Yin

To my wife, Ai Ling, and my son, Chao, for giving me a wonderful family full of harmony and happiness, like a peaceful harbor in a turbulent sea that makes me safe and sound

Ji-Feng Zhang

To my wife, Ying, for the magic of turning everything into happiness and for her support when I considered the topics included in this book during my overseas visits. To my family in my hometown for their selfless love, and, especially, in memory of my grandfather

Yanlong Zhao

Contents

Preface

This book concerns the identification of systems in which only quantized output observations are available, due to sensor limitations, signal quantization, or coding for communications. Although there are many excellent treaties in system identification and its related subject areas, a systematic study of identification with quantized data is still in its early stage. This book presents new methodologies that utilize quantized information in system identification and explores their potential in extending control capabilities for systems with limited sensor information or networked systems.

The book is an outgrowth of our recent research on quantized identification; it offers several salient features. From the viewpoint of targeted plants, it treats both linear and nonlinear systems, and both time-invariant and time-varying systems. In terms of noise types, it includes independent and dependent noises, stochastic disturbances and deterministic bounded noises, and noises with unknown distribution functions. The key methodologies of the book combine empirical measures and information-theoretic approaches to cover convergence, convergence rate, estimator efficiency, input design, threshold selection, and complexity analysis. We hope that it can shed new insights and perspectives for system identification.

The book is written for systems theorists, control engineers, applied mathematicians, as well as practitioners who apply identification algorithms in their work. The results presented in the book are also relevant and useful to researchers working in systems theory, communication and computer networks, signal processing, sensor networks, mobile agents, data fusion, remote sensing, tele-medicine, etc., in which noise-corrupted quan-

tized data need to be processed. Selected materials from the book may also be used in a graduate-level course on system identification.

This book project could not have been completed without the help and encouragement of many people. We first recognize our institutions and colleagues for providing us with a stimulating and supportive academic environment. We thank the series editor Tamer Başar for his encouragement and consideration. Our thanks also go to Birkhäuser editor Tom Grasso for his assistance and help, and to the production manager, and the Birkhäuser professionals for their work in finalizing the book. Our appreciation also goes to three anonymous reviewers, who read an early version of the book and offered many insightful comments and suggestions. During the years of study, our research has been supported in part by the National Science Foundation and the National Security Agency of the United States, and by the National Natural Science Foundation of China. Their support is greatly appreciated. We are deeply indebted to many researchers in the field for insightful discussions and constructive criticisms, and for enriching us with their expertise and enthusiasm. Most importantly, we credit our families for their unconditional support and encouragement even when they question our wisdom in working so tirelessly on mathematics symbols.

Detroit Le Yi Wang
Detroit George Yin
Beijing Ji-Feng Zhang
Beijing Yanlong Zhao

Conventions

This book uses a chapter-indexed numbering system for equations, theorems, etc., divided into three groups: (1) equations; (2) definitions, theorems, lemmas, corollaries, examples, propositions, and remarks; (3) assumptions. Each group uses its own number sequencing. For example, equation (3.10) indicates the tenth equation in Chapter 3. Similarly, group 2 entries are sequenced as Definition 4.1, Theorem 4.2, Corollary 4.3, Remark 4.4, Example 4.5 in Chapter 4. Assumptions are marked with the prefix "A" such as (A6.1), which indicates the first-listed assumption in Chapter 6.

In this book, the subscript is mainly used as a time index or iteration number for a sequence, such as y_k for signals at time k, a_l for the lth value of the system impulse response, and θ_N for the estimate at the Nth iteration. We limit the usage of superscripts whenever possible to avoid confusion with polynomial powers, or double subscripts to reduce convoluted notation, and will confine them in local sections when they must be used. The further dependence of a symbol on other variables such as vector or matrix indices, parameters, data length, etc., will be included in parentheses. For example, for a vector process y_k, $y_k^{\{i\}}$ denotes its ith element; $M(i,j)$ or $M^{\{i,j\}}$ represents the element at the ith row and jth column of a matrix M; $e_N(\theta,\mu)$ or $M(i,j;\theta,\mu)$ indicates their dependence on parameters θ and μ, although such a dependence will be suppressed when it becomes clear from the context. For a quick reference, in what follows we provide a glossary of symbols used throughout the book.

Glossary of Symbols

A'	transpose of a matrix or a vector A
A^{-1}	inverse of a matrix A
$\mathrm{Ball}_p(c, r)$	closed ball of center c and radius r in the l^p norm
$\mathbb{C}$	space of complex numbers
C_i	ith threshold of a quantized sensor
CR lower bound	Cramér–Rao lower bound
$E\xi$	expectation of a random variable ξ
$F(\cdot)$	probability distribution function
$\mathcal{F}$	σ-algebra
$\{\mathcal{F}_t\}$	filtration $\{\mathcal{F}_t, t \geq 0\}$
$G(v)$	componentwise operation of a scalar function G on a vector $v = [v^{\{1\}}, \ldots, v^{\{n\}}]'$, $G(v) = \left[G(v^{\{1\}}), \ldots, G(v^{\{n\}})\right]'$
$G^{-1}(v)$	componentwise inverse of a scalar invertible function G on a vector v: $G^{-1}(v) = \left[G^{-1}(v^{\{1\}}), \ldots, G^{-1}(v^{\{n\}})\right]'$
$H(e^{i\omega})$	scalar or vector complex function of ω
I	identity matrix of suitable dimension
I_A	indicator function of the set A
$\mathbb{N}$	set of natural numbers
ODE	ordinary differential equation
QCCE	quasi-convex combination estimate
$O(y)$	function of y satisfying $\sup_{y \neq 0} \lvert O(y)\rvert / \lvert y\rvert < \infty$
$\mathbb{R}^n$	n-dimensional real-valued Euclidean space

$\mathrm{Rad}_p(\Omega)$	radius of an uncertainty set Ω in l^p
$\mathcal{S}$	binary-valued or quantized sensor
T	Toeplitz matrix
$\mathbb{Z}_+$	set of positive integers
a^+	$= \max\{a, 0\}$ for a real number a
a^-	$= -\max\{-a, 0\}$ for a real number a
a.e.	almost everywhere
a.s.	almost sure
$\mathrm{diag}(A^1, \ldots, A^l)$	diagonal matrix of blocks $A^1, \ldots, A^l$
$f(\cdot)$	probability density function $f(x) = dF(x)/dx$
g_x or $\nabla_x g$	gradient of a function g with respect to (w.r.t.) x
$\mathtt{i}$	pure imaginary number with $\mathtt{i}^2 = -1$
i.i.d.	independent and identically distributed
$\ln x$	natural logarithm of x
$\log x$ or $\log_2 x$	base 2 logarithm of x
$o(y)$	function of y satisfying $\lim_{y\to 0} o(y)/\lvert y\rvert = 0$
q	one-step delay operator: $qx_k = x_{k-1}$
$s_k = \mathcal{S}(y_k)$	output of a sensor, either scalar or vector
$\mathrm{tr}(A)$	trace of the matrix A
$v^{\{i\}}$	ith component of the vector v
w.p.1	with probability one
$\lceil x \rceil$	ceiling function: the smallest integer that is $\geq x$
$\lfloor x \rfloor$	floor function: the largest integer that is $\leq x$
$\lVert x \rVert_p$	l^p norm of a sequence of real numbers $x = \{x_k; k \in \mathbb{N}\}$
Φ_N	$= [\phi_0, \ldots, \phi_{N-1}]'$, regression matrix at iteration N
$(\Omega, \mathcal{F}, P)$	probability space
$\varrho(c, r)$	neighborhood about c of radius r
θ	system parameter (column) vector of the modeled part
$\tilde{\theta}$	system parameter vector of the unmodeled dynamics, usually infinite dimensional
θ_N	estimate of θ at iteration step N
ϕ_k	regression (column) vector of θ at time k
$\tilde{\phi}_k$	regression vector of $\tilde{\theta}$ at time k, usually infinite dimensional
$:=$ or $\stackrel{\text{def}}{=}$	defined to be
$\mathbb{1}$	column vector with all elements equal to one
$\square$	end of a proof
$\lvert \cdot \rvert$	absolute value of a scalar or the Euclidean norm
$\lVert \cdot \rVert$	of a vector largest singular value of a matrix

Part I

Overview

1
Introduction

This book studies the identification of systems in which only quantized output observations are available. The corresponding problem is termed *quantized identification*.

Sampling and quantization were initially introduced into control systems as part of the computer revolution when controllers became increasingly implemented in computers. When computers had very limited memory and low speeds, errors due to sampling and quantization were substantial. Dramatic improvements in computer memory, speed, precision, and computational power made these considerations more an academic delicacy than a practical constraint. This seems to be the case even for wired and dedicated computer networks for which fiber optic networks can carry large quantities of data with lightning speed.

The recent explosive development in computer and communication networks has ushered in a new era of information processing. Thousands, even millions, of computers are interconnected using a heterogeneous network of wireless and wired systems. Due to fundamental limitations on bandwidth resources in wireless communications and large numbers of customers who share network resources, bandwidths have become a bottleneck for nearly all modern networks. Similar concerns arise in special-purpose networks such as smart sensors, MEMS (micro electromechanical systems), sensor networks, mobile agents, distributed systems, and remote-controlled systems, which have very limited power for communications and whose dataflow rates carry significant costs and limitations. These developments have made the issue of sampling and quantization once again fundamental for theoretical development and practical applications [13, 61, 80, 81, 82].

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_1,

Consider, for example, computer information processing of a continuous-time system whose output is sampled with a sampling rate of N Hz and quantized with a precision word-length of B bits. Its output observations carry the data-flow rate of NB bits per second (bps). For a typical 16-bit precision and 2-KHz sampling rate, a $32K$-bps bandwidth of data transmission resource is required, on observations of one output alone. This is a significant resource demand, especially when wireless communications of data are involved.

Additionally, quantized sensors are commonly employed in practical systems [10, 42, 49, 73, 90, 98, 99]. Usually they are more cost-effective than regular sensors. In many applications, they are the only ones available during real-time operations. There are numerous examples of binary-valued or quantized observations such as switching sensors for exhaust gas oxygen, ABS (anti-lock braking systems), and shift-by-wire in automotive applications; photoelectric sensors for positions, and Hall-effect sensors for speed and acceleration for motors; chemical process sensors for vacuum, pressure, and power levels; traffic condition indicators in the asynchronous transmission mode (ATM) networks; and gas content sensors (CO, CO_2, H_2, etc.) in the gas and oil industries. In medical applications, estimation and prediction of causal effects with dichotomous outcomes are closely related to binary sensor systems. The following examples represent some typical scenarios.

1.1 Motivating Examples

ATM ABR Traffic Control

An ATM network, depicted in Figure 1.1, consists of sources, switches, and destinations. Due to variations in other higher-priority network traffic such as constant bit rate (CBR) and variable bit rate (VBR), an available bit rate (ABR) connection experiences significant uncertainty on the available bandwidth during its operation. A physical or logical buffer is used in a switch to accommodate bandwidth fluctuations. The actual amount of bandwidth an ABR connection receives is provided to the source using rate-based closed-loop feedback control. One typical technique for providing traffic information is relative rate marking, which uses two fields in the Resource Management (RM) cell—the No Increase (NI) bit and the Congestion Indication (CI) bit. The NI bit is set when the queue reaches length C_1, and the CI bit is set when the queue length reaches C_2 ($C_2 > C_1$).

In this system, the queue length is not directly available for traffic control. The NI and CI bits indicate merely that it takes values in one of the three uncertainty sets $[0, C_1]$, $(C_1, C_2]$, and (C_2, ∞). This can be represented by two binary sensors. It is noted that the desired queue length is usually a value different than C_1 or C_2.

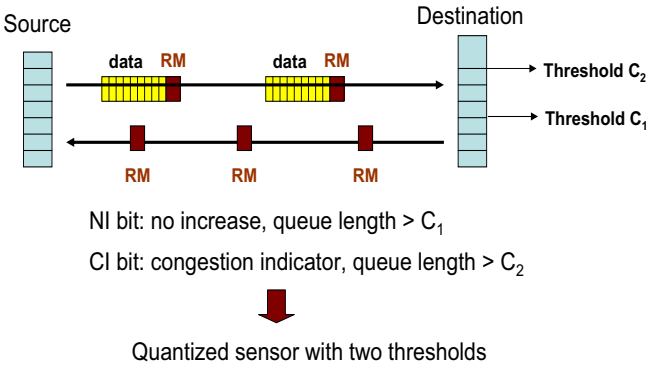


FIGURE 1.1. ATM network control

LNT and Air-to-Fuel Ratio Control with an EGO Sensor

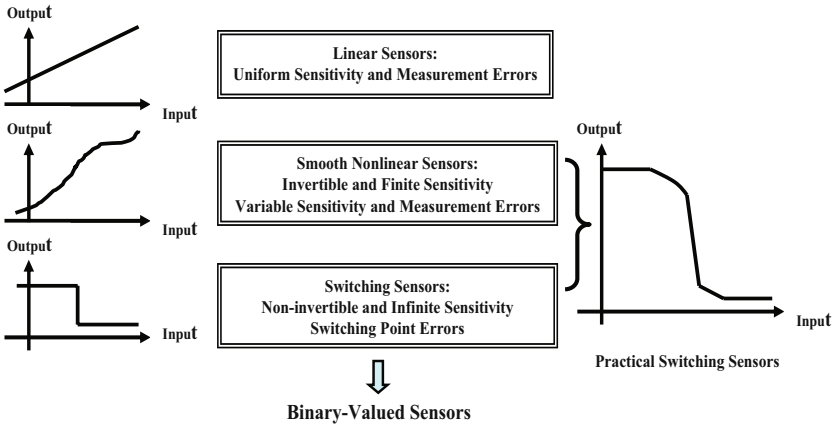


FIGURE 1.2. Sensor types

In automotive and chemical process applications, oxygen sensors are widely used for evaluating gas oxygen contents. Inexpensive oxygen sensors are switching types that change their voltage outputs sharply when excess oxygen in the gas is detected; see Figure 1.2. In particular, in automotive emission control, the exhaust gas oxygen sensor (EGO or HEGO) will switch its outputs when the air-to-fuel ratio in the exhaust gas crosses the stoichiometric value. To maintain the conversion efficiency of the three-way catalyst or to optimize the performance of a lean NOx trap (LNT), it is essential to estimate the internally stored NOx and oxygen. In this case, the switching point of the sensor has no direct bearing on the control target. The idea of using the switching sensor for identification purposes, rather

than for control only, can be found in [98, 99, 112].

Identification of Binary Perceptrons
There is an interesting intersection between quantized identification and statistical learning theory in neural networks. Consider an unknown binary perceptron depicted in Figure 1.3 that is used to represent a dynamic relationship:

$$y(t) = \mathcal{S}(w_1 x_1 + w_2 x_2 + \cdots + w_n x_n - C + d),$$

where C is the known neuron firing threshold, $w_1, \ldots, w_n$ are the weighting coefficients to be learned, and $\mathcal{S}(\cdot)$ is a binary-valued function switching at 0. This learning problem can be formulated as a special case of binary sensor identification without unmodeled dynamics. Traditional neural models, such as the McCulloch–Pitts and Nagumo–Sato models, contain a neural firing threshold that naturally introduces a binary function [13, 38, 42, 73]. Fundamental stochastic neural learning theory studies the stochastic updating algorithms for neural parameters [94, 95, 96].

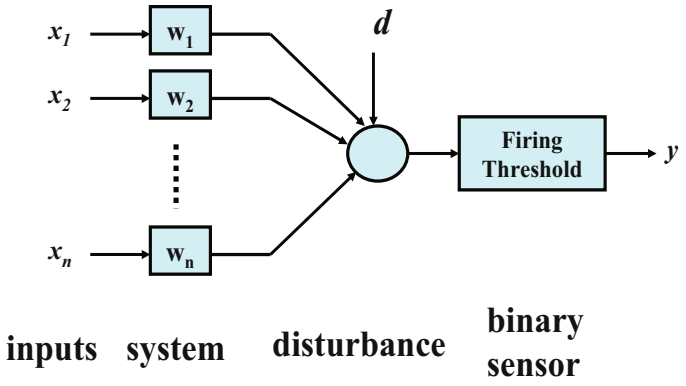


FIGURE 1.3. Artificial neural networks

Networked Systems
In a networked system, see Figure 1.4, signals must be transmitted through communication channels. Since communications make quantization mandatory, it is essential to understand identification with quantized observations. Unlike physical sensors whose characteristics such as switching thresholds cannot be altered during identification experiments, quantization for communication may be viewed generally as a partition of the signal range into a collection of subsets. Consequently, quantization thresholds may be selected to reduce identification errors, leading to the problems on threshold selection. Furthermore, source coding and channel coding after quantization play an important role in identification error characterization when communication channels are noisy.

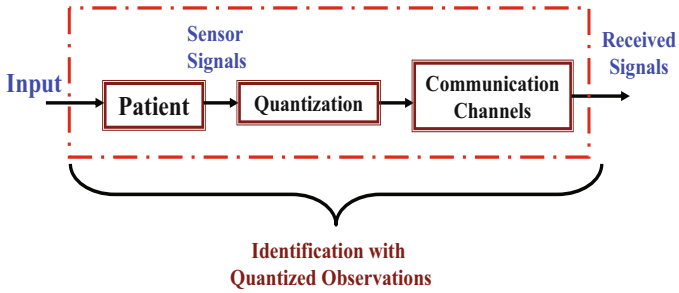


FIGURE 1.4. Communication networks

1.2 System Identification with Quantized Observations

The most common example of quantization is to divide the output range into equal-length intervals. This book considers a general class of quantized observations that allows a partition of the output range into a finite collection of subsets. The subsets may have unequal lengths or be unbounded, may be fixed due to sensor configuration limitations, or may be design variables such as quantization or coding in communication systems. This subject is of importance in understanding the modeling capability for systems with limited sensor information, establishing relationships between communication resource limitations and identification complexity, and studying sensor networks [1, 10, 13, 34, 61, 80, 81, 90, 112, 115].

The use of quantized observations introduces substantial difficulties since only very limited information is available for system modeling, identification, and control. Since switching sensors are nonlinear components, studies of their roles and impact on systems are often carried out in nonlinear system frameworks, such as sliding mode control, describing function analysis, switching control, hybrid control, etc. In these control schemes, the switching thresholds of the sensors are directly used to define a control target. However, their fundamental impact on system modeling and identification is a relatively new area. This book presents recent developments on the inherent consequences of using quantized observations in system identification as well as methods and algorithms that use quantized observations effectively to extend control capabilities.

The main motivation is embodied in many applications in which modeling of such systems is of great importance in performing model predictive control, model-based diagnosis, outcome predictions, optimal control strategy development, control adaptation, etc. When inputs can be arbitrarily selected within certain bounds and outputs are measured by regular sensors, system identification problems have been studied extensively in the traditional setup under the frameworks of either stochastic systems or

worst-case identification. The issues of identification accuracy, convergence, model complexity, time complexity, input design, persistent excitation, identification algorithms, etc. have been pursued by many researchers. A vast literature is now available on this topic; see [17, 55, 62, 67], among others.

It should be clarified that the treatments of this book will be meaningful only for the application problems in which quantization levels carry a substantial system cost or operational penalty. If an application can use cheaper sensors of high precision and data-flow rates do not carry a cost, traditional system identification methods will suffice. On the other hand, when a sensor of lower precision is cheaper than a higher-precision sensor, it is important to understand what the performance penalty will be if the cheaper sensor is used. Similarly, when communication bandwidths are limited, the reduction of quantization levels will save communication resources. Intrinsically, one may choose to use coarse quantization (lower space complexity) so that more data points can be transmitted (higher time complexity) with the same bandwidth resource demand. This tradeoff between sampling rates and quantization accuracy is a fundamental issue in complexity analysis for understanding the impact of communication channels on system performance.

Some fundamental issues emerge when the output observation is quantized: How accurately can one identify the parameters of the system? How fast can one reduce uncertainty on model parameters? What are the optimal inputs for fast identification? What are the conditions that ensure the convergence of the identification algorithms? What are the impacts of unmodeled dynamics and disturbances on identification accuracy and time complexity? In contrast to classical system identification, answers to these familiar questions under switching sensors differ substantially from the traditional setup.

This book demonstrates that quantized observations increase time complexity significantly; the optimal inputs differ from those in traditional identification; identification characteristics depart significantly between stochastic and deterministic noise representations; and unmodeled dynamics have a fundamental influence on identification accuracy of the modeled part. In contrast to traditional system identification, in which the individual merits of stochastic versus worst-case frameworks are sometimes debated, it is beneficial to combine these two frameworks in quantized identification problems.

1.3 Outline of the Book

This book is organized into five parts as follows: I. Overview; II. Stochastic Methods for Linear Systems; III. Deterministic Methods for Linear Systems; IV. Identification of Nonlinear and Switching Systems; V. Complexity Analysis.

Part I is an overview that provides motivational applications for system identification with quantized observations in Chapter 1 and that introduces the common systems settings for the entire book in Chapter 2. After a general introduction of the problems in Chapter 1, generic system settings are formulated in Chapter 2. Special cases of systems are then delineated, such as gain systems, finite impulse-response systems, rational systems, and nonlinear systems. The main issues of system identification are further explained, including typical inputs and signal ranks. System uncertainties considered in this book consist of unmodeled dynamics for linear dynamics, model mismatch for nonlinear functions, and disturbances in either a stochastic or deterministic worst-case description. Identification in different system configurations is outlined, in which distinct issues arising from open-loop and closed-loop systems, and input and actuator noises are further discussed.

In this book, unmodeled dynamics are always described as a bounded but unknown uncertainty. In contrast, disturbances are modeled either in a stochastic framework, or as an unknown but bounded uncertainty. Since these two frameworks require vastly different input design, employ distinct analysis methods, and entail diversified convergence properties, they are presented in their elementary but fundamental forms in Parts II and III, respectively.

Part II covers stochastic methods for linear systems with quantized observations. It presents identification algorithms, properties of estimators, and various convergence results in a stochastic system framework. Chapter 3 introduces the main methodology of empirical measures and derives convergence properties, including strong convergence, convergence in distribution, and mean-square convergence. When noise is modeled as a stochastic process and the system is subject to unmodeled dynamics, it becomes mandatory to deal with a combined framework, which is investigated in Chapter 4. Upper and lower error bounds are derived.

When dealing with a complicated system structure, a fundamental idea is to reduce the identification of its parameters to a finite set of identification of gains. This is achieved by employing some intrinsic properties of periodic inputs and invertible mappings. In Chapter 3, we show how a full-rank periodic input can reduce the problem of identifying a finite impulse-response system to a number of core identification problems of gains. When the system is rational, the problem becomes much more difficult. We show in Chapter 5 that this difficulty can be overcome by full-rank periodic inputs when the rational model is co-prime.

The convergence and efficiency of estimators in quantized identification are studied in Chapter 6. When the observation becomes quantized with a finite number of thresholds, an algorithm, termed optimal quasi-convex combination estimation (optimal QCCE), is introduced to derive an estimate from multiple thresholds. The resulting estimate is shown to be asymptotically efficient by comparing its convergence speed to the Cramér–Rao (CR) lower bound.

The utility of full-rank periodic inputs is further investigated in Chapter 7 in typical system configurations such as cascade and feedback connections. It is revealed that periodicity and full-rankness of a signal are two fundamental input properties that are preserved after the signal passes through a stable system with some mild constraints. Consequently, it becomes clear that the input design, identification algorithms, and convergence properties are inherently invariant under open-loop and closed-loop configurations. Furthermore, we present in Chapter 8 the joint identification of system parameters, unknown thresholds, and unknown noise distribution functions. The main idea is to use signal scaling to excite further information on sensor thresholds and noise distribution functions.

Part III is concerned with deterministic methods for linear systems. Here the emphasis is shifted to the deterministic framework for disturbances. Under this framework, noise is modeled as unknown but bounded. Our exploration starts in Chapter 9 with the case of binary-valued observations. Input design that utilizes the idea of bisection is shown to reduce uncertainty exponentially. This idea is employed when both observation noise and unmodeled dynamics are present. Explicit error bounds are derived. Chapter 10 considers the case of quantized observations. The utility of multiple thresholds in accelerating convergence speed is investigated.

Part IV concentrates on the identification of nonlinear and switching systems. The first concentration is on Wiener and Hammerstein systems in which the nonlinearity is confined to be memoryless. The algorithms for identifying such nonlinear systems closely follow the ideas of signal scaling in Chapter 8 to separate the identification of the linear dynamics and nonlinear function and to extract information on the nonlinear part. This is especially apparent in Chapter 11 for Wiener systems. Hammerstein systems are treated in Chapter 12. Although there are similarities between Wiener and Hammerstein systems, input design is more stringent in Hammerstein systems. Some new concepts of input ranks are introduced. Systems with switching parameters are discussed in Chapter 13. In such systems, parameters are themselves Markov chains. Two essential cases are included. In the first case, parameters switch their values much faster than identification convergence speeds. Consequently, it is only feasible to identify the average of the switching parameters. On the other hand, if the parameter jumps occur infrequently with respect to identification speeds, parameter tracking by identification algorithms can be accomplished. Algorithms, convergence, and convergence rates toward an irreducible uncertainty set are established.

Part V explores fundamental complexity issues in system identification with quantized observations. The main tool is the asymptotic efficiency that defines the impact of observation time complexity (data length) and space complexity (number of thresholds) on identification accuracy. The tradeoff between time complexity and space complexity points to a broad utility in threshold selection, optimal resource allocations, and communication quantization design. These discussions are contained in Chapter 14. This

understanding is further utilized to study the impact of communication channels on system identification in Chapter 15. The concept of the Fisher information ratio is introduced.

In addition to the aforementioned chapters and an extensive list of references at the end of the book, each chapter (except for Chapter 1) has a section of notes in which historical remarks, developments of related work and references, and possible future study topics are presented and discussed.

2

System Settings

This chapter presents basic system structures, sensor representations, input types and characterizations, system configurations, and uncertainty types for the entire book. This chapter provides a problem formulation, shows connections among different system settings, and demonstrates an overall picture of the diverse system identification problems that will be covered in this book. Other than a few common features, technical details are deferred to later chapters.

Section 2.1 presents the basic system structure and its special cases of FIR (finite impulse response), IIR (infinite impulse response), rational, and nonlinear systems that will be discussed in detail in later chapters. Quantized observations and their mathematical representations are described in Section 2.2. Essential properties of periodic input signals that are critical for quantized identification are established in Section 2.3. When a system is embedded in a larger configuration, its input and output are further confined by the system structure, introducing different identification problems. Section 2.4 shows several typical system configurations and their corresponding unique features in system identification. There are many types of uncertainties that can potentially impact system identification. These are summarized in Section 2.5.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_2,

2.1 Basic Systems

The basic system structure under consideration is a single-input–single-output stable system in its generic form

$$y_k = G(U_k, \theta) + \Delta(U_k, \widetilde{\theta}) + d_k, \quad k = 0, 1, 2, \ldots, \tag{2.1}$$

where $U_k = \{u_j, 0 \le j \le k\}$ is the input sequence up to the current time k, $\{d_k\}$ is a sequence of random variables representing disturbance, θ is the vector-valued parameter to be identified, and $\widetilde{\theta}$ represents the unmodeled dynamics. All systems will assume zero initial conditions, which will not be mentioned further in this book. We first list several typical cases of the system in (2.1).

1. **Gain Systems:**

   $$y_k = au_k + d_k.$$

   Hence, $\theta = a$, $G(U_k, \theta) = au_k$, and $\Delta(U_k, \widetilde{\theta}) = 0$.

2. **Finite Impulse Response (FIR) Models:**

   $$y_k = a_0 u_k + \cdots + a_{n_0-1} u_{k-n_0+1} + d_k.$$

   This is usually written in a regression form

   $$G(U_k, \theta) = a_0 u_k + \cdots + a_{n_0-1} u_{k-n_0+1} = \phi_k' \theta,$$

   where $\theta = [a_0, \ldots, a_{n_0-1}]'$ is the unknown parameter vector and $\phi_k' = [u_k, \ldots, u_{k-n_0+1}]$ is the regressor. In this case, the model order is n_0, which is sometimes used as a measure of model complexity. Again, $\Delta(U_k, \widetilde{\theta}) = 0$.

3. **Infinite Impulse Response (IIR) Models:**

   $$y_k = \sum_{n=0}^{\infty} a_n u_{k-n} + d_k,$$

   where the system parameters satisfy the bounded-input–bounded-output (BIBO) stability constraint

   $$\sum_{n=0}^{\infty} |a_n| < \infty.$$

   For system identification, this model is usually decomposed into two parts:

   $$\sum_{n=0}^{n_0-1} a_n u_{k-n} + \sum_{n=n_0}^{\infty} a_n u_{k-n} = \phi_k' \theta + \widetilde{\phi}_k' \widetilde{\theta}, \tag{2.2}$$

where $\theta = [a_0, \ldots, a_{n_0-1}]'$ is the modeled part and $\widetilde{\theta} = [a_{n_0}, a_{n_0+1}, \ldots]'$ is the unmodeled dynamics, with corresponding regressors

$$\phi_k' = [u_k, \ldots, u_{k-n_0+1}] \quad \text{and}$$
$$\widetilde{\phi}_k' = [u_{k-n_0}, u_{k-n_0-1}, \ldots],$$

respectively. In this case, the model order is n_0. For the selected n_0, we have

$$G(U_k, \theta) = \phi_k'\theta; \quad \Delta(U_k, \widetilde{\theta}) = \widetilde{\phi}_k'\widetilde{\theta}.$$

4. **Rational Transfer Functions:**

$$y_k = G(q, \theta)u_k + d_k. \tag{2.3}$$

Here q is the one-step shift operator $qu_k = u_{k-1}$ and $G(q)$ is a stable rational function[2.1] of q:

$$G(q) = \frac{B(q)}{1 - A(q)} = \frac{b_1 q + \cdots + b_{n_0} q^{n_0}}{1 - (a_1 q + \cdots + a_{n_0} q^{n_0})}.$$

In this case, the model order is n_0 and the system has $2n_0$ unknown parameters $\theta = [a_1, \ldots, a_{n_0}, b_1, \ldots, b_{n_0}]'$. Note that in this scenario, the system output is nonlinear in parameters. To relate it to sensor measurement errors in practical system configurations, we adopt the output disturbance setting in (2.3), rather than the equation disturbance structure in

$$y_k + a_1 y_{k-1} + \cdots + a_{n_0} y_{k-n_0} = b_1 u_{k-1} + \cdots + b_{n_0} u_{k-n_0} + d_k,$$

which is an autoregressive moving average (ARMA) model structure. The ARMA model structure is more convenient for algorithm development. But output measurement noises in real applications occur in the form of (2.3).

5. **Wiener Models:**

$$G(U_k, \theta) = H(G_0(q, \theta_1)u_k, \beta),$$

or in a more detailed expression

$$x_k = \sum_{n=0}^{n_0-1} a_n u_{k-n}, \quad y_k = H(x_k, \beta) + d_k.$$

[2.1] When $G(q, \theta)$ is used in a closed-loop system, it will be allowed to be unstable, but is assumed to be stabilized by the feedback loop.

Here, β is the parameter (column) vector of the output memoryless nonlinear function H and $\theta_1 = [a_0, \ldots, a_{n_0-1}]'$ is the parameter vector of the linear part. The combined unknown parameters are $\theta = [\theta_1', \beta']'$.

6. **Hammerstein Models:**

$$G(U_k, \theta) = G_0(q, \theta_1) H(u_k, \beta)$$

or

$$y_k = \sum_{n=0}^{n_0-1} a_n x_{k-n} + d_k, \quad x_k = H(u_k, \beta).$$

Here, β is the parameter vector of the input memoryless nonlinear function H and $\theta_1 = [a_0, \ldots, a_{n_0-1}]'$ is the parameter vector of the linear part. The combined unknown parameters are $\theta = [\theta_1', \beta']'$.

2.2 Quantized Output Observations

Let us begin with Figure 2.1. The output y_k in (2.1) is measured by a

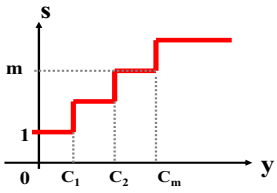


FIGURE 2.1. Quantized observations

sensor of m_0 thresholds $-\infty < C_1 < \ldots < C_{m_0} < \infty$. The sensor can be represented by a set of m_0 indicator functions $s_k = [s_k(1), \ldots, s_k(m_0)]'$, where $s_k(i) = I_{\{-\infty < y_k \le C_i\}}$, $i = 1, \ldots, m_0$, and

$$I_{\{y_k \in A\}} = \begin{cases} 1, & \text{if } y_k \in A, \\ 0, & \text{otherwise.} \end{cases}$$

In such a setting, the sensor is modeled as m_0 binary-valued sensors with overlapping switching intervals, which imply that if $s_k(i) = 1$, then $s_k(j) = 1$ for $j \ge i$. An alternative representation of the sensor is by defining $\widetilde{s}_k(i) = I_{\{C_{i-1} < y_k \le C_i\}}$ with $C_0 = -\infty$, and $C_{m_0+1} = \infty$ with the interval (C_{m_0}, ∞). This representation employs distinct switching intervals. Consequently, only one $s_k(i) = 1$ at any k.

Under a quantized sensor of m_0 thresholds, each sample of the signal can be represented by a code of length $\log_2 m_0$ bits. This will be viewed as the space complexity of the signal measurements.

2.3 Inputs

In this book, we use extensively periodic input signals in identification experiments under a stochastic framework. A signal v_k is said to be n_0-periodic if $v_{k+n_0} = v_k$. We first establish some essential properties of periodic signals, which will play an important role in the subsequent development.

Toeplitz Matrices

Recall that an $n_0 \times n_0$ Toeplitz matrix [37] is any matrix with constant values along each (top-left to bottom-right) diagonal. That is, a Toeplitz matrix has the form

$$T = \begin{bmatrix} v_{n_0} & \dots & v_2 & v_1 \\ v_{n_0+1} & \ddots & \ddots & v_2 \\ \vdots & \ddots & \ddots & \vdots \\ v_{2n_0-1} & \dots & v_{n_0+1} & v_{n_0} \end{bmatrix}.$$

It is clear that a Toeplitz matrix is completely determined by its entries in the first row and the first column $\{v_1, \dots, v_{2n_0-1}\}$, which is referred to as the symbol of the Toeplitz matrix.

Circulant Toeplitz Matrices and Periodic Signals

A Toeplitz matrix T is said to be circulant if its symbol satisfies $v_k = v_{k-n_0}$ for $k = n_0 + 1, \dots, 2n_0 - 1$; see [25]. A circulant matrix [57] is completely determined by its entries in the first row $[v_{n_0}, \dots, v_1]$, so we denote it as $T([v_{n_0}, \dots, v_1])$. Moreover, T is said to be a generalized circulant matrix if $v_k = \rho v_{k-n_0}$ for $k = n_0 + 1, \dots, 2n_0 - 1$, where $\rho > 0$, which is denoted by $T(\rho, [v_{n_0}, \dots, v_1])$ and

$$T(\rho, [v_{n_0}, \dots, v_1]) = \begin{bmatrix} v_{n_0} & \dots & v_2 & v_1 \\ \rho v_1 & \ddots & \ddots & v_2 \\ \vdots & \ddots & \ddots & \vdots \\ \rho v_{n_0-1} & \dots & \rho v_1 & v_{n_0} \end{bmatrix}.$$

Definition 2.1. An n_0-periodic signal generated from its one-period values $v = (v_1, \dots, v_{n_0})$ is said to be full rank if $T([v_{n_0}, \dots, v_1])$, the circulant matrix, is full rank.

An important property of circulant matrices is the following frequency-domain criterion.

Lemma 2.2. *If $T = T(\rho, [v_{n_0}, \ldots, v_1])$ is a generalized circulant matrix, then the eigenvalues of T are $\{\rho\gamma_k, k = 1, \ldots, n_0\}$ and the determinant of T is $\det(T) = \prod_{k=1}^{n_0} \rho\gamma_k$, where γ_k is the discrete Fourier transform (DFT) of $v_j\rho^{-(j/n_0)}$, $j = 1, \ldots, n_0$:*

$$\gamma_k = \sum_{j=1}^{n_0} v_j \rho^{-\frac{j}{n_0}} e^{-i\omega_k j}, \quad \omega_k = \frac{2\pi k}{n_0}, \ k = 1, \ldots, n_0.$$

Hence, T is full rank if and only if $\gamma_k \neq 0$, $k = 1, \ldots, n_0$.

Proof. Let

$$M = \begin{bmatrix} 0 & I_{n_0-1} \\ \rho & 0 \end{bmatrix},$$

whose characteristic polynomial is $\lambda^{n_0} - \rho$ and eigenvalues are $\rho^{-(j/n_0)}e^{i\omega_k}$, $k = 1, \ldots, n_0$. Then, T can be represented as $T = \sum_{j=1}^{n_0} v_j M^{n_0-j}$. For $k = 1, \ldots, n_0$, if x_k is the corresponding eigenvector of the eigenvalue $\rho^{-(1/n_0)}e^{i\omega_k}$ for M, then

$$\begin{aligned} Tx_k &= \sum_{j=1}^{n_0} v_j M^{n_0-j} x_k \\ &= \sum_{j=1}^{n_0} v_j (\rho^{\frac{1}{n_0}} e^{i\omega_k})^{n_0-j} x_k \\ &= \rho\gamma_k x_k. \end{aligned}$$

Therefore, $\rho\gamma_k$ is an eigenvalue of T and the expression for $\det(T)$ is confirmed. By hypothesis, $\rho > 0$. Hence, T is full rank if and only if $\gamma_k \neq 0$, $k = 1, \ldots, n_0$. □

For the special case when $\rho = 1$, we have the following property.

Corollary 2.3. *An n_0-periodic signal generated from $v = (v_1, \ldots, v_{n_0})$ is full rank if and only if its discrete Fourier transform $\gamma_k = \sum_{j=1}^{n_0} v_j e^{-i\omega_k j}$ is nonzero at $\omega_k = 2\pi k/n_0$, $k = 1, \ldots, n_0$.*

Recall that $\Gamma = \{\gamma_1, \ldots, \gamma_{n_0}\}$ are the frequency samples of the n_0-periodic signal v. Hence, Definition 2.1 may be equivalently stated as "an n_0-periodic signal v is said to be full rank if its frequency samples do not contain 0." In other words, the signal contains n_0 nonzero frequency components.

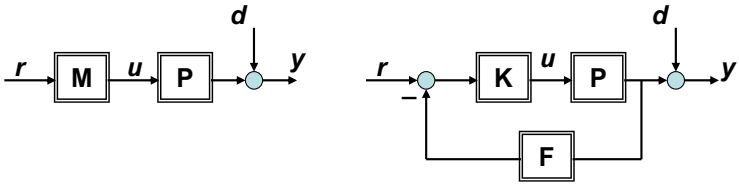
2.4 System Configurations

The basic system (2.1) is a typical open-loop identification setting in which the input can be selected directly by the user and the output noise is the

only disturbance. Practical systems are far more complicated in which a system to be identified is often a subsystem interconnected in different system configurations. Consequently, the input to the plant may not be directly accessible for design and there may be multiple noise corruptions. In this book, we recognize and treat some of these configurations.

2.4.1 Filtering and Feedback Configurations

Consider the system configurations in Figure 2.2. The filtering configuration is an open-loop system where M is linear, time invariant, and stable but may be unknown. The feedback configuration is a general structure of two-degree-of-freedom controllers where K and F are linear, time invariant, may be unstable, but are stabilizing for the closed-loop system. The mapping from r to u is the stable system $M = K/(1 + PKF)$. When $K = 1$, it is a regulator structure, and when $F = 1$, it is a servo-mechanism or tracking structure. Note that system components M, K, F are usually designed for achieving other goals and cannot be tuned for identification experiment design.



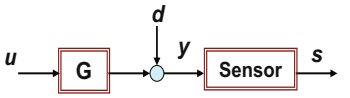
(a) Filtering Configuration **(b) Feedback Configuration**

FIGURE 2.2. Typical system configurations

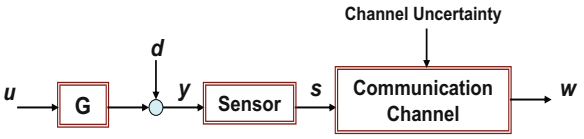
In these configurations, the input u to the plant P may be measured but cannot be directly selected. Only the external input r can be designed.

2.4.2 Systems with Communication Channels

The parameters of the system G in Figure 2.3 are to be identified. Two scenarios of system configuration are considered. System identification with quantized sensors is depicted in Figure 2.3(a) in which the observations on u_k and s_k are used. On the other hand, when sensor outputs of a system are transmitted through a communication channel and observed after transmission, the system parameters must be estimated by observing u_k and w_k, as shown in Figure 2.3(b). Since communication channels are subject to channel uncertainties, such as channel noises, the identification of system parameters is influenced by channel descriptions. Furthermore, when the channel contains unknown parameters such as unknown noise distribution functions, they must be incorporated into the overall identification problem.



(a) System identification with quantized observations



(b) System identification with communication channels

FIGURE 2.3. System configurations

2.5 Uncertainties

The main purpose of system identification is to reduce uncertainty on the system by using information from input-output observations. There are multiple sources of uncertainty in system configuration, modeling, and environments that will have a significant bearing on system identification. In addition to the unknown model parameters that are to be identified, we list below some of the uncertainties to be considered in this book.

2.5.1 System Uncertainties: Unmodeled Dynamics

In practical applications, a dynamic system is usually infinite dimensional. For cost reduction on system analysis, design, and implementation, it is desirable to use a low-order model to represent the system. A typical case is the IIR system (2.2)

$$\sum_{n=0}^{n_0-1} a_n u_{k-n} + \sum_{n=n_0}^{\infty} a_n u_{k-n} = \phi_k'\theta + \widetilde{\phi}_k'\widetilde{\theta}.$$

Here, the unmodeled dynamics $\widetilde{\theta}$ impact the system output through the term $\widetilde{\phi}_k'\widetilde{\theta}$. The unmodeled dynamics are characterized by certain saline features: (1) They are unknown but bounded when the system is stable. (2) They are better modeled as a deterministic uncertainty since they usually do not change randomly with time. (3) It is a multiplicative uncertainty,

in contrast to observation noise, which is commonly additive. While an additive uncertainty can be reduced by employing larger signals (namely a larger signal-to-noise ratio), a multiplicative uncertainty cannot be reduced by signal scaling. Consequently, its fundamental impact and reduction must be analyzed through different means. (4) The size of the unmodeled dynamics, $\sum_{n=n_0}^{\infty} |a_n|$, is a monotonically decreasing function of the model complexity n_0.

2.5.2 System Uncertainties: Function Mismatch

When a system model involves a nonlinear function $g(x)$, such as in the Wiener and Hammerstein models, the nonlinear function is often parameterized by $g(x;\mu)$ with a finite parameter vector μ. This parameterization introduces model mismatch:

$$\delta(x;\mu) = g(x) - g(x;\mu).$$

Function mismatch is similar to multiplicative uncertainty, albeit in a nonlinear form, in which the reduction of $\delta(x,\mu)$ cannot be achieved in general by signal scaling.

2.5.3 Sensor Bias and Drifts

A quantized sensor is characterized by its thresholds. In many applications, sensor thresholds may not be exactly known or change with time. System identification in which sensor thresholds are unknown must consider the thresholds as part of unknown parameters to be identified. Including thresholds in parameter vectors inevitably leads to a nonlinear structure.

2.5.4 Noise

The additive observation noise d_k in (2.1) may be modeled either as an unknown-but-bounded noise in a deterministic framework or as a random process in a stochastic framework.

In a deterministic framework, prior information on $\{d_k\}$ is limited to its uniform bound $|d_k| \le \delta$. In other words, the uncertainty set is $\Gamma_d = \{d_k \in \mathbb{R} : |d_k| \le \delta\}$. Identification errors resulting from this type of observation noise are characterized by the worst-case bound over all possible noises in Γ_d.

In contrast, in a stochastic framework, $\{d_k\}$ is described as a random process. Typical cases include independent and identically distributed (i.i.d.) processes whose probability distribution function of d_1 is $F(\cdot)$, and mixing processes for dependent noises.

2.5.5 Unknown Noise Characteristics

Identification methodologies and algorithms in this book utilize extensively information on noise distribution functions $F(\cdot)$ or density functions $f(\cdot)$. Such functions may not be available or may be subject to deviations and drifting. Unknown noise distribution functions compel their inclusion in identification. While it is possible to model distribution and density functions either parametrically or nonparametrically, the parameterization approach is more consistent with the methods of this book. As a result, in this book, $F(\cdot)$ is represented by a parameterized model $F(\cdot;\mu)$ when it is unknown.

2.5.6 Communication Channel Uncertainties

In networked systems, sensor outputs are not directly measured, but rather are transmitted through a communication channel, shown in Figure 2.3(b). When s_k is transmitted through a communication channel, the received sequence w_k is subject to channel noise and other uncertainties. When identification must be performed with observations on w_k, instead of s_k, channel noise and uncertainties will directly influence identification accuracy and convergence rates.

2.6 Notes

The basic system configurations presented in this chapter are representative in control systems, although they are not exhaustive. There are many other system settings that can be considered when essential issues are understood from the basic configurations. Quantization is an essential part of digital signals that have been extensively studied, mostly in uniformly spaced quantization. Its theoretical foundation and main properties have been studied in a different context beyond sampled date systems; see [1, 34, 80]. Quantized information processing occurs in many different applications beyond system identification [13, 38, 39, 73, 96]. Identification input design is an integral part of an identification experiment. This book is mostly limited to periodic signals due to their unique capability in providing input richness, in simplifying identification problems, and in their invariance when passing through systems. Most textbooks on system identification under stochastic formulations contain some discussions on uncertainties; see, for example, [17, 62]. The worst-case identification under set membership uncertainty is covered comprehensively in [66, 67, 68]. Models of noisy communication channels and their usage in information theory can be found in [22].

Part II

Stochastic Methods for Linear Systems

3

Empirical-Measure-Based Identification: Binary-Valued Observations

This chapter presents a stochastic framework for systems identification based on empirical measures that are derived from binary-valued observations. This scenario serves as a fundamental building block for subsequent studies on quantized observation data.

In a stochastic framework, we work with a probability space $(\Omega, \mathcal{F}, P)$. Our study of system identification for binary-valued and quantized observations is largely based on the application of empirical measures. It is deeply rooted in the law of large numbers and the central limit theorem. To a large extent, Part II is concerned with empirical processes associated with the identification task, their convergence, their moment behaviors, and associated efficiency issues.

The rest of the chapter is arranged as follows. We begin in Section 3.1 with an overview of our key methodology, its fundamental capability, and the main issues that limit its applications and their remedies. Section 3.2 extends discussions on empirical measures and their utility in system identification with binary-valued observations. Then Section 3.3 provides the almost sure convergence of the empirical processes for our identification problems. Section 3.4 is concerned with the scaled sequence of the empirical processes and their limits. Section 3.5 focuses on mean-square convergence of our identification problems with binary observations. Empirical measure based identification algorithms are strongly convergent under a broad class of noises. Section 3.6 summarizes and illustrates convergence under dependent noises of the basic algorithm of this chapter. To assist the reading and to streamline the presentation, some technical proofs are relegated to Section 3.7.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_3,

3.1 An Overview of Empirical-Measure-Based Identification

We first outline the basic ideas of using empirical measures for identifying a constant. The detailed results will be presented later. Consider a noise-corrupted sequence $\{y_k\}$:

$$y_k = \theta + d_k,$$

where θ is the parameter of interest, and $\{d_k\}$ is a sequence of independent and identically distributed random variables with a known distribution function $F(\cdot)$ whose density function $f(x) \neq 0$ and is continuously differentiable. This implies that F has a continuous inverse. y_k is observed via a binary-valued sensor of threshold C, and $s_k = \mathcal{S}(y_k)$.

Since

$$p = P\{s_k = 1\} = P\{\theta + d_k \leq C\} = P\{d_k \leq C - \theta\} = F(C - \theta),$$

and F is invertible, we have a relationship

$$\theta = C - F^{-1}(p).$$

Consequently, if $\widehat{p}$ is an estimate of p, then

$$\widehat{\theta} = C - F^{-1}(\widehat{p})$$

is a natural choice for an estimate of θ. In our method, $\widehat{p}$ is obtained by the empirical measure

$$\xi_N = \frac{1}{N} \sum_{k=0}^{N-1} s_k,$$

which leads to

$$\widehat{\theta}_N = C - F^{-1}(\xi_N).$$

Since ξ_N converges to p w.p.1, and F^{-1} is continuous, it is expected that $\widehat{\theta}_N$ will converge to θ w.p.1.

Implementation of this algorithm encounters a technical issue. For any finite N, ξ_N takes the value 0 or 1 with a positive probability. Since $F(\cdot)$ is usually not invertible at these two points, the algorithm must be modified. It is noted that if the prior information on θ includes its upper and lower bounds, then the upper and lower bounds on $C - \theta$ are known. If $0 < F(C - \theta_{\max}) < F(C - \theta_{\min}) < 1$, noting that F is monotone, there exists a small constant z for which $0 < z < F(C - \theta) < 1 - z < 1$ for all θ within

the bounds. This enables us to modify the algorithm by defining

$$\xi_N = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{k=0}^{N-1} s_k, & \text{if } z \le \dfrac{1}{N}\displaystyle\sum_{k=0}^{N-1} s_k \le 1-z, \\ z, & \text{if } \dfrac{1}{N}\displaystyle\sum_{k=0}^{N-1} s_k < z, \\ 1-z, & \text{if } \dfrac{1}{N}\displaystyle\sum_{k=0}^{N-1} s_k > 1-z. \end{cases} \tag{3.1}$$

With this modification,

$$\widehat{\theta}_N = C - F^{-1}(\xi_N)$$

is well defined for all N.

Since $\xi_N \to p$ w.p.1 and $z < p < 1-z$, the bounds in (3.1) will not affect the sample path convergence. As a result, the strong convergence $\widehat{\theta}_N \to \theta$ w.p.1 is ensured; see Section 3.3. In addition, we will show in Sections 3.4 and 3.5 that $\widehat{\theta}_N$ also converges to θ in the mean-square sense and

$$NE(\widehat{\theta}_N - \theta)^2 \to \frac{F(C-\theta)(1-F(C-\theta))}{f^2(C-\theta)}.$$

We will establish in Chapter 6 that

$$\frac{F(C-\theta)(1-F(C-\theta))}{Nf^2(C-\theta)} \tag{3.2}$$

is the Cramér–Rao (CR) lower bound. Hence, this algorithm is asymptotically efficient.

Of course, these desirable convergence properties are so far only valid for identification of a constant. The above results, however, are fundamental. We will show in the subsequent chapters that in many system configurations, the identification of a complicate system, such as FIR systems in this chapter, rational systems in Chapter 5, systems with unknown noise distributions in Chapter 8, and Wiener and Hammerstein systems in Chapters 11 and 12, that involve many parameters and nonlinearity, can often be reduced to a finite set of identification problems for constants. Furthermore, strong convergence and asymptotical efficiency for identification of constants carry over to identification of the parameters for the more complicated systems. The main task in this endeavor is to find suitable input signals that can generate an invertible mapping that relates the parameters to the constants. Input design, establishment of the invertible mappings, and convergence properties constitute the main tasks of the remaining chapters.

Since the CR lower bound $F(C-\theta)(1-F(C-\theta))/(Nf^2(C-\theta))$ is independent of algorithms, it reveals the fundamental information on θ

that is contained in $\{s_k\}$. A closer look at the CR lower bound tells us the following informational aspects of quantized identification problems.

First, if $C - \theta$ lies outside the support of the density function $f(x)$ (the support of f is the set $\{x : f(x) \neq 0\}$), then $f(C - \theta) = 0$ and $F(C - \theta)(1 - F(C - \theta))/(Nf^2(C - \theta)) = \infty$. In other words, the data do not contain any *statistical information* on θ. When this happens, $s_k = 1$ for all k or $s_k = 0$ for all k. This, however, does not imply that the data do not contain other *nonstatistical information* on θ. By carefully designing the input, the nonstatistical information can be used to reduce uncertainty on θ, just as statistical information does, albeit in different senses. This leads to deterministic frameworks for quantized identification in Chapters 9 and 10. A typical scenario for this to happen is when the disturbance is bounded $|d_k| \leq \delta$, such as a uniform random variable. This issue becomes especially relevant when the disturbance is bounded and small. Interestingly, under this scenario the deterministic approach, which uses input design to extract nonstatistical information on the unknown parameters, becomes very effective, which will be explored in Part III. On the other hand, deterministic worst-case identification will leave an irreducible error and convergence will be lost. It is then intuitive that in such a scenario, it might be best to combine these two frameworks so that both nonstatistical and statistical information can be used collaboratively, to reduce uncertainty on θ first by a deterministic method, and then switch to a stochastic method to achieve convergence. This will be discussed in Chapter 9.

Even when $f(C - \theta) \neq 0$, one must pay attention to the actual values of the CR bound. For example, if the noise is a standard Gaussian noise (mean zero and variance one), the CR bounds as a function of $C - \theta$ are listed in the following table.

$x = C - \theta$	0	1	2	3	4	5	6
$\frac{F(x)(1-F(x))}{f^2(x)}$	1.57	2.28	7.63	69	1768	1.3×10^5	2.7×10^7

Only when the threshold C is close to the true parameter θ, the convergence rate becomes fast enough for the data to be practically useful for parameter estimation. This implies that the threshold C must be correctly designed from the outset or adapted during identification. While detailed analysis and algorithm development are postponed until Chapter 14, we illustrate the idea of threshold adaptation by an example.

Denote

$$v = C - \theta.$$

To have the best CR lower bound in (3.2), we choose

$$v^* = \arg\min_v \frac{F(v)(1 - F(v))}{f^2(v)}. \tag{3.3}$$

Since F and f are known, v^* can be calculated off-line. The optimal threshold is related to v^* by $C^* = v^* + \theta$. Consequently, when C is adapted with value C_k at time k, one may simply choose

$$\widehat{\theta}_k = C_k - v^*.$$

Define $\mu = F(v^*)$. In particular, if d_k is Gaussian distributed with mean 0, then $v^* = 0$ and $\mu = 0.5$.

Consider the following stochastic approximation algorithm with a constant step size β:

$$\begin{aligned} s_k &= \begin{cases} 1, & \theta + d_k \le C_k, \\ 0, & \theta + d_k > C_k, \end{cases} \\ \xi_{k+1} &= \xi_k + \beta(s_k - \xi_k), \\ C_{k+1} &= C_k + \beta(v^* - F^{-1}(\xi_k)), \\ \widehat{\theta}_k &= C_k - v^*. \end{aligned} \tag{3.4}$$

The first line is the sensor characterization with a time-varying threshold C_k; the second is for empirical measure update in a recursive form with step size β; the third is threshold adaptation; and the last line is for parameter estimation. Convergence properties of this class of algorithms with threshold adaptation will be discussed in Chapter 14.

Example 3.1. Suppose $\theta = 100$. The noise is normally distributed with mean zero and variance $\sigma^2 = 100$. In this case, $v^* = 0$ and the optimal threshold is $C = \theta = 100$. The initial threshold is set at $C_0 = 40$. Under the constant step size $\beta = 0.05$, Figure 3.1 shows how the threshold is adapted and the parameter estimates move toward the true value.

3.2 Empirical Measures and Identification Algorithms

We begin with a single-input–single-output (SISO), linear time-invariant (LTI), stable, and discrete-time system

$$y_k = \sum_{i=0}^{\infty} a_i u_{k-i} + d_k, \qquad k = 1, 2, \ldots, \tag{3.5}$$

where $\{d_k\}$ is a sequence of random noises. System parameters a_i satisfy $\sum_{i=0}^{\infty} |a_i| < \infty$. The input u is uniformly bounded $|u_k| \le u_{\max}$ but can be arbitrarily selected otherwise. The output $\{y_k\}$ is measured by a binary-valued sensor with threshold C. We assume that for a given model order

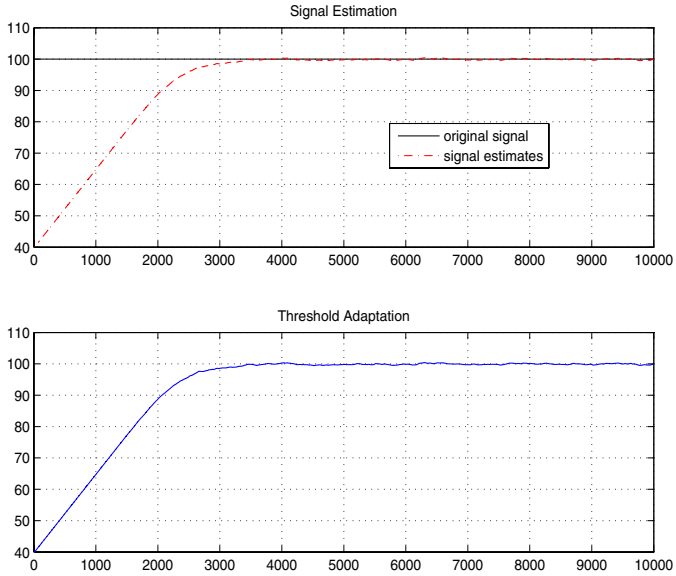


FIGURE 3.1. Threshold adaptation and estimation of a constant when $\beta = 0.05$. The x-axis is the number of algorithm iterations

n_0, the system parameters can be decomposed into the modeled part

$$\theta = [a_0, \ldots, a_{n_0-1}]'$$

and the unmodeled dynamics

$$\widetilde{\theta} = [a_{n_0}, a_{n_0+1}, \ldots]'.$$

Using such a setting, the system's input–output relationship becomes

$$y_k = \phi_k'\theta + \widetilde{\phi}_k'\widetilde{\theta} + d_k, \tag{3.6}$$

where

$$\phi_k = [u_k, u_{k-1}, \ldots, u_{k-(n_0-1)}]' \quad \text{and} \quad \widetilde{\phi}_k = [u_{k-n_0}, u_{k-n_0-1}, \ldots]'.$$

Under a selected input sequence $\{u_k\}$, the output s_k is measured. We would like to estimate θ on the basis of this input–output observation. To develop an estimator of the parameter θ and to analyze its asymptotic properties, we pose the following conditions.

(A3.1) $\{d_k\}$ is a sequence of i.i.d. random variables whose distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are twice continuously differentiable and known.

(A3.2) $\|\widetilde{\theta}\|_1 \leq \varepsilon_u$, where $\|\cdot\|_1$ is the l_1 norm (the subscript u in ε_u stands for "unmodeled").

For notational simplicity, assume the total data length is Nn_0 for some integer $N > 0$. As a result, we can group the input–output equations into N blocks of size n_0:

$$Y_l = \Phi_l\theta + \widetilde{\Phi}_l\widetilde{\theta} + D_l, \quad l = 0, 1, \ldots, N-1,$$

where

$$\begin{aligned} Y_l &= [y_{k_0+ln_0+1}, \ldots, y_{k_0+ln_0+n_0}]', \\ \Phi_l &= [\phi_{k_0+ln_0+1}, \ldots, \phi_{k_0+ln_0+n_0}]', \\ \widetilde{\Phi}_l &= [\widetilde{\phi}_{k_0+ln_0+1}, \ldots, \widetilde{\phi}_{k_0+ln_0+n_0}]', \\ D_l &= [d_{k_0+ln_0+1}, \ldots, d_{k_0+ln_0+n_0}]'. \end{aligned}$$

In particular, if the input is n_0-periodic, i.e., $u_{k+n_0} = u_k$, $\forall k$, we have

$$\Phi_l = \Phi_0 \quad \text{and} \quad \widetilde{\Phi}_l = \widetilde{\Phi}_0 = [\Phi_0, \Phi_0, \ldots] = \Phi_0[I, I, \ldots], \ l = 0, \ldots, N-1.$$

Moreover, if the n_0-period input is full rank (see Definition 2.1), then Φ_0 is invertible. As a result,

$$Y_l = \Phi_0\theta + \Phi_0[I, I, \ldots]\widetilde{\theta} + D_l = \Phi_0\widehat{\theta} + D_l,$$

where $\widehat{\theta} = \theta + [I, I, \ldots]\widetilde{\theta}$. In what follows, a scalar function applied to a vector will mean componentwise operation of the function.

For each θ (fixed but unknown) and $\widetilde{\theta}$, define

$$\xi_N^{\{i\}} = \frac{1}{N}\sum_{l=0}^{N-1} s_{k_0+ln_0+i}, \quad i = 0, \ldots, n_0-1, \tag{3.7}$$

$\xi_N = [\xi_N^{\{0\}}, \ldots, \xi_N^{\{n_0-1\}}]'$. Note that the event $\{y_{k_0+ln_0+i} \le C\}$ is the same as the event $\{d_{k_0+ln_0+i} \le v^{\{i\}}\}$, where $v^{\{i\}} = C - \gamma^{\{i\}}$ and $\gamma^{\{i\}}$ is the ith component of $\Phi_0\widehat{\theta}$. Denote

$$\begin{aligned} \gamma &= [\gamma^{\{0\}}, \gamma^{\{n_0-1\}}]' \in \mathbb{R}^{n_0}, \\ v &= [v^{\{0\}}, \ldots, v^{\{n_0-1\}}]' = C\mathbb{1} - \gamma. \end{aligned} \tag{3.8}$$

Then $\xi_N^{\{i\}}$ is precisely the value of the N-sample empirical distribution, denoted by $F_N(x)$, of the noise d_1 at $x = v^{\{i\}}$ and

$$\begin{aligned} \xi_N &= [F_N(v^{\{0\}}), \ldots, F_N(v^{\{n_0-1\}})]' \\ &= [F_N(C - \gamma^{\{0\}}), \ldots, F_N(C - \gamma^{\{n_0-1\}})]'. \end{aligned} \tag{3.9}$$

The identification algorithm for θ is given by

$$\begin{aligned} \zeta_N^{\{i\}} &= F^{-1}(\xi_N^{\{i\}}), \\ \zeta_N &= [\zeta_N^{\{0\}}, \ldots, \zeta_N^{\{n_0-1\}}]', \\ \widehat{\theta}_N &= \Phi_0^{-1}(C\mathbb{1} - \zeta_N). \end{aligned} \tag{3.10}$$

3.3 Strong Convergence

We are in a position to present a strong convergence result here. Recall that $F_N(x)$ denotes the N-sample empirical distribution. We first derive a general result of strong convergence of $F_N(\cdot)$, and then use it to obtain the strong convergence of the parameter estimator $\widehat{\theta}_N$ by noting the relationship given in (3.9).

Theorem 3.2. *Under condition* (A3.1)*, for any compact set* $S \subset \mathbb{R}$ *as* $N \to \infty$*,*

$$\sup_{x \in S} |F_N(x) - F(x)| \to 0 \quad w.p.1. \tag{3.11}$$

This theorem will be proved by using a result of the law of large numbers type. Let us first consider $\{Z_k\}$, a real-valued sequence of i.i.d. random variables with $E|Z_1| < \infty$. For each $z \in \mathbb{R}$, set

$$G_N(z) = \frac{1}{N} \sum_{i=1}^{N} I_{\{Z_i \le z\}}. \tag{3.12}$$

The dependence of $G_N(z)$ on the sample point $\omega \in \Omega$ may be written as $G_N(z, \omega)$. However, we often suppress the ω-dependence. The independence of $\{Z_k\}$ implies that of $\{I_{\{Z_k \le z\}}\}$. It is easily seen that the sequence has a finite mean. In fact, it is a sequence of i.i.d. Bernoulli random variables with $EI_{\{Z_k \le z\}} = G(z)$ the distribution function of Z_k evaluated at z. Thus, the well-known strong law of large numbers due to Kolmogorov [19, p. 123] yields that

$$G_N(z) \to G(z) \quad \text{w.p.1 as} \quad N \to \infty.$$

For a fixed z, our quest would stop here. Nevertheless, since z takes values in $(-\infty, \infty)$, naturally, we further seek uniform convergence of $G_N(\cdot) \to G(\cdot)$. As a prelude, we first present a proposition, which is known as the Glivenko–Cantelli theorem [8, p. 103]. It is also referred to by Loéve [63] as the "fundamental theorem of statistics." The proof of the result is postponed until Section 3.7.

Proposition 3.3. *Suppose that* $\{Z_k\}$ *is a sequence of i.i.d. random variables. Then, for any compact set* $S \subset \mathbb{R}$*,*

$$\sup_{z \in S} |G_N(z) - G(z)| \to 0 \quad w.p.1 \text{ as } \quad N \to \infty. \tag{3.13}$$

Proof of Theorem 3.2. It follows immediately from Proposition 3.3. □

Example 3.4. To illustrate convergence of empirical measures, consider a uniformly distributed noise on $[-1.2, 1.2]$. The actual distribution function is $F(z) = (z + 1.2)/2.4$. For different values of z, Figure 3.2 shows convergence of the empirical measures at several points in $[-1.2, 1.2]$ when the sample size is increased gradually from $N = 20$ to $N = 1000$.

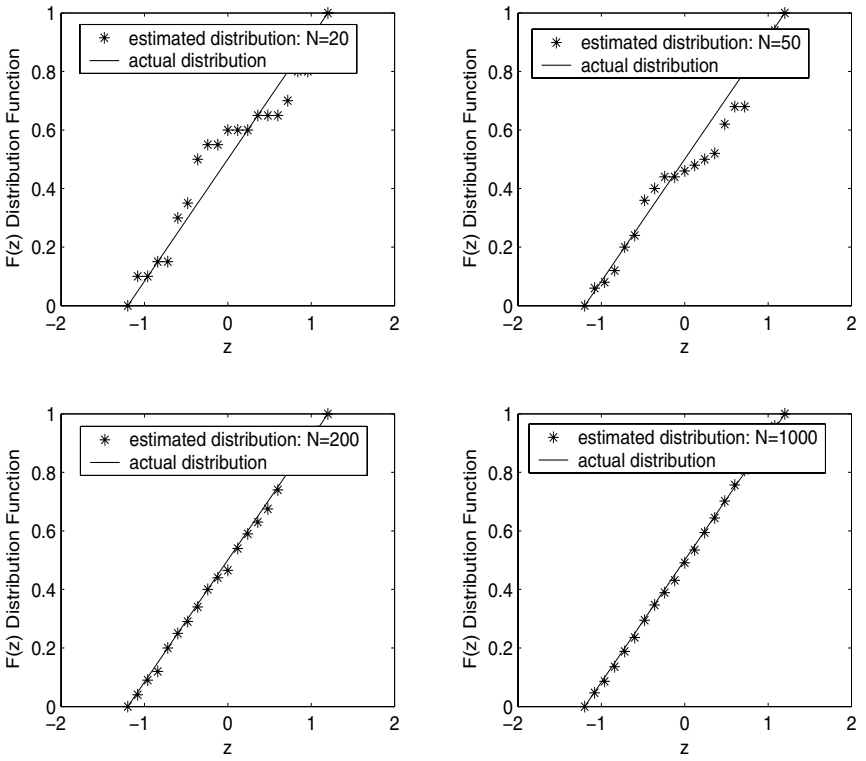


FIGURE 3.2. Convergence of empirical measures

With the preparations thus far, we proceed to analyze convergence properties of the estimator in (3.10). The next theorem establishes strong convergence of $\widehat{\theta}_N$. Recall that

$$\widehat{\theta} = \theta + [I, I, \ldots]\widetilde{\theta}.$$

Theorem 3.5. *Under Assumptions* (A3.1) *and* (A3.2)*, and assuming that the input* $\{u_k\}$ *is* n_0*-periodic and full rank, then*

$$\widehat{\theta}_N \to \widehat{\theta} \quad \textit{w.p.1 and} \quad \|\widehat{\theta} - \theta\|_1 \leq \varepsilon_u, \tag{3.14}$$

where $\|\cdot\|_1$ *denotes the* l_1 *norm. If* $\varepsilon_u = 0$*, i.e., no unmodeled dynamics, then the estimator is strongly consistent in that* $\widehat{\theta}_N \to \theta$ *w.p.1.*

Proof. By virtue of Theorem 3.2, as $N \to \infty$,

$$\xi_N \to [F(C - \gamma^{\{0\}}), \ldots, F(C - \gamma^{\{n_0-1\}})]' \quad \text{w.p.1.}$$

By the continuity of $F^{-1}(\cdot)$, we have that $F^{-1}(\xi_N^{\{i\}}) \to C - \gamma^{\{i\}}$ w.p.1. Due to the periodicity of the input u_k, γ is given as in (3.8). It follows that

$$\zeta_N \to C\mathbb{1} - \gamma = C\mathbb{1} - (\Phi_0\theta + \widetilde{\Phi}_0\widetilde{\theta}) \quad \text{w.p.1.}$$

Note that by the periodicity of u_k,

$$\widetilde{\Phi}_0 = [\widetilde{\phi}_1, \ldots, \widetilde{\phi}_{n_0}]' = [\Phi_0, \Phi_0, \ldots] = \Phi_0[I, I, \ldots].$$

That is,

$$\widehat{\theta}_N = \Phi_0^{-1}(C\mathbb{1} - \zeta_N) \to \theta + \Phi_0^{-1}\widetilde{\Phi}_0\widetilde{\theta} = \widehat{\theta} \quad \text{w.p.1.}$$

Moreover, by Assumption (A3.2),

$$\|\widehat{\theta} - \theta\|_1 = \|[I, I, \ldots]\widetilde{\theta}\|_1 \le \varepsilon_u.$$

The assertion is proved. □

3.4 Asymptotic Distributions

For the empirical measures defined in (3.7), we present a result of asymptotic distributions for a centered and scaled sequence. Define a sequence of scaled and centered estimation errors as

$$B_N(x) = \sqrt{N}(F_N(x) - F(x)), \quad \text{for each} \quad x \in \mathbb{R}. \tag{3.15}$$

To begin our query about asymptotic distributions of the empirical measures, let us first recall a few definitions and basic results. A random vector $x = [x^{\{1\}}, \ldots, x^{\{r\}}]'$ is said to be Gaussian if its characteristic function is given by

$$\phi(y) = \exp\left(\mathtt{i}\langle y, \mu\rangle - \frac{1}{2}\langle \Sigma y, y\rangle\right),$$

where $\mu \in \mathbb{R}^r$ is a constant vector, $\langle y, \mu\rangle$ is the usual inner product, $\mathtt{i}$ denotes the pure imaginary number satisfying $\mathtt{i}^2 = -1$, and Σ is a symmetric nonnegative definite $r \times r$ matrix. μ and Σ are the mean vector and covariance matrix of x, respectively.

Suppose that $X(t)$, $t \ge 0$, is a stochastic process (either real-valued, or vector-valued, or defined in a more general measurable space). By finite-dimensional distributions, we mean the joint distributions of

$$(X(t_1), X(t_2), \ldots, X(t_k))$$

for any choice of $0 \le t_1 < t_2 < \cdots < t_k$ and any $k = 1, 2, \ldots$

We say that a stochastic process $X(t)$, $t \ge 0$, is a Gaussian process, if any finite-dimensional distribution of $(X(t_1), X(t_2), \ldots, X(t_k))$ is Gaussian. A random process $X(\cdot)$ has *independent increments* if, for any $0 \le t_1 < t_2 < \cdots < t_k$ and $k = 1, 2, \ldots$,

$$X(t_1) - X(0),\ X(t_2) - X(t_1),\ \ldots,\ X(t_k) - X(t_{k-1})$$

are independent.

A sufficient condition for a process to be Gaussian (see Skorohod [85, p.7]) is: Suppose that the process $X(\cdot)$ has independent increments and continuous sample paths with probability one. Then $X(\cdot)$ is a Gaussian process.

An $\mathbb{R}^r$-valued random process $B(t)$ for $t \geq 0$ is a Brownian motion, if

- $B(0) = 0$ w.p.1;
- $B(\cdot)$ is a process with independent increments;
- $B(\cdot)$ has continuous sample paths with probability one;
- the increments $B(t)-B(s)$ have a Gaussian distribution with $E(B(t)-B(s)) = 0$ and $\text{Cov}(B(t), B(s)) = \Sigma|t-s|$ for some nonnegative definite $r \times r$ matrix Σ, where $\text{Cov}(B(t), B(s))$ denotes the covariance.

A process $B(\cdot)$ is said to be a standard Brownian motion if $\Sigma = I$. Note that a Brownian motion is necessarily a Gaussian process. For an $\mathbb{R}^r$-valued Brownian motion $B(t)$, the σ-algebra filtration generated by $B(\cdot)$ is denoted by $\mathcal{F}_t = \sigma\{B(s) : s \leq t\}$.

We next recall the notion of a Brownian bridge. Rather than proving its existence, we simply construct it from a Brownian motion. The definition is given below.

Definition 3.6. Let $B(\cdot)$ be a standard real-valued Brownian motion. A process $B^0(t)$ is a Brownian bridge process or tied-down Brownian motion if

$$B^0(t) = B(t) - tB(1), \ 0 \leq t \leq 1. \tag{3.16}$$

Note that a Brownian bridge is a function of a Brownian motion defined on $[0, 1]$. Loosely speaking, it is a Brownian motion tied down at both endpoints of the interval $[0, 1]$. Between the two endpoints, the process evolves just as a Brownian motion. Sometimes, when t is allowed to take real values outside $[0, 1]$, the Brownian bridge is said to be a stretched one. The terminology "stretched Brownian bridge" follows from that of [76]. Throughout the rest of the book, we will not use the modifier "stretched" but will still call it a Brownian bridge for simplicity.

We note the following properties of a Brownian bridge:

$$\begin{aligned} &B^0(0) = B(0) = 0 \ \text{ w.p.1}, \\ &E[B^0(t) - B^0(s)]^2 = (t-s)(1-(t-s)), \ \text{ if } \ s \leq t, \\ &E[B^0(s_2) - B^0(s_1)][B^0(t_2) - B^0(t_1)] \\ &\qquad = -(s_2 - s_1)(1-(t_2-t_1)), \ \text{ if } \ s_1 \leq s_2 \leq t_1 \leq t_2. \end{aligned} \tag{3.17}$$

To study the asymptotic distribution, we need to use the notion of weak convergence of probability measures. Let X_N and X be $\mathbb{R}^r$-valued random

variables. We say that X_N converges weakly to X if and only if for any bounded and continuous function $b(\cdot)$,

$$Eb(X_N) \to Eb(X).$$

$\{X_N\}$ is said to be tight if and only if for each $\eta > 0$, there is a compact set K_η such that

$$P(X_N \in K_\eta) \geq 1 - \eta \quad \text{for all} \quad N.$$

The definitions of weak convergence and tightness extend to random variables in a metric space. The notion of weak convergence is a substantial generalization of convergence in distribution. It implies much more than just convergence in distribution since $b(\cdot)$ can be chosen in many interesting ways. On a complete and separable metric space, the notion of tightness is equivalent to sequential compactness. This is known as Prohorov's theorem. By virtue of this theorem, we are able to extract convergent subsequences once tightness is verified. Let $D^r[0, \infty)$ denote the space of $\mathbb{R}^r$-valued functions that are right continuous and have left-hand limits, endowed with the Skorohod topology. For various notations and terminologies in the weak convergence theory such as the Skorohod topology, Skorohod representation etc., we refer the reader to [28, 55] and references therein; see also Section A.3 in the appendix of this book for a summary of results on weak convergence.

Remark 3.7. In the analysis to follow for empirical processes, one may focus on sequences of i.i.d. random variables with a uniform $[0, 1]$ distribution. The rationale is that arbitrary distributions can be treated by using the approach of inverse transformations. In fact, for a random variable with an arbitrary distribution function $G(\cdot)$, we can define a left-continuous inverse to $G(\cdot)$ as

$$G^{-1}(t) = \inf\{s : G(s) \geq t\} \quad \text{for} \quad 0 < t < 1.$$

Then $t \leq G(s)$ if and only if $G^{-1}(t) \leq s$. Now consider a sequence of i.i.d. random variables $\{Z_k\}$ for which Z_1 has distribution $G(\cdot)$. Let $\{Y_k\}$ be a sequence of i.i.d. uniform $[0, 1]$ random variables. Since

$$P(G^{-1}(Y_k) \leq t) = P(Y_k \leq G(t)) = G(t),$$

we may represent Z_k as $Z_k = G^{-1}(Y_k)$. Therefore, without loss of generality, suppose that $\{Z_k\}$ is a sequence of uniform $[0, 1]$ random variables in what follows.

Proposition 3.8. *Let $\{Z_k\}$ be a sequence of i.i.d. random variables uni-*

formly distributed on $[0,1]$. *For any* $t \in [0,1]$, *define*

$$\begin{aligned} G_N(t) &= \frac{1}{N}\sum_{i=1}^{N} I_{\{Z_i \le t\}}, \\ \widehat{B}_N(t) &= \sqrt{N}(G_N(t) - G(t)) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}[I_{\{Z_i \le t\}} - G(t)], \end{aligned} \tag{3.18}$$

where $G(\cdot)$ *is the distribution function of* Z_1. *Then* $\widehat{B}_N(\cdot)$ *converges weakly to* $B^0(\cdot)$, *the Brownian bridge process, with*

$$EB^0(t) = 0, \quad E[B^0(t)B^0(s)] = s(1-t), \ s \le t. \tag{3.19}$$

Remark 3.9. It is easily seen that the function $g(x) = \sup_{t\in[0,1]} |x(t)|$ is a continuous function. Thus, Proposition 3.8 yields that $\sup_{t\in[0,1]} |\widehat{B}_N(t)|$ converges in distribution to $\sup_{t\in[0,1]} |B^0(t)|$. Now using the property of a Brownian bridge (see (11.39) in [8]), we obtain that

$$P\left(\sup_{t\in[0,1]} |\widehat{B}_N(t)| \le a\right) \to 1 - 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp\left(-2k^2a^2\right), \quad \forall a \ge 0.$$

The proof of Proposition 3.8 is postponed until Section 3.7. With the help of this proposition and the inverse transformation mentioned in Remark 3.7, we can obtain the following result for our identification problems, whose proof is omitted.

Theorem 3.10. *Define* $B_N(x) = \sqrt{N}(F_N(x) - F(x))$. *Under condition* (A3.1), *for each* $v^{\{i\}}$ *defined in* (3.8) *with* $i = 0, \ldots, n_0 - 1$, $B_N(v^{\{i\}})$ *converges in distribution to* $B^0(v^{\{i\}})$, *where* $B^0(\cdot)$ *is the Brownian bridge process such that the covariance of* $B^0(\cdot)$ *is given by*

$$EB^0(x_1)B^0(x_2) = \min(F(x_1), F(x_2)) - F(x_1)F(x_2), \quad \forall x_1, x_2. \tag{3.20}$$

Remark 3.11. Note that $B_N(\cdot)$ converges weakly to $B^0(\cdot)$. By virtue of the Skorohod representation [55, p. 230] (with a slight abuse of notation), we may assume that $B_N(\cdot) \to B^0(\cdot)$ w.p.1 and the convergence is uniform on any compact set.

3.5 Mean-Square Convergence

This section focuses on the mean-square convergence of the parameter estimator $\widehat{\theta}_N - \widehat{\theta}$. Similarly to Theorem 3.5, we have the following result. With $\widehat{\theta}_N$ satisfying (3.14), denote the ith component of $\widehat{\theta}_N$ by $\widehat{\theta}_N^{\{i\}}$, of θ by $\theta^{\{i\}}$, and of $\widehat{\theta}$ by $\widehat{\theta}^{\{i\}}$, for $i = 0, \ldots, n_0 - 1$.

Theorem 3.12. *In addition to the assumptions of Theorem* 3.5*, suppose that the true parameter θ belongs to a compact set. Assume that for each $v^{\{i\}} = C - \gamma^{\{i\}}$ defined in* (3.8) *with $i = 0, \dots, n_0 - 1$, there is a neighborhood $\varrho(v^{\{i\}}, \varepsilon_i)$ (centered at $v^{\{i\}}$ with radius ε_i) of $v^{\{i\}}$ in which the second derivative of the distribution function $F(\cdot)$ exists and is continuous, and there is a neighborhood $\varrho(\xi^{\{i\}}, R_i)$ with $\xi^{\{i\}} = F(v^{\{i\}})$ in which the inverse $F^{-1}(\cdot)$ exists together with its first and second derivatives such that $(d^2/dv^2)F^{-1}(v^{\{i\}})$ is bounded. Then as $N \to \infty$,*

$$E(\sqrt{N}(\widehat{\theta}_N^{\{i\}} - \widehat{\theta}^{\{i\}}))^2 \to \frac{dF^{-1}(\xi^{\{i\}})}{d\xi^{\{i\}}} F(v^{\{i\}})(1 - F(v^{\{i\}})). \tag{3.21}$$

In particular, in case the unmodeled dynamics are absent, we have

$$E(\sqrt{N}(\widehat{\theta}_N^{\{i\}} - \theta^{\{i\}}))^2 \to \frac{dF^{-1}(\xi^{\{i\}})}{d\xi^{\{i\}}} F(v^{\{i\}})(1 - F(v^{\{i\}})).$$

This theorem will be proved when we invoke the following assertion. Since the following result will also be used later, we present it as a proposition.

Proposition 3.13. *Let $\{Z_k\}$ be a sequence of i.i.d. random variables with a common distribution function $G(\cdot)$. Let t_0 be fixed but otherwise arbitrary, and $t_0 \in K_c \subset \mathbb{R}$, where K_c is a compact subset of $\mathbb{R}$. Denote $p = G(t_0)$ and $q = 1 - p$. Assume that there is a neighborhood $\varrho(t_0, \epsilon_0)$ in which the second derivative of the distribution function $G(\cdot)$ exists and is continuous and that there is a neighborhood $\varrho(p, R_0)$ in which the inverse $G^{-1}(\cdot)$ exists together with its first and second derivatives such that $(d/dt)G^{-1}(\cdot)$ and $(d^2/dt^2)G^{-1}(\cdot)$ are bounded there. Define*

$$G_N = \frac{1}{N} \sum_{k=1}^{N} I_{\{Z_k \le t_0\}}, \quad x_N = G^{-1}(G_N). \tag{3.22}$$

Let $x = G^{-1}(p)$ belong to a compact subset of $\mathbb{R}$. Then

$$NE(x_N - x)^2 \to \frac{dG^{-1}(p)}{dp} pq \quad as \quad N \to \infty.$$

Proof. We divide the proof into several steps.

Step 1: Since $\{I_{\{Z_k \le t_0\}}\}$ is a sequence of i.i.d. Bernoulli random variables, clearly

$$EI_{\{Z_k \le t_0\}} = G(t_0) = p \quad \text{and} \quad E(I_{\{Z_k \le t_0\}} - p)^2 = pq.$$

Then the law of large numbers implies that $G_N \to p$ w.p.1. Moreover, $EG_N = p$ and $E(G_N - p)^2 = pq/N$. By the well-known central limit theorem,

$$\sqrt{N}(G_N - p)/\sqrt{pq} \to N(0, 1) \text{ in distribution as } N \to \infty.$$

By independence,

$$\sup_N E[\sqrt{N}(G_N - p)]^2 \quad = \sup_N N\frac{1}{N^2}\sum_{k=1}^{N} E[I_{\{Z_k \le t_0\}} - p]^2 = pq < \infty. \tag{3.23}$$

Thus $\{\sqrt{N}(G_N - p)\}$ is uniformly integrable. Likewise, it can be shown that $\{[\sqrt{N}(G_N - p)]^l\}$ is also uniformly integrable for each positive integer l.

Step 2: Note that x is in a compact set $K_c \subset [a, b]$ for some $a, b \in \mathbb{R}$. The monotonicity yields that for any $\eta > 0$,

$$0 < \eta < F(a) \le p = F(x) \le F(b) < 1 - \eta < 1.$$

In reference to the x_N defined in (3.22), define

$$\widetilde{G}_N = \begin{cases} G_N, & \eta \le G_N \le 1 - \eta, \\ \eta, & G_N < \eta, \\ 1 - \eta, & G_N > 1 - \eta. \end{cases} \tag{3.24}$$

Step 3: We demonstrate that $P(G_N \ne \widetilde{G}_N)$ is exponentially small. To do so, denote $X = (I_{\{Z_1 \le t_0\}} - p)/\sqrt{pq}$. Note that $EX = 0$ and $EX^2 = 1$. By the i.i.d. assumption, it is easily seen that the moment generating function of $\sqrt{N}(G_N - p)/\sqrt{pq}$ is

$$M_N(y) = E\left(\exp\left(\frac{yX}{\sqrt{N}}\right)\right)^N.$$

Taking a Taylor expansion of $M_N(z)$, we obtain

$$\begin{aligned} M_N(y) &= \left[E[1 + \frac{yX}{\sqrt{N}} + \frac{y^2X^2}{2N} + O(N^{-3/2})]\right]^N \\ &= \left[1 + \frac{y^2}{2N} + O(N^{-(3/2)})\right]^N. \end{aligned}$$

In the above, we have used the uniform integrability of $[\sqrt{N}(G_N - p)]^l$ provided in Step 1. Consequently, for any $\tau \in \mathbb{R}$,

$$\begin{aligned} &\inf_y [\exp(-y\tau)M_N(y)] \\ &= \inf_y \left[\exp(-y\tau)[1 + \frac{y^2}{2N} + O(N^{-(3/2)})]\right]^N \\ &\le K \exp\left(-\frac{\tau^2}{2}\right), \end{aligned} \tag{3.25}$$

where $K > 0$ is a positive constant.

By means of Chernoff's bound [83, p. 326], for any τ satisfying $-\infty < \tau \le p$,

$$\begin{aligned} P(G_N \le \tau) &= P\left(S_N \le N\frac{(\tau-p)}{\sqrt{pq}}\right) \\ &\le \left(\inf_y \left[\exp\left(-\frac{y(\tau-p)}{\sqrt{pq}}\right) M_N(y)\right]\right)^N, \end{aligned} \tag{3.26}$$

where

$$S_N = \sum_{k=1}^{N} \frac{I_{\{Z_k \le t_0\}} - p}{\sqrt{pq}}.$$

Then, for any $p \le \tau < \infty$,

$$P(G_N \ge \tau) \le \left(\inf_y \left[\exp\left(-\frac{y(\tau-p)}{\sqrt{pq}}\right) M_N(y)\right]\right)^N. \tag{3.27}$$

By (3.25), (3.26), and (3.27),

$$\begin{aligned} P(\widetilde{G}_N \ne G_N) &= P(\widetilde{G}_N \ne G_N;\ G_N \le \eta) + P(\widetilde{G}_N \ne G_N;\ G_N \ge 1-\eta) \\ &\le P(G_N \le \eta) + P(G_N \ge 1-\eta) \\ &\le K\exp\left(-\frac{(\eta-p)^2 N}{2pq}\right) + K\exp\left(-\frac{(1-\eta-p)^2 N}{2pq}\right). \end{aligned} \tag{3.28}$$

This verifies the claim in Step 3.

Step 4: Recall that $x_N = G^{-1}(\widetilde{G}_N)$. Since G^{-1} is continuous and $G_N \to p$ w.p.1, $x_N \to x$ w.p.1. Since $(d/dx)G^{-1}(x)$ and $(d^2/dx^2)G^{-1}(x)$ are bounded and continuous in the neighborhood $\varrho(p, R_0)$,

$$\begin{aligned} \sup_{x\in\varrho(p,R_0)} \left|\frac{dG^{-1}(x)}{dx}\right| &= \beta < \infty, \\ \sup_{x\in\varrho(p,R_0)} \left|\frac{d^2G^{-1}(x)}{dx^2}\right| &= \gamma < \infty. \end{aligned}$$

Then,

$$\begin{aligned} x_N &= EG^{-1}(\widetilde{G}_N) \\ &= \left[G^{-1}(p) + \frac{dG^{-1}(p)}{dp}(\widetilde{G}_N - p) + \frac{1}{2}\frac{d^2G^{-1}(\alpha_N)}{dp^2}(\widetilde{G}_N - p)^2\right] \\ &= G^{-1}(p) + \frac{dG^{-1}(p)}{dp}(G_N - p) \\ &\quad + \frac{1}{2}\frac{d^2G^{-1}(\alpha_N(\widetilde{G}_N))}{dp^2}(G_N - p)^2 + o(1), \end{aligned}$$

where α_N lies between $\widetilde{G}_N$ and p, and $o(1) \to 0$ w.p.1 as $N \to \infty$. Moreover, by Step 3, $P(\widetilde{G}_N \neq G_N) \to 0$ exponentially fast. Thus,

$$P(\alpha_N \notin [\eta, 1-\eta]) \to 0,$$

and as a result,

$$P(\alpha_N \notin [\eta, 1-\eta] \cap \varrho(p, R_0)) \to 0.$$

In addition,

$$\begin{aligned} & E\left|\frac{d^2 G^{-1}(\alpha_N)}{dx^2}(G_N - p)^2\right| \\ & \quad \leq KE(G_N - p)^2 \\ & \quad = \frac{1}{N} KEN(G_N - p)^2 \to 0 \end{aligned}$$

since $EN(G_N - p)^2 \to pq$. These together with the dominated convergence theorem imply the desired result. □

3.6 Convergence under Dependent Noise

Since it appears to be best to convey the main ideas without undue notational and technical complexity, most of the results in this book are stated under i.i.d. noises with finite variances, rather than considering sequences of random variables in most general forms or weakest conditions. For example, in the previous sections, we developed the convergence and rates of convergence (asymptotic distribution of estimation errors) of the empirical-measure-based algorithms under "white noise" sequences.

However, most results can be extended to much more general noise processes under much weaker conditions, without changing the algorithms. We shall illustrate such extensions by using a typical case of dependent noises, namely, the dependence asymptotically diminishing.

In lieu of (A3.1), assume that $\{d_k\}$ is a stationary sequence whose distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are twice continuously differentiable and known, and that $\{d_k\}$ is ϕ-mixing with mixing measure $\check{\psi}_k$; see [8, 28], among others, for a definition of ϕ-mixing processes. Some common examples of ϕ-mixing sequences are listed below.

- Recall that $\{d_k\}$ is said to be m-dependent if the random vectors $(d_i, d_{i+1}, \ldots, d_k)$ and $(d_{i+n}, d_{i+n+1}, \ldots, d_l)$ are independent whenever $n > m$. A particular example is

$$d_k = \sum_{i=0}^{m} c_i w_{k-i},$$

where c_i are constants and $\{w_k\}$ is a sequence satisfying (A3.1). For such a $\{d_k\}$, it is ϕ-mixing with mixing rate $\check{\psi}_k$ such that $\check{\psi}_k = 0$ for $k > m$.

- Suppose that $\{d_k\}$ is a finite state Markov chain with state space $\mathcal{M} = \{1, \ldots, m_0\}$ and a one-step transition matrix P such that the chain or its transition matrix is irreducible and aperiodic. Then the sequence is also ϕ-mixing. In fact, it is known that such a Markov chain is ergodic with stationary measure $\nu = (\nu_1, \ldots, \nu_{m_0})$. In this case, we have $|P^n - \nu| \leq \lambda^n$ for some $0 < \lambda < 1$. This spectrum gap conditions yields that the Markov chain $\{d_k\}$ is ϕ-mixing with exponential mixing rate.

Strong convergence of the empirical measure based identification algorithms holds under ϕ-mixing noises. In view of the discussions in Karlin and Taylor [47, p. 488], a stationary ϕ-mixing sequence is strongly ergodic. Consequently, a strong law of large numbers holds for the sequence. As a result, ξ_N defined in Section 3.1 is still strongly convergent. Together with the smoothness of the distribution function and its inverse mentioned above, the strong consistency of the parameter estimator of θ introduced in Section 3.1 remains true. In fact, the Glivenko–Cantelli theorem continues to hold, yielding the desired uniform convergence as before, so does the weak convergence to the Brownian bridge process.

To illustrate further, we can follow the ideas presented in Section 3.3, Section 3.4, and Section 3.5, with the modification of dependent noises. For any x, consider $F_N(x)$, the N-sample empirical distribution. Then, for any compact set $S \subset \mathbb{R}$, $\sup_{x \in S} |F_N(x) \to F(x)| \to 0$ w.p.1 as $N \to \infty$. To explore the rates of convergence (or to study the scaled estimation errors), for any $x \in \mathbb{R}$, write

$$s_{k,x} = I_{\{d_k \leq x\}},$$

denote

$$h_{k,x} = s_{k,x} - F(x),$$

and define $B_N(x) = \sqrt{N}(F_N(x) - F(x))$ as in (3.15). Then we can use the techniques as in [8, p. 197] to show that $B_N(\cdot)$ converges weakly to a Brownian bridge $B(\cdot)$, whose covariance is given by

$$EB(u)B(t) = E[h_{0,u}h_{0,t}] + \sum_{k=1}^{\infty} E[h_{0,u}h_{k,t}] + \sum_{k=1}^{\infty} E[h_{k,u}h_{0,t}],$$

provided the mixing measure satisfies

$$\sum_{k=0}^{\infty} k^2 \check{\psi}_k^{1/2} < \infty.$$

With such a framework of strong convergence and asymptotic distributions, we can proceed to the detailed studies as done in Sections 3.3–3.5.

There are no essential difficulties in applying the identification algorithms under the ϕ-mixing processes and/or other mixing processes as defined in

[58]; see also [41]. Furthermore, we may even extend the results to certain nonstationary processes as long as the underlying sequence verifies an ergodicity condition, the dependence and the correlation are asymptotically diminishing, and there is a limit function $F(\cdot)$ together with its inverse F^{-1} such that both F and F^{-1} are suitably smooth.

3.7 Proofs of Two Propositions

In this section, we provide a couple of technical complements. The proof of Proposition 3.3 is essentially in [65]. The reader is also referred to [19, 21, 84], among others, for additional reading. We adopt the usual definition that the distribution $F(\cdot)$ of a random variable is a right-continuous function with a left-hand limit. Note that in [19], a distribution function is defined to be left continuous, however.

Proof of Proposition 3.3

We use $G(t-)$ and $G(t+)$ to denote the limit from the left $[\lim_{x\to t-0} G(x)]$ and the limit from the right $[\lim_{x\to t+0} G(x)]$, respectively. By the strong law of large numbers, for each fixed $t \in \mathbb{R}$,

$$G_N(t) \to G(t) \ \text{ w.p.1 and } \ G_N(t-) \to G(t-) \ \text{ w.p.1.}$$

Let $t_{j,k}$ be the smallest real number t that satisfies

$$G(t-) \le k/j \le G(t+) = G(t) \ \text{ for } \ k = 1, 2, \ldots, j.$$

Then, the w.p.1 convergence implies that

$$G_N(t_{j,k}) \to G(t_{j,k}) \ \text{ w.p.1 and } \ G_N(t_{j,k}-) \to G(t_{j,k}-) \ \text{ w.p.1.}$$

Define

$$\begin{aligned} J_{j,k} &= \{\omega \in \Omega : G_N(t_{j,k}) \to G(t_{j,k}) \ \text{ as } \ N \to \infty\}, \\ J_{j,k}^- &= \{\omega \in \Omega : G_N(t_{j,k}-) \to G(t_{j,k}-) \ \text{ as } \ N \to \infty\}, \\ J_j &= \bigcap_{k=1}^{j} J_{j,k}, \ J_j^- = \bigcap_{k=1}^{j} J_{j,k}^-. \end{aligned}$$

The w.p.1 convergence implies that $P(J_j) = P(J_j^-) = 1$ for all j.

Set

$$J = \bigcap_{j=1}^{\infty} J_j, \ J^- = \bigcap_{j=1}^{\infty} J_j^-, \ J^0 = J \cap J^-.$$

We claim that $P(J^0) = 1$. First consider J. Note that for each j, J_j^c, the complement of J_j, has probability 0. Thus, for $J^c = \bigcup_{j=1}^{\infty} J_j^c$,

$$P(J^c) = P\Big(\bigcup_{j=1}^{\infty} J_j^c\Big) \le \sum_{j=1}^{\infty} P(J_j^c) = 0.$$

It follows that $P(J^c) = 0$, and hence, $P(J) = 1$. Likewise, $P(J^-) = 1$ and $P(J^0) = 1$.

Choose t and k such that $t_{j,k} \le t < t_{j,k+1}$. Then

$$\begin{aligned} &G_N(t_{j,k}) \le G_N(t) \le G(t_{j,k+1}-),\\ &G(t_{j,k}) \le G(t) \le G(t_{j,k+1}-),\\ &G(t_{j,k+1}-) - G(t_{j,k}) \le \frac{1}{j}. \end{aligned}$$

Set

$$X_{j,N} = \max_{1\le k\le j} \{|G_N(t_{j,k}) - G(t_{j,k})|, |G_N(t_{j,k}-) - G(t_{j,k}-)|\}.$$

Then

$$\begin{aligned} G_N(t) - G(t) &\le G_N(t_{j,k+1}-) - G(t_{j,k})\\ &\le G_N(t_{j,k+1}-) - G(t_{j,k+1}-) + \frac{1}{j}\\ &\le G_N(t_{j,k}-) - G(t_{j,k}-) + \frac{1}{j}. \end{aligned}$$

Thus, for each $t \in \mathbb{R}$,

$$|G_N(t) - G(t)| \le X_{j,N} + \frac{1}{j}. \tag{3.29}$$

Finally, let $X_N = \sup_{t\in\mathbb{R}} |G_N(t) - G(t)|$. Note that for each j, the definitions of $J_{j,k}$ and $J_{j,k}^-$ imply that

$$J_j = \left\{\omega \in \Omega : \max_{|G_N(t_{j,k}) - G(t_{j,k})| \to 0}\right\}$$

and

$$J_j^- = \left\{\omega \in \Omega : \max_{|G_N(t_{j,k}-) - G(t_{j,k}-)| \to 0}\right\}.$$

Then, the event $J \cap J^-$ happens if and only if $X_{j,N} \to 0$ for each j. Moreover, if $\omega \in J \cap J^-$, then $\omega \in \{\omega : X_N \to 0\}$ by (3.29). Thus,

$$1 = P(J \cap J^-) \le P(\{\omega : X_N \to 0\}).$$

The desired result then follows. □

Proof of Proposition 3.8

This section provides a proof of the weak convergence of a scaled sequence of empirical processes. We will prove the assertion by carrying out the following steps. In the first step, we show that $G_N(\cdot)$ is tight; in the second step, we show the finite-dimensional distributions converge and identify the limit process.

Step 1: Tightness. Note that $t = G(t) = EI_{\{Z_k \le t\}}$ for each $t \in [0,1]$, and that for each $\Delta > 0$ and $0 \le s \le \Delta$, $G(t+s) - G(t) = s$. By the independence of the sequence $\{Z_k\}$ and hence that of $I_{\{Z_k \le t\}}$, we have

$$\begin{aligned} & E|\widehat{B}_N(t+s) - \widehat{B}_N(t)|^2 \\ &= E\left|\frac{1}{\sqrt{N}}\sum_{k=1}^{N}(I_{\{Z_k \le t+s\}} - I_{\{Z_k \le t\}} - s)\right|^2 \\ &= \frac{1}{N}\sum_{k=1}^{N} E[(I_{\{Z_k \le t+s\}} - I_{\{Z_k \le t\}})^2 - 2E(I_{\{Z_k \le t+s\}} - I_{\{Z_k \le t\}})s + s^2] \\ &= s(1-s) \le \Delta. \end{aligned} \tag{3.30}$$

Thus,

$$\lim_{\Delta \to 0} \limsup_{N} E|\widehat{B}_N(t+s) - \widehat{B}_N(t)|^2 = \lim_{\Delta \to 0} \Delta = 0.$$

This together with the tightness criterion (see [53, p. 47] and also [55, Chapter 7]) yields that $\widehat{B}_N(\cdot)$ is tight.

Step 2: Convergence of finite-dimensional distributions. First, observe that $NG_N(t)$ represents the number of points among $Z_1, \ldots, Z_N$ satisfying $Z_i \le t$. That is, $NG_N(t)$ is a frequency count. Let $0 = t_0 < t_1 < \cdots < t_k = 1$ be an arbitrary partition of $[0,1]$. Then for $i = 1, \ldots, k$, $NG_N(t_i) - NG_N(t_{i-1})$ follow a multinomial distribution with parameters N and $p_i = t_i - t_{i-1}$. The well-known central limit theorem then implies that the vector-valued sequence

$$\widetilde{B}_N = \begin{pmatrix} \widehat{B}_N(t_1) - \widehat{B}_N(t_0) \\ \widehat{B}_N(t_2) - \widehat{B}_N(t_1) \\ \cdots \\ \widehat{B}_N(t_k) - \widehat{B}_N(t_{k-1}) \end{pmatrix}$$

converges in distribution to a normal random vector. Let us examine its ith component. We have for each $i = 1, \ldots, k$,

$$e_i' \widetilde{B}_N = \widehat{B}_N(t_i) - \widehat{B}_N(t_{i-1}) = \frac{1}{\sqrt{N}}(NG_N(t_i) - NG_N(t_{i-1}) - Np_i),$$

where e_i denotes the ith standard unit vector. It follows that $e_i'\widehat{B}_N$ converges in distribution to a normal random variable with mean 0 and variance $p_i(1-p_i)$ and covariance $-p_ip_j$. Thus, Definition 3.6 and (3.17) imply that the limit is $B^0(t_i) - B^0(t_{i-1})$.

It is also easily seen that $E\widehat{B}_N(t) = 0$ for $t \in [0,1]$ and similarly to (3.30), for $s, t \in [0,1]$ with $s \le t$,

$$\begin{aligned}\mathrm{Cov}[\widehat{B}_N(t), \widehat{B}_N(s)] &= E[\widehat{B}_N(t)\widehat{B}_N(s)] \\ &= \frac{1}{N}\sum_{k=1}^{N} E(I_{\{Z_k \le t\}} - (t))(I_{\{Z_k \le s\}} - s) \\ &= s - ts = s(1-t).\end{aligned}$$

As a result, the limit of the covariance is also given by $s(1-t)$. Thus, the desired Brownian bridge limit is obtained. □

3.8 Notes

The subject matter of this chapter is about empirical measures. An excellent survey containing many results can be found in Shorack and Wellner in [84]. Related work can also be found in [76]. Extensive studies on weak convergence are contained in [8, 28, 55] and references therein. Related results on almost sure convergence can also be found in [87].

The study for identification of systems with binary observations was initiated in our work [111]. This line of research has been continued in [108, 109, 110]. Subsequently, quantized observations are treated. In [104], efficiency of empirical measure-based algorithms was established.

For some related but different identification algorithms such as binary reinforcement and some applications, the reader is referred to [2, 13, 18, 27, 29, 33, 73, 119]. The main tools for stochastic analysis and identification methodologies can be found in [8, 17, 30, 31, 55, 62, 76, 83].

When the disturbance has a finite support, i.e., the density $f_d(x) = 0$, $x < -\kappa$ or $x > \kappa$ with a finite κ, the corresponding $F(x)$ is not invertible outside the interval $[-\kappa, \kappa]$. The results in this section are not applicable if $C - \theta \notin [-\kappa, \kappa]$. Consequently, the identification capability of the binary sensor will be reduced. In other words, it is possible that for a selected input, s_k is a constant (0 or 1) for all k; hence, no information is obtained. One possible remedy for this situation is to add a dither to the sensor input. Hence, assume the disturbance d_k contains two parts: $d_k = d_k^0 + h_k$, where d_k^0 is an i.i.d. disturbance with density f_0 and h_k is an i.i.d. stochastic dither, independent of d^0, with density f_h. In this case, the density f_d of d is the convolution: $f_d = f_0 * f_h$. By choosing an appropriate f_h, f_d will have a larger support and possess the desired properties for system identification. Another approach is to combine deterministic and stochastic

methods, which will be covered in Chapter 9. Furthermore, one may adapt C to make $C - \theta \in (-\kappa, \kappa)$. This is discussed in Chapter 14.

4

Estimation Error Bounds: Including Unmodeled Dynamics

In Chapter 3, we derived convergent estimators of the system parameters using binary-valued observations. Our aim here is to obtain further bounds on estimation errors from unmodeled dynamics. In this book, unmodeled dynamics are treated as a deterministic uncertainty which is unknown but has a known bound in an appropriate space. Due to the coexistence of deterministic uncertainty from unmodeled dynamics and stochastic disturbances, we are treating necessarily a mixed environment. Consequently, estimation error characterization has a probabilistic measure that is compounded with a worst-case scenario over unmodeled dynamics, an idea introduced in our earlier work [101, 102].

In Section 4.1, we formulate system identification problems with binary-valued observations and under mixed deterministic and stochastic uncertainties. An identification error characterization is defined. Section 4.2 derives upper bounds on estimation errors. Here one of the techniques used is the large deviations type of upper bounds. Based on the bounds, we obtain further properties of the estimation algorithms. Section 4.3 is devoted to lower bounds on estimation errors. Lower bounds from normal distributions are used first. Then a class of identification problems is treated using the central limit theorem and normal approximation techniques. The lower bounds provide some basic understanding of the information contents and complexity aspects of our identification problems.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_4,

4.1 Worst-Case Probabilistic Errors and Time Complexity

Recall the linear system given in (3.5):

$$y_k = \sum_{i=0}^{\infty} a_i u_{k-i} + d_k = \phi_k' \theta + \widetilde{\phi}_k' \widetilde{\theta} + d_k, \qquad k = 0, 1, 2, \ldots, \tag{4.1}$$

where d_k and $\widetilde{\theta}$ satisfy Assumptions (A3.1) and (A3.2).

The following framework was introduced in [101, 102]. It treats unmodeled dynamics as an unknown-but-bounded uncertainty. On the other hand, random disturbances are stochastic processes. Consequently, a worst-case probability measure is used to evaluate error bounds. For a given set $\mathbb{L}(k_0, u, N)$ of admissible estimates $\widehat{\theta}_N$ of the true parameter θ, on the basis of N measurements on s_k starting at k_0 with input u_k, and an error tolerance level ε, we define

$$\lambda_N(\varepsilon) = \inf_{\|u\|_\infty \le u_{\max}} \sup_{k_0} \inf_{\widehat{\theta}_N \in \mathbb{L}(k_0,u,N)} \sup_{\|\widetilde{\theta}\|_1 \le \eta \varepsilon_u} P\left(\|\widehat{\theta}_N - \theta\|_1 \ge \varepsilon \right). \tag{4.2}$$

This is the optimal (over the input u and admissible estimate $\widehat{\theta}_N$) worst-case (over $\widetilde{\theta}$ and k_0) probability of errors larger than the given level ε. By considering the worst case of k_0, we are dealing with a "persistent identification" problem, a concept introduced in [97]. Then, for a given confidence level $\alpha \in [0, 1)$,

$$N_\alpha(\varepsilon) = \min \{N : \ \lambda_N(\varepsilon) \le \alpha\} \tag{4.3}$$

is the probabilistic time complexity. It is noted that if $\alpha = 0$, $N_\alpha(\varepsilon)$ is reduced to (modulo a set of probability measure 0) deterministic worst-case time complexity for achieving estimation accuracy ε. We will derive upper and lower bounds on $\lambda_N(\varepsilon)$ and $N_\alpha(\varepsilon)$ in subsequent sections.

4.2 Upper Bounds on Estimation Errors and Time Complexity

To derive an upper bound, we select a specific input, usually an n_0-periodic input of full rank. Under such an input, $\widehat{\theta} = \theta + [I, I, \ldots]\widetilde{\theta}$ and the identification algorithm (3.10) will be used. We shall establish bounds on identification errors and time complexity for a finite N. For a fixed $N > 0$,

$$\begin{aligned} \widehat{\theta}_N - \widehat{\theta} &= \Phi_0^{-1} \left(C\mathbb{1} - \zeta_N \right) - \left(\theta + [I, I, \ldots]\widetilde{\theta} \right) \\ &= \Phi_0^{-1} \left(C\mathbb{1} - \zeta_N - \left(\Phi_0 \theta + \widetilde{\Phi}_0 \widetilde{\theta} \right) \right) \\ &= \Phi_0^{-1} \left(v - \zeta_N \right), \end{aligned} \tag{4.4}$$

where ζ_N is defined in (3.10) and $v = C\mathbb{1} - (\Phi_0\theta + \widetilde{\Phi}_0\widetilde{\theta})$. Since

$$\|\widehat{\theta}_N - \widehat{\theta}\|_1 \le \|\Phi_0^{-1}\|_I \|v - \zeta_N\|_1,$$

where $\|\cdot\|_I$ is the l_1-induced operator norm, for any $\varepsilon_1 > 0$,

$$\begin{aligned} P\left(\|\widehat{\theta}_N - \widehat{\theta}\|_1 \ge \varepsilon_1\right) &\le P\left(\|\zeta_N - v\|_1 \ge \frac{\varepsilon_1}{\|\Phi_0^{-1}\|_I}\right) \\ &\le P\left(\|\zeta_N - v\|_\infty \ge \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right) \\ &\le P\left(\bigcup_{i=0}^{n_0-1}\left\{|\zeta_N^{\{i\}} - v^{\{i\}}| \ge \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right\}\right) \\ &\le \sum_{i=0}^{n_0-1} P\left(|\zeta_N^{\{i\}} - v^{\{i\}}| \ge \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right). \end{aligned}$$

The inequality

$$|\zeta_N^{\{i\}} - v^{\{i\}}| \ge \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}$$

is equivalent to

$$\zeta_N^{\{i\}} \ge v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I} \quad \text{or} \quad \zeta_N^{\{i\}} \le v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}. \tag{4.5}$$

Note that $\zeta_N^{\{i\}} = F^{-1}(\xi_N^{\{i\}})$. Since $F^{-1}(\cdot)$ is monotone,

$$\zeta_N^{\{i\}} \ge v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I} \quad \Leftrightarrow \quad \xi_N^{\{i\}} \ge F\left(v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right) \tag{4.6}$$

and

$$\zeta_N^{\{i\}} \le v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I} \quad \Leftrightarrow \quad \xi_N^{\{i\}} \le F\left(v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right). \tag{4.7}$$

It follows that

$$\begin{aligned} P\left(\left|\widehat{\theta}_N - \widehat{\theta}\right| \ge \varepsilon_1\right) &\le \sum_{i=0}^{n_0-1} P\left(|\zeta_N^{\{i\}} - v^{\{i\}}| \ge \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right) \\ &\le \sum_{i=0}^{n_0-1} P\left(\xi_N^{\{i\}} \ge F\left(v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right) \\ &\quad + \sum_{i=0}^{n_0-1} P\left(\xi_N^{\{i\}} \le F\left(v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right). \end{aligned} \tag{4.8}$$

For simplicity, we adopt the short-hand notation $s_l^{\{i\}} = s_{k_0+ln_0+i}$. Since $\{d_k\}$ is a sequence of i.i.d. random variables, for each $i = 0, 1, \ldots, n_0 - 1$,

$\{s_l^{\{i\}}\}$ is also a sequence of i.i.d. random variables with respect to l. Denote the moment generating function of $s_0^{\{i\}}$ by $G_i(z) = E\exp(zs_0^{\{i\}})$ with $z \in \mathbb{R}$. Let

$$g_i(t) = \inf_z E\exp(z(s_0^{\{i\}} - t)) = \inf_z \exp(-zt)G_i(z).$$

By the definition of $s_0^{\{i\}}$, $Es_0^{\{i\}} = F(v^{\{i\}})$. By the monotonicity of $F(\cdot)$, we have

$$F\left(v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right) > F\left(v^{\{i\}}\right)$$

and

$$F\left(v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right) \le F(v^{\{i\}}).$$

Consequently, an application of Chernoff's inequality ([83, p. 326]) yields

$$P\left(\xi_N^{\{i\}} \ge F\left(v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right) \le \left(g_i\left(v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right)^k \quad (4.9)$$

and

$$P\left(\xi_N^{\{i\}} \le F\left(v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right) \le \left(g_i\left(v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right)^k. \quad (4.10)$$

Combining (4.8), (4.9), and (4.10), we obtain the following upper bounds.

Theorem 4.1. *For any $\varepsilon_1 > 0$,*

$$P\left(\left|\widehat{\theta}_N - \widehat{\theta}\right| \ge \varepsilon_1\right) \le H_{\varepsilon_1,N}, \quad (4.11)$$

where

$$H_{\varepsilon_1,N} = \sum_{i=0}^{n_0-1}\left[\left(g_i\left(v^{\{i\}} + \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right)^N + \left(g_i\left(v^{\{i\}} - \frac{\varepsilon_1}{n_0\|\Phi_0^{-1}\|_I}\right)\right)^N\right].$$

Corollary 4.2. *For any $\varepsilon > \varepsilon_u > 0$ with ε_u given in Assumption* (A3.2)*, we have*

(a)

$$\lambda_N(\varepsilon) \le H_{\varepsilon-\varepsilon_u,N}, \quad (4.12)$$

where $\lambda_N(\varepsilon)$ is defined in (4.2) *and $H_{\varepsilon-\varepsilon_u,N}$ in Theorem 4.1.*

(b)

$$N_\alpha(\varepsilon) \le n_0 \min\{N : H_{\varepsilon-\varepsilon_u,N} \le \alpha\}. \quad (4.13)$$

Proof. To prove part (a), by Theorem 4.1 the selected input and the estimate $\widehat{\theta}_N$ defined in (3.10) guarantee that

$$\begin{aligned} P(\|\widehat{\theta}_N - \theta\|_1 \geq \varepsilon) &\leq P\left(\|\widehat{\theta}_N - \widehat{\theta}\|_1 + \|\widehat{\theta} - \theta\|_1 \geq \varepsilon\right) \\ &\leq P\left(\|\widehat{\theta}_N - \widehat{\theta}\|_1 \geq \varepsilon - \varepsilon_u\right) \\ &\leq H_{\varepsilon-\varepsilon_u,N}. \end{aligned}$$

Since this is valid for all k_0 and $\widetilde{\theta}$, (4.12) follows.

Now, for part (b),

$$\begin{aligned} N_\alpha(\varepsilon) &= \min\{N : \lambda_N(\varepsilon) \leq \alpha\} \\ &\leq \min\{Nn_0 : H_{\varepsilon-\varepsilon_u,N} \leq \alpha\} \\ &\leq n_0 \min\{N : H_{\varepsilon-\varepsilon_u,N} \leq \alpha\}, \end{aligned}$$

which yields (4.13). □

4.3 Lower Bounds on Estimation Errors

To obtain lower bounds on the estimation error when the above full-rank periodic input is used, we use a similar argument as that of the upper bound case.

From $\Phi_0(\widehat{\theta}_N - \widehat{\theta}) = v - \zeta_N$, we have $\|v - \zeta_N\|_1 \leq \|\Phi_0\|_I \|\widehat{\theta}_N - \widehat{\theta}\|_1$. In view of (4.4), the independence of $\xi_N^{\{i\}}$ for $i = 0, \ldots, n_0 - 1$ implies that for any $\varepsilon_1 > 0$,

$$\begin{aligned} &P\left(\left|\widehat{\theta}_N - \widehat{\theta}\right| \geq \varepsilon_1\right) \\ &\quad \geq P\left(\|\zeta_N - v\|_1 \geq \varepsilon_1 \|\Phi_0\|_I\right) \\ &\quad \geq P\left(\bigcap_{i=0}^{n_0-1} \left\{|\zeta_N^{\{i\}} - v^{\{i\}}| \geq \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right\}\right) \\ &\quad \geq \prod_{i=0}^{n_0-1} P\left(|\zeta_N^{\{i\}} - v^{\{i\}}| \geq \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right) \\ &\quad \geq \prod_{i=0}^{n_0-1} P\left(\xi_N^{\{i\}} \geq F\left(v^{\{i\}} + \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right)\right) \\ &\qquad + \prod_{i=0}^{n_0-1} P\left(\xi_N^{\{i\}} \leq F\left(v^{\{i\}} - \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right)\right). \end{aligned} \tag{4.14}$$

Our approach of obtaining the lower bounds involves two steps. First, if the random variables are normally distributed, the lower bounds can

be obtained using an inequality in [30] together with the properties of a normal distribution. The second step deals with the situation in which the noises are not normal, but are approximately normal by the Barry–Esseen estimate.

Assume that $\{d_k\}$ is a sequence of normally distributed random variables with mean 0 and variance σ^2. Suppose that $Z(x)$ is the distribution of the standard normal random variable, i.e.,

$$Z(x) = \int_{-\infty}^{x} z(\zeta) d\zeta,$$

where

$$z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \ -\infty < x < \infty.$$

It was shown in [30, Lemma 2, p. 175] that

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) z(x) < 1 - Z(x) < \frac{1}{x} z(x), \quad \text{for } x > 0. \tag{4.15}$$

Since d_k is normally distributed with mean zero and variance σ^2, $\xi_N^{\{i\}}$ is also normally distributed with mean $F(v^{\{i\}})$ and variance $F(v^{\{i\}})(1 - F(v^{\{i\}}))/N$. Therefore,

$$\sqrt{N}(\xi_N^{\{i\}} - F(v^{\{i\}}))/\sqrt{F(v^{\{i\}})(1 - F(v^{\{i\}}))}$$

is normally distributed with mean 0 and variance 1. As a result, to obtain the desired lower bounds using (4.14), for any $\varepsilon_1 > 0$, it suffices to consider

$$P\left(\xi_N^{\{i\}} \geq F\left(v^{\{i\}} + \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right)\right)$$

and

$$P\left(\xi_N^{\{i\}} \leq F\left(v^{\{i\}} - \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right)\right).$$

Denote

$$\widetilde{\alpha}_i^+ = \widetilde{\alpha}_i^+(\varepsilon_1) = \frac{\sqrt{N}\left(F\left(v^{\{i\}} + \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right) - F(v^{\{i\}})\right)}{\sqrt{F(v^{\{i\}})(1 - F(v^{\{i\}}))}}.$$

Then

$$\begin{aligned} &P\left(\xi_N^{\{i\}} \geq F\left(v^{\{i\}} + \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right)\right) \\ &\quad = P\left(\frac{\sqrt{N}(\xi_N^{\{i\}} - F(v^{\{i\}}))}{\sqrt{F(v^{\{i\}})(1 - F(v^{\{i\}}))}} \geq \widetilde{\alpha}_i^+\right). \end{aligned}$$

Therefore, by (4.15),

$$\begin{aligned} & P\left(\xi_N^{\{i\}} \geq F\left(v^{\{i\}} + \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right)\right) \\ & = 1 - Z(\widetilde{\alpha}_i^+) \\ & \geq \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\widetilde{\alpha}_i^+} - \left(\frac{1}{\widetilde{\alpha}_i^+}\right)^3\right) \exp\left(-\frac{(\widetilde{\alpha}_i^+)^2}{2}\right). \end{aligned} \tag{4.16}$$

Likewise, denote

$$\widetilde{\alpha}_i^- = \widetilde{\alpha}_i^-(\varepsilon_1) = \frac{\sqrt{N}\left(F\left(v^{\{i\}} - \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right) - F(v^{\{i\}})\right)}{\sqrt{F(v^{\{i\}})(1 - F(v^{\{i\}}))}}.$$

Note that $\widetilde{\alpha}_i^-(\varepsilon_1) < 0$. We obtain

$$\begin{aligned} & P\left(\xi_N^{\{i\}} \leq F\left(v^{\{i\}} - \frac{\varepsilon_1 \|\Phi_0\|_I}{n_0}\right)\right) \\ & = 1 - Z(-\widetilde{\alpha}_i^-) \\ & \geq 1 + \frac{1}{\sqrt{2\pi}} \frac{1}{\widetilde{\alpha}_i^-} \exp\left(-\frac{(\widetilde{\alpha}_i^-)^2}{2}\right). \end{aligned} \tag{4.17}$$

Combining (4.16) and (4.17), we obtain the following lower bounds.

Theorem 4.3. *For any $\varepsilon_1 > 0$,*

$$\begin{aligned} & P(||\widehat{\theta}_N - \widehat{\theta}||_1 \geq \varepsilon_1) \\ & \geq \prod_{i=0}^{n_0-1} \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\widetilde{\alpha}_i^+(\varepsilon_1)} - \left(\frac{1}{\widetilde{\alpha}_i^+(\varepsilon_1)}\right)^3\right) e^{-\frac{(\widetilde{\alpha}_i^+(\varepsilon_1))^2}{2}} \\ & \quad + \prod_{i=0}^{n_0-1}\left(1 + \frac{1}{\sqrt{2\pi}} \frac{1}{\widetilde{\alpha}_i^-(\varepsilon_1)} e^{-\frac{(\widetilde{\alpha}_i^-(\varepsilon_1))^2}{2}}\right). \end{aligned}$$

Furthermore, we also obtain the following corollary with $\varepsilon_1 = \varepsilon + \varepsilon_u$.

Corollary 4.4. *Setting $\varepsilon_1 = \varepsilon + \varepsilon_u$ in Theorem 4.3, we have*

$$\begin{aligned} & P(||\widehat{\theta}_k - \theta||_1 \geq \varepsilon + \varepsilon_u) \\ & \geq \prod_{i=0}^{n_0-1} \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\widetilde{\alpha}_i^+(\varepsilon + \varepsilon_u)} - \left(\frac{1}{\widetilde{\alpha}_i^+(\varepsilon + \varepsilon_u)}\right)^3\right) e^{-\frac{(\widetilde{\alpha}_i^+(\varepsilon+\varepsilon_u))^2}{2}} \\ & \quad + \prod_{i=1}^{n_0}\left(1 + \frac{1}{\sqrt{2\pi}} \frac{1}{\widetilde{\alpha}_i^-(\varepsilon + \varepsilon_u)} e^{-\frac{(\widetilde{\alpha}_i^-(\varepsilon+\varepsilon_u))^2}{2}}\right). \end{aligned}$$

Lower Bounds Based on Asymptotic Normality

The idea here is to approximate the underlying distribution by a normal random variable. It is easily seen that

$$\rho_N^{\{i\}} := \frac{\sqrt{N}(\xi_N^{\{i\}} - F(v^{\{i\}}))}{\sqrt{F(v^{\{i\}})(1 - F(v^{\{i\}}))}}$$

converges in distribution to the standard normal random variable. By virtue of the Barry–Esseen estimate [31, Theorem 1, p. 542], the following lemma is in force.

Lemma 4.5. *$|P(\rho_N^{\{i\}} \le z) - P(Z \le z)| \le \Delta_N$, where $\Delta_N = O(1/\sqrt{N})$ as $N \to \infty$ and Z is the standard normal random variable.*

Theorem 4.6. *The following lower bounds hold:*

$$\begin{aligned} &P(||\widehat{\theta}_N - \theta||_1 \ge \varepsilon + \varepsilon_u) \\ &\ge \prod_{i=0}^{n_0-1} \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\widetilde{\alpha}_i^+(\varepsilon + \varepsilon_u)} - \left(\frac{1}{\widetilde{\alpha}_i^+(\varepsilon + \varepsilon_u)} \right)^3 \right) e^{-\frac{(\widetilde{\alpha}_i^+(\varepsilon+\varepsilon_u))^2}{2}} \\ &\quad + \prod_{i=1}^{n_0} \left(1 + \frac{1}{\sqrt{2\pi}} \frac{1}{\widetilde{\alpha}_i^-(\varepsilon + \varepsilon_u)} e^{-\frac{(\widetilde{\alpha}_i^-(\varepsilon+\varepsilon_u))^2)}{2}} \right) + \Delta_N, \end{aligned}$$

where $\Delta_N = O(1/\sqrt{N})$ as $N \to \infty$.

Proof. Note that by Lemma 4.5,

$$\begin{aligned} &P\left(\xi_N^{\{i\}} \ge F\left(v^{\{i\}} + \frac{(\varepsilon + \varepsilon_u)\|\Phi_0\|_I}{n_0} \right) \right) \\ &\quad = P(\rho_N^{\{i\}} \ge \widetilde{\alpha}_i^+(\varepsilon + \varepsilon_u)) \\ &\quad \ge P(Z \ge \widetilde{\alpha}_i^+(\varepsilon + \varepsilon_u)) - \Delta_N. \end{aligned}$$

Similarly,

$$\begin{aligned} &P\left(\xi_N^{\{i\}} \le F\left(v^{\{i\}} - \frac{(\varepsilon + \varepsilon_u)\|\Phi_0\|_I}{n_0} \right) \right) \\ &\quad = P(\rho_N^{\{i\}} \le \widetilde{\alpha}_i^-(\varepsilon + \varepsilon_u)) \\ &\quad \ge P(Z \le \widetilde{\alpha}_i^-(\varepsilon + \varepsilon_u)) - \Delta_N. \end{aligned}$$

Using the estimates of lower bounds as in Theorem 4.3 for the normal random variable Z, the desired result then follows. □

4.4 Notes

This chapter is based on our work [111], which initiated the work on identification with binary-valued observations. The main focus of this chapter is

on the derivation of the upper and lower bounds of estimation errors, which were also obtained in [111]. The framework for combining stochastic and deterministic approaches for identification was introduced in [101, 102]. For some related but different identification algorithms, we refer the reader to [2, 13, 18, 27, 29, 33, 38, 73, 96, 119] and references therein.

5
Rational Systems

The systems in Chapters 3 and 4 are finite impulse-response models. Due to nonlinearity in output observations, switching or nonsmooth nonlinearity enters the regressor for rational models. A common technique for solving this problem is to use the methods of nonlinear filtering. One difficulty of using such nonlinear filtering is that it normally leads to infinite-dimensional filters. To overcome this difficulty, a two-step identification procedure is introduced that employs periodic signals, empirical measures, and identifiability features so that rational models can be identified without resorting to complicated nonlinear search algorithms. Identification errors and input design are examined in a stochastic information framework.

This chapter begins with a description of the problems in Section 5.1. Our development starts in Section 5.2 with estimation of plant outputs. Section 5.3 establishes the identifiability of plant parameters. A basic property of rational systems is established. It shows that if the input is periodic and full rank, system parameters are uniquely determined by its periodic outputs. Consequently, under such inputs, the convergence of parameter estimates can be established when the convergence results of Section 5.2 are utilized.

5.1 Preliminaries

Consider the system

$$y_k = G(q)u_k + d_k = x_k + d_k, \tag{5.1}$$

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_5,

which is in an output error form. Here q is the one-step backward shift operator $qu_k = u_{k-1}$; $\{d_k\}$ is a sequence of random sensor noises; $x_k = G(q)u_k$ is the noise-free output of the system; $G(q)$ is a stable rational function of q:

$$G(q) = \frac{B(q)}{1 - A(q)} = \frac{b_1 q + \cdots + b_{n_0} q^{n_0}}{1 - (a_1 q + \cdots + a_{n_0} q^{n_0})}.$$

The observation $\{y_k\}$ is measured by a binary-valued sensor of threshold $C > 0$, and the parameters

$$\theta = [a_1, \ldots, a_{n_0}, b_1, \ldots, b_{n_0}]'$$

are to be identified.

For system identification, the system (5.1) is commonly expressed in its regression form

$$y_k = A(q)y_k + B(q)u_k + (1 - A(q))d_k = \psi_k'\theta + \widetilde{d}_k, \tag{5.2}$$

where

$$\psi_k = [y_{k-1}, \ldots, y_{k-n_0}, u_{k-1}, \ldots, u_{k-n_0}]',$$

and $\widetilde{d}_k = (1 - A(q))d_k$. The sequence $\{\widetilde{d}_k\}$ may not be independent even if $\{d_k\}$ is.

Most identification algorithms, especially recursive ones, have been developed from the observation structure (5.2). Direct application of this structure in our problem encounters a daunting difficulty since y_k is not directly measured. Using s_k in this structure introduces nonlinearities that make it more difficult to design feasible algorithms or to establish their fundamental properties such as convergence and accuracy. In addition, use of the indicator function makes the problem nonsmooth.

In this chapter, we consider a two-step approach:

(i) First, x_k in (5.1) is estimated on the basis of s_k;

(ii) θ is identified from the input u_k and the estimated x_k, using the structure (5.2).

The first step is accomplished by using periodic inputs and empirical measures. The second step is validated by using identifiability arguments and computed by recursive algorithms. Convergence of the algorithms will be derived.

5.2 Estimation of x_k

To estimate x_k, select u_k to be $2n_0$-periodic. Then the noise-free output $x_k = G(q)u_k$ is also $2n_0$-periodic, after a short transient duration. Since

the system is assumed to be stable, all transient modes decay exponentially, much faster than the convergence rates of empirical measures. As a consequence, their impact is negligible and will be ignored in the analysis. Hence, $x_{j+2ln_0} = x_j$, for any positive integer l. The $2n_0$-periodic sequence $\{x_k\}$ is determined entirely by its first $2n_0$ unknown real numbers $\gamma^{\{j\}}, j = 1, \ldots, 2n_0$;

$$x_j = \gamma^{\{j\}}, \quad j = 1, \ldots, 2n_0; \tag{5.3}$$

and $\Gamma = [\gamma^{\{1\}}, \ldots, \gamma^{\{2n_0\}}]'$ are to be estimated. For each $j = 1, \ldots, 2n_0$, the observations can be expressed as

$$y_{j+2ln_0} = x_{j+2ln_0} + d_{j+2ln_0} = \gamma^{\{j\}} + d_{j+2ln_0}, \quad l = 0, 1, \ldots \tag{5.4}$$

Note that (5.4) indicates that for a fixed j, $\gamma^{\{j\}}$ is an unknown constant, and empirical measures can be calculated with respect to the index l. Let the observation length be $2n_0N$ for some positive integer N. For a given $j = 1, \ldots, 2n_0$, define

$$\xi_N^{\{j\}} = \frac{1}{N} \sum_{l=0}^{N-1} s_{j+2ln_0}. \tag{5.5}$$

The event $\{s_{j+2ln_0} = 1\} = \{y_{j+2ln_0} \le C\}$ is the same as that of $\{d_{j+2ln_0} \le C - \gamma^{\{j\}}\}$. Then $\xi_N^{\{j\}}$ is precisely the value of the N-sample empirical distribution $F_N(z)$ of the noise d at $z = C - \gamma^{\{j\}}$. The well-known Glivenko–Cantelli theorem presented in Chapter 3 guarantees the convergence of $\xi_N^{\{j\}}$; see Theorem 3.2 and Theorem 3.10.

We note that the pointwise convergence of $F_N(z)$ at $z = C - \gamma^{\{j\}}$, $j = 1, \ldots, 2n_0$, will suffice for our purpose. To proceed, we first construct an estimate of $\gamma^{\{j\}}$, which will then be used to identify the system parameter θ. Since $F(\cdot)$ is invertible and known, we define

$$\widehat{\gamma}_N^{\{j\}} = C - F^{-1}(\xi_N^{\{j\}}). \tag{5.6}$$

Theorem 5.1. *Under Assumption* (A3.1),

$$\widehat{\gamma}_N^{\{j\}} \to \gamma^{\{j\}} \quad \textit{w.p.1 as} \quad N \to \infty.$$

Proof. By Theorem 3.2, as $N \to \infty$,

$$\xi_N^{\{j\}} \to F(C - \gamma^{\{j\}}) \ \text{ w.p.1.}$$

Hence, the continuity of $F^{-1}(\cdot)$ implies that $F^{-1}(\widehat{F}_N(C - \gamma^{\{j\}})) \to C - \gamma^{\{j\}}$ w.p.1. Therefore,

$$F^{-1}(\xi_N^{\{j\}}) \to C - \gamma^{\{j\}} \ \text{ w.p.1,}$$

or, equivalently, $C - F^{-1}(\xi_N^{\{j\}}) \to \gamma^{\{j\}}$ w.p.1. □

Example 5.2. Consider the case $\gamma^{\{j\}} = 2.1$: $y_k = 2.1 + d_k$ and the sensor threshold $C = 3.5$. The disturbance is uniformly distributed in $[-2, 2]$. Figure 5.1 shows estimates of $\gamma^{\{j\}}$ as a function of sample sizes.

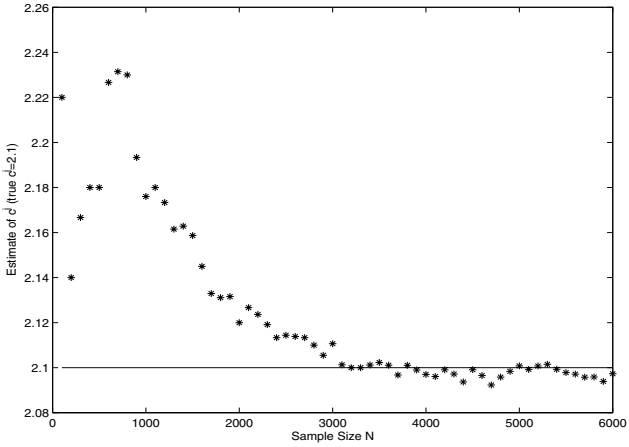


FIGURE 5.1. Convergence of Estimates of $\gamma^{\{j\}}$

5.3 Estimation of Parameter θ

Under a sequence of periodic inputs $\{u_k\}$, the one-to-one mapping between θ and the periodic output x_k of the system G will first be established. This relationship will be used to derive an estimate of θ from that of x_k.

5.3.1 Parameter Identifiability

Recall that the noise-free system output is

$$x_k = G(q)u_k = \frac{b_1 q + \cdots + b_{n_0} q^{n_0}}{1 - (a_1 q + \cdots + a_{n_0} q^{n_0})} u_k$$

or in a regression form

$$x_k = \phi_k' \theta, \tag{5.7}$$

where

$$\phi_k = [x_{k-1}, \ldots, x_{k-n_0}, u_{k-1}, \ldots, u_{k-n_0}]',$$
$$\theta = [a_1, \ldots, a_{n_0}, b_1, \ldots, b_{n_0}]'.$$

Then under a $2n_0$-periodic input, the noise-free output x and system parameters θ are related by $X = \Phi\theta$ with

$$\begin{aligned} X &= [x_{k_0}, \ldots, x_{k_0+2n_0-1}]', \\ \Phi &= [\phi_{k_0}, \ldots, \phi_{k_0+2n_0-1}]'. \end{aligned} \tag{5.8}$$

Apparently, if Φ is full rank, then there is a one-to-one correspondence between X and θ.

Since Φ contains both input u_k and output x_k, in general, the invertibility of Φ depends on both u_k and x_k, hence on the true (but unknown) plant $G(q)$. Furthermore, the invertibility may also vary with the starting time k_0. However, it will be shown that such complications dissipate when u_k is $2n_0$-periodic.

Theorem 5.3. *Suppose that the pair $D(q) = 1 - A(q)$ and $B(q)$ are coprime. If u_k is $2n_0$-periodic and full rank, then*

(a) *Φ given by* (5.8) *is invertible for all k_0;*

(b) *$\|\Phi^{-1}\|$ is independent of k_0, where $\|\cdot\|$ is the largest singular value. Hence, $\mu = \|\Phi^{-1}\| < \infty$ is a constant for all k_0.*

Proof. (a) The proof will follow from certain arguments of identifiability. The true plant $G(q)$ is of order n_0 with transfer function

$$G(q) = \frac{b_1 q + \cdots + b_{n_0} q^{n_0}}{1 - a_1 q - \cdots - a_{n_0} q^{n_0}} = \frac{B(q)}{D(q)},$$

where $D(q)$ and $B(q)$ are coprime polynomials. The observation equation is $X = \Phi\theta$. Note that Φ is invertible if and only if θ can be uniquely determined from the observation equation. Assume that there exists another n_0th-order system $\widetilde{G}(q) = \widetilde{B}(q)/\widetilde{D}(q)$, with $\widetilde{D}$ and $\widetilde{B}$ coprime and $\widetilde{D}(0) = 1$, also satisfying the observation. In particular, $\widetilde{x}_k = (\widetilde{G}u)_k = x_k$, for $k = 1, \ldots, 2n_0$. Define

$$\Delta(q) = G(q) - \widetilde{G}(q) = \frac{B(q)\widetilde{D}(q) - \widetilde{B}(q)D(q)}{D(q)\widetilde{D}(q)} := \frac{qN(q)}{R(q)},$$

where $R(q)$ is a polynomial of order $2n_0$ and $N(q)$ a polynomial of order $2n_0 - 1$.

For the given $2n_0$-periodic input u, by hypothesis we have $h_k = (\Delta u)_k = 0$, $k = 1, \ldots, 2n_0$. It follows that

$$\widetilde{H}(\omega) = \frac{1}{\sqrt{2n_0}} \sum_{k=1}^{2n_0} h_k e^{-i\omega k} = 0.$$

On the other hand, by frequency-domain analysis,

$$\widetilde{H}(\omega) = \Delta(e^{i\omega})U(\omega) + Q(\omega),$$

where

$$U(\omega) = \frac{1}{\sqrt{2n_0}} \sum_{k=1}^{2n_0} u_k e^{-i\omega k}$$

and $Q(\omega) = 0$, for $\omega = 2\pi j/(2n_0)$, $j = 1, \ldots, 2n_0$. By the hypothesis, $U(\omega) \neq 0$, for $\omega = 2\pi j/(2n_0)$, $j = 1, \ldots, 2n_0$. Hence,

$$\Delta(e^{i\omega}) = 0, \text{ for } \omega = 2\pi j/(2n_0), \ j = 1, \ldots, 2n_0.$$

However, since $N(q)$ is of order $2n_0 - 1$, if $\Delta \not\equiv 0$, $\Delta(e^{i\omega})$ can have a maximum of $2n_0 - 1$ finite zeros. Consequently, $\Delta(q) \equiv 0$, i.e., $G(q) \equiv \widetilde{G}(q)$. Now, this equality, together with the coprimeness of $G(q)$ and $\widetilde{G}(q)$, implies that there exists a constant c for which $B(q) = c\widetilde{B}(q)$ and $D(q) = c\widetilde{D}(q)$. Finally, $D(0) = \widetilde{D}(0) = 1$ implies $c = 1$. Therefore, $B(q) = \widetilde{B}(q)$, $D(q) = \widetilde{D}(q)$. Namely, $B(q)$ and $D(q)$, or equivalently θ, are uniquely determined by the observation equation.

(b) For Φ given in (5.8), to emphasize the dependence of Φ on k_0, we write it as $\Phi(k_0)$. To show that $\|\Phi^{-1}(k_0)\|$ is independent of k_0, we observe that since both u_k and x_k are $2n_0$-periodic, $\Phi(k_0 + 1) = J\Phi(k_0)$, where

$$J = \begin{bmatrix} 0 & I_{(2n_0-1)\times(2n_0-1)} \\ 1 & 0 \end{bmatrix}$$

is a $(2n_0) \times (2n_0)$ unitary matrix obtained by permuting the rows of the identity matrix. As a result,

$$\|\Phi^{-1}(k_0)\| = \|J\Phi^{-1}(k_0 + 1)\| = \|\Phi^{-1}(k_0 + 1)\|$$

since the norm $\|\cdot\|$ is unitary-invariant. □

Example 5.4. Suppose that the true system has the transfer function $G(p) = (q + 0.5q^2)/(1 - 0.5q + 0.2q^2)$. Hence, the true plant has the regression model

$$x_k = 0.5x_{k-1} - 0.2x_{k-2} + u_{k-1} + 0.5u_{k-2}.$$

Since the order of the system is $n_0 = 2$, we select the input to be 4-periodic with $u_1 = 1, u_2 = -0.2, u_3 = 1.5, u_4 = -0.1$. For a selected $k_0 = 20$,

$$\Phi = \begin{bmatrix} -1.4884 & -0.6889 & -0.1000 & 1.5000 \\ -1.2564 & -1.4884 & 1.0000 & -0.1000 \\ -1.2805 & -1.2564 & -0.2000 & 1.0000 \\ -0.6890 & -1.2805 & 1.5000 & -0.2000 \end{bmatrix},$$

$$\Phi^{-1} = \begin{bmatrix} -0.8079 & -1.9384 & 1.3004 & 1.4118 \\ 1.0345 & 0.7749 & -1.6066 & -0.6619 \\ 0.5624 & -0.4417 & -0.7069 & 0.9044 \\ 0.3776 & -1.5969 & 0.5053 & 1.1572 \end{bmatrix},$$

and $\|\Phi^{-1}\| = 3.8708$. It can be verified that for different k_0, Φ will be different only by permutation of its rows. Consequently, $\|\Phi^{-1}\| = 3.8708$ for all k_0.

5.3.2 Identification Algorithms and Convergence Analysis

For each $j = 1, \ldots, 2n_0$, the estimate $\widehat{\gamma}_N^{\{j\}}$ of $\gamma^{\{j\}}$ can be written as

$$\widehat{\gamma}_N^{\{j\}} = \gamma^{\{j\}} + e_N^{\{j\}},$$

where, by Theorem 5.1, $e_N^{\{j\}} \to 0$ w.p.1 as $N \to \infty$.

Define an estimated $2n_0$-periodic output sequence of $G(q)$ by periodic extension

$$\widehat{x}_{j+2ln_0} = \widehat{x}_j = \widehat{\gamma}_N^{\{j\}}, \tag{5.9}$$

for $j = 1, \ldots, 2n_0$ and $l = 1, \ldots, N-1$. Then

$$\widehat{x}_{j+2ln_0} = x_{j+2ln_0} + e_N^{\{j\}}, \quad j = 1, \ldots, 2n_0.$$

To estimate the parameter θ, we use $\widehat{x}_k$ in place of x_k in (5.7),

$$\widehat{x}_k = \widehat{\phi}_k' \widehat{\theta}_N,$$

where $\widehat{\phi}_k = [\widehat{x}_{k-1}, \ldots, \widehat{x}_{k-n_0}, u_{k-1}, \ldots, u_{k-n_0}]'$. Then

$$\widehat{X}_N = \widehat{\Phi}_N \widehat{\theta}_N \tag{5.10}$$

for the estimated system, where

$$\widehat{X}_N = [\widehat{x}_{k_0}, \ldots, \widehat{x}_{k_0+2n_0-1}]',$$
$$\widehat{\Phi}_N = [\widehat{\phi}_{k_0}, \ldots, \widehat{\phi}_{k_0+2n_0-1}]'.$$

Since $\widehat{\Phi}_N' \widehat{\Phi}_N$ is invertible w.p.1, the estimate $\widehat{\theta}_N$ is calculated from

$$\widehat{\theta}_N = (\widehat{\Phi}_N' \widehat{\Phi}_N)^{-1} \widehat{\Phi}_N' \widehat{X}_N \quad \text{w.p.1.} \tag{5.11}$$

Since $\widehat{\Phi}_N$ is a square matrix, one may also write $\widehat{\theta}_N = \widehat{\Phi}_N^{-1} \widehat{X}_N$, but (5.11) is the standard least-squares expression. We proceed to establish the convergence of $\widehat{\theta}_N$ to θ.

Theorem 5.5. *Suppose that $D(q)$ and $B(q)$ are coprime. If $\{u_k\}$ is $2n_0$-periodic and full rank, then*

$$\widehat{\theta}_N \to \theta \quad \textit{w.p.1 as} \quad N \to \infty.$$

Proof. From $\widehat{x}_{j+2ln_0} = x_{j+2ln_0} + e_N^{\{j\}}$, (5.10) can be expressed as

$$X_N + E_N = (\Phi_N + \varsigma(E_N))\widehat{\theta}_N, \tag{5.12}$$

where both E_N and $\varsigma(E_N)$ are perturbation terms, $E_N \to 0$ w.p.1 as $N \to \infty$, and $\varsigma(\cdot)$ is a continuous function of its argument satisfying $\varsigma(E_N) \to 0$ as $E_N \to 0$.

Since Φ_N has a bounded inverse and $\varsigma(E_N) \to 0$, w.p.1, $\Phi_N + \varsigma(E_N)$ is invertible w.p.1 for sufficiently large N. It follows that for sufficiently large N, by (5.12),

$$\Phi'_N X_N + \Phi'_N E_N = (\Phi'_N \Phi_N + \Phi'_N \varsigma(E_N))\widehat{\theta}_N.$$

This implies that

$$\begin{aligned} \widehat{\theta}_N &= (\Phi'_N \Phi_N + \Phi'_N \varsigma(E_N))^{-1}(\Phi'_N X_N + \Phi'_N E_N) \\ &\to (\Phi'\Phi)^{-1}\Phi' X = \theta \end{aligned}$$

w.p.1 as $N \to \infty$. □

5.4 Notes

This chapter is a continuation of the early chapters by generalizing the simpler FIR models to rational systems. It is based on our work [108]. The idea of persistent identification, which considers worst-case identification errors over all possible starting times, can be found in [97]. When sensors are nonlinear and nonsmooth such as the switching sensors investigated in this book, system identification for plants in ARMA structures usually becomes difficult, due to the lack of constructive and convergent identification algorithms. Our two-step approach employs the periodic signals to avoid this difficulty.

6

Quantized Identification and Asymptotic Efficiency

Up to this point, we have been treating binary-valued observations. The fundamental principles and basic algorithms for binary-valued observations can be modified to handle quantized observations as well. One way to understand the connection is to view a quantized observation as a vector-valued binary observation in which each vector component represents the output of one threshold, which is a binary-valued sensor. The dimension of the vector is the number of the thresholds in the quantized sensor.

However, since a binary-valued sensor is already sufficient for achieving strong convergence and mean-square convergence for the parameter estimates, and weak convergence of the centered and scaled estimation errors, it is natural to ask: Why do we need quantized observations? What benefits can be gained from using more complicated sensors? To answer these questions, we study the efficiency issue that is characterized by the Cramér–Rao (CR) lower bounds. We are seeking algorithms that utilize all statistical information in the data about the unknown parameters, in the sense that they achieve asymptotically the best convergence rates given by the CR lower bound. Consequently, the CR lower bounds become a characterization of the system complexity in this problem. Comparisons of such complexities among sensors of different thresholds permit us to answer the above questions rigorously and completely.

Section 6.1 begins with basic identification algorithms and their convergence properties. To utilize information from all thresholds collaboratively, Section 6.2 introduces an algorithm named by the authors as the quasi-convex combination estimator (QCCE). Expressions for identification errors are derived. Section 6.3 develops some important expressions of

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_6,

identification error covariances that are essential for deriving the efficiency of the optimal QCCE algorithms. The main results of this chapter are contained in Section 6.4, in which the asymptotic efficiency of the optimal QCCE is established.

6.1 Basic Algorithms and Convergence

Consider a single-input–single-output, linear, time-invariant, stable, discrete-time system G given by

$$y_k = Gu_k + d_k,\ k = 1, 2, \ldots, \tag{6.1}$$

where u_k is the input, d_k is the disturbance, and G is either a rational transfer function or an FIR (finite impulse response) system, or in the simplest case, a gain. The output y_k is measured by a sensor of m_0 thresholds $-\infty < C_1 < \cdots < C_{m_0} < \infty$. The sensor is represented by a set of m_0 indicator functions $s_k = [s_k^{\{1\}}, \ldots, s_k^{\{m_0\}}]'$, where $s_k^{\{i\}} = I_{\{-\infty < y_k \le C_i\}}$, $i = 1, \ldots, m_0$.

First, consider the simplest case of identifying a constant θ:

$$y_k = \theta + d_k.$$

Under Assumption (A3.1), $\{d_k\}$ is a sequence of i.i.d. random variables with distribution function $F(\cdot)$. Thus, for each threshold C_i, $\{s_k^{\{i\}}\}$ is also an i.i.d. sequence. Then

$$p_i = E(s_k^{\{i\}}) = F(C_i - \theta) := F_i(\theta).$$

Since $F_i(\theta)$ is invertible, we denote its inverse by $G_i(\cdot)$, and hence $G_i(p_i) = \theta$. Define

$$\xi_N^{\{i\}} = \frac{1}{N}\sum_{k=0}^{N-1} s_k^{\{i\}}; \quad \theta_N^{\{i\}} = G_i(\xi_N^{\{i\}}).$$

By virtue of the results in Chapter 3, $\theta_N^{\{i\}}$ is asymptotically unbiased. Let $\theta_N^{\{i\}}$, $i = 1, \ldots, m_0$, be m_0 asymptotically unbiased estimators of θ based on samples of size N. Denote

$$\begin{aligned}
&\Theta_N = [\theta_N^{\{1\}}, \ldots, \theta_N^{\{m_0\}}]',\\
&e_N^{\{i\}} = \theta_N^{\{i\}} - \theta,\\
&e_N = [e_N^{\{1\}}, \ldots, e_N^{\{m_0\}}]',\\
&\mathbb{1} = [1, 1, \ldots, 1]' \in \mathbb{R}^{m_0}.
\end{aligned}$$

Then $e_N = \Theta_N - \theta\mathbb{1}$. Define

$$V_N(\theta) = Ee_N e_N'. \tag{6.2}$$

Note that $Ee_N \to 0$ as $N \to \infty$, and that $V_N(\theta)$ is a covariance matrix of e_N that is positive semidefinite. Although the calculation of $V_N(\theta)$ may be cumbersome, the following asymptotic result shows that $V_N(\theta)$ can be approximated by a computable function.

From $p_i = F_i(\theta)$, define

$$h_i(\theta) = \partial F_i(\theta)/\partial\theta.$$

Then

$$\partial G_i(p_i)/\partial p_i = 1/h_i(\theta).$$

Denote

$$\begin{aligned} &p = [p_1, \ldots, p_{m_0}]', \\ &h(\theta) = [h_1(\theta), \ldots, h_{m_0}(\theta)]', \\ &G(p) = [G_1(p_1), \ldots, G_{m_0}(p_{m_0})]', \end{aligned} \tag{6.3}$$

and

$$\begin{aligned} M &= \begin{bmatrix} p_1 & p_1 & \cdots & p_1 \\ p_1 & p_2 & \cdots & p_2 \\ \vdots & & & \vdots \\ p_1 & p_2 & \cdots & p_{m_0} \end{bmatrix}, \\ U &= \operatorname{diag}\left(\frac{1}{h_1(\theta)}, \ldots, \frac{1}{h_{m_0}(\theta)}\right). \end{aligned} \tag{6.4}$$

Theorem 6.1. *As* $N \to \infty$,

$$NV_N(\theta) \to U(M - pp')U := \Psi(\theta). \tag{6.5}$$

Proof. As mentioned in Remark 3.11, the weak convergence of $B_N(\cdot)$ and the Skorohod representation (without changing notations) enable us to assume that $B_N(\cdot) \to B^0(\cdot)$ w.p.1. Let $\xi_N = [\xi_N^{\{1\}}, \ldots, \xi_N^{\{m_0\}}]'$. Consider

$$e_N = \Theta_N - \theta\mathbb{1} = G(\xi_N) - G(p).$$

By Theorem 3.12,

$$\Upsilon_N^{\{i\}} = \sqrt{N}e_N^{\{i\}} = \sqrt{n}(G_i(\xi_N^{\{i\}}) - G_i(p_i))$$

converges to

$$\Upsilon_N^{\{i\}} \to \Upsilon^{\{i\}} = \frac{\partial G_i(p_i)}{\partial p_i} B^0(v^{\{i\}}) = \frac{B^0(v^{\{i\}})}{h_i(\theta)} \quad \text{w.p.1 as} \quad N \to \infty. \tag{6.6}$$

As a result, $\Upsilon = U(B^0)'$ with

$$\begin{aligned}\Upsilon &= [\Upsilon^{\{1\}}, \ldots, \Upsilon^{\{m_0\}}]',\\ B^0 &= [B^0(v^{\{1\}}), \ldots, B^0(v^{\{m_0\}}]',\\ U &= [\frac{\partial G_1(p_1)}{\partial p_1}, \ldots, \frac{\partial G_{m_0}(p_{m_0})}{\partial p_{m_0}}]'.\end{aligned}$$

It follows from Theorem 3.12 that

$$E\Upsilon_N\Upsilon_N' \to EUB^0(B^0)'U$$

and

$$\begin{aligned}EB^0(B^0)' &= \begin{bmatrix} p_1 - p_1^2 & p_1 - p_1p_2 & \cdots & p_1 - p_1p_{m_0} \\ p_1 - p_1p_2 & p_2 - p_2^2 & \cdots & p_2 - p_2p_{m_0} \\ \vdots & & & \vdots \\ p_1 - p_{m_0}p_1 & p_2 - p_{m_0}p_2 & \cdots & p_{m_0} - p_{m_0}^2 \end{bmatrix}\\ &= M - pp^T.\end{aligned}$$

Therefore,

$$E\Upsilon_N\Upsilon_N' \to U(M - pp')U.$$

The proof of the theorem is thus completed. □

The covariance in Theorem 6.1 reflects identification errors from each threshold and their correlations. However, it does not consider the unique feature here that all estimates are for the same parameter θ. Combining these estimates will eventually lead to an efficient algorithm.

6.2 Quasi-Convex Combination Estimators (QCCE)

Define $\beta = [\beta_1, \ldots, \beta_{m_0}]'$ such that $\beta_1 + \cdots + \beta_{m_0} = 1$. One can construct an estimator $\widehat{\theta}_N$ of θ by

$$\widehat{\theta}_N = \sum_{i=1}^{m_0} \beta_i \theta_N^{\{i\}} = \beta'\Theta_N. \tag{6.7}$$

$\widehat{\theta}_N$ is called a *quasi-convex combination estimator* (*QCCE*). The term "quasi-convex" is used since β_i need not be nonnegative. Since $\theta_N^{\{i\}}$ is asymptotically unbiased,

$$E\widehat{\theta}_N = \beta' E\Theta_N \to \beta'\theta\mathbb{1} = \theta \quad \text{as} \quad N \to \infty.$$

Hence, $\widehat{\theta}_N$ is an asymptotically unbiased estimate of θ. Moreover, the variance of the estimation error $\widehat{\theta}_N - \theta$ is given by

$$\begin{aligned}\overline{\sigma}_N^2 : \ &= E(\beta'\Theta_N - \theta)^2 = E(\beta'\Theta_N - \beta'\theta\mathbb{1})^2 \\ &= \beta' E e_N e_N' \beta = \beta' V_N(\theta)\beta.\end{aligned}$$

That is, the variance is in a quadratic form with respect to the vector β.

The estimator that minimizes $\overline{\sigma}_N^2$ is called *the optimal quasi-convex combination estimator* (*optimal QCCE*), which is obtained from

$$\sigma_N^2 = \min_{\beta,\beta'\mathbb{1}=1} \overline{\sigma}_N^2 = \min_{\beta,\beta'\mathbb{1}=1} \beta' V_N(\theta)\beta. \tag{6.8}$$

Theorem 6.2. *Under Assumption* (A3.1) *and assuming* $V_N(\theta)$ *is positive definite, the optimal QCCE can be obtained by choosing*

$$\beta^* = \frac{V_N^{-1}(\theta)\mathbb{1}}{\mathbb{1}' V_N^{-1}(\theta)\mathbb{1}}, \quad \widehat{\theta}_N = (\beta^*)'\Theta_N, \tag{6.9}$$

and the minimal variance is

$$\sigma_N^2 = \frac{1}{\mathbb{1}' V_N^{-1}(\theta)\mathbb{1}}. \tag{6.10}$$

Proof. The estimator that solves (6.8) is in fact the Gauss–Markov estimator [64, p. 84] (linear minimum variance unbiased estimator) and (6.9) follows directly. For an elementary derivation, one defines the Hamiltonian

$$H(\beta, \lambda) = \beta' V_N(\theta)\beta + \lambda(1 - \beta'\mathbb{1}),$$

where λ is a Lagrange multiplier. Using standard techniques in optimization (see [64, Chapter 10]) yields the stationary point (λ^*, β^*) of $H(\beta, \lambda)$ with $\lambda^* = 2/(\mathbb{1}' V_N^{-1}(\theta)\mathbb{1})$ and β^* given in (6.9). It can be verified that the stationary point is indeed a minimum. Substituting the above solutions into $\overline{\sigma}_N^2$, we obtain the optimal variance as in (6.10). □

Remark 6.3. The optimal QCCE naturally gives more weights on the thresholds that provide more accurate information. To gain insights, suppose hypothetically that the estimators $\theta_N^{\{i\}}$, $i = 1, \ldots, m_0$, are independent. Then V_N is diagonal: $V_N = \operatorname{diag}(v_N^1, \ldots, v_N^{m_0})$. It follows that the optimal weighting is

$$\beta_N = \frac{[(v_N^1)^{-1}, \ldots, (v_N^{m_0})^{-1}]'}{(v_N^1)^{-1} + \cdots + (v_N^{m_0})^{-1}}. \tag{6.11}$$

In other words, the estimators with smaller variances will be more heavily weighted.

6.3 Alternative Covariance Expressions of Optimal QCCEs

Recall from Section 2.2 that $\widetilde{s}_k^{\{i\}} = I_{\{C_{i-1}<y_k\le C_i\}}$, $i = 1, \ldots, m_0+1$, with $C_0 = -\infty$ and $\widetilde{s}_k(m_0+1) = I_{\{C_{m_0}<y_k<\infty\}}$. Let

$$\begin{aligned}\widetilde{p}_i &= P\{\widetilde{s}_k^{\{i\}} = 1\} = P\{C_{i-1} < y_k \le C_i\} \\ &= F(C_i - \theta) - F(C_{i-1} - \theta) := \widetilde{F}_i(\theta).\end{aligned}$$

Define

$$\widetilde{h}_i(\theta) = \frac{\partial \widetilde{F}_i(\theta)}{\partial \theta} = -f(C_i - \theta) + f(C_{i-1} - \theta).$$

The following relationships between $\widetilde{p}_i, \widetilde{h}_i$ and p_i, h_i in (6.3) can be easily established:

$$\begin{aligned}&\sum_{i=1}^{m_0+1} \widetilde{p}_i = 1, \quad \sum_{i=1}^{m_0+1} \widetilde{h}_i = 0, \\ &p_j = \sum_{i=1}^{j} \widetilde{p}_i, \quad h_j = \sum_{i=1}^{j} \widetilde{h}_i, j = 1, \ldots, m_0, \\ &\widetilde{p}_1 = p_1, \quad \widetilde{p}_i = p_i - p_{i-1}, i = 2, \ldots, m_0, \\ &\widetilde{p}_{m_0+1} = 1 - p_{m_0}.\end{aligned}$$

Denote

$$\begin{aligned}&\widetilde{p}(\theta) = [\widetilde{p}_1, \ldots, \widetilde{p}_{m_0}]', \\ &\widetilde{U}(\theta) = \mathrm{diag}(1/\widetilde{h}_1, \ldots, 1/\widetilde{h}_{m_0}), \\ &\widetilde{M}(\theta) = \mathrm{diag}(\widetilde{p}_1, \ldots, \widetilde{p}_{m_0}),\end{aligned}$$

and

$$\widetilde{\Psi}(\theta) = \widetilde{U}(\widetilde{M} - \widetilde{p}\widetilde{p}')\widetilde{U}. \tag{6.12}$$

For notational simplicity, we suppress the θ dependence in the expressions. Assume that $\widetilde{p}_i \neq 0$, $i = 1, \ldots, m_0+1$. If $\widetilde{p}_j = 0$, the threshold C_j can be eliminated since the interval $(C_{i-1}, C_j]$ contains no useful information on θ. Then, $\widetilde{M}$ is invertible. Let

$$\widetilde{W} = \widetilde{M}^{1/2} = \mathrm{diag}(\sqrt{\widetilde{p}_1}, \ldots, \sqrt{\widetilde{p}_{m_0}}).$$

We now prove an important lemma that will be essential for establishing the efficiency of the optimal QCCE algorithms in the next section.

Lemma 6.4. $\mathbb{1}'\widetilde{\Psi}^{-1}\mathbb{1} = \sum_{i=1}^{m_0+1} \widetilde{h}_i^2/\widetilde{p}_i.$

Proof. Note that

$$\begin{aligned}\widetilde{\Psi}^{-1} &= (\widetilde{U}(\widetilde{W}^2 - \widetilde{p}\widetilde{p}')\widetilde{U})^{-1} \\ &= (\widetilde{W}\widetilde{U})^{-1}(I - \widetilde{W}^{-1}\widetilde{p}\widetilde{p}'\widetilde{W}^{-1})^{-1}(\widetilde{U}\widetilde{W})^{-1}.\end{aligned}$$

By the well-known matrix inversion lemma ([62, p. 306, eq. (11.10)]),

$$(I - \widetilde{W}^{-1}\widetilde{p}\widetilde{p}'\widetilde{W}^{-1})^{-1} = I + \frac{\widetilde{W}^{-1}\widetilde{p}\widetilde{p}'\widetilde{W}^{-1}}{1 - \widetilde{p}'\widetilde{M}^{-1}\widetilde{p}}.$$

Observe that

$$\begin{aligned}1 - \widetilde{p}'\widetilde{M}^{-1}\widetilde{p} &= 1 - \sum_{i=1}^{m_0}\widetilde{p}_i = \widetilde{p}_{m_0+1}, \\ \widetilde{p}'\widetilde{W}^{-1}(\widetilde{U}\widetilde{W})^{-1}\mathbb{1} &= [\sqrt{\widetilde{p}_1}, \ldots, \sqrt{\widetilde{p}_{m_0}}]\Big[\frac{\widetilde{h}_1}{\sqrt{\widetilde{p}_1}}, \ldots, \frac{\widetilde{h}_{m_0}}{\sqrt{\widetilde{p}_{m_0}}}\Big]' = \sum_{i=1}^{m_0}\widetilde{h}_i,\end{aligned}$$

and

$$\mathbb{1}'(\widetilde{W}\widetilde{U})^{-1}(\widetilde{U}\widetilde{W})^{-1}\mathbb{1} = \sum_{i=1}^{m_0}\frac{\widetilde{h}_i^2}{\widetilde{p}_i}.$$

Consequently,

$$\begin{aligned}\mathbb{1}'\widetilde{\Psi}^{-1}\mathbb{1} &= \mathbb{1}'(\widetilde{U}(\widetilde{W}^2 - \widetilde{p}\widetilde{p}')\widetilde{U})^{-1}\mathbb{1} \\ &= \mathbb{1}'(\widetilde{W}\widetilde{U})^{-1}(I - \widetilde{W}^{-1}\widetilde{p}\widetilde{p}'\widetilde{W}^{-1})^{-1}(\widetilde{U}\widetilde{W})^{-1}\mathbb{1} \\ &= \sum_{i=1}^{m_0}\frac{\widetilde{h}_i^2}{\widetilde{p}_i} + \frac{(\sum_{i=1}^{m_0}\widetilde{h}_i)^2}{\widetilde{p}_{m_0+1}}.\end{aligned}$$

However, from $\sum_{i=1}^{m_0+1}\widetilde{p}_i = 1$, we have

$$\sum_{i=1}^{m_0+1}\widetilde{h}_i = \sum_{i=1}^{m_0+1}\frac{\partial\widetilde{p}_i}{\partial\theta} = 0,$$

or $\widetilde{h}_{m_0+1} = -\sum_{i=1}^{m_0}\widetilde{h}_i$. It follows that

$$\frac{(\sum_{i=1}^{m_0}\widetilde{h}_i)^2}{\widetilde{p}_{m_0+1}} = \frac{\widetilde{h}_{m_0+1}^2}{\widetilde{p}_{m_0+1}}.$$

Therefore,

$$\mathbb{1}'\widetilde{\Psi}^{-1}\mathbb{1} = \sum_{i=1}^{m_0+1}\frac{\widetilde{h}_i^2}{\widetilde{p}_i}, \tag{6.13}$$

which completes the proof. □

We have introduced $\widetilde{\Psi}$ to provide a convenient way of expressing the CR lower bound in the next section. Next, we establish the connection between $\widetilde{\Psi}$ in (6.12) and Ψ in (6.5). Denote $\widetilde{h} = [\widetilde{h}_1, \ldots, \widetilde{h}_{m_0}]'$, and refer to (6.3)–(6.5) for related expressions.

Lemma 6.5. $\mathbb{1}'\Psi^{-1}\mathbb{1} = \mathbb{1}'\widetilde{\Psi}^{-1}\mathbb{1}$.

Proof. First, note that

$$\mathbb{1}'\Psi^{-1}\mathbb{1} = h'(M - pp')^{-1}h,$$
$$\mathbb{1}'\widetilde{\Psi}^{-1}\mathbb{1} = \widetilde{h}'(\widetilde{M} - \widetilde{p}\widetilde{p}')^{-1}\widetilde{h}.$$

Let V_1 be the matrix

$$V_1 = \begin{bmatrix} 1 & -1 & \dots & -1 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix},$$

which is formed by subtracting columns $i = 2, \dots, m_0$ from the first column of the identity matrix. Then,

$$h'(M - pp')^{-1}h = h'V_1(V_1'MV_1 - V_1'pp'V_1)^{-1}V_1'h.$$

It is easy to verify that

$$V_1'MV_1 = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 - p_1 & \dots & p_2 - p_1 \\ \vdots & & & \vdots \\ 0 & p_2 - p_1 & \dots & p_{m_0} - p_1 \end{bmatrix} := M_1,$$

$$V_1'p = [p_1, p_2 - p_1, \dots, p_{m_0} - p_1]' := W_1,$$
$$V_1'h = [h_1, h_2 - h_1, \dots, h_{m_0} - h_1]' := H_1.$$

Hence,

$$\mathbb{1}'\Psi^{-1}\mathbb{1} = H_1'(M_1 - W_1W_1')H_1.$$

This process can be repeated, but restricting to the lower right $(m_0 - 1) \times (m_0 - 1)$ submatrix of M_1. In other words, using

$$V_2 = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & -1 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

we obtain

$$\begin{aligned}\mathbb{1}'\Psi^{-1}\mathbb{1} &= H_1'V_2(V_2'M_1V_2 - V_2'W_1W_1'V_2)^{-1}V_2'H_1\\ &= H_2'(M_2 - W_2W_2')H_2.\end{aligned}$$

After $(m_0 - 1)$ such elementary operations, we obtain

$$M_{m_0-1} = \begin{bmatrix} p_1 & 0 & 0 & \dots & 0\\ 0 & p_2 - p_1 & 0 & \dots & 0\\ 0 & 0 & p_3 - p_2 & \dots & 0\\ \vdots & & & & \vdots\\ 0 & 0 & 0 & \dots & p_{m_0} - p_{m_0-1}\end{bmatrix} = \widetilde{M},$$

$$W_{m_0-1} = [p_1, p_2 - p_1, \dots, p_{m_0} - p_{m_0-1}]' = \widetilde{p},$$

$$H_{m_0-1} = [h_1, h_2 - h_1, \dots, h_{m_0} - h_{m_0-1}]' = \widetilde{h},$$

since $\widetilde{p}_1 = p_1$, $\widetilde{p}_i = p_i - p_{i-1}$, $i = 2, \dots, m_0$. Consequently,

$$\mathbb{1}'\Psi^{-1}\mathbb{1} = \widetilde{h}'(\widetilde{M} - \widetilde{p}\widetilde{p}')^{-1}\widetilde{h} = \mathbb{1}'\widetilde{\Psi}^{-1}\mathbb{1}.$$

□

6.4 Cramér–Rao Lower Bounds and Asymptotic Efficiency of the Optimal QCCE

We first recall the notion of efficiency from estimation theory. Suppose that $X_1, \dots, X_N$ is a random sample of size N from a distribution with probability density function $f(x; \vartheta)$, where ϑ is an unknown parameter. For two unbiased estimators $\widehat{\vartheta}_i$ of ϑ, with $E\widehat{\vartheta}_i^2 < \infty$ $(i = 1, 2)$, we say that $\widehat{\vartheta}_1$ is more efficient than $\widehat{\vartheta}_2$ if the relative efficiency $\mathrm{eff}_\vartheta(\widehat{\vartheta}_1|\widehat{\vartheta}_2) < 1$, where

$$\mathrm{eff}_\vartheta(\widehat{\vartheta}_1|\widehat{\vartheta}_2) = \frac{\mathrm{var}_\vartheta(\widehat{\vartheta}_1)}{\mathrm{var}_\vartheta(\widehat{\vartheta}_2)}.$$

Let $\widehat{\vartheta} = \vartheta(X_1, \dots, X_m)$ be an unbiased estimator of ϑ. Under certain regularity conditions (that are usually fulfilled for the identification problems we are working with), the well-known Cramér–Rao (CR) bound states that the variance $\sigma^2_{\widehat{\vartheta}}$ of the estimator $\widehat{\vartheta}$ is bounded below by

$$\sigma^2_{\widehat{\vartheta}} \geq \frac{1}{NE[(\partial/\partial\vartheta)f(X;\vartheta)]^2}. \tag{6.14}$$

We say that an estimator $\widehat{\vartheta}$ is efficient or most efficient if the CR bound is attained. For any unbiased estimator $\widehat{\vartheta}_1$, the efficiency of the estimator is defined as $\mathrm{eff}_\vartheta(\widehat{\vartheta}_1|\widehat{\vartheta})$, where $\widehat{\vartheta}$ is an efficient estimator. An estimator $\widehat{\vartheta}_1$ with sample size N is asymptotically efficient if

(i) $\widehat{\vartheta}_1$ is at least asymptotically unbiased in the sense that $E\widehat{\vartheta}_1 \to \vartheta$ as $N \to \infty$, and

(ii) $\lim_N \mathrm{eff}_\vartheta(\widehat{\vartheta}_1|\widehat{\vartheta}) = 1$.

For further discussion on related issues, we refer the reader to [78, Section 8.5] and the references therein.

Let $\widetilde{\xi}_N^{\{i\}} = \frac{1}{N}\sum_{k=1}^N \widetilde{s}_k^{\{i\}}$, which is the sample relative frequency of y_k taking values in $(C_{i-1}, C_i]$. In the statement of the following lemma, we note that information contained in $\{s_k\}$ is the same as that in $\{\widetilde{s}_k\}$.

Lemma 6.6. *The CR lower bound for estimating θ based on observations of $\{s_k\}$ is*

$$\sigma^2_{\mathrm{CR}}(N, m_0) = \Big(N \sum_{i=1}^{m_0+1} \frac{\widetilde{h}_i^2}{\widetilde{p}_i}\Big)^{-1}. \tag{6.15}$$

Proof. Augment s_k to $s_{ak} = [s_k', 1]'$, where the added element represents $1 = P\{-\infty < y_k < \infty\}$. Let $x_k \in \mathbb{R}^{m_0+1}$ be some possible sample values of s_{ak}. Noting the i.i.d. assumption, the likelihood function of $s_{a1}, \ldots, s_{aN}$ taking values $x_1, \ldots, x_N$, conditioned on θ, is given by

$$\begin{aligned} \ell(x_1, \ldots, x_N; \theta) &= P\{s_{a1} = x_1, \ldots, s_{aN} = x_N; \theta\} \\ &= \prod_{k=1}^N P\{s_{ak} = x_k; \theta\}. \end{aligned}$$

Due to the sensor structure, x_k always takes the form of $[0, \ldots, 0, 1, 1, \ldots, 1]'$. Let $i_0(k)$ be the index of the first 1 in x_k. Then

$$P\{s_{ak} = x_k; \theta\} = P\{\widetilde{s}_k(i_0(k)) = 1; \theta\} = \widetilde{p}_{i_0(k)}.$$

Hence,

$$\ell(x_1, \ldots, x_N; \theta) = \prod_{k=1}^N \widetilde{p}_{i_0(k)}. \tag{6.16}$$

Replace the particular realizations x_k by their corresponding random elements v_k, and denote the resulting quantity by $\ell = \ell(v_1, \ldots, v_N; \theta)$. Note that ℓ is random by its definition. In (6.16), for a given i, $i_0(k) = i$ if and only if $\widetilde{s}_k^{\{i\}} = 1$. Consequently, for a given sample path, the number of occurrences of a particular $\widetilde{p}_i$ in (6.16) is $\sum_{k=1}^N \widetilde{s}_k^{\{i\}} = \widetilde{\xi}_N^{\{i\}} N$. As a result,

by grouping all occurrences of $\widetilde{p}_i$ in (6.16), we have

$$\ell = \prod_{i=1}^{m_0+1} \widetilde{p}_i(\, \widetilde{\xi}_N^{\{i\}} N).$$

Consequently, $\log \ell = N \sum_{i=1}^{m_0+1} \widetilde{\xi}_N^{\{i\}} \log \widetilde{p}_i$, and

$$\frac{\partial \log \ell}{\partial \theta} = N \sum_{i=1}^{m_0+1} \widetilde{\xi}_N^{\{i\}} \frac{1}{\widetilde{p}_i} \widetilde{h}_i,$$
$$\frac{\partial^2 \log \ell}{\partial \theta^2} = N \sum_{i=1}^{m_0+1} \widetilde{\xi}_N^{\{i\}} \left[\frac{-1}{\widetilde{p}_i^2} \widetilde{h}_i^2 + \frac{1}{\widetilde{p}_i} \frac{\partial^2 \widetilde{p}_i}{\partial \theta^2} \right].$$

Since $\sum_{i=1}^{m_0+1} \widetilde{p}_i = 1$, we have

$$\sum_{i=1}^{m_0+1} \frac{\partial^2 \widetilde{p}_i}{\partial \theta^2} = 0.$$

As a result,

$$E \sum_{i=1}^{m_0+1} \frac{\widetilde{\xi}_N^{\{i\}}}{\widetilde{p}_i} \frac{\partial^2 \widetilde{p}_i}{\partial \theta^2} = \sum_{i=1}^{m_0+1} \frac{\partial^2 \widetilde{p}_i}{\partial \theta^2} = 0.$$

Hence,

$$E \frac{\partial^2 \log \ell}{\partial \theta^2} = -N \sum_{i=1}^{m_0+1} \frac{\widetilde{h}_i^2}{\widetilde{p}_i}.$$

The CR lower bound is then given by

$$\sigma_{\mathrm{CR}}^2(N, m_0) = -\Big(E \frac{\partial^2 \log \ell}{\partial \theta^2} \Big)^{-1} = \Big(N \sum_{i=1}^{m_0+1} \frac{\widetilde{h}_i^2}{\widetilde{p}_i} \Big)^{-1}. \tag{6.17}$$

□

Recall that from Theorem 6.2, the variance of the optimal QCCE is

$$\sigma_N^2 = \frac{1}{\mathbb{1}' V_N^{-1}(\theta) \mathbb{1}}.$$

One of the main results of this chapter is the following theorem, which reveals that the optimal QCCE is asymptotically efficient.

Theorem 6.7. *The optimal QCCE is asymptotically efficient in the sense that*

$$N\sigma_N^2 - N\sigma_{\mathrm{CR}}^2(N, m_0) \to 0 \quad \textit{as } N \to \infty.$$

Proof. By Theorems 6.1 and 6.2, the variance of the optimal QCCE satisfies

$$\begin{aligned} N\sigma_N^2 &= N\frac{1}{\mathbb{1}'V_N^{-1}(\theta)\mathbb{1}} \\ &= \frac{1}{\mathbb{1}'N^{-1}V_N^{-1}(\theta)\mathbb{1}} \\ &\to \frac{1}{\mathbb{1}'\Psi^{-1}(\theta)\mathbb{1}} \quad \text{as} \quad N\to\infty, \end{aligned}$$

where $\Psi^{-1}(\theta)$ is the limit of $N^{-1}V_N^{-1}(\theta)$. On the other hand, by Lemma 6.6,

$$N\sigma_{\mathrm{CR}}^2(N,m_0) = \Big(\sum_{i=1}^{m_0+1}\frac{\widetilde{h}_i^2}{\widetilde{p}_i}\Big)^{-1}. \tag{6.18}$$

Now, Lemmas 6.4 and 6.5 yield

$$\mathbb{1}'\Psi^{-1}\mathbb{1} = \sum_{i=1}^{m_0+1}\frac{\widetilde{h}_i^2}{\widetilde{p}_i},$$

which leads to the desired result. □

Remark 6.8. Expression (6.18) delineates the contribution of each sensor interval $(C_{i-1}, C_i]$ to the reduction of identification errors as $\widetilde{h}_i^2(\theta)/\widetilde{p}_i(\theta)$. The smaller the sensitivity $\widetilde{h}_i^2/\widetilde{p}_i$, the less useful is the threshold interval $(C_{i-1}, C_i]$ in error reduction. This may be used as a guide in deciding if increasing the quantization accuracy (that is, adding more thresholds) is worthwhile. Furthermore, the quantity $\sum_{i=1}^{m_0+1}\widetilde{h}_i^2/\widetilde{p}_i$ can be used for threshold selection.

A numerically less complex implementation of the optimal QCCE algorithm is to use the sample mean and sample covariance in place of θ and V_N. From the estimates $\{\Theta_j = G(\xi_j), j=1,\dots,N\}$, we compute its arithmetic average $\overline{\Theta}_N = \sum_{j=1}^N \Theta_j/N$. Since Θ_N is asymptotically unbiased, by elementary analysis, we also have $\sum_{j=1}^N E\Theta_j/N \to \theta\mathbb{1}$ as $N\to\infty$. This leads to the following algorithm:

$$\begin{aligned} \overline{\Theta}_N &= \sum_{j=1}^N \Theta_j/N, \\ \widehat{V}_N &= \frac{1}{N-1}\sum_{j=1}^N(\Theta_j-\overline{\Theta}_N)(\Theta_j-\overline{\Theta}_N)', \\ \beta_N &= \frac{\widehat{V}_N^{-1}\mathbb{1}}{\mathbb{1}'\widehat{V}_N^{-1}\mathbb{1}}, \\ \widehat{\theta}_N &= (\beta_N)'\Theta_N. \end{aligned} \tag{6.19}$$

This algorithm can be written recursively as

$$\begin{aligned}\overline{\Theta}_N &= \overline{\Theta}_{N-1} - \frac{1}{N}\overline{\Theta}_{N-1} + \frac{\Theta_N}{N},\\ \widehat{V}_N &= \widehat{V}_{N-1} - \frac{1}{N-1}\widehat{V}_{N-1} + \frac{(\Theta_N - \overline{\Theta}_N)(\Theta_N - \overline{\Theta}_N)'}{N-1}.\end{aligned} \tag{6.20}$$

It can be shown that

$$\widehat{V}_N - V_N(\theta) \to 0, \quad \widehat{V}_N^{-1} - V_N^{-1}(\theta) \to 0, \text{ as } N \to \infty. \tag{6.21}$$

If we work with the implementable estimates defined in (6.19), we can define $\widehat{\sigma}_N^2 = 1/(\mathbb{1}'\widehat{V}_N^{-1}\mathbb{1})$, and obtain the following result.

Corollary 6.9. *For the estimators given in* (6.19)*,*

$$N\widehat{\sigma}_N^2 - N\sigma_{\text{CR}}^2(N, m_0) \to 0 \quad \textit{as } N \to \infty.$$

Proof. By virtue of Theorem 6.7, it suffices to show that $N\widehat{\sigma}_N^2 - N\sigma_N^2 \to 0$ as $N \to \infty$. In fact, a simple calculation shows that

$$\frac{1}{\mathbb{1}'\widehat{V}_N^{-1}\mathbb{1}} - \frac{1}{\mathbb{1}'V_N^{-1}\mathbb{1}} = \frac{\mathbb{1}'(V_N^{-1} - \widehat{V}_N^{-1})\mathbb{1}}{(\mathbb{1}'\widehat{V}_N^{-1}\mathbb{1})(\mathbb{1}'V_N^{-1}\mathbb{1})} \to 0 \text{ as } N \to \infty. \tag{6.22}$$

Hence, the result follows. □

6.5 Notes

This chapter introduces the optimal quasi-convex combination estimators and establishes their asymptotic efficiency. These optimality results lay a foundation in which complexity issues in system identification with quantized observations can be rigorously investigated. This chapter follows [104].

Space complexity (the number of intervals in quantization, or the word length of each measurement) is a relatively new paradigm in system identification. Traditional quantization uses uniform quantization intervals and ubiquitously employs quantization errors in analysis. However, when the signal range is large or even unbounded, such uniform quantization suffers from high or infinite space complexity. The results of this chapter provide a foundation to evaluate if finite quantization levels are sufficient. Studies of the impact of quantization errors can also be conducted in a worst-case or probabilistic framework, depending on how quantization errors are modeled [1, 2, 39, 34, 80].

7

Input Design for Identification in Connected Systems

Input design is of essential importance in system identification to provide sufficient probing capabilities to guarantee the convergence of parameter estimators to their true values; namely, the estimators are consistent. Input conditions for consistent estimation depend on sensor characteristics, system configurations, noise locations and distributions, and identification algorithms. The previous chapters consider only the basic formulation in which the input u_k can be directly designed. This chapter covers input design in more general system configurations.

The system configurations illustrated in Figure 2.2 represent typical scenarios in which identification experiments must be performed. They introduce challenges in input design, signal measurements, and interaction with control tasks. In these configurations, the input u to the plant P, which is either FIR or rational with n_0 parameters, may be measured with noise corruption but cannot be directly selected. Only the external input r can be designed. In these configurations, u_k is the output of a possibly unknown stable system with input r_k. This chapter resolves several issues that are critical for applying the methods of this book to system identification in filtering and closed-loop systems.

Section 7.1 establishes conditions under which a periodic and full-rank signal will retain these features after passing through a stable system, even when the system is unknown. As a result, the design of external probing signals or dithers can be easily accomplished. Section 7.2 details such a design, especially for the typical case of tracking control. In general, input noises, including input measurement noises and actuator noises, will affect signal rank and introduce identification bias. Sections 7.3 and 7.4 are

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_7,

devoted to developing input design principles and modified algorithms to recover signal richness and gain convergence.

7.1 Invariance of Input Periodicity and Rank in Open- and Closed-Loop Configurations

A condition for u_k to provide sufficient probing capability for the convergence of parameter estimates is that u is n_0-periodic and full rank. In this chapter, such conditions will be called "sufficient richness" conditions, to avoid confusion with the typical input "persistent excitation" conditions. Here we would like to establish relationships between periodicity and rank properties of the external signal r and those of u.

Let H be a linear time-invariant and stable system with impulse response $\{h_k\}$. Suppose that $u = Hr$, or in the time domain

$$u_k = \sum_{l=0}^{\infty} h_l r_{k-l}. \tag{7.1}$$

Suppose that the discrete Fourier transform (DFT) of H is

$$H(e^{i\omega}) = \sum_{l=0}^{\infty} h_l e^{-i\omega l}.$$

Theorem 7.1. *Suppose that r is n_0-periodic and full rank. Then u is also n-periodic and full rank if and only if $H(e^{i\omega}) \neq 0$, for $\omega = \omega_k := (2\pi k/n_0), k = 1, \ldots, n_0$.*

Proof. Since r is n_0-periodic and full rank, by Corollary 2.3, the frequency samples of r are nonzero, $R_k = \sum_{l=1}^{n_0} r_l e^{-i\omega_k l} \neq 0$, $k = 1, \ldots, n_0$. Since r is n_0-periodic, and H is LTI and stable, u is also n_0-periodic after a short transient. Furthermore, the frequency samples U_k of u are related to R_k by

$$\begin{aligned} U_k &= \sum_{l=1}^{n_0} u_l e^{-i\omega_k l} = \sum_{l=1}^{n_0} \sum_{t=0}^{\infty} h_t r_{l-t} e^{-i\omega_k l} \\ &= \sum_{t=0}^{\infty} h_t e^{-i\omega_k t} \sum_{l=1}^{n_0} r_{l-t} e^{-i\omega_k (l-t)} = H(e^{i\omega_k}) R_k. \end{aligned}$$

Here, the cyclic property of the DFT is applied:

$$R_k = \sum_{l=1}^{n_0} r_l e^{-i\omega_k l} = \sum_{l=1}^{n_0} r_{l-t} e^{-i\omega_k (l-t)}.$$

By Corollary 2.3, u is full rank if and only if $U_k \neq 0$, $k = 1, \ldots, n_0$. However, by hypothesis, $R_k \neq 0$, $k = 1, \ldots, n_0$. As a result, $U_k \neq 0$ if and only if $H(e^{i\omega_k}) \neq 0$, $k = 1, \ldots, n_0$. □

Example 7.2. The necessity of the condition of Theorem 7.1 can be verified by examining the following second-order system: $u_k = r_k + r_{k-1}$. When r is a 2-periodic signal and full rank, u_k is a constant and hence is not rank 2. This is due to the fact that $H(e^{i\omega}) = 1 + e^{i\omega}$ and for $\omega = \omega_1 = 2\pi/2 = \pi$, $H(e^{i\omega_1}) = 0$.

Remark 7.3. Theorem 7.1 claims that for any system H not having annihilating zeros at n_0 points $e^{i\omega_k}$, $\omega_k = (2\pi k/n_0), k = 1, \ldots, n_0$, on the unit circle, sufficient richness capability of the signal r is always preserved after passing through H. In particular, for the feedback configuration in Figure 2.2, we have the following result indicating that input richness properties are invariant under a feedback mapping.

(A7.1) Consider the feedback configuration in (b) of Figure 2.2. Assume that for $\omega_k = (2\pi k/n_0), k = 1, \ldots, n_0$, $K(e^{i\omega})$ does not have zeros at ω_k; and $P(e^{i\omega})$ and $F(e^{i\omega})$ do not have singularities (such as poles) at ω_k.

Corollary 7.4. *Under Assumption* (A7.1), $M = K/(1 + PKF)$ *does not have annihilating zeros at* $\omega_k = 2\pi k/n_0, k = 1, \ldots, n_0$. *As a result, if* r *is* n_0*-periodic and full rank, so is* u.

Proof. From

$$M(e^{i\omega}) = \frac{K(e^{i\omega})}{1 + P(e^{i\omega})K(e^{i\omega})F(e^{i\omega})},$$

it is clear that the zeros of M are either the zeros of K or the singularities (such as poles) of P or F. By assumption (A7.1), $K(e^{i\omega_k}) \neq 0$, and ω_k is not a singularity point of $P(e^{i\omega})$ or $F(e^{i\omega})$. Hence, $M(e^{i\omega_k}) \neq 0$, $k = 1, \ldots, n_0$. Now by Theorem 7.1, u is n_0-periodic and full rank whenever r is n_0-periodic and full rank. □

7.2 Periodic Dithers

Consider the tracking configuration in Figure 7.1. When the desired output is r_0, usually $r = r_0$ is the set point. However, a constant $r_0 \neq 0$ is 1-periodic. It is only good for the identification of a gain system (namely, $n_0 = 1$). This is an indication that the goals of control and identification are usually not consistent.

To enhance probing capability, one may add a small n_0-periodic dither ϖ_k to r_0, leading to $r_k = \varpi_k + r_0$. Since

$$u_k = Mr_k = M\varpi_k + Mr_0 = v_k + \mu,$$

where v_k is an n_0-periodic signal and $\mu = Mr_0$ a constant, we need to establish rank conditions on u_k.

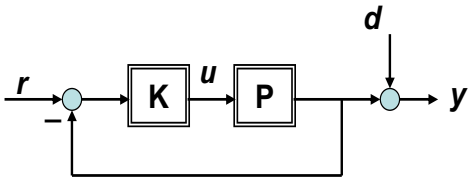


FIGURE 7.1. Tracking configuration

More generally, consider an input signal u: $u_k = v_k + e_k$, which is a perturbation of v. Suppose that v_k is n_0-periodic and full rank. We would like to establish conditions under which u_k is also n_0-periodic and full rank.

(A7.2) Both v_k and e_k are n_0-periodic.

Under Assumption (A7.2), the Toeplitz matrices for v, e, and u, denoted by Φ_v, Φ_e, and Φ_u, respectively, are circulant matrices. Let their corresponding frequency samples be

$$\Gamma^u = \mathcal{F}[u] = \{\gamma_k^u, k = 1, \ldots, n_0\},$$
$$\Gamma^v = \mathcal{F}[v] = \{\gamma_k^v, k = 1, \ldots, n_0\},$$
$$\Gamma^e = \mathcal{F}[e] = \{\gamma_k^e, k = 1, \ldots, n_0\}.$$

Theorem 7.5. *Under Assumption* (A7.2)*, u is full rank if and only if $\gamma_k^v + \gamma_k^e \neq 0$, $k = 1, \ldots, n_0$.*

Proof. This follows immediately from $\gamma_k^u = \gamma_k^v + \gamma_k^e$ and the fact that Φ_u is full rank if and only if its frequency samples do not contain 0. □

We now consider the special case when $e_k \equiv \mu$, which is a typical case in tracking problems as shown above.

Corollary 7.6. *Suppose v_k is n_0-periodic and full rank and $e_k = \mu$. Then u_k is n_0-periodic. Let $\eta = \frac{1}{n_0}\sum_{j=1}^{n_0} v_j$. u is full rank if and only if $\mu \neq -\eta$.*

Proof. Since v_k is full rank, by Corollary 2.3 we have $\gamma_k^v \neq 0$, $k = 1, \ldots, n_0$. In particular,

$$\gamma_n^v = \sum_{j=1}^{n_0} v_j = n_0\eta.$$

Moreover, the frequency samples of $e_k \equiv \mu$ are

$$\gamma_k^e = 0, \ k = 1, \ldots, n_0 - 1, \text{ and } \gamma_{n_0}^e = n_0\mu.$$

Consequently, by Theorem 7.5, u_k is full rank if and only if

$$\gamma_{n_0}^v + \gamma_{n_0}^e \neq 0.$$

That is, $n_0\eta + n_0\mu \neq 0$, or $\mu \neq -\eta$, as claimed. □

Corollary 7.6 may be verified directly by matrix manipulations. Toeplitz matrices Φ_u, Φ_v, and Φ_e for u, v, and e, respectively, are

$$\begin{aligned}\Phi_u &= \Phi_v + \Phi_e \\ &\sim \begin{bmatrix} n_0\eta + n_0\mu & 0 & \dots & 0 \\ v_1 + \mu & v_{n_0} - v_1 & \ddots & v_2 - v_1 \\ \vdots & \ddots & \ddots & \vdots \\ v_{n_0-1} + \mu & v_{n_0-2} - v_{n_0-1} & \dots & v_{n_0} - v_{n_0-1} \end{bmatrix},\end{aligned}$$

by adding the second to nth rows to the first row, followed by subtracting the first column from the second to nth columns. The last matrix is full rank since $\eta+\mu \neq 0$ and the lower right $(n_0-1)\times(n_0-1)$ submatrix, which is obtained by performing elementary operations from Φ_v, is full rank.

7.3 Sufficient Richness Conditions under Input Noise

Under the system configurations in Figure 2.2, $u = Mr$ is generated from r by a possibly unknown system M. In the previous sections, u is assumed to be accurately measured. When u is further corrupted by noise, it can no longer be exactly measured. Furthermore, the actual values of u cannot be directly derived from r since M is unknown. Sufficient richness conditions and identification algorithms under this scenario will be explored in this section.[7.1]

We will consider two cases of input noises shown in Figure 7.2:

1. Input measurement noise ε_k: When u is measured by a regular sensor, the measured values are related to u by $w_k = u_k + \varepsilon_k$, where ε_k is the measurement noise.

2. Actuator noise e_k: In this case, the actual input to the plant is $u_k = v_k + e_k$, where $v_k = Mr$, and e_k is the actuator noise.

As a result, the measured input is $w_k = v_k + e_k + \varepsilon_k$ and identification of the plant must be performed from the observation data on w_k and $s_k = \mathcal{S}(y_k)$.

[7.1] Input noises cause errors in the regressors, leading to a case of errors-in-variables. It is well known in statistical analysis that errors-in-variables will introduce estimation bias. The discussions here involve further complications since the input u cannot be directly designed and measurements are binary valued.

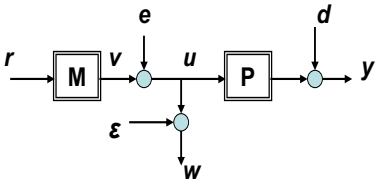


FIGURE 7.2. Input noise configuration

(A7.3) $\{v_k\}$ is n_0-periodic and full rank. $\{e_k\}$ and $\{\varepsilon_k\}$ are sequences of i.i.d. random variables with zero mean and finite variances such that $\{e_k\}$ and $\{\varepsilon_k\}$ are independent.

Denote the $n_0 \times n_0$ Toeplitz matrices for w and v by

$$\Phi_l^w = \begin{bmatrix} w_{ln_0} & w_{ln_0-1} & \cdots & w_{ln_0-n_0+1} \\ w_{ln_0+1} & w_{ln_0} & \ddots & w_{ln_0-n_0+2} \\ \vdots & \ddots & \ddots & \vdots \\ w_{ln_0+n_0-1} & w_{ln_0+n_0-2} & \cdots & w_{ln_0} \end{bmatrix},$$

$$\Phi^v = \begin{bmatrix} v_{n_0} & v_{n_0-1} & \cdots & v_1 \\ v_1 & v_{n_0} & \ddots & v_2 \\ \vdots & \ddots & \ddots & \vdots \\ v_{n_0-1} & v_{n_0-2} & \cdots & v_{n_0} \end{bmatrix}.$$

Although Φ_l^w is not circulant and varies with l, the limit of their averages is a full-rank circulant matrix.

Lemma 7.7. *Under Assumption* (A7.3), $\sum_{l=1}^{N} \Phi_l^w / N \to \Phi^v$ *w.p.1 as* $N \to \infty$.

Proof. This follows directly from the strong law of large numbers, applied to each element of the matrices. □

We consider first the case of measurement noise only. Actuator noises will be discussed in the next section. In this case, $e_k = 0$, for all k. Hence, $u_k = v_k$, $w_k = u_k + \varepsilon_k$, and $\Phi^u = \Phi^v := \Phi$. Due to measurement noise, the actual u_k is unknown. As a result, Φ is unknown and cannot be used directly in identification algorithms. However, by Lemma 7.7 it can be estimated asymptotically by averaging. The following algorithm utilizes this idea to estimate θ.

We shall use the FIR model for discussion here,

$$y_k = \phi_k' \theta + d_k. \tag{7.2}$$

Assume that y_k is measured by a binary-valued sensor of threshold C. We use the following notation for elementwise vector functions. For the distribution function $F(\cdot)$ and a vector $x = [x_1, \ldots, x_{n_0}]' \in \mathbb{R}^{n_0}$, we define

$$\begin{aligned} F(x) &= [F(x_1), \ldots, F(x_{n_0})]' \in \mathbb{R}^{n_0}, \\ G(x) &= [F^{-1}(x_1), \ldots, F^{-1}(x_{n_0})]' \in \mathbb{R}^{n_0}. \end{aligned} \tag{7.3}$$

Similarly, for

$$\alpha = [\alpha_1, \ldots, \alpha_{m_0}]' \text{ and } c = [c_1, \ldots, c_{m_0}]' \in \mathbb{R}^{n_0},$$

write

$$\mathbf{I}_{\{\alpha \le c\}} = [I_{\{\alpha_1 \le c_1\}}, \ldots, I_{\{\alpha_{m_0} \le c_{m_0}\}}]'.$$

We use $\mathbb{1}_\ell$ and $\mathbf{0}_\ell \in \mathbb{R}^\ell$ to denote column vectors of dimension ℓ with all components being 1 and 0, respectively.

Recall that if Φ were known, a consistent estimator of θ would be

$$\theta_N = \Phi^{-1}(C\mathbb{1} - G(\xi_N)).$$

Define $s_l = [s_{ln_0}, \ldots, s_{ln_0+n_0-1}]'$. This estimator is no longer causal since it employs the unknown Φ in computing θ_N. In other words, one needs the future information on the sequence $\{w_k\}$ in computing θ_N. The following algorithm replaces the future information Φ by a sample average.

Let

$$\Phi_N = \frac{1}{N} \sum_{l=1}^{N} \Phi_l^w.$$

When Φ_N is nonsingular, define

$$\theta_N = \Phi_N^{-1}(C\mathbb{1} - G(\xi_N)).$$

This estimator can be recursively defined as follows.

1. Initial conditions: $\xi_1 = s_1$, $\Phi_1 = \Phi_1^w$ is generated from initial data on w, $\theta_1 = 0$.

2. Recursion: Suppose that at N, ξ_N, Φ_N, and θ_N have been obtained. Then at $N+1$, we update

$$\begin{aligned} \xi_{N+1} &= \xi_N - \frac{1}{N+1}\xi_N + \frac{1}{N+1} s_{N+1}, \\ \Phi_{N+1} &= \Phi_N - \frac{1}{N+1}\Phi_N + \frac{1}{N+1}\Phi_{N+1}^w, \\ \theta_{N+1} &= \begin{cases} \Phi_{N+1}^{-1}(C\mathbb{1} - G(\xi_{N+1})), \text{ if } \Phi_{N+1} \text{ is nonsingular,} \\ \theta_N, \text{ if } \Phi_{N+1} \text{ is singular.} \end{cases} \end{aligned} \tag{7.4}$$

Theorem 7.8. *Under Assumption* (A7.3), $\theta_N \to \theta$ *w.p.1 as* $N \to \infty$.

Proof. Since the true input to the plant is u, $\xi_N \to \xi = F(C\mathbb{1} - \Phi\theta)$ w.p.1. Then $\theta_N - \theta = \Phi_N^{-1}(G(\xi) - G(\xi_N)) + (\Phi_N^{-1} - \Phi^{-1})(C\mathbb{1} - G(\xi))$. By the strong law of large numbers, the convergence $\theta_N - \theta \to 0$ w.p.1 follows from $\Phi_N \to \Phi$ and $\xi_N \to \xi$ w.p.1, the continuity of F^{-1}, and the invertibility of Φ. □

7.4 Actuator Noise

Unlike measurement noise ε_k that affects measured input values but does not enter the plant, actuator noise e_k affects the output y_k of the plant. Consider the case $u_k = v_k + e_k$ and $w_k = u_k$. To understand the impact of e_k, we express the regressor in (7.2) by ϕ_k^u or ϕ_k^v, depending on which signal is used in the regressor. Under Assumption (A7.3), v is n_0-periodic and full rank, but u is not periodic. However, by Lemma 7.7,

$$\frac{1}{N}\sum_{l=1}^{N} \Phi_l^u \to \Phi^v \quad \text{w.p.1 as} \quad N \to \infty.$$

Since $u_k = v_k + e_k$, we have

$$\begin{aligned} y_k &= (\phi_k^u)'\theta + d_k = (\phi_k^v)'\theta + (\phi_k^e)'\theta + d_k \\ &= (\phi_k^v)'\theta + z_k. \end{aligned}$$

Observe that the equivalent noise z_k is

$$\begin{aligned} z_k &= (\phi_k^e)'\theta + d_k \\ &= a_0 e_k + \cdots + a_{n_0-1} e_{k-n_0+1} + d_k. \end{aligned}$$

Under Assumption (A7.3), although $\{z_k\}$ may not be independent, it is strictly stationary. Recall that $\{z_k\}$ is strictly stationary if for any positive integer ν, points $t_1, \ldots, t_\nu \in \mathbb{Z}_+$ and $l \in \mathbb{Z}_+$, the joint distribution of $\{z_{t_1}, \ldots, z_{t_\nu}\}$ is the same as that of $\{z_{t_1+l}, \ldots, z_{t_\nu+l}\}$ (i.e., its finite-dimensional distributions are translation invariant; see [47, p. 443]). Denote the distribution function by $F_z(x;\theta)$. A moment of reflection reveals that the sequence is $(n_0 - 1)$-dependent. A precise definition of $(n_0 - 1)$-dependence can be found in [8, p. 167, Example 1]. Since an $(n_0 - 1)$-dependent sequence belongs to the class of ϕ-mixing signals, whose remote past and distant future are asymptotically independent, the sequence is strongly ergodic [47, p. 488]. That is, a strong law of large numbers still holds.

Following (7.4), define

$$\xi_N = \frac{1}{N}\sum_{j=1}^{N} s_j.$$

Let θ_N be the solution to

$$\xi_N = F_z(C\mathbb{1} - \Phi\theta_N; \theta_N). \tag{7.5}$$

For any ϑ, define the Jacobian matrix

$$J(\vartheta) = \frac{\partial F_z(C\mathbb{1} - \Phi\vartheta; \vartheta)}{\partial \vartheta}.$$

A sufficient condition for invertibility of the function in (7.5) is that $J(\theta_N)$ is full rank. In this case, by denoting the inverse of $\xi = F_z(C\mathbb{1} - \Phi\vartheta; \vartheta)$ as $\vartheta = H(\xi)$, the estimate θ_N in (7.5) may be symbolically written as $\theta_N = H(\xi_N)$.

Proposition 7.9. *If $H(\cdot)$ exists and is continuous, then $\theta_N \to \theta$ w.p.1 as $N \to \infty$.*

Proof. By the strong law of large numbers, $\xi_N \to \xi = F_z(C\mathbb{1} - \Phi\theta; \theta)$ w.p.1. Since $H(\cdot)$ exists and is continuous, $\theta_N = H(\xi_N) \to H(\xi) = \theta$ w.p.1. □

For a given ϑ, denote the inverse of $F_z(x; \vartheta)$ (with respect to x) by

$$G_z(x; \vartheta) = F_z^{-1}(x; \vartheta). \tag{7.6}$$

Computationally, it is observed that for a given ξ, the implicit function $\xi = F_z(C\mathbb{1} - \Phi\vartheta; \vartheta)$ of ϑ may be expressed as a fixed-point equation $\vartheta = \Phi^{-1}(C\mathbb{1} - G_z(\xi; \vartheta))$.

Next, a special case will be considered. Suppose that $\{e_k\}$ is a sequence of i.i.d. normal random variables with zero mean and variance σ_e^2, and $\{d_k\}$ is a sequence of i.i.d. normal random variables with zero mean and variance σ_d^2. Then,

$$z = a_0 e_k + \cdots + a_{n_0-1} e_{k-n_0+1} + d_k$$

is also normally distributed and has zero mean and variance

$$\sigma_z^2(\theta) = (a_0^2 + \cdots + a_{n_0-1}^2)\sigma_e^2 + \sigma_d^2 = \sigma_e^2\|\theta\|^2 + \sigma_d^2.$$

Let $F_0(x)$ be the normal distribution function of zero mean and variance 1. Then $F_z(x; \vartheta) = F_0(x/\sigma_z(\vartheta))$. It follows that

$$F_z(C\mathbb{1} - \Phi\vartheta; \vartheta) = F_0\left(\frac{C\mathbb{1} - \Phi\vartheta}{\sigma_z(\vartheta)}\right),$$

and the Jacobian matrix is

$$\begin{aligned} J(\vartheta) &= \frac{dF_z(C\mathbb{1} - \Phi\vartheta; \vartheta)}{d\vartheta} \\ &= -\frac{1}{\sigma_z}\frac{dF_0}{dx}\left[\Phi\left(I_{n_0} - \frac{\sigma_e^2\vartheta\vartheta'}{\sigma_z^2}\right) + \frac{\sigma_e^2 C\mathbb{1}\vartheta'}{\sigma_z^2}\right], \end{aligned}$$

where

$$x = \frac{C\mathbb{1} - \Phi\vartheta}{\sigma_z(\vartheta)}.$$

Since

$$\frac{dF_0}{dx} = \operatorname{diag}(f_z(C - \phi_1'\vartheta), \ldots, f_z(C - \phi_{n_0}'\vartheta))$$

is full rank, where f_z is the density function of F_z, the Jacobian matrix $J(\vartheta)$ is full rank if and only if

$$\Phi(I_{n_0} - \sigma_e^2\vartheta\vartheta'/\sigma_z^2) + \sigma_e^2 C\mathbb{1}\vartheta'/\sigma_z^2$$

is full rank.

Remark 7.10. It is easily verified that if A is an n_0-dimensional square matrix with $\|A\| < 1$, then $I_{n_0} + A$ is invertible, where I_{n_0} denotes the $n_0 \times n_0$ identity matrix. Moreover, if A is an n_0-dimensional invertible matrix and $\|B\| < \|A^{-1}\|^{-1}$, then $A + B$ is invertible.

Theorem 7.11. *If*

$$\|\Phi^{-1}\| < \frac{2\sigma_d^3}{C\sigma_e\sqrt{n_0}(\sigma_e^2\|\theta\|^2 + \sigma_d^2)}, \tag{7.7}$$

then $\theta_N = H(\xi_N) \to \theta$ *w.p.1 as* $N \to \infty$.

Proof. Noting that

$$\left\|\frac{\sigma_e^2\theta\theta'}{\sigma_z^2}\right\| = \left\|\frac{\sigma_e^2\theta\theta'}{\sigma_e^2\theta'\theta + \sigma_d^2}\right\| = \frac{\sigma_e^2\theta'\theta}{\sigma_e^2\theta'\theta + \sigma_d^2} < 1,$$

by Remark 7.10, $I_{n_0} - \sigma_e^2\theta\theta'/\sigma_z^2$ is full rank. Since

$$\left\|\frac{\sigma_e^2 C\mathbb{1}\theta'}{\sigma_z^2}\right\| \le \frac{\sigma_e^2\|C\mathbb{1}\|\|\theta\|}{\sigma_e^2\theta'\theta + \sigma_d^2} \le \frac{\sigma_e^2 C\sqrt{n}\|\theta\|}{2\sigma_e\sigma_d\|\theta\|} = \frac{\sigma_e C\sqrt{n}}{2\sigma_d},$$

we have

$$\begin{aligned} &\left\|\frac{\sigma_e^2 C\mathbb{1}\theta'}{\sigma_z^2}\left(I_{n_0} - \frac{\sigma_e^2\theta\theta'}{\sigma_z^2}\right)^{-1}\right\| \\ &= \left\|\frac{\sigma_e^2 C\mathbb{1}\theta'}{\sigma_z^2}\sum_{i=0}^{\infty}\left(\frac{\sigma_e^2\theta\theta'}{\sigma_z^2}\right)^i\right\| < \|\Phi^{-1}\|^{-1}. \end{aligned}$$

By Remark 7.10,

$$\Phi + \frac{\sigma_e^2 C \mathbb{1}\theta'}{\sigma_z^2}\left(I_{n_0} - \frac{\sigma_e^2 \theta\theta'}{\sigma_z^2}\right)^{-1}$$

is invertible. Then,

$$\Phi\left(I_{n_0} - \frac{\sigma_e^2 \theta\theta'}{\sigma_z^2}\right) + \frac{\sigma_e^2 C \mathbb{1}\theta'}{\sigma_z^2}$$

is invertible. So, $J(\theta)$ is invertible. Hence, Proposition 7.9 confirms that $\theta_N \to \theta$ w.p.1. □

Remark 7.12. Equation (7.7) can be used to design input signals. Indeed, suppose that the prior information on the unknown parameters is that $\|\theta\| \leq \beta$. By using β^2 in place of $\|\theta\|^2$, one can design an input such that Φ satisfies (7.7). Consequently, consistency of the estimates will be guaranteed for any $\theta \in \{\vartheta : \|\vartheta\| \leq \beta\}$.

Example 7.13. Suppose the true system is $y_k = 0.9u_k + 1.1u_{k-1} + d_k$. Hence, the true parameters are $\theta = [0.9, 1.1]'$ and $\|\theta\|^2 = 1.93$. Assume that the prior information on θ is that $\|\theta\|^2 \leq 2$. The output measurement noise d_k is i.i.d. normally distributed with zero mean and variance $\sigma_d^2 = 4$. The input signal $u_k = v_k + e_k$, where v_k is 2-periodic with its one-period values $v_1 = 3, v_2 = 15$, and e_k is a sequence of i.i.d., normally distributed noise of zero mean and variance $\sigma_e^2 = 1$. By direct calculation, $\|\Phi^{-1}\| = 0.083$. For $C = 20$, and the prior information $\|\theta\|^2 \leq 2$, the right-hand side of (7.7) is 0.094. Hence, the input satisfies condition (7.7). In fact, under this input, (7.7) is satisfied for all $\theta \in \{\vartheta : \|\vartheta\|^2 \leq 2\}$.

An identification algorithm is devised for this example. At each step N, ξ_N is calculated from (7.4). Then the estimate θ_N is derived by solving (7.5). The inverse function of normal distribution is calculated by the Matlab function *norminv*. The simulation in Figure 7.3 illustrates the convergence of parameter estimates. The relative estimation error $\|\theta_N - \theta\|/\|\theta\|$ is used to evaluate the accuracy and convergence of the estimates. Figure 7.3 shows the parameter convergence of this algorithm.

7.5 Notes

This chapter presents conditions on input ensembles that provide sufficiently rich probing power for convergence of parameter estimates. It is an extension of the basic input design covered in the previous chapters and is based on [109]. We summarize here several reasons that periodic signals are of essential importance for quantized identification.

The classical control theory of Bode and Nyquist characterizes systems by using periodic input signals (frequency responses). They are relatively easy

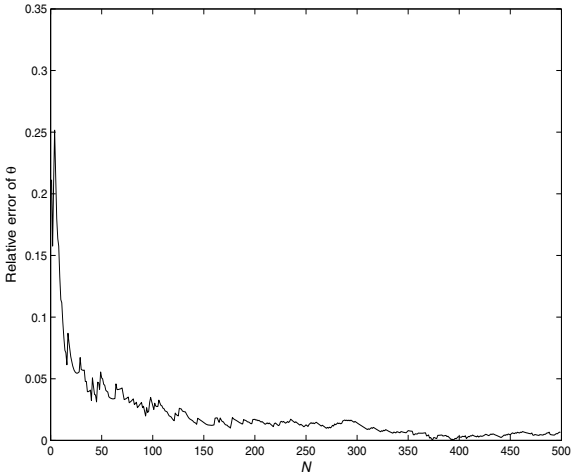


FIGURE 7.3. Relative errors of parameter estimates

to apply and there are many special devices for obtaining system frequency responses. Periodic inputs are especially useful for quantized identification for several technical reasons:

(1) Periodic inputs are uniformly bounded. In contrast, to identify stochastic systems, a typical method uses Gaussian-distributed signals that are unbounded and more difficult to apply in practical systems. The truncation of unbounded signals due to input saturation may cause bias in system identification.

(2) Essential features for a periodic signal to be rich for identification are certain rank conditions, rather than the magnitudes of the signals. As a result, one may use small probing inputs for identification with the benefit of contained perturbation to system operations.

(3) Periods and ranks of periodic signals are shift invariant. As such, they are natural choices for achieving "persistent identification" for time-varying systems [97, 103].

(4) As established in this chapter, periods and ranks of periodic signals are invariant after passing through a linear-time-invariant system (with some mild conditions). Consequently, an externally applied periodic signal can be easily designed for identification of a plant in a closed-loop setting [103].

(5) As shown in the previous chapters, under periodic inputs, the identification of a system with multiple parameters under quantized sensors can often be reduced to a number of greatly simplified identification problems for gains.

(6) Under periodic inputs, our algorithms have been shown to be asymptotically optimal, since they achieve the CR lower bounds asymptotically.

8

Identification of Sensor Thresholds and Noise Distribution Functions

The developments in the early chapters rely on the knowledge of the distribution function $F(\cdot)$ or its inverse, as well as the threshold C. However, in many applications, the noise distributions are not known, or only limited information is available. On the other hand, input–output data from the system contain information about the noise distribution. By viewing unknown distributions and system parameters jointly as uncertainties, we develop a methodology of joint identification.

Section 8.1 deals with unknown sensor thresholds. The threshold C is added as an additional unknown parameter to be identified together with the primary system parameters. Unknown distribution functions are more difficult to handle and are treated in the remaining sections. Section 8.2 discusses parameterization of distribution functions so that joint identification remains parametric. Joint identification problems are formulated in Section 8.3. Section 8.4 delineates input conditions that will render the system identifiable. The main algorithms are introduced in Section 8.5, whose convergence properties are derived in Section 8.6. For practical implementations, Section 8.7 introduces some recursive algorithms, followed by a graphical summary of the algorithms in Section 8.8 and some illustrative examples in Section 8.9.

8.1 Identification of Unknown Thresholds

The main relationship in computing estimates is the equation $\xi = F(C\mathbb{1} - \Phi\theta)$. When C is unknown, this relationship is not sufficient to determine θ

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_8,

and C, since it has n_0 equations but $n_0 + 1$ unknowns. We introduce the following modified algorithm to estimate C and θ collectively.

We carry out our discussions under both input and output measurement noises. Namely, $w_k = u_k + \varepsilon_k$, and $y_k = \phi_k'\theta + d_k$, where $\{d_k\}$ is a sequence of random variables satisfying (A3.1).

8.1.1 Sufficient Richness Conditions

(A8.1) Suppose that $\{u_k\}$ is $(n_0 + 1)$-periodic and full rank, and that $\{\varepsilon_k\}$ is an i.i.d. sequence with zero mean.

From $y_k = \phi_k'\theta + d_k$, $k = 1, 2, \ldots$, define

$$\begin{aligned}
\widetilde{Y}_j &= [y_{(j-1)(n_0+1)+1}, \ldots, y_{j(n_0+1)}]' \in \mathbb{R}^{(n_0+1)},\\
\widetilde{\Phi}_j &= [\phi_{(j-1)(n_0+1)+1}, \ldots, \phi_{j(n_0+1)}]' \in \mathbb{R}^{(n_0+1)\times n_0},\\
\widetilde{D}_j &= [d_{(j-1)(n_0+1)+1}, \ldots, d_{j(n_0+1)}]' \in \mathbb{R}^{(n_0+1)},\\
\widetilde{S}_j &= [s_{(j-1)(n_0+1)+1}, \ldots, s_{j(n_0+1)}]' \in \mathbb{R}^{(n_0+1)}.
\end{aligned}$$

Then,

$$\widetilde{Y}_j = \widetilde{\Phi}_j\theta + \widetilde{D}_j, \quad \text{for} \quad j = 1, 2, \ldots$$

Note that $\{\widetilde{\Phi}_j\}$ is a sequence of $(n_0 + 1) \times n_0$ matrices, generated from u. Due to measurement noise, the actual u_k is unknown and only w_k can be used in algorithms. Define

$$\begin{aligned}
\widetilde{\xi}_N &= \frac{1}{N}\sum_{l=1}^{N} \widetilde{S}_l,\\
\widetilde{\Psi}_N^w &= \frac{1}{N}\sum_{l=1}^{N} \widetilde{\Phi}_l^w,
\end{aligned}$$

where

$$\widetilde{\Phi}_l^w = \begin{bmatrix}
w_{l(n_0+1)} & w_{l(n_0+1)-1} & \cdots & w_{l(n_0+1)-n_0+1}\\
w_{l(n_0+1)+1} & w_{l(n_0+1)} & \ddots & w_{l(n_0+1)-n+2}\\
\vdots & \ddots & \ddots & \vdots\\
w_{l(n_0+1)+n_0} & w_{l(n_0+1)+n_0-1} & \cdots & w_{l(n_0+1)+1}
\end{bmatrix}.$$

Under Assumption (A8.1), $\widetilde{\Psi}_N^w \to \widetilde{\Psi}$ w.p.1, where

$$\widetilde{\Psi} = \begin{bmatrix} u_{n_0+1} & u_{n_0} & \dots & u_2 \\ u_1 & u_{n_0+1} & \ddots & u_3 \\ \vdots & \ddots & \ddots & \vdots \\ u_{n_0} & u_{n_0-1} & \dots & u_1 \end{bmatrix}.$$

Define $\overline{\Psi}_N^w = [\mathbb{1}_{n_0+1}, -\widetilde{\Psi}_N^w]$ and $\overline{\Psi} = [\mathbb{1}_{n_0+1}, -\widetilde{\Psi}]$. Note that $\overline{\Psi}$ is an $(n_0+1) \times (n_0+1)$ matrix.

Lemma 8.1. *The following assertions hold.*

(i) *Under Assumption* (A8.1), $\overline{\Psi}$ *is full rank.*

(ii) *Conversely, if* u_k *is* (n_0+1)*-periodic but not full rank, and* $\sum_{j=1}^{n_0+1} u_j \neq 0$*, then* $\overline{\Psi}$ *is not full rank.*

Proof.

(i) $\widetilde{\Psi}$ is the first n_0 columns of the $(n_0+1) \times (n_0+1)$ circulant matrix $T = T([u_{n_0+1}, \dots, u_1])$. Since $\{u_j, j = 1, \dots, n_0+1\}$ is full rank, T is full rank and $\sum_{j=1}^{n_0+1} u_j \neq 0$. Adding the first n_0 columns to the last column, transferring the last column to be the first one, and dividing the first column by $\sum_{j=1}^{n_0+1} u_j$ result in

$$\begin{aligned} T \sim & \begin{bmatrix} u_{n_0+1} & u_{n_0} & \dots & u_2 & \sum_{j=1}^{n_0+1} u_j \\ u_1 & u_{n_0+1} & \dots & u_3 & \sum_{j=1}^{n_0+1} u_j \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ u_{n_0} & u_{n_0-1} & \dots & u_1 & \sum_{j=1}^{n_0+1} u_j \end{bmatrix} \\ \sim & \begin{bmatrix} \sum_{j=1}^{n_0+1} u_j & u_{n_0+1} & u_{n_0} & \dots & u_2 \\ \sum_{j=1}^{n_0+1} u_j & u_1 & u_{n_0+1} & \ddots & u_3 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \sum_{j=1}^{n_0+1} u_j & u_{n_0} & u_{n_0-1} & \dots & u_1 \end{bmatrix} \sim \overline{\Psi}. \end{aligned} \tag{8.1}$$

This implies that $\overline{\Psi}$ is full rank.

(ii) Conversely, if u_k is not full rank, T is not full rank. Since $\sum_{j=1}^{n_0+1} u_j \neq 0$, (8.1) is valid. It follows that $\overline{\Psi}$ is not full rank.

□

By Lemma 8.1, under Assumption (A8.1), $\overline{\Psi}$ is invertible. Define an augmented parameter vector $\Theta = [C, \theta']'$. Let $\Theta_N = (\overline{\Psi}_N^w)^{-1} G(\widetilde{\xi}_N)$, where $G(x) = F^{-1}(x)$.

Theorem 8.2. *The following assertions hold.*

(i) *Suppose that $\{d_k\}$ is a sequence of i.i.d. random variables whose distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are twice continuously differentiable and known, and that* (A8.1) *holds. Then $\Theta_N \to \Theta$ w.p.1 as $N \to \infty$. This implies that $u = \{u_k\}$ is sufficiently rich.*

(ii) *Conversely, if u_k is (n_0+1)-periodic but not full rank, and $\sum_{j=1}^{n_0+1} u_j \neq 0$, then u_k is not sufficiently rich.*

Proof.

(i) Recall that

$$\overline{\Psi}_N^w = [\mathbb{1}_{n_0+1}, -\widetilde{\Psi}_N^w].$$

Under Assumption (A8.1), $\overline{\Psi}_N^w \to \overline{\Psi}$ w.p.1. Under Assumption (A3.1), by the strong law of large numbers,

$$\widetilde{\xi}_N \to \widetilde{\xi} = F(\overline{\Psi}\Theta) \quad \text{w.p.1 as} \quad N \to \infty.$$

This implies, by the continuity of $F^{-1}(\cdot)$,

$$G(\widetilde{\xi}_N) \to \overline{\Psi}\Theta \quad \text{w.p.1 as} \quad N \to \infty.$$

As a result, by Lemma 8.1, $\Theta_N = (\overline{\Psi}_N^w)^{-1} G(\widetilde{\xi}_N) \to \Theta$ w.p.1 as $N \to \infty$.

(ii) Under the hypothesis, by Lemma 8.1, $\overline{\Psi}$ is not full rank. Hence, there exists $\delta \neq 0$ such that $\overline{\Psi}\delta = 0$. Suppose C_1 and θ_1 are true parameters, and $[C_2, \theta_2']' = [C_1, \theta_1']' + \delta$. Then

$$y_k(\theta_1) = \phi_k'\theta_1 + d_k \leq C_1$$

if and only if

$$y_k(\theta_2) = \phi_k'\theta_2 + d_k \leq C_2, \ \forall k.$$

It follows that the output sequences satisfy $s_k(C_1, \theta_1) = s_k(C_2, \theta_2)$. In other words, u_k is information insufficient, which implies that u_k is not sufficiently rich.

□

8.1.2 Recursive Algorithms

A causal and recursive algorithm for computing Θ_N can be constructed as follows.

1. Initialization: $\widetilde{\xi}_1 = \widetilde{S}_1$, $\widetilde{\Psi}_1 = \widetilde{\Phi}_1^w$, $\Theta_1 = 0$.

2. Recursion: Suppose that at N, $\widetilde{\xi}_N$, $\widetilde{\Psi}_N^w$, and $\Theta_N = [C_N, \theta_N']'$ have been obtained. Then at $N+1$, we update

$$\begin{aligned}
\widetilde{\xi}_{N+1} &= \widetilde{\xi}_N - \frac{1}{N+1}\widetilde{\xi}_N + \frac{1}{N+1}\widetilde{S}_{N+1}, \\
\widetilde{\Psi}_{N+1}^w &= \widetilde{\Psi}_N^w - \frac{1}{N+1}\widetilde{\Psi}_N^w + \frac{1}{N+1}\widetilde{\Phi}_{N+1}^w, \\
\overline{\Psi}_{N+1}^w &= [\mathbb{1}_{n_0+1}, -\widetilde{\Psi}_{N+1}^w], \\
\Theta_{N+1} &= \begin{cases} \Theta_N, & \text{if } \overline{\Psi}_{N+1}^w \text{ is singular,} \\ (\overline{\Psi}_{N+1}^w)^{-1} G(\widetilde{\xi}_{N+1}), & \text{otherwise.} \end{cases}
\end{aligned}$$

   The following theorem claims convergence of Θ_N, whose proof is similar to that of Theorem 7.8 and is omitted.

Theorem 8.3. *Under the assumptions of Theorem* 8.2, $\Theta_N \to \Theta$ *w.p.1 as* $N \to \infty$.

8.2 Parameterized Distribution Functions

To estimate the distribution function $\eta = F(\lambda)$, one needs interpolation data in the form of $\eta^{\{i\}} = F(\lambda^{\{i\}})$, $i = 1, 2, \ldots, l$. When $F(\cdot)$ is not parameterized, the estimation of F can become sufficiently accurate only if the data points $\{\lambda^{\{i\}}\}$ are sufficiently dense, rendering an estimation problem of high complexity. Consequently, we adopt a parameterization approach for treating $F(\cdot)$.

Our approach involves three key ideas.

(a) $F(\cdot)$ is approximately parameterized by a model with unknown parameter vector α.

(b) We have shown that the empirical measure $\xi_N^{\{j\}}$ is an approximation of $F(C - \gamma^{\{j\}})$, where $\gamma^{\{j\}}$ is to be estimated as well. Since the underlying system is linear, when the input u_k is scaled to ρu_k and the threshold C is shifted to C_i, we shift the data point from $F(C - \gamma^{\{j\}})$ to $F(C_i - \rho\gamma^{\{j\}})$. This allows us to generate more data points for the estimation of F.

(c) Since $\gamma^{\{j\}}$ is also unknown, we jointly estimate $\gamma^{\{j\}}$ and α. As a result, we can simultaneously estimate $\gamma^{\{j\}}$ for system identification and α for distribution functions.

Suppose that the unknown noise distribution function is $F(\cdot)$, which is approximated by a parameterized model $\widehat{F}(x,\alpha)$, where $\alpha = [\alpha_1, \ldots, \alpha_L]'$ is the unknown model parameter vector of size L. For a given class $\mathbb{F}$ of possible distribution functions, the representation error of $F(x) \in \mathbb{F}$ by $\widehat{F}(x,\alpha)$ is

$$\varepsilon = \sup_{F\in\mathbb{F}} \inf_{\alpha} \sup_{x\in\mathcal{X}} |F(x) - \widehat{F}(x,\alpha)|, \tag{8.2}$$

where $\mathcal{X}$ is a domain for function approximation. For a given $F \in \mathbb{F}$, if the corresponding minimizer of (8.2) is α, then

$$F(x) = \widehat{F}(x,\alpha) + \Delta(x) \tag{8.3}$$

with $|\Delta(x)| \leq \varepsilon, \forall x \in \mathcal{X}$. When α is estimated from the data, its estimate $\widehat{\alpha}$ induces an estimated distribution function $\widehat{F}(x,\widehat{\alpha})$. The overall representation error becomes

$$F(x) - \widehat{F}(x,\widehat{\alpha}) = \widehat{F}(x,\alpha) - \widehat{F}(x,\widehat{\alpha}) + \Delta(x).$$

When a class $\mathbb{F}$ of distribution functions is given, explicit structures of the parameterization may become apparent. As an explanation, we note the following two cases.

1. If $\mathbb{F}$ is the class of normal distributions with unknown mean μ and variance σ^2, then $F(x) = F_0((x-\mu)/\sigma)$, where $F_0(x)$ is the standard normal distribution of $\mu = 0$ and $\sigma^2 = 1$. In this case, $\mathbb{F}$ can be parameterized by $\alpha = [\mu, \sigma^2]'$ with $\varepsilon = 0$.

2. Suppose F is a uniform distribution of a fixed but unknown interval $\mathcal{X} = [a,b]$. For $x \in \mathcal{X}$, $F(x)$ is completely parameterized by $F(x) = \alpha_1 + \alpha_2 x$ with $\varepsilon = 0$. On the other hand, if the uniform distribution is known to have zero mean, then $\mathcal{X} = [-\delta, \delta]$ and

$$F(x) = \frac{x}{2\delta} + \frac{1}{2}, \quad x \in \mathcal{X}.$$

In these examples, the parameterization $\widehat{F}(x,\alpha)$ comes naturally and represents $F(x)$ precisely for all x. However, in general, one may need to use more generic structures of parameterization. For example, for computational convenience, it is common to use a set of L base functions $b_j(\cdot)$, $j = 1, \ldots, L$, to represent $F(\cdot)$. Then

$$\widehat{F}(x,\alpha) = \sum_{j=1}^{L} \alpha_j b_j(x) = b'(x)\alpha, \tag{8.4}$$

where $b(x) = [b_1(x), \ldots, b_L(x)]'$.

It is noted that some routine modifications to (8.4) may be needed. For example, suppose polynomials of x are used as base functions. If $F(x)$ is

a normal distribution, then for any finite L, $F(x)$ cannot be well approximated by $\widehat{F}(x,\alpha)$ over all $x \in \mathbb{R}$. In this case, one may limit (8.4) to a finite interval $[a,b]$ and modify $\widehat{F}(x,\alpha)$ for $x \notin [a,b]$ so that $\widehat{F}(x,\alpha)$ decreases toward 0 for $x \to -\infty$ and $\widehat{F}(x,\alpha)$ increases toward 1 when $x \to \infty$. Since these techniques are standard in function approximations, they will not be discussed further.

8.3 Joint Identification Problems

The main idea of our approach is to explore input scaling, possibly together with threshold shifting, to provide joint information on the unknown distribution function and system parameters. Due to parameterization of the uncertainty set $\mathbb{F}$, the identification of F is reduced to parameter estimation of α.

To be more specific, the n_0-periodic full-rank input u employed in the previous chapters for parameter identification can be expanded by scaling. Let ρ_i, $i = 1, \ldots, \kappa$, be κ nonzero scaling factors. Define $u^{\{i\}} = \rho_i u$. Note that by linearity of the system, when the input is $u^{\{i\}}$, for a given $j = 1, \ldots, n_0$, the corresponding output becomes

$$y^{\{i\}}_{ln_0+j} = \rho_i \gamma^{\{j\}} + d_{j+ln_0}, \quad l = 0, 1, \ldots, N-1.$$

In addition, the threshold C may also be shifted to C_i.

Under the periodic signal u, scaling factors ρ_i, and thresholds C_i, let the corresponding sequences of the sensor output be $\{s^{\{i\}}_k\}$. Now, the empirical measures

$$\begin{aligned} \xi^{\{i,j\}}_N &= \frac{1}{N}\sum_{l=0}^{N-1} s^{\{i\}}_{ln_0+j} \to \xi^{\{i,j\}} \quad \text{w.p.1.} \\ &= F(C_i - \rho_i \gamma^{\{j\}}). \end{aligned} \tag{8.5}$$

The limit of the empirical measures satisfies, for a given $j = 1, \ldots, n_0$,

$$\xi^{\{i,j\}} = \widehat{F}(C_i - \rho_i\gamma^{\{j\}}, \alpha) + [F(C_i - \rho_i\gamma^{\{j\}}) - \widehat{F}(C_i - \rho_i\gamma^{\{j\}}, \alpha)],\ i = 1, \ldots, \kappa. \tag{8.6}$$

Our task is to calculate $\Gamma = [\gamma^{\{1\}}, \ldots, \gamma^{\{n_0\}}]'$ and α. When u_k is n_0-periodic and full rank, θ can be identified from γ. As a result, joint identification of α and θ is reduced to joint identification of α and Γ.

8.4 Richness Conditions for Joint Identification

An essential property for identifying α and Γ is that the systems of equations (8.6) have a unique solution. It is noted that for a given j, the κ

equations

$$\widetilde{\xi}^{\{i,j\}} = F(C_i - \rho_i \gamma^{\{j\}}, \alpha),\ i = 1, \ldots, \kappa, \tag{8.7}$$

contain $L+1$ unknowns: $\gamma^{\{j\}}$ and α. Hence, we should take $\kappa \geq L+1$. If C_i and ρ_i are selected such that (8.7) has a unique solution α and γ, then by repeating the procedure for $\gamma = \gamma^{\{j\}}$, $j = 1, \ldots, n_0$, (8.6) will have a unique solution α and Γ. Denote $\Lambda = \{(C_i, \rho_i), i = 1, \ldots, \kappa\}$. Suppose the prior information on α and γ is that $[\alpha', \gamma]' \in \Omega \subseteq \mathbb{R}^{L+1}$.

Definition 8.4. Given a parameterization $\widehat{F}(x, \alpha)$, a set of pairs $\Lambda = \{(C_i, \rho_i), i = 1, \ldots, \kappa\}$ is said to be *sufficiently rich* for joint identification of α and Γ if, under Λ, (8.7) has a unique solution α and γ in Ω.

Remark 8.5. A sufficient condition for Λ to be sufficiently rich is that the $\kappa \times (L+1)$ Jacobian matrix

$$J = \begin{bmatrix} \frac{\partial \widehat{F}(C_1 - \rho_1 \gamma, \alpha)}{\partial \alpha} & -\rho_1 \frac{\partial \widehat{F}(C_1 - \rho_1 \gamma, \alpha)}{\partial (C_1 - \rho_1 \gamma)} \\ \vdots & \vdots \\ \frac{\partial \widehat{F}(C_\kappa - \rho_\kappa \gamma, \alpha)}{\partial \alpha} & -\rho_\kappa \frac{\partial \widehat{F}(C_\kappa - \rho_\kappa \gamma, \alpha)}{\partial (C_\kappa - \rho_\kappa \gamma)} \end{bmatrix}$$

is full rank for all $[\alpha', \gamma']' \in \Omega$.

Example 8.6. Suppose F is a normal distribution function with unknown μ and σ, $F(x) = F_0((x-\mu)/\sigma)$, where F_0 is the normal distribution of $\mu = 0$ and $\sigma = 1$. Then for $\kappa = L + 1 = 3$, (8.7) becomes

$$\xi^{\{i\}} = F_0((C_i - \rho_i \gamma - \mu)/\sigma),\ i = 1, 2, 3.$$

Define $x_i = F_0^{-1}(\xi^{\{i\}})$, $i = 1, 2, 3$. Then we have

$$x_i = \frac{C_i - \rho_i \gamma - \mu}{\sigma},\ i = 1, 2, 3,$$

or

$$\begin{bmatrix} C_1 & -\rho_1 & -1 \\ C_2 & -\rho_2 & -1 \\ C_3 & -\rho_3 & -1 \end{bmatrix} \begin{bmatrix} 1/\sigma \\ \gamma/\sigma \\ \mu/\sigma \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

In this case, (8.7) has a unique solution if the matrix

$$M = \begin{bmatrix} C_1 & -\rho_1 & -1 \\ C_2 & -\rho_2 & -1 \\ C_3 & -\rho_3 & -1 \end{bmatrix}$$

is full rank. For example, if $C_1 = 1$, $C_2 = 2$, $C_3 = 4$, $\rho_1 = 1$, $\rho_2 = 3$, and $\rho_3 = 5$, then it can be calculated that

$$M = \begin{bmatrix} 1 & -1 & -1 \\ 2 & -3 & -1 \\ 4 & -5 & -1 \end{bmatrix},$$

which is full rank for any μ, σ, γ.

In this example, it is easy to verify that shifting the threshold is necessary for M to be full rank. Indeed, if $C_1 = C_2 = C_3 = C$, then M is not full rank. In fact, the expression

$$\frac{C_i - \rho_i\gamma - \mu}{\sigma} = \frac{C - \mu}{\sigma} - \rho_i\frac{\gamma}{\sigma}$$

cannot be used to determine the three parameters μ, σ, γ.

On the other hand, if it is known that the noise is zero mean, namely, $\mu = 0$, then one may use a fixed threshold. In this case, we have

$$x_1 = \frac{C}{\sigma} - \rho_1\frac{\gamma}{\sigma}, \quad x_2 = \frac{C}{\sigma} - \rho_2\frac{\gamma}{\sigma}.$$

γ and σ can be solved uniquely if $\rho_1 \neq \rho_2$.

8.5 Algorithms for Identifying System Parameters and Distribution Functions

Note that the event $\{y^{\{i\}}_{j+ln_0} \leq C_i\}$ is the same as $\{d_{j+ln_0} \leq C_i - \rho_i\gamma^{\{j\}}\}$. Then $\xi^{\{i,j\}}_N$ is precisely the value of the N-sample empirical distribution $F_N(x)$ of the noise d at $x = C_i - \rho_i\gamma^{\{j\}}$: $\xi^{\{i,j\}}_N = F_N(C_i - \rho_i\gamma^{\{j\}})$. Consequently, in consideration of parameterized models of F, over κ input sequences, we obtain the following κ-sample values of $\widehat{F}(x, \alpha)$ at

$$\xi^{\{i,j\}}_N = \widehat{F}(C_i - \rho_i\gamma^{\{j\}}, \alpha) + e^{\{i,j\}}_N + \Delta, \quad i = 1, \ldots, \kappa,$$

where $e^{\{i,j\}}_N$ is the identification error and Δ is the representation error in (8.3).

For a fixed j, let $\gamma = \gamma^{\{j\}}$. Consider

$$\xi^{\{i,j\}}_N = \widehat{F}(C_i - \rho_i\gamma, \alpha) + e^{\{i,j\}}_N + \Delta, \quad i = 1, \ldots, \kappa.$$

In general, parameterization $\widehat{F}(x, \alpha)$ is a nonlinear mapping with respect to α. Consequently, the nonlinear equations $\xi^{\{i,j\}}_N = \widehat{F}(C_i - \rho_i\gamma^{\{j\}}, \widehat{\alpha})$

will be used to derive an estimate $\widehat{\alpha}$. For simplicity of discussions, we will present our algorithms for linear parameterization since it renders a simpler sequential procedure.

By the linear representation of $F(x)$ in (8.4), we have that for $i = 1, \ldots, \kappa$,

$$\xi_N^{\{i,j\}} = b'(C_i - \rho_i\gamma)\alpha + \Delta(C_i - \rho_i\gamma) + e_N^{\{i,j\}}.$$

By defining

$$\begin{aligned} \Xi_N &= [\xi_N(1,j), \ldots, \xi_N(\kappa,j)]', \\ B(\gamma) &= [b(C_1 - \rho_1\gamma), \ldots, b(C_\kappa - \rho_\kappa\gamma)]', \\ \Delta &= [\Delta(C_1 - \rho_1\gamma), \ldots, \Delta(C_\kappa - \rho_\kappa\gamma)]', \\ \widetilde{E}_N &= [e_N(1,j), \ldots, e_N(\kappa,j)]', \end{aligned}$$

we obtain the relationship

$$\Xi_N = B(\gamma)\alpha + \Delta + \widetilde{E}_N. \tag{8.8}$$

The goal here is to select α and γ to minimize $\|\Xi_N(\gamma,\alpha) - B(\gamma)\alpha\|_2^2$, where $\|\cdot\|_2$ is the Euclidean norm. The following joint identification algorithm is introduced.

We shall write (8.8) as

$$\Xi = B(\gamma)\alpha + \Delta + \widetilde{E}.$$

For any given γ, if the corresponding $B(\gamma)$ is full rank, the optimal least-squares estimation error for α is

$$V(\gamma) = \|(I - B(\gamma)(B'(\gamma)B(\gamma))^{-1}B'(\gamma))\Xi\|_2^2.$$

Then the following optimal line search optimization is conducted:

$$\min_{\gamma} V(\gamma). \tag{8.9}$$

Denote the optimal solution by $\widehat{\gamma}$. Then

$$\widehat{\alpha} = (B'(\widehat{\gamma})B(\widehat{\gamma}))^{-1}B'(\widehat{\gamma})\Xi. \tag{8.10}$$

This algorithm is based on the consecutive marginal optimization

$$\inf_{\gamma}\left(\inf_{\alpha}\|\Xi - B(\gamma)\alpha\|_2^2\right) = \inf_{\gamma} V(\gamma). \tag{8.11}$$

Observe that, in general, the joint identification

$$\inf_{\alpha,\gamma}\|\Xi - B(\gamma)\alpha\|_2^2 \tag{8.12}$$

is a nonlinear optimization problem, which bears a higher computational complexity. Although for a finite observation, the consecutive optimization (8.11) may not be equivalent to the estimates from the joint optimization (8.12), convergence results on $\widehat{\gamma}$ and $\widehat{\alpha}$ can be established. Note that for algorithm execution, (8.11) will be repeated for $\gamma = \gamma^{\{j\}}$, $j = 1, \ldots, n_0$.

8.6 Convergence Analysis

We now derive convergence properties of $\widehat{\gamma}$ and $\widehat{\alpha}$.

(A8.2) $\{d_k\}$ is a sequence of i.i.d. random variables whose distribution function $F(\cdot)$ and inverse $F^{-1}(\cdot)$ are twice continuously differentiable. $F(\cdot)$ is unknown but belongs to a class $\mathbb{F}$.

Theorem 8.7. *Suppose that* $\Lambda = \{(C_i, \rho_i), i = 1, \ldots, \kappa\}$ *is sufficiently rich. Then under the representation error bound* (8.2)*, there is an* $\widehat{\alpha}_N$ *such that for any compact subset* $\mathcal{X} \subset \mathbb{R}$*,*

$$\limsup_{N\to\infty} \sup_{x\in\mathcal{X}} |\widehat{F}(x, \widehat{\alpha}_N) - F(x)| \le c\varepsilon \quad w.p.1,$$

for some constant $c \ge 0$*, where* ε *is defined in* (8.2)*.*

Proof. From (8.3) and (8.4), we have that for $i = 1, \ldots, \kappa$,

$$F(C_i - \rho_i\gamma) = b'(C_i - \rho_i\gamma)\alpha + \Delta(C_i - \rho_i\gamma),$$

in a vector form

$$\Xi = B\alpha + \Delta.$$

Since Λ is sufficiently rich, $B'B$ is invertible. Hence, by the least-squares method, we obtain

$$\widehat{\alpha}_N = (B'B)^{-1}B'\Xi$$

and

$$\widehat{\alpha}_N - \alpha = (B'B)^{-1}B'\Delta,$$

which implies

$$\|\widehat{\alpha}_N - \alpha\|_2 \le \beta_1\varepsilon,$$

for some constant $\beta_1 > 0$. Consequently, for some $\beta_2 > 0$,

$$|\widehat{F}(x, \widehat{\alpha}_N) - F(x)| \le |b'(x)(\alpha - \widehat{\alpha}_N) + \Delta(x)| \le \beta_2\beta_1\varepsilon + \varepsilon = \beta\varepsilon. \qquad \square$$

Theorem 8.8. *Under Assumption* (A8.2) *and the representation of error bound* (8.2)*, if* (8.11) *holds, then the* $\widehat{\gamma}_N$ *obtained by solving* (8.9) *satisfies*

$$\limsup_{N\to\infty} |\widehat{\gamma}_N - \gamma| \le \beta_0\varepsilon \quad w.p.1,$$

for some constant $\beta_0 > 0$*.*

Proof. From (8.3) and (8.4), we have

$$\begin{aligned}
&F(C_i - \rho_i\gamma) - F(C_i - \rho_i\widehat{\gamma}_N) \\
&= [F(C_i - \rho_i\gamma) - F_N(C_i - \rho_i\gamma)] + [F_N(C_i - \rho_i\gamma) - b'(C_i - \rho_i\widehat{\gamma}_N)\alpha] \\
&\quad + [b'(C_i - \rho_i\widehat{\gamma}_N)\alpha - F(C_i - \rho_i\widehat{\gamma}_N)] \\
&= [F(C_i - \rho_i\gamma) - F_N(C_i - \rho_i\gamma)] + [F_N(C_i - \rho_i\gamma) - b'(C_i - \rho_i\widehat{\gamma}_N)\alpha] \\
&\quad - \Delta(C_i - \rho_i\widehat{\gamma}_N),
\end{aligned}$$

which leads to

$$\begin{aligned} &|F(C_i - \rho_i\gamma) - F(C_i - \rho_i\widehat{\gamma}_N)| \\ &\le |F(C_i - \rho_i\gamma) - F_N(C_i - \rho_i\gamma)| \\ &\quad + |F_N(C_i - \rho_i\gamma) - b'(C_i - \rho_i\widehat{\gamma}_N)\alpha| + \varepsilon. \end{aligned} \tag{8.13}$$

Note that by (8.11),

$$\begin{aligned} &|F_N(C_i - \rho_i\gamma) - b'(C_i - \rho_i\widehat{\gamma}_N)\alpha| \\ &\quad \le \|\Xi - B(C_i - \rho_i\widehat{\gamma}_N)\alpha\|_2 \\ &\quad \le \|\Xi - B(C_i - \rho_i\gamma)\alpha\|_2 = \|\Delta\|_2. \end{aligned}$$

Then from

$$\lim_{N\to\infty} |F(C_i - \rho_i\gamma) - F_N(C_i - \rho_i\gamma)| = 0$$

and (8.13), we have

$$|F(C_i - \rho_i\gamma) - F(C_i - \rho_i\widehat{\gamma}_N)| \le \beta_1\varepsilon$$

for some constant β_1. Thus, by the differentiability of $F^{-1}(\cdot)$, we conclude that

$$|(C_i - \rho_i\widehat{\gamma}_N) - (C_i - \rho_i\gamma)| \le \beta_1|\dot{F}^{-1}(\xi)|\varepsilon$$

for some value

$$\xi \in [\min\{F(C_i - \rho_i\gamma), F(C_i - \rho_i\widehat{\gamma}_N)\},\ \max\{F(C_i - \rho_i\gamma), F(C_i - \rho_i\widehat{\gamma}_N)\}].$$

This together with the continuity of $\dot{F}^{-1}(\cdot)$ implies that

$$\limsup_{N\to\infty} |\widehat{\gamma}_N - \gamma| \le \beta_0\varepsilon,$$

as stated. □

In particular, if F is well represented by the parameterized model, i.e., $\varepsilon \to 0$, then $\widehat{\gamma}_N \to \gamma$ w.p.1 as $N \to \infty$.

8.7 Recursive Algorithms

In this section, we develop a class of recursive algorithms for estimating α and γ. In lieu of the line search (8.9) and least squares procedure (8.10), the estimate $\widehat{\alpha}_N$ will be constructed via an adaptive filtering algorithm to reduce the computational complexity, and the estimate $\widehat{\gamma}_N$ will be recursified. This section is divided into three parts: First, we present the algorithms. Then, we establish the convergence of the schemes. Finally, we make some additional remarks on alternatives.

The identification problem involves several indices which can be confusing in our recursive algorithms: (a) the time index k. (b) The time-block index N. Iteration from N to $N+1$ represents an acquisition of n_0 observation points on s_k. (c) The cyclic index $j = 1, \ldots, n_0$. This index indicates rotation of parameters $\gamma^{\{j\}}$ in identification, in other words, indicating one of the sequential optimization problems. (d) The index i in ρ_i, $i = 1, \ldots, \kappa$. This represents the ith scaling factor ρ_i is applied at input.

For example, due to the cyclic nature, $\widehat{\gamma}^{\{j\}}$ can only be updated once every n_0 data points. As a result, it is indexed as $\widehat{\gamma}_N^{\{j\}}$. On the other hand, all data points contain information on α. Hence, it can be indexed as $\widehat{\alpha}_k$. In case we choose to update $\widehat{\alpha}$ at the same time as updating $\gamma^{\{j\}}$, we shall use $\widehat{\alpha}_N$ instead.

8.7.1 Recursive Schemes

The following two typical classes of recursive algorithms will be considered.

(A) Adaptive Filtering Algorithms

For each $i = 1, \ldots, \kappa$ of scaling values at the input, and $j = 1, \ldots, n_0$,

$$\begin{cases} \xi_{N+1}^{\{i,j\}} = \xi_N^{\{i,j\}} - \dfrac{1}{N+1}[\xi_N^{\{i,j\}} - s_{j+(N+1)n_0}^{\{i\}}], \\ \widehat{\alpha}_{N+1}^{\{i,j\}} = \widehat{\alpha}_N^{\{i,j\}} + \dfrac{1}{N+1} b_N [\xi_N^{\{i,j\}} - b_N' \widehat{\alpha}_N^{\{i,j\}}], \\ \widehat{\gamma}_{N+1}^{\{i,j\}} = \widehat{\gamma}_N^{\{i,j\}} + \dfrac{1}{N+1}\left[\widehat{\gamma}_N^{\{i,j\}} - \dfrac{C_i - \widehat{F}^{-1}(\xi_N^{\{i,j\}}, \widehat{\alpha}_N^{\{i,j\}})}{\rho_i}\right], \end{cases} \tag{8.14}$$

where

$$b_N = b(C_i - \rho_i \widehat{\gamma}_N^{\{i,j\}}),$$

with $b(\cdot)$ given in (8.4), and $\widehat{F}^{-1}(z, \widehat{\alpha})$ denotes the inverse of $\widehat{F}(z, \alpha)$ when α is fixed. Note that, in fact, b_N is j-dependent, so it should have been written as $b_N^{\{j\}}$. We have suppressed j-dependence for notational simplicity.

(B) Combined Adaptive Filtering and Least-Squares Algorithm

For each $i = 1, \ldots, \kappa$ of scaling values at the input, $j = 1, \ldots, n_0$,

$$\begin{cases} \xi_{N+1}^{\{i,j\}} = \xi_N^{\{i,j\}} - \dfrac{1}{N+1}[\xi_N^{\{i,j\}} - s_{j+(N+1)n_0}^{\{i\}}], \\ \widehat{\alpha}_{N+1}^{\{i,j\}} = \widehat{\alpha}_N^{\{i,j\}} + a_N \Psi_N b_N [\xi_N^{\{i,j\}} - b_N' \widehat{\alpha}_N^{\{i,j\}}], \\ \Psi_{N+1} = \Psi_N - a_N \Psi_N b_N b_N^T \Psi_N, \\ a_N = (1 + b_N' \Psi_N b_N)^{-1}, \\ \widehat{\gamma}_{N+1}^{\{i,j\}} = \widehat{\gamma}_N^{\{i,j\}} + \dfrac{1}{N+1}\left[\widehat{\gamma}_N^{\{i,j\}} - \dfrac{C_i - \widehat{F}^{-1}(\xi_N^{\{i,j\}}, \widehat{\alpha}_N^{\{i,j\}})}{\rho_i}\right]. \end{cases} \tag{8.15}$$

8.7.2 Asymptotic Properties of Recursive Algorithm (8.14)

In what follows, we present asymptotic properties of the algorithms given in (8.14) and (8.15). To proceed, we need some conditions, which are listed below.

(A8.3) The following system of differential equations

$$\begin{cases} \dfrac{d}{dt}\alpha^{\{i,j\}}(t) &= b(C_i - \gamma^{\{i,j\}}(t))F(C_i - \gamma^{\{i,j\}}) \\ &\quad -b(C_i - \gamma^{\{i,j\}}(t))b'(C_i - \gamma^{\{i,j\}}(t))\alpha^{\{i,j\}}(t), \\ \dfrac{d}{dt}\gamma^{\{i,j\}}(t) &= \gamma^{\{i,j\}}(t) \\ &\quad -\dfrac{C_i - \widehat{F}^{-1}(F(C_i - \gamma^{\{i,j\}}\rho_i), \alpha^{\{i,j\}}(t))}{\rho_i}, \end{cases} \tag{8.16}$$

has a unique solution for each initial condition. In addition, (8.16) has a unique asymptotically stable point $(\alpha^{\{i,j\},0}, \gamma^{\{i,j\},0})$ in the sense of Lyapunov.

(A8.4) The following conditions hold.

- The sequences $\{\widehat{\gamma}_N^{\{i,j\}}\}$ for $j = 1, \ldots, n_0$ are bounded w.p.1.
- Denoting $A^j = b(C_i - \rho_i\gamma^{\{i,j\},0})b'(C_i - \rho_i\gamma^{\{i,j\},0})$, A^j is symmetric and positive definite.
- Both $b(\cdot)$ and $\widehat{F}^{-1}(\cdot,\cdot)$ are continuous.

Remark 8.9. To ensure the boundedness, we can use a projection algorithm

$$\widehat{\gamma}_{N+1}^{\{i,j\}} = \Pi_G\left[\widehat{\gamma}_N^{\{i,j\}} + \frac{\widehat{\gamma}_N^{\{i,j\}} - \frac{[C_i - \widehat{F}^{-1}(\xi_N^{\{i,j\}}, \widehat{\alpha}_N^{\{i,j\}})]}{\rho_i}}{N+1}\right],$$

where Π_G is the projection operator onto the bounded set G (see [55] for more details). Owing to the use of $\{s_N^{\{i\}}\}$, $\{\xi_N^{\{i,j\}}\}$ is bounded. Note that we can choose G to be as simple as a box, and choose it to be large enough so that it contains the true parameter $\gamma^{\{j\}}$. However, to simplify notation and for convenience, we have assumed that the boundedness of $\{\widehat{\gamma}_N^{\{j\}}\}$ in Assumption (A8.4). The continuity of $b(\cdot)$ implies that $\widehat{F}(\cdot,\cdot)$ is also continuous since it is linear in α. We require that the matrix A^j be positive definite, which is essentially a solvability or identifiability condition.

We claim that $\{\widehat{\alpha}_N^{\{i,j\}}\}$ is bounded w.p.1 uniformly in N. To see this, write

$$\begin{aligned}\widehat{\alpha}_{N+1}^{\{i,j\}} = A_{N,0}\widehat{\alpha}_0^{\{i,j\}} &+ \sum_{l=0}^{N} \frac{1}{l+1} A_{N,l}[A^j - b_l b_l^T]\widehat{\alpha}_l^{\{i,j\}} \\ &+ \sum_{l=0}^{N} \frac{1}{l+1} A_{N,l} b_l \xi_l^{\{j\}},\end{aligned} \tag{8.17}$$

where

$$A_{N,l} = \begin{cases} \displaystyle\prod_{i=l+1}^{N} \Big(I - \frac{1}{i+1}A^j\Big), & l < N, \\ I, & l = N. \end{cases}$$

Note that following the convention for b_N, we suppressed the j-dependence in the notation $A_{N,l}$. Thus, taking the norm in (8.17), an application of the Gronwall's inequality yields that

$$|\widehat{\alpha}_{N+1}^{\{i,j\}}| \le K_{1,N} \exp(K_{2,N}), \tag{8.18}$$

where

$$\begin{aligned} K_{1,N} &= |A_{N,0}\widehat{\alpha}_0^{\{i,j\}}| + \sum_{l=0}^{N} \frac{1}{l+1}|A_{N,l}||b_l \xi_l^{\{i,j\}}|, \\ K_{2,N} &= \sum_{l=0}^{N} \frac{1}{l+1}|A_{N,l}||A^j - b_l b_l'|, \end{aligned}$$

with

$$a_{N,l} = \begin{cases} \displaystyle\prod_{i=l+1}^{N} \Big(1 - \frac{1}{i+1}\lambda\Big), & l < N, \\ 1, & l = N, \end{cases}$$

where λ is the minimal eigenvalues of A^j,

$$\begin{aligned}\sum_{l=0}^{N} \frac{1}{l+1}|A_{N,l}| &\le \sum_{l=0}^{N} \frac{1}{l+1} a_{N,l} \\ &= \frac{K_0}{\lambda} \sum_{l=0}^{N} [a_{N,l+1} - a_{N,l}] = K_0(1 - a_{N,0}) < \infty.\end{aligned} \tag{8.19}$$

Note that in the above, we used K_0 as a generic positive constant whose value may change for different appearances. The bound in (8.19) together with (8.17), the boundedness of $\{\xi_N^{\{i,j\}}\}$ and $\{\widehat{\gamma}_N^{\{i,j\}}\}$, implies that $K_{1,N}$ is bounded w.p.1 uniformly in N, so is $K_{2,N}$. The w.p.1 boundedness (uniform in N) of $\{\widehat{\alpha}_N^{\{i,j\}}\}$ then follows from (8.18).

Next, consider the joint process $z_N^{\{i,j\}} = (\widehat{\alpha}_N^{\{i,j\}}, \widehat{\gamma}_N^{\{i,j\}})'$. Set

$$t_N = \sum_{l=0}^{N-1} \frac{1}{l+1} \quad \text{and} \quad m(t) = \max\{N : t_N \le t\}.$$

Denote $z_N = (z_N^{\{i,j\}} : \ i = 1, \dots, \kappa, \ j = 1, \dots, n_0)$ and define the piecewise-constant interpolation

$$z^0(t) = z_N \quad \text{for} \quad t \in [t_N, t_{N+1}) \quad \text{and} \quad z^N(t) = z^0(t + t_N).$$

Note that $z^N(\cdot)$ is a shifted sequence for bringing the asymptotic properties of the sequence to the foreground. We also define the component of the interpolation $z^{N,\{i,j\}}(\cdot)$ as $\alpha^{N,\{i,j\}}(\cdot)$ and $\gamma^{N,\{i,j\}}(\cdot)$. The boundedness on $\{\xi_N^{\{i,j\}}, \widehat{\alpha}_N^{\{i,j\}}, \widehat{\gamma}_N^{\{i,j\}}\}$ yields that $z^{N,\{i,j\}}(\cdot)$ is uniformly bounded. The continuity condition in Assumption (A8.4) and the continuity of the distribution function and its inverse imply $z^{N,\{i,j\}}(\cdot)$ is equicontinuous in the extended sense as defined in [55, p. 102]. By the Arzelà–Ascoli theorem [55, p. 102] applied to a sequence of equicontinuous functions (in the extended sense), we can extract a convergent subsequence $z^{N_1,\{i,j\}}(\cdot)$ such that $z^{N_1,\{i,j\}}(\cdot)$ converges to $z^{\{i,j\}}(\cdot)$ w.p.1 and the convergence is uniform on any bounded interval. For convenience, in what follows, we simply write N_1 as N. Using the usual ODE approach (see [55]), we can show that $(\alpha^{N,\{i,j\}}(\cdot), \gamma^{N,\{i,j\}}(\cdot)) \to (\alpha^{\{i,j\}}(\cdot), \gamma^{\{i,j\}}(\cdot))$. Then $\alpha^{\{i,j\}}(\cdot)$ and $\gamma^{\{i,j\}}(\cdot)$ satisfy the differential equation in (8.16).

Next, let $\{\tau_N\}$ be a sequence of positive real numbers satisfying $\tau_N \to \infty$ as $N \to \infty$. Then it can be shown (see [55] for more details) that $(\alpha^{N,\{i,j\}}(\cdot + \tau_N), \gamma^{N,\{i,j\}}(\cdot + \tau_N)) \to (\alpha^{\{i,j\},0}, \gamma^{\{i,j\},0})$ as $N \to \infty$. Thus, $(\widehat{\alpha}_N^{\{i,j\}}, \widehat{\gamma}_N^{\{i,j\}}) \to (\alpha^{\{i,j\},0}, \gamma^{\{i,j\},0})$ w.p.1 as $N \to \infty$.

Note that the stationary point $\gamma^{\{i,j\},0}$ is given by

$$\gamma^{\{i,j\},0} = [C_i - \widehat{F}^{-1}(F(C_i - \gamma^{\{i,j\},0}\rho_i), \alpha^{\{i,j\},0})]/\rho_i.$$

As in the previous section, it can be shown that for some $\beta_0 > 0$,

$$\limsup_{N\to\infty} |C_i - F^{-1}(\xi_N^j) - \gamma^{\{i,j\},0}| \le \beta_0 \varepsilon. \tag{8.20}$$

Summarizing what has been proved, we obtain the following theorem.

Theorem 8.10. *Assume* (A8.1)–(A8.4). *Then,* $\{\xi_N^{\{i,j\}}, \widehat{\alpha}_N^{\{i,j\}}, \widehat{\gamma}_N^{\{i,j\}}\}$, *the sequence of triples converges w.p.1. Moreover, we have the following upper bound on the deviation* $\widehat{\gamma}_N^{\{i,j\}} - \gamma^{\{j\}}$*:*

$$\limsup_{N\to\infty} |\widehat{\gamma}_N^{\{i,j\}} - \gamma^{\{j\}}| \le \beta_0 \varepsilon, \quad \textit{w.p.1 for some} \ \ \beta_0 > 0.$$

8.8 Algorithm Flowcharts

Our algorithms for joint identification of system parameters and noise distributions are summarized in Figure 8.1. We now use an example to demon-

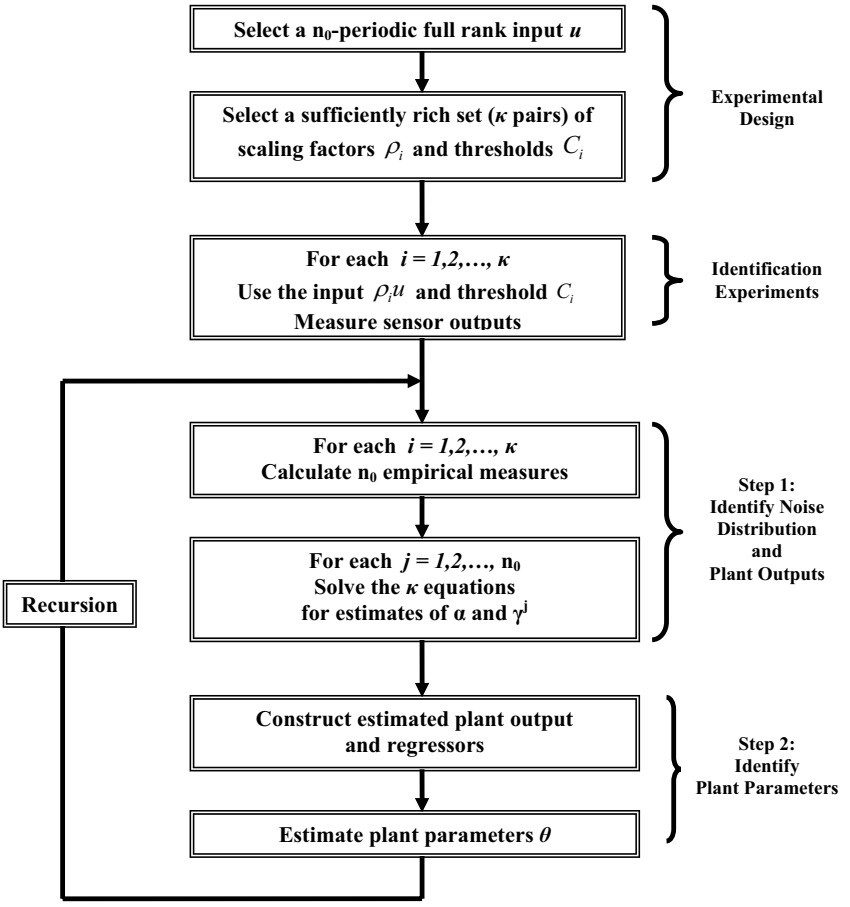


FIGURE 8.1. An algorithm flowchart

strate the identification algorithms presented so far. Suppose that the true plant is a first-order system

$$x_k = -a_0 x_{k-1} + b_0 u_{k-1}, \quad y_k = x_k + d_k,$$

where $a_0 = 0.4$, and $b_0 = 1.6$. $\{d_k\}$ is an i.i.d. sequence, uniformly distributed on $[-1.2, 1.2]$. Hence, the true distribution function is $\xi = F(x) =$

$(1/2.4)x + 0.5$ for $x \in [-1.2, 1.2]$. The true system parameters and the distribution function interval are unknown.

1. Experimental design: Here we need to select the parameterization of the unknown distribution function, input signal, scaling, and threshold selections.

   (a) For this example, we assume the linear function parameterization of F: $\xi = \widehat{F}(x, \alpha) = \alpha_1 + \alpha_2 x$. Since this is a correct parameterization, the function representation error satisfies $\varepsilon = 0$.

   (b) To identify the two system parameters a_0 and b_0, the base input is 2-periodic with $u_1 = 0.7$ and $u_2 = 0.2$, which is full rank.

   (c) Signal scaling factors ρ_i and thresholds C_i are to be selected such that (8.7) can be solved uniquely for α and γ. In this example, we have

   $$\begin{aligned} \xi^{\{1\}} &= \alpha_1 + \alpha_2(C_1 - \rho_1\gamma), \\ \xi^{\{2\}} &= \alpha_1 + \alpha_2(C_2 - \rho_2\gamma), \\ \xi^{\{3\}} &= \alpha_1 + \alpha_2(C_3 - \rho_3\gamma). \end{aligned}$$

   This system has a unique solution if

   $$M = \begin{bmatrix} 1 & C_1 & \rho_1 \\ 1 & C_2 & \rho_2 \\ 1 & C_3 & \rho_3 \end{bmatrix}$$

   is full rank. For example, we use the following three sets of values: $\rho_1 = 0.3, C_1 = -0.4$; $\rho_2 = 0.5, C_2 = 0.4$; $\rho_3 = 0.8, C_3 = 0.8$. This leads to

   $$M = \begin{bmatrix} 1 & -0.4 & 0.3 \\ 1 & 0.4 & 0.5 \\ 1 & 0.8 & 0.8 \end{bmatrix},$$

   which is full rank.

2. Identification: Identify α and θ.

   (a) The system output y_k is simulated by

   $$y_k = -a_0 x_{k-1} + b_0 u_{k-1} + d_k,$$

   for a total of 900 sample steps.

   (b) The sensor outputs are observed and empirical measures are calculated.

(c) The recursive identification algorithms (8.9) and (8.10) are applied to identify the plant parameters and distribution function simultaneously.

3. Evaluation: The plant parameter estimates are compared to the true values $a_0 = 0.4, b_0 = 1.6$, and the distribution function parameters $[\alpha_1, \alpha_2]$ are compared to their true values $[0.5, 1/2.4]$. The results are shown in Figure 8.2, where relative errors are plotted as a function of sample sizes.

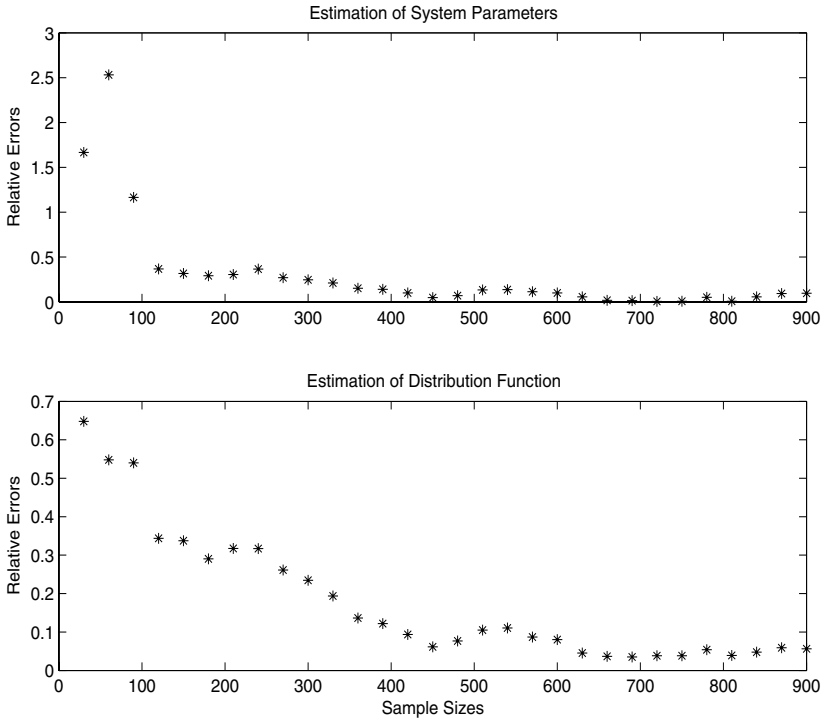


FIGURE 8.2. Joint identification of distribution functions and plant parameters

8.9 Illustrative Examples

In this section, we use two examples to demonstrate the algorithms developed in this chapter. Example 8.11 illustrates the case when the switching threshold is unknown. It shows that when the input is $n_0 + 1$ full rank, both the threshold C and system parameters θ can be estimated simultaneously. Example 8.12 covers the scenario of unknown noise distributions.

The input design and joint identification algorithms are shown to lead to consistent estimates.

Example 8.11. Suppose that the threshold C is unknown and that the input has measurement noise. Consider a third-order system: $y_k = \phi_k'\theta + d_k$, where the output is measured by a binary-valued sensor with unknown threshold C. Suppose the true parameters are $C = 28$ and $\theta = [2.1, 2.7, 3.6]'$, and that $\{d_k\}$ is a sequence of i.i.d. normal variables with mean zero and variance $\sigma_d^2 = 4$. The noise-free input v is 4-periodic with one-period values $(3.1, 4.3, 2.3, 3.5)$, which is full rank. The actual input is $u_k = v_k + \varepsilon_k$, where $\{\varepsilon_k\}$ is a sequence of i.i.d. normal variables with mean zero and variance $\sigma_\varepsilon^2 = 1$.

For $n = 3$, define

$$\begin{aligned}
\widetilde{Y}_j &= [y_{4(j-1)+1}, \ldots, y_{4j}]' \in \mathbb{R}^4, \\
\widetilde{\Phi}_j &= [\phi_{4(j-1)+1}, \ldots, \phi_{4j}]' \in \mathbb{R}^{4\times 3}, \\
\widetilde{D}_j &= [d_{4(j-1)+1}, \ldots, d_{4j}]' \in \mathbb{R}^4, \\
\widetilde{S}_j &= [s_{4(j-1)+1}, \ldots, s_{4j}]' \in \mathbb{R}^4.
\end{aligned}$$

It follows that $\widetilde{Y}_j = \widetilde{\Phi}_j\theta + \widetilde{D}_j$, for $j = 1, 2, \ldots$ Since $\{\widetilde{D}_j\}$ is a sequence of i.i.d. normal variable vectors, we have $\widetilde{\xi}_N = \frac{1}{N}\sum_{j=1}^N \widetilde{S}_j \to F(\overline{\Psi}\Theta)$. Since v is full rank, $\overline{\Psi}$ is invertible. If $\overline{\Psi}$ is known, by the continuity of F and G, an estimate of θ can be constructed as $(\overline{\Psi})^{-1}G(\widetilde{\xi}_N) \to \Theta$ w.p.1. Due to the input measure noise, $\overline{\Psi}$ is not measured directly. What we can use is $\widetilde{\Psi}_N^w$. Theorem 8.3 confirms that $\Theta_N = (\overline{\Psi}_N^w)^{-1}G(\widetilde{\xi}_N)$ will be a consistent estimate of Θ.

Set initial conditions as $\widetilde{\xi}_1 = \widetilde{S}_1 = [1, 1, 1, 1]'$, $\widetilde{\Psi}_1 = \widetilde{\Phi}_1$, $\Theta_1 = \mathbf{0}$. We construct a causal and recursive algorithm as in Section 8.1.2. The relative estimation error $\|\Theta_N - \Theta\|/\|\Theta\|$ is used to evaluate the accuracy and convergence of the estimates. Figure 8.3 shows that Θ_N converges to the true parameters $\Theta = [C, \theta']'$.

Example 8.12. When the noise distribution function is unknown, joint identification is used to estimate the system parameters and noise distribution function jointly. Consider a gain system ($n = 1$): $y_k = au_k + d_k$, where the true value $a = 2$, and $\{d_k\}$ is a sequence of i.i.d. normal variables. The sensor has threshold $C = 12$. Let $F_0(x)$ be the normal distribution function of zero mean and variance 1, and let $G_0(x)$ be the inverse of $F_0(x)$. Then the distribution function of d_k is

$$F(x; [\mu, \sigma]) = F_0((x - \mu)/\sigma).$$

Let $\mu = 3$ be given, and the true value of variance $\sigma = 3$. If μ is known, $F(x; [\mu, \sigma])$ is jointly identifiable. Let $v = 4$. For $k = 1, 2, \ldots$, the scaled

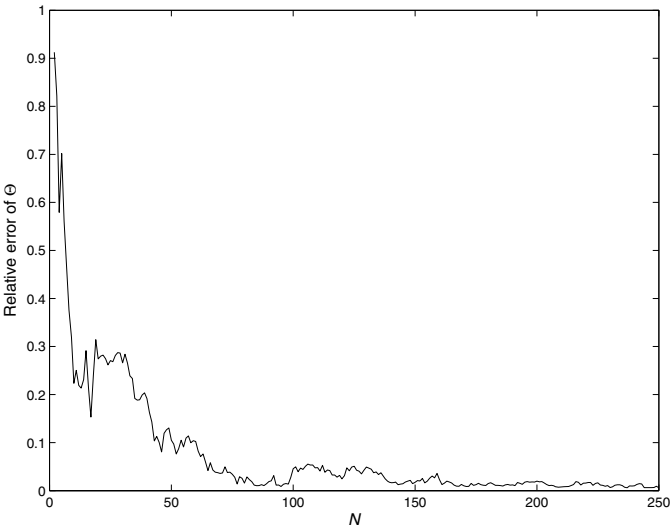


FIGURE 8.3. Recursive algorithm to estimate the parameters when C is unknown

input is defined as $u_{2k-1} = v$; $u_{2k} = qv$, where $q = 1.05$. It is easy to verify that u is an exponentially scaled full-rank signal (see Definition 11.2). Set $U = [4, 4.2]'$ and

$$\xi_N = \Big[\frac{1}{N}\sum_{i=1}^{N} s_{2i-1}, \frac{1}{N}\sum_{i=1}^{N} s_{2i}\Big]'.$$

Then $\xi_N \to \xi$ w.p.1 and $G_0(\xi_N) \to [(C-\mu)\mathbb{1}_2 - aU]/\sigma$. Since $\widehat{F}(x,\alpha)$ is jointly identifiable, we obtain the estimates of a and σ: $[\widehat{a}_N, \hat{\sigma}_N]' = [U, G_0(\xi_N)]^{-1}[8,8]'$. Figure 8.4 illustrates that the estimated values of the system parameter and distribution function parameter converge to the actual ones.

8.10 Notes

The material in this chapter is based on the results in [108]. Unknown noise distribution functions introduce a static nonlinearity as part of the unknowns. Combined with the linear system, we are in fact dealing with an identification of a Wiener structure, using binary-valued or quantized observations. The main idea for dealing with this more complicated problem remains the same: Find a suitable input signal under which the identification of this nonlinear system with many unknown parameters can be transformed into a finite set of simple identification problems for gains. The input turns out to be a scaled periodic signal and the nonlinear mapping for the transformation is derived. Consequently, the basic convergence

conclusions of Chapter 3 can be applied. This idea will also be employed in later chapters when we deal with nonlinear systems.

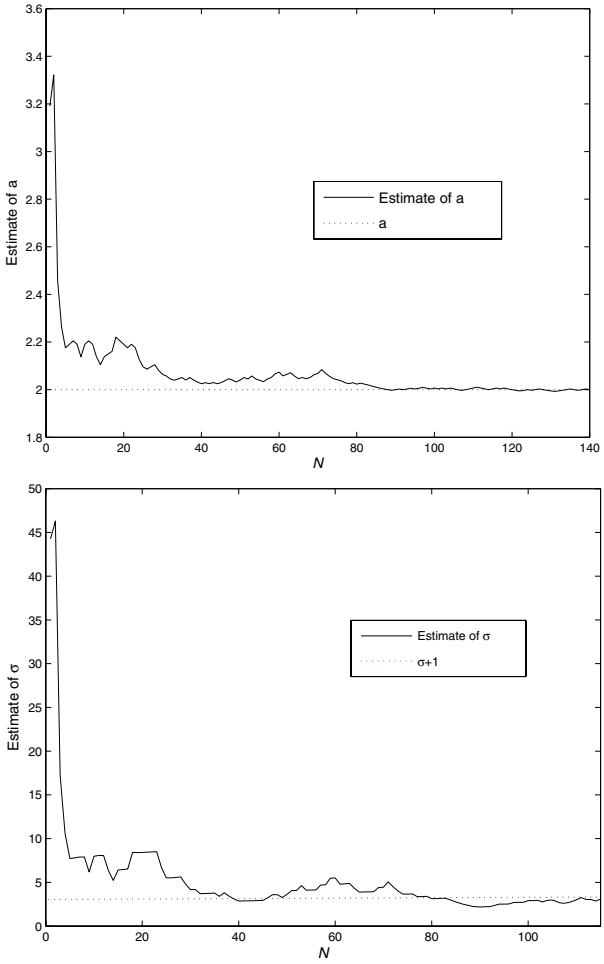


FIGURE 8.4. Joint identification of system parameter a and distribution function parameter σ

Part III

Deterministic Methods for Linear Systems

9

Worst-Case Identification under Binary-Valued Observations

This chapter focuses on the identification of systems where the disturbances are formulated in a deterministic framework as unknown but bounded. Different from the previous chapters, here the identification error is measured by the radius of the set that the unknown parameters belong to, which is a worst-case measure of the parameter uncertainties. By considering several different combinations of the disturbances and unmodeled dynamics, a number of fundamental issues are studied in detail: When only binary-valued observations are available, how accurately can one identify the parameters of the system? How fast can one reduce uncertainty on model parameters? What are the optimal inputs for fast identification? What is the impact of unmodeled dynamics and disturbances on identification accuracy and time complexity?

The rest of this chapter is arranged as follows. In Section 9.1, the problem is formulated and the main conditions on the disturbances and unmodeled dynamics are given. In Section 9.2, lower bounds on the identification errors and time complexity of the identification algorithms are established, underscoring an inherent relationship between identification time complexity and the Kolmogorov ε-entropy. Identification input design and upper bounds on identification errors are then derived in Section 9.3, demonstrating that the Kolmogorov ε-entropy indeed defines the complexity rates. For the single parameter case, the results are refined further in Section 9.4. Section 9.5 presents a comparison between the stochastic and deterministic frameworks. In contrast to the common perception that these two are competing frameworks, we show that they complement each other in binary sensor identification.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_9,

9.1 Worst-Case Uncertainty Measures

Recall the linear system defined in Chapter 2 and further detailed in Chapter 4,

$$y_k = \sum_{i=0}^{\infty} a_i u_{k-i} + d_k, \qquad k = k_0 + 1, k_0 + 2, \ldots,$$

where $\{d_k\}$ is a sequence of disturbances, $\{u_k\}$ is the input with $u_k = 0$, $k < k_0$, and $a = \{a_i, i = 0, 1, \ldots\}$, satisfying $\|a\|_1 = \sum_{i=0}^{\infty} |a_i| < \infty$, is the vector-valued parameter.

As in Chapter 4, for a given model order n_0, the system parameters can be decomposed into the modeled part $\theta = [a_0, \ldots, a_{n_0-1}]'$ and the unmodeled dynamics $\widetilde{\theta} = [a_{n_0}, a_{n_0+1}, \ldots]'$, and the system input–output relationship can be expressed as

$$y_k = \phi_k'\theta + \widetilde{\phi}_k'\widetilde{\theta} + d_k, \quad k = k_0 + 1, k_0 + 2, \ldots, \tag{9.1}$$

where

$$\phi_k = [u_k, u_{k-1}, \ldots, u_{k-n_0+1}]',$$
$$\widetilde{\phi}_k = [u_{k-n_0}, u_{k-n_0-1}, \ldots]'.$$

Under a selected input sequence u_k, the output s_k from a binary-valued sensor of threshold C is measured for $k = k_0+1, \ldots, k_0+N$. We would like to estimate θ on the basis of input–output observations on u_k and s_k. The issues of identification accuracy, time complexity, and input design will be discussed.

Because some results in this chapter will be valid under any l^p norm, the following assumption is given in a generic l^p norm. The norm will be further specified if certain results are valid only for some p values.

(A9.1) For a fixed $p \geq 1$, to be specified later,

1. the unmodeled dynamics $\widetilde{\theta}$ is bounded in the l^p norm by $\|\widetilde{\theta}\|_p \leq \eta$;
2. the disturbance d is uniformly bounded in the l^∞ norm by $\|d\|_\infty \leq \delta$;
3. the prior information on θ is given by $\Omega_0 = \mathrm{Ball}_p(\theta_0, \varepsilon_0) \subset \mathbb{R}^{n_0}$ for $\theta_0 \in \mathbb{R}^{n_0}$ and $\varepsilon_0 > 0$.

For a selected input sequence u_k, let $s = \{s_k,\ k = k_0 + 1, \ldots, k_0 + N\}$ be the observed output. Define

$$\Omega_N(k_0, u, s) = \{\theta :\ \ s_k = I_{\{\phi_k'\theta + \widetilde{\phi}_k'\widetilde{\theta} + d_k \leq C\}} \text{ for some } \|\widetilde{\theta}\|_p \leq \eta, \\ \|d\|_\infty \leq \delta \text{ and } k = k_0 + 1, \ldots, k_0 + N\}$$

and

$$e_N = \inf_{\|u\|_\infty \leq u_{\max}} \sup_{k_0} \sup_{s} \mathrm{Rad}_p \left(\Omega_N(k_0, u, s) \cap \mathrm{Ball}_p(\theta_0, \varepsilon_0)\right),$$

where $\mathrm{Rad}_p(\Omega)$ is the radius of Ω in the l_p norm. e_N is the optimal (in terms of the input design) worst-case (with respect to initial time, unmodeled dynamics, and disturbances) uncertainty after N steps of observations. For a given desired identification accuracy ε, the time complexity of $\mathrm{Ball}_p(\theta_0, \varepsilon_0)$ is defined as

$$N(\varepsilon) = \min\{N : \ e_N \leq \varepsilon\}.$$

We will characterize e_N, determine optimal or suboptimal inputs u, and derive bounds on time complexity $N(\varepsilon)$.

9.2 Lower Bounds on Identification Errors and Time Complexity

We will show in this section that identification time complexity is bounded below by the Kolmogorov entropy of the prior uncertainty set.

Noise-Free and No Unmodeled Dynamics

Theorem 9.1. *Assume Assumption* (A9.1). *Let* $\delta = 0$ *and* $\eta = 0$. *Suppose that for a given* $p \geq 1$ *the prior uncertainty* $\Omega_0 = \mathrm{Ball}_p(\theta_0, \varepsilon_0)$. *Then, for any* $\varepsilon < \varepsilon_0$, *the time complexity* $N(\varepsilon)$ *is bounded below by* $N(\varepsilon) \geq n_0 \log(\varepsilon_0/\varepsilon)$.

Proof. $\mathrm{Ball}_p(c, \varepsilon)$ in $\mathbb{R}^{n_0}$ has volume $a_{p,n_0}\varepsilon^{n_0}$, where the coefficient a_{p,n_0} is independent of ε. To reduce the identification error from ε_0 to below ε, the volume reduction must be at least

$$a_{p,n_0}\varepsilon^{n_0}/(a_{p,n_0}\varepsilon_0^{n_0}) = (\varepsilon/\varepsilon_0)^{n_0}.$$

Each binary sensor observation defines a hyperplane in the parameter space $\mathbb{R}^{n_0}$. The hyperplane divides an uncertainty set into two subsets, with the volume of the larger subset at least half of the volume of the original set. As a result, in a worst-case scenario, one binary observation can reduce the volume of a set by $1/2$ at best. Hence, the number N of observations required to achieve the required error reduction is at least

$$(1/2)^N \leq (\varepsilon/\varepsilon_0)^{n_0}, \quad \text{or } N \geq n_0 \log(\varepsilon_0/\varepsilon).$$

The proof is concluded. □

It is noted that $n \log(\varepsilon_0/\varepsilon)$ is precisely the Kolmogorov ε-entropy of the prior uncertainty set Ω_0 [50, 125]. Hence, Theorem 9.1 provides an interesting new interpretation of the Kolmogorov entropy in system identification, beyond its application in characterizing model complexity [125]. Theorem 9.1 establishes a lower bound of exponential rates of time complexity. Upon obtaining an upper bound of the same rates in the next section, we will

show that the Kolmogorov ε-entropy indeed defines the time complexity rates in this problem. Next, we present an identifiability result, which is limited to $p = 1$.

Proposition 9.2. *The uncertainty set* $\mathrm{Ball}_1(0, C/u_{\max})$ *is not identifiable.*

Proof. For any $\theta \in \mathrm{Ball}_1(0, C/u_{\max})$, the output

$$y_k = \phi_k'\theta \le \|\phi_k\|_\infty \|\theta\|_1 \le u_{\max}\frac{C}{u_{\max}} = C.$$

It follows that $s_k = 1, \forall k$. Hence, the observations could not provide further information to reduce uncertainty. □

Bounded Disturbances

In the case of noisy observations, the input–output relationship becomes

$$y_k = \phi_k'\theta + d_k, \quad s_k = I_{\{y_k \le C\}}, \tag{9.2}$$

where $|d_k| \le \delta$. For any given ϕ_k, an observation on s_k from (9.2) defines, in a worst-case sense, two possible uncertainty half-planes:

$$\Omega_1 = \{\theta \in \mathbb{R}^{n_0} : \ \phi_k'\theta \le C + \delta\}, \quad s_k = 1,$$
$$\Omega_2 = \{\theta \in \mathbb{R}^{n_0} : \ \phi_k'\theta > C - \delta\}, \quad s_k = 0.$$

Uncertainty reduction via observation is possible only if the uncertainty set before observation is not a subset of each half-plane (so that the intersection of the uncertainty set and the half-plane results in a smaller set).

Theorem 9.3. *If* $\varepsilon \le \delta/u_{\max}$*, then for any* $\theta_0 \in \mathbb{R}^{n_0}$*, either* $\mathrm{Ball}_1(\theta_0, \varepsilon) \subseteq \Omega_1$ *or* $\mathrm{Ball}_1(\theta_0, \varepsilon) \subseteq \Omega_2$*. Consequently, in a worst-case sense,* $\mathrm{Ball}_1(\theta_0, \varepsilon)$ *is not identifiable.*

Proof. Suppose that $\mathrm{Ball}_1(\theta_0, \varepsilon) \not\subseteq \Omega_1$. Then, there exists $\theta_1 \in \mathrm{Ball}_1(\theta_0, \varepsilon)$ such that $\phi_k'\theta_1 > C + \delta$. $\theta \in \mathrm{Ball}_1(\theta_0, \varepsilon)$ satisfies $\|\theta - \theta_1\|_1 \le 2\varepsilon$. We have

$$\begin{aligned}\phi_k'\theta &= \phi_k'\theta_1 + \phi_k'(\theta - \theta_1)\\ &> C + \delta + \phi_k'(\theta - \theta_1)\\ &\ge C + \delta - u_{\max}2\varepsilon \ge C - \delta,\end{aligned}$$

for any $\theta \in \mathrm{Ball}_1(\theta_0, \varepsilon)$. This implies that $\mathrm{Ball}_1(\theta_0, \varepsilon) \subseteq \Omega_2$. Likewise, we can show that if $\mathrm{Ball}_1(\theta_0, \varepsilon) \not\subset \Omega_2$, then it is contained in Ω_1. □

Theorem 9.3 shows that worst-case disturbances introduce irreducible identification errors of size at least $\delta/u_{\max}$. This is a general result. A

substantially higher lower bound can be obtained in the special case of $n_0 = 1$.

Consider the system $y_k = au_k + d$. Suppose that at time k the prior information on a is that $a \in \Omega = [\underline{a}, \overline{a}]$ with $\underline{a} > C/u_{\max}$ for identifiability (see Proposition 9.2). The uncertainty set has center $a_0 = (\underline{a} + \overline{a})/2$ and radius $\varepsilon = (\overline{a} - \underline{a})/2$. To minimize the posterior uncertainty in the worst-case sense, the optimal u_k can be easily obtained as $u_k = C/a_0$.

Theorem 9.4. *If $\delta < C$, then the uncertainty set $[\underline{a}, \overline{a}]$ cannot be reduced if*

$$\varepsilon \le \frac{\delta/u_{\max}}{1 - \delta/C}.$$

Proof. Let $\varepsilon = \frac{\delta/u_{\max}}{1-\delta/C}$. Then, $\delta = \frac{\varepsilon C}{C/u_{\max}+\varepsilon}$. For any $a \in [\underline{a}, \overline{a}]$, noting $a_0 = \underline{a} + \varepsilon$, we have $|a - a_0| \le \varepsilon$, and

$$\begin{aligned} au_k &= a\frac{C}{a_0} = (a_0 + (a - a_0))\frac{C}{a_0} = C + (a - a_0)\frac{C}{a_0} \\ &\le C + \frac{\varepsilon C}{\underline{a} + \varepsilon} < C + \frac{\varepsilon C}{C/u_{\max} + \varepsilon} = C + \delta. \end{aligned}$$

Hence, the observation $s_k = 1$ does not provide any information. Similarly, if $s_k = 0$, we can show that all $\theta \in [\underline{a}, \overline{a}]$ will result in $au_k > C - \delta$. Again, the observation does not reduce uncertainty. □

At present, it remains an open question if Theorem 9.4 holds for higher-order systems.

Unmodeled Dynamics

When the system contains unmodeled dynamics, the input–output relationship becomes

$$y_k = \phi_k'\theta + \widetilde{\phi}_k'\widetilde{\theta}, \qquad s_k = I_{\{y_k \le C\}}, \tag{9.3}$$

where $\|\widetilde{\theta}\|_1 \le \eta$. We will show that unmodeled dynamics will introduce an irreducible identification error on the modeled part.

For any $\widetilde{\phi}_k$, the set $\{\widetilde{\phi}_k'\widetilde{\theta} : \|\widetilde{\theta}\|_1 \le \eta\} = [-\eta m_k, \eta m_k]$, where $m_k = \|\widetilde{\phi}_k\|_\infty$.

Theorem 9.5. *If $\varepsilon \le \eta$, then in a worst-case sense, for any θ_0,* $\mathrm{Ball}_1(\theta_0, \varepsilon)$ *is not identifiable.*

Proof. Under (9.3), an observation on s_k provides observation information

$$\begin{aligned} \Omega_1 &= \{\theta \in \mathbb{R}^{n_0} : \phi_k'\theta \le C + \eta m_k\}, \quad s_k = 1, \\ \Omega_2 &= \{\theta \in \mathbb{R}^{n_0} : \phi_k'\theta > C - \eta m_k\}, \quad s_k = 0. \end{aligned}$$

In the worst-case sense, $\mathrm{Ball}_1(\theta_0, \varepsilon)$ can be reduced by this observation only if $\mathrm{Ball}_1(\theta_0, \varepsilon)$ is a subset of neither Ω_1 nor Ω_2.

Suppose that $\mathrm{Ball}_1(\theta_0, \varepsilon) \not\subseteq \Omega_2$. We will show that $\mathrm{Ball}_1(\theta_0, \varepsilon) \subseteq \Omega_1$. Indeed, in this case, there exists $\theta_1 \in \mathrm{Ball}_1(\theta_0, \varepsilon)$ such that $\phi_k'\theta_1 \leq C - \eta m_k$. Since any $\theta \in \mathrm{Ball}_1(\theta_0, \varepsilon)$ satisfies $\|\theta - \theta_1\|_1 \leq 2\varepsilon$, we have

$$\begin{aligned} \phi_k'\theta &= \phi_k'\theta_1 + \phi_k'(\theta - \theta_1) \\ &\leq C - \eta m_k + \phi_k'(\theta - \theta_1) \\ &\leq C - \eta m_k + m_k 2\varepsilon \\ &\leq C + \eta m_k. \end{aligned}$$

This implies $\mathrm{Ball}_1(\theta_0, \varepsilon) \subseteq \Omega_1$. □

9.3 Upper Bounds on Time Complexity

In this subsection, general upper bounds on identification errors or time complexity will be established.

For a fixed $p \geq 1$, suppose that the prior information on θ is given by $\mathrm{Ball}_p(\theta_0, \varepsilon_0)$. For identifiability, assume that the signs of a_i have been detected and

$$\underline{a} = \min\{|\underline{a}_i|, i = 1, \ldots, n\} > \frac{C}{u_{\max}}.$$

The sign of a_i can be obtained easily by choosing an initial testing sequence of u. Also, those parameters with $|\underline{a}_i| < C/u_{\max}$ can be easily detected. Since uncertainty on these parameters cannot be further reduced (see Proposition 9.2), they will be left as remaining uncertainty. $\underline{a}$ defined here will be applied to the rest of the parameters. The detail is omitted for brevity. Denote

$$\overline{a} = \max_{\theta \in \mathrm{Ball}_p(\theta_0, \varepsilon_0)} \|\theta\|_\infty.$$

We will establish upper bounds on the time complexity $N(\varepsilon)$ to reduce the size of the uncertainty from ε_0 to ε, in the l^p norm.

Noise-Free and No Unmodeled Dynamics

Let $\eta = 0$ and $\delta = 0$ and consider $y_k = \phi_k'\theta$.

Theorem 9.6. *Suppose that $u_{\max} > C/\underline{a}$. Then the time complexity to reduce the uncertainty from ε_0 to ε is bounded by*

$$N(\varepsilon) \leq ({n_0}^2 - n_0 + 1) \left\lceil \frac{1}{p} \log n_0 + \log \frac{\varepsilon_0}{\varepsilon} \right\rceil. \tag{9.4}$$

Since n_0 is a constant independent of N, this result, together with Theorem 9.1, confirms that the Kolmogorov entropy defines the time complexity rates in binary sensor identification. The accurate calculation for $N(\varepsilon)$ remains an open and difficult question, except for $n_0 = 1$ (gain uncertainty), which is discussed in the next section.

The proof of Theorem 9.6 utilizes the following lemma. Consider the first-order system $y_k = au_k$, $s_k = I_{\{y_k \le C\}}$, where $a \in [\underline{a}, \overline{a}]$ and $\underline{a} > C/u_{\max} > 0$. Let $\varepsilon_0 = (\overline{a} - \underline{a})/2$.

Lemma 9.7. *There exists an input sequence u such that N observations on s_k can reduce the radius of uncertainty to $\varepsilon = 2^{-N}\varepsilon_0$.*

Proof. Let $[\underline{a}_k, \overline{a}_k]$ be the prior uncertainty before a measurement on s_k. Then $\varepsilon_k = (\overline{a}_k - \underline{a}_k)/2$. By choosing $u_k = C/(\underline{a}_k + \varepsilon_k)$, the observation on s_k will determine uniquely either $a \in [\underline{a}_k, \underline{a}_k + \varepsilon_k]$ if $s_k = 1$; or $a \in [\overline{a}_k - \varepsilon_k, \overline{a}_k]$ if $s_k = 0$. In either case, the uncertainty is reduced by half. Iterating on the number of observations leads to the conclusion. □

The proofs of this section rely on the following idea. Choose $u_k = 0$ except those with index $j({n_0}^2 - n_0 + 1) + i$, $i = 1, n_0 + 1, \ldots, (n_0 - 1)n_0 - n_0 + 3$, $j = 0, 1, \ldots$ This input design results in a specific input–output relationship:

$$\begin{cases} y_{j({n_0}^2 - n_0 + 1) + n_0} = a_{n_0 - 1} u_{j({n_0}^2 - n_0 + 1) + 1}, \\ y_{j({n_0}^2 - n_0 + 1) + n_0 + 1} = a_0 u_{j({n_0}^2 - n_0 + 1) + n_0 + 1}, \\ \quad \vdots \\ y_{j({n_0}^2 - n_0 + 1) + (n_0 - 1)n_0 + 1} = a_{n_0 - 2} u_{j({n_0}^2 - n_0 + 1) + (n_0 - 1)n_0 - n_0 + 3}. \end{cases} \tag{9.5}$$

In other words, within each block of ${n_0}^2 - n_0 + 1$ observations, each model parameter can be identified individually once. Less conservative inputs can be designed. However, they are more problem-dependent and ad hoc, and will not be presented here.

Proof of Theorem 9.6. By Lemma 9.7, the uncertainty radius on each parameter can be reduced by a factor of 2^{-N_1} after N_1 observations. This implies that by using the input (9.5), after $N = ({n_0}^2 - n_0 + 1)N_1$ observations, the uncertainty radius can be reduced to

$$\begin{aligned} \mathrm{rad}_p(\Omega_N) &\le {n_0}^{1/p} \mathrm{rad}_\infty(\Omega_N) \le {n_0}^{1/p} 2^{-\frac{N}{{n_0}^2 - n_0 + 1}} \mathrm{rad}_\infty(\Omega_0) \\ &\le {n_0}^{1/p} 2^{-\frac{N}{{n_0}^2 - n_0 + 1}} \mathrm{rad}_p(\Omega_0) = {n_0}^{1/p} 2^{-\frac{N}{{n_0}^2 - n_0 + 1}} \varepsilon_0. \end{aligned}$$

Hence, for

$${n_0}^{1/p} 2^{-\frac{N}{{n_0}^2 - n_0 + 1}} \varepsilon_0 \le \varepsilon,$$

it suffices to have

$$N = ({n_0}^2 - n_0 + 1) \left\lceil \frac{1}{p} \log n + \log \frac{\varepsilon_0}{\varepsilon} \right\rceil.$$

The desired result follows. □

Bounded Disturbances

Consider $y_k = \phi_k'\theta + d_k$, where $|d_k| \le \delta$.

Theorem 9.8. *Suppose $\delta < C$. Let*

$$\beta = \frac{\delta}{C}, \quad \rho = \frac{1}{2}(1-\beta), \quad \textit{and} \quad \sigma = \frac{\delta\overline{a}}{2C(1-\rho)} = \frac{\overline{a}\beta}{1+\beta}.$$

If $\varepsilon_0 > \varepsilon > \sigma$ and $u_{\max} > C/\underline{a}$, then the time complexity $N(\varepsilon)$ to reduce the uncertainty from ε_0 to ε is bounded in the l^p norm by

$$N(\varepsilon) \le ({n_0}^2 - n_0 + 1)\left\lceil \frac{1}{p}\log n_0 + \frac{\log\frac{\varepsilon-\sigma}{\varepsilon_0-\sigma}}{\log\rho}\right\rceil. \tag{9.6}$$

Proof. Using the input in (9.5), the identification of the n parameters a_0, $\ldots$, a_{n_0-1} is reduced to identifying each parameter individually. Now for identification of a single parameter $y_k = au_k + d_k$, we can derive the following iterative uncertainty reduction relationship. If the prior uncertainty at k is $[a_k - \varepsilon_k, a_k + \varepsilon_k]$, then the optimal worst-case input u_k can be shown as $u_k = C/a_k$. (More detailed derivations are given in the next section.) The posterior uncertainty will be either $[a_k - \varepsilon_k, (1+\beta)a_k]$, if $s_k = 1$; or $[(1-\beta)a_k, a_k + \varepsilon_k]$, if $s_k = 0$. Both have the radius

$$\varepsilon_{k+1} = \frac{1}{2}(\varepsilon_k + \beta a_k) = \frac{1-\beta}{2}\varepsilon_k + \frac{\beta}{2}(a_k + \varepsilon_k) \le \rho\varepsilon_k + \frac{\beta\overline{a}}{2}.$$

Starting from ε_0, after N_1 observations, we have

$$\begin{aligned}\varepsilon(N_1) &\le \rho^{N_1}\varepsilon_0 + \frac{\beta\overline{a}}{2}\sum_{i=0}^{N_1-1}\rho^i = \rho^{N_1}\varepsilon_0 + \frac{\beta\overline{a}}{2}\frac{1-\rho^{N_1}}{1-\rho}\\ &= \rho^{N_1}\varepsilon_0 + \sigma(1-\rho^{N_1}) = \rho^{N_1}(\varepsilon_0 - \sigma) + \sigma.\end{aligned}$$

To achieve $\varepsilon(N_1) \le \varepsilon$, it suffices that

$$\rho^{N_1}(\varepsilon_0 - \sigma) + \sigma \le \varepsilon \quad \text{or} \quad N_1 \ge \frac{\log\frac{\varepsilon-\sigma}{\varepsilon_0-\sigma}}{\log\rho}.$$

Following the same arguments as in the proof of Theorem 9.6, we conclude that

$$N = ({n_0}^2 - n_0 + 1)\left\lceil \frac{1}{p}\log n_0 + \frac{\log\frac{\varepsilon-\sigma}{\varepsilon_0-\sigma}}{\log\rho}\right\rceil$$

suffices to reduce the uncertainty from ε_0 to ε in the l^p norm. □

Unmodeled Dynamics

Consider $y_k = \phi_k'\theta + \widetilde{\phi}_k'\widetilde{\theta}$. The results of this case hold for $p = 1$ only. The unmodeled dynamics introduce an uncertainty on the observation on y_k: $\{\widetilde{\phi}_k'\widetilde{\theta} : \|\widetilde{\theta}\|_1 \le \eta\} = [-\eta m_k, \eta m_k]$, $m_k = \|\phi_k\|_\infty$.

Theorem 9.9. *Suppose* $0 < \eta < C/u_{\max}$. *Let*

$$\rho_1 = \frac{1}{2}\left(1 - \frac{\eta u_{\max}}{C}\right), \quad \sigma_1 = \frac{\eta u_{\max}\overline{a}}{2C(1-\rho_1)}.$$

Then

$$N(\varepsilon) \le ({n_0}^2 - n_0 + 1)\left\lceil \log n_0 + \frac{\log \frac{\varepsilon - \sigma_1}{\varepsilon_0 - \sigma_1}}{\log \rho_1} \right\rceil. \tag{9.7}$$

Proof. By using the input (9.5), the identification of θ is reduced to each of its components. For a scalar system $y_k = au_k + \widetilde{\phi}_k'\widetilde{\theta}$, since $|\widetilde{\phi}_k'\widetilde{\theta}| \le \eta u_{\max}$, we can apply Theorem 9.8 with δ replaced by $\eta u_{\max}$. Inequality (9.7) then follows from Theorem 9.8. □

9.4 Identification of Gains

In the special case $n = 1$, explicit results and tighter bounds can be obtained. When $n = 1$, the observation equation becomes

$$y_k = au_k + \widetilde{\phi}_k'\widetilde{\theta} + d_k.$$

Assume that the initial information on a is that $\underline{a}_0 \le a \le \overline{a}_0$, $\underline{a}_0 \ne 0$, $\overline{a}_0 \ne 0$, with radius $\varepsilon_0 = (\overline{a}_0 - \underline{a}_0)/2$.

Case 1: $y_k = au_k$

It is noted that this is a trivial identification problem when regular sensors are used: After one input $u_0 \ne 0$, a can be identified uniquely.

Theorem 9.10. *The following assertions hold.*

(1) *Suppose the sign of* a *is known, say,* $\underline{a}_0 > 0$, *and* $u_{\max} \ge C/\underline{a}_0$. *Then the optimal identification error is* $e_N = 2^{-N}e_0$ *and the time complexity is* $N(\varepsilon) = \lceil \log(\varepsilon_0/\varepsilon) \rceil$.

If, at $k-1$, *the information on* a *is that* $a \in [\underline{a}_{k-1}, \overline{a}_{k-1}]$, *then the one-step optimal* u_k *is*

$$u_k = \frac{2C}{\underline{a}_{k-1} + \overline{a}_{k-1}}, \tag{9.8}$$

where $\underline{a}_k$ *and* $\overline{a}_k$ *are updated by*

$$\underline{a}_k = \begin{cases} (\underline{a}_{k-1} + \overline{a}_{k-1})/2, & \text{if } s_k = 0, \\ \underline{a}_{k-1}, & \text{if } s_k = 1; \end{cases}$$

$$\overline{a}_k = \begin{cases} \overline{a}_{k-1}, & \text{if } s_k = 0, \\ (\underline{a}_{k-1} + \overline{a}_{k-1})/2, & \text{if } s_k = 1. \end{cases}$$

(2) *If* $\underline{a}_0$ *and* $\overline{a}_0$ *have opposite signs and*

$$\delta_l = \max\left\{\underline{a}_0, -\frac{C}{u_{\max}}\right\}, \quad \delta_h = \min\left\{\overline{a}_0, \frac{C}{u_{\max}}\right\},$$

then the uncertainty interval (δ_l, δ_h) *is not identifiable. Furthermore, in the case of* $\underline{a}_0 \le \delta_l$ *and* $\overline{a}_0 \ge \delta_h$*, if* $\delta_h - \delta_l \le \varepsilon$ *and* $\varepsilon_0 \ge 2\varepsilon$*, then the time complexity* $N(\varepsilon)$ *is bounded by*

$$\left\lceil \log \frac{\varepsilon_0}{\varepsilon} \right\rceil \le N(\varepsilon) \le \left\lceil \log \frac{\varepsilon_0 - (\delta_h - \delta_l)}{\varepsilon} \right\rceil + 2.$$

Proof. The proof is divided into a couple of steps.

(1) The identification error and time complexity follow directly from Theorems 9.1 and 9.6 with $n = 1$. As for the optimal input, note that starting from the uncertainty $[\underline{a}_k, \overline{a}_k]$, an input u_k defines a testing point C/u_k on a. The optimal worst-case input is then obtained by placing the testing point at the middle. That is,

$$\frac{C}{u_k} = \frac{1}{2}(\underline{a}_k + \overline{a}_k),$$

which leads to the optimal input and results in posterior uncertainty sets.

(2) When the input is bounded by $u_k \in [-u_{\max}, u_{\max}]$, the testing points cannot be selected in the interval $[-C/u_{\max}, C/u_{\max}]$. Consequently, this uncertainty set cannot be further reduced by identification. Furthermore, by using $u_1 = -u_{\max}$ and $u_2 = u_{\max}$ as the first two input values, a can be determined as belonging uniquely to one of the three intervals:

$$[\underline{a}_0, -C/u_{\max}),\ [-C/u_{\max}, C/u_{\max}],\ [C/u_{\max}, \overline{a}_0].$$

By taking the worst-case scenario of

$$\overline{a}_0 - C/u_{\max} = \varepsilon_0 - (\delta_h - \delta_l),$$

the time complexity for reducing the remaining uncertainty to ε is $\left\lceil \log \frac{\varepsilon_0 - (\delta_h - \delta_l)}{\varepsilon} \right\rceil$. This leads to the upper bound on $N(\varepsilon)$. The lower bound follows from Theorem 9.1 with $n = 1$.

□

In this special case, the actual value $C > 0$ does not affect the identification accuracy. This is due to noise-free observation. The value C will become essential in deriving optimal identification errors when observation noises are present. $C = 0$ is a singular case in which the uncertainty on a cannot be reduced (in the sense of the worst-case scenario). Indeed, in this case, one can only test the sign of a. It is also observed that the optimal u_k depends on the previous observation s_{k-1}. As a result, u_k can be constructed causally and sequentially, but not off-line.

Case 2: $y_k = au_k + d_k$

Here we assume $|d_k| \le \delta < C$. The prior information on a is given by $a \in \Omega_0 = [\underline{a}_0, \overline{a}_0]$, and $\underline{a}_0 > 0$.

Theorem 9.11. *Suppose that*

$$u_{\max} \ge \frac{C}{\underline{a}_0} \quad \textit{and} \quad \frac{\overline{a}_0}{\underline{a}_0} \ge \frac{1+\beta}{1-\beta}.$$

Then

(1) *the optimal input u_k is given by the causal mapping from the available information at $k-1$:*

$$u_k = \frac{2C}{\underline{a}_{k-1} + \overline{a}_{k-1}}.$$

The optimal identification error satisfies the iteration equation

$$e_k = \frac{1}{2} e_{k-1} + \frac{1}{2}\beta(\overline{a}_{k-1} + \underline{a}_{k-1}), \tag{9.9}$$

where $\overline{a}_k$ and $\underline{a}_k$ are updated by the rules

$$\begin{aligned} \overline{a}_k &= \overline{a}_{k-1}, \quad \underline{a}_k = \frac{C-\delta}{u_k}, \quad \text{if } s_k = 0, \\ \underline{a}_k &= \underline{a}_{k-1}, \quad \overline{a}_k = \frac{C+\delta}{u_k}, \quad \text{if } s_k = 1. \end{aligned}$$

(2)

$$\frac{\overline{a}(k)}{\underline{a}(k)} \ge \frac{1+\beta}{1-\beta} \quad \textit{for all} \quad k \ge 1;$$

$\{\underline{a}_k\}$ is monotonically increasing, $\{\overline{a}_k\}$ and $\left\{\frac{\overline{a}_k}{\underline{a}_k}\right\}$ are monotonically decreasing;

$$\lim_{k\to\infty} \frac{\overline{a}_k}{\underline{a}_k} = \frac{1+\beta}{1-\beta}.$$

(3) *At each time k, uncertainty reduction is possible if and only if*

$$\frac{\overline{a}_{k-1}}{\underline{a}_{k-1}} > \frac{1+\beta}{1-\beta}.$$

Proof. (1) Since $u_k > 0$, the relationship (9.2) can be written as $a = \frac{y_k - d_k}{u_k}$. The observation outcome $y_k \geq C$ will imply that

$$a \geq \frac{C - d_k}{u_k} \geq \frac{C - \delta}{u_k},$$

which will reduce uncertainty from $a \in [\underline{a}_{k-1}, \overline{a}_{k-1}]$ to $[\frac{C-\delta}{u_k}, \overline{a}_{k-1}]$ with error $e_1(k) = \overline{a}_{k-1} - \frac{C-\delta}{u_k}$. Similarly, $y < C$ implies $a < \frac{C+\delta}{u_k}$ and $a \in [\underline{a}_{k-1}, \frac{C+\delta}{u_k}]$ with $e_2(k) = \frac{C+\delta}{u_k} - \underline{a}_{k-1}$. In a worst-case scenario,

$$e_k = \max\{e_1(k), e_2(k)\}.$$

Consequently, the optimal u_k can be derived from $\inf_{u_k} \ e_k$. Hence, the optimal u_k is the one that causes $e_1(k) = e_2(k)$, namely,

$$\frac{C+\delta}{u_k} - \underline{a}_{k-1} = \overline{a}_{k-1} - \frac{C-\delta}{u_k},$$

or

$$u_k = \frac{2C}{\underline{a}_{k-1} + \overline{a}_{k-1}}.$$

The optimal identification error is then

$$\begin{aligned} e_k &= \frac{(C+\delta)(\overline{a}_{k-1} + \underline{a}_{k-1})}{2C} - \underline{a}_{k-1} \\ &= \left(\frac{1}{2} + \frac{\beta}{2}\right)(\overline{a}_{k-1} + \underline{a}_{k-1}) - \underline{a}_{k-1} \\ &= \frac{1}{2} e_{k-1} + \frac{\beta}{2}(\overline{a}_{k-1} + \underline{a}_{k-1}). \end{aligned}$$

(2) We prove $\frac{\overline{a}_k}{\underline{a}_k} \geq \frac{1+\beta}{1-\beta}$ by induction. Suppose that $\frac{\overline{a}_{k-1}}{\underline{a}_{k-1}} \geq \frac{1+\beta}{1-\beta}$. Then we have $u_k \overline{a}_{k-1} \geq C + \delta$ and $u_k \underline{a}_{k-1} \leq C - \delta$, which, respectively, leads to $\frac{\overline{a}_k}{\underline{a}_k} = \frac{u_k \overline{a}_{k-1}}{C-\delta} \geq \frac{1+\beta}{1-\beta}$ in the case of $s_k = 1$, and $\frac{\overline{a}_k}{\underline{a}_k} = \frac{C+\delta}{u_k \underline{a}_{k-1}} \geq \frac{1+\beta}{1-\beta}$ in the case of $s_k = 0$. Thus, by the initial condition that $\varepsilon_0 \geq \frac{1+\beta}{1-\beta}$, we have $\frac{\overline{a}_k}{\underline{a}_k} \geq \frac{1+\beta}{1-\beta}$ for all $k \geq 1$.

By $\frac{\overline{a}_{k-1}}{\underline{a}_{k-1}} \geq \frac{1+\beta}{1-\beta}$, we have $u_k \underline{a}_{k-1} \leq C - \delta$ and $u_k \overline{a}_{k-1} \geq C + \delta$, which gives $\overline{a}_k = \overline{a}_{k-1}$ and $\frac{\underline{a}_k}{\underline{a}_{k-1}} = \frac{C-\delta}{u_k \underline{a}_{k-1}} \geq 1$ in the case of $s_k = 1$, and $\underline{a}_k = \underline{a}_{k-1}$ and $\frac{\overline{a}_k}{\overline{a}_{k-1}} = \frac{C+\delta}{u_k \overline{a}_{k-1}} \leq 1$ in the case of $s_k = 0$. Thus, $\{\underline{a}_k\}$ is monotonically increasing and $\{\overline{a}_k\}$ is monotonically decreasing.

Furthermore, by $\frac{\underline{a}_k}{\underline{a}_{k-1}} \geq 1$ and $\frac{\overline{a}_{k-1}}{\overline{a}_k} \geq 1$, we obtain $\frac{\underline{a}_k \overline{a}_{k-1}}{\overline{a}_k \underline{a}_{k-1}} \geq 1$, i.e., $\frac{\overline{a}_{k-1}}{\underline{a}_{k-1}} \geq \frac{\overline{a}_k}{\underline{a}_k}$. Hence, $\left\{\frac{\overline{a}_k}{\underline{a}_k}\right\}$ is monotonically decreasing.

The dynamic expression (9.9) can be modified as

$$e_k = \frac{1}{2}(1-\beta)\, e_{k-1} + \beta \overline{a}_{k-1}, \tag{9.10}$$

or

$$e_k = \frac{1}{2}(1+\beta)\, e_{k-1} + \beta \underline{a}_{k-1}. \tag{9.11}$$

By taking $k \to \infty$ on both sides of (9.10) and (9.11), we obtain $\overline{a}(\infty) = \frac{C+\delta}{2\delta} e(\infty)$ and $\underline{a}(\infty) = \frac{C-\delta}{2\delta} e(\infty)$. This leads to $\lim_{k\to\infty} \frac{\overline{a}_k}{\underline{a}_k} = \frac{1+\beta}{1-\beta}$.

(3) From (9.9) it follows that the uncertainty is reducible if and only if

$$\beta(\overline{a}_{k-1} + \underline{a}_{k-1}) < e_{k-1} = \overline{a}(k-1) - \underline{a}_{k-1}.$$

This is equivalent to

$$\frac{\overline{a}_{k-1}}{\underline{a}_{k-1}} > \frac{1+\beta}{1-\beta}.$$

□

Theorem 9.12. *Let*

$$\alpha_1 = \frac{1}{2}(1-\beta), \quad \alpha_2 = \frac{1}{2}(1+\beta).$$

Then under the conditions and notation of Theorem 9.11*,*

(1) *for $k \geq 1$, the optimal identification error e_k is bounded by*

$$\begin{aligned} \alpha_1^k e_0 + \beta \frac{a(1-\alpha_1^k)}{\alpha_2} &\leq \alpha_1^k e_0 + \beta \frac{\overline{a}_{k-1}(1-\alpha_1^k)}{\alpha_2} \\ &\leq e_k \leq \alpha_2^k e_0 + \beta \frac{\underline{a}_{k-1}(1-\alpha_2^k)}{\alpha_1} \\ &\leq \alpha_2^k e_0 + \beta \frac{a(1-\alpha_2^k)}{\alpha_1}; \end{aligned} \tag{9.12}$$

(2) *let $\varepsilon_0 = e_0/2$ and $\varepsilon_0 > \varepsilon > \frac{\beta a}{\alpha_1} = \frac{2\beta a}{1-\beta}$. Then the time complexity $N(\varepsilon)$ for reducing the uncertainty from ε_0 to ε is bounded by*

$$\left\lceil \frac{\log \frac{\varepsilon - \frac{\beta a}{\alpha_2}}{\varepsilon_0 - \frac{\beta a}{\alpha_2}}}{\log \alpha_1} \right\rceil \leq N \leq \left\lceil \frac{\log \frac{\varepsilon - \frac{\beta a}{\alpha_1}}{\varepsilon_0 - \frac{\beta a}{\alpha_1}}}{\log \alpha_2} \right\rceil;$$

(3) *there exists an irreducible relative error*

$$\frac{2\beta}{1+\beta} \leq \frac{e(\infty)}{a} \leq \frac{2\beta}{1-\beta}; \tag{9.13}$$

(4) *the parameter estimation error is bounded by*

$$0 \leq \frac{\overline{a}(\infty) - a}{\overline{a}(\infty)} \leq \frac{2\beta}{1+\beta}, \quad 0 \leq \frac{a - \underline{a}(\infty)}{\underline{a}(\infty)} \leq \frac{2\beta}{1-\beta}. \tag{9.14}$$

Proof. We prove the assertions step by step as follows.

(1) From (9.10) and the monotonically decreasing property of $\overline{a}_k$, we have

$$e_k \geq \alpha_1^k e_0 + \frac{\delta \overline{a}_{k-1}}{C} \sum_{i=0}^{k-1} \alpha_1^i,$$

and from (9.11) and the monotonically increasing property of $\underline{a}_k$,

$$e_k \leq \alpha_2^k e_0 + \frac{\delta \underline{a}_{k-1}}{C} \sum_{i=0}^{k-1} \alpha_2^i.$$

The results follow from $\sum_{i=0}^{k-1} \alpha_1^i = \frac{1-\alpha_1^k}{1-\alpha_1}$, $\sum_{i=0}^{k-1} \alpha_2^i = \frac{1-\alpha_2^k}{1-\alpha_2}$, $1-\alpha_1 = \alpha_2$, and $\underline{a}_k \leq a \leq \overline{a}_k$.

(2) From item (2) of Theorem 9.11, it follows that the error $e_k = \overline{a}_k - \underline{a}_k$ is monotonically decreasing. Thus, the upper bound on the time complexity is obtained by solving the inequality for the smallest N satisfying

$$e_N \leq \alpha_2^N \varepsilon_0 + \frac{\beta a(1-\alpha_2^N)}{\alpha_1} \leq \varepsilon.$$

Similarly, the lower bound can be obtained by calculating the largest N satisfying

$$\varepsilon \leq \alpha_1^N \varepsilon_0 + \frac{\beta a(1-\alpha_1^N)}{\alpha_2} \leq e_k.$$

(3) This follows from (9.12) and item (2) of Theorem 9.11, which implies the existence of $\lim_{t\to\infty} e_k$.

(4) From the last two lines of the proof of item (2) of Theorem 9.11, it follows that $\overline{a}(\infty) = \frac{C+\delta}{2\delta} e(\infty)$ and $\underline{a}(\infty) = \frac{C-\delta}{2\delta} e(\infty)$. This, together with (9.13), gives (9.14).

□

Remark 9.13. It is noted that the bounds in item (2) of Theorem 9.12 can be easily translated to sequential information bounds by replacing a with the on-line inequalities $\underline{a}_{k-1} \leq a \leq \overline{a}_{k-1}$.

Case 3: $y_k = au_k + \widetilde{\phi}_k'\widetilde{\theta}$

Let $u_k = \{u_\tau,\ \tau \le k\}$. Then $\|u_k\|_\infty$ is the maximum $|u_\tau|$ up to time k. Since we assume no information on $\widetilde{\theta}$, except that $\|\widetilde{\theta}\|_1 \le \eta$, it is clear that $\sup_{\|\widetilde{\theta}\|_1 \le \eta} |\widetilde{\phi}_k'\widetilde{\theta}| = \eta m_k$, where $m_k = \|\widetilde{\phi}_k\|_\infty$. Let $w_k = \widetilde{\phi}_k'\widetilde{\theta}$. Then

$$\{\widetilde{\phi}_k'\widetilde{\theta}:\ \|\widetilde{\theta}\|_1 \le \eta\} = \{w_k:\ |w_k| \le \eta m_k\}.$$

Theorem 9.14. *Suppose that $\underline{a}_0 > 0$, $u_{\max} \ge C/\underline{a}_0$, $\eta < \underline{a}_0$. Then*

(1) *the optimal input u_k, which minimizes the worst-case uncertainty at k, is given by the causal mapping from the available information at $k-1$:*

$$u_k = \frac{2C}{\underline{a}_{k-1} + \overline{a}_{k-1}}. \tag{9.15}$$

The optimal identification error at k satisfies the iteration equation

$$e_k = \frac{1}{2}e_{k-1} + \frac{\eta m_k}{2C}(\overline{a}_{k-1} + \underline{a}_{k-1}), \tag{9.16}$$

where $\overline{a}_k$ and $\underline{a}_k$ are updated by the rules

$$\begin{aligned} \overline{a}_k &= \overline{a}_{k-1}, \quad \underline{a}_k = \frac{C - \eta m_k}{u_k}, \quad \text{if } s_k = 1, \\ \underline{a}_k &= \underline{a}_{k-1}, \quad \overline{a}_k = \frac{C + \eta m_k}{u_k}, \quad \text{if } s_k = 0; \end{aligned}$$

(2) *the uncertainty is reducible if and only if $\overline{a}_{k-1} > \underline{a}_{k-1} + 2\eta$;*

(3) *for $k \ge 1$, the optimal identification error e_k is bounded by*

$$\begin{aligned} &\left(\prod\nolimits_{j=1}^{k}\beta_1(j)\right)e_0 + \frac{\eta a}{C}\sum_{i=1}^{k} m_i \prod\nolimits_{j=i+1}^{k}\beta_1(j) \\ &\quad \le e_k \le \left(\prod\nolimits_{j=1}^{k}\beta_2(j)\right)e_0 + \frac{\eta a}{C}\sum_{i=1}^{k} m_i \prod\nolimits_{j=i+1}^{k}\beta_2(j), \end{aligned} \tag{9.17}$$

where $\beta_1(k) = \frac{1}{2}\left(1 - \frac{\eta m_k}{C}\right)$ and $\beta_2(k) = \frac{1}{2}\left(1 + \frac{\eta m_k}{C}\right)$;

(4) *let $\varepsilon_0 = e_0/2$ and $\varepsilon_0 > \varepsilon > \frac{2\eta\overline{a}(0)}{\underline{a}_0 - \eta}$. Also, denote $\beta_1 = \frac{1}{2}\left(1 - \frac{\eta}{\underline{a}_0}\right)$, $\beta_2 = \frac{1}{2}\left(1 + \frac{\eta}{\underline{a}_0}\right)$. Then the time complexity $N(\varepsilon)$ for reducing the uncertainty from ε_0 to ε is bounded by*

$$\left\lceil \frac{\log \frac{\varepsilon - \frac{\eta a}{\overline{a}_0\beta_2}}{\varepsilon_0 - \frac{\eta a}{\overline{a}_0\beta_2}}}{\log \beta_1} \right\rceil \le N(\varepsilon) \le \left\lceil \frac{\log \frac{\varepsilon - \frac{\eta a}{\underline{a}_0\beta_1}}{\varepsilon_0 - \frac{\eta a}{\underline{a}_0\beta_1}}}{\log \beta_2} \right\rceil. \tag{9.18}$$

Proof. The proof is arranged as follows.

(1) The results follow from the definition of m_k and Theorem 9.12, with δ replaced by ηm_k.

(2) From (9.16) and (9.15), it follows that the uncertainty is reducible if and only if $\frac{\eta m_k}{u_k} < \frac{1}{2}e_{k-1} = \frac{1}{2}(\overline{a}_{k-1} - \underline{a}_{k-1})$. This is equivalent to $\eta < \frac{1}{2}(\overline{a}_{k-1} - \underline{a}_{k-1})$ or $\overline{a}_{k-1} > \underline{a}_{k-1} + 2\eta$, since $\frac{m_k}{u_k} \geq 1$.

(3) By (9.16), we have

$$e_k = \frac{1}{2}\left(1 + \frac{\eta_k}{C}\right) e_{k-1} + \frac{\eta m_k}{C}\underline{a}_{k-1} \tag{9.19}$$

and

$$e_k = \frac{1}{2}\left(1 - \frac{\eta m_k}{C}\right) e_{k-1} + \frac{\eta m_k}{C}\overline{a}_{k-1}. \tag{9.20}$$

Furthermore, from $\underline{a}_k \leq a \leq \overline{a}_k$ for all $k \geq 0$,

$$e_k \leq \beta_2(k)e_{k-1} + \frac{\eta m_k}{C}a$$

and

$$e_k \geq \beta_1(k)e_{k-1} + \frac{\eta m_k}{C}a.$$

Then, the inequalities in (9.17) can be obtained by iterating the above two inequalities in k.

(4) Since for all $k \geq 1$, $\overline{a}_0 \geq \overline{a}_k \geq \underline{a}_k \geq \underline{a}_0$,

$$u_k = \frac{2C}{\underline{a}_{k-1} + \overline{a}_{k-1}} \leq \frac{C}{\underline{a}_0},$$

which implies that $\frac{C}{\overline{a}_0} \leq u_k \leq \frac{C}{\underline{a}_0}$. This leads to

$$\beta_1(k) \geq \beta_1 = \frac{1}{2}(1 - \frac{\eta}{\underline{a}_0})$$

and

$$\beta_2(k) \leq \beta_2 = \frac{1}{2}\left(1 + \frac{\eta}{\underline{a}_0}\right).$$

Hence,

$$\beta_1 e_{k-1} + \frac{\eta a}{\overline{a}_0} \leq e_k \leq \beta_2 e_{k-1} + \frac{\eta a}{\underline{a}_0} \quad \text{for all } k \geq 1. \tag{9.21}$$

As a result, the inequalities of Theorem 9.12 can be adopted here to get (9.18).

□

Note that $\beta_2(k) \geq \beta_1(k)$ and $\beta_1(k) + \beta_2(k) = 1$; and $\beta_1 \to \beta_2$ as $\eta \to 0$, uniformly in k.

9.5 Identification Using Combined Deterministic and Stochastic Methods

This section highlights the distinctive underlying principles used in designing inputs and deriving posterior uncertainty sets in the stochastic and deterministic information frameworks.

In the deterministic worst-case framework, the information on noise is limited to its magnitude bound. Identification properties must be evaluated against worst-case noise sample paths. As shown earlier, the input is designed on the basis of choosing an optimal worst-case testing point (a hyperplane) for the prior uncertainty set. When the prior uncertainty set is large, this leads to an exponential rate of uncertainty reduction. However, when the uncertainty set is close to its irreducible limits due to disturbances or unmodeled dynamics, its rate of uncertainty reduction decreases dramatically due to its worst-case requirements. Furthermore, when the disturbance magnitude is large, the irreducible uncertainty will become too large for identification error bounds to be practically useful.

In contrast, in a stochastic framework, noise is modeled by a stochastic process and identification errors are required to be small with a large probability. Binary sensor identification in this case relies on the idea of averaging. Typically, in identification under stochastic setting, the input is designed to provide sufficient excitation for asymptotic convergence, rather than fast initial uncertainty reduction. Without effective utilization of prior information in designing the input during the initial time interval, the initial convergence can be slow. This is especially a severe problem in binary sensor identification since a poorly designed input may result in a very imbalanced output of the sensor in its 0 or 1 values, leading to a slow convergence rate. In the case of large prior uncertainty, the selected input may result in nonswitching at the output, rendering the stochastic binary-sensor identification inapplicable. On the other hand, averaging disturbances restores estimation consistency and overcomes a fundamental limitation of the worst-case identification.

Consequently, it seems a sensible choice of using the deterministic framework initially to achieve fast uncertainty reduction when the uncertainty set is large, then using the stochastic framework to modify estimation consistency. In fact, we shall demonstrate by an example that these two frameworks complement each other precisely, in the sense that when one framework fails, the other starts to be applicable.

9.5.1 Identifiability Conditions and Properties under Deterministic and Stochastic Frameworks

We first establish identifiability conditions of the two frameworks for a gain system

$$y_k = au_k + d_k, \quad k = 1, 2, \ldots, \tag{9.22}$$

where $\{d_k\}$ is a sequence of disturbances, and a is an unknown parameter. The prior information on a is given by $a \in [\underline{a}, \overline{a}]$, with $0 < \underline{a} \leq \overline{a} < \infty$. $u_k > 0$ is the input. The output y_k is measured by a binary-valued sensor with threshold C.

Deterministic Framework.

The idea of deterministic framework is to reduce the parameter uncertainty based on the bound of disturbances. Denote $r_k = \underline{a}_k/\overline{a}_k$ as the relative error.

Starting from the initial uncertainty $\Omega_0 = [\underline{a}, \overline{a}]$ and input $u_0 = u^*$, we check the output of binary sensor. If $s_1 = 0$, which means $au^* + d_1 > 0$, we obtain $au^* + \delta \geq au^* + d_1 > C$. Hence, $a > (C - \delta)/u^*$ and $e_1 = \overline{a} - (C - \delta)/u^* < e_0$. This means the parameter bound is reducible if $\underline{a} < (C - \delta)/u^*$. Otherwise, we have $s_1 = 1$. Then, $au^* - \delta \leq au^* + d_1 \leq C$; hence, $a \leq (C + \delta)/u^*$ and $e_1 = (C + \delta)/u^* - \underline{a} < e_0$ if $\overline{a} > (C + \delta)/u^*$. So, the parameter bound is reducible if

$$\underline{a} < \frac{C - \delta}{u^*} \quad \text{and} \quad \overline{a} > \frac{C + \delta}{u^*}$$

in the worst-case sense, or equivalently,

$$r_0 < \frac{C - \delta}{C + \delta} := \Delta. \tag{9.23}$$

Furthermore, by the above analysis, we arrive at the new uncertainty set

$$e_1 = \max\left\{\overline{a} - \frac{C - \delta}{u^*}, \frac{C + \delta}{u^*} - \underline{a}\right\}$$

in the worst-case sense. The uncertainty set is minimized at the optimal input

$$u_1^* = \frac{2C}{\underline{a} + \overline{a}} \quad \text{and} \quad e_1^* = \frac{(1 + \beta)\overline{a} - (1 - \beta)\underline{a}}{2} \tag{9.24}$$

with $\beta = \delta/C$.

The one-step optimal input design and parameter error was introduced in [111]. This, however, is not an overall optimal design if N steps are considered. The N-step optimal input design was developed in [14].

Theorem 9.15 [14]. *For binary observations with threshold C, the optimal parameter bound is*

$$e_N^* = 2\beta \frac{\overline{a}(1 + \beta)^{(2^N - 1)} - \underline{a}(1 - \beta)^{(2^N - 1)}}{(1 + \beta)^{(2^N)} - (1 - \beta)^{(2^N)}} \tag{9.25}$$

and the optimal inputs are

$$u_k^* = \frac{C}{\widetilde{a}_{k|N}}, \quad k = 1, 2, \ldots, N,$$

where

$$\widetilde{a}_{k|N} = \frac{\overline{a}_{k-1}(1+\beta)^{(2^{N-k}-1)} + \underline{a}_{k-1}(1-\beta)^{(2^{N-k}-1)}}{(1+\beta)^{(2^{N-k})} + (1-\beta)^{(2^{N-k})}}.$$

Example 9.16. For system (9.22) with $C = 40$, $\delta = 0.5$, $\underline{a} = 1$, and $\overline{a} = 10$, the optimal error provided by (9.25) is shown in Figure 9.1. It is shown that at first the uncertainty is reduced very fast, but uncertainty reduction gradually slows down toward an irreducible error bound.

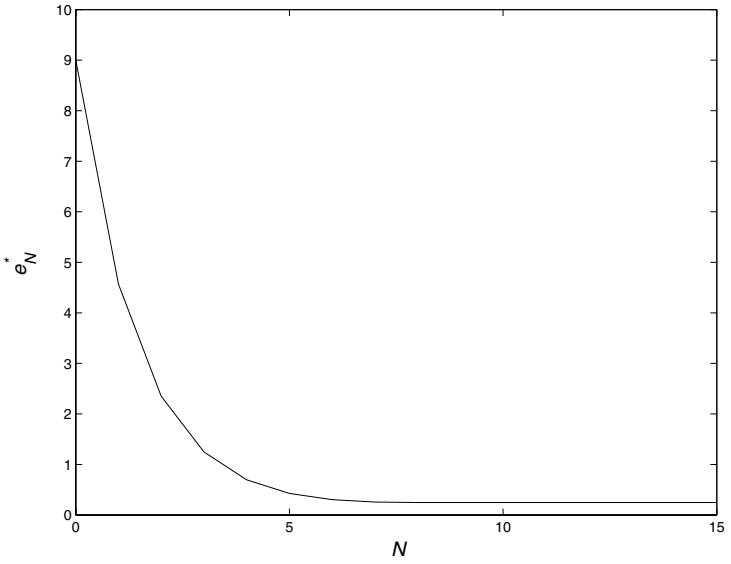


FIGURE 9.1. Optimal parameter error

Stochastic Framework

The essence of stochastic framework is to utilize the probabilistic properties of disturbances. Define the empirical measure $\xi_N^0 = \sum_{k=1}^N s_k/N$. If there exists u^* such that $C - au^*$ is on the support of $F(\cdot)$, which means

$$-\delta < C - au^* < \delta, \tag{9.26}$$

then ξ_N^0 is the empirical measure of $F(\cdot)$ at $C - au^*$, and

$$\xi_N^0 \to F(C - au^*), \quad \text{w.p.1.} \tag{9.27}$$

(A9.2) The noise $\{d_k\}$ is a sequence of i.i.d. random variables bounded by $|d_k| \le \delta$ whose distribution function $F(x)$, $x \in (-\delta, \delta)$, and its inverse $F^{-1}(\cdot)$ are twice continuously differentiable in $(-\delta, \delta)$ and known.

Throughout the rest of the chapter, we assume Assumption (A9.2) also holds. Note that F is a monotone function in view of Assumption (A9.2). If a is bounded, then there exists $z > 0$ such that $p = F(C - au^*)$ is bounded by

$$z < p < 1 - z. \tag{9.28}$$

Since $F(\cdot)$ is not invertible at 0 and 1, we modify ξ_N^0 to avoid these points as in (3.1):

$$\xi_N = \begin{cases} \xi_N^0, & \text{if} \quad z \le \xi_N^0 \le 1 - z, \\ z, & \text{if} \quad \xi_N^0 < z, \\ 1 - z, & \text{if} \quad \xi_N^0 > 1 - z. \end{cases} \tag{9.29}$$

As shown in Chapter 3, $\xi_N \to p$ w.p.1. Define

$$\widehat{a}_N = (C - F^{-1}(\xi_N))/u^*. \tag{9.30}$$

Then

$$\widehat{a}_N \to a \text{ w.p.1.}$$

For $a \in [\underline{a}, \overline{a}]$, the identifiability condition (9.26) becomes

$$-\delta < C - \overline{a}u^* \le C - \underline{a}u^* < \delta.$$

So for a given threshold C, u^* can be chosen to construct the estimation algorithm if and only if

$$r_0 > \Delta, \tag{9.31}$$

which complements exactly (9.23) for the deterministic framework. Under (9.31) and $C > \delta$, the admissible input set is

$$u^* \in \Gamma = \left(\frac{C - \delta}{\underline{a}}, \frac{C + \delta}{\overline{a}} \right). \tag{9.32}$$

By Chapter 6, for a given u^*, the optimal CR lower bound with binary-valued observations is

$$\eta_N^*(a, u^*) = E\left(\widehat{a}_N^* - a\right)^2 = \frac{F(C - au^*)(1 - F(C - au^*))}{N(u^*)^2 f^2(C - au^*)} \tag{9.33}$$

and $N(\eta_N - \eta_N^*(a, u^*)) = N[E(\widehat{a}_N - a)^2 - \eta_N^*] \to 0$ as $N \to \infty$, which means the algorithm (9.30) of the stochastic framework is asymptotically efficient.

Remark 9.17. The foregoing analysis indicates that the identifiability condition for the deterministic framework is that $r_0 < \Delta$ in the worst case and $r_0 > \Delta$ for the stochastic framework. Due to the strict inequalities, there is a dividing line $r_0 = \Delta$ between the two frameworks under binary observations. A key problem in combining the two frameworks is to find a way to connect the two sets of identifiability regions.

9.5.2 Combined Deterministic and Stochastic Identification Methods

In this subsection, we introduce a method to connect the two frameworks and develop the criteria for switching from one framework to another.

Connection of Two Frameworks by Input Design

Since there is no intersection between the two identifiability sets (9.23) and (9.31), one cannot design a strategy to switch from one framework to another. Consequently, it is necessary to find an approach to connect the sets. Here, we modify the stochastic methods by using two input values, rather than one. Since each input value creates one identifiability set, by choosing the inputs appropriately, we can create a scenario that these two sets collectively intersect to the identifiability set of the deterministic method.

For the initial uncertainty $[\underline{a}, \overline{a}]$, let $b \in (\underline{a}, \overline{a})$. Then, b divides the interval into two parts, $[\underline{a}, b]$ and $(b, \overline{a}]$. For $a \in [\underline{a}, b]$, the identifiability condition (9.31) becomes $\underline{a}/b > \Delta$. Similarly, for $a \in (b, \overline{a}]$, the requirement is $b/\overline{a} > \Delta$. Since

$$\max_b \min \left\{ \frac{\underline{a}}{b}, \frac{b}{\overline{a}} \right\} = \sqrt{\frac{\underline{a}}{\overline{a}}}$$

with $b^* = \sqrt{\underline{a}\overline{a}}$, the parameter can be estimated if $r_0 > \Delta^2$. Since $\Delta = (C - \delta)/(C + \delta) < 1$, we have $\Delta^2 < \Delta$; thus, there is an intersection between the identifiability sets of two frameworks.

This analysis indicates that it is possible to connect the two frameworks if two input values are used for the stochastic framework. We discuss next the switching strategies. This will be done by using convergence speeds. We first use an example to illustrate the basic ideas.

Example 9.18. Consider the one-step optimal worst-case error in Theorem 9.15

$$\frac{e_1^*}{e_0^*} = \frac{(1+\beta)\overline{a} - (1-\beta)\underline{a}}{2e_0} = \frac{1-\beta}{2} + \frac{\beta}{1-r_0},$$

which decreases with r_0. For the same system as in Example 9.16, the ratio is plotted as a function of r_0 in Figure 9.2 with $\beta = 0.2$. We can see that the ratio goes to 1 when r_0 approaches $(1-\beta)/(1+\beta)$, which means the uncertainty is almost irreducible.

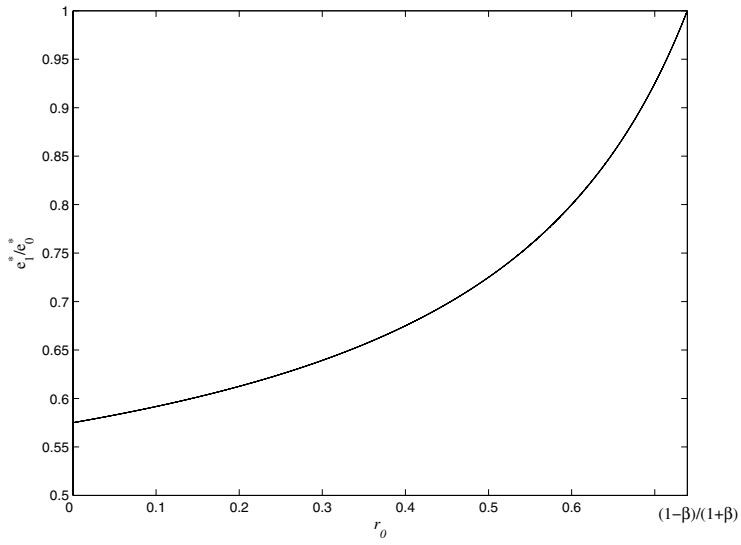


FIGURE 9.2. Optimal parameter reduction ratio

Consider the identifiability condition of the stochastic framework $r_0 > \Delta$. The convergence speed may be slow in the worst case as r_0 is small and close to Δ. The optimal covariance of the stochastic method with threshold C and constant input u^* is

$$\eta_N^*(a,u) = \frac{F(C-au^*)(1-F(C-au^*))}{N(u^*)^2 f^2(C-au^*)}.$$

Let

$$\eta^*(\Omega_0,u^*) = \sup_{a\in\Omega_0} \frac{F(C-au^*)(1-F(C-au^*))}{(u^*)^2 f^2(C-au^*)},$$

and

$$\eta^*(\Omega_0) = \inf_{u^*} \eta^*(\Omega_0,u^*). \tag{9.34}$$

Then, the optimal convergence speed by designing an optimal input value can be derived as

$$\eta_N^*(\Omega_0) = \eta^*(\Omega_0)/N. \tag{9.35}$$

For Ω_0, if we use the $\eta_N^*(\Omega_0)$, we can first design identification algorithms for $\Psi_1 = [\underline{a}, b^*]$ and $\Psi_2 = [b^*, \overline{a}]$, and then calculate $\eta_N^*(\Psi_1)$ and $\eta_N^*(\Psi_2)$. Hence, the switch time N_s can be decided by the following rule:

$$N_s = \min_N \left\{ \varepsilon_{2N}^2 > \min\{\eta_N^*([\underline{a}, \sqrt{\underline{a}\overline{a}})), \eta_N^*([\sqrt{\underline{a}\overline{a}}, \overline{a}])\} \right\}. \tag{9.36}$$

With this switching rule, the joint identification algorithm can be constructed as follows:

1. In the case of $r_k < \Delta^2$, apply the deterministic method.
2. Denote the first time that $r_k \geq \Delta^2$ as K, and calculate N_s by (9.36) with the information of parameter uncertainty lower and upper bounds at that time K.
3. Keep using deterministic methods for another $N_s - 1$ steps. Then get the parameter lower and upper bounds, namely, $\underline{a}_s$ and $\overline{a}_s$.
4. Switch to the stochastic method.

9.5.3 Optimal Input Design and Convergence Speed under Typical Distributions

We now solve (9.34) concretely for some typical noise distribution functions. We will derive specific expressions for the uniform distribution and truncated normal distribution. For other distributions, similar methods can be used, although derivation details may vary. For simplification, let $\eta_N^* = \eta_N^*(\Omega_0)$ and $\eta^*(u^*) = \eta^*(\Omega_0, u^*)$.

Uniform Distribution
Suppose that the density function of d_k is $f(x) = 1/(2\delta)$ for the support set (i.e., strictly positive) in $(-\delta, \delta)$. Then, $F(x) = \frac{\delta+x}{2\delta}$. We have

$$\begin{aligned} \eta^*(u^*) &= \sup_{a\in\Omega_0} \{\delta^2 - (C - au^*)^2\}, \\ &= \begin{cases} \delta^2, & \text{if } u^* \in \Gamma_1 = (\frac{C}{\overline{a}}, \frac{C}{\underline{a}}), \\ \delta^2 - (C - \overline{a}u^*)^2, & \text{if } u^* < \frac{C}{\overline{a}}, \\ \delta^2 - (C - \underline{a}u^*)^2, & \text{if } u^* > \frac{C}{\underline{a}}. \end{cases} \end{aligned} \tag{9.37}$$

Theorem 9.19. *Suppose d_k has a uniform distribution on $(-\delta, \delta)$. Then for Ω_0, η^* defined in* (9.34) *can be expressed as*

$$\eta^* = \begin{cases} \frac{\delta^2\overline{a}^2}{(C+\delta)^2}, & \text{if } r \leq \frac{C}{C+\delta}, \\ \frac{\delta^2\underline{a}^2}{C^2}, & \text{if } r > \frac{C}{C+\delta} \text{ and } C > \delta, \\ \frac{(\overline{a}-\underline{a})[(C+\delta)\underline{a}-(C-\delta)\overline{a}]}{C+\delta}, & \text{if } r > \frac{C}{C+\delta} \text{ and } C \leq \delta, \end{cases} \tag{9.38}$$

and the optimal input can be derived concretely by the above cases, respectively.

Proof. By (9.32), the feasible input set is $u^* \in \Gamma$. The set is nonempty if and only if $r > \Delta$ in case of $C > \delta$.

Case (i): In case of $\Delta < r \le \frac{C-\delta}{C}$, since $\frac{C-\delta}{C} \le \frac{C}{C+\delta}$, we have $\frac{C-\delta}{\underline{a}} \ge \frac{C}{\overline{a}}$ and $\frac{C+\delta}{\overline{a}} \le \frac{C}{\underline{a}}$, namely, $\Gamma \subset \Gamma_1 = (\frac{C}{\overline{a}}, \frac{C}{\underline{a}})$. So for $\forall u^* \in \Gamma$, there exists $a \in \Omega_0$ such that $a = C/u^*$, which induces $\eta^*(u^*) = \delta^2$. Hence,

$$\eta^* = \inf_{u^* \in \Gamma} \frac{\delta^2}{(u^*)^2} = \frac{\delta^2 \overline{a}^2}{(C+\delta)^2}$$

with $u^* = (C+\delta)/\overline{a}$.

Case (ii): In case of $\frac{C-\delta}{C} < r \le \frac{C}{C+\delta}$, we have $\frac{C-\delta}{\underline{a}} < \frac{C}{\overline{a}}$ and $\frac{C+\delta}{\overline{a}} \le \frac{C}{\underline{a}}$. For $u^* \in \Gamma_2 = (\frac{C}{\overline{a}}, \frac{C+\delta}{\overline{a}})$, we have $\eta^*(u^*) = \delta^2$. So

$$\inf_{u^* \in \Gamma_2} \frac{\delta^2}{(u^*)^2} = \frac{\delta^2 \overline{a}^2}{(C+\delta)^2}.$$

For $u^* \in \Gamma_3 = (\frac{C-\delta}{\underline{a}}, \frac{C}{\overline{a}}]$, notice that $C - au^* > C - a\frac{C}{\overline{a}} \ge 0$, which means $a \le C/u^*$. So $\eta^*(u^*) = \delta^2 - (C - \overline{a}u^*)^2$ for $u^* \in \Gamma_3$. Since

$$\inf_{u^* \in \Gamma_3} \frac{\delta^2 - (C - \overline{a}u^*)^2}{(u^*)^2} = \inf_{u^* \in \Gamma_3} \left\{ \frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\overline{a}C}{u^*} - \overline{a}^2 \right\}$$

and $C > \delta$, $\frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\overline{a}C}{u^*} - \overline{a}^2$, as a function of $1/u^*$, is symmetric about $1/u^* = \frac{\overline{a}C}{C^2 - \delta^2}$.

Since

$$r = \frac{\underline{a}}{\overline{a}} \le \frac{C}{C+\delta} \le \frac{C^2 + \delta^2}{C(C+\delta)},$$

we have

$$\left(\frac{\underline{a}}{C-\delta} - \frac{\overline{a}C}{C^2 - \delta^2} \right) - \left(\frac{\overline{a}C}{C^2 - \delta^2} - \frac{\overline{a}}{C} \right) = \frac{\underline{a}C(C+\delta) - \overline{a}(C^2 + \delta^2)}{C(C^2 - \delta^2)} \le 0.$$

As a result,

$$\frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\overline{a}C}{u^*} - \overline{a}^2$$

is minimized at $u^* = C/\overline{a}$ on Γ_3, namely,

$$\inf_{u^* \in \Gamma_3} \left\{ \frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\overline{a}C}{u^*} - \overline{a}^2 \right\} = \frac{\delta^2 \overline{a}^2}{C^2}.$$

Hence, we have

$$\eta^* = \min \left\{ \frac{\delta^2 \overline{a}^2}{(C+\delta)^2}, \frac{\delta^2 \overline{a}^2}{C^2} \right\} = \frac{\delta^2 \overline{a}^2}{(C+\delta)^2}.$$

Case (iii): In the case of $r > \frac{C}{C+\delta}$, we have $\frac{C-\delta}{\underline{a}} < \frac{C}{\overline{a}}$ and $\frac{C+\delta}{\overline{a}} > \frac{C}{\underline{a}}$. For $u^* \in \Gamma_1$, we have $\eta^*(u^*) = \delta^2$ and $\eta^*(u^*) = \delta^2 - (C - \overline{a}u^*)^2$ for $u^* \in \Gamma_3$. So

$$\inf_{u^* \in \Gamma_1} \frac{\delta^2}{(u^*)^2} = \frac{\delta^2 \underline{a}^2}{C^2} \quad \text{and} \quad \inf_{u^* \in \Gamma_3} \frac{\delta^2 - (C - \overline{a}u^*)^2}{(u^*)^2} = \frac{\delta^2 \overline{a}^2}{C^2}.$$

For $u^* \in \Gamma_4 = [C/\underline{a}, \frac{C+\delta}{\overline{a}})$, note that $C - au^* \leq C - a\frac{C}{\underline{a}} \leq 0$, which means $a \geq C/u^*$. So $\eta^*(u^*) = \delta^2 - (C - \underline{a}u^*)^2$ for $u^* \in \Gamma_4$. The minimization problem is

$$\inf_{u^* \in \Gamma_4} \frac{\delta^2 - (C - \underline{a}u^*)^2}{(u^*)^2} = \inf_{u^* \in \Gamma_4} \left\{ \frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\underline{a}C}{u^*} - \underline{a}^2 \right\}.$$

Since $C > \delta$, $\frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\underline{a}C}{u^*} - \underline{a}^2$, as a function of $1/u^*$, is symmetric about $1/u^* = \frac{\underline{a}C}{C^2 - \delta^2}$. Since

$$r > \frac{C}{C+\delta} \geq \frac{C-\delta}{C} \geq \frac{C(C-\delta)}{C^2 + \delta^2},$$

we have

$$\begin{aligned} &\left(\frac{\underline{a}}{C} - \frac{\underline{a}C}{C^2 - \delta^2} \right) - \left(\frac{\underline{a}C}{C^2 - \delta^2} - \frac{\overline{a}}{C+\delta} \right) \\ &= \frac{\overline{a}C(C-\delta) - \underline{a}(C^2 + \delta^2)}{C(C^2 - \delta^2)} \geq 0. \end{aligned}$$

As a result,

$$\frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\underline{a}C}{u^*} - \underline{a}^2$$

is minimized at $u^* = C/\underline{a}$ on Γ_4, namely,

$$\inf_{u^* \in \Gamma_4} \left\{ \frac{\delta^2 - C_1^2}{(u^*)^2} + \frac{2\underline{a}C}{u^*} - \underline{a}^2 \right\} = \frac{\delta^2 \underline{a}^2}{C^2}.$$

Hence, $\eta^* = \delta^2 \underline{a}^2 / C^2$.

The proof for $C < \delta$ is similar and omitted. □

Truncated Normal Distribution

Suppose d_k has a truncated normal distribution with probability density function

$$f_\sigma(x) = \frac{\frac{1}{\sigma}\lambda\left(\frac{x}{\sigma}\right)}{\Lambda\left(\frac{\delta}{\sigma}\right) - \Lambda\left(\frac{-\delta}{\sigma}\right)},$$

where $x \in (-\delta, \delta)$, $\lambda(\cdot)$ is the probability density function of the standard normal distribution, and $\Lambda(\cdot)$ its cumulative distribution function. Here,

we discuss the case of $\sigma = 1$; general cases can be derived similarly. Then, we have the density function

$$f(x) = \frac{\lambda(x)}{\Lambda(\delta) - \Lambda(-\delta)}$$

and the distribution function given by

$$F(x) = \frac{\Lambda(x) - \Lambda(-\delta)}{\Lambda(\delta) - \Lambda(-\delta)}.$$

Denote

$$\lambda_1(x) = \lambda(x)(1 - 2\Lambda(x)),$$
$$\lambda_2(x) = (\Lambda(x) - \Lambda(-\delta))(\Lambda(\delta) - \Lambda(x)),$$

and

$$G(x) = \frac{\lambda_2(x)}{\lambda^2(x)}.$$

Hence,

$$\eta^*(u^*) = \sup_{C - \bar{a}u^* \le x \le C - \underline{a}u^*} G(x)$$

and

$$\eta^* = \inf_{u^*} \eta^*(u^*).$$

First, we analyze the property of $G(x)$. The derivative of $G(x)$ can be written as

$$G'(x) = \frac{\lambda_1(x) + 2x\lambda_2(x)}{\lambda^2(x)}.$$

Let

$$g_1(x) = \lambda_1(x) + 2x\lambda_2(x).$$

Then, we have $g_1(0) = 0$ and $g_1(\delta) = \lambda(\delta)(1 - 2\Lambda(\delta)) < 0$.

Note that

$$g_2(x) = g_1'(x) = x\lambda_1(x) - 2\lambda^2(x) + 2\lambda(x).$$

Then

$$g_2(\delta) = \delta\lambda(\delta)(1 - 2\Lambda(\delta)) - 2\lambda^2(\delta) < 0$$

and

$$g_2(0) = 2\left(\Lambda(\delta) - \frac{1}{2}\right)^2 - 2\lambda^2(0) < 0$$

in the case of $\Lambda(\delta) < \frac{1}{2} + \lambda(0)$, and $g_2(0) \ge 0$ in the case of

$$\Lambda(\delta) \ge \frac{1}{2} + \lambda(0).$$

Lemma 9.20. *$g_2(x) < 0$ on $(0, \delta)$ for $\Lambda(\delta) \le \frac{1}{2} + \lambda(0)$. And for $\Lambda(\delta) > \frac{1}{2} + \lambda(0)$, there exists exactly one $x_2 \in (0, \delta)$ such that $g_2(x_2) = 0$, $g_2(x) > 0$ on $(0, x_2)$, and $g_2(x) < 0$ on (x_2, δ).*

Theorem 9.21. *$G'(x) < 0$ on $(0, \delta)$ for $\Lambda(\delta) \le \frac{1}{2} + \lambda(0)$. In addition, for $\Lambda(\delta) > \frac{1}{2} + \lambda(0)$, there exists $x_3 \in (0, \delta)$ such that $G_1'(x_3) = 0$, $G'(x) > 0$ on $(0, x_3)$, and $G'(x) < 0$ on (x_3, δ).*

Proof. Note that $G'(x) = g(x)/\lambda^2(x)$, so we need only prove the same conclusion for $g_1(x)$. By Lemma 9.20, $g_2(x) < 0$ on $(0, \delta)$ for $\Lambda(\delta) \le \frac{1}{2} + \lambda(0)$. In addition, $g_1(0) = 0$, and we have $g_1(x) < 0$ on $(0, \delta)$. For $\Lambda(\delta) > \frac{1}{2} + \lambda(0)$, $g_2(x) > 0$ on $(0, x_2)$ and $g_2(x) < 0$ on (x_2, δ) by Lemma 9.20, so $g_1(x_2) > g_1(0) = 0$. Since $g_1(\delta) < 0$ and $g_2(x) < 0$ on (x_2, δ), the second part is true. □

Here, we only derive the case of $C > \delta$, and $\Lambda(\delta) \le \frac{1}{2} + \lambda(0)$. We can discuss other cases similarly. Recall (9.32); the feasible input set is $u^* \in \Gamma = \left(\frac{C-\delta}{\underline{a}}, \frac{C+\delta}{\overline{a}}\right)$ and the set is nonempty if and only if $r > \Delta$. Then, we have the following theorem.

Theorem 9.22 *Suppose d has a truncated normal distribution on $(-\delta, \delta)$. Then for Ω_0, η^* defined in* (9.34) *can be expressed as*

$$\eta^* = \begin{cases} \frac{G(0)\overline{a}^2}{(C+\delta)^2}, & \text{if } \Delta < r \le \frac{C-\delta}{C}, \\ \min\{\frac{G(0)\overline{a}^2}{(C+\delta)^2}, H(C - \frac{C-\delta}{\underline{a}}\overline{a})\overline{a}^2\}, & \text{if } \frac{C-\delta}{C} < r \le \frac{C}{C+\delta}, \\ \min\{\frac{G(0)\underline{a}^2}{C^2}, H(0)\overline{a}^2, H(C - \frac{C-\delta}{\underline{a}}\overline{a})\overline{a}^2, \underline{a}^2 H(C - \frac{C+\delta}{\overline{a}}\underline{a})\}, & \\ & \text{if } r > \frac{C}{C+\delta}, \end{cases} \tag{9.39}$$

where

$$H(t) = \frac{\lambda_2(t)}{(C-t)^2\lambda^2(t)}.$$

For system (9.22) with $C = 40$, $\underline{a} = 1$, $\overline{a} = 50$, and the actual parameter $a = 15$. The disturbance has a uniform distribution on $(-\delta, \delta)$ with $\delta = 6$, by the algorithm developed in Section 9.5.2:

We have $K = 2$, $\underline{a}_K = 9.8$, and $\overline{a}_K = 15$. Then, we calculate $N_s = 1$ by (9.36). We turn to stochastic method and get $\widetilde{a}$.

It is shown that the parameter uncertainty is reduced to a certain bound using the deterministic method in the first stage and convergent to the real parameter using the stochastic method afterwards.

9.6 Notes

The material in this chapter is derived mostly from [111]. This chapter presents input design, uncertainty reduction rates, and time complexity

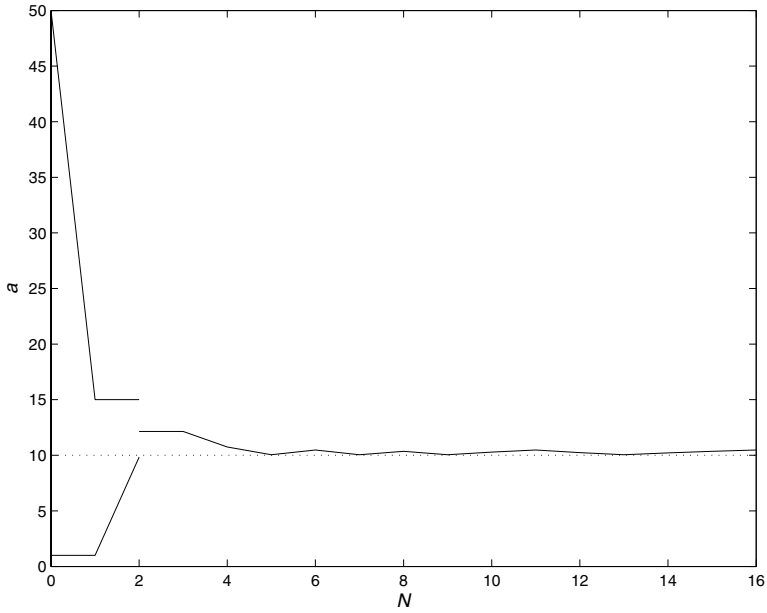


FIGURE 9.3. Simulation on the combined deterministic and stochastic identification methods

for system identification under binary-valued observations. This chapter deals with nonstatistical information from the observed data. We show that to enhance the nonstatistical information, the input must be properly designed.

Our input design is based on the idea of one-step optimal design: From the current uncertainty set on the unknown parameter, we select the best input value of the input such that the next uncertainty set can be maximally reduced, assuming no further information toward the future. Casini, Garulli, and Vicino have shown in [14] that if one has additional information on the number N of remaining steps toward the end of the identification data window, a better input design can be achieved. A dynamic programming method was introduced to optimize such an input design. It can be shown that in that case, the one-step optimal input design employed in Section 9.4 is no longer optimal for this N-step optimal input design. On the other hand, to achieve convergence with growing data sizes (namely, $N \to \infty$, rather than a fixed integer from the outset), the one-step design is a simple choice to achieve exponential convergence toward the irreducible uncertainty set.

The deterministic approaches are subject to an irreducible identification error; hence convergence is lost. They work well when the magnitude of the noise error bounds is relatively small since the irreducible set is a function

of the size of the noise. Also, the input design can achieve exponential convergence rates toward the irreducible set, which is much faster than the polynomial rates of convergence in a stochastic framework. On the other hand, stochastic information in the data can produce a convergent estimator. A combined identification algorithm that employs the input design first to reduce the parameter uncertainty set exponentially, followed by a statistical averaging approach to achieve convergence with periodic inputs, seems to be the best choice in overcoming the shortcomings of each individual framework.

10

Worst-Case Identification Using Quantized Observations

In this chapter, the parameter identification problem under unknown-but-bounded disturbances and quantized output sensors is discussed. In Chapter 9, an input sequence in (9.5) was used to generate observation equations in which only one parameter appears, reducing the problem to the identification of gain systems. A more general input design method is introduced in this chapter to achieve parameter decoupling that transforms a multi-parameter model into a single-parameter model. The input sequence with the shortest length that accomplishes parameter decoupling is sought. Identification algorithms are introduced, and their convergence, convergence rates, and time complexity for achieving a predefined estimation accuracy are investigated.

Section 10.1 formulates the problem and derives a lower bound on identification errors. Section 10.2 studies the input design that achieves parameter decoupling. Parameter decoupling reduces the original identification problem to a one-parameter identification problem. Section 10.3 presents algorithms for identifying one parameter with quantized observations. Time-complexity issues are further studied in Section 10.4. Finally, Section 10.5 illustrates the identification algorithms and their convergence properties by several examples.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_10,

10.1 Worst-Case Identification with Quantized Observations

Consider an FIR system

$$y_k = \sum_{i=0}^{n_0-1} a_i u_{k-i} + d_k, \quad k = k_0, k_0+1, \ldots, \tag{10.1}$$

where $\{d_k\}$ is a sequence of disturbances, $\{a_i\}$ are unknown system parameters, and the input $u_k \in \mathbb{U} = \{u_k : 0 < u_k \le u_{\max}, k = k_0, k_0+1, \ldots\}$. The output y_k is measured by a quantized sensor with m_0 thresholds $C_1 < C_2 < \cdots < C_{m_0}$. Namely, the sensor $s = \mathcal{S}(y)$ is represented by the indicator function

$$s_k = \mathcal{S}(y_k) = \sum_{i=1}^{m_0} i I_{\{y_k \in (C_i, C_{i+1}]\}}, \tag{10.2}$$

where $i = 1, \ldots, m_0$ with $C_0 = -\infty$ and $C_{m_0+1} = \infty$. Although a sum is presented in (10.2), at any time k, only one term is nonzero. Hence, $s_k = j$, $j = 0, 1, \ldots, m_0$, implies that $y_k \in (C_j, C_{j+1}]$.

Define $\theta = [a_0, \ldots, a_{n_0-1}]'$. Then the system input–output relationship becomes

$$y_k = \phi_k' \theta + d_k, \tag{10.3}$$

where $\phi_k = [u_k, u_{k-1}, \ldots, u_{k-n_0+1}]'$. The system will be studied under the following assumptions.

(A10.1) For a fixed $p \ge 1$,

(i) the sequence of disturbances $d = \{d_k : k \ge 0\}$ is bounded by $\|d\|_\infty \le \delta$;

(ii) the prior information on θ is given by $\Omega_0 = \mathrm{Ball}_p(\theta_0, e_0) \subset \mathbb{R}^{n_0}$ for some known $\theta_0 \in \mathbb{R}^{n_0}$ and $e_0 > 0$.

For a selected input sequence u_k, let $s = \{s_k, k = k_0, \ldots, k_0 + N - 1\}$ be the observed output. Define

$$\Omega_N(k_0, u, s) = \Big\{\theta : \ s_k = \sum_{i=0}^{m_1} i I_{\{\phi_k'\theta + d_k \in (C_i, C_{i+1}]\}} \text{ for some } |d_k| \le \delta, \ k = k_0, \ldots, k_0 + N - 1\Big\},$$

and the optimal worst-case uncertainty after N steps of observations as

$$e_N = \inf_{\|u\|_\infty \le u_{\max}} \sup_{k_0} \sup_{s} \mathrm{Rad}_p(\Omega_N(k_0, u, s) \cap \mathrm{Ball}_p(\theta_0, e_0)).$$

Proposition 10.1. *Assuming Assumption* (A10.1)*, for* (10.3)*, the uncertainty set* $\text{Ball}_1(0, (C_1 - \delta)/u_{\max})$ *is not identifiable.*

Proof. For any $\theta \in \text{Ball}_p(0, (C_1 - \delta)/u_{\max})$, the output

$$\begin{aligned} y_k &= \phi_k'\theta + d_k \le \|\phi(k)\|_\infty \|\theta\|_1 + \delta \\ &\le u_{\max}\frac{C_1 - \delta}{u_{\max}} + \delta = C_1. \end{aligned}$$

It follows that $s_k = 1, \forall k$. Hence, the observations could not provide further information to reduce uncertainty. □

10.2 Input Design for Parameter Decoupling

In order to simplify the problem, an input design method was introduced in [111] to decouple system parameters for identification. We are seeking the shortest input sequence lengths to decouple (10.1) into n_0 single-parameter observation equations.

Definition 10.2. An input sequence $\{u_i, i = k_0, \ldots, k_0 + N_0 - 1\}$ is said to be n_0*-parameter-decoupling* if, for each $j = 0, \ldots, n_0 - 1$, there exists u_{k_j} such that $y_k = a_j u_{k_j} + d_k$ for some $k_0 \le k \le k_0 + N_0 - 1$ and without considering inputs before time k_0. In other words, the N_0 observation equations contain at least one single-parameter observation equation for each parameter a_j. The input sequence is called the shortest n_0-parameter-decoupling sequence if N_0 is minimal.

Example 10.3. The shortest n_0-parameter-decoupling input segment is not unique. For example, when $n_0 = 3$ and $k_0 = 0$, input $u = \{u_1, 0, 0, u_4, 0, u_6, 0\}$, we have

$$y_5 = a_1 u_4, \quad y_8 = a_2 u_6, \quad y_4 = a_3 u_1.$$

That is, u is a three-parameter-decoupling input segment. By exhaustive testing, we can verify that u is shortest. It can be easily checked that $\{0, u_2^*, 0, u_4^*, 0, 0, u_7^*\}$ is also three-parameter decoupling.

Since parameter decoupling is independent of the actual values of u_{k_j}, for simplicity we will always use $u_{k_j} = 1$ to represent the nonzero input value at k_j.

Definition 10.4. A vector is called a $\{0, 1\}$-vector if its components are 1 or 0.

Definition 10.5. A $\{0, 1\}$-vector is said to contain the jth row of the $n_0 \times n_0$ identity matrix if the jth row of the $n_0 \times n_0$ identity matrix is a block of it. In this case, the 1 in the block is said to map to the jth row of the $n_0 \times n_0$ identity matrix.

Definition 10.6. A {0,1}-vector is said to be complete if

(i) it contains every row of the $n_0 \times n_0$ identity matrix;

(ii) each 1 maps to at most one row of the $n_0 \times n_0$ identity matrix.

Two complete {0,1}-vectors are equivalent if they have the same length.

Definition 10.7. For a given $\{0,1\}$-vector $b = [b_1, b_2, \ldots, b_N]'$ and $c = [c_1, c_2, \ldots, c_N]'$, if $c_i = b_{N-i+1}$ for $i = 1, 2, \ldots, N$, then b and c are said to be converse to each other.

Definition 10.8. Assume a {0,1}-vector $b = (b_1, b_2, \ldots, b_N)$ is complete, and b_l maps to the first row of the $n_0 \times n_0$ identity matrix. If $c = (c_1, \ldots, c_N)$ satisfies $c_i = b_{l+i-1}$ for $i = 1, \ldots, N - n_0 + 1$ and $c_{N-l+j+2} = b_j$ for $j = 1, \ldots, l-1$, then the transfer from b to c is called initial-1.

Proposition 10.9. *For a complete* $\{0,1\}$*-vector* $b = (b_1, \ldots, b_N)$*, after converse and/or initial-*1 *transfers, the new vector is equivalent to* b*.*

Proof. Assume that $b_{k_i} (i = 1, \ldots, n_0)$ maps to the ith row of the $n_0 \times n_0$ identity matrix.

Converse: Denote $c = (b_N, \ldots, b_1)$ as the vector that is converse to b. By definition, b_{N-k_i+1} is the component of c that maps to the ith row of the $n_0 \times n_0$ identify matrix, so c has property (i) in Definition 10.6. Since $k_i \neq k_j$ for $i \neq j$, we have $N - k_i + 1 \neq N - k_j + 1$. Hence, c has property (ii). In addition, b and c have the same length. So b is equivalent to c.

It is similar to prove for the initial-1 transfer. □

Definition 10.10. Two 1's in a {0,1}-vector are called neighbors if there's no 1 between them.

Proposition 10.11. *For a complete* $\{0,1\}$*-vector, if there are more than* (n_0-1) *0s between two neighboring 1s, then keeping only* (n_0-1) *0s between them will not change its completeness.*

Lemma 10.12. *The shortest length of complete* $\{0,1\}$*-vectors is*

$$\nu(n_0) = \begin{cases} \frac{{n_0}^2+2n_0-1}{2}, & \text{if } n_0 \text{ is odd,} \\ \frac{{n_0}^2+2n_0-2}{2}, & \text{if } n_0 \text{ is even.} \end{cases} \tag{10.4}$$

Proof. By Propositions 10.11 and 10.9, any {0,1}-vector is equivalent to the one with first $\nu(n_0)$ components:

$$\overbrace{1,0,\ldots,0,1}^{n_0+1},\overbrace{0,1,0,\ldots,0,1}^{n_0+1},\ldots,\overbrace{0,\ldots,0,1,\underbrace{0,\ldots,0}_{n_0-l},1}^{n_0+1},\overbrace{0,\ldots,0}^{l-1}, \tag{10.5}$$

where $l = \frac{n_0+1}{2}$ (or $\frac{n_0}{2}$) when l is odd (or even). For $k \le n_0$, let

$$l_k = \begin{cases} \frac{k+1}{2}, & \text{if } k \text{ is odd,} \\ n_0 - \frac{k}{2} + 1, & \text{if } k \text{ is even.} \end{cases}$$

Then, the kth 1 in (10.5) maps to the l_kth row of the $n_0 \times n_0$ identity matrix.

Hence, the first to the $\nu(n_0)$th components of the vector in (10.5) is complete. □

Theorem 10.13. *For system* (10.1)*, the length of the shortest* n_0*-parameter-decoupling input segment is* $\nu(n_0)$*. Furthermore,*

(i) *if* n_0 *is even,* $u_k = 0$ *for all* k *except* $k = k_0 + i(n_0+2), k_0 + (i+1)(n_0+1) - 1$*, or, for all* k *except* $k = k_0 + \nu(n_0) - i(n_0+2) - 1, k_0 + \nu(n_0) - (i+1)(n_0+1)$*, where* $i = 0, 1, \ldots, \frac{n_0}{2} - 1$*;*

(ii) *if* n_0 *is odd,* $u_k = 0$ *for all except* $k = k_0 + i(n_0+2) + 1, k_0 + (i+1)(n_0+1)$*, and* $k_0 + \frac{n_0^2+n_0}{2}$*; or, all except* $k_0 + \nu(n_0) - i(n_0+2), k_0 + \nu(n_0) - (i+1)(n_0+1) + 1$*, and* $k_0 + \frac{n_0-1}{2}$*, where* $i = 0, 1, \ldots, \frac{n_0-3}{2}$*.*

Proof. Suppose $u = [u_{k_0}, u_{k_0+1}, \ldots, u_{k_0+N-1}]'$ is a shortest n_0-parameter-decoupling input segment. By definition, there exists k_1 such that

$$u_{k_1} \neq 0 \text{ and } u_k = 0 \quad \text{for} \quad k = k_1 - n_0 + 1, \ldots, k_1 - 1. \tag{10.6}$$

So we have $y_{k_1} = a_0 u_{k_1}$, and hence a_0 is decoupled.

Consider the nonzero components of u as 1, u becomes a $\{0,1\}$-vector. Then, (10.6) can be considered as the n_0th row of the $n_0 \times n_0$ identity matrix. Namely, the $\{0,1\}$-vector u must satisfy condition (i) of Lemma 10.12. Furthermore, by the definition of the shortest parameter-decoupling input vector, $k_i \neq k_j$ for $i \neq j$, So the $\{0,1\}$-vector u is required with condition (ii) of Lemma 10.12. Lemma 10.12 confirms the first part of Theorem 10.13. The second part follows the proof of Lemma 10.12. □

10.3 Identification of Single-Parameter Systems

In this section, the identification of single-parameter systems is studied. The conditions of identification using quantized sensors are given, and the effect of threshold values to identification is discussed.

Consider the single-parameter system

$$y_k = a u_k + d_k, \quad k = k_0, k_1 + 2, \ldots, \tag{10.7}$$

where $a \in [\underline{a}_0, \overline{a}_0]$ and $\underline{a}_0 > 0$, $d = \{d_{k_0}, d_{k_0+1}, \ldots\}$ is the sequence of disturbances satisfying $\|d\|_\infty \le \delta$, u_k is the input, and the output y_k is measured by the quantized sensor (10.2).

Remark 10.14. Proposition 10.1 confirms that the uncertainty is irreducible when $|a| < (C_1 - \delta)/u_{\max}$, but in order to investigate the relationship between identification and all threshold values, we assume $|a| > (C_{m_0} - \delta)/u_{\max}$. In this case, input $u_0 = \frac{C_{m_0}+\delta}{C_1-\delta} u_{\max}$. If $s(0) = 0$, $y(0) \le C_1$, which indicates $a > 0$; else, $s(0) \neq 0$ indicates $a < 0$. Thus, the sign of a is known. Without loss of generality, we assume $\underline{a}_0 > (C_{m_0} - \delta)/u_{\max}$.

For simplification, the following symbols will be used in this chapter:

1. $\underline{a}_k, \overline{a}_k$: the uncertainty upper and lower bounds of a at time k;
2. $e_k = \overline{a}_k - \underline{a}_k$;
3. $L_k = e_k/e_{k-1}$;
4. $r_k = \underline{a}_k/\overline{a}_k$;
5. $\Delta C_l = C_{l+1} - C_l$, $l = 0, \ldots, m_0 - 1$;
6. $\max \Delta C_l = \max_{1\le l\le m_0-1} \Delta C_l$ and $\min \Delta C_l = \min_{1\le l\le m_0-1} \Delta C_l$;
7. $R = \frac{C_1 - \max \Delta C_l - \delta}{C_{m_0} + \max \Delta C_l + \delta}$;
8. $P_k(i,j) = \frac{(C_i+\delta)\overline{a}_{k-1} - (C_j-\delta)\underline{a}_{k-1}}{(C_i+C_j)e_{k-1}}$, $i \le j$;
9. $Q_k(i,j) = \frac{\overline{a}_{k-1}(\max_{i\le l\le j-1} \Delta C_l + 2\delta)}{(C_j + \max_{i\le l\le j-1} \Delta C_l + \delta)e_{k-1}}$, $i \le j$;
10. $\vee\{x_1, x_2\} = \max\{x_1, x_2\}$ and $\wedge\{x_1, x_2\} = \min\{x_1, x_2\}$;
11. $L_k(i,j) = \wedge\{\vee\{P_k(i,j), Q_k(i,j)\}, 1\}$;
12. $\widetilde{L}_k(i) = \wedge\{P_k(i,i), 1\}$.

10.3.1 General Quantization

A quantized sensor is binary when $m_0 = 1$. By Chapter 9, when the decoupled observations (10.7) are measured by a binary sensor with threshold C_1, in the worst case, the minimum of L_k is $\widetilde{L}_k(1)$ and the optimal input is $u_k = \frac{2C_1}{\overline{a}_{k-1} + \underline{a}_{k-1}}$. Subsequently, we will consider $m_0 \ge 2$.

Theorem 10.15. *For* (10.7)*, when* $e_{k-1} < (\min \Delta C_l + 2)/u_{\max}$*, the minimum of* L_k *is* $L_k = \widetilde{L}_k(m_0)$ *in the worst-case sense.*

Proof. By (10.7), we have $a = (y_k - d_k)/u_k$. The necessary condition of $L_k < 1$ is that for all $u_k \in \mathbb{U}$, so there exists some threshold C_i such that

$$\underline{a}_{k-1} \le \frac{C_i - \delta}{u_k} \le \frac{C_i + \delta}{u_k} \le \overline{a}_{k-1}. \tag{10.8}$$

Suppose C_{i_0} satisfies (10.8), since

$$e_{k-1} < (\min \Delta C_l + 2\delta)/u_{\max} \le (C_{i_0} - C_{i_0-1} + 2\delta)/u_{\max},$$

we have

$$C_{i_0-1} - \delta < C_{i_0} + \delta - u_{\max}(\overline{a}_{k-1} - \underline{a}_{k-1}).$$

By (10.8), $C_{i_0} + \delta < \overline{a}_{k-1}u_{\max}$, so

$$\frac{C_{i_0-1} - \delta}{u_k} < \underline{a}_{k-1}.$$

Since $C_1 < C_2 < \cdots < C_{m_0}$, for $i < i_0$, C_i does not satisfy (10.8). Similarly, for $i > i_0$, C_i does not satisfy (10.8). Namely, for a given $u_k \in \mathbb{U}$, there exists one threshold at most, which satisfies (10.8). By Chapter 9, in a worst-case sense, the minimum of L_k is $\widetilde{L}_{i_0}(k)$ when only C_{i_0} satisfies (10.8), and the optimal input is $u_k = 2C_{i_0}/(\overline{a}_{k-1} + \underline{a}_{k-1})$.

Furthermore, since $\widetilde{L}_i(k)$ is monotonically decreasing for i and $C_1 < C_2 < \cdots < C_{m_0}$, the minimum of L_k is $\widetilde{L}_1(k)$ and the optimal input is $u_k = 2C_1/(\overline{a}_{k-1} + \underline{a}_{k-1})$. □

In order to design input with quantized sensors to make L_k less than $\widetilde{L}_k(C_1)$, we now describe how to identify systems with quantized observations.

Theorem 10.16. *Assume* $\delta < \min \Delta C_l/2$, $u_{\max} \ge (C_{m_0} - \delta)/\underline{a}_0$, *and* $e_{k-1} \ge (C_{m_0} - C_1 + 2\delta)/u_{\max}$. *Then*

$$L_k = L_k(1, m_0). \tag{10.9}$$

Furthermore,

(i) *if* $r_{k-1} \le R$, *then* $L_k = P_k(1, m)$ *and the optimal input is* $u_k = \frac{C_1 + C_{m_0}}{\overline{a}_{k-1} + \underline{a}_{k-1}}$;

(ii) *if* $r_{k-1} > R$, *then* $L_k = Q_k(1, m)$ *and the optimal input is* $u_k = \frac{C_{m_0} + \max \Delta C_+ \delta}{\overline{a}_{k-1}}$.

Proof. Since $e_{k-1} \ge (C_{m_0-1} - C_1 + 2\delta)/u_{\max}$, there exists $u_k \in \mathbb{U}$ such that

$$\frac{C_1 + \delta}{u_k} \ge \underline{a}_{k-1}, \quad \frac{C_{m_0} + \delta}{u_k} \le \overline{a}_k,$$

which together with $\delta < \min \Delta C_l/2$ gives

$$C_{l+1} - \delta \ge C_l + \delta, \quad \text{for} \quad l = 1, 2, \ldots, m_0 - 1.$$

Hence, there exists $u_k \in \mathbb{U}$ such that

$$\underline{a}_{k-1} \le \frac{C_1 - \delta}{u_k} \le \frac{C_1 + \delta}{u_k} \le \cdots \le \frac{C_{m_0} - \delta}{u_k} \le \frac{C_{m_0} + \delta}{u_k} \le \overline{a}_{k-1}. \tag{10.10}$$

By (10.4), $a = (y_k - d_k)/u_k$. For the input in (10.10), consider s_k: If $s_k = m_0$, then $a > (C_{m_0} - \delta)/u_k$, and hence,

$$\underline{a}_k = \vee\{(C_{m_0} - \delta)/u_k, \underline{a}_{k-1}\} = (C_{m_0} - \delta)/u_k, \quad \overline{a}_k = \overline{a}_{k-1}.$$

Denote $\gamma_k = u_k e_{k-1}$. Then, we have

$$L_k = (\overline{a}_{k-1}u_k - (C_{m_0} - \delta))/\gamma_k.$$

If $s_k = j$, $j = 1, \ldots, m_0 - 1$, then $(C_j - \delta)/u_k < a \le (C_{j+1} + \delta)/u_k$, and hence,

$$\underline{a}_k = (C_j - \delta)/u_k, \quad \overline{a}_k = (C_{j+1} + \delta)/u_k, \quad L_k = (\Delta C_j + 2\delta)/\gamma_k.$$

If $s_k = 0$, then $a \le (C_1 - \delta)u_k$, and hence,

$$\underline{a}_k = \underline{a}_{k-1}, \quad \overline{a}_k = (C_1 + \delta)/u_k, \quad L_k = (C_1 + \delta - \underline{a}_{k-1}u_k)/\gamma_k.$$

So, in the worst case, we have

$$\begin{cases} L_k \ge (\overline{a}_{k-1}u_k - (C_{m_0} - \delta))/\gamma_k, \\ L_k \ge (\Delta C_j + 2\delta)/\gamma_k, \\ L_k \ge (C_1 + \delta - \underline{a}_{k-1}u_k)/\gamma_k. \end{cases} \tag{10.11}$$

Since $u_k > 0$, (10.11) is equivalent to

$$\begin{cases} u_k & \le \frac{C_{m_0} - \delta}{\overline{a}_{k-1} - L_k e_{k-1}}, \\ u_k & \ge \frac{\Delta C_j + 2\delta}{L_k e_{k-1}}, \\ u_k & \ge \frac{C_1 + \delta}{\underline{a}_{k-1} + L_k e_{k-1}}, \end{cases}$$

which means that there exists $u_k \in \mathbb{U}$ satisfying (10.10) such that

$$L_k \ge P_k(1, m_0), \quad L_k \ge Q_k(1, m_0), \tag{10.12}$$

and the equalities in (10.12) are achieved in the case of

$$u_k = \frac{C_{m_0} + C_1}{\overline{a}_{k-1} + \underline{a}_{k-1}} \quad \text{and} \quad u_k = \frac{C_{m_0} + \max\{C_i - C_{i+1}\} + d}{\overline{a}_{k-1}},$$

respectively. So, (10.9) is true.

Furthermore, when $r_{k-1} \le R$, we have $P_k(1, m_0) \ge Q_k(1, m_0)$. Let $u_k = \frac{C_1 + C_{m_0}}{\overline{a}_{k-1} - \underline{a}_{k-1}}$. Then, by

$$u_k = \frac{C_1 + C_{m_0}}{\overline{a}_{k-1} + \underline{a}_{k-1}} \le \frac{C_{m_0} - \delta + C_1 + \delta}{2\underline{a}_{k-1}} \le \frac{C_{m_0} - \delta}{\underline{a}_{k-1}} \le u_{\max},$$

$u_k \in \mathbb{U}$. From (10.11), one can get $L_k = P_k(1, m_0)$.

Similarly, when $r_{k-1} > R$, let $u_k = \frac{C_{m_0} + \max \Delta C_l + \delta}{\overline{a}_{k-1}}$. Then, $u_k \in \mathbb{U}$ and $L_k = Q_k(1, m_0)$. □

Theorem 10.16 shows that when $r_{k-1} \le R$, which means that the uncertainty bound is large, $\overline{a}_{k-1} - (C_{m_0} - \delta)/u_k$ and $(C_1 + \delta)/u_k - \underline{a}_{k-1}$ are larger than $(\max \Delta C_l + 2\delta)/u_k$. So, L_k is determined by C_1 and C_{m_0}. When $\underline{a}_{k-1}/\overline{a}_{k-1} > R$, which means the uncertainty bound is small, $\overline{a}_{k-1} - (C_{m_0} - \delta)/u_k$ and $(C_1 + \delta)/u_k - \underline{a}_{k-1}$ are less than $(\max \Delta C_l + 2\delta)/u_k$. Then L_k is determined by the two neighboring thresholds with the maximum space between them. As a result, L_k derived from $m_0 - 1$ thresholds may be less than the one derived from m_0 thresholds.

Example 10.17. Let $C_1 = 85$, $C_2 = 94$, $C_3 = 97$, $C_4 = 100$, and $\delta = 1$. Compare L_k derived from $\{C_1, C_2, C_3, C_4\}$ and $\{C_2, C_3, C_4\}$.

1. For $\{C_1, C_2, C_3, C_4\}$, if $r_0 \le \frac{C_1 - (C_2 - C_1) - \delta}{C_4 + (C_2 - C_1) + \delta} = \frac{15}{22}$, then

$$L_1 = P_1(1, 4) = \frac{86\overline{a}_0 - 99\underline{a}_0}{185 e_0}.$$

If $r_0 > \frac{15}{22}$, then

$$L_1 = Q_1(1, 4) = \frac{\overline{a}_0}{10 e_0}.$$

2. For $\{C_2, C_3, C_4\}$, if $r_0 \le \frac{C_2 - (C_3 - C_2) - \delta}{C_4 + (C_3 - C_2) + \delta} = \frac{45}{52}$, then

$$L_1 = P_1(2, 4) = \frac{95\overline{a}_0 - 99\underline{a}_0}{194 e_0}.$$

If $r_0 > \frac{45}{52}$, then

$$L_1 = Q_1(2, 3) = \frac{\overline{a}_0 (C_4 - C_3 + 2\delta)}{(2C_4 - C_3 + \delta) e_0} = \frac{5\overline{a}_0}{104 e_0}.$$

The curves of L_1 with r_0 are plotted in Figure 10.1. It is shown that we have $L_1(1, 4) \le L_1(2, 4)$ as $r_0 \le 0.77$ and $L_1(1, 4) > L_1(2, 4)$ as $r_0 > 0.77$. The reason is that the space between C_1 and C_2 is larger than others, so $\max_{1 \le l \le 3} \Delta C_l > \max_{2 \le l \le 3} \Delta C_l$, which causes the result in Figure 10.1 by (10.9).

Before studying the general case, we start from the case $m_0 = 2$.

Corollary 10.18. *For system* (10.7) *with* $m_0 = 2$ *and* $u_k \in \mathbb{U}$, *in the worst case, the minimum value of* L_k *is*

$$L_k = \wedge\{\vee\{P_k(1, 2), Q_k(1, 2)\}, \widetilde{L}_k(1)\}. \tag{10.13}$$

Furthermore,

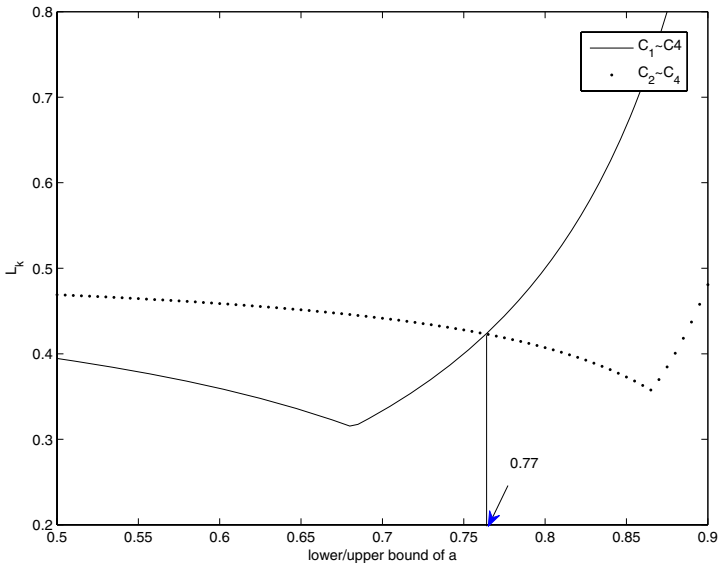


FIGURE 10.1. L_k derived from $C_1 \sim C_4$ and $C_2 \sim C_4$

(i) *if*

$$r_{k-1} \leq \frac{2C_1 - C_2 - \delta}{2C_2 - C_1 + \delta},$$

then $L_k = P_1(1,2)$ *and the optimal input is*

$$u_k = \frac{C_1 + C_2}{\overline{a}_{k-1} + \underline{a}_{k-1}};$$

(ii) *if*

$$\frac{2C_1 - C_2 - \delta}{2C_2 - C_1 + \delta} \leq r_{k-1} \leq \frac{C_1 - \delta}{2C_2 - C_1 + \delta},$$

then $L_k = Q_k(1,2)$ *and the optimal input is*

$$u_k = \frac{2C_2 - C_1 + \delta}{\overline{a}_{k-1}};$$

(iii) *if*

$$\frac{C_1 - \delta}{2C_2 - C_1 + \delta} < r_{k-1} \leq \frac{C_2 - \delta}{C_2 + \delta},$$

then $L_k = \widetilde{L}_k(1)$ *and the optimal input is*

$$u_k = \frac{2C_2}{\overline{a}_{k-1} + \underline{a}_{k-1}};$$

(iv) *if*

$$r_{k-1} > \frac{C_2 - \delta}{C_2 + \delta},$$

then $L_k = 1$ *for any* $u_k \in \mathbb{U}$.

Proof. Since $m_0 = 2$, we can construct inputs with one or two thresholds. By Chapter 9, the minimum L_k with a single threshold is $L_k = \widetilde{L}_k(2)$. In addition to Theorem 10.16, (10.13) is true.

For (i), we have $P_k(1,2) \geq Q_k(1,2)$, which together with $C_1 < C_2$ gives

$$P_k(1,2) = \frac{(C_1 + \delta)\overline{a}_{k-1} - (C_2 - \delta)\underline{a}_{k-1}}{(C_1 + C_2)e_{k-1}} \leq \widetilde{L}_k(2).$$

Hence, $L_k = \widetilde{L}_k(C_2)$.

For (ii), we have $P_k(1,2) \leq Q_k(1,2)$, which together with $r_{k-1} \leq \frac{C_1 - \delta}{2C_2 - C_1 + \delta}$ gives $Q_k(1,2) \leq \widetilde{L}_k(2)$. Hence, $L_k = Q_k(1,2)$. By Theorem 10.16, one can get the optimal input of (i)–(iii).

For (iii), we have $L_k = \widetilde{L}_k(m_0)$. By Chapter 9, the optimal input is $u_k = \frac{2C_2}{\overline{a}_{k-1} + \underline{a}_{k-1}}$.

For (iv), by Chapter 9 we have $L_k = 1$ for any inputs. □

10.3.2 Uniform Quantization

Consider that L_k is affected by C_1, C_{m_0}, and $\max \Delta C_l$, and $Q_k(1, m_0)$ decreases with decreasing $\max \Delta C_l$. In order to minimize $\max \Delta C_l$, we assume

$$C_2 - C_1 = C_3 - C_2 = \cdots = C_{m_0} - C_{m_0 - 1} := \Delta C.$$

Lemma 10.19. *Consider* (10.7). *Assume that* $\delta < \Delta C/2$, $u_{\max} \geq (C_{m_0} - \delta)/\underline{a}_0$, *and* $e_{k-1} \geq (C_{m_0} - C_1 + 2\delta)/u_{\max}$. *For* $2 \leq i \leq j \leq m_0$, *denote*

$$\widetilde{L}_k(i,j) = \vee\{P_k(i,j), Q_k(i-1,i)\}.$$

In the worst case, we have

$$\widetilde{L}_k(1, m_0) \leq \wedge\{\widetilde{L}_k(1, m_0 - 1), \widetilde{L}_k(C_2, C_{m_0})\}. \tag{10.14}$$

Proof. By $C_1 < C_2$, we have $P_k(1, m_0) \leq Q_k(1, m_0 - 1)$ and $\widetilde{L}_k(1, m_0) \leq \widetilde{L}_k(1, m_0 - 1)$. To prove (10.14), we need only show that $\widetilde{L}_k(1, m_0) \leq \widetilde{L}_k(2, m_0)$.

For $r_{k-1} \leq \frac{C_1 - \Delta C - \delta}{C_i + \Delta C + \delta}$, $i = m_0 - 1, m_0$, we have $P_k(1,i) \leq Q_k(i-1,i)$, and hence, $\widetilde{L}_k(1,i) = P_k(1,i)$. Similarly, for $r_{k-1} > \frac{C_1 - \Delta C - \delta}{C_i + \Delta C + \delta}$, we have $\widetilde{L}_k(1,i) = Q_k(i-1,i)$.

Considering that $\widetilde{L}_k(1,i)$ is piecewise about $\underline{a}_{k-1}/\overline{a}_{k-1}$, we study the following three cases:

(i) When $r_{k-1} \le \frac{C_1-\Delta C-\delta}{C_{m_0}+\Delta C+\delta}$, we have

$$\widetilde{L}_k(1,i) = P_k(1,i), \quad i = m_0-1,\ m_0.$$

Noting that

$$\begin{aligned} P_k(1,i) &= \frac{(C_1+\delta)\overline{a}_{k-1} - (C_i-\delta)\underline{a}_{k-1}}{(C_1+C_i)e_{k-1}} \\ &= \frac{(C_1+\delta)(\overline{a}_{k-1}+\underline{a}_{k-1})}{(C_1+C_i)e_{k-1}} - \frac{\underline{a}_{k-1}}{e_{k-1}}, \end{aligned}$$

and $C_{m_0} \ge C_2$, we have $P_k(1,m_0) \le P_k(2,m_0)$, and hence, $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(C_2,C_{m_0})$.

(ii) When $r_{k-1} \ge \frac{C_1-\Delta C-\delta}{C_{m_0-1}+\Delta C+\delta}$, we have

$$\widetilde{L}_k(1,i) = Q_k(i,i+1), \quad i = m_0-1, m_0.$$

Noticing that

$$Q_k(i,i+1) = \frac{\overline{a}_{k-1}(\Delta C+2\delta)}{(C_i+\Delta C+\delta)e_{k-1}} = \frac{\overline{a}_{k-1}(\Delta C+2\delta)}{(C_i+\Delta C+\delta)e_{k-1}}$$

and $C_{m_0-1} < C_{m_0}$, we have $Q_k(m_0-2,m_0-1) \le Q_k(m_0-1,m_0)$, and hence, $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(1,m_0-1)$.

(iii) When $\frac{C_1-\Delta C-\delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} < \frac{C_1-\Delta C-\delta}{C_{m_0-1}+\Delta C+\delta}$, we have

$$\widetilde{L}_k(1,m_0) = Q_k(m_0-1,m_0) \quad \text{and} \quad \widetilde{L}_k(1,C_{m_0-1}) = P_k(1,m_0-1).$$

By (ii), we have $r_{k-1} = \frac{C_1-\Delta C-\delta}{C_2+\Delta C+\delta}$, and hence, $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(1,m_0-1)$. Notice that $\widetilde{L}_k(1,m_0) = Q_k(m_0-1,m_0)$ increases about $\underline{a}_{k-1}/\overline{a}_{k-1}$. Then $\widetilde{L}_k(1,m_0-1) = Q_k(m_0-1,m_0)$ decreases about r_{k-1}. Thus, for case (iii), we have $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(1,m_0-1)$. To summarize, (10.14) is true. □

Theorem 10.20. *Consider* (10.7). *Assume* $\delta < \Delta C/2$ *and* $e_{k-1} \ge (C_{m_0} - C_1 + 2\delta)/u_{\max}$. *Then for* $u_k \in \mathbb{U}$, *in the worst case, the minimum* L_k *is that*

$$L_k = \wedge\{\vee\{P_k(1,m_0), Q_k(m_0-1,m_0)\}, \widetilde{L}_k(m_0)\}. \tag{10.15}$$

Furthermore,

(i) *for* $r_{k-1} \le \frac{C_{m_0}-m_0\Delta C-\delta}{C_{m_0}+\Delta C+\delta}$, $L_k = P_k(1,m_0)$ *and the optimal input is* $u_k = \frac{C_1+C_{m_0}}{\overline{a}_{k-1}+\underline{a}_{k-1}}$;

(ii) *for* $\frac{C_{m_0}-m_0\Delta C-\delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} \le \frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta}$, $L_k = Q_k(m_0-1, m_0)$ *and the optimal input is* $u_k = \frac{C_{m_0}+\Delta C+\delta}{\overline{a}_{k-1}}$;

(iii) *for* $\frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} \le \frac{C_{m_0}-\delta}{C_{m_0}+\delta}$, $L_k = \widetilde{L}_k(m_0)$ *and the optimal input is* $u_k = \frac{2C_{m_0}}{\overline{a}_{k-1}+\underline{a}_{k-1}}$;

(iv) *for* $r_{k-1} > \frac{C_{m_0}-\delta}{C_{m_0}+\delta}$, $L_k = 1$ *for any* $u_k \in \mathbb{U}$.

Proof. By Corollary 10.18, (10.15) is true for $m_0 = 2$. Assume that for $i \ge 2$, (10.15) is true for $m_0 = i$. Then for $m_0 = i+1$, we have

$$L_k = \wedge\{\widetilde{L}_k(1, i+1), \widetilde{L}_k(1, i), \widetilde{L}_k(2, i+1), \widetilde{L}_k(i+1)\}.$$

By Lemma 10.19, we have $\widetilde{L}_k(1, i+1) \le \wedge\{\widetilde{L}_k(1, i), \widetilde{L}_k(2, i+1)\}$, which implies that (10.15) is true for $m = i+1$. Thus, by induction we have (10.15).

The rest can be obtained by comparing $P_k(1, m_0)$, $Q_k(m_0-1, m_0)$, and $\widetilde{L}_k(m_0)$. □

By Theorem 10.20, $L_k < \widetilde{L}_k(m_0)$ is equivalent to $\widetilde{L}_k(1, m_0) < \widetilde{L}_k(m_0)$, or equivalently, $r_{k-1} < \frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta}$.

Theorem 10.21. *Consider* (10.7). *Assume* $\delta < \Delta C/2$ *and denote*

$$\mathcal{K} = \{k : e_{k-1} \ge (C_{m_0} - C_1 + 2\delta)/u_{\max}\}.$$

For each k, *choose the optimal* u_k *to minimize* L_k; *then in the worst case, there exists at most one* $k \in \mathcal{K}$ *such that* $L_k = Q_k(m_0-1, m_0)$.

Proof. Since $\underline{a}_k$ and $\overline{a}_k$ are increasing and decreasing with respect to k, by Theorem 10.20, for

$$\frac{C_{m_0} - m_0\Delta C - \delta}{C_{m_0} + \Delta C + \delta} < r_{k-1} \le \frac{C_{m_0-1} - \delta}{C_{m_0} + \Delta C + \delta},$$

the minimum of L_k is $L_k = Q_k(m_0-1, m_0)$. Hence, if there exist two k's in $\mathcal{K}$ such that $L_k = Q_k(m_0-1, m_0)$, then they must be neighbors. Denote them as k_0, k_0+1, respectively. Then,

$$\frac{L_{k_0+1}}{L_{k_0}} = \frac{\overline{a}_{k_0}}{\overline{a}_{k_0-1}} \frac{e_{k_0-1}}{e_{k_0}} = \frac{\overline{a}_{k_0}}{\overline{a}_{k_0-1}} \frac{1}{L_{k_0}},$$

or equivalently,

$$L_{k_0+1} = \frac{\overline{a}_{k_0}}{\overline{a}_{k_0-1}}.$$

Since $\overline{a}_{k_0} = \overline{a}_{k_0+1}$ in the worst case, we have $L_{k_0+1} = 1$, which contradicts the fact that $L_{k_0+1} < \widetilde{L}_{k_0+1}(m_0) < 1$ for

$$\frac{C_{m_0} - m_0\Delta C - \delta}{C_{m_0} + \Delta C + \delta} < \frac{\underline{a}_{k_0}}{\overline{a}_{k_0}} \leq \frac{C_{m_0-1} - \delta}{C_{m_0} + \Delta C + \delta}.$$

□

By Theorem 10.21, we can figure out how many k's there exist such that $L_k < \widetilde{L}_k(m_0)$.

Theorem 10.22. *Consider* (10.7). *Assume* $\delta < \Delta C/2$ *and denote* $\mathcal{K} = \{k : e_{k-1} \geq (C_{m_0} - C_1 + 2\delta)/u_{\max}\}$. *In the worst case, choose the optimal input to minimize* L_k. *Then, the total number* K *of the elements* $k \in \mathcal{K}$ *such that* $L_k < \widetilde{L}_k(m_0)$ *satisfies*

$$\frac{\ln \alpha_4 - \ln(1-\alpha_4)}{\ln \alpha_3} \leq K \leq \frac{\ln \alpha_2 - \ln(1-\alpha_2)}{\ln \alpha_1} + 1, \tag{10.16}$$

where

$$\begin{aligned} \alpha_1 &= \frac{C_1 + \delta}{C_1 + C_{m_0}}, \quad \alpha_2 = \frac{(C_{m_0} - C_1 - 2\delta)\underline{a}_0}{(C_{m_0} - \delta)}, \\ \alpha_3 &= \frac{C_{m_0} - \delta}{C_1 + C_{m_0}}, \quad \alpha_4 = \frac{(C_{m_0} - C_1 - 2\delta)\overline{a}_0}{(C_{m_0} + \delta)}. \end{aligned}$$

Proof. By Theorem 10.20, $L_k = \wedge\{\vee\{Q_k(1, m_0), Q_k(m_0-1, m_0)\}, \widetilde{L}_k(m_0)\}$ for $k \in \mathcal{K}$ and $L_k < \widetilde{L}_k(m_0)$ only for $\frac{\underline{a}_{k-1}}{\overline{a}_{k-1}} < \frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta}$.

Furthermore, for $r_{k-1} \leq \frac{C_{m_0}-m_0\Delta C-\delta}{C_{m_0}+\Delta C+\delta}$, we have

$$L_k = P_k(1, m_0), \tag{10.17}$$

and for $\frac{C_{m_0}-m_0\Delta C-\delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} \leq \frac{C_2-\delta}{C_{m_0}+\Delta C+\delta}$,

$$L_k = Q_k(m_0 - 1, m_0). \tag{10.18}$$

Thus, by Theorem 10.21, there exists at most one k satisfying (10.18).

From (10.17), one obtains

$$e_k = \alpha_1 e_{k-1} - \frac{(C_{m_0} - C_1 - 2\delta)}{(C_1 + C_{m_0})}\underline{a}_{k-1} \tag{10.19}$$

and

$$e_k = \alpha_3 e_{k-1} - \frac{(C_{m_0} - C_{m_0} - 2\delta)}{(C_1 + C_{m_0})}\overline{a}_{k-1}. \tag{10.20}$$

By (10.19) and the fact that $\underline{a}_i$ decreases about i, we have

$$\begin{aligned} e_k &= \alpha_1^k e_0 - \frac{(C_{m_0} - C_1 - 2)}{(C_1 + C_{m_0})} \sum_{i=0}^{k-1} \alpha_1^i \underline{a}_{k-i-1} \\ &\leq \alpha_1^k e_0 - \alpha_2(1 - \alpha_1^k) \\ &= (1 - \alpha_2)\alpha_1^k - \alpha_2. \end{aligned}$$

By $e_k > 0$, or equivalently, $(1-\alpha_2)\alpha_1^k - \alpha_2 > 0$, we have

$$k \le \frac{\ln \dfrac{\alpha_2}{1-\alpha_2}}{\ln \alpha_1}.$$

Similarly,

$$k \ge \frac{\ln \alpha_4 - \ln(1-\alpha_4)}{\ln \alpha_3}.$$

Thus, by (10.18), we have (10.16). □

Remark 10.23. In Theorem 10.22, K is estimated by using the prior information of parameters. Since the bound of the unknown parameters decreases, the estimate of K is more accurate if updated parameter bounds are used.

10.4 Time Complexity

Section 10.3.2 shows that for uniform quantization with m_0 thresholds, in the worst case, the minimum L_k is

$$L_k = \wedge\{\vee\{P_k(1, m_0), Q_k(m_0 - 1, m_0)\}, \widetilde{L}_k(m_0)\}.$$

Considering that $\widetilde{L}_k(m_0)$ is about the largest threshold, and

$$r_{k-1} < \frac{C_{m_0-1} - \delta}{C_{m_0} + \Delta C + \delta},$$

we have $L_k < \widetilde{L}_k(m_0)$. Furthermore, by Theorem 10.21, in the worst case, there exists at most one $k \in \mathcal{K} = \{k : e_{k-1} \ge (C_{m_0} - C_1 + 2\delta)/u_{\max}\}$ such that $L_k = Q_k(m_0 - 1, m_0)$. So, we will study the effect of $P_k(1, m_0)$ on the parameter estimation and its time complexity.

(A10.2) $C_2 - C_1 = C_3 - C_2 = \cdots = C_{m_0} - C_{m_0-1} := \Delta C$, $\delta < \Delta C/2$, and $e > (C_{m_0} - C_1 + 2\delta)/u_{\max}$.

Lemma 10.24. *Consider* (10.7). *Let*

$$\sigma = \frac{m_0 \Delta C + 2\delta}{C_{m_0} + \Delta C - \delta} \overline{a}_0.$$

For $e \in (\sigma, e_0)$, $N(e)$ *is the time complexity of the parameter's unknown bound from* e_0 *to* e. *Choose optimal inputs in each time* k. *Then,*

$$N(e) \le \frac{\ln(\alpha_2 + e) - \ln[(1-\alpha_2)e_0]}{\ln \alpha_1}, \tag{10.21}$$

where

$$\alpha_1 = \frac{C_1 + \delta}{C_1 + C_{m_0}}, \quad \alpha_2 = \frac{(C_{m_0} - C_1 - 2\delta)\underline{a}_0}{(C_{m_0} - \delta)}.$$

Proof. Since $e_{k-1} > e > \sigma$ and $\overline{a}_0 \geq \overline{a}_{k-1}$,

$$r_{k-1} \leq \frac{C_{m_0} - m_0 \Delta C - \delta}{C_{m_0} + \Delta C + \delta}.$$

By Theorem 10.20, the minimum of L_k is $P_k(1, m_0)$. Hence,

$$\begin{aligned} e_k &= \alpha_1 e_{k-1} - \frac{C_{m_0} - C_1 - 2\delta}{C_1 + C_{m_0}} \underline{a}_{k-1} \\ &\leq \alpha_1^k e_0 - \frac{(C_{m_0} - C_1 - 2\delta)\underline{a}_0}{C_1 + C_{m_0}} \sum_{i=0}^{k-1} \alpha_1^i \\ &= (1 - \alpha_2)\alpha_1^k e_0 - \alpha_2. \end{aligned}$$

In the worst case, a necessary and sufficient condition of $e_k \leq e$ is

$$(1 - \alpha_2)\alpha_1^k e_0 - \alpha_2 \leq e,$$

or equivalently,

$$k \geq \frac{\ln \dfrac{\alpha_2 + e}{(1 - \alpha_2)e_0}}{\ln \alpha_1}.$$

Thus, (10.21) is true. □

Theorem 10.25. *For* (10.3), *let*

$$\begin{aligned} \overline{\theta}_0 &= \max_{1 \leq i \leq n_0} \overline{a}_0(i), & \underline{\theta}_0 &= \min_{1 \leq i \leq n_0} \underline{\theta}_0(i), \\ \sigma &= \frac{m_0 \Delta C + 2\delta}{C_{m_0} + \Delta C - \delta} \overline{a}_0, & e_0 &= \mathrm{Rad}_p(\Omega_0). \end{aligned}$$

Then, for any given $e \in (\sigma, e_0)$, *we have*

$$N(e) \leq \nu(n_0) \frac{\ln(\alpha_2 + e/n_0^{1/p}) - \ln[(1 - \alpha_2)e_0]}{\ln \alpha_1}, \tag{10.22}$$

where

$$\alpha_1 = \frac{C_1 + \delta}{C_1 + C_{m_0}} \quad \textit{and} \quad \alpha_2 = \frac{(C_{m_0} - C_1 - 2\delta)\underline{\theta}_0}{(C_{m_0} - \delta)}.$$

Proof. For the inputs constructed in Theorem 10.13, after $N = \nu(n_0)N_1$ steps, each parameter bound can be updated N_1 times. In addition to Lemma 10.24, we have

$$\begin{aligned} \mathrm{Rad}_p(\Omega_N) &\leq n_0^{1/p} \mathrm{Rad}_\infty(\Omega_N) \\ &\leq n_0^{1/p}[(1 - \alpha_2)\alpha_1^{N/\nu(n_0)} \mathrm{Rad}_\infty(\Omega_0) - \alpha_2] \\ &\leq n_0^{1/p}[(1 - \alpha_2)\alpha_1^{N/\nu(n_0)} \mathrm{Rad}_p(\Omega_0) - \alpha_2] \\ &= n_0^{1/p}[(1 - \alpha_2)\alpha_1^{N/\nu(n_0)} e_0 - \alpha_2]. \end{aligned}$$

So, (10.22) is true. □

Replace the l^p norm in Theorem 10.25 by the l^∞ norm. Then the following theorem can be derived.

Corollary 10.26. *For* (10.3), *let* $\overline{\theta}_0 = \max_{1\le i\le n_0} \overline{a}_i$, $\underline{\theta}_0 = \min_{1\le i\le n_0} \underline{a}_{is}$, $\sigma = \frac{m_0\Delta C+2\delta}{C_{m_0}+\Delta C-\delta}\overline{\theta}_0$, *and* $e_0 = \mathrm{Rad}_\infty(\Omega_0)$. *For any given* $e \in (\sigma_0, e)$, *we have*

$$N(e) \le \nu(n_0)\frac{\ln(\alpha_2+e) - \ln[(1-\alpha_2)e_0]}{\ln\alpha_1},$$

where α_1 *and* α_2 *were introduced in Theorem* 10.25.

10.5 Examples

Two examples are used to demonstrate the methods developed in this chapter. In Example 10.27, the parameters are estimated, and then the input is designed to track a target output. In Example 10.28, the difference between the estimation of quantized observations ($m_0 > 2$) and binary observations is discussed.

Example 10.27 (Tracking problem). For given target output y^*, design inputs such that $y_k = y^*$. It is worth mentioning that the y_k is measured by a quantized sensor with thresholds $C_i,\ i = 1, 2, \ldots, m_0$. For the classical tracking problem, the target output y^* must be equal to some threshold, and for $y^* \ne C_i$, $i = 1, 2, \ldots, m_0$, the tracking problem cannot be solved by classical methods. However, by using the results developed in this chapter, the unknown parameter can be estimated first, then the inputs can be designed to track y^*.

Consider

$$y_k = a_1u_k + a_2u_{k-2} + a_3u_{k-3} + d_k,\ k = 3, 4, \ldots,$$

where $d \le 1$; the real a_1, a_2, a_3 are 12, 10, 5, but unknown; the prior information is: $a_i \in [1, 70]$, $i = 1, 2, 3$, $u_{\max} = 30$, $y^* = 70$, and y_k is measured by a two-threshold sensor with $C_1 = 70$, $C_2 = 80$.

By Theorem 10.13, let

$$u = \{u_0, 0, 0, u_3, 0, u_5, 0, u_7, 0, u_9, 0, 0, u_8, \ldots\};$$

then

$$\begin{aligned}
y_{13i+3} &= a_3u_{13i}, & y_{13i+12} &= a_3u_{13i+9}, & i &= 1, 2, \ldots,\\
y_{13i+4} &= a_1u_{13i+3}, & y_{13i+13} &= a_1u_{13i+12}, & i &= 1, 2, \ldots,\\
y_{13i+7} &= a_2u_{13i+5}, & y_{13i+9} &= a_2u_{13i+7} & i &= 1, 2, \ldots
\end{aligned}$$

Hence, the parameters are decoupled and then estimated. The estimate is aimed at reducing the unknown bound of each parameter to be less than 1 (see Figure 10.2).

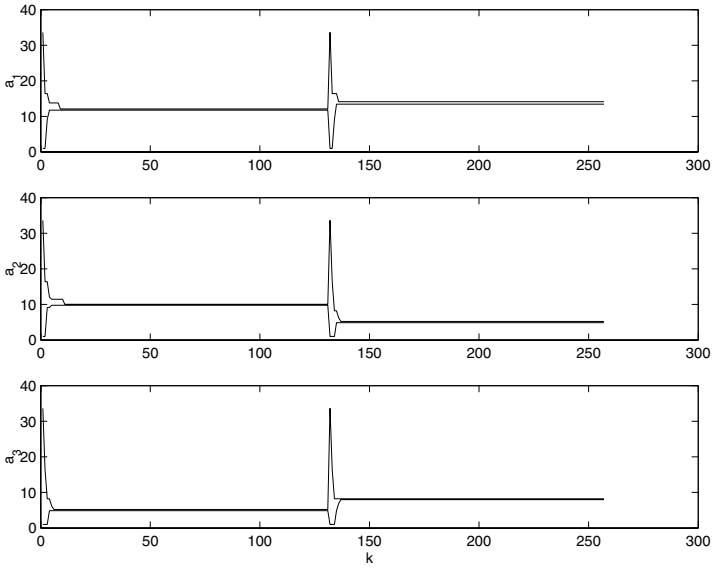


FIGURE 10.2. Reducing unknown parameter bound

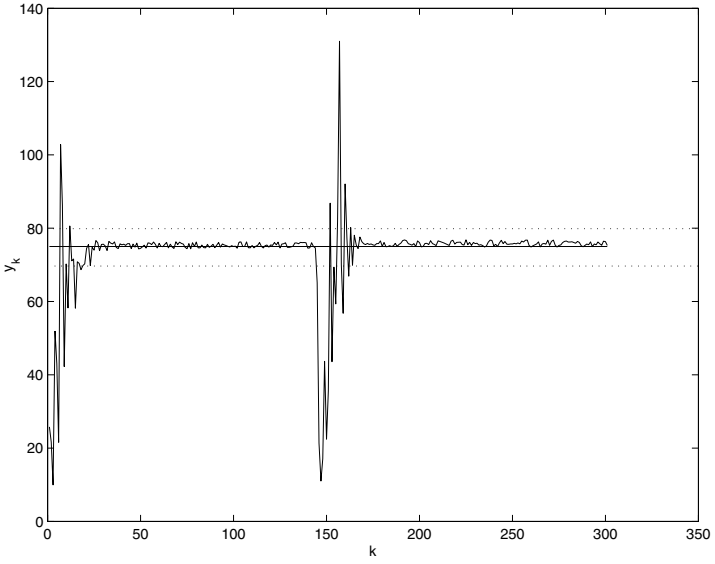


FIGURE 10.3. Tracking target output y^*

TABLE 10.1. Two-threshold sensor

Item	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
$\underline{a}_k$	1	1	1	8.26	11.35
$\overline{a}_k$	60	30.03	15.28	15.28	12.19
e_k	59	29.03	14.28	7.01	0.8
L_k	1	0.49	0.49	0.49	0.12

TABLE 10.2. Binary sensor

Item	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
$\underline{a}_k$	1	1	1	8.45	8.45	10.31	11.23	11.69
$\overline{a}_k$	60	30.82	16.08	16.08	12.39	12.39	12.39	12.39
e_k	59	29.82	15.08	7.63	3.94	2.08	1.16	0.70
L_k	1	0.51	0.51	0.51	0.52	0.53	0.56	0.61

Consequently, consider the center of each unknown parameter interval $\widehat{a}_1, \widehat{a}_2, \widehat{a}_3$ as the estimate of a_1, a_2, a_3. For y^*, let

$$u_k = \frac{1}{\widehat{a}_1}(y^* - \widehat{a}_2 u_k - \widehat{a}_3 u_{k-2}).$$

At some k, change the parameters to $a_1 = 14, a_2 = 5, a_3 = 8$. Then y_k no longer tracks y_k. However, the quantized sensor can detect it and the parameters can be estimated again (Figures 10.2 and 10.3).

Example 10.28 (Comparison of estimations with quantized and binary sensors). Consider

$$y_k = au_k + d_k, \tag{10.23}$$

where the real parameter $a = 12$ but is unknown. The prior information is that $a \in [1, 60]$, $\delta = \|dl\|_\infty \leq 1$, and $u_{\max} = 30$. There are two-threshold sensors with $C_1 = 95$, $C_2 = 100$ and a binary sensor with $C = 95$. The estimation aim is to reduce the unknown parameter bound to be less than 1.

Then,

(i) increasing thresholds makes the estimation faster (Tables 10.1, 10.2, and Figure 10.4);

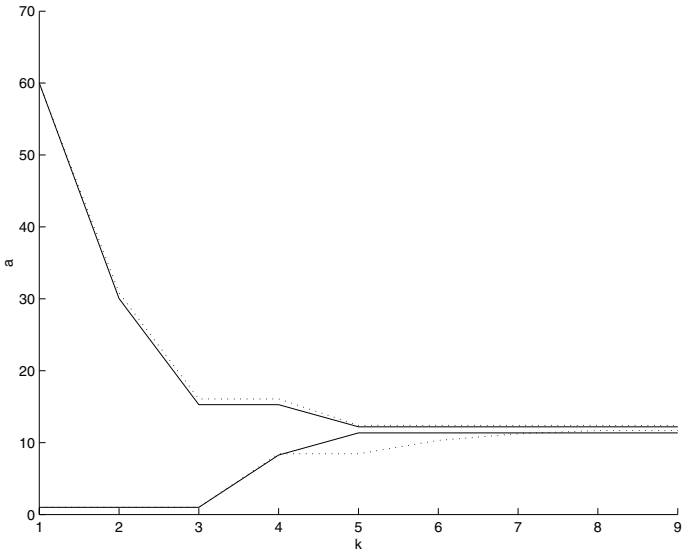


FIGURE 10.4. Comparison of estimations with two thresholds and binary sensors

(ii) from Table 10.1, we can find that $L_k < 1/m_0$, because the m_0 thresholds divided the parameter space into $m_0 + 1$ intervals, and L_k is decided by the longest interval in the worst case. However, for a given question, L_k may be decided by a smaller interval.

10.6 Notes

Based on the shortest decoupling input and analysis of gain system, the identification of an FIR system with quantized observations is studied.

In this chapter, the shortest decoupling input starting from k_0 is defined and designed without utilizing the information of inputs before k_0 and the length is $\nu(n_0)$. In some special cases assuming $u_k = 0$ for $k = k_0 - n_0 + 2, \ldots, k_0$, the shortest length of inputs to decouple each parameter once can be designed to be $n_0(n_0 + 1)/2$, which can be generalized to decouple each parameter p times with only $pn_0(n_0 + 1)/2$ (see [14]). For example, if $n_0 = 3$ and $u_k = 0$ for $k = -1, 0$, by input $u = \{u_1, u_2, 0, 0, u_5, 0\}$, one can get

$$y_2 = a_1 u_1, \quad y_5 = a_3 u_2, \quad y_7 = a_2 u_5.$$

That is, u is a three-parameter-decoupling input segment and the length is $n_0(n_0 + 1)/2 = 6$.

However, the condition that $u_k = 0$ for $k = k_0 - n_0 + 2, \ldots, k_0$ may not be reasonable. For example, after the input segments, in the above example,

we have $k_0' = 6$, but the condition that $u_k = 0$ for $k = k_0' - 1, k_0'$ is not satisfied. The decoupling method developed in this chapter has a benefit of being independent of such conditions, and it is the optimal one in this case.

Part IV

Identification of Nonlinear and Switching Systems

11

Identification of Wiener Systems with Binary-Valued Observations

This chapter studies the identification of Wiener systems whose outputs are measured by binary-valued sensors. The system consists of a linear FIR (finite impulse response) subsystem of known order, followed by a nonlinear function with a known parameterization structure. The parameters of both linear and nonlinear parts are unknown. Input design, identification algorithms, and their essential properties are presented under the assumptions that the distribution function of the noise is known and the nonlinearity is continuous and invertible. We show that under scaled periodic inputs, the identification of Wiener systems can be decomposed into a finite number of core identification problems. The concept of joint identifiability of the core problem is introduced to capture the essential conditions under which the Wiener system can be identified with binary-valued observations. Under scaled full-rank conditions and joint identifiability, a strongly convergent algorithm is constructed. The algorithm is shown to be asymptotically efficient for the core identification problem, hence achieving asymptotic optimality in its convergence rate. For computational simplicity, recursive algorithms are also developed.

This chapter is organized as follows. The structure of Wiener systems using binary-valued observations is presented in Section 11.1. Section 11.2 shows that under scaled periodic inputs, the identification of Wiener systems can be decomposed into a number of core identification problems. Basic properties of periodic signals and the concepts of joint identifiability are introduced in Section 11.3. Based on the algorithms for the core problems, Section 11.4 presents the main identification algorithms for Wiener systems. Under scaled full-rank inputs and joint identifiability, the identification

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_11,

algorithms for Wiener systems are shown to be strongly convergent. Identification algorithms for the core problems are constructed in Section 11.5. By comparing the estimation errors with the CR lower bound, the algorithms are shown to be asymptotically efficient, hence achieving asymptotically optimal convergence speed. For simplicity, recursive algorithms are discussed in Section 11.6 that can be used to find system parameters under certain stability conditions. Illustrative examples are presented in Section 11.7 to demonstrate input design, identification algorithms, and convergence results of the methodologies discussed in this chapter.

11.1 Wiener Systems

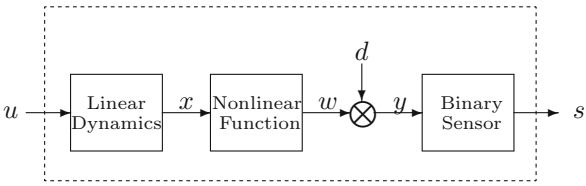


FIGURE 11.1. Wiener systems with binary-valued observations

Consider the system in Figure 11.1, in which

$$\begin{cases} y_k = H(x_k, \eta) + d_k, \\ x_k = \sum_{i=0}^{n_0-1} a_i u_{k-i}, \end{cases} \tag{11.1}$$

where u_k is the input, x_k the intermediate variable, and d_k the measurement noise. $H(\cdot, \eta)$: $\mathcal{D}_H \subseteq \mathbb{R} \to \mathbb{R}$ is a parameterized static nonlinear function with domain $\mathcal{D}_H$ and vector-valued parameter $\eta \in \Omega_\eta \subseteq \mathbb{R}^{m_0}$. Both n_0 and m_0 are known. By defining $\phi_k = [u_k, \ldots, u_{k-n_0+1}]'$ and $\theta = [a_0, \ldots, a_{n_0-1}]'$, the linear dynamics can be expressed compactly as $x_k = \phi_k'\theta$. The output y_k is measured by a binary sensor with threshold C.

(A11.1) The noise $\{d_k\}$ is a sequence of i.i.d. random variables whose distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are twice continuously differentiable and known. For any given $\eta \in \Omega_\eta$, $H(x, \eta)$ is bounded for any finite x, continuous and invertible in x.

Parameterization of the static nonlinear function $H(\cdot, \eta)$ depends on specific applications. Often, the structures of actual systems can provide guidance in selecting function forms whose parameters carry physical

meanings ([103, 107]). On the other hand, when a black-box approach is employed, namely, the models represent input–output relationships based on data only, one may choose some generic structures for theoretical and algorithmic development. For instance, a common structure is $H(x,\eta) = \sum_{i=0}^{m_0-1} b_i h_i(x)$, where $h_i(x)$, $i = 0,\ldots,m_0-1$, are base functions and $\eta = [b_0, b_1, \ldots, b_{m_0-1}]' \in \mathbb{R}^{m_0}$ is a vector of m_0 unknown parameters. For example, the typical polynomial structure is

$$H(x,\eta) = \sum_{i=0}^{m_0-1} b_i x^i. \tag{11.2}$$

In this chapter, we discuss input design, derive joint estimators of θ and η, and establish their identifiability, convergence, convergence rates, and efficiency (optimality in convergence rate).

11.2 Basic Input Design and Core Identification Problems

We first outline the main ideas of using $2n_0(m_0+1)$-scaled periodic inputs and empirical measures to identify Wiener systems under binary-valued observations. It will be shown that this approach leads to a core identification problem, for which identification algorithms and their key properties will be established.

The input signal u to be used to identify the system is $2n_0(m_0+1)$-periodic, whose one-period values are $(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_{m_0} v, \rho_{m_0} v)$, where $v = (v_1, \ldots, v_{n_0})$ is to be specified. (The reason for repeating each scaled vector, such as $\rho_0 v, \rho_0 v$, etc., is to simplify the algorithm development and convergence analysis, not a fundamental requirement.) The scaling factors $\rho_0, \rho_1, \ldots, \rho_{m_0}$ are assumed to be nonzero and distinct.

If, under the $2n_0$ input values $u = (v, v)$, the linear subsystem has the following n_0 consecutive output values at $n_0, \ldots, 2n_0 - 1$:

$$\delta_i = a_0 u_{n_0+i} + \cdots + a_{n_0-1} u_{1+i}, \quad i = 0, \ldots, n_0 - 1,$$

then the output under the scaled input (qv, qv) is

$$x_{n_0+i} = q\delta_i, \quad i = 0, \ldots, n_0 - 1.$$

Without loss of generality, assume $\delta_0 \neq 0$. (When v is full rank, there exists at least one i such that $\delta_i \neq 0$.) The output of the linear subsystem contains the following (m_0+1)-periodic subsequence with its one-period values $\{\rho_0\delta_0, \rho_1\delta_0, \ldots, \rho_{m_0}\delta_0\}$:

$$x_{n_0} = \rho_0\delta_0,\ x_{3n_0} = \rho_1\delta_0, \ldots,$$
$$x_{(2m_0+1)n_0} = \rho_{m_0}\delta_0, \ldots$$

By concentrating on this subsequence of x_k, under a new index l with $l = 1, 2, \ldots$, the corresponding output of the nonlinear part may be rewritten as

$$\begin{aligned} \widetilde{Y}_{l(m_0+1)} &= H(\rho_0\delta_0, \eta) + \widetilde{D}_{l(m_0+1)}, \\ \widetilde{Y}_{l(m_0+1)+1} &= H(\rho_1\delta_0, \eta) + \widetilde{D}_{l(m_0+1)+1}, \\ &\vdots \\ \widetilde{Y}_{l(m_0+1)+m_0} &= H(\rho_{m_0}\delta_0, \eta) + \widetilde{D}_{l(m_0+1)+m_0}. \end{aligned} \tag{11.3}$$

The equations in (11.3) form the basic observation relationship for identifying η and δ_0.

For $\rho = [\rho_0, \ldots, \rho_{m_0}]'$ and a scalar δ, denote

$$\mathbf{H}(\rho\delta, \eta) = [H(\rho_0\delta, \eta), \ldots, H(\rho_{m_0}\delta, \eta)]'. \tag{11.4}$$

Then, (11.3) can be expressed as

$$\widetilde{Y}_l = H(\rho\delta, \eta) + \widetilde{D}_l, \quad l = 0, 1, \ldots, \tag{11.5}$$

where $\delta \neq 0$,

$$\begin{aligned} &\widetilde{Y}_l = [\widetilde{Y}_{l(m_0+1)}, \ldots, \widetilde{Y}_{l(m_0+1)+m_0}]' \text{ and} \\ &\widetilde{D}_l = [\widetilde{D}_{l(m_0+1)}, \ldots \widetilde{D}_{l(m_0+1)+m_0}]'. \end{aligned}$$

Correspondingly, the outputs of the binary-valued sensor on $\widetilde{Y}_l$ are $\widetilde{S}_l = \mathcal{S}(\widetilde{Y}_l)$, $l = 0, 1, \ldots$ Let $\tau = [\tau_0, \ldots, \tau_{m_0}]' := [\delta, \eta']'$. We introduce the following core identification problem.

Core Identification Problem

Consider the problem of estimating the parameter τ from observations on $\widetilde{S}_l$. Denote $\zeta_i = H(\rho_i\delta, \eta), i = 0, 1, \ldots, m_0$. Then $\zeta = [\zeta^{\{1\}}, \ldots, \zeta^{\{m_0+1\}}]' = H(\rho\delta, \eta)$ and (11.5) can be rewritten as

$$\widetilde{Y}_l = \zeta + \widetilde{D}_l, \quad l = 0, 1, \ldots \tag{11.6}$$

The main idea of solving the core identification problem is first to estimate ζ, and then to solve the interpolation equations

$$\zeta_i = H(\rho_i\delta, \eta), \quad i = 0, 1, \ldots, m_0, \tag{11.7}$$

for δ and η. The basic properties on signals and systems that ensure solvability of the core identification problem will be discussed next.

11.3 Properties of Inputs and Systems

We first establish some essential properties of periodic signals and present the idea of joint identifiability, which will play an important role in the subsequent development. Some related ideas can be found in [43, 57, 108].

Generalized Circulant Matrices and Periodic Inputs

An $n_0 \times n_0$ generalized circulant matrix ([57])

$$T = \begin{bmatrix} v_{n_0} & v_{n_0-1} & v_{n_0-2} & \cdots & v_1 \\ qv_1 & v_{n_0} & v_{n_0-1} & \ddots & v_2 \\ qv_2 & qv_1 & v_{n_0} & \ddots & v_3 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ qv_{n_0-1} & qv_{n_0-2} & qv_{n_0-3} & \cdots & v_{n_0} \end{bmatrix} \tag{11.8}$$

is completely determined by its first row $[v_{n_0}, \ldots, v_1]$ and q, which will be denoted by $T(q, [v_{n_0}, \ldots, v_1])$. In the special case of $q = 1$, the matrix in (11.8) is a circulant matrix and will be denoted by $T([v_{n_0}, \ldots, v_1])$.

Definition 11.1. A $2n_0(m_0 + 1)$-periodic signal u is called a scaled full-rank signal if its one-period values are $(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_{m_0} v, \rho_{m_0} v)$, and $v = (v_1, \ldots, v_{n_0})$ is full rank; $\rho_j \neq 0$, $j = 1, \ldots, m_0$, and $\rho_i \neq \rho_j$, $i \neq j$. We use $\mathcal{U}$ to denote the class of such signals.

Definition 11.2. An $n_0(m_0 + 1)$-periodic signal u is called an exponentially scaled full-rank signal if its single-period values are $(v, qv, \ldots, q^{m_0} v)$ with $q \neq 0$ and $q \neq 1$, and $v = (v_1, \ldots, v_{n_0})$ is full rank. We use $\mathcal{U}_e$ to denote this class of input signals.

Joint Identifiability

Joint identifiability conditions mandate that the unknown parameters δ and η can be uniquely and jointly determined by the interpolation conditions (11.7).

Prior Information. The prior information on the unknown parameters $\tau = [\delta, \eta']'$ for the core identification problem is $\tau \in \Omega \subseteq \mathbb{R}^{m_0+1}$. Denote $\mathbb{R}_d^{m_0} = \{\rho = [\rho_1, \ldots, \rho_{m_0}]' \in \mathbb{R}^{m_0} : \rho_j \neq 0, \forall j; \rho_i \neq \rho_j, i \neq j\}$, namely, the set of all vectors in $\mathbb{R}^{m_0}$ that contain nonzero and distinct elements.

Definition 11.3. Suppose that $\Upsilon \subseteq \mathbb{R}_d^{m_0+1}$. $H(x; \eta)$ is said to be *jointly identifiable* in Ω with respect to Υ if, for any $\rho = [\rho_0, \ldots, \rho_{m_0}]' \in \Upsilon$, $H(\rho\delta; \eta)$ is invertible in Ω; namely, $\zeta = H(\rho\delta; \eta)$ has a unique solution $\tau \in \Omega$. In this case, elements in Υ are called *sufficiently rich scaling factors.*

Depending on the parametric function forms $H(\cdot, \eta)$ and the domain $\mathcal{D}_H$, the set of sufficiently rich scaling factors can vary significantly. For example, the polynomial class of functions of a fixed order has a large set Υ. The polynomial class has been used extensively as the nonlinear part of Wiener systems and their approximations in [15, 72, 113].

When the base functions are polynomials of order m_0, $H(x, \eta)$ can be expressed as

$$H(x, \eta) = \sum_{j=0}^{m_0} b_j x^j, \text{ with } b_{m_0} \neq 0.$$

Then $H(\rho_i \delta, \eta) = \sum_{j=0}^{m_0} b_j \delta^j \rho_i^j$, $i = 0, 1, \ldots, m_0$. Apparently, one cannot uniquely determine m_0+2 parameters $\delta, b_0, \ldots, b_{m_0}$ from m_0+1 coefficients of the polynomial. A typical remedy to this well-known fact is normalization of the parameter set by assuming one parameter, say, $b_l = 1$ for some l. In this case, the coefficient equations become $b_j \delta^j = c_j, j \neq l$, and $\delta^l = c_l$. For a given c_j, to ensure uniqueness of solutions $b_j, j \neq l$, and δ to the equations, l must be an odd number.

We now show that $H(x, \eta)$ satisfying Assumption (A11.1) contains at least one nonzero odd-order term. Indeed, if $H(x, \eta)$ contains only even-order terms, it must be an even function. It follows that $H(x, \eta) = H(-x, \eta)$; namely, it is not an invertible function. This contradicts Assumption (A11.1).

Since $H(x, \eta)$ contains at least one nonzero odd-order term $b_l x^l$ for some odd integer l, without loss of generality we assume $b_l = 1$. The reduced-parameter vector is $\eta_0 = [b_0, \ldots, b_{l-1}, b_{l+1}, \ldots, b_{m_0}]'$, which contains only m_0 unknowns. Such polynomials will be called normalized polynomial functions of order m_0.

Proposition 11.4. *Under Assumption* (A11.1)*, all normalized polynomial functions of order m_0 are jointly identifiable with respect to $\mathbb{R}_d^{m_0}$.*

Proof. For any given $\rho = [\rho_0, \ldots, \rho_{m_0}] \in \mathbb{R}_d^{m_0+1}$, the interpolation equations

$$\sum_{j=0}^{m_0} c_j \rho_i^j = \zeta_i, \quad i = 0, \ldots, m_0,$$

can be rewritten as $\Re c = \zeta$, where ζ is defined in (11.6) and

$$\Re = \begin{pmatrix} 1 & \rho_0 & \ldots & \rho_0^{m_0} \\ 1 & \rho_1 & \ddots & \rho_1^{m_0} \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \rho_{m_0} & \ldots & \rho_{m_0}^{m_0} \end{pmatrix}, \; c = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{m_0} \end{pmatrix}.$$

Since the determinant of the Vandermonde matrix is

$$\det \Re = \prod_{0 \le i < j \le m_0 - 1} (\rho_j - \rho_i) \neq 0$$

for distinct $\rho_i, i = 0, \ldots, m_0 - 1$, we have $c = \Re^{-1}\zeta$. Furthermore, the equation $\delta^l = c_l$ yields the unique solution $\delta = (c_l)^{1/l} \neq 0$ by hypothesis. Then, $b_j = c_j/\delta^j$, $j \neq l$; solve uniquely for the remaining parameters. Consequently, $H(\rho\delta; \eta_0)$ is invertible as a joint function of δ and η_0. This implies that $H(x, \eta_0)$ is jointly identifiable with respect to any vector in $\mathbb{R}_d^{m_0}$. □

Other bases can also be used. For instance, $H(x, \eta) = \eta + e^x$, where $\eta \neq 0$. Under the prior information $\Omega = \{[\delta, \eta']' : \delta > 0, \eta \neq 0\}$, consider $\Upsilon = \{(\rho_0, \rho_1) : \rho_0 > 0, \rho_1 < 0\}$. The interpolation equations are

$$\begin{cases} \eta + e^{\rho_0 \delta} = \zeta^{\{1\}}, \\ \eta + e^{\rho_1 \delta} = \zeta_1. \end{cases} \tag{11.9}$$

These imply

$$e^{\rho_0 \delta} - e^{\rho_1 \delta} = \zeta^{\{1\}} - \zeta_1. \tag{11.10}$$

It is easily seen that for $\rho_0 > 0$ and $\rho_1 < 0$, the derivative of (11.10) is

$$\frac{\mathrm{d}(e^{\rho_0 \delta} - e^{\rho_1 \delta})}{\mathrm{d}\delta} = \rho_0 e^{\rho_0 \delta} - \rho_1 e^{\rho_1 \delta} > 0.$$

Hence, (11.10) has a unique solution, which indicates that $H(x, \eta)$ is jointly identifiable with respect to Υ.

Joint identifiability is not a trivial condition. For the above function form $H(x, \eta) = \eta + e^x$, Υ cannot be expanded to $\mathbb{R}_d^2$. Indeed, if one selects $\rho_0 = -2$, $\rho_1 = -1$, $\zeta^{\{1\}} = 1.075$, $\zeta_1 = 1.2$, then (11.9) becomes

$$\begin{cases} \eta + e^{-2\delta} = 1.075, \\ \eta + e^{-\delta} = 1.2. \end{cases}$$

Both $\delta = 1.921$, $\eta = 1.054$ and $\delta = 0.158$, $\eta = 0.346$ solve the equations. By definition, $H(x, \eta) = \eta + e^x$ is not jointly identifiable with respect to $\mathbb{R}_d^2$.

11.4 Identification Algorithms

Based on periodic inputs and joint identifiability, we now derive algorithms for parameter estimates and prove their convergence.

(A11.2) The following conditions hold.

(i) The prior information on θ and η is that $\theta \neq 0$, $\eta \neq 0$, $\theta \in \Omega_\theta$, and $\eta \in \Omega_\eta$ such that under Ω_θ and Ω_η, the set Υ of sufficiently rich scaling factors is nonempty. $C - y_k$ lies within the support of the noise density $f(\cdot)$ for $k = 1, 2, \ldots$

(ii) $H(x, \eta)$ is jointly identifiable with respect to Υ and continuously differentiable with respect to both x and η.

By using the vector notation, for $j = 1, 2, \ldots$,

$$\begin{aligned} X_j &= [x_{2(j-1)(m_0+1)n_0+n_0}, \ldots, x_{2j(m_0+1)n_0+n_0-1}]', \\ Y_j &= [y_{2(j-1)(m_0+1)n_0+n_0}, \ldots, y_{2j(m_0+1)n_0+n_0-1}]', \\ \widetilde{\Phi}_j &= [\phi_{2(j-1)(m_0+1)n_0+n_0}, \ldots, \phi_{2j(m_0+1)n_0+n_0-1}]', \\ D_j &= [d_{2(j-1)(m_0+1)n_0+n_0}, \ldots, d_{2j(m_0+1)n_0+n_0-1}]', \\ S_j &= [s_{2(j-1)(m_0+1)n_0+n_0}, \ldots, s_{2j(m_0+1)n_0+n_0-1}]', \end{aligned} \tag{11.11}$$

the observations can be rewritten in block form as

$$\begin{cases} Y_j = H(X_j, \eta) + D_j, \\ X_j = \widetilde{\Phi}_j \theta. \end{cases}$$

The input is a scaled $2n_0(m_0 + 1)$-periodic signal with single-period values $(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_{m_0} v, \rho_{m_0} v)$ with $v = (v_1, \ldots, v_{n_0})$ full rank.

By periodicity, $\widetilde{\Phi}_j = \widetilde{\Phi}$, for all j and $\widetilde{\Phi}$, can be decomposed into $2(m_0+1)$ submatrices Φ_i, $i = 1, \ldots, 2(m_0 + 1)$, of dimension $n_0 \times n_0$:

$$\widetilde{\Phi} = [\Phi_1', \Phi_2', \ldots, \Phi_{2(m_0+1)}']'.$$

Denote the $n_0 \times n_0$ circulant matrix $\Phi = T([v_{n_0}, \ldots, v_1])$. Then the odd-indexed block matrices satisfy the simple scaling relationship

$$\Phi_1 = \rho_0 \Phi, \quad \Phi_3 = \rho_1 \Phi, \quad \ldots, \quad \Phi_{2m_0+1} = \rho_{m_0} \Phi. \tag{11.12}$$

Note that the even-indexed block matrices are not used in the proof.

Remark 11.5. In $(\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v, \ldots, \rho_{m_0} v, \rho_{m_0} v)$, there are always two identical subsequences $\rho_i v, i = 0, \ldots, m_0$, appearing consecutively. The main reason for this input structure is to generate block matrices that satisfy the above scaling relationship (11.12).

Algorithms for the Core Identification Problem

For the core problem (11.6), let

$$\begin{aligned} \widetilde{\xi}_N^0 &= \frac{1}{N} \sum_{l=0}^{N-1} \widetilde{S}_l \\ &= \frac{1}{N} \sum_{l=0}^{N-1} I_{\{\widetilde{D}_l \leq C\mathbb{1} - H(\rho\delta, \eta)\}}, \end{aligned}$$

which is the empirical distribution of $\widetilde{D}_k$ at $C\mathbb{1} - H(\rho\delta, \eta)$. Then, by the strong law of large numbers,

$$\widetilde{\xi}_N^0 \to p = F(C\mathbb{1} - H(\rho\delta, \eta)) \quad \text{w.p.1.}$$

Note that F is a monotone function and $H(x, \eta)$ is bounded by Assumption (A11.1). Then, there exists $\widetilde{z} > 0$ such that

$$\widetilde{z}\mathbb{1} \le \widetilde{p} := F(C\mathbb{1} - H(\rho\delta, \eta)) \le (1 - \widetilde{z})\mathbb{1}.$$

Since $F(\cdot)$ is not invertible at 0 and 1, we modify $\widetilde{\xi}_N^0$ to avoid these points. Let

$$\widetilde{\xi}_N = \begin{cases} \widetilde{\xi}_N^0, & \text{if} \quad \widetilde{z}\mathbb{1} \le \widetilde{\xi}_N^0 \le (1 - \widetilde{z})\mathbb{1}, \\ \widetilde{z}\mathbb{1}, & \text{if} \quad \widetilde{\xi}_N^0 < \widetilde{z}\mathbb{1}, \\ (1 - \widetilde{z})\mathbb{1}, & \text{if} \quad \widetilde{\xi}_N^0 > (1 - \widetilde{z})\mathbb{1}. \end{cases} \tag{11.13}$$

Then,

$$\widetilde{\xi}_N \to p = F(C\mathbb{1} - H(\rho\delta, \eta)) \quad \text{w.p.1.} \tag{11.14}$$

Hence,

$$\begin{aligned} \zeta_N &= C\mathbb{1} - F^{-1}(\widetilde{\xi}_N) \\ &\to \zeta = C\mathbb{1} - F^{-1}(\widetilde{p}) = H(\rho\delta, \eta) \quad \text{w.p.1.} \end{aligned}$$

By Assumption (A11.2), H is invertible as a function of $\tau = [\delta, \eta']'$. As a result, $\tau_N = H^{-1}(\zeta_N) \to \tau$ w.p.1. In summary, we have the following theorem.

Theorem 11.6. *Under Assumptions* (A11.1) *and* (A11.2)*, let*

$$\tau_N = H^{-1}(\zeta_N) = H^{-1}(C\mathbb{1} - F^{-1}(\widetilde{\xi}_N)).$$

Then

$$\tau_N \to \tau \quad \textit{w.p.1 as} \quad N \to \infty. \tag{11.15}$$

Proof. Under Assumption (A11.1), H^{-1} and F^{-1} are continuous. By the above analysis, we have

$$\begin{aligned} \tau_N &= H^{-1}(C\mathbb{1} - F^{-1}(\widetilde{\xi}_N)) \\ &\to H^{-1}(C\mathbb{1} - F^{-1}(\widetilde{p})) = H^{-1}(\zeta) = \tau \quad \text{w.p.1 as} \quad N \to \infty. \square \end{aligned}$$

Parameter Estimates of the Original Problem

Parameter estimates are generated as follows. Define $\xi_N^0 = \frac{1}{N}\sum_{l=0}^{N-1} S_l$ and

$$\xi_N = \begin{cases} \xi_N^0, & \text{if} \quad z\mathbb{1} \le \xi_N^0 \le 1, \\ z\mathbb{1}, & \text{if} \quad \xi_N^0 > z\mathbb{1}, \\ (1 - z)\mathbb{1}, & \text{if} \quad \xi_N^0 < (1 - z)\mathbb{1}. \end{cases} \tag{11.16}$$

Then, we have

$$\xi_N \to \xi = F(C\mathbb{1} - H(\widetilde{\Phi}\theta, \eta)) \text{ w.p.1 as } N \to \infty. \tag{11.17}$$

Denote $\xi^{\{i:j\}}$ as the vector of the ith to jth components of ξ. Then, equations in (11.17) for system (11.1) contain the following equations by extracting the odd-indexed blocks:

$$H(\rho_j \Phi\theta; \eta) = C\mathbb{1} - F^{-1}(\xi^{\{2jn_0+1:2jn_0+n_0\}}),\ j = 0, \ldots, m_0. \tag{11.18}$$

We now show that this subset of equations is sufficient to determine θ and η uniquely.

Theorem 11.7. *Suppose $u \in \mathcal{U}$. Under Assumptions* (A11.1) *and* (A11.2)*,*

$$F(C\mathbb{1} - H(\widetilde{\Phi}\theta, \eta)) = \xi \tag{11.19}$$

has a unique solution (θ^*, η^*).

Proof. Consider the first block $\Phi_1\theta$ of $\widetilde{\Phi}\theta$. Since v is full rank, Φ_1 is a full-rank matrix. It follows that for any nonzero θ, $\Phi_1\theta \neq 0_{n_0}$. Without loss of generality, suppose that the i^*th element δ of $\Phi_1\theta$ is nonzero. By construction of $\widetilde{\Phi}$, we can extract the following m_0 nonzero elements from $\widetilde{\Phi}\theta$: The $(2nl + i^*)$th element, $l = 0, \ldots, m_0$, is $\rho_l\delta$. Extracting these rows from the equation $H(\widetilde{\Phi}\theta, \eta) = C\mathbb{1} - F^{-1}(\xi)$ leads to a core problem

$$H(\rho\delta, \eta) = C\mathbb{1} - F^{-1}(\widetilde{\xi}), \tag{11.20}$$

where $\rho = [\rho_0, \rho_1, \ldots, \rho_{m_0}]'$. Since $\delta \neq 0$ and ρ has distinct elements, $\rho\delta$ has distinct elements. By hypothesis, $H(x; \eta)$ is jointly identifiable. It follows that (11.20) has a unique solution (δ^*, η^*).

From the derived η^*, we denote the first n_0 equations of $H(\widetilde{\Phi}\theta, \eta) = C\mathbb{1} - F^{-1}(\xi)$ by

$$H(\Phi\theta, \eta^*) = C\mathbb{1} - F^{-1}(\xi^{\{1:n_0\}}). \tag{11.21}$$

By Assumption (A11.1), $H^{-1}(x; \eta^*)$ exists (as a function of x). Since v is full rank, $\Phi = T([v_{n_0}, \ldots, v_1])$ is invertible. As a result,

$$\theta^* = \Phi^{-1} H^{-1}(C\mathbb{1} - F^{-1}(\xi^1), \eta^*)$$

is the unique solution to (11.21). This completes the proof. □

A particular choice of the scaling factors ρ_j is $\rho_j = q^j$, $j = 0, 1, \ldots, m_0$, for some $q \neq 0$ and $q \neq 1$. In this case, the period of input u can be shortened to $n_0(m_0 + 1)$ under a slightly different condition.

Let ξ_N be defined as in (11.16), with dimension changed from $2n_0(m_0+1)$ to $n_0(m_0 + 1)$. By the strong law of large numbers, as $N \to \infty$,

$$\xi_N^e \to \xi^e = F(C\mathbb{1} - H(\Phi^e\theta, \eta)) \text{ w.p.1}$$

for some $(n_0(m_0+1))\times n_0$ matrix Φ^e. Partition Φ^e into (m_0+1) submatrices Φ_i^e, $i=1,\ldots,m_0+1$, of dimension $n_0\times n_0$:

$$\Phi^e=[(\Phi_1^e)',(\Phi_2^e)',\ldots,(\Phi_{m_0+1}^e)']'.$$

If $u\in\mathcal{U}_e$, then it can be directly verified that

$$\Phi_{l+1}^e=q^l\Phi^e=q^lT(q,[v_{n_0},\ldots,v_1]),\ l=0,1,\ldots,m_0.$$

We have the following result, whose proof is similar to that of Theorem 11.7 and hence is omitted.

Theorem 11.8. *Suppose $u\in\mathcal{U}_e$. Under Assumptions* (A11.1) *and* (A11.2)*,*

$$F(C\mathbb{1}-H(\Phi^e\theta,\eta))=\xi^e$$

has a unique solution (θ^*,η^*).

Identification Algorithms and Convergence

The $\xi_N=[\xi_N^{\{1\}},\ldots,\xi_N^{\{2n_0(m_0+1)\}}]'$ in (11.16) has $2n_0(m_0+1)$ components for a scaled full-rank signal $u\in\mathcal{U}$, but there are only n_0+m_0 unknown parameters. Consider $\Phi\theta=[\delta_0,\ldots,\delta_{n_0-1}]'$. We separate the components to n_0 groups, for $i=1,\ldots,n_0$, $\varepsilon_N(i)=[\xi_N^{\{i\}},\xi_N^{\{i+2n_0\}},\ldots,\xi_N^{\{i+2n_0m_0\}}]'$. Let $\delta_N(i)$ and $\eta_N(i)$ satisfy

$$\begin{aligned}\varepsilon_N(i)&=[\varepsilon_N^{\{1\}}(i),\ldots,\varepsilon_N^{\{m_0+1\}}(i)]'\\&=F(C\mathbb{1}-H(\rho\delta_N(i),\eta_N(i))).\end{aligned}\tag{11.22}$$

Then, by (11.17), we have

$$\varepsilon_N(i)\to\varepsilon_i=F(C\mathbb{1}-H(\delta(i)\rho,\eta)).\tag{11.23}$$

If $\delta(i)\neq 0$, (11.23) becomes a core identification problem. Furthermore, since $\theta\neq\mathbf{0}_{n_0}$ and Φ is full rank, there exists i^* such that $\delta(i^*)\neq 0$. The identification algorithms include the following steps:

1. Calculate $i^*=\text{argmax}_i|\delta_i|$ to choose nonzero $\delta(i^*)$. If there exists $j\neq k$ such that $\varepsilon_N^{\{j\}}(i)=\varepsilon_N^{\{k\}}(i)$, then let $\delta_N(i)=0$ and $\eta_N(i)=0_{m_0}$. Otherwise, $\delta_N(i)$ and $\eta_N(i)$ are solved from (11.22). Let $i_N^*=\text{argmax}_i|\delta_N(i)|$, where "argmax" means the argument of the maximum.

2. Estimate η from the core identification problem, $\eta_N=\eta_N(i^*)$.

3. Estimate θ: $\theta_N=\Phi^{-1}H^{-1}(C\mathbb{1}-F^{-1}(\xi_N^*),\eta_N)$, where $\xi_N^*=[\xi_N^{\{1\}},\xi_N^{\{2\}},\ldots,\xi_N^{\{n_0\}}]'$.

Theorem 11.9. *Suppose* $u \in \mathcal{U}$. *Under Assumptions* (A11.1) *and* (A11.2),

$$\theta_N \to \theta \quad \textit{and} \quad \eta_N \to \eta \quad \textit{w.p.1 as} \quad N \to \infty.$$

Proof. By Assumption (A11.1), $\delta_N(i)$ and $\eta_N(i)$ can be solved from step 1. By core identification problems, if $\delta(i) \neq 0$, $\delta_N(i) \to \delta(i)$ w.p.1 as $N \to \infty$. Hence,

$$i_N^* = \text{argmax}_i |\delta_N(i)| \to i^* = \text{argmax}_i |\delta(i)| \quad \text{w.p.1.}$$

Since there exists $\delta(i) \neq 0$, we have $\delta_{i^*} \neq 0$. By (11.15), we have $\delta_N \to \delta(i^*)$, $\eta_N \to \eta$, as w.p.1 as $N \to \infty$. For $\xi_N^* = [\xi_N^{*\{1\}}, \xi_N^{*\{2\}}, \ldots, \xi_N^{*\{n_0\}}]'$, $\xi_N^* \to \xi^* = F(C\mathbb{1} - H(\Phi\theta, \eta))$ w.p.1, so as $N \to \infty$,

$$\begin{aligned} \theta_N &= \Phi^{-1} H^{-1}(C\mathbb{1} - F^{-1}(\xi_N^*), \eta_N) \\ &\to \Phi^{-1} H^{-1}(C\mathbb{1} - F^{-1}(\xi^*), \eta) = \theta \quad \text{w.p.1.} \end{aligned}$$

□

Similarly, for an exponentially scaled full-rank signal $u \in \mathcal{U}_e$, the identification algorithms can be constructed and its convergence can be derived similarly.

11.5 Asymptotic Efficiency of the Core Identification Algorithms

The identification of the core problem uses the main idea of the algorithms constructed in Section 11.4. In this section, the efficiency of the core identification algorithms will be established by comparing the error variance with the CR lower bound.

Asymptotic Analysis of Identification Errors

The following analysis of identification errors is generic, and hence is described without reference to specific algorithms. For simplicity, for $x \in \mathbb{R}$, denote $B(x) = C - F^{-1}(x)$. Then, by (11.14), we have

$$p = [p^{\{1\}}, \ldots, p^{\{m_0\}}]' = F(C\mathbb{1} - \zeta) = B^{-1}(\zeta), \tag{11.24}$$

where ζ is denoted as $\zeta = [\zeta^{\{1\}}, \ldots, \zeta^{\{m_0+1\}}]'$. Let $g(\zeta) = [g_0(\zeta), \ldots, g_{m_0}(\zeta)]' = H^{-1}(\zeta)$. Then, ζ_N, τ_N in Theorem 11.6 and $\tau = [\tau^{\{1\}}, \ldots, \tau^{\{m_0+1\}}]'$ can be written as

$$\zeta_N = B(\widetilde{\xi}_N), \ \tau_N = g(\zeta_N), \ \tau = g(B(p)). \tag{11.25}$$

The estimation error for τ is $e_N = [e_N^{\{1\}}, \ldots, e_N^{\{m_0+1\}}]' = \tau_N - \tau$.

For $\tau = g(\zeta)$, the Jacobian matrix is

$$J(g(\zeta)) = \frac{\partial g(\zeta)}{\partial \zeta} = \begin{bmatrix} \frac{\partial g_0(\zeta)}{\partial \zeta^{\{1\}}} & \cdots & \frac{\partial g_0(\zeta)}{\partial \zeta^{\{m_0+1\}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_{m_0}(\zeta)}{\partial \zeta^{\{1\}}} & \cdots & \frac{\partial g_{m_0}(\zeta)}{\partial \zeta^{\{m_0+1\}}} \end{bmatrix},$$

and for $\zeta = H(\tau)$,

$$J(H(\tau)) = \frac{\partial H(\tau)}{\partial \tau} = \begin{bmatrix} \frac{\partial h_0(\tau)}{\partial \tau^{\{1\}}} & \cdots & \frac{\partial h_0(\tau)}{\partial \tau^{\{m_0+1\}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{m_0}(\tau)}{\partial \tau^{\{1\}}} & \cdots & \frac{\partial h_{m_0}(\tau)}{\partial \tau^{\{m_0+1\}}} \end{bmatrix}.$$

Since $\zeta = H(\tau)$, we have

$$J(g(\zeta))J(H(\tau)) = \frac{\partial g(\zeta)}{\partial \zeta}\frac{\partial H(\tau)}{\partial \tau} = \frac{\partial g(H(\tau))}{\partial \tau} = \frac{\partial \tau}{\partial \tau} = I_{m_0+1}.$$

As a result, $J(g(\zeta)) = J(H(\tau))^{-1}$. From (11.24), we have $\zeta^{\{i\}} = B(p^{\{i\}})$, $i = 0, 1, \ldots, m_0$. It follows that the Jacobian matrix for $\zeta = B(p)$ is

$$J(B(p)) = \operatorname{diag}\left(\frac{\partial B(p^{\{1\}})}{\partial p^{\{1\}}}, \ldots, \frac{\partial B(p^{\{m_0+1\}})}{\partial p^{\{m_0+1\}}}\right),$$

and for $p = B^{-1}(\zeta)$,

$$J(B^{-1}(p)) = \operatorname{diag}\left(\frac{\partial B^{-1}(\zeta^{\{1\}})}{\partial \zeta^{\{1\}}}, \ldots, \frac{\partial B^{-1}(\zeta^{\{m_0+1\}})}{\partial \zeta^{\{m_0+1\}}}\right).$$

Theorem 11.10. *Under Assumptions* (A11.1) *and* (A11.2),

$$N\sigma^2(e_N) = NEe_N e_N' \to \Lambda \quad as \quad N \to \infty, \tag{11.26}$$

where $\Lambda = WVW'$ *with* $W = J(g(\zeta))J(B(p))$ *and*

$$V = \operatorname{diag}(p^{\{1\}}(1 - p^{\{1\}}), \ldots, p^{\{m_0\}}(1 - p^{\{m_0\}})).$$

Proof. Consider

$$e_N^{\{i+1\}} = \tau_N^{\{i+1\}} - \tau^{\{i+1\}} = g_i(\zeta_N) - g_i(\zeta), \quad i = 0, \ldots, m_0,$$

where $\zeta_N = [\zeta_N^{\{1\}}, \ldots, \zeta_N^{\{m_0+1\}}]'$, $\tau(N) = [\tau_N^{\{1\}}, \ldots, \tau_N^{\{m_0+1\}}]'$, and τ and ζ are given by (11.6) and (11.25), respectively. Denote

$$\begin{aligned} \Omega_N = \; & [\min\{\zeta_N^{\{1\}}, \zeta^{\{1\}}\}, \max\{\zeta_N^{\{2\}}, \zeta^{\{2\}}\}] \times \cdots \\ & \times [\min\{\zeta_N^{\{m_0+1\}}, \zeta^{\{m_0+1\}}\}, \max\{\zeta_N^{\{m_0+1\}}, \zeta^{\{m_0+1\}}\}] \end{aligned}$$

as the Cartesian product ([79, p. 3]) of the sets $[\min\{\zeta_N^{\{i\}}, \zeta^{\{i\}}\}, \max\{\zeta_N^{\{i\}}, \zeta^{\{i\}}\}]$, for $i = 1, \ldots, m_0 + 1$.

For $j = 1, \ldots, m_0$, denote

$$\widetilde{\zeta}_N(j) = [\zeta^{\{1\}}, \ldots, \zeta^{\{j\}}, \zeta_N^{\{j+1\}}, \ldots, \zeta_N^{\{m_0+1\}}]',$$

$\widetilde{\zeta}_N(0) = [\zeta_N^{\{1\}}, \ldots, \zeta_N^{\{m_0+1\}}]'$, and $\widetilde{\zeta}_N(m_0 + 1) = \zeta$. Then

$$\begin{aligned} e_N^{\{i\}} &= g_i(\zeta_N) - g_i(\zeta) \\ &= \sum_{j=-1}^{m_0-1} [g_i(\widetilde{\zeta}_N(j)) - g_i(\widetilde{\zeta}_N(j+1))]. \end{aligned}$$

Since $H(\cdot)$ is continuous, by the well-known mean-value theorem, there exists $\lambda_N(i, j) \in \Omega_N$ for $j = 0, \ldots, m_0$ such that

$$g_i(\widetilde{\zeta}_N(j)) - g_i(\widetilde{\zeta}_N(j+1)) = \frac{\partial g_i(\lambda_N(i,j))}{\partial \zeta^{\{j\}}}(\zeta_N^{\{j\}} - \zeta^{\{j\}}),$$

which implies

$$\begin{aligned} e_N^{\{i\}} &= \sum_{j=0}^{m_0} \frac{\partial g_i(\lambda_N(i,j)),}{\partial \zeta^{\{j\}}}(\zeta_N^{\{j\}} - \zeta^{\{j\}}) \\ &= \left[\frac{\partial g_i(\lambda_N(i,0))}{\partial \zeta^{\{1\}}}, \ldots, \frac{\partial g_i(\lambda_N(i,m_0))}{\partial \zeta^{\{m_0+1\}}}\right](\zeta_N - \zeta). \end{aligned}$$

Thus,

$$e_N = L_N(\zeta_N - \zeta), \tag{11.27}$$

where

$$L_N = \begin{bmatrix} \frac{\partial g_0(\lambda_N(0,0))}{\partial \zeta^{\{1\}}} & \cdots & \frac{\partial g_0(\lambda_N(0,m_0))}{\partial \zeta^{\{m_0+1\}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_{m_0}(\lambda_N(m_0,0))}{\partial \zeta^{\{1\}}} & \cdots & \frac{\partial g_{m_0}(\lambda_N(m_0,m_0))}{\partial \zeta^{\{m_0+1\}}} \end{bmatrix}.$$

Since $\zeta_N^{\{i\}} = B(\widetilde{\xi}_N(i))$, $i = 1, \ldots, m_0 + 1$, by the mean-value theorem, there exists $\kappa_N(i)$ on the line segment $\widetilde{\xi}_N(i)$ and $p^{\{i\}}$ such that

$$\zeta_N - \zeta = \operatorname{diag}\left(\frac{\partial B(\kappa_N(1))}{\partial p^{\{1\}}}, \ \ldots, \ \frac{\partial B(\kappa_N(m_0+1))}{\partial p^{\{m_0+1\}}}\right)(\widetilde{\xi}_N - p). \tag{11.28}$$

Moreover, as $N \to \infty$,

$$L_N = \operatorname{diag}\left(\frac{\partial B(\kappa_N(1))}{\partial p^{\{1\}}}, \ \ldots, \ \frac{\partial B(\kappa_N(m_0+1))}{\partial p^{\{m_0+1\}}}\right) \to W \quad \text{w.p.1.} \tag{11.29}$$

Hence, (11.26) is true. □

CR Lower Bound and Asymptotic Efficiency

Consider N blocks of m_0+1 observations for the core identification problem. We first derive the CR lower bound based on these $N(m_0+1)$ observations. The CR lower bound is denoted as $\sigma^2_{\rm CR}(N)$. To proceed, we first derive a lemma and then Theorem 11.12 follows.

Lemma 11.11. *The CR lower bound for estimating the parameter τ, based on observations of $\{\widetilde{S}_k\}$, is $\sigma^2_{\rm CR}(N) = \Lambda/N$.*

Proof. Let x_k take values in $\{0,1\}$. The likelihood function, which is the joint distribution of $\widetilde{S}_1, \ldots, \widetilde{S}_{N(m_0+1)}$, depending on

$$\tau = [\tau^{\{1\}}, \ldots, \tau^{\{m_0+1\}}]' = [\delta, \eta']',$$

is given by

$$\begin{aligned} l_N &= P\{\widetilde{S}_1 = x_1, \ldots, \widetilde{S}_{N(m_0+1)} = x_{N(m_0+1)}; \tau\} \\ &= \prod_{k=1}^{m_0+1} P\{\widetilde{S}_{kN+1} = x_{kN+1}, \ldots, \widetilde{S}_{kN+m_0+1} = x_{(k+1)N}; \tau\}. \end{aligned}$$

Replace the x_k's by their corresponding random elements $\widetilde{S}_k$'s, and denote the resulting quantity by l in short. Then, we have

$$\begin{aligned} \log l_N &= \log \left[\prod_{k=1}^{m_0+1} p(k,\tau)^{N\widetilde{\xi}_N(k)} (1-p(k,\tau))^{N(1-\widetilde{\xi}_N(k))} \right] \\ &= N \sum_{k=1}^{m_0+1} [\widetilde{\xi}_N(k) \log p(k,\tau) + (1-\widetilde{\xi}_N(k)) \log(1-p(k,\tau))], \\ \frac{\partial \log l_N}{\partial \tau^{\{i\}}} &= N \sum_{k=1}^{m_0+1} \left(\frac{\widetilde{\xi}_N(k)}{p^{\{k\}}} - \frac{1-\widetilde{\xi}_N(k)}{1-p^{\{k\}}} \right) \frac{\partial p^{\{k\}}}{\partial \zeta^{\{k\}}} \frac{\partial \zeta^{\{k\}}}{\partial \tau^{\{i\}}}, \\ \frac{\partial \log l_N}{\partial \tau} &= \Big[\frac{\partial \log l_N}{\partial \tau^{\{1\}}}, \ldots, \frac{\partial \log l_N}{\partial \tau^{\{m_0+1\}}} \Big]'. \end{aligned}$$

Furthermore, for $i, j = 0, \ldots, m_0$,

$$\begin{aligned} \frac{\partial^2 \log l_N}{\partial \tau^{\{i\}} \partial \tau^{\{j\}}} = \; & N \sum_{k=1}^{m_0+1} \Big[\Big(- \frac{\widetilde{\xi}_N(k)}{(p^{\{k\}})^2} - \frac{1-\widetilde{\xi}_N(k)}{(1-p^{\{k\}})^2} \Big) \frac{\partial p^{\{k\}}}{\partial \tau^{\{i\}}} \frac{\partial p^{\{k\}}}{\partial \tau^{\{j\}}} \\ & + \Big(\frac{\widetilde{\xi}_N(k)}{p^{\{k\}}} - \frac{1-\widetilde{\xi}_N(k)}{1-p^{\{k\}}} \Big) \frac{\partial^2 p^{\{k\}}}{\partial \tau^{\{i\}} \partial \tau^{\{j\}}} \Big]. \end{aligned}$$

As a result,

$$\begin{aligned} E\frac{\partial^2 \log l_N}{\partial\tau^{\{i\}}\partial\tau^{\{j\}}} &= NE\sum_{k=1}^{m_0+1}\Big[\Big(-\frac{\widetilde{\xi}_N(k)}{(p^{\{k\}})^2}-\frac{1-\widetilde{\xi}_N(k)}{(1-p^{\{k\}})^2}\Big)\frac{\partial p^{\{k\}}}{\partial\tau^{\{i\}}}\frac{\partial p^{\{k\}}}{\partial\tau^{\{j\}}} \\ &\quad +\Big(\frac{\widetilde{\xi}_N(k)}{p^{\{k\}}}-\frac{1-\widetilde{\xi}_N(k)}{1-p^{\{k\}}}\Big)\frac{\partial^2 p^{\{k\}}}{\partial\tau^{\{i\}}\partial\tau^{\{j\}}}\Big] \\ &= -N\sum_{k=1}^{m_0+1}\frac{1}{p^{\{k\}}(1-p^{\{k\}})}\frac{\partial p^{\{k\}}}{\partial\tau^{\{i\}}}\frac{\partial p^{\{k\}}}{\partial\tau^{\{j\}}} \\ &= -N\sum_{k=1}^{m_0+1}\frac{1}{p^{\{k\}}(1-p^{\{k\}})}\Big(\frac{\partial p^{\{k\}}}{\partial\zeta^{\{k\}}}\Big)^2\frac{\partial\zeta^{\{k\}}}{\partial\tau^{\{i\}}}\frac{\partial\zeta^{\{k\}}}{\partial\tau^{\{j\}}}, \end{aligned}$$

and

$$E\frac{\partial^2 \log l_N}{\partial\tau\partial\tau} = -NW^{-1}V^{-1}(W')^{-1}.$$

The CR lower bound is then given by

$$\sigma^2_{\rm CR}(N) = -\Big(E\frac{\partial^2 \log l_N}{\partial\tau\partial\tau}\Big)^{-1} = \frac{WVW'}{N} = \frac{\Lambda}{N}.$$

□

Theorem 11.12. *Under Assumptions* (A11.1) *and* (A11.2),

$$N[\sigma^2(e_N) - \sigma^2_{\rm CR}(N)] \to 0 \quad as \quad N\to\infty.$$

Proof. This follows directly from Theorem 11.10 and Lemma 11.11. □

11.6 Recursive Algorithms and Convergence

This section develops a recursive algorithm for estimating (θ^*, η^*). The essence is to treat the parameters (θ, η) jointly. Define $\Theta = [\theta', \eta']' \in \mathbb{R}^{(n_0+m_0)\times 1}$. For an $(n_0 + m_0) \times 2n_0(m_0 + 1)$ matrix M, and for each $\widetilde{\xi}$, define

$$G(\Theta, \widetilde{\xi}) = M[\widetilde{\xi} - F(C\mathbb{1}_{2n_0(m_0+1)} - H(\widetilde{\Phi}\theta, \eta)]. \tag{11.30}$$

It is easily seen that the purpose of the matrix M is to make the function under consideration "compatible" with the dimension of the vector Θ. We use the following recursive algorithm for parameter estimation:

$$\begin{cases} \xi_{k+1} = \xi_k - \dfrac{1}{k+1}\xi_k + \dfrac{1}{k+1}S_{k+1}, \\ \Theta_{k+1} = \Theta_k + \beta_k G(\Theta_k, \xi_k), \quad k = 0, 1, \ldots, \end{cases} \tag{11.31}$$

where S_{k+1} is defined in (11.11). In the above algorithm, β_k is a sequence of step sizes satisfying $\beta_k \geq 0$, $\sum_{k=1}^{\infty} \beta_k = \infty$, $\beta_k \to 0$, and

$$\frac{\beta_k - \beta_{k+1}}{\beta_k} = O(\beta_k) \quad \text{as} \quad k \to \infty. \tag{11.32}$$

Take, for instance, $\beta_k = 1/k^\alpha$ with $0 < \alpha \leq 1$. Then, condition (11.32) is satisfied. Commonly used step sizes include $\beta_k = O(1/k^\alpha)$ with $\alpha \in (1/2, 1]$.

Associated with (11.31), consider an ordinary differential equation (ODE)

$$\dot{\Theta} = \overline{G}(\Theta), \tag{11.33}$$

where $\overline{G}(\Theta) = M(\xi - F(C\mathbb{1}_{2(m_0+1)n_0} - H(\widetilde{\Phi}\theta, \eta))$. Θ^* is the unique stationary point of (11.33). To proceed, we assume the following assumption holds.

(A11.3) The ODE (11.33) has a unique solution for each initial condition; $\Theta^* = (\theta^*, \eta^*)$ is an asymptotically stable point of (11.33); $H(\cdot, \cdot)$ is continuous in its arguments together with its inverse.

Remark 11.13. A sufficient condition to ensure the asymptotic stability of (11.33) can be obtained by linearizing $M[\xi - F(C\mathbb{1}_{2n_0(m_0+1)} - H(\widetilde{\Phi}\theta, \eta))]$ about its stationary point Θ^*. Under this linearization, if the Jacobian matrix $-M(\partial F(C\mathbb{1} - H(\widetilde{\Phi}\theta^*, \eta^*))/\partial\Theta)$ is a stable matrix (that is, all of its eigenvalues are on the left-hand side of the complex plane), the required asymptotic stability follows.

Theorem 11.14. *Under Assumptions* (A11.1)–(A11.3), $\xi_k \to \xi$ *and* $\Theta_k \to \Theta^*$ *w.p.1 as* $k \to \infty$.

Proof. Note that we have already proved that $\xi_k \to \xi$ w.p.1. Thus, to obtain the desired result, we need only establish the convergence of $\{\Theta_k\}$. To this end, we use the ODE methods to complete the proof.

We will use the basic convergence theorem ([55, Theorem 6.1.1, p. 166]). Thus, we need only verify the conditions in the aforementioned theorem hold. Note that we do not have a projection now, but in our recursion $\mathbf{F}$ is used and is uniformly bounded. In view of assumptions (A11.1)–(A11.3), as explained in [55, Section 6.2, p. 170], to verify the conditions in the theorem, we need only show that a "rate of change" condition (see [55, p. 137], for a definition) is satisfied. Thus, the remaining proof is to verify this condition.

Define $t_0 = 0$, $t_k = \sum_{i=0}^{k-1} \beta_i$, and let $m(t)$ be the unique value k such that $t_k \leq t < t_{k+1}$ when $t \geq 0$, and set $m(t) = 0$ when $t < 0$. Define the piecewise-constant interpolation as $\Theta^0(t) = \Theta_k$ for $t_k \leq t < t_{k+1}$, and define the shifted sequence by $\Theta^k(t) = \Theta^0(t + t_k)$, $t \in (-\infty, \infty)$. Using the ODE methods, we can show that the sequence of functions $\Theta^k(\cdot)$ converges

to the solution of desired limit ODE. For $m = 1, 2, \ldots$, and a fixed Θ, denote

$$\Xi(m) = \sum_{i=0}^{m-1} [G(\Theta, \xi(i)) - \overline{G}(\Theta)],$$

and $\Xi_0 = 0$. In view of (11.30), $G(\cdot, \cdot)$ is a continuous function in both variables.

We note that by a partial summation, for any $m, j \geq 0$,

$$\sum_{i=j}^{m} \beta_i [G(\Theta, \xi(i)) - \overline{G}(\Theta)]$$
$$= \beta_m \Xi(m+1) - \beta_m \Xi(j) + \sum_{i=j}^{m-1} [\Xi(i+1) - \Xi_j](\beta_i - \beta_{i+1}).$$

Taking $m = m(t) - 1$ and $j = 0$, and recalling $\Xi_0 = 0$, we obtain

$$\sum_{i=0}^{m(t)-1} \beta_i [G(\Theta, \xi(i)) - \overline{G}(\Theta)]$$
$$= \beta_{m(t)} \Xi(m(t)) + \sum_{i=0}^{m(t)-2} \Xi(i+1) \frac{\beta_i - \beta_{i+1}}{\beta_i} \beta_i.$$

It is readily seen that as $k \to \infty$, $\beta_k \Xi_k \to 0$ w.p.1. Thus, the asymptotic rate of change of $\sum_{i=0}^{m(t)-1} \beta_i [G(\Theta, \xi_i) - \overline{G}(\Theta)]$ is zero w.p.1. Then by virtue of Theorem 6.1.1 in [55], the limit ODE is precisely (11.33). The asymptotic stability of the ODE then leads to the desired result. □

Remark 11.15. Note that in (11.31), we could include additional random noises (representing the measurement noise and other external noise). The treatment remains essentially the same. We choose the current setup for notational simplicity.

11.7 Examples

In this section, we illustrate the convergence of estimates from the algorithms developed in this chapter. The noise is Gaussian with zero mean and known variance, although the algorithms are valid for all distribution functions that satisfy Assumption (A3.1). The identification algorithm of Section 11.4 is shown in Example 11.16, and the asymptotic efficiency is also illustrated for the core identification problem. Example 11.17 illustrates the recursive algorithm. The estimates of parameters are shown to be convergent in both cases.

Example 11.16. Consider

$$\begin{cases} y_k = H(x_k, \eta) + d_k = b_0 + e^{x_k} + d_k, \\ x_k = a_1 u_{k-1} + a_2 u_{k-2}, \end{cases} \tag{11.34}$$

where the noise $\{d_k\}$ is a sequence of i.i.d. normal random variables with $Ed_1 = 0$, $\sigma_d^2 = 1$. For a normal distribution, the support is $(-\infty, \infty)$. The output is measured by a binary-valued sensor with threshold $C = 3$. The linear subsystem has order $n_0 = 2$. The nonlinear function is parameterized as $b_0 + e^x$. The prior information on b_0 and a_i, $i = 1, 2$, is that $b_0, a_i \in [0.5, 5]$. Suppose the true values of the unknown parameters are $\theta = [a_1, a_2] = [0.7, 0.63]$ and $\eta = b_0 = 1.1$.

For $n_0 = 2$ and $m_0 = 1$, the input should be $2n_0(m_0 + 1) = 8$-periodic with single period $u = [\rho_0 v, \rho_0 v, \rho_1 v, \rho_1 v]$. Using the results of the section on joint identifiability, $H(x, \eta)$ is jointly identifiable with respect to $\Upsilon = \{(\rho_0, \rho_1) : \rho_0 > 0, \rho_1 < 0\}$. Let $v = [1, 1.2]$, $\rho_0 = 1$, and $\rho_1 = -1$. Define the block variables $X_j, Y_j, \widetilde{\Phi}_j, D_j$, and S_j, in the case of an 8-periodic input, $\widetilde{\Phi}_j = \widetilde{\Phi} = [\Phi_1', \ldots, \Phi_4']'$, where $\Phi_1 = \rho_0 \Phi = \Phi = \begin{bmatrix} v_2 & v_1 \\ v_1 & v_2 \end{bmatrix}$ and $\Phi_3 = \rho_1 \Phi$. Using (11.16), we can construct the algorithms as described above Theorem 11.9.

The estimates of θ and η are shown in Figure 11.2, where the errors are measured by the Euclidean norm. The algorithms are simulated five times. It is shown that both parameter estimates of the linear and nonlinear subsystems converge to their true values. In this simulation η_N demonstrates a higher convergence speed than θ_N. A possible explanation is that η_N is updated first, and then used to obtain θ_N. As a result, convergence of θ_N can occur only after the error $\eta_N - \eta$ is reduced.

To understand the reliability of the estimation schemes, the estimation algorithms are performed 500 times of total data length 2000 each. Estimation errors for each run are recorded at $N = 500$, $N = 1000$, and $N = 2000$. The error distributions are calculated by histograms in Figure 11.3, which illustrate improved estimation accuracy with respect to data length N and are consistent with the theoretical analysis.

Consider the core identification problem of (11.34),

$$\widetilde{Y}_l = H(\rho\delta, \eta) + \widetilde{D}_l = b_0 \mathbb{1}_2 + e^{\rho\delta} + \widetilde{D}_l,$$

where $\delta = a_0 v_2 + a_1 v_1 \neq 0$ and $\rho = [\rho_0, \rho_1]'$. The convergence of $N[\sigma^2(e_N) - \sigma_{\text{CR}}^2(N)]$ in Theorem 11.12 is shown in Figure 11.4, where the error is measured by the Frobenius norm.

Example 11.17. We use the same system and inputs as in Example 11.16. The recursive algorithms in Section 11.6 are now used.

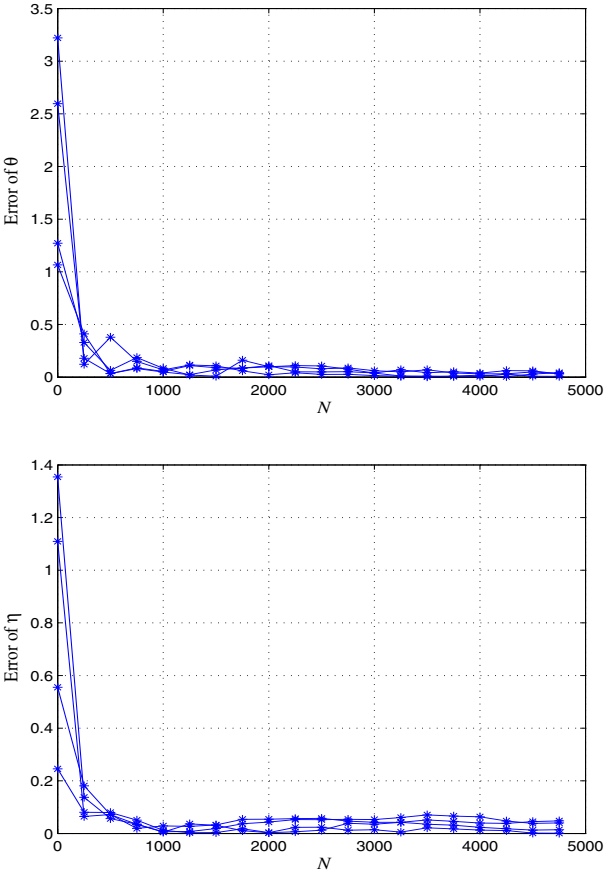


FIGURE 11.2. Joint identification errors of θ and η

Let $\rho_1 = 0.5$ and $\Theta = [\theta', \eta']'$. For system (11.34), the ODE (11.33) becomes

$$\dot{\Theta} = M[\xi - F((C-b)\mathbb{1}_8 - \exp(\widetilde{\Phi}\theta))].$$

Choose $\beta_k = 1/k$ and

$$M = -\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

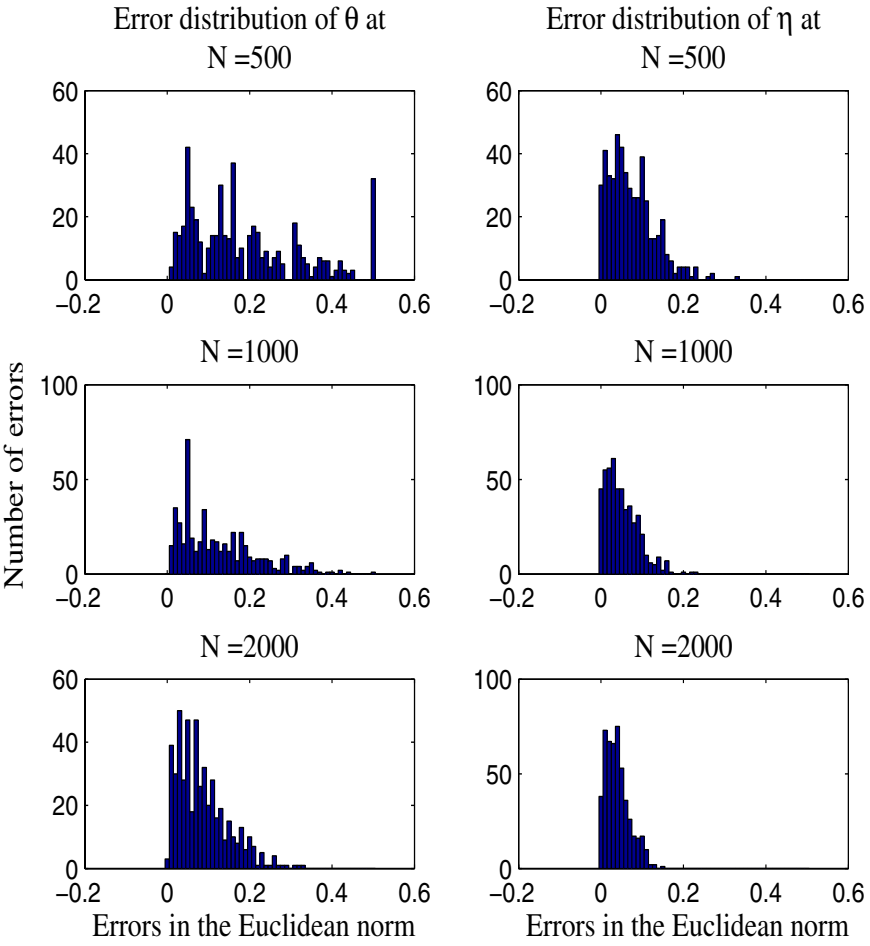


FIGURE 11.3. Estimation error distributions

Then the Jacobian matrix can be calculated to be

$$\begin{aligned} J(\Theta) &= -M[\partial F((C-b)\mathbb{1}_8 - \exp(\widetilde{\Phi}\theta))/\partial\Theta] \\ &= \begin{bmatrix} -0.660 & -0.247 & -0.429 \\ -0.242 & -0.645 & -0.434 \\ -0.210 & -0.079 & -0.397 \end{bmatrix}. \end{aligned}$$

The eigenvalues of $J(\Theta)$ are $[-1.08, -0.402, -0.220]$, which are all less than 0. As a result, the Jacobian matrix $J(\eta)$ is stable.

Let $\Theta_k = [\theta_k', \eta_k']'$ be the estimates of $\Theta = [\theta', \eta']'$. Then the recursive algorithms can be constructed as follows: First, set $\beta_k = 1/k$, $\Theta(1) =$

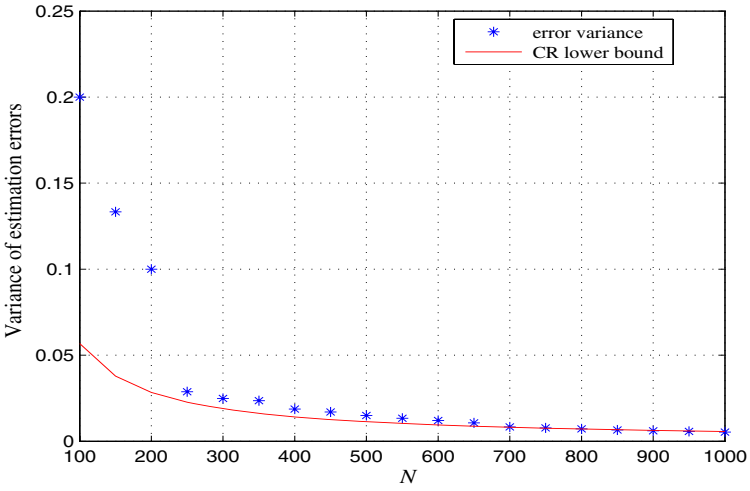


FIGURE 11.4. Asymptotic efficiency

$[1.5, 1.5, 1.5]'$, and $\xi_1 = \mathbf{0}_8$. The estimates are then updated according to (11.31). Convergence of Θ is shown in Figure 11.5, where the errors are measured by the Euclidean norm and the algorithms are simulated five times.

11.8 Notes

In this chapter, the identification of Wiener systems with binary-valued output observations is studied. There is an extensive literature on Wiener model identification under regular sensors. In a way, a binary-valued output sensor becomes an added output nonlinearity to the Wiener model, leading to a different structure than the traditional Wiener model identification. We refer the reader to [3, 15, 16, 44, 51, 56, 72, 113] for more detailed information on typical Wiener model identification methodologies. The material of this chapter is from [127]. Unlike traditional approximate gradient methods or covariance analysis, we employ the methods of empirical measures. Under assumptions of known noise distribution function, invertible nonlinearity, and joint identifiability, identification algorithms, convergence properties, and identification efficiency are derived.

We have assumed that the structure and order of the linear dynamics and nonlinear function are known. The issues of unmodeled dynamics (for the linear subsystem when the system order is higher than the model order) and model mismatch (for the nonlinear part when the nonlinear function does not belong to the model class) are not included in this chapter. Irre-

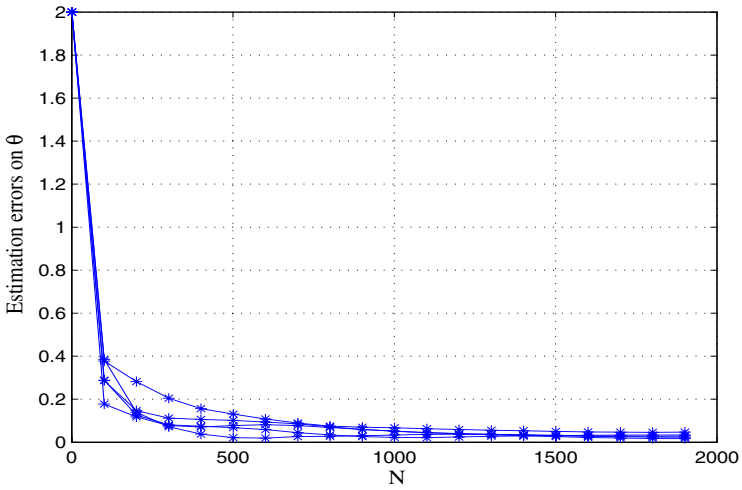


FIGURE 11.5. Estimation errors of Θ using recursive algorithms

ducible identification errors due to unmodeled dynamics were characterized in Chapter 4.

12

Identification of Hammerstein Systems with Quantized Observations

This chapter concerns the identification of Hammerstein systems whose outputs are measured by quantized sensors. The system consists of a memoryless nonlinearity that is polynomial and possibly noninvertible, followed by a linear subsystem. The parameters of linear and nonlinear parts are unknown but have known orders. We present input design, identification algorithms, and their essential properties under the assumptions that the distribution function of the noise and the quantization thresholds are known. Also introduced is the concept of strongly scaled full-rank signals to capture the essential conditions under which the Hammerstein system can be identified with quantized observations. Then under strongly scaled full-rank conditions, we construct an algorithm and demonstrate its consistency and asymptotic efficiency.

The structure of Hammerstein models using quantized observations is formulated in Section 12.1. The concepts of strongly full-rank signals and their essential properties are introduced in Section 12.2. Under strongly full rank inputs, estimates of unknown parameters based on individual thresholds are constructed in Section 12.3. Estimation errors for these estimates are established. The estimates are integrated in an optimal quasi-convex combination estimator (QCCE) in Section 12.4. The resulting estimates are shown to be strongly convergent. Their efficiency is also investigated. The algorithms are expanded in Section 12.5 to derive identification algorithms for both the linear and nonlinear parts. Illustrative examples are presented in Section 12.6 on input design and convergence properties of the methodologies and algorithms.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_12,

12.1 Problem Formulation

Consider the system in Figure 12.1, in which

$$\begin{cases} y_k = \sum_{i=0}^{n_0-1} a_i x_{k-i} + d_k, \\ x_k = b_0 + \sum_{j=1}^{q_0} b_j u_k^j, \quad b_{q_0} = 1, \end{cases}$$

where u_k is the input, x_k the intermediate variable, and d_k the measurement noise. Both n_0 and q_0 are known.

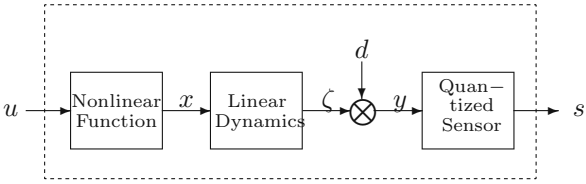


FIGURE 12.1. Hammerstein systems with quantized observations

The output y_k is measured by a sensor, which is represented by the indicator functions

$$s_k^{\{i\}} = I_{\{y_k \le C_i\}}, \quad i = 1, \ldots, m_0,$$

where C_i for $i = 1, \ldots, m_0$ are the thresholds. Denote $\theta = [a_0, \ldots, a_{n_0-1}]'$, $\phi_k^0 = [1, \ldots, 1]'$, and $\phi_k^j = [u_k^j, \ldots, u_{k-n_0+1}^j]'$, $j = 1, \ldots, q_0$. Then

$$\begin{aligned} y_k &= \sum_{i=0}^{n_0-1} a_i (b_0 + \sum_{j=1}^{q_0} b_j u_{k-i}^j) + d_k \\ &= b_0 \sum_{i=0}^{n_0-1} a_i + \sum_{j=1}^{q_0} b_j \sum_{i=0}^{n_0-1} a_i u_{k-i}^j + d_k \\ &= \sum_{j=0}^{q_0} b_j (\phi_k^j)' \theta + d_k. \end{aligned} \tag{12.1}$$

By using the vector notation, for $k = 1, 2, \ldots$,

$$\begin{aligned} Y_l &= [y_{2(l-1)n_0(q_0+1)+n_0}, \ldots, y_{2ln_0(q_0+1)+n_0-1}]' \in \mathbb{R}^{2n_0(q_0+1)}, \\ \Phi_l^j &= [\phi^j_{2(l-1)n_0(q_0+1)+n_0}, \ldots, \phi^j_{2ln_0(q_0+1)+n_0-1}]' \in \mathbb{R}^{2n_0(q_0+1)\times n_0}, \\ & \qquad\qquad j = 0, \ldots, q_0, \\ D_l &= [d_{2(l-1)n_0(q_0+1)+n_0}, \ldots, d_{2ln_0(q_0+1)+n_0-1}]' \in \mathbb{R}^{2n_0(q_0+1)}, \\ S_l^{\{i\}} &= [s^{\{i\}}_{2(l-1)n_0(q_0+1)+n_0}, \ldots, s^{\{i\}}_{2ln_0(q_0+1)+n_0-1}]' \in \mathbb{R}^{2n_0(q_0+1)}, \\ & \qquad\qquad i = 1, \ldots, m_0, \quad l = 1, 2, \ldots, \end{aligned} \tag{12.2}$$

we can rewrite (12.1) in block form as

$$Y_l = \sum_{j=0}^{q_0} b_j \Phi_l^j \theta + D_l, \quad l = 1, 2, \ldots \tag{12.3}$$

We proceed to develop identification algorithms of parameters θ and $\eta = [b_0, \ldots, b_{q_0-1}]'$ with the information of the input u_k and the output s_k of the quantized sensor.

The input signal, which will be used to identify the system, is a $2n_0(q_0 + 1)$-periodic signal u whose one-period values are

$$(v, v, \rho_1 v, \rho_1 v, \ldots, \rho_{q_0} v, \rho_{q_0} v),$$

where the base vector $v = (v_1, \ldots, v_{n_0})$ and the scaling factors are to be specified. The scaling factors 1, ρ_1, ..., ρ_{q_0} are assumed to be nonzero and distinct. Under $2n_0(q_0 + 1)$-periodic inputs, we have

$$\Phi_l^j = \Phi_1^j := \Phi^j, \quad l = 1, 2, \ldots$$

Thus, (12.3) can be written as

$$Y_l = \sum_{j=0}^{q_0} b_j \Phi^j \theta + D_l := \zeta + D_l. \tag{12.4}$$

The identification algorithm will be divided into two steps: (i) to estimate ζ (which can be reduced to estimation of gain systems), and (ii) to estimate θ from the estimated ζ.

12.2 Input Design and Strong-Full-Rank Signals

This section is to introduce a class of input signals, called strongly full-rank signals, which will play an important role in what follows. First, some basic properties of periodic signals will be derived.

Recall that an $n_0 \times n_0$ generalized circulant matrix

$$T = \begin{bmatrix} vn_0 & v_{n_0-1} & v_{n_0-2} & \dots & v_1 \\ \lambda v_1 & vn_0 & v_{n_0-1} & \ddots & v_2 \\ \lambda v_2 & \lambda v_1 & vn_0 & \ddots & v_3 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \lambda v_{n_0-1} & \lambda v_{n_0-2} & \lambda v_{n_0-3} & \dots & vn_0 \end{bmatrix} \tag{12.5}$$

is completely determined by its first row $[vn_0, \dots, v_1]$ and λ, which will be denoted by $T(\lambda, [vn_0, \dots, v_1])$. In the special case of $\lambda = 1$, the matrix in (12.5) is called a circulant matrix and will be denoted by $T([vn_0, \dots, v_1])$.

Definition 12.1. An n_0-periodic signal generated from its one-period values $(v_1, \dots, vn_0)$ is said to be *strongly full rank with order* m_0 if the circulant matrices $T([v^i n_0, \dots, v_1^i])$ are all full rank for $i = 1, \dots, m_0$.

Obviously, an n_0-periodic signal generated from $v = (v_1, \dots, vn_0)$ is strongly full rank with order m_0 if it is strongly $m_0 + 1$ full rank. An important property of circulant matrices is the following frequency-domain criterion. By Lemma 2.2, we have the following theorem.

Lemma 12.2. *An n_0-periodic signal generated from $v = (v_1, \dots, vn_0)$ is strongly full rank with order m_0 if and only if for $l = 1, 2, \dots, m_0$,*

$$\gamma_{k,l} = \sum_{j=1}^{n_0} v_j^l e^{-\mathrm{i}\omega_k j}$$

are nonzero at $\omega_k = 2\pi k/n_0$, $k = 1, \dots, n_0$.

Proposition 12.3. *For $n_0 = 1, 2$, an n_0-periodic signal u generated from $v = (v_1, \dots, vn_0)$ is strongly full rank with order m_0 if and only if it is full rank.*

Proof. For $n_0 = 1$, by Corollary 2.3, u is full rank if and only if $\gamma_1 = v_1 \neq 0$. By Lemma 12.2, u is strongly full rank with order m_0 if and only if $\gamma_{1,l} = v_1^l \neq 0, \forall l$. So, $\gamma_1 \neq 0$ is equivalent to $\gamma_{1,l} \neq 0$.

For $n_0 = 2$, by Corollary 2.3, u is full rank if and only if $\gamma_1 = v_2 - v_1 \neq 0$ and $\gamma_2 = v_2 + v_1 \neq 0$, that is, $v_2 \neq \pm v_1$. By Lemma 12.2, u is strongly full rank with order m_0 if and only if

$$\gamma_{1,l}\gamma_{2,l} = (v_1^l e^{-i\pi} + v_2^l e^{-i2\pi})(v_1^l e^{-i2\pi} + v_2^l e^{-i4\pi}) = v_2^{2l} - v_1^{2l}.$$

Thus, we have $\gamma_1\gamma_2 = v_2^2 - v_1^2 \neq 0$ if and only if u is full rank. □

Remark 12.4. For $n_0 > 2$, the conditions of strongly full rank with order m_0 may be different from the conditions of full rank. For example, for $n_0 = 3$ and $l = 1, \ldots, m_0$,

$$\gamma_{1,l}\gamma_{2,l}\gamma_{3,l} = (v_3^l + v_2^l + v_1^l)\Big[(v_2^l - \frac{1}{2}(v_3^l + v_1^l))^2 + \frac{3}{4}(v_3^l - v_1^l)^2\Big] \neq 0$$

is not equivalent to

$$\gamma_1\gamma_2\gamma_3 = (v_3 + v_2 + v_1)\Big[(v_2 - \frac{1}{2}(v_3 + v_1))^2 + \frac{3}{4}(v_3 - v_1)^2\Big] \neq 0$$

except $m_0 = 1$.

Definition 12.5. A $2n_0(m_0 + 1)$-periodic signal u is strongly scaled m_0 full rank if its one-period values are $(v, v, \rho_1 v, \rho_1 v, \ldots, \rho_{m_0} v, \rho_{m_0} v)$, where $v = (v_1, \ldots, vn_0)$ is strongly full rank with order m_0, i.e., $0 \notin \mathcal{F}[v]$; $\rho_j \neq 0$, $\rho_j \neq 1$, $j = 1, \ldots, m_0$, and $\rho_i \neq \rho_j$, $i \neq j$. We use $\mathcal{U}(n_0, q_0)$ to denote the class of such signals.

Definition 12.6. An $n_0(m_0+1)$-periodic signal u is exponentially strongly scaled full rank with order m_0 signal if its one-period values are $(v, \lambda v, \ldots, \lambda^{m_0} v)$, where $\lambda \neq 0$ and $\lambda \neq 1$, and $T_j = T_j(\lambda^j, [v^j n_0, \ldots, v_1^j])$ are all full rank for $j = 1, \ldots, m_0$. We use $\mathcal{U}_\lambda(n_0, q_0)$ to denote this class of input signals.

By Definition 12.6 and Lemma 2.2, we have the following result.

Lemma 12.7. *An $n_0(q_0+1)$-periodic signal u with one-period values $(v, \lambda v, \ldots, \lambda^{m_0} v)$ is exponentially strongly scaled full rank with order m_0 if $\lambda \neq 0$, $\lambda \neq 1$, and for $l = 1, \ldots, m_0$,*

$$\gamma_{k,l} = \sum_{j=1}^{n_0} v_j^l \lambda^{-\frac{jl}{n_0}} e^{-\mathrm{i}\omega_k j}$$

are nonzero at $\omega_k = (2\pi k)/n_0$, $k = 1, \ldots, n_0$.

Remark 12.8. Definitions 12.5 and 12.6 require that $T(\lambda^i, [v^i n_0, \ldots, v_1^i])$, $i = 1, \ldots, m_0$, are all full rank for $\lambda = 1$ and $\lambda \neq 0, 1$, respectively. However, since the event of singular random matrices has probability zero, if v is chosen randomly, almost all v will satisfy the conditions in Definitions 12.5 and 12.6, which will be shown in the following example.

Example 12.9. For $n_0 = 4$, $m_0 = 4$, $\lambda = 0.9$, $v = (0.5997, 0.9357, 0.9841, 1.4559)$ is generated randomly by Matlab, v is strongly 4 full rank since

$$\det(T([v_4, v_3, v_2, v_1])) = 0.4041, \quad \det(T([v_4^2, v_3^2, v_2^2, v_1^2])) = 2.4823,$$
$$\det(T([v_4^3, v_3^3, v_2^3, v_1^3])) = 7.7467, \quad \det(T([v_4^4, v_3^4, v_2^4, v_1^4])) = 19.8312.$$

Furthermore, for $\lambda = 0.9$,

$$\det(T(\lambda, [v_4, v_3, v_2, v_1])) = 0.3796, \quad \det(T(\lambda^2, [v_4^2, v_3^2, v_2^2, v_1^2])) = 1.7872,$$
$$\det(T(\lambda^3, [v_4^3, v_3^3, v_2^3, v_1^3])) = 4.2853, \quad \det(T(\lambda^4, [v_4^4, v_3^4, v_2^4, v_1^4])) = 8.5037.$$

v is generated randomly 10000 times, it is shown that all $T([v_4^i, v_3^i, v_2^i, v_1^i])$ and $T(\lambda^i, [v_4^i, v_3^i, v_2^i, v_1^i])$, $i = 1, \ldots, 4$, are nonsingular.

12.3 Estimates of ζ with Individual Thresholds

Based on strongly scaled full-rank signals, we now derive the estimation algorithms for ζ and analyze their convergence. To this end, estimation algorithms based on the information of individual thresholds are first investigated.

(A12.1) The noise $\{d_k\}$ is a sequence of i.i.d. random variables whose distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are twice continuously differentiable and known.

(A12.2) The prior information on $\theta = [a_0, \ldots, a_{n_0-1}]'$ and $\eta = [b_0, \ldots, b_{q_0-1}]'$ is that $\sum_{i=0}^{n_0-1} a_i \neq 0$, $b_{q_0} = 1$, $\eta \neq 0$, $\theta \in \Omega_\theta$, and $\eta \in \Omega_\eta$, where Ω_θ and Ω_η are known compact sets.

The input is a scaled $2n_0(q_0 + 1)$-periodic signal with one-period values

$$(v, v, \rho_1 v, \rho_1 v, \ldots, \rho_{q_0} v, \rho_{q_0} v),$$

where $v = (v_1, \ldots, v_{n_0})$ is strongly q_0 full rank.

By periodicity, $\Phi_l^j = \Phi^j$ for $j = 0, \ldots, n_0$, and Φ^j can be decomposed into $2(q_0+1)$ submatrices $\Phi^j(i)$, $i = 1, \ldots, 2(q_0+1)$, of dimension $n_0 \times n_0$: $\Phi^j = [(\Phi^j(1))', (\Phi^j(2))', \ldots, (\Phi^j(2(q_0 + 1)))']'$. Actually, for $k = 1, \ldots, 2(q_0 + 1)$,

$$\Phi^j(k) = \left[\ \phi^j_{kn_0}, \phi^j_{kn_0+1}, \ldots, \phi^j_{kn_0+n_0-1} \ \right]'.$$

Denote the $n_0 \times n_0$ circulant matrices

$$V^0 = T([1, \ldots, 1]), \quad \text{and} \quad V^j = T([v^j_{n_0}, \ldots, v^j_1]), \quad j = 1, \ldots, q_0.$$

Then, for $j = 0, \ldots, q_0$, the odd-indexed block matrices satisfy the simple scaling relationship

$$\Phi^j(1) = V^j, \quad \Phi^j(3) = \rho_1^j V^j, \quad \ldots, \quad \Phi^j(2q_0 + 1) = \rho_{q_0}^j V^j, \tag{12.6}$$

and the even-indexed block matrices are

$$\Phi^j(2l) = \rho_{l-1}^j T((\rho_l/\rho_{l-1})^j, [v_{n_0}, v_{n_0-1}, \ldots, v_1]), \quad l = 1, \ldots, q_0 + 1,$$

where $\rho_0 = \rho_{q_0+1} = 1$. Denote

$$\tau^{\{j\}} = [\tau^{\{j,1\}}, \ldots, \tau^{\{j,n_0\}}]' = V^j\theta,\ j = 0, \ldots, q_0. \tag{12.7}$$

Then, we have

$$\Phi^j(1)\theta = \tau^{\{j\}},\quad \Phi^j(3)\theta = \rho_1^j\tau^{\{j\}},\quad \ldots,\quad \Phi^j(2q_0+1)\theta = \rho_{q_0}^j\tau^{\{j\}}. \tag{12.8}$$

Let

$$\Psi_\theta = [\Phi^0\theta, \Phi^1\theta, \ldots, \Phi^{q_0}\theta].$$

Then, from (12.4), we have

$$Y_l = \Psi_\theta[\eta', 1]' + D_l = \zeta + D_l. \tag{12.9}$$

Remark 12.10. In $(v, v, \rho_1 v, \rho_1 v, \ldots, \rho_{q_0} v, \rho_{q_0} v)$, there are always two identical subsequences $\rho_i v, i = 1, \ldots, q_0$, appearing consecutively. The main reason for this input structure is to generate block matrices that satisfy the above scaling relationship (12.6).

For (12.9) and $i = 1, \ldots, m_0$, let

$$\begin{aligned}\mu_N^{\{i\}} &= [\mu_N^{\{i,1\}}, \ldots, \mu_N^{\{i,2n_0(q_0+1)\}}]' \\ &= \frac{1}{N}\sum_{k=1}^N S_k^{\{i\}} = \frac{1}{N}\sum_{k=1}^N I\{D_k \le C_i\mathbb{1}_{2n_0(q_0+1)} - \Psi_\theta[\eta', 1]'\},\end{aligned}$$

which is the empirical distribution of D_l at

$$C_i\mathbb{1}_{2n_0(q_0+1)} - \zeta = C_i\mathbb{1}_{2n_0(q_0+1)} - \Psi_\theta[\eta', 1]'.$$

Then, by the strong law of large numbers,

$$\mu_N^{\{i\}} \to p^{\{i\}} = F(C_i\mathbb{1}_{2n_0(q_0+1)} - \Psi_\theta[\eta', 1]') \text{ w.p.1.}$$

Denote $S_N^{\{i\}} = [S_N^{\{i,1\}}, \ldots, S_N^{\{i,2n_0(q_0+1)\}}]'$, where $S_N^{\{i\}}$ is as defined in (12.2) and $S_N^{\{ij\}}$ denotes its jth component. By Assumption (A12.1), for each $i = 1, \ldots, m_0$, $\{S_k^{\{i\}}\}$ is an i.i.d. sequence. Since $j = 1, \ldots, 2n_0(q_0+1)$, $ES_k^{\{i,j\}} = p^{\{i,j\}} = F(C_i - \zeta_j)$ and

$$E(S_k^{\{i,j\}} - p^{\{i,j\}})^2 = p^{\{i,j\}}(1 - p^{\{i,j\}}) := \Delta_{i,j}^2.$$

Define $z_N^{\{ij\}} = \sum_{k=1}^N S_k^{\{ij\}}/N$. Then,

$$Ez_N^{\{i,j\}} = \frac{1}{N}\sum_{k=1}^N ES_k^{\{i,j\}} = p^{\{i,j\}},$$

$$E(\mu_N^{\{i,j\}} - p^{\{i,j\}})^2 = \frac{\Delta_{i,j}^2}{N}. \tag{12.10}$$

Note that F is a monotone function by Assumption (A12.1), and Ω_θ and Ω_η are bounded by Assumption (A12.2). Then, there exists $z > 0$ such that

$$z \le p^{\{i,j\}} = F(C_i - \zeta_j) \le 1 - z, \quad i = 1, \ldots, m_0, \quad j = 1, \ldots, 2n_0(q_0 + 1).$$

Since $F(\cdot)$ is not invertible at 0 and 1, we modify $\mu_N^{\{i,j\}}$ to avoid this "singularity." Let

$$\xi_N^{\{i,j\}} = \begin{cases} \mu_N^{\{i,j\}}, & \text{if} \quad z \le \mu_N^{\{i,j\}} \le 1 - z, \\ z, & \text{if} \quad \mu_N^{\{i,j\}} < z, \\ 1 - z, & \text{if} \quad \mu_N^{\{i,j\}} > 1 - z. \end{cases} \tag{12.11}$$

Since $\mu_N^{\{i,j\}} \to p^{\{i,j\}}$, w.p.1 and $z < p^{\{i,j\}} < 1-z$, we have $\xi_N^{\{i,j\}} \to p^{\{i,j\}}$, w.p.1. Denote

$$\xi_N^{\{i\}} = [\xi_N^{\{i,1\}}, \ldots, \xi_N^{\{i,2n_0(q_0+1)\}}]'. \tag{12.12}$$

By Assumption (A12.1), F has a continuous inverse. Hence, for each $i = 1, \ldots, m_0$,

$$\begin{aligned} \zeta_N^{\{i\}} &= [\zeta_N^{\{i,1\}}, \ldots, \zeta_N^{\{i,2n_0(q_0+1)\}}]' \\ &:= C_i \mathbb{1}_{2n_0(q_0+1)} - F^{-1}(\xi_N^{\{i\}}) \\ &\to C_i \mathbb{1}_{2n_0(q_0+1)} - F^{-1}(p_i) = \Psi_\theta[\eta', 1]' \ \text{ as } \ N \to \infty \\ &= \zeta = [\zeta_1, \ldots, \zeta_{2n_0(q_0+1)}]' \ \text{ w.p.1.} \end{aligned} \tag{12.13}$$

12.4 Quasi-Convex Combination Estimators of ζ

Since $\zeta_N^{\{i\}}$ is constructed from each individual threshold C_i, this enables us to treat the coefficients of the quasi-convex combination as design variables such that the resulting estimate has the minimal variance. This resulting estimate is exactly the optimal QCCE in Chapter 6.

For $j = 1, \ldots, 2n_0(q_0 + 1)$, define $\zeta_N(j) = [\zeta_N^{\{1,j\}}, \ldots, \zeta_N^{\{m_0,j\}}]'$ and $c_N(j) = [c_N(j,1), \ldots, c_N(j,m_0)]'$ with $c_N(j,1) + \cdots + c_N(j,m_0) = 1$. Construct an estimate of ζ_j by defining

$$\widehat{\zeta}_N(j) = c_N'(j)\zeta_N(j) = \sum_{k=1}^{m_0} c_N(j,k)\zeta_N^{\{k,j\}}.$$

Denote $c(j) = [c(j,1), \ldots, c(j,m_0)]'$ such that $c_N(j) \to c(j)$. Then $c(j,1) + \cdots + c(j,m_0) = 1$, and by (12.13),

$$\widehat{\zeta}_N(j) = \sum_{k=1}^{m_0} c_N(j,k)\zeta_N^{\{k,j\}} \to \zeta_j \sum_{k=1}^{m_0} c(j,k) = \zeta_j.$$

Denote the estimation errors

$$e_N(j) = \widehat{\zeta}_N(j) - \zeta_j,$$
$$\varepsilon_N(j) = \zeta_N(j) - \zeta_j \mathbb{1}_{m_0},$$

and their covariances

$$\sigma_N^2(j) = Ee_N(j)e_N'(j), \quad Q_N(j) = E\varepsilon_N(j)\varepsilon_N'(j),$$

respectively. Then the covariance of estimation error is

$$\begin{aligned}\sigma_N^2(j) &:= E\left(\widehat{\zeta}_N(j) - \zeta_j\right)^2 = E\left(\sum_{k=1}^{m_0} c_N(j,k)(\zeta_N^{\{k,j\}} - \zeta_j)\right)^2 \\ &= c_N'(j)E\varepsilon_N(j)\varepsilon_N'(j)c_N(j) = c_N'(j)Q_N(j)c_N(j).\end{aligned}$$

That is, the variance is a quadratic form with respect to the variable $c(j)$. To obtain the quasi-convex combination estimate, we choose $c(j)$ to

$$\text{minimize } \sigma_N^2(j), \text{ subject to the constraint } c_N'(j)\mathbb{1}_{m_0} = 1.$$

Theorem 12.11. *Under Assumptions* (A12.1) *and* (A12.2)*, suppose* $u \in \mathcal{U}_{q_0}$ *and* $R_N(j) = NQ_N(j) = NE\varepsilon_N(j)\varepsilon_N'(j)$ *for* $j = 1, \ldots, 2n_0(q_0+1)$ *is positive definite. Then, the quasi-convex combination estimate can be obtained by choosing*

$$c_N^*(j) = \frac{R_N^{-1}(j)\mathbb{1}_{m_0}}{\mathbb{1}_{m_0}' R_N^{-1}(j)\mathbb{1}_{m_0}}, \quad \widehat{\zeta}_N(j) = \sum_{i=1}^{m_0} c^*(j,i)\zeta_N^{\{i,j\}}, \tag{12.14}$$

and the minimal variance satisfies

$$N\sigma_N^{2*}(j) = \frac{1}{\mathbb{1}_{m_0}' R_N^{-1}(j)\mathbb{1}_{m_0}}. \tag{12.15}$$

Consistency and Efficiency

From (12.14), $\widehat{\zeta}(j)$ can be regarded as an estimate of ζ_j. In this subsection, consistency and efficiency properties of this estimate will be analyzed.

By Assumption (A12.1), $G(x) = F^{-1}(x)$ is continuous on $(0,1)$. As a result, $G(x)$ is bounded on the compact set $[z, 1-z]$. Since $\zeta_N^{\{i,j\}} = C_i - G(\xi_N^{\{i,j\}}) \to \zeta^{\{i,j\}}$ w.p.1, we have $\zeta_N^{\{i,j\}} \to \zeta^{\{i,j\}}$ in probability. Furthermore, by the Lebesgue dominated convergence theorem [19, p. 100], $E\zeta_N^{\{i,j\}} \to \zeta^{\{i,j\}}$. Hence,

$$E\widehat{\zeta}_N(j) = E\sum_{k=1}^{m_0} c_N(j,k)\zeta_N^{\{k,j\}} \to \zeta_j \quad \text{as} \quad N \to \infty,$$

which means the estimate of ζ_j is asymptotically unbiased.

Subsequently, the efficiency of the estimate will be studied. To this end, the properties of $\xi_N^{\{i,j\}}$ in (12.11) will be introduced first.

Lemma 12.12. *Suppose* $u \in \mathcal{U}(n_0, q_0)$, *where* $\mathcal{U}(n_0, q_0)$ *is defined in Definition* 12.5. *Under Assumptions* (A12.1) *and* (A12.2), *there exist* $K_{i,j} \in (0, \infty)$ *and* $L_{i,j} \in (0, \infty)$, $i = 1, \ldots, m_0$, $j = 1, \ldots, 2n_0(q_0 + 1)$, *such that*

$$P\{\xi_N^{\{i,j\}} \neq \mu_N^{\{i,j\}}\} \leq K_{i,j} e^{-L_{i,j} N}. \tag{12.16}$$

Proof. Denote $X^{\{i,j\}} = (S_1^{\{i,j\}} - p^{\{i,j\}})/\Delta_{i,j}$. Note that $EX^{\{i,j\}} = 0$ and $E(X^{\{i,j\}})^2 = 1$. By the i.i.d. assumption, taking a Taylor expansion of $M_N^{\{i,j\}}(h) = [E \exp(hX^{\{i,j\}}/\sqrt{N})]^N$, the moment generating function of $\sqrt{N}(\mu_N^{\{i,j\}} - p^{\{i,j\}})/\Delta_{i,j}$, we obtain

$$\begin{aligned} M_N^{\{i,j\}}(h) &= \left[E[1 + \frac{hX^{\{i,j\}}}{\sqrt{N}} + \frac{h^2 (X^{\{i,j\}})^2}{2N} + O(N^{-3/2})]\right]^N \\ &= \left[1 + \frac{h^2}{2N} + O(N^{-(3/2)})\right]^N. \end{aligned}$$

Consequently, for any $t \in \mathbb{R}$,

$$\inf_h e^{-ht} M_N^{\{i,j\}}(h) = \inf_h e^{-ht} \left[1 + \frac{h^2}{2N} + O(N^{-(3/2)})\right]^N \leq K e^{-\frac{t^2}{2}}, \tag{12.17}$$

where $K > 0$ is a positive constant.

By means of the Chernoff bound [83, p. 326], for any $t \in (-\infty, p^{\{i,j\}}]$,

$$\begin{aligned} P\left\{\mu_N^{\{i,j\}} \leq t\right\} &= P\left\{\sum_{k=1}^N (S_k^{\{i,j\}} - p_{i,j}) \leq N \frac{(t - p_{i,j})}{\Delta_{i,j}}\right\} \\ &\leq \left\{\inf_h \left[e^{-\frac{h(t-p_{i,j})}{\Delta_{i,j}}} M_N^{\{i,j\}}(h)\right]\right\}^N \end{aligned} \tag{12.18}$$

and for any $p_{i,j} \leq t < \infty$,

$$P\{\mu_N^{\{i,j\}} \geq t\} \leq \left\{\inf_h \left[e^{-\frac{h(t-p_{i,j})}{\Delta_{i,j}}} M_N^{\{i,j\}}(h)\right]\right\}^N. \tag{12.19}$$

Considering

$$P\{\xi_{i,j}(N) \neq \mu_N^{\{i,j\}}\} = P(\mu_N^{\{i,j\}} \leq z) + P(\mu_N^{\{i,j\}} \geq 1 - z)$$

and (12.17)–(12.19), (12.16) is true. □

Theorem 12.13. *Under the conditions of Lemma* 12.12, *we have*

$$NE(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2 \to \Delta_{i,j}^2 \quad as \quad N \to \infty, \tag{12.20}$$

and

$$NE|(\xi_N^{\{i,j\}} - p^{\{i,j\}})|^{q_0} \to 0 \quad as \quad N \to \infty, \; q_0 = 3, 4, \ldots \tag{12.21}$$

Proof. (i) By Theorem 12.12, there exist $K_{i,j} \in (0,\infty)$ and $L_{i,j} \in (0,\infty)$ such that

$$\begin{aligned} EN(\xi_N^{\{i,j\}} - \mu_N^{\{i,j\}})^2 &\le NzP\{\xi_N^{\{i,j\}} \neq \mu_N^{\{i,j\}}\} \\ &\le zK_{i,j}Ne^{-L_{i,j}N} \to 0. \end{aligned}$$

This together with

$$\begin{aligned} &EN(\mu_N^{\{i,j\}} - p^{\{i,j\}})(\xi_N^{\{i,j\}} - \mu_N^{\{i,j\}}) \\ &\quad \le \sqrt{EN(\mu_N^{\{i,j\}} - p^{\{i,j\}})^2 EN(\xi_N^{\{i,j\}} - \mu_N^{\{i,j\}})^2} \\ &\quad = \Delta_{i,j}\sqrt{EN(\xi_N^{\{i,j\}} - \mu_N^{\{i,j\}})^2} \end{aligned}$$

implies that

$$\begin{aligned} &EN(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2 - EN(\mu_N^{\{i,j\}} - p^{\{i,j\}})^2 \\ &\quad = 2EN(\mu_N^{\{i,j\}} - p^{\{i,j\}})(\xi_N^{\{i,j\}} - \mu_N^{\{i,j\}}) \\ &\qquad + EN(\xi_N^{\{i,j\}} - \mu_N^{\{i,j\}})^2 \\ &\quad \to 0 \quad \text{as} \quad N \to \infty. \end{aligned} \tag{12.22}$$

Thus, by (12.10), we obtain (12.20).

(ii) Similarly, for $q_0 = 3, 4, \ldots$, one obtains

$$NE|(\xi_N^{\{i,j\}} - p^{\{i,j\}})|^{q_0} - NE|(\mu_N^{\{i,j\}} - p^{\{i,j\}})|^{q_0} \to 0.$$

By Hölder's inequality,

$$NE|\mu_N^{\{i,j\}} - p^{\{i,j\}}|^{q_0} \le \Delta_{i,j}\sqrt{NE(\mu_N^{\{i,j\}} - p^{\{i,j\}})^{2(q_0-1)}}. \tag{12.23}$$

Notice that for each i, j, $S_k^{\{i,j\}}$ is i.i.d. Then, we have

$$\begin{aligned} NE(\mu_N^{\{i,j\}} - p^{\{i,j\}})^{2(q_0-1)} &= NE\Big[\frac{1}{N}\sum_{k=1}^{N}(S_k^{\{i,j\}} - p^{\{i,j\}})\Big]^{2(q_0-1)} \\ &= N^{-2(m_0-2)}E(S_1^{\{i,j\}} - p^{\{i,j\}})^{2(q_0-1)} \\ &\le N^{-2(q_0-2)}, \end{aligned}$$

which together with (12.23) results in

$$NE|\mu_N^{\{i,j\}} - p^{\{i,j\}}|^{q_0} \le \Delta_{i,j}N^{-(q_0-2)} \to 0.$$

Hence, (12.21) is obtained. □

From (12.15), the covariance of the estimation $\widehat{\zeta}_N(j)$ is decided by $R_N(j)$.

Theorem 12.14. *Suppose $u \in \mathcal{U}(n_0, q_0)$. If, in addition to Assumptions* (A12.1) *and* (A12.2)*, the density function $f(x)$ is continuously differentiable, then as $N \to \infty$,*

$$R_N(j) := NQ_N(j) = NE\varepsilon_N(j)\varepsilon_N'(j) \to \Lambda(j)W(j)\Lambda(j) := R(j), \quad (12.24)$$

where

$$\varepsilon_N(j) = \zeta_N(j) - \zeta_j \mathbb{1}_{m_0},$$
$$\Lambda(j) = \text{diag}^{-1}\{f(C_1 - \zeta_j), \ldots, f(C_{m_0} - \zeta_j)\},$$

and

$$W(j) = \begin{bmatrix} p^{\{1,j\}}(1-p^{\{1,j\}}) & \ldots & p^{\{1,j\}}(1-p^{\{m_0,j\}}) \\ \vdots & \ddots & \vdots \\ p^{\{1,j\}}(1-p^{\{m_0,j\}}) & \ldots & p^{\{m_0,j\}}(1-p^{\{m_0,j\}}) \end{bmatrix}. \quad (12.25)$$

Proof. Denote $\varepsilon_N(j,i)$ as the ith component of $\varepsilon_N(j)$, $\dot{G}(x) = dG(x)/dx$, and $\ddot{G}(x) = d\dot{G}(x)/dx$. Then

$$\begin{aligned} \dot{G}(x) &= \frac{dG(x)}{dx} = \frac{dG(x)}{dF(G(x))} = \frac{1}{f(G(x))}, \\ \ddot{G}(x) &= \frac{d\dot{G}(x)}{dx} = -\frac{1}{f^2(G(x))}\dot{f}(G(x))\dot{G}(x). \end{aligned}$$

Since $\dot{f}(x)$ is continuous, by Assumption (A12.1), both $\dot{G}(x)$ and $\ddot{G}(x)$ are continuous, and hence bounded in $[z, 1-z]$. Let

$$\beta^{\{i,j\}} = \sup_{x\in[z,1-z]}\{|\dot{G}(x)|\} \text{ and } \gamma^{\{i,j\}} = \sup_{x\in[z,1-z]}\{|\ddot{G}(x)|\}.$$

Then, there exists a number $\lambda_N^{\{i,j\}}$ between $p^{\{i,j\}}$ and $\xi_N^{\{i,j\}}$ such that

$$\begin{aligned} \varepsilon_N(j,i) &= \zeta_N^{\{i,j\}} - \zeta_j = G(\xi_N^{\{i,j\}}) - G(p^{\{i,j\}}) \\ &= \dot{G}(p^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}}) + \frac{1}{2}\ddot{G}(\lambda_N^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2. \end{aligned}$$

This implies that for $i, k = 1, \ldots, m_0$,

$$\begin{aligned} &NE\varepsilon_N(j,i)\varepsilon_N(j,k) \\ &= NE(\zeta_N^{\{i,j\}} - \zeta_j)(\zeta_N^{\{k,j\}} - \zeta_j) \\ &= N\dot{G}(p^{\{i,j\}})\dot{G}(p^{\{k,j\}})E(\xi_N^{\{i,j\}} - p^{\{i,j\}})(\xi_N^{\{k,j\}} - p^{\{k,j\}}) \\ &\quad + NE\dot{G}(p^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}})(\xi_N^{\{k,j\}} - p^{\{k,j\}})^2\ddot{G}(\lambda_N^{\{k,j\}}) \\ &\quad + NE\ddot{G}(\lambda_N^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2(\xi_N^{\{k,j\}} - p^{\{k,j\}})\dot{G}(p^{\{k,j\}}) \\ &\quad + NE\ddot{G}(\lambda_N^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2(\xi_N^{\{k,j\}} - p^{\{k,j\}})^2\dot{G}(\lambda_N^{\{k,j\}}). \end{aligned} \quad (12.26)$$

By Hölder's inequality and Theorem 12.13, we have

$$\begin{aligned}
&|NE\dot{G}(p^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}})(\xi_N^{\{k,j\}} - p^{\{k,j\}})^2\ddot{G}(\lambda_N^{\{k,j\}})| \\
&\quad \le \beta^{\{i,j\}}\gamma^{\{i,j\}}\sqrt{NE(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2 NE(\xi_N^{\{k,j\}} - p^{\{k,j\}})^4|} \\
&\quad \le \beta^{\{i,j\}}\gamma^{\{i,j\}}\Delta_{i,j}\sqrt{NE(\xi_N^{\{k,j\}} - p^{\{k,j\}})^4|} \to 0.
\end{aligned} \tag{12.27}$$

Similarly,

$$|NE\ddot{G}(\lambda_N^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2(\xi_N^{\{k,j\}} - p^{\{k,j\}})\dot{G}(p^{\{k,j\}})| \to 0, \tag{12.28}$$

$$|NE\ddot{G}(\lambda_N^{\{i,j\}})(\xi_N^{\{i,j\}} - p^{\{i,j\}})^2(\xi_N^{\{k,j\}} - p^{\{k,j\}})^2\dot{G}(\lambda_N^{\{k,j\}})| \to 0. \tag{12.29}$$

Thus, similarly to (12.22), we have

$$\begin{aligned}
N\Big[&E(\xi_N^{\{i,j\}} - p^{\{i,j\}})(\xi_N^{\{k,j\}} - p^{\{k,j\}}) \\
&- E(\mu_N^{\{i,j\}} - p^{\{i,j\}})(\mu_N^{\{k,j\}} - p^{\{k,j\}})\Big] \to 0.
\end{aligned} \tag{12.30}$$

Since d_k, $k = 1, 2, \ldots$, are i.i.d.,

$$\begin{aligned}
&NE(\mu_N^{\{i,j\}} - p^{\{i,j\}})(\mu_N^{\{k,j\}} - p^{\{k,j\}}) \\
&= \frac{1}{N}E\Big[\Big(\sum_{l_1=1}^{N} I\{d_{l_1} \le p^{\{i,j\}}\} - p^{\{i,j\}}\Big)\Big(\sum_{l_2=1}^{N} I\{d_{l_2} \le p^{\{k,j\}}\} - p^{\{k,j\}}\Big)\Big] \\
&= \frac{1}{N}E\sum_{l_1=1}^{N} I\{d_{l_1} \le p^{\{i,j\}}\}I\{d_{l_1} \le p^{\{k,j\}}\} - p^{\{i,j\}}p^{\{k,j\}} \\
&= p^{\{\min\{i,k\},j\}} - p^{\{i,j\}}p^{\{k,j\}}
\end{aligned} \tag{12.31}$$

and

$$\dot{G}(p^{\{i,j\}}) = \frac{1}{f(G(p^{\{i,j\}}))} = \frac{1}{f(C_i - \zeta_j)}. \tag{12.32}$$

Therefore, (12.24) follows from (12.26)–(12.32). □

Proposition 12.15. *$R(j)$, $j = 1, \ldots, 2n_0(q_0 + 1)$, defined by (12.24), is positive definite, and*

$$\mathbb{1}'_{m_0} R^{-1}(j)\,\mathbb{1}_{m_0} = \sum_{k=1}^{m_0+1} \frac{h^2(j,k)}{\widetilde{p}^{\{k,j\}}}, \tag{12.33}$$

where

$$\begin{aligned}
\widetilde{p}^{\{i,j\}} &= F(C_i - \zeta_j) - F(C_{i-1} - \zeta_j), \\
h(j,i) &= f(C_{i-1} - \zeta_j) - f(C_i - \zeta_j),
\end{aligned}$$

with $C_0 = -\infty$ and $C_{l+1} = \infty$.

Proof. Since

$$R_N(j) = NE\varepsilon_N(j)\varepsilon'_N(j) \geq 0,$$

so is $R(j)$. Noting

$$\begin{aligned} R(j) &= \Lambda(j)W(j)\Lambda(j), \\ \Lambda(j) &= \mathrm{diag}^{-1}\{f(C_1 - \zeta_j), \ldots, f(C_{m_0} - \zeta_j)\}, \end{aligned}$$

and $f(C_i - \zeta_j) > 0$, $i = 1, \ldots, m_0$, we need only to show that $W(j)$ is positive definite.

From (12.25),

$$\begin{aligned} \det(W(j)) &= \begin{vmatrix} p^{\{1,j\}}(1-p^{\{1,j\}}) & \ldots & p^{\{1,j\}}(1-p^{\{m_0,j\}}) \\ \vdots & \ddots & \vdots \\ p^{\{1,j\}}(1-p^{\{m_0,j\}}) & \ldots & p^{\{m_0,j\}}(1-p^{\{m_0,j\}}) \end{vmatrix} \\ &= p^{\{1,j\}} \begin{vmatrix} 1-p^{\{1,j\}} & p^{\{1,j\}}-p^{\{2,j\}} & \ldots & p^{\{1,j\}}-p^{\{m_0,j\}} \\ 1-p^{\{2,j\}} & 0 & & p^{\{2,j\}}-p^{\{m_0,j\}} \\ \vdots & & \ddots & \vdots \\ 1-p^{\{m_0,j\}} & 0 & \cdots & 0 \end{vmatrix} \\ &= p^{\{1,j\}}(p^{\{1,j\}}-p^{\{2,j\}})\ldots(p^{\{m_0,j\}}-p^{\{j\}}_{m_0-1})(1-p^{\{m_0,j\}}) \neq 0. \end{aligned}$$

Thus, $R(j) > 0$. Furthermore, by Lemma 6.4,

$$\mathbb{1}'R^{-1}(j)\mathbb{1} = \sum_{k=1}^{m_0+1} \frac{h^2(j,k)}{\widetilde{p}^{\{k,j\}}}.$$

Thus, (12.33) is also true. □

Lemma 12.16. *The Cramér–Rao lower bound for estimating ζ_j based on $\{s_k\}$ is*

$$\sigma^2_{\mathrm{CR}}(N,j) = \left(N \sum_{j=1}^{m_0+1} \frac{h^2(j,i)}{\widetilde{p}^{\{i,j\}}} \right)^{-1}.$$

Next, we demonstrate that the aforementioned algorithms are asymptotically efficient based on the following theorem.

Theorem 12.17. *Under the conditions of Theorem* 12.14, *for $j = 1, \ldots, 2n_0(q_0+1)$,*

$$\lim_{N\to\infty} N\left(\sigma^{2*}_N(j) - \sigma^2_{\mathrm{CR}}(N,j)\right) = 0 \quad as \quad N \to \infty.$$

Proof. This theorem can be proved directly by Theorem 12.14, Proposition 12.15, and Lemma 12.16.

Recursive Quasi-Convex Combination Estimates

Since $\sigma_N^2(j) = E\varepsilon_N(j)\varepsilon_j'(N)$ contains an unknown parameter ζ_j, it cannot be directly computed. As a result, the quasi-convex combination estimate $\zeta_N(j)$ in (12.14) cannot be computed. In this section, we will derive computable estimates. The basic idea is to employ a recursive structure in which the unknown ζ_j is replaced by the current estimate $\widehat{\zeta}_N(j)$. Convergence of the algorithms will be established.

For $i = 1, \ldots, m_0$ and $j = 1, \ldots, 2n_0(q_0 + 1)$, let $\xi_0(i) = 0_{2n_0(q_0+1)}$, $\widehat{c}_0(j) = 0_{q_0}$, $\widehat{R}_0(j) = 0_{q_0 \times q_0}$, and $\widehat{\zeta}_0(j) = 0_{2n_0(q_0+1)}$. Suppose that at step $N-1$ ($N \geq 1$), $\xi_{N-1}(i)$, $c_{N-1}(j)$, and $\widehat{R}_{N-1}(j)$ have been obtained. Then the estimation algorithms can be constructed as follows.

(i) Calculate the sample distribution values

$$\xi_N^{\{i\}} = \frac{1}{N} S_N^{\{i\}} + \frac{N-1}{N} \xi_{N-1}^{\{i\}}.$$

(ii) Calculate the data points

$$\zeta_N^{\{i\}} = F^{-1}(\xi_N^{\{i\}}).$$

Let

$$\zeta_N(j) = [\zeta_N^{\{1,j\}}, \ldots, \zeta_N^{\{q_0,j\}}]', \ j = 1, \ldots, 2n_0(q_0 + 1).$$

(iii) Calculate each covariance estimate $R_N(j)$.

Let

$$p_N^{\{i,j\}} = F(C_i - \zeta_{N-1}^{\{i,j\}}),$$
$$\widehat{\Lambda}_N(j) = \mathrm{diag}^{-1}\{f(p_N^{\{1,j\}}), \ldots, f(p_N^{\{1,m_0\}})\},$$
$$W_N(j) = \begin{bmatrix} p_N^{\{1,j\}}(1 - p_N^{\{1,j\}}) & \cdots & p_N^{\{1,j\}}(1 - p_N^{\{1,m_0\}}) \\ \vdots & \ddots & \vdots \\ p_N^{\{1,j\}}(1 - p_N^{\{1,m_0\}}) & \cdots & p_N^{\{1,m_0\}}(1 - p_N^{\{1,m_0\}}) \end{bmatrix}.$$

Calculate $R_N(j)$ by

$$\widehat{R}_N(j) = \widehat{\Lambda}_N(j) W_N(j) \widehat{\Lambda}_N(j).$$

(iv) If $\widehat{R}_N(j)$ is nonsingular, then let

$$\widehat{c}_N(j) = \frac{\widehat{R}_j^{-1}(N)\mathbb{1}}{\mathbb{1}'\widehat{R}_N^{-1}(j)\mathbb{1}},$$

and compute

$$\widehat{\zeta}_N^{\{j\}} = \widehat{c}_N'(j)([C_1, \ldots, C_{m_0}]' - \zeta_N(j)).$$

Otherwise, $\widehat{\zeta}_N^{\{j\}} = \widehat{\zeta}_{N-1}^{\{j\}}$.

(v) Let $\widehat{\zeta}_N = [\widehat{\zeta}_N^{\{1\}}, \ldots, \widehat{\zeta}_N^{\{2n_0(q_0+1)\}}]'$. Go to step 1.

This algorithm depends only on sample paths. At each step, it minimizes the estimation variance based on the most recent information on the unknown parameter. In addition, the following asymptotic properties hold.

Theorem 12.18. *Under the conditions of Theorem* 12.14*, for* $j = 1, \ldots, 2n_0(q_0 + 1)$*, the above recursive algorithms have the following properties:*

$$\lim_{N\to\infty} \widehat{\zeta}_N(j) = \zeta_j \quad w.p.1, \tag{12.34}$$

$$\lim_{N\to\infty} \widehat{R}_N(j) = R(j) \quad w.p.1, \tag{12.35}$$

$$\lim_{N\to\infty} NE(\widehat{\zeta}_N(j) - \zeta_j)^2 = \frac{1}{\mathbb{1}' R^{-1}(j)\mathbb{1}} \quad w.p.1. \tag{12.36}$$

Proof. Note that $\xi_N^{\{i\}} \to F(C_i \mathbb{1}_{2n_0(q_0+1)} - \zeta)$ w.p.1 and the convergence is uniform in $C_i \mathbb{1}_{2n_0(q_0+1)} - \zeta$. Since $F(\cdot)$ and $F^{-1}(\cdot)$ are both continuous,

$$\begin{aligned} \zeta_N^{\{i\}} &= C_i \mathbb{1}_{2n_0(q_0+1)} - F^{-1}(\xi_N^{\{i\}}) \\ &\to C_i \mathbb{1}_{2n_0(q_0+1)} - F^{-1}(F(C_i \mathbb{1}_{2n_0(q_0+1)} - \zeta) = \zeta \end{aligned}$$

w.p.1 as $N \to \infty$. Thus, the quasi-convex combination $\widehat{\zeta}_N(j)$ converges to ζ w.p.1. That is, (12.34) holds.

By Assumption (A12.1), $F(\cdot)$ and $f(\cdot)$ are both continuous. Hence,

$$\widehat{\Lambda}_N(j) \to \Lambda(j) \quad \text{and} \quad W_N(j) \to W_j.$$

As a result, (12.35) holds, and by (12.15),

$$E(\widehat{\zeta}_N(j) - \zeta_j)^2 = \frac{1}{\mathbb{1}'_{m_0} \widehat{R}_N^{-1}(j) \mathbb{1}_{m_0}} \to \frac{1}{\mathbb{1}'_{m_0} R^{-1}(j) \mathbb{1}_{m_0}},$$

which results in (12.36). □

12.5 Estimation of System Parameters

Identification algorithms of the system parameters will be constructed based on the estimate of ζ. The parameters of the linear part are first estimated, then the nonlinearity is identified.

Identifiability of the Unknown Parameters

Theorem 12.19. *Suppose $u \in \mathcal{U}(n_0, q_0)$. Then,*

$$\Psi_\theta[\eta', 1]' = \zeta$$

has a unique solution (θ^*, η^*).

Proof. (i) To obtain θ^*.

By the first component of (12.13), we have $\zeta = [\zeta_1, \ldots, \zeta_{2n_0(q_0+1)}]'$, and

$$b_0\tau^{\{0,1\}} + b_1\tau^{\{1,1\}} + \cdots + b_{q_0}\tau^{\{q_0,1\}} = \zeta_1.$$

From (12.8), the $2in_0 + 1$ $(i = 1, \ldots, q_0)$ component of (12.13) turns out to be

$$b_0\tau^{\{0,1\}} + \rho_i b_1\tau^{\{1,1\}} + \cdots + \rho_i^{q_0} b_{q_0}\tau^{\{q_0,1\}} = \zeta_{2in_0+1},$$

or equivalently,

$$\Re \begin{bmatrix} b_0\tau^{\{0,1\}} \\ b_1\tau^{\{1,1\}} \\ \vdots \\ b_{q_0}\tau^{\{q_0,1\}} \end{bmatrix} = \begin{bmatrix} \zeta_1 \\ \zeta_{2n_0+1} \\ \vdots \\ \zeta_{2q_0n_0+1} \end{bmatrix}, \quad \text{where} \quad \Re = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & \rho_1 & \ldots & \rho_1^{q_0} \\ \vdots & \cdots & \ldots & \vdots \\ 1 & \rho_{q_0} & \ldots & \rho_{q_0}^{q_0} \end{bmatrix}.$$

Since $\rho_j \neq 0$, $\rho_j \neq 1$, $j = 1, \ldots, q_0$, and $\rho_i \neq \rho_j$, the determinant of the Vandermonde matrix

$$\det \Re = \prod_{0 \leq i < j \leq q_0 - 1} (\rho_j - \rho_i) \neq 0 \quad \text{with} \quad \rho_0 = 1.$$

Hence, $b_j\tau^{\{j,1\}}$, $j = 0, \ldots, q_0$, can be solved by

$$\begin{bmatrix} b_0\tau^{\{0,1\}} \\ b_1\tau^{\{1,1\}} \\ \vdots \\ b_{q_0}\tau^{\{q_0,1\}} \end{bmatrix} = \Re^{-1} \begin{bmatrix} \zeta_1 \\ \zeta_{2n_0+1} \\ \vdots \\ \zeta_{2q_0n_0+1} \end{bmatrix}.$$

Similarly, we have

$$\Gamma = \Re^{-1}\Xi, \tag{12.37}$$

where

$$\Gamma = \begin{bmatrix} b_0\tau^{\{0,1\}} & b_0\tau^{\{0,2\}} & \dots & b_0\tau^{\{0,n_0\}} \\ b_1\tau^{\{1,1\}} & b_1\tau^{\{1,2\}} & \dots & b_1\tau^{\{1,n_0\}} \\ & \vdots & & \\ b_{q_0}\tau^{\{q_0,1\}} & b_{q_0}\tau^{\{q_0,2\}} & \dots & b_{q_0}\tau^{\{q_0,n_0\}} \end{bmatrix},$$

$$\Xi = \begin{bmatrix} \zeta_1 & \zeta_2 & \dots & \zeta_{n_0} \\ \zeta_{2n_0+1} & \zeta_{2n_0+2} & \dots & \zeta_{3n_0} \\ & \vdots & & \\ \zeta_{2q_0n_0+1} & \zeta_{2q_0n_0+2} & \dots & \zeta_{2q_0+1)n_0} \end{bmatrix}.$$

Denote $r(i)$ as the ith column of $(\Re^{-1})'$. Then, by $b_{q_0} = 1$, we have

$$\tau^{\{q_0\}} = [\tau^{\{q_0,1\}}, \dots, \tau^{\{q_0,n_0\}}]' = \Xi' r(q_0).$$

Note that $u \in \mathcal{U}(n_0, q_0)$ implies that V^{q_0} is full rank. Then, by (12.7), one can get $\theta^* = V_{q_0}^{-1}\tau^{\{q_0\}}$.

(ii) To obtain η^*.

By Assumption (A12.2), $\sum_{i=0}^{n_0-1} a_i \neq 0$, or $V^0\theta \neq \mathbf{0}_{n_0}$. For $u \in \mathcal{U}(n_0, q_0)$ and $j = 1, \dots, q_0$, $V^j = T([v_{n_0}^j, \dots, v_1^j])$ is full rank by Definition 12.1, and so $V^j\theta \neq \mathbf{0}_{n_0}$. Thus, for each $j = 0, \dots, q_0$, $\tau^{\{j\}} = V^j\theta$ has a nonzero component $\tau^{\{j,i_N^*(j)\}}$. For any given positive integer k and $j = 1, \dots, k$, let $\beta_j(k)$ be a k-dimensional vector with all components being zero except the jth being 1, that is,

$$\beta_j(k) = [\underbrace{0, \dots, 0}_{j-1}, 1, \underbrace{0, \dots, 0}_{k-j}]'.$$

Then, from (12.37), we have

$$b_j\tau^{\{j,i_N^*(j)\}} = \beta_j'(m_0+1)\Re^{-1}\Xi\beta_{i^*(j)}(n_0), \quad j = 0, \dots, q_0,$$

which gives b_j, $j = 0, \dots, q_0$, since $\tau^{\{j,i_N^*(j)\}}$ can be calculated from V^j and θ^* via (12.7). Thus, η^* is obtained. □

A particular choice of the scaling factors ρ_j is $\rho_j = \lambda^j$, $j = 0, 1, \dots, q_0$, for some $\lambda \neq 0$ and $\lambda \neq 1$. In this case, the period of input u can be shortened to $n_0(q_0+2)$ under a slightly different condition.

Identification Algorithms and Convergence Properties

The $\zeta_N = [\zeta_N^{\{1\}}, \dots, \zeta_N^{\{2n_0(q_0+1)-1\}}]'$ in (12.12) has $2n_0(q_0+1)$ components for a strongly scaled q_0 full-rank signal $u \in \mathcal{U}(n_0, q_0)$.

Let

$$V_{q_0} = T([v_{n_0}^{q_0}, \dots, v_1^{q_0}]), \qquad [r_1, \dots, r(q_0)] := (\Re')^{-1},$$

$$\Xi_N = \begin{bmatrix} \zeta_N^{\{1\}} & \zeta_N^{\{2\}} & \cdots & \zeta_N^{\{n_0\}} \\ \zeta_N^{\{2n_0+1\}} & \zeta_N^{\{2n_0+2\}} & \cdots & \zeta_N^{\{3n_0\}} \\ & \vdots & & \\ \zeta_N^{\{2q_0n_0+1\}} & \zeta_N^{\{2q_0n_0+2\}} & \cdots & \zeta_N^{\{(2q_0+1)n_0\}} \end{bmatrix}.$$

Then, we have the following identification algorithm:

(i) *Estimate* θ. The estimate of θ is taken as

$$\theta_N = V_{q_0}^{-1} \Xi_N' r(q_0). \tag{12.38}$$

(ii) *Estimate* η. Let $b_0(j) = 0$ and

$$b_N(j) = \begin{cases} [\zeta_N^{\{i_N^*(j)\}}, \dots, \zeta_N^{\{2q_0n_0+i_N^*(j)\}}] r_N(i_N^*(j)) / \tau_N^{\{j,i_N^*(j)\}}, \\ \qquad\qquad \text{if } \tau^{\{j,i_N^*(j)\}} \neq 0, \\ b_{N-1}(j), \qquad\qquad \text{if } \tau^{\{j,i_N^*(j)\}} = 0, \end{cases}$$

where

$$i_N^*(j) = \min\{\text{argmax}_{1 \le i \le n_0} |\tau^{\{j,i\}}|\}, \quad j = 0, 1, \dots, q_0 - 1; \tag{12.39}$$

$r(i_N^*(j))$ is the $i_N^*(j)$th column of $(\Re')^{-1}$, and $\tau^{\{j,i_N^*(j)\}}$ is the $i_N^*(j)$-th component of $\tau_N^{\{j\}} = V^j \theta_N$. Then, the estimate of η is taken as

$$\eta_N = [b_N(0), \dots, b_N(q_0 - 1)]'. \tag{12.40}$$

Theorem 12.20. *Suppose* $u \in \mathcal{U}(n_0, q_0)$. *Then, under Assumptions* (A12.1) *and* (A12.2),

$$\theta_N \to \theta \quad \textit{and} \quad \eta_N \to \eta \quad \textit{w.p.1 as} \quad N \to \infty.$$

Proof. By (12.13), $\zeta_N \to \zeta$ w.p.1. as $N \to \infty$. So,

$$\theta_N = V_{q_0}^{-1} \Xi_N' r_{q_0} \to V_{q_0}^{-1} \Xi' r(q_0) = \theta,$$

which in turn leads to

$$\tau_N^{\{j\}} = [\tau_N^{\{j,1\}}, \dots, \tau_N^{\{j\}}(n_0)]' := V^j \theta_N \to V^j \theta = \tau^{\{j\}} \quad \text{w.p.1.}$$

and $\tau^{\{j,i\}} \to \tau^j(i)$ w.p.1 for $i = 1, \dots, n_0$. Thus, for $j = 0, \dots, q_0 - 1$, we have

$$i_N^*(j) = \min\{\text{argmax}_{1 \le i \le n_0} |\tau^{\{j,i\}}|\} \to \min\{\text{argmax}_{1 \le i \le n_0} |\tau^j(i)|\} := i^*(j),$$

and

$$\tau_N^{\{j,i_N^*(j)\}} \to \tau^{\{j,i_N^*(j)\}} \neq 0.$$

This means that with probability 1, there exists $N_0 > 0$ such that

$$\tau_N^{\{j,i_N^*(j)\}} \neq 0, \ \forall N \geq N_0.$$

Let

$$b_N(j) = \frac{r(i_N^*(j))}{\tau_N^{\{j,i_N^*(j)\}}}[\zeta_N(i_N^*(j)), \zeta_N(2n_0 + i_N^*(j)), \ldots, \zeta_N(2q_0 n_0 + i_N^*(j)].$$

Then by

$$\begin{aligned} b_N(j)\tau_N^{\{j,i_N^*(j)\}} &\to [\zeta_{i^*(j)}, \zeta_{2n_0+i^*(j)}, \ldots, \zeta_{2q_0 n_0+i^*(j)}]r(i^*(j)) \\ &= b_j \tau^{\{j,i_N^*(j)\}}, \end{aligned}$$

we have $b_N(j) \to b_j$ for $j = 0, \ldots, q_0 - 1$. Hence, $\eta_N \to \eta$ w.p.1 as $N \to \infty$. □

Algorithms under Exponentially Scaled Inputs

Let u be $n_0(q_0+2)$-periodic with one-period values $(v, \lambda v, \ldots, \lambda^{q_0} v, \lambda^{q_0+1} v)$. The $\overline{\zeta}_N = [\overline{\zeta}_N^{\{1\}}, \ldots, \overline{\zeta}_N^{\{q_0+1\}}]'$ can be estimated by the algorithms in Section 12.4 with dimension changed from $2n_0(q_0 + 1)$ to $n_0(q_0 + 2)$, and

$$\overline{\zeta}_N \to \overline{\zeta} = \sum_{j=0}^{q_0} b_j \overline{\Phi}^j \theta.$$

Partition $\overline{\Phi}^j$ into (q_0+2) submatrices $\overline{\Phi}^j(i)$, $i = 1, \ldots, q_0+2$, of dimension $n_0 \times n_0$:

$$\overline{\Phi}^j = [(\overline{\Phi}^j(1))', (\overline{\Phi}^j(2))', \ldots, (\overline{\Phi}^j(q_0 + 2))']'.$$

If $u \in \mathcal{U}_\lambda(n_0, q_0)$, then it can be directly verified that

$$\begin{aligned} &\overline{\Phi}^j(l+1) = \lambda^{jl} \overline{\tau}^{\{j\}}, \quad l = 0, 1, \ldots, q_0, \\ &\overline{\Phi}^j(q_0 + 2) = \lambda^{j(q_0+2)} T(\lambda^{-j(q_0+2)}, [v_{n_0}, \ldots, v_1]), \end{aligned}$$

where $\overline{\tau}^{\{j\}} = T(\lambda^j, [v_{n_0}^j, \ldots, v_1^j])$. With these notations, we have the following result, whose proof is similar to that of Theorem 12.19, and hence, is omitted.

Theorem 12.21. *Suppose $u \in \mathcal{U}_\lambda(n_0, q_0)$. Then, under Assumptions* (A12.1) *and* (A12.2)*,*

$$\overline{\Psi}_\theta [\eta', 1]' = \overline{\zeta}$$

has a unique solution (θ^*, η^*), where

$$\overline{\Phi}_\theta = [\overline{\Phi}(0)\theta, \overline{\Phi}(1)\theta, \ldots, \overline{\Phi}(q_0)\theta].$$

Let

$$\overline{\zeta}_N = [\overline{\zeta}_N^{\{1\}}, \ldots, \overline{\zeta}_N^{\{n_0(q_0+1)\}}]',$$

and

$$\overline{V}^{q_0} = T(\lambda^{q_0}, [v_{n_0}^{q_0}, \ldots, v_1^{q_0}]), \qquad [r_1, \ldots, r(q_0)] := (\Re')^{-1},$$

$$\overline{\Xi}_N = \begin{bmatrix} \overline{\zeta}_N^{\{1\}} & \overline{\zeta}_N^{\{2\}} & \cdots & \overline{\zeta}_N^{\{n_0\}} \\ \overline{\zeta}_N^{\{n_0+1\}} & \overline{\zeta}_N^{\{n_0+2\}} & \cdots & \overline{\zeta}_N^{\{2n_0\}} \\ & & \vdots & \\ \overline{\zeta}_N^{\{q_0 n_0+1\}} & \overline{\zeta}_N^{\{q_0 n_0+2\}} & \cdots & \overline{\zeta}_N^{\{n_0(q_0+1)\}} \end{bmatrix}.$$

Then, we have the following identification algorithm:

(i) *Estimate* θ. The estimate of θ is taken as

$$\theta_N^e = (\overline{\Phi}(q_0))^{-1}(\overline{\Xi}_N)' r_N(q_0).$$

(ii) *Estimate* η. Let $b_0^e(j) = 0$ and

$$b_N^e(j) = \begin{cases} [\overline{\zeta}_N^{\{i_N^e(j)\}}, \overline{\zeta}_N^{2n_0+i_N^e(j)}, \ldots, \\ \overline{\zeta}_N^{\{2q_0 n_0+i_N^e(j)\}}] r_N(i_N^e(j))/\overline{\tau}_N^{\{j,i_N^e(j)\}}, & \text{if } \overline{\tau}_N^{\{j,i_N^e(j)\}} \neq 0, \\ b_{N-1}^e(j), & \text{if } \overline{\tau}_N^{\{j,i_N^e(j)\}} = 0, \end{cases}$$

where

$$i_N^e(j) = \min\{\text{argmax}_{1 \le i \le n_0} |\overline{\tau}_N^{\{j\}}(i)|\}, \quad j = 0, 1, \ldots, q_0 - 1,$$

$r(i_N^e(j))$ is the $i_N^e(j)$-th column of $(\Re')^{-1}$, and $\overline{\tau}_N^{\{j,i_N^e(j)\}}$ is the $i_N^e(j)$-th component of

$$\overline{\tau}_N(j) = \overline{\tau}^{\{j\}}\theta_N.$$

Then, the estimate of η is taken as

$$\eta_N^e = [b_N^e(0), \ldots, b_N^e(q_0 - 1)]'.$$

Theorem 12.22. *Suppose* $u \in \mathcal{U}_\lambda(n_0, q_0)$. *Then, under Assumptions* (A12.1) *and* (A12.2),

$$\theta_N^e \to \theta \quad \textit{and} \quad \eta_N^e \to \eta \quad \textit{w.p.1 as} \quad N \to \infty.$$

12.6 Examples

In this section, we illustrate the convergence of the estimates given by the algorithms described above. The noise is Gaussian with known mean and variance. In Example 12.23, the identification algorithm with quantized sensors is shown. Example 12.24 concerns the identification of systems with non-monotonic nonlinearities. Example 12.25 illustrates an algorithm based on the prior information, which is more simplified than the one described by (12.38)–(12.40). The parameter estimates are shown to be convergent in all cases.

Example 12.23. Consider a gain system $y_k = a + d_k$. Here the actual value of the unknown a is 5. The disturbance is a sequence of i.i.d. Gaussian variables with zero mean and standard deviation $\sigma = 5$. The sensor has three switching thresholds, $C_1 = 2$, $C_2 = 6$, and $C_3 = 10$. Then, the recursive algorithm in Section 12.4 is used to generate quasi-convex combination estimates. For comparison, estimates derived by using each threshold individually (i.e., binary-valued sensors) are also calculated. Figure 12.2 compares quasi-convex combination estimates to those using each threshold. It is shown that the estimate with three thresholds converges faster than the ones with each threshold individually. The weights of the estimates of each threshold are shown in Figure 12.3, which illustrates that the weights are not sure to be positive.

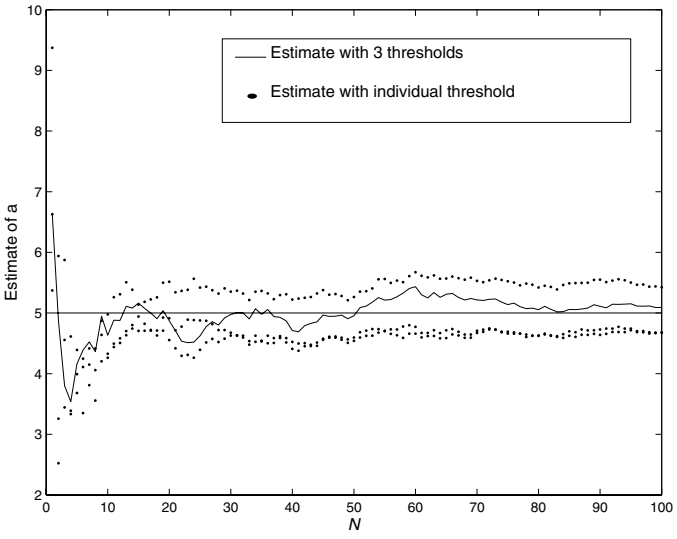


FIGURE 12.2. Identification with quantized output observations

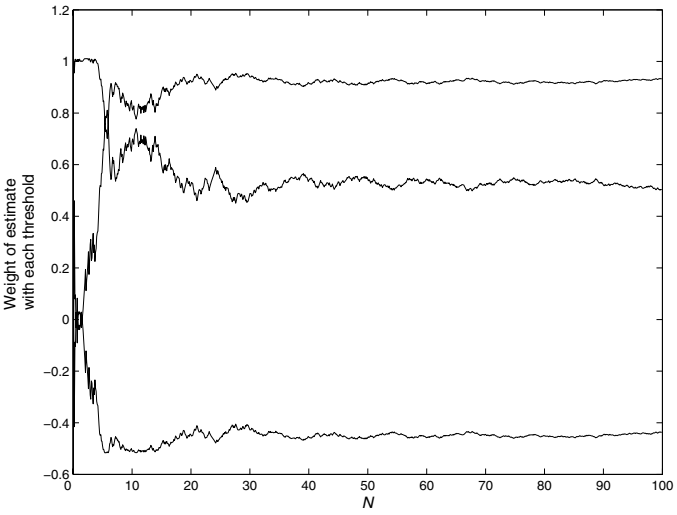


FIGURE 12.3. Weights of estimates with each threshold

Example 12.24. Consider

$$\begin{cases} y_k = a_0 x_k + a_1 x_{k-1} + d_k, \\ x_k = b_0 + b_1 u_k + b_2 u_k^2 + u_k^3, \end{cases}$$

where the noise $\{d_k\}$ is a sequence of i.i.d. Gaussian variables with $Ed_1 = 0$, $\sigma_d^2 = 1$. The output is measured by a binary-valued sensor with threshold $C = 13$. The linear subsystem has order $n_0 = 2$, and the nonlinear function has order $q_0 = 2$. The prior information on a_i, $i = 0, 1$, is that $a_i \in [0.5, 5]$. Suppose the true values of unknown parameters are $\theta = [a_0, a_1]' = [1.31, 0.85]'$ and $\eta = [b_0, b_1, b_2]' = [4, 1.4, -3]'$. The nonlinearity is not monotone, which is illustrated in Figure 12.4. It is shown that not all values of $v, \rho_1 v, \rho_2 v, \rho_3 v$ are situated in the same monotone interval of the nonlinearity.

The input is chosen to be $2n_0(q_0 + 1) = 12$-periodic with one period $(v, v, \rho_1 v, \rho_1 v,\ \rho_2 v, \rho_2 v)$, where $v = [1.2, 0.85]$, $\rho_1 = 0.5$, $\rho_2 = 1.65$, and $\rho_3 = 0.75$. Define the block variables X_l, Y_l, Φ_l^j, D_l, and S_l, in the case of a six-periodic input. Using (12.12), we can construct the algorithms (12.38)–(12.40) to identify θ and η.

The estimation errors of θ and η are illustrated in Figure 12.5, where the errors are measured by the Euclidean norm. Both parameter estimates of the linear and nonlinear subsystems converge to their true values, despite the nonlinearity being non-monotonic.

Example 12.25. For some prior information, algorithms (12.38)–(12.40) can be simplified. For example, the estimation algorithms of η can be simplified when the prior information on θ is known to be positive and the

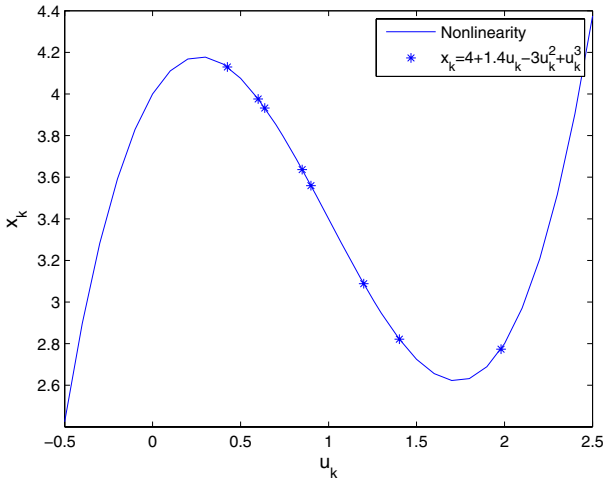


FIGURE 12.4. Nonmonotonic nonlinearity

periodic input u is positive. Both the mean and the variance of disturbance are not zero in this example.

Consider

$$\begin{cases} y_k = a_0 x_k + a_1 x_{k-1} + d_k, \\ x_k = b_0 + b_1 u_k + u_k^2, \end{cases}$$

where the noise $\{d_k\}$ is a sequence of i.i.d. Gaussian variables with $Ed_1 = 2$, $\sigma_d^2 = 4$. The output is measured by a binary-valued sensor with threshold $C = 13$. The linear subsystem has order $n_0 = 2$, and the nonlinear function has order $q_0 = 2$. The prior information on a_i, $i = 0, 1$, is that $a_i \in [0.5, 5]$. Suppose the true values of the unknown parameters are $\theta = [a_0, a_1]' = [1.17, 0.95]'$ and $\eta = [b_0, b_1]' = [3, 1.3]'$.

The input is 12-periodic with one period $(v, v, \rho_1 v, \rho_1 v, \rho_2 v, \rho_2 v)$, where $v = [1.2, 0.85]$, $\rho_1 = 0.65$, and $\rho_2 = 1.25$. Define the block variables X_l, Y_l, Φ_l^j, D_l and S_l, in the case of a 12-periodic input. Using (12.12), we can construct the algorithms (12.38)–(12.40) to identify θ.

Considering the prior information on θ, a more simplified algorithm can be constructed to identify η than the one given by (12.38)–(12.40). Note that $a_i \in [0.5, 5]$, $i = 1, 2$, and u is positive. Then, $\tau^{\{j,1\}}$, the first component of $V^j \theta$, is

$$\tau^{\{j,1\}} = a_0 v_2^2 + a_1 v_1^2 \geq 0.5(v_2^2 + v_1^2) \neq 0,$$

where the last inequality is derived from the fact that v is strongly 2 full rank. So, it is not necessary to calculate $i_N^*(j)$ in (12.39), which aims to

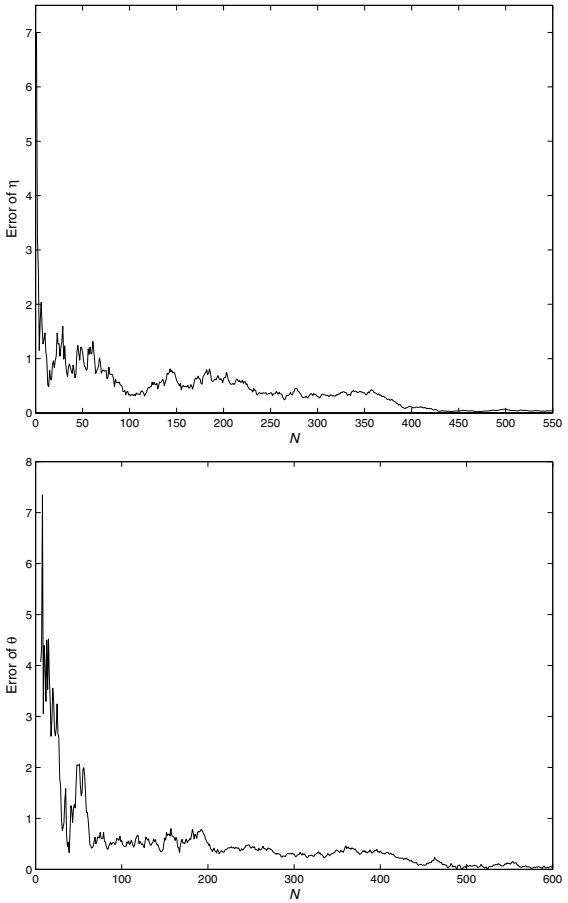


FIGURE 12.5. Identification errors of θ and η with nonmonotonic nonlinearity

find the nonzero component of $\tau^{\{j\}}$. And η can be estimated as follows:

$$\eta_0 = 0$$

$$\eta_N = \begin{cases} \Lambda_N \Re^c [\zeta_N^{\{1\}}, \zeta_N^{\{2n_0+1\}}, \ldots, \zeta_N^{\{2q_0 n_0+1\}}]', & \text{if } \prod_{j=0}^{q_0-1} \tau_N^{\{j,1\}} \neq 0, \\ \eta_{N-1}, & \text{if } \prod_{j=0}^{q_0-1} \tau_N^{\{j,1\}} = 0, \end{cases}$$

where $\Lambda_N = \text{diag}^{-1}(\tau_N^{\{0,1\}}, \ldots, \tau_N^{\{q_0-1,1\}})$, $\Re^c$ is a $q_0 \times (q_0+1)$ matrix containing the first to q_0-1th rows of $\Re^{-1}$, and $\tau_N^{\{j,1\}}$ is the first component of $\tau_N^{\{j\}} = V^j \theta_N$.

The estimation errors of θ and η are shown in Figure 12.6, where the

errors are measured by the Euclidean norm. Both parameter estimates of the linear and nonlinear subsystems converge to their true values.

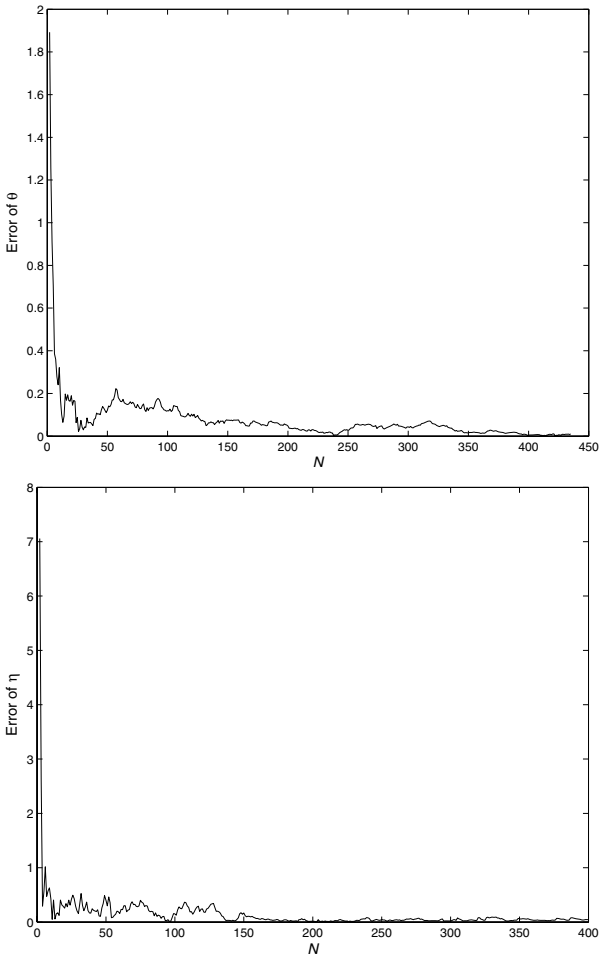


FIGURE 12.6. Identification errors of θ and η

12.7 Notes

In this chapter, the identification of Hammerstein systems with quantized output observations is studied. The development follows [128]. Hammerstein systems have been used to model practical systems with memoryless actuators and have been studied extensively in system identification, see for example [3, 45, 69, 71].

Structurally, a Hammerstein system with a quantized sensor may be viewed as Hammerstein–Wiener system which contains both input and output nonlinearity. However, our approaches are quite different from typical studies of such nonlinear system identification problems in which the output nonlinearities usually contain some sections of smooth functions. Unlike traditional approximate gradient methods or covariance analysis, we employ the methods of empirical measures and parameter mappings. Under assumptions of known noise distribution functions and strongly scaled full-rank inputs, identification algorithms, convergence properties, and the estimation efficiency are derived.

13

Systems with Markovian Parameters

This chapter concerns the identification of systems with time-varying parameters. The parameters are vector-valued and take values in a finite set. As in the previous chapters, only binary-valued observations are available.

Our study is motivated by applications in the areas of smart sensors, sensor networks, networked mobile agents, distributed power generation networks, etc. For instance, consider an array of mobile sensors being dispatched to survey an area for potential land contamination. Each sensor travels along a trajectory, measures a surface, and communicates the measured values via a wireless network to the command center. Some of the features include

(1) The parameter of interest takes only a few possible values representing regions such as "no contamination," "low contamination," and "high contamination."

(2) When the sensor travels, the parameter values switch randomly depending on the actual contamination.

(3) Due to communication limitations, only quantized measurements are available. Here when the sensor moves slowly, the parameter values switch infrequently. This problem may be described as a system with an unknown parameter that switches over finite possible values randomly. This application represents problems in ocean survey, detection of water pipe safety, mobile robots for bomb, chemical, biological threats, etc.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_13,

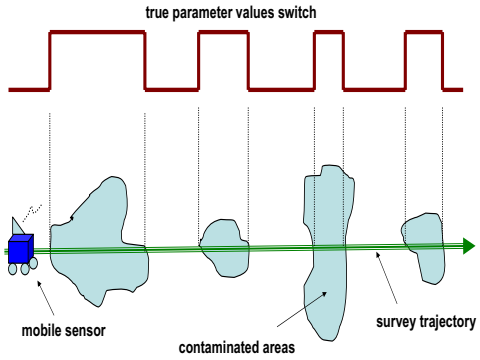


FIGURE 13.1. Mobile sensor systems for area survey

To capture the essence of such problems as those above, we formulate a class of identification problems with randomly switching parameters and binary-valued observations. We shall focus on time-varying parameters modeled by a discrete-time Markov chain with a finite state space. The limited information due to binary-valued sensors makes identification a difficult task. Our approach for identifying regime-switching systems with binary observations relies on the basic idea of Wonham-type filters. Based on the key ideas of such filters, we derive mean-square estimators and analyze their error bounds. To obtain the error bounds for mean-square estimators, we utilize asymptotic distributional results. We first establish weak convergence of functional central limit results, followed by strong approximation of the scaled sequences. Then these distributional results are used to obtain error bounds.

In applications, the frequency of the switching processes plays a crucial role. Consider two typical cases for tracking and identification. The first case is concerned with Markov chains whose switching movements occur infrequently. Here, the time-varying parameter takes a constant value for a relatively long time and switches to another value at a random time. The jumps happen relatively infrequently. We develop maximum posterior (MAP) estimators and obtain bounds on estimation or tracking errors based on Wonham filters. We also point out that a simplified estimator can be developed using empirical measures. The second class of systems aims at treating fast-switching systems. One motivation of such systems is the discretization of a fast-varying Markovian system in continuous time. Suppose the precise transition probabilities are unknown. When parameters frequently change their values, the system becomes intractable if one insists on tracking instantaneous changes. In fact, if the jump parameter switches too frequently, it would be impossible to identify the instantaneous jumps even with regular linear sensors, let alone binary observations. As a result, an alternative approach is suggested. Instead of tracking the moment-by-moment changes, we examine the averaged behavior of the sys-

tem. The rationale is as follows: Because the Markov chain varies at a fast pace, within a short period of time, it should settle down at a stationary or steady state. In the steady state, the underlying system is a weighted average with the weighting factors the components of the stationary distribution of the Markov chains.

Section 13.1 begins with the setup of the tracking and identification problem with a Markov parameter process. Section 13.2 presents Wonham-type filters for the identification problem. Section 13.3 concerns mean-square criteria. Section 13.4 proceeds with the study of infrequently switching systems. Section 13.5 takes up the issue of fast-switching systems.

13.1 Markov Switching Systems with Binary Observations

Consider a single-input–single-output (SISO), discrete-time system represented by

$$y_k = \phi_k' \theta_k + d_k, \tag{13.1}$$

where $\phi_k = (u_k, \ldots, u_{k-n_0+1})'$ and $\{d_k\}$ is a sequence of random disturbances. The θ_k is a Markov chain that takes m_0 possible vector values $\theta^{(j)} \in \mathbb{R}^{n_0}$, $j = 1. \ldots, m_0$. y_k is measured by a binary-valued sensor with the known threshold C. After applying an input u, the output s_k is measured for $k = 0, \ldots, N-1$ with observation length $N \geq n_0$. We will use the following assumptions throughout this chapter.

(A13.1) The time-varying process $\{\theta_k\}$ is a discrete-time Markov chain with a transition probability matrix P and a finite state space $\mathcal{M} = \{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m_0)}\}$.

(A13.2) The $\{d_k\}$ is a sequence of i.i.d. random variables with a continuously differentiable distribution function $F(\cdot)$ whose density function is denoted by $f(\cdot)$. The inverse $F^{-1}(\cdot)$ exists and is continuous, and the moment generating function of d_k exists.

13.2 Wonham-Type Filters

Tracking θ_k or identifying the system under binary-valued observations is a nonlinear filtering problem. A crucial step toward this goal is to build a good estimator of the probability distribution given the observations. The identification problem may be stated as follows.

Denote the observation data up to k by $S_k = \{s_l, l \leq k\}$ and the sequence of increasing σ-algebras of the observations up to time k by $\mathcal{F}_{S_k} = \sigma\{s_l : l \leq k\}$. Note that $\mathcal{F}_{S_0} \subset \mathcal{F}_{S_1} \cdots \subset \mathcal{F}_{S_k}$. Similarly, denote the sequence of

σ-algebras generated by θ_k as $\mathcal{F}_{\Theta_k} = \sigma\{\theta_l : l \leq k\}$, and the σ-algebras generated by the observation noise as $\mathcal{F}_{D_k} = \sigma\{d_l : l \leq k\}$. We wish to find the probabilities

$$w_N^{\{j\}} = P(\theta_N = \theta^{(j)} | \mathcal{F}_{S_N}), \quad N \geq 0, \ j = 1, \ldots, m_0. \tag{13.2}$$

Denote the initial probability distribution by $p_0^{\{j\}} = P(\theta_0 = \theta^{(j)})$. Recall that $P = (p^{\{ij\}}) \in \mathbb{R}^{m_0 \times m_0}$, with

$$p^{\{ij\}} = P(\theta_N = \theta^{(j)} | \theta_{N-1} = \theta^{(i)}), \ i, j = 1, \ldots, m_0,$$

are the entries in the transition matrix P. The development uses the Wonham filter techniques in [59], which is a discrete version of the original Wonham filter in [114]. Nevertheless, in our case, we only have binary-valued observations. The noise does not appear additive either. For each $j = 1, \ldots, m_0$, we denote

$$\begin{aligned} G^{\{j\}}(s_N) \ &:= P(s_N | \theta_N = \theta^{(j)}) = I_{\{s_N = 1\}} F(C - \phi_N' \theta^{(j)}) \\ &\quad + [1 - I_{\{s_N = 1\}}](1 - F(C - \phi_N' \theta^{(j)}), \end{aligned} \tag{13.3}$$

which is a function of the random variable s_N.

Theorem 13.1. *Assume* (A13.1) *and* (A13.2). *The Wonham-type filter for the binary-valued observations can be constructed as*

$$w_0^{\{j\}} = \frac{p_0^{\{j\}} G^{\{j\}}(s_0)}{\sum_{j_1=1}^{m_0} p_0^{\{j_1\}} G^{\{j_1\}}(s_0)}, \ j = 1, \ldots, m_0 \tag{13.4}$$

and

$$w_N^{\{j\}} = \frac{G^{\{j\}}(s_N) \sum_{i=1}^{m_0} p^{\{ij\}} w_{N-1}^{\{i\}}}{\sum_{i=1}^{m_0} \sum_{j_1=1}^{m_0} G^{\{j_1\}}(s_N) p^{\{ij_1\}} w_{N-1}^{\{j_1\}}}, \ j = 1, \ldots, m_0. \tag{13.5}$$

Proof. To verify (13.4), applying Bayes' theorem leads to

$$w_0^{\{j\}} = \frac{P(s_0 | \theta_0 = \theta^{(j)}) P(\theta_0 = \theta^{(j)})}{\sum_{j_1=1}^{m_0} P(s_0, \theta_N = \theta^{(j_1)})} = \frac{p_0^{\{j\}} G^{\{j\}}(s_0)}{\sum_{j_1=1}^{m_0} p_0^{\{j_1\}} G^{\{j_1\}}(s_0)}.$$

To prove (13.5), we first introduce the one-step prediction

$$w_{N|N-1}^{\{j\}} = P(\theta_N = \theta^{(j)} | \mathcal{F}_{S_{N-1}}). \tag{13.6}$$

Since $\{d_k\}$ is a sequence of i.i.d. random variables and $\{\theta_N\}$ is Markovian, we have

$$P(\theta_N = \theta^{(j)} | \theta_{N-1} = \theta^{(j_1)}, \mathcal{F}_{S_{N-1}}) = \theta^{(j_1)}) = p^{\{j_1 j\}}.$$

By the law of total probability,

$$E(w_N^{\{j\}} | \mathcal{F}_{S_{N-1}}) = P(\theta_N = \theta^{(j)} | \mathcal{F}_{S_{N-1}}) = \sum_{j_1=1}^{m_0} p^{\{j_1 j\}} w_{N-1}^{\{j_1\}}. \tag{13.7}$$

Now, by Bayes' theorem and (13.7),

$$\begin{aligned} w_N^{\{j\}} &= P(\theta_N = \theta^{(j)} | s_N, \mathcal{F}_{S_{N-1}}) \\ &= \frac{P(s_N | \theta_N = \theta^{(j)}, \mathcal{F}_{S_{N-1}}) P(\theta_N = \theta^{(j)} | \mathcal{F}_{S_{N-1}})}{\sum_{j_1=1}^{m_0} P(s_N | \theta_N = \theta^{(j_1)}, \mathcal{F}_{S_{N-1}}) P(\theta_N = \theta^{(j_1)} | \mathcal{F}_{S_{N-1}})} \\ &= \frac{G^{\{j\}}(s_N) \sum_{i=1}^{m_0} p^{\{ij\}} w_{N-1}^{\{i\}}}{\sum_{j_1=1}^{m_0} G^{j_1}(s_N) \sum_{i=1}^{m_0} p^{\{ik\}} w_{N-1}^{\{i\}}}. \end{aligned}$$

The last line above follows from (13.6). □

13.3 Tracking: Mean-Square Criteria

Based on Wonham-type filters, under different criteria, we may develop several different estimators. First, consider the following optimization problem: Choose θ to minimize the mean-square errors conditioned on the information up to time N. That is, find θ to minimize $\min_\theta E(|\theta_N - \theta|^2 | S_N)$. Just as in the usual argument for Kalman filters, bearing in mind the use of conditional expectation, we obtain the minimizer of the cost, which leads to the following mean-square estimator:

$$\widehat{\theta}_N = E(\theta_N | \mathcal{F}_{S_n}) = \sum_{j=1}^{m_0} \theta^{(j)} w_N^{\{j\}}.$$

To derive the error estimates of $\widehat{\theta}_N - \theta_N$, we need the associated asymptotic distribution for

$$e_N = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} (\theta_k - \widehat{\theta}_k).$$

Note that $\{e_N\}$ is a sequence of centered and scaled deviations of the Markov chain from its mean-square tracker with a scaling factor $\sqrt{N}$. For future use, we note that

$$\begin{aligned} e_N &= \frac{1}{\sqrt{N}} \sum_{j\in\mathcal{M}} \sum_{k=0}^{N-1} \theta^{(j)} \{[I_{\{\theta_k=\theta^{(j)}\}} - P(\theta_k = \theta^{(j)})] \\ &\quad + [P(\theta_k = \theta^{(j)}) - w_k^{\{j\}}]\}. \end{aligned} \tag{13.8}$$

To examine the deviation, in lieu of working with a discrete-time formula directly, we focus on a continuous-time interpolation of the form

$$v_N(t) = \frac{1}{\sqrt{N}} \sum_{j\in\mathcal{M}} \sum_{k=0}^{\lfloor Nt\rfloor-1} \theta^{(j)} [I_{\{\theta_k=\theta^{(j)}\}} - w_k^{\{j\}}],\ t \in [0,1], \tag{13.9}$$

where $\lfloor z\rfloor$ denotes the integer part of $z \in \mathbb{R}$. We shall show that the limit of $v_N(\cdot)$ is a Brownian motion, whose properties help us to derive the desired error bounds. Obtaining the weak convergence to the Brownian motion requires verifying that the sequence under consideration is tight (or compact). Then we characterize the limit by means of martingale problem formulation.

To use weak convergence theory, it is common and more convenient to use the so-called D space, which is a space of functions that are right continuous and have left limits, with a topology weaker than uniform convergence, known as the Skorohod topology. The main advantage of using such a setup is that it enables one to verify the tightness or compactness relatively easily. The exact definitions of these are somewhat technical; we refer the reader to [55, Chapter 7] for further reference.

Lemma 13.2. *Assume the conditions of Theorem* 13.1*, and suppose the Markov chain is irreducible.*

(a) *Then for each $\delta > 0$, each $t \geq 0$, and each $s > 0$ with $0 \leq s \leq \delta$,*

$$\sup_N E|v_N(t+s) - v_N(t)|^2 \leq Ks, \tag{13.10}$$

for some $K > 0$.

(b) *The sequence $v_n(\cdot)$ is tight in $D([0,1];\mathbb{R}^{m_0})$, the space of $\mathbb{R}^{m_0}$-valued functions that are right continuous, have left limits, and are endowed with the Skorohod topology.*

Proof. We first prove (a). In view of the second line of (13.8), for each $\delta > 0$, for $t > 0$, $s > 0$ satisfying $0 \leq s \leq \delta$, we have

$$E|v_N(t+s) - v_N(t)|^2 \leq L_{N,1} + L_{N,2},$$

where

$$
\begin{aligned}
L_{N,1} &= \frac{2}{N} \sum_{j_1,j_2 \in \mathcal{M}} \sum_{k=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \sum_{i=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \theta^{(j_1),\prime} \theta^{(j_2)} E I_k^{\{j_1\}} I_i^{\{j_2\}}, \\
L_{N,2} &= \frac{2}{N} \sum_{j_1,j_2 \in \mathcal{M}} \sum_{k=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \sum_{i=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \theta^{(j_1),\prime} \theta^{(j_2)} E \check{w}_k^{\{j_1\}} \check{w}_i^{\{j_2\}}, \\
I_k^{\{j\}} &= I_{\{\theta_k = \theta^{(j)}\}} - P(\theta_k = \theta^{(j)}), \\
\check{w}_k^{\{j\}} &= P(\theta_k = \theta^{(j)}) - w_k^{\{j\}}, \\
\widetilde{w}_k^{\{j\}} &= I_{\{\theta_k = \theta^{(j)}\}} - w_k^{\{j\}}.
\end{aligned}
\tag{13.11}
$$

To proceed, we estimate $L_{N,1}$ and $L_{N,2}$. We need only look at a fixed pair j_1 and j_2. First, consider $L_{N,1}$ without the first sum. Without loss of generality, assume $k \geq i$. Then we obtain that for fixed j_1 and $j_2 \in \mathcal{M}$,

$$
\begin{aligned}
&\left| \sum_{i=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \sum_{k=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \theta^{(j_1),\prime} \theta^{(j_2)} E I_k^{\{j_1\}} I_i^{\{j_2\}} \right| \\
&\quad \leq 2 \sum_{i=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \sum_{i \leq k}^{\lfloor N(t+s) \rfloor - 1} \theta^{(j_1),\prime} \theta^{(j_2)} \left| E[I_i^{\{j_2\}} E_i I_k^{\{j_1\}}] \right|,
\end{aligned}
\tag{13.12}
$$

where E_l denotes the expectation conditioned on $\mathcal{F}_l = \sigma\{d_k, \theta_k : k \leq l\}$, the past information up to l. Since the Markov chain is irreducible, it is ergodic. That is, there is a row vector ν, the stationary distribution of the Markov chain such that

$$
|P^N - \mathbb{1}\nu| \leq K\lambda^N \quad \text{for some} \quad 1 > \lambda > 0,
$$

where $\mathbb{1}$ is a column vector with all its component being 1. Using this spectrum gap estimate,

$$
|E_i I_k^{\{j_1\}}| \leq \lambda^{k-i}.
$$

It then follows that for the term in (13.12), we have

$$
\begin{aligned}
&\left| \sum_{i=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \sum_{k=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \theta^{(j_1),\prime} \theta^{(j_2)} E I_k^{\{j_1\}} I_i^{\{j_2\}} \right| \\
&\quad \leq K \left(\lfloor N(t+s) \rfloor - \lfloor Nt \rfloor \right).
\end{aligned}
$$

Dividing the above by N leads to $\sup_N L_{N,1} \leq Ks$. As for the terms

involved in $L_{N,2}$,

$$\left| \sum_{i=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \sum_{i \le k}^{\lfloor N(t+s) \rfloor - 1} \theta^{(j_1),\prime} \theta^{(j_2)} E \check{w}_k^{\{j_1\}} \check{w}_i^{\{j_2\}} \right|$$
$$\le K \sum_{i=\lfloor Nt \rfloor}^{\lfloor N(t+s) \rfloor - 1} \sum_{k \ge i}^{\lfloor N(t+s) \rfloor - 1} \lambda^{k-i} \le KNs.$$

Dividing the above by N and taking $\sup_N$ yields $\sup_N L_{N,2} \le Ks$. Thus, (a) is true.

By using (a), with arbitrary $\delta > 0$ and the chosen t and s, we have

$$\lim_{\delta \to 0} \limsup_{N \to \infty} E|v_N(t+s) - v_N(t)|^2 \le \lim_{\delta \to 0} K\delta = 0.$$

Thus, the tightness follows from the criterion [53, Theorem 3, p. 47]. The lemma is proved. □

To proceed, let us point out:

(i) Since $v_N(\cdot)$ is tight, we can extract weakly convergent subsequences by means of Prohorov's theorem (see [55, Chapter 7]). Loosely, sequential compactness enables us to extract convergent subsequences. Without loss of generality, still index the selected subsequence by N, and assume $v_N(\cdot)$ itself is the weakly convergent subsequence. Denote the limit by $v(\cdot)$. We shall characterize the limit process.

(ii) From the defining relationship of $v_N(t)$, it is readily seen that

$$\begin{aligned} Ev_N(t) &= \frac{1}{\sqrt{N}} \sum_{j \in \mathcal{M}} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \theta^{(j)} \left\{ E\left[I_{\{\theta_k = \theta^{(j)}\}} - P\left(\theta_k = \theta^{(j)}\right) \right] \right. \\ &\quad \left. + E\left[P\left(\theta_k = \theta^{(j)}\right) - w_k^{\{j\}} \right] \right\} = 0 \quad \text{for each} \quad t \ge 0. \end{aligned} \tag{13.13}$$

To determine the limit process, we consider a vector-valued process $\widetilde{v}_N(t) = (\widetilde{v}_N^1(t), \ldots, \widetilde{v}_N^{m_0}(t)) \in \mathbb{R}^{m_0}$, where

$$\widetilde{v}_N^{\{i\}}(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \theta^{(i)} \left[I_{\{\theta_k = \theta^{(i)}\}} - w_k^{\{i\}} \right].$$

Define $\Sigma_N(t) = (\Sigma^{\{ij\}}(t)) = E\widetilde{v}_N(t)\widetilde{v}_N'(t)$, where $\Sigma_N^{\{ij\}}(t)$ denotes the ijth entry of the partitioned matrix $\Sigma_N(t)$, namely,

$$\Sigma_N^{\{ij\}}(t) = \frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \sum_{l=0}^{\lfloor Nt \rfloor - 1} E\zeta_k^{\{i\}} \zeta_l^{\{j\},\prime}, \tag{13.14}$$

where

$$\zeta_k^{\{i\}} = \theta^{(i)} \left[I_{\{\theta_k = \theta^{(i)}\}} - w_k^{\{i\}} \right] \in \mathbb{R}^{m_0}.$$

Using the notation of $\widetilde{v}_N(t)$, we can rewrite $v_N(t)$ as $v_N(t) = \mathbb{1}'_{m_0} \widetilde{v}_N(t)$, where $\mathbb{1}'_{m_0} = (1, \ldots, 1) \in \mathbb{R}^{1 \times m_0}$. To proceed, we first determine the limit covariance function of $Ev_N(t)v'_N(t) = \mathbb{1}'_{m_0}[E\widetilde{v}_N(t)\widetilde{v}'_N(t)]\mathbb{1}_{m_0}$. From the above expression, it is seen that to accomplish this goal, we need only consider the limit covariance of $E\widetilde{v}_N(t)\widetilde{v}'_N(t)$. The following lemma details the calculation of the asymptotic covariance.

Lemma 13.3. *Assume the conditions of Lemma* 13.2. *Then*

(a) *the limit covariance of* $\widetilde{v}_N(t)$ *is given by*

$$\begin{aligned} &\lim_{N\to\infty} \Sigma_N(t) = t\Sigma_0, \quad \Sigma_0 = \operatorname{diag}\left(\Sigma_0^{\{11\}}, \ldots, \Sigma_0^{\{m_0 m_0\}}\right), \\ &\Sigma_0^{\{i\}} \stackrel{\text{def}}{=} \Sigma_0^{\{ii\}} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} E\zeta_k^{\{i\}} \zeta_k^{\{i\},\prime}, \ i \in \mathcal{M}; \end{aligned} \tag{13.15}$$

(b) *as* $N \to \infty$,

$$Ev_N(t)v'_N(t) \to t\overline{\Sigma} = t \sum_{i=1}^{m_0} \Sigma_0^{\{i\}}.$$

Proof. To prove (a), it suffices to work with the partitioned matrix $\Sigma_N^{\{ij\}}(t)$. Note that

$$\sum_{k<l}^{\lfloor Nt \rfloor - 1} E\zeta_k^{\{i\}} \zeta_l^{\{j\},\prime} = \sum_{k<l}^{\lfloor Nt \rfloor - 1} E\left[\zeta_k^{\{i\}} E_k \zeta_l^{\{j\},\prime}\right] = 0 \ \text{ for } \ i \neq j.$$

Likewise,

$$\sum_{l<k}^{\lfloor Nt \rfloor - 1} E\zeta_k^{\{i\}} \zeta_l^{\{j\},\prime} = 0 \ \text{ for } \ i \neq j.$$

This leads to

$$\begin{aligned} \Sigma_N^{\{ij\}}(t) &= \delta_{ij} \frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} E\zeta_k^{\{i\}} \zeta_k^{\{j\},\prime} + \frac{1}{N} \sum_{l=0}^{\lfloor Nt \rfloor - 1} \sum_{k<l}^{\lfloor Nt \rfloor - 1} E\zeta_k^{\{i\}} \zeta_l^{\{j\},\prime} \\ &\quad + \frac{1}{N} \sum_{k=0}^{\lfloor Nt \rfloor - 1} \sum_{l<k}^{\lfloor Nt \rfloor - 1} E\zeta_k^{\{i\}} \zeta_l^{\{j\},\prime} \\ &= \delta_{ij} \frac{\lfloor Nt \rfloor}{N} \frac{1}{\lfloor Nt \rfloor} \sum_{k=0}^{\lfloor Nt \rfloor - 1} E\zeta_k^{\{i\}} \zeta_k^{\{j\},\prime} \\ &\to \begin{cases} t\Sigma_0^{\{i\}}, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{13.16}$$

as $N \to \infty$, where $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ otherwise, and $\Sigma_0^{\{ij\}}$ denotes the ijth partitioned matrix in Σ_0.

Finally, (b) is a direct consequence of Lemma 13.3. The lemma is thus proved. □

Note that by (13.15), $\Sigma_0^{\{i\}} = \Sigma_0^{\{ii\}}$, the partitioned matrix of Σ_0. To proceed, we prove that $v_N(\cdot)$ converges weakly to a Brownian motion. We characterize the limit process by means of identifying the limit covariance function. The analysis is carried out by using the martingale problem formulation. For a twice continuously differentiable function $h : \mathbb{R}^{m_0} \mapsto \mathbb{R}$, define an operator as

$$\mathcal{L}h(v) = \frac{1}{2}\text{tr}\left[\overline{\Sigma}h_{vv}(v)\right], \tag{13.17}$$

where h_{vv} denotes the Hessian matrix (the second partial derivatives with respect to v). We have the following result.

Theorem 13.4. *Assume the conditions of Lemma* 13.2. *Then*

(a) $v_N(\cdot)$ *converges weakly to* $v(\cdot)$, *which is a Brownian motion with covariance* $\overline{\Sigma}t$;

(b) $v_N(1)$ *converges in distribution to a normal random variable with mean 0 and covariance* $\overline{\Sigma}$.

Proof. Part (b) is a direct consequence of (a). Thus, we need only prove (a). Since $v_N(\cdot)$ converges weakly, there is a convenient device known as the Skorohod representation (see [55, Chapter 7]) that enables us to work with w.p.1 convergence on an enlarged space. Without loss of generality and with a slight abuse of notation, we may assume $v_N(\cdot) \to v(\cdot)$ in the sense of w.p.1. We want to show that $v(\cdot)$ is a solution to the martingale problem with operator $\mathcal{L}$ defined in (13.17). To this end, it suffices to show that

$$h(v(t)) - h(v(0)) - \int_0^t \mathcal{L}h(v(\rho))d\rho \text{ is a martingale.}$$

To verify the above, it only needs to be shown (see [55]) that for any bounded and continuous function $H(\cdot)$, any $t, s > 0$, any integers ℓ, and any $t_\iota \le t$,

$$\begin{aligned} EH(v(t_\iota) : \iota \le \ell)[\, & h(v(t+s)) - h(v(t)) \\ & - \int_t^{t+s} \mathcal{L}h(v(\rho))d\rho] = 0. \end{aligned} \tag{13.18}$$

To verify (13.18), use $v_n(\cdot)$. By the weak convergence and the Skorohod representation, as $N \to \infty$,

$$\begin{aligned} & EH(v_N(t_\iota) : \iota \le \ell)[h(v_N(t+s)) - h(v_N(t))] \\ & \quad \to EH(v(t_\iota) : \iota \le \ell)[h(v(t+s)) - h(v(t))]. \end{aligned} \tag{13.19}$$

On the other hand, direct computation reveals that

$$\begin{aligned}
&\lim_{N\to\infty} EH(v_N(t_\iota):\iota\le\ell)[h(v_N(t+s))-v_N(t))]\\
&=\lim_{N\to\infty} EH(v_N(t_\iota):\iota\le\ell)\left[\frac{1}{\sqrt{N}}\sum_{i\in\mathcal{M}} h_v'(v_N(t))\theta^{(i)}\sum_{k=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\widetilde{w}_k^{\{i\}}\right.\\
&\quad\left.+\frac{1}{2N}\sum_{i\in\mathcal{M}}\sum_{i_1\in\mathcal{M}}\mathrm{tr}[h_{vv}(v_N(t))\theta^{(i)}\theta^{(i_1)}]\sum_{k=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\sum_{l=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\widetilde{w}_k^{\{i\}}\widetilde{w}_l^{i_1}\right],
\end{aligned}\tag{13.20}$$

where $\widetilde{w}_k^{\{i\}}$ is given by (13.11). Using nested expectation and inserting $E_{\lfloor Nt\rfloor}$, since $v_N(t_\iota):\iota\le\ell$ and $v_N(t)$ are all $\mathcal{F}_{\lfloor Nt\rfloor}$-measurable, by inserting $E_{\lfloor Nt\rfloor}$ we have

$$\begin{aligned}
&EH(v_N(t_\iota):\iota\le\ell)\left[\frac{1}{\sqrt{N}}\sum_{i\in\mathcal{M}} h_v'(v_N(t))\theta^{(i)}E_{\lfloor Nt\rfloor}\sum_{k=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\widetilde{w}_k^{\{i\}}\right]\\
&\quad\to 0\ \text{ as }\ N\to\infty.
\end{aligned}$$

Likewise,

$$\begin{aligned}
&EH(v_N(t_\iota):\iota\le\ell)\left[\frac{1}{2N}\sum_{i\in\mathcal{M}}\sum_{i_1\in\mathcal{M}}\mathrm{tr}[h_{vv}(v_N(t))\theta^{(i)}\theta^{(i_1)}]\right.\\
&\qquad\left.\times\sum_{k=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\sum_{l=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\widetilde{w}_k^{\{i\}}\widetilde{w}_l^{i_1}\right]\\
&=EH(v_N(t_\iota):\iota\le\ell)\left[\frac{1}{2N}\sum_{i\in\mathcal{M}}\sum_{i_1\in\mathcal{M}}\mathrm{tr}[h_{vv}(v_N(t))\theta^{(i)}\theta^{(i_1)}]\right.\\
&\qquad\left.\times E_{\lfloor Nt\rfloor}\sum_{k=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\sum_{l=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\widetilde{w}_k^{\{i\}}\widetilde{w}_l^{i_1}\right].
\end{aligned}$$

Dividing the cases into $l\le k$ and $k<l$, we can handle the last equation above as in the proof of Lemma 13.3 by inserting E_l and E_k, respectively. It follows that the double summations above reduce to a single one. The last two equations together with (13.20) then imply that

$$\begin{aligned}
&EH(v_N(t_\iota):\iota\le\ell)\Big[\frac{1}{2N}\sum_{i\in\mathcal{M}}\sum_{i_1\in\mathcal{M}}\mathrm{tr}[h_{vv}(v_N(t))\theta^{(i)}\theta^{(i_1)}]\\
&\qquad\times E_{\lfloor Nt\rfloor}\sum_{k=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\sum_{l=\lfloor Nt\rfloor}^{\lfloor N(t+s)\rfloor-1}\widetilde{w}_k^{\{i\}}\widetilde{w}_l^{i_1}\Big]\\
&\quad\to EH(v(t_\iota):\iota\le\ell)\Big[\int_t^{t+s}\mathcal{L}h(\rho)d\rho\Big]\ \text{ as }\ N\to\infty.
\end{aligned}$$

This establishes the desired theorem. □

By virtue of Theorem 13.4, we further obtain a strong approximation result. This strong approximation will aid us in obtaining error bounds in what follows.

Lemma 13.5. *Under the conditions of Theorem* 13.4*, there is a constant* $\gamma > 0$ *such that*

$$\sup_{0\le t\le 1} |v_N(t) - v(t)| = o(N^{-\gamma}) \quad w.p.1.$$

Proof. Note that

$$E_{k-1}\zeta_k^{\{i\}} = E_{k-1}[I_{\{\theta_k=\theta^{(i)}\}} - w_k^{\{i\}}] = 0.$$

Thus, it is a martingale difference sequence. Using the martingale version of the Skorohod representation (see [41, p. 269]), we can establish the result. The details are omitted. □

We next show that the tracking error in the average sense is exponentially small. The result is based on part (b) in Theorem 13.4 and large deviations for normal random variables. There are different ways to obtain the error bounds. We do one as follows, whose proof is also in the appendix.

Theorem 13.6. *Under the conditions of Lemma* 13.2*,*

$$\begin{aligned} P\left(\frac{1}{\sqrt{N}}\Big|\sum_{i\in\mathcal{M}}\sum_{k=0}^{N}\theta^{(i)}[I_{\{\theta_k=\theta^{(i)}\}} - w_k^{\{i\}}]\Big|_1 \ge \varepsilon\right) \\ \le 2m_0 \exp\left(-\frac{N\varepsilon^2}{2{m_0}^2\sigma^2_{v^{\{i\}}(1)}}\right), \end{aligned} \tag{13.21}$$

where $|\cdot|_1$ *denotes the* l_1 *norm.*

Proof. Note that $v^{\{i\}}(1) = e_i'v(1)$, where e_i is the ith standard unit vector. Note also that $v^{\{i\}}(1)$ is normally distributed with mean 0 and variance $\sigma^2_{v^{\{i\}}(1)} = e_i'\overline{\Sigma}e_i$. We then have that for any $\alpha > 0$,

$$\begin{aligned} P\left(\frac{1}{\sqrt{N}}|v^{\{i\}}(1)| \ge \frac{\varepsilon}{m_0}\right) &\le \exp\left(-\frac{\alpha\varepsilon}{m_0}\right) E\exp\left(\frac{\alpha|v^{\{i\}}(1)|}{\sqrt{N}}\right) \\ &\le 2\exp\left(-\frac{\alpha\varepsilon}{m_0} + \frac{\sigma^2_{v^{\{i\}}(1)}\alpha^2}{2N}\right). \end{aligned} \tag{13.22}$$

Choosing the α to minimize the index in the exponent leads to

$$\alpha = (N\varepsilon/(m_0\sigma^2_{v^{\{i\}}(1)})).$$

Using this in (13.22) yields the upper bound

$$P\left(\frac{1}{\sqrt{N}}|v^{\{i\}}(1)| \geq \frac{\varepsilon}{m_0}\right) \leq 2\exp\left(-\frac{N\varepsilon^2}{2{m_0}^2\sigma^2_{v^{\{i\}}(1)}}\right).$$

Thus,

$$P\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{m_0}|v^{\{i\}}(1)| \geq \varepsilon\right) \leq 2m_0\exp\left(-\frac{N\varepsilon^2}{2{m_0}^2\sigma^2_{v^{\{i\}}(1)}}\right). \tag{13.23}$$

Note that

$$\begin{aligned} &P\left(\frac{1}{\sqrt{N}}\left|\sum_{i\in\mathcal{M}}\sum_{k=0}^{N}\theta^{(i)}\widetilde{w}_k^{\{i\}}\right|_1 \geq \varepsilon\right) \\ &\quad = P\left(\exp\left(\frac{\alpha}{\sqrt{N}}\left|\sum_{i\in\mathcal{M}}\sum_{k=0}^{N}\theta^{(i)}\widetilde{w}_k^{\{i\}}\right|_1\right) \geq \exp(\alpha\varepsilon)\right). \end{aligned}$$

We can approximate the

$$(\alpha/\sqrt{N})\sum_{k=0}^{N}\theta^{(i)}\widetilde{w}_k^{\{i\}}$$

by $v^{\{i\}}(t)$ by using Lemma 13.5. Adding and subtracting $v^{\{i\}}(t)$ in the above and using the triangle inequality yield that

$$\begin{aligned} &P\left(\frac{1}{\sqrt{N}}\left|\sum_{i\in\mathcal{M}}\sum_{k=0}^{N}\theta^{(i)}\widetilde{w}_k^{\{i\}}\right|_1 \geq \varepsilon\right) \\ &\leq \exp\left(-\frac{\alpha\varepsilon}{m_0}\right)E\exp\left(\frac{\alpha}{\sqrt{N}}\left|\sum_{i\in\mathcal{M}}\sum_{k=0}^{N}\{\theta^{(i)}\widetilde{w}_k^{\{i\}} - v^{\{i\}}(t)\}\right|_1\right. \\ &\qquad\qquad \left. + \frac{\alpha}{\sqrt{N}}\sum_{i\in\mathcal{M}}|v^{\{i\}}(t)|_1\right) \\ &\leq \exp\left(-\frac{\alpha\varepsilon}{m_0}\right)E\exp(o(N^{-\gamma}))\exp\left(\frac{\alpha}{\sqrt{N}}\sum_{i\in\mathcal{M}}|v^{\{i\}}(t)|_1\right). \end{aligned}$$

Using (13.23) in the above estimate, the desired result then follows. □

13.4 Tracking Infrequently Switching Systems: MAP Methods

Here, we construct a sequence of estimates of the Markov chain by maximizing the *a posterior* probabilities. The estimator is given by

$$\widehat{\theta}_N = \theta^{(j_N)}, \quad j_N = \mathrm{argmax}_{j\in\mathcal{M}}w_N^{\{j\}}. \tag{13.24}$$

Our goal is to derive an error bound on $P(\widehat{\theta}_N \neq \theta_N)$. We are interested in the case that for each $i \in \mathcal{M}$, $\sum_{j \neq i} p_{ij} = \varepsilon$ and $p_{ii} = 1 - \varepsilon$, for ε sufficiently small. One such model assumes the transition probability of the Markov chain to be $P^\varepsilon = I + \varepsilon Q$, where Q is a generator of a continuous-time Markov chain. It indicates that most of the time, the system will remain at a constant value, but it has infrequent jumps from one parameter value to another at random times. This is a class of "infrequent switching" systems. It is intuitively understood that the data size n should be neither too small for lack of information from data nor too large since old data will contain diminishing information about the current θ_N. It is also conceivable that for smaller ε, a larger N may be used. It is our desire to establish a concrete relationship between N and ε to guarantee a desired accuracy of identification. For a selected N, the implementation is the standard moving-window method: To identify θ_k, the data in the time window $l = k - N$, $k - N + 1$, ..., k, will be used. It is noted for a large window size N, the initial distribution of θ_{k-N} will have diminishing effects on the MAP estimates $\widehat{\theta}_N$, at the end of the window $k - 1 \leq l \leq N$. Consequently, one may choose any initial distribution, such as the uniform distribution, to start the MAP algorithm. The following discussion is generic for a given moving window with a chosen initial distribution. For simplicity, we make the following assumption.

(A13.3) $y_k = \theta_k + d_k$ and the initial probability distribution of the Markov chain satisfies $p^{\{j\}}(0) > 0$ for $j = 1, \ldots, m_0$.

In what follows, we choose ε to be sufficiently small and N sufficiently large. Let $p_j = F(C - \theta^{(j)})$ and $\delta = \min_{i \neq j} |p_i - p_j|$. Define

$$\xi_{N+1} = \frac{1}{N+1} \sum_{k=0}^{N} s_k, \tag{13.25}$$

and denote the data set by $S_N = \{s_k, k = 0, \ldots, N\}$. For a given $\beta < \delta/2$, define

$$M_N^{\{j\}} = \{s_k : 0 \leq k \leq N,\ |\xi_{N+1} - p_j| < \beta\},$$
$$M_N = \bigcup_{j=1}^{m_0} M_N^{\{j\}}.$$

Lemma 13.7. *For sufficiently small ε and sufficiently large N and some constant $c > 0$,*

$$P(M_N) \geq (1 - e^{-N\beta^2 c})(1 - \varepsilon)^N, \tag{13.26}$$

which implies

$$\lim_{N \to \infty} \lim_{\varepsilon \to 0} P(M_N) = 1.$$

Proof. Note that

$$\begin{aligned} P(M_N^{\{j\}}) &= \sum_{i=1}^{m_0} P(M_N^{\{j\}}|\theta_0 = \theta^{(i)}) p_0^{\{i\}} \\ &= \left[p_0^{\{j\}} P(M_N^{\{j\}}|\theta_0 = \theta^{(j)}) + \sum_{i \neq j} P(M_N^{\{j\}}|\theta_0 = \theta^{(i)}) p_0^{\{i\}} \right] \\ &\geq p_0^{\{j\}} P(M_N^{\{j\}}|\theta_0 = \theta^{(j)}). \end{aligned}$$

Then,

$$\begin{aligned} P(M_N^{\{j\}}) &\geq p_0^{\{j\}} P(M_N^{\{j\}}|\theta_k = \theta^{(j)}, k = 0, \ldots, N) \\ &\quad \times P(\theta_k = \theta^{\{j\}}, k = 1, \ldots, N|\theta_0 = \theta^{(j)}) \\ &= p_0^{\{j\}} P(M_N^{\{j\}}|\theta_k = \theta^{(j)}, k = 0, \ldots, N)(1-\varepsilon)^N \\ &= p_0^{\{j\}} P(|\xi_{N+1} - p_j| \leq \beta|\theta_k = \theta^{(j)}, k = 0, \ldots, N)(1-\varepsilon)^N. \end{aligned}$$

By the large deviations principle,

$$P(|\xi_{N+1} - p_j| \leq \beta|\theta_k = \theta^{(j)}, k = 0, \ldots, N) \geq 1 - e^{-N\beta^2 c}$$

for some $c > 0$. This implies

$$P(M_N^{\{j\}}) \geq p_0^{\{j\}}(1 - e^{-N\beta^2 c})(1-\varepsilon)^N.$$

Since $\beta < \delta/2$, $M_N^{\{j\}}$, $j = 1, \ldots, m_0$, are disjoint. Hence,

$$P(M_N) = \sum_{j=1}^{m_0} P(M_N^{\{j\}})$$

and (13.26) follows. This completes the proof. □

A sequence in $\{M_N\}$ is called a *typical sequence*. Lemma 13.7 indicates that for small ε and large N, the probability for a sequence to be typical is nearly 1. For this reason, to derive error bounds in probability, we may consider only the data set in M_N.

Lemma 13.8. *For sufficiently small ε and β, and sufficiently large N, if $S_N \in M_N^{\{j\}}$, then $\widehat{\theta}_N = \theta^{(j)}$.*

Proof. Suppose $S_N \in M_N^{\{j\}}$. Using the MAP estimator, $\widehat{\theta}_N = \theta^{(j)}$ if and only if

$$P(\theta_N = \theta^{(j)}|S_N) > P(\theta_N = \theta^{(i)}|S_N),\ i \neq j.$$

Since

$$P(\theta_N = \theta^{(i)}|S_N) = \frac{P(\theta_N = \theta^{(i)}, S_N)}{P(S_N)}$$

the conclusion is true if

$$P(\theta_N = \theta^{(j)}, S_N) > P(\theta_N = \theta^{(i)}, S_N),\ i \neq j.$$

In the following derivation, K_1, K_2, and K are some positive constants. Now

$$\begin{aligned}
&P(\theta_N = \theta^{(i)}, S_N) \\
&= \sum_{l=1}^{m_0} P(\theta_N = \theta^{(i)}, S_N | \theta_0 = \theta^{(l)}) p_0^l \\
&= p_0^{\{i\}} P(\theta_N = \theta^{(i)}, S_N | \theta_0 = \theta^{(i)}) + \varepsilon K_1 \\
&= p_0^{\{i\}} P(S_N | \theta_k = \theta^{(i)}, k = 0, \ldots, N) \\
&\quad \times P(\theta_k = \theta^{(i)}, k = 1, \ldots, N-1 | \theta_0 = \theta^{(i)}) + \varepsilon K_2 + \varepsilon K_1 \\
&= p_0^{\{i\}} P(S_N | \theta_k = \theta^{(i)}, k = 0, \ldots, N)(1-\varepsilon)^N + \varepsilon K.
\end{aligned}$$

By the definition of ξ_{N+1}, S_N contains $(N+1)\xi_{N+1}$ of 1's and $(N+1)(1-\xi_{N+1})$ of 0's:

$$P(S_N | \theta_k = \theta^{(i)}, k = 0, \ldots, N) = p_i^{(N+1)\xi_{N+1}} (1-p_i)^{(N+1)(1-\xi_{N+1})}.$$

Consequently,

$$\begin{aligned}
&P(\theta_N = \theta^{(i)}, S_N) \\
&= \frac{1}{m_0} p_i^{(N+1)\xi_{N+1}} (1-p_i)^{(N+1)(1-\xi_{N+1})} (1-\varepsilon)^N + \varepsilon K.
\end{aligned} \tag{13.27}$$

For sufficiently small ε, the first term is dominant. As a result, to prove

$$P(\theta_N = \theta^{(j)}, S_N) > P(\theta_N = \theta^{(i)}, S_N),\ i \neq j,$$

we need only show

$$\begin{aligned}
&p_j^{(N+1)\xi_{N+1}} (1-p_j)^{(N+1)(1-\xi_{N+1})} \\
&\quad > p_i^{(N+1)\xi_{N+1}} (1-p_i)^{(N+1)(1-\xi_{N+1})},
\end{aligned}$$

or equivalently, if

$$\begin{aligned}
&\xi_{N+1} \log p_j + (1-\xi_{N+1}) \log(1-p_j) \\
&\quad > \xi_{N+1} \log p_i + (1-\xi_{N+1}) \log(1-p_i).
\end{aligned} \tag{13.28}$$

Since $S_N \in M_N^{\{j\}}$, $p_j - \beta \leq \xi_{N+1} \leq p_j + \beta$. Now, the convex inequality [74, p. 643], which is, in fact, the relative entropy or Kullback–Leibler distance [22, p. 18],

$$p_j \log p_j + (1-p_j) \log(1-p_j) > p_j \log p_i + (1-p_j) \log(1-p_i),\ p_i \neq p_j,$$

and the continuity imply that for sufficiently small β, (13.28) holds. This concludes the proof. □

We now derive error bounds on the MAP algorithm.

Theorem 13.9. *Let $0 < \beta < \delta/2$ be a sufficiently small constant. For sufficiently small ε and sufficiently large N,*

$$P(\widehat{\theta}_N = \theta_N) \geq (1-\beta)2^{-(N+1)\beta}(1-\varepsilon)^N. \tag{13.29}$$

Proof. By Lemma 13.7, we may focus on $S_N \in M_N$. The probability of correct identification of θ_N is

$$\begin{aligned} P(\widehat{\theta}_N = \theta_N) &= \sum_{S_N} P(\widehat{\theta}_N = \theta_N | S_N) P(S_N) \\ &\geq \sum_{S_N \in M_N} P(\widehat{\theta}_N = \theta_N | S_N) P(S_N) \\ &= \sum_{j=1}^{m_0} \sum_{S_N \in M_N^{\{j\}}} P(\widehat{\theta}_N = \theta_N | S_N) P(S_N). \end{aligned}$$

By Lemma 13.8, for $S_N \in M_N^{\{j\}}$, $\widehat{\theta}_N = \theta^{(j)}$, which implies

$$P(\widehat{\theta}_N = \theta_N) \geq \sum_{j=1}^{m_0} \sum_{S_N \in M_N^{\{j\}}} P(\theta_N = \theta^{(j)} | S_N) P(S_N).$$

By (13.27),

$$\begin{aligned} &P(\theta_N = \theta^{(j)} | S_N) P(S_N) = P(\theta_N = \theta^{(j)}, S_N) \\ &\quad = p_0^{\{j\}} p_j^{(N+1)\xi_{N+1}} (1-p_j)^{(N+1)(1-\xi_{N+1})} (1-\varepsilon)^N + \varepsilon K. \end{aligned}$$

Let $\lambda_N^{\{j\}}$ be the cardinality of $M_N^{\{j\}}$. Then,

$$\begin{aligned} &\sum_{S_N \in M_N^{\{j\}}} p_j^{(N+1)\xi_{N+1}} (1-p_j)^{(N+1)(1-\xi_{N+1})} \\ &\qquad = \lambda_N^{\{j\}} p_j^{(N+1)\xi_{N+1}} (1-p_j)^{(N+1)(1-\xi_{N+1})}. \end{aligned}$$

By [22, Theorem 3.1.2, p. 51], for sufficiently small ε and sufficiently large N,

$$\lambda_N^{\{j\}} \geq (1-\beta)2^{(N+1)(H(p_j)-\beta)}$$

for any small β, where

$$H(p_j) = -p_j \log p_j - (1-p_j)\log(1-p_j)$$

is the entropy of p_j. Denote

$$H(p_j, \xi_{N+1}) = -\xi_{N+1} \log p_j - (1 - \xi_{N+1}) \log(1 - p_j).$$

Since $\xi_N \to p_j$ w.p.1 as $N \to \infty$, $H(p_j, \xi_{N+1}) \to H(p_j)$ w.p.1 as $N \to \infty$. It follows that

$$\begin{aligned} &\lambda_N^{\{j\}} p_j^{(N+1)\xi_{N+1}} (1 - p_j)^{(N+1)(1-\xi_{N+1})} \\ &\quad \geq (1 - \beta) 2^{(N+1)(H(p_j) - H(p_j, \xi_{N+1}) - \beta)} \\ &\quad = (1 - \beta) 2^{-N\beta} 2^{-\beta + o(1)}, \end{aligned}$$

where $o(1) \to 0$ w.p.1 as $N \to \infty$. For sufficiently small β and sufficiently large N, $2^{-\beta + o(1)} > 1 - \beta$ since $\beta > 1 - 2^{-\beta}$. As a result, for sufficiently small ε and sufficiently large N,

$$\sum_{S_N \in M_N^{\{j\}}} P(\theta_N = \theta^{(j)} | S_N) P(S_N) > p_0^{\{j\}} (1 - \beta)^2 2^{-N\beta} (1 - \varepsilon)^N.$$

Therefore, (13.29) is obtained. □

Theorem 13.9 provides a guideline for window size selection. To achieve a required estimation accuracy for $0 < \eta < 1$, we may select the window size to be $(1 - \beta)^2 2^{-N\beta} (1 - \varepsilon)^N = \eta$ provided that ε is sufficiently small and β is sufficiently small.

13.5 Tracking Fast-Switching Systems

We begin this section by considering the scenario that the process $\theta(t)$ is a continuous-time Markov chain whose states vary on a fast pace. It is now understood that depending on the actual scenarios, only when the speed or frequency of state variations is relatively small, one expects to track the time-varying parameters with reasonable accuracy [6]. For instance, consider a continuous-time system whose observation is given by

$$y(t) = \varphi'(t)\theta(t) + w(t),$$

where $\varphi(t)$ is the input. Suppose that the parameter process $\theta(t)$ is a continuous-time Markov chain with a finite state space $\mathcal{M}$ and generator $Q^\eta = Q/\eta$ with $\eta > 0$ a small parameter. With Q being irreducible, when $\eta \to 0$, within a very short period of time $\theta(t)$ reaches its stationary distribution. In this case, it is virtually impossible to track the instantaneous variation of the process from observations of binary-valued outputs $s(t) = I_{\{y(t) \leq C\}}$. For such systems, the main goal becomes identifying an averaged system (averaging with respect to the stationary measure of the Markov chain). The main reason for focusing on the averaged system is

the following: When a system performance is measured by some averaged outputs, as in most performance indices for optimal or adaptive control, the net effect of fast-switching parameters on the system performance can be approximated by using their average values.

The development of this section is motivated by the following scenario: For the above parameter process $\theta(t)$, denote its transition matrix by $P(t) = P^\eta(t)$. Then $P^\eta(t)$ satisfies the forward equation

$$\dot{P}^\eta(t) = P^\eta(t)Q/\eta.$$

A change of variables $\tau = t/\eta$ and $P(\tau) = P^\eta(t)$ leads to

$$\frac{d}{d\tau}P(\tau) = P(\tau)Q.$$

Discretizing the equation with a step size $h > 0$, we obtain a discrete matrix recursion

$$P^{k+1} = P^k[I + hQ].$$

By choosing $h > 0$ properly, $I+hQ$ becomes a one-step transition matrix of a discrete-time Markov chain θ_N and P^k represents the kth-step transition probability. In terms of the original fast-changing $\theta(t)$, we see that θ_N is corresponding to $\theta(N\eta h)$. When η is small, for a fixed time t, we have $N = t/(\eta h)$. That is, for the discrete-time system, we need to look at its property for N being large enough. We call such a chain a fast switching discrete-time Markov chain. Consequently, estimation of $\theta(t)$ for small η is reduced to estimation of θ_N for large n. For the problem treated in this section, in addition to the conditions posed previously, we make the following additional assumptions.

(A13.4) The Markov chain $\{\theta_n\}$ is irreducible and aperiodic.

It is observed that under Assumptions (A13.1), (A13.2), and (A13.4), if both $\mathcal{M}$ and P are unknown, then the stationary distribution $\nu = (\nu_1, \nu_2, \ldots, \nu_{m_0})$ can be derived from P and the average w.r.t. the stationary measure can be calculated directly from

$$\overline{\theta} = \sum_{j=1}^{m_0} \nu_j \theta^{(j)}.$$

We will develop algorithms that estimate $\overline{\theta}$ without prior knowledge on P. Hence, we assume that $\mathcal{M}$ is known, but P is unknown. In this case, the goal is to identify ν from which $\overline{\theta}$ can be calculated.

13.5.1 Long-Run Average Behavior

Since $\nu_{m_0} = 1-(\nu_1+\cdots+\nu_{m_0-1})$, we need only identify m_0-1 parameters. For simplicity, we consider the observation horizon L with $L = N(m_0 - 1)$

for some positive integer N. Denote by $\mathbb{N}_0$ the following class of input signals:

$$\mathbb{N}_0 := \{u \in l^\infty : |u|_\infty \le K_u, u \text{ is } (m_0-1)-\text{periodic and full rank}\}.$$

Define the $(m_0-1)\times(m_0-1)$ matrix $\widetilde{M} = (\widetilde{m}_{ij})$, where

$$\widetilde{m}_{ij} = F(C-\phi_i'\theta^{(j)}) - F(C-\phi_i'\theta^{(m_0)}).$$

Let $\mathbb{N} := \{u \in \mathbb{N}_0\text{: } \widetilde{M} \text{ is full rank}\}$, define

$$\xi_N^{\{i\}} = \frac{1}{N}\sum_{l=0}^{N-1} s_{l(m_0-1)+i},\ i = 1,\dots,m_0-1, \tag{13.30}$$

and denote $\xi_N = (\xi_N^{\{1\}},\dots,\xi_N^{\{m_0-1\}})'$. It is easy to verify that

$$p_i = E\xi_N^{\{i\}} = \sum_{j=1}^{m_0-1}\nu_j(F(C-\phi_i'\theta^{(j)}) - F(C-\phi_i'\theta^{(m_0)})) + F(C-\phi_i'\theta^{(m_0)}).$$

Hence, $\xi_N^{\{i\}}$ represents the empirical measure of p_i. By defining

$$p = [p_1,\dots,p_{m_0-1}]',$$
$$b = [F(C-\phi_1'\theta^{(m_0)}),\dots,F(C-\phi_{m_0-1}'\theta^{(m_0)})]',$$

and

$$\widetilde{\nu} = (\nu_1,\dots,\nu_{m_0-1}) \in \mathbb{R}^{1\times(m_0-1)},$$

we obtain $p = \widetilde{M}\widetilde{\nu} + b$. This implies a relationship between p and $\widetilde{\nu}$, $\widetilde{\nu} = \widetilde{M}^{-1}(p-b)$. Since $\widetilde{M}$ and b are known from the input, this relationship implies that an estimate of $\widetilde{\nu}$ can be derived from the empirical measures of p, $\widehat{\nu}_N = \widetilde{M}^{-1}(\xi_N - b)$. From

$$\widehat{\nu}_N - \widetilde{\nu} = \widetilde{M}^{-1}(\xi_N - p),$$

the analysis of error bounds, convergence, and convergence rates of $\widehat{\nu}_N$ can be directly derived from that of ξ_N. For this reason, the remaining part of this section is devoted to the analysis of error bounds on empirical measures.

Example 13.10 The selection of inputs that will make the matrix M full rank is not difficult. For instance, suppose that the distribution is uniform with support on $[-30, 30]$. Let the threshold be $C = 10$. The system has three states: $\theta_1 = [1,3]'$, $\theta_2 = [5,-3]'$, and $\theta_3 = [10,2]'$. The input is

randomly selected to generate three regressors: $\phi_1' = [0.4565, 0.0185]$, $\phi_2' = [0.8214, 0.4447]$, and $\phi_3' = [0.6154, 0.7919]$. The M matrix becomes

$$\widetilde{M} = \begin{pmatrix} 0.6581 & 0.6296 & 0.5900 \\ 0.6307 & 0.6205 & 0.5149 \\ 0.6168 & 0.6550 & 0.5377 \end{pmatrix},$$

which is full rank.

13.5.2 Empirical Measure-Based Estimators

One immediate question is, what can one say about the asymptotic properties of the empirical measures defined above? From the well-known result of the Glivenko–Cantelli theorem ([8, p. 103]), in the usual empirical measure setup, the law of large numbers yields the convergence to the distribution function of the noise process if no switching is present. However, in the current setup, the empirical measures are coupled by a Markov chain. Intuitively, one would not doubt the existence of a limit. However, the additional random elements due to the Markov chain make the identification of the limit a nontrivial task. Corresponding to the above-mentioned law of large numbers, we first obtain the following result.

Theorem 13.11. *Under* (A13.1), (A13.2), *and* (A13.4),

$$\xi_N^{\{i\}} \to \sum_{j=1}^{m_0} \nu_j F(C - \phi_i' \theta^{(j)})$$

in probability as $N \to \infty$ *uniformly in* $i = 0, 1, 2, \ldots, m_0 - 1$.

Proof. For each $i = 0, 1, 2, \ldots, m_0 - 1$, the equalities $\phi_{lm_0+i} = \phi_i$ and $\widetilde{\phi}_{lm_0+i} = \widetilde{\phi}_i$ hold for all $l = 0, 1, \ldots, N-1$ due to the periodicity of the inputs, so

$$\begin{aligned} s_{lm_0+i} &= I_{\{y_{lm_0+i} \le C\}} = I_{\{\phi'_{lm_0+i}\theta_{lm_0+i} + d_{lm_0+i} \le C\}} \\ &= \sum_{j=1}^{m_0} I_{\{d_{lm_0+i} \le C - \phi_i'\theta^{(j)}\}} I_{\{\theta_{lm_0+i} = \theta^{(j)}\}} \\ &= \sum_{j=1}^{m_0} [I_{\{d_{lm_0+i} \le C - \phi_i'\theta^{(j)}\}} - \nu_j] I_{\{\theta_{lm_0+i} = \theta^{(j)}\}} \\ &\quad + \sum_{j=1}^{m_0} \nu_j I_{\{\theta_{lm_0+i} = \theta^{(j)}\}}. \end{aligned}$$

By virtue of the same argument as that of [122, p. 74], we have

$$E\left| \frac{1}{N} \sum_{l=0}^{N-1} I_{\{d_{lm_0+i} \le C - \phi_i'\theta^{(j)}\}} (I_{\{\theta_{lm_0+i} = \theta^{(j)}\}} - \nu_j) \right|^2 \to 0 \tag{13.31}$$

as $N \to \infty$, so

$$\frac{1}{N}\sum_{l=0}^{N-1} I_{\{d_{lm_0+i}\le C-\phi_i'\theta^{(j)}\}}(I_{\{\theta_{lm_0+i}=\theta^{(j)}\}}-\nu_j)\to 0$$

in probability and in the second moment as $N\to\infty$. Thus, it follows that

$$\xi_N^{\{i\}} = \sum_{j=1}^{m_0}\frac{1}{N}\sum_{l=0}^{N-1} I_{\{d_{lm_0+i}\le C-\phi_i'\theta^{(j)}\}}\nu_j + o(1), \tag{13.32}$$

where $o(1)\to 0$ in probability as $N\to\infty$. Note that

$$\frac{1}{N}\sum_{l=0}^{N-1} I_{\{d_{lm_0+i}\le C-\phi_i'\theta^{(j)}\}}$$

is the empirical distribution of the noise $\{d_N\}$ at $x = C-\phi_i'\theta^{(j)}$. Thus, by virtue of the well-known Glivenko–Cantelli theorem, for each $j=1,\dots,m_0$ and $i=0,1,\dots,m_0-1$,

$$\frac{1}{N}\sum_{l=0}^{N-1} I_{\{d_{lm_0+i}\le C-\phi_i'\theta^{(j)}\}}\to F(C-\phi'(t_0+i)\theta^{(j)}) \quad \text{as} \quad N\to\infty.$$

Thus, the desired result follows from the familiar Slutsky's result. □

The above result may be considered as the first approximation of the empirical measures to the weighted average of the distribution functions. Naturally, one would also like to know how fast the convergence will take place. This is presented in Theorem 13.12, which entails the study of the asymptotics of a centered and scaled sequence of errors or deviations. Compared with the results with a fixed parameter, it can be viewed as a hybrid coupling of discrete events with the normalized deviations. Conceptually, one expects that a rescaled sequence of the empirical measures should converge to a Brownian bridge suitably combined or coupled owing to the Markov chain in the original observation. In view of the above law of large numbers for empirical measures, one expects that the weak limit of the rescaled sequence should also be suitably combined by the stationary distributions of the Markov chain. Thus, it is not difficult to guess the limit. However, verifying this limit is not at all trivial. To illustrate, if we have two sequences $X_N^{\{1\}}$ and $X_N^{\{2\}}$ satisfying $X_N^{\{i\}}\to X^{\{i\}}$, $i=1,2$, in distribution as $N\to\infty$, we cannot conclude $X_N^{\{1\}}+X_N^{\{2\}}\to X^{\{1\}}+X^{\{2\}}$ in distribution generally. In our case, the difficulties are incurred by the presence of the Markov chain. In the proof of Theorem 13.12, we overcome the difficulty by establishing several claims. We first derive its asymptotic equivalence by bringing out the important part and discarding the asymptotically negligible part. We may call this step the decorrelation step. Next, we consider

a suitably scaled sequence with a fixed-θ process. That is, we replace the "random jump" process with a fixed value. This replacement enables us to utilize a known result on the empirical process with a fixed parameter. The third step is to use finite-dimensional distributions convergence due to weak convergence of the empirical measures to treat an m-tuple

$$\left(\eta_N^{\{1\}}(\theta^{(1)}), \ldots, \eta_N^{\{2\}}(\theta^{(m_0)})\right)$$

[the definition of $\eta_N^{\{i\}}(\theta)$ is given in (13.34) in what follows]. Finally, we use a Wold's device [8, p. 52] to finish the proof. The proof itself is interesting in its own right.

Theorem 13.12. *Assume the conditions of Theorem* 13.11. *The sequence*

$$\sqrt{N}\ \left[\xi_N^{\{i\}} - \sum_{j=1}^{m_0} \nu_j F(C - \phi_i' \theta^{(j)})\right]$$

converges weakly to

$$\sum_{j=1}^{m_0} \nu_j B(C - \phi_i' \theta^{(j)}),$$

where $B(\cdot)$ *is a Brownian bridge process such that the covariance of* $B(\cdot)$ *(for* $x_1,\ x_2 \in \mathbb{R}$*) is given by*

$$\begin{aligned} &EB(x_1)B(x_2) \\ &\quad = \min(F(x_1), F(x_2)) - F(x_1)F(x_2). \end{aligned}$$

Proof. Step 1: Asymptotic equivalence: By virtue of [122, p. 74], similarly to (13.31), we can show that for each $j = 1, \ldots, m_0$ and $i = 0, 1, \ldots, m_0 - 1$,

$$\begin{aligned} &\frac{1}{N} \sum_{l=0}^{N-1} I_{\{d_{lm_0+i} \le C - \phi_i' \theta^{(j)}\}} I_{\{\theta_{lm_0+i} = \theta^{(j)}\}} \\ &\quad = \frac{1}{N} \sum_{l=0}^{N-1} I_{\{d_{lm_0+i} \le C - \phi_i' \theta^{(j)}\}} \nu_j + o(1), \end{aligned}$$

where $o(1) \to 0$ in probability (and also in the second moment) as $N \to \infty$. This together with (13.32) leads to

$$\begin{aligned} &\sqrt{N} \left[\xi_N^{\{i\}} - \sum_{j=1}^{m_0} \nu_j F(C - \phi_i' \theta^{(j)})\right] \\ &= \sum_{j=1}^{m_0} \frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} [I_{\{d_{lm_0+i} \le C - \phi_i' \theta^{(j)}\}} - F(C - \phi_i' \theta^{(j)})] \nu_j + o(1), \end{aligned} \tag{13.33}$$

where $o(1) \to 0$ in probability as $N \to \infty$.

Step 2: Convergence in distribution of a fixed-θ process: Consider now a typical term in the last line of (13.33). For convenience, for a fixed θ, define

$$\eta_N^{\{i\}}(\theta) = \frac{1}{\sqrt{N}} \sum_{l=0}^{N-1} \Big[I_{\{d_{lm_0+i} \le C - \phi_i'\theta\}} - F(C - \phi_i'\theta) \Big]. \tag{13.34}$$

It is readily seen that $\eta_N^{\{i\}}(\theta)$ is a centered empirical measure (with a fixed θ) re-scaled by $\sqrt{N}$. The results on empirical measures (see [8, p. 141], [76], and also [84]) then imply that $\eta_N^{\{i\}}(\cdot)$ converges weakly to a Brownian bridge process $B(C - \phi_i' \cdot)$ with mean 0 and covariance

$$\begin{aligned} &EB(C - \phi_i'\theta_1)B(C - \phi_i'\theta_2) \\ &\quad = \min(F(C - \phi_i'\theta_1), F(C - \phi_i'\theta_2)) - F(C - \phi_i'\theta_1)F(C - \phi_i'\theta_2). \end{aligned}$$

Step 3: Convergence of finite-dimensional distributions: Since $\eta_N^{\{i\}}(\cdot)$ converges weakly to $B(C - \phi_i' \cdot)$, its finite-dimensional distributions converge. That is, for any integer p and any $(x_1, \ldots, x_p)$, $(\eta_N^{\{i\}}(x_1), \ldots, \eta_N^{\{i\}}(x_p))$ converges in distribution to $(B(C - \phi_i' x_1), \ldots, B(C - \phi_i' x_p))$. In particular, we have that $(\eta_N^{\{i\}}(\theta^{(1)}), \ldots, \eta_N^{\{i\}}(\theta^{(p)}))$ converges in distribution to $(B(C - \phi_i'\theta^{(1)}), \ldots, B(C - \phi_i'\theta^{(p)}))$.

Step 4: The weak convergence and the form of the finite-dimensional distributional convergence in Step 3 imply that

$$(\nu_1, \ldots, \nu_{m_0})' \Big(\eta_N^{\{i\}}(\theta^{(1)}), \ldots, \eta_N^{\{i\}}(\theta^{(m_0)}) \Big)$$

converges in distribution to $\sum_{j=1}^{m_0} \nu_j B(C - \phi_i'\theta^{(j)})$ by Wold's device [8, p. 52]. Finally, putting all the steps together, the desired result follows. □

Note that a Brownian bridge is a Brownian motion tied down at both ends. Here we emphasize that the process considered is allowed to take values not just in $[0, 1]$, but in the entire real line; thus, what we have is a "stretched" Brownian bridge as discussed in Chapter 3; see also [76]. Similarly to Lemma 13.5, the next lemma provides a strong approximation result for empirical measures. Its detailed proof is omitted for brevity.

Lemma 13.13 *Under the conditions of Theorem* 13.4*, there is a constant* $\gamma > 0$ *such that*

$$\begin{aligned} &\sup_{0 \le i \le m_0 - 1} \left| \sqrt{N} \left[\xi_N^{\{i\}} - \sum_{j=1}^{m_0} \nu_j F(C - \phi_i'\theta^{(j)}) \right] - \sum_{j=1}^{m_0} \nu_j B(C - \phi_i'\theta^{(j)}) \right| \\ &\quad = o(N^{-\gamma}) \ \textit{w.p.1.} \end{aligned}$$

13.5.3 Estimation Errors on Empirical Measures: Upper and Lower Bounds

In the context of system identification, estimation error bounds are of crucial importance. This section obtains such bounds for the fast-varying systems. As a preparation, we first present a lemma, which is an exponential estimate for a Gaussian process.

Lemma 13.14 *Under the assumptions of Theorem* 13.11*, for N sufficiently large and for each $j = 1, \ldots, m_0$,*

$$P\left(\left|\frac{1}{\sqrt{N}}B(C-\phi_i'\theta^{(j)})\right| \geq \frac{\varepsilon}{m_0 M}\right) \leq 2\exp\left(-\frac{2N\varepsilon^2}{{m_0}^2 M^2}\right), \qquad (13.35)$$

where $M = \max\{\nu_1, i \leq m_0\}$ and $B(\cdot)$ is given by Theorem 13.12*.*

Proof. Let $\sigma_{ij}^2 = \mathrm{Var}(B(C-\phi_i'\theta^{(j)}))$. By direct computation, one can show that for any $\alpha > 0$,

$$E\exp(\alpha|\frac{1}{\sqrt{N}}B(C-\phi_i'\theta^{(j)})|) \leq 2\exp(\frac{\alpha^2\sigma_{ij}^2}{2N}).$$

Thus,

$$\begin{aligned} &P\left(\alpha\left|\frac{1}{\sqrt{N}}B(C-\phi_i'\theta^{(j)})\right| \geq \frac{\alpha\varepsilon}{m_0 M}\right) \\ &\quad\leq \exp\left(-\frac{\alpha\varepsilon}{m_0 M}\right) E\exp\left(\alpha\left|\frac{1}{\sqrt{N}}B(C-\phi_i'\theta^{(j)})\right|\right) \\ &\quad\leq 2\exp\left(\frac{\sigma_{ij}^2}{2N}\alpha^2 - \frac{\varepsilon}{m_0 M}\alpha\right). \end{aligned}$$

Choose $\alpha = N_\varepsilon/(m_0 M\sigma_{ij}^2)$ to minimize the quadratic form in the exponent above and note that

$$\sigma_{ij}^2 = F(C-\phi_i'\theta^{(j)})(1-F(C-\phi_i'\theta^{(j)})) \leq \frac{1}{4}.$$

Then the upper bound is obtained. □

It can be seen that Lemma 13.14 derives an exponential type of upper bound on the estimation errors. To some extent, it is a large deviations result. Note that $B(\cdot)$, the Brownian bridge process, is a Gaussian process. The deviation given above indicates that the "tail" probabilities of deviations of the order $O(\sqrt{N})$ will be exponentially small. With this lemma, we can proceed to obtain the "large deviations" of the empirical measures from those of the averaged distribution functions (average with respect to the stationary distributions of the Markov chain).

Theorem 13.15. *Under the assumptions of Theorem* 13.11*, for N large enough and for any $\varepsilon > 0$,*

$$P\left(\left|\xi_N^{\{i\}} - \sum_{j=1}^{m_0} \nu_j F(C - \phi_i' \theta^{(j)})\right| \geq \varepsilon\right) \leq 2m_0 \exp\left(-\frac{2N\varepsilon^2}{{m_0}^2 M^2}\right). \tag{13.36}$$

Proof. It is easy to see that

$$P\left(\left|\frac{1}{\sqrt{N}} \sum_{j=1}^{m_0} \nu_j B(C - \phi_i' \theta^{(j)})\right| \geq \varepsilon\right) \leq P\left(\sum_{j=1}^{m_0} \left|\frac{1}{\sqrt{N}} B(C - \phi_i' \theta^{(j)})\right| \geq \frac{\varepsilon}{M}\right).$$

Recall that $\widetilde{c}_{ij}$ is given in Theorem 13.16. Observe that

$$\begin{aligned}
&\left\{(\tilde{c}_{i1}, \ldots, \tilde{c}_{im_0})' : \sum_{j=1}^{m_0} \left|\frac{1}{\sqrt{N}} B(\tilde{c}_{ij})\right| \geq \frac{\varepsilon}{M}\right\} \\
&\quad \subseteq \left\{(\tilde{c}_{i1}, \ldots, \tilde{c}_{im_0})' : \left|\frac{1}{\sqrt{N}} B(\tilde{c}_{ij})\right| \geq \frac{\varepsilon}{m_0 M} \text{ for some } j\right\} \\
&\quad \subseteq \bigcup_{j=1}^{m_0} \left\{(\tilde{c}_{i1}, \ldots, \tilde{c}_{im})' : \left|\frac{1}{\sqrt{k}} B(\tilde{c}_{ij})\right| \geq \frac{\varepsilon}{m_0 M}\right\}.
\end{aligned}$$

Thus, by (13.36),

$$\begin{aligned}
&P\left(\left|\frac{1}{\sqrt{N}} \sum_{j=1}^{m_0} \nu_j B(C - \phi_i' \theta^{(j)})\right| \geq \varepsilon\right) \\
&\quad \leq \sum_{j=1}^{m_0} P\left(\left|\frac{1}{\sqrt{N}} B(\tilde{c}_{ij})\right| \geq \frac{\varepsilon}{m_0 M}\right) \leq 2m_0 \exp\left(-\frac{2N\varepsilon^2}{{m_0}^2 M^2}\right).
\end{aligned}$$

For sufficiently large N, the desired result follows from Lemma 13.13 and the same kind of argument as in the proof of Theorem 13.6. □

Next, we proceed to obtain lower bounds on the estimation error when full-rank periodic inputs are used.

Theorem 13.16. *Denote*

$$\widetilde{c}_{ij} = C - \phi_i' \theta^{(j)}$$

for each $i = 0, \ldots, m_0 - 1$ and $j = 1, \ldots, m_0$, and denote the matrix

$$\Sigma = (\tilde{\sigma}_{i_1, i_2} : i_1, i_2 = 1, \ldots, m_0),$$

where for $i_1, i_2 = 1, \ldots, m_0$,

$$\tilde{\sigma}_{i_1,i_2} = \min(F(\tilde{c}_{ii_1}), F(\tilde{c}_{ii_2})) - F(\tilde{c}_{ii_1})F(\tilde{c}_{ii_2}).$$

Let λ *be the minimum eigenvalue of the covariance matrix* Σ. *Under the assumptions of Theorem* 13.11, *for sufficiently large* N *and for any* $\varepsilon > 0$,

$$\begin{aligned} &P\left(\left|\xi_N^{\{i\}} - \sum_{j=1}^{m_0} \nu_j F(C - \phi_i' \theta^{(j)})\right| \geq \varepsilon\right) \\ &\quad \geq \sqrt{\frac{2}{\pi}} \left(\frac{\sqrt{\lambda}|\nu|}{\sqrt{N}\varepsilon} - \left(\frac{\sqrt{\lambda}|\nu|}{\sqrt{N}\varepsilon}\right)^3\right) \exp\left(-\frac{\varepsilon^2}{2\lambda|\nu|^2} N\right). \end{aligned} \tag{13.37}$$

Proof. Note that $(B(\tilde{c}_{i1}), \ldots, B(\tilde{c}_{im_0}))$ can be regarded as a multinormal distributed vector with mean 0 and covariance Σ. Recall that

$$\nu = (\nu_1, \ldots, \nu_{m_0}) \quad \text{and} \quad \check{S} = \sum_{j=1}^{m_0} \nu_j B(\tilde{c}_{ij}).$$

Then $\check{S}$ is a one-dimensional Gaussian random variable with mean 0 and variance $\nu' \Sigma \nu$. Direct computation yields that

$$P\left(\left|\frac{1}{\sqrt{N}} \sum_{j=1}^{m_0} \nu_j B(\tilde{c}_{ij})\right| \geq \varepsilon\right) = P\left(|Z| \geq \frac{\sqrt{N}\varepsilon}{\sqrt{\nu' \Sigma \nu}}\right),$$

where $Z = \check{S}/(\sqrt{\nu' \Sigma \nu})$. Then

$$\begin{aligned} &P\left(\left|\frac{1}{\sqrt{N}} \sum_{j=1}^{m_0} \nu_j B(\tilde{c}_{ij})\right| \geq \varepsilon\right) \geq 2P\left(Z \geq \frac{\sqrt{N}\varepsilon}{\sqrt{\lambda}|\nu|}\right) \\ &= 2\left(1 - \Phi\left(\frac{\sqrt{N}\varepsilon}{\sqrt{\lambda}|\nu|}\right)\right) \geq 2\left(\frac{\sqrt{\lambda}|\nu|}{\sqrt{N}\varepsilon} - \left(\frac{\sqrt{\lambda}|\nu|}{\sqrt{N}\varepsilon}\right)^3\right) \check{\varphi}\left(\frac{\sqrt{N}\varepsilon}{\sqrt{\lambda}|\nu|}\right), \end{aligned}$$

where $\Phi(\cdot)$ and $\check{\varphi}(\cdot)$ are the cumulative distribution and density function of the standard normal variable, respectively. Thus, for sufficiently large N,

$$\begin{aligned} &P\left(\left|\xi_N^{\{i\}} - \sum_{j=1}^{m_0} \nu_j F(C - \phi_i' \theta^{(j)})\right| \geq \varepsilon\right) \\ &\quad \geq \sqrt{\frac{2}{\pi}} \left(\frac{\sqrt{\lambda}|\nu|}{\sqrt{N}\varepsilon} - \left(\frac{\sqrt{\lambda}|\nu|}{\sqrt{N}\varepsilon}\right)^3\right) \exp\left(-\frac{\varepsilon^2}{2\lambda|\nu|^2} N\right). \end{aligned}$$

The proof is concluded. □

13.6 Notes

Although the problems considered in this chapter are centered around switching systems with binary observations, the main ideas and results can be generalized to quantized observations. Regime-switching systems often appear as integrated parts of hybrid systems, discrete-event systems, logic-based systems, finite automata, hierarchical systems, and complex systems. Consequently, our results may have potential applications in these areas as well. More information on regime-switching systems can be found in [123] and its references.

The framework here is based on our recent work [120] for tracking Markovian parameters with binary-valued observations. Several directions may be pursued. The inclusion of unmodeled dynamics is a worthwhile research direction. Quantized sensors may be treated. Optimal sensor placement in conjunction with the filters developed in this chapter may be considered. Optimal selection of the threshold values and sensor locations is an important issue. Complexity is another direction for further investigation.

Part V

Complexity Analysis

14

Space and Time Complexities, Threshold Selection, Adaptation

The number m_0 of thresholds is a measure of space complexity, whereas the observation length N is a measure of time complexity that quantifies how fast uncertainty can be reduced. The significance of understanding space and time complexities can be illustrated by the following example. For computer information processing of a continuous-time system, its output must be sampled (e.g., with a sampling rate N Hz) and quantized (e.g., with a precision word-length of B bits). Consequently, its output observations carry the data-flow rate of NB bits per second (bps). For instance, for 8-bit precision and a 10-KHz sampling rate, an $80K$-bps bandwidth of data transmission resource is required. In a sensor network in which a large number of sensors must communicate within the network, such resource demand is overwhelming especially when wireless communications of data are involved.

The problem is generic since any computerized information processing for analog signals will inherently encounter the problem of data precision and sampling rates. However, this problem is not acute when data-flow bandwidths are not limited, such as in wired systems with fast computers. New technology developments in smart sensors, MEMS (micro electromechanical systems), sensor networks, computer communication systems, wireless systems, mobile agents, distributed systems, and remote-controlled systems have ushered in new paradigms in which data-flow rates carry significant costs and limitations.

Conceptually, it is well understood that increasing precision levels is desirable for enhancing accuracy in information processing. Similarly, increasing data size can be potentially useful for reducing identification errors.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_14,

However, these will jointly demand more resources. A fundamental question must be answered: Is such resource demand necessary for achieving a required identification requirement? To answer this question, a framework is required that can facilitate the analysis of both time complexity (such as the sampling rate) and space complexity (such as the number of subsets for output partition). This chapter addresses the following key issues:

(1) What are the main benefits in increasing the space complexity defined by the number of output observation subsets, in terms of identification accuracy?

(2) What is the relationship between the space complexity (measurement precision) and the time complexity (speed of uncertainty reduction)?

(3) How should the output range be partitioned for the best identification accuracy?

(4) What is the optimal resource allocation when communication channels provide only limited bandwidths?

Section 14.1 defines the concepts of space complexity and time complexity and derives some basic properties of space complexity. Information contents of the observed data can be improved by input and threshold design, which are covered in Sections 14.2, 14.3, and 14.4 with binary-valued observations. It starts in Section 14.2 with a feasibility study on worst-case threshold selection, followed by robust threshold selection in Section 14.3. Adaptive algorithms for thresholds are presented in Section 14.4. Section 14.5 generalizes the results of 14.2 to quantized observations. Based on these results, Section 14.6 discusses the utility of relationships between space and time complexities in network design.

14.1 Space and Time Complexities

In an information processing problem that involves computer networks, or communication data transmission, or wireless networks, the required resource is usually represented by bandwidths in bits per second. In our identification problems, if identification must be accomplished in τ seconds to facilitate subsequent tasks (control, prediction, diagnosis, etc.), then the time complexity N is translated into N samples per τ seconds. Correspondingly, the required bandwidth will be $R = N \log(m_0+1)$ bits per τ seconds, or R/τ bps. Throughout the chapter, the simplified notation $\log = \log_2$ is used. Since τ is an external constant, we shall simply view R as the required bandwidth on communication channels. For an available total resources R, one may choose to assign more resources to space complexity (increasing m_0) or to time complexity (increasing N, i.e., the rate of data acquisition). The overall goal is to achieve the best uncertainty reduction for a

given resource, or to achieve minimum resource utilization for a given level of uncertainty reduction. This section presents basic properties of space complexity.

Separation of Time and Space Complexities

To understand the impact of sensor threshold values and the number of thresholds on identification error variance, we use $\sigma^2_{\rm CR}(N, m_0, \theta)$ in Lemma 6.6. We will interpret m_0 exchangeably as space complexity, the number of binary-valued sensors, or the number of sensor thresholds. Let us first fix an integer m_0. $\sigma^2_{\rm CR}(N, m_0, \theta)$ indicates a basic relationship that delineates a fundamental property of asymptotic separation of space and time complexity in variance reduction.

From the Cramér–Rao (CR) lower bound in Chapter 6, we have the following conclusion.

Corollary 14.1.

$$\eta(m_0, \theta) := N\sigma^2_{\rm CR}(N, m_0, \theta) = \left(\sum_{i=1}^{m_0+1} \frac{\widetilde{h}_i^2}{\widetilde{p}_i} \right)^{-1} \tag{14.1}$$

is independent of N.

Remark 14.2. Corollary 14.1 shows that asymptotically the optimal variance $\sigma^2_{\rm CR}(N, m_0, \theta)$ is reduced in the rate of $1/N$ in terms of its time complexity. Its reduction by increasing space complexity m_0 is characterized entirely by $\eta(m_0, \theta)$, which is independent of N. This separation of space and time complexity in their capability for identification error reduction provides a convenient foundation for complexity analysis. Consequently, $\eta(m_0, \theta)$ will be used for the analysis of space complexity.

Monotonicity of Space Complexity

To proceed, we make a couple of definitions first. Suppose that $[y_{\min}, y_{\max}]$ is the range of y_k. A placement of m_0 sensor thresholds is a partition $y_{\min} < C_1 < \cdots < C_{m_0} < y_{\max}$ of the interval $[y_{\min}, y_{\max}]$. In what follows, we also use the notation

$$\mathcal{T}_{m_0} = \{y_{\min}, C_1, \ldots, C_{m_0}, y_{\max}\}$$

to denote the set of points of the partition.

Definition 14.3. Suppose $m_1 < m_2$ are two positive integers, and $\mathcal{T}_{m_1} = \{y_{\min}, C^1_1, \ldots, C^1_{m_1}, y_{\max}\}$ and $\mathcal{T}_{m_2} = \{y_{\min}, C^2_1, \ldots, C^2_{m_2}, y_{\max}\}$ are two placements of sensor thresholds. We say that $\mathcal{T}_{m_2}$ is a *refinement* of $\mathcal{T}_{m_1}$ if

$$\{y_{\min}, C^1_1, \ldots, C^1_{m_1}, y_{\max}\} \text{ is a subset of } \{y_{\min}, C^2_1, \ldots, C^2_{m_2}, y_{\max}\}.$$

Remark 14.4. In the definition of placement of sensors, $[y_{\min}, y_{\max}]$ can be either finite or infinite. In case one of these values is ∞ or $-\infty$, it is understood that we work with the extended real number system. For practical utility, we have assumed that no sensor is placed at either $y_{\min}$ or $y_{\max}$; otherwise, they do not provide any useful information for system identification.

Note that the statement "$\mathcal{T}_{m_2}$ is a refinement of $\mathcal{T}_{m_1}$" means that $\mathcal{T}_{m_2}$ can be obtained by starting with the threshold points $C_1^1 < \cdots < C_{m_1}^1$ and interposing $m_2 - m_1$ points between them to form a finer subdivision.

Theorem 14.5. *Suppose that $\mathcal{T}_{m_1}$ and $\mathcal{T}_{m_2}$ are two placements of sensor thresholds such that $\mathcal{T}_{m_2}$ is a refinement of $\mathcal{T}_{m_1}$. Then $\eta(m_0, \theta)$ given in* (14.1) *satisfies*

$$\eta(m_2, \theta) \le \eta(m_1, \theta).$$

Remark 14.6. Theorem 14.1 is a result about the monotonicity of space complexity. It indicates that a reduction in error variance is achieved by increasing space complexity.

Proof. The extra thresholds in $\mathcal{T}_{m_2}$ that are not in $\mathcal{T}_{m_1}$ can be added one at a time. Hence, we need only show that if one additional threshold C is added to $\mathcal{T}_{m_1}$, we have $\eta(m_1 + 1, \theta) \le \eta(m_1, \theta)$.

Suppose that one additional threshold C is inserted in $\mathcal{T}_{m_1}$, between $C_1, C_2 \in \mathcal{T}_{m_1}$, $C_1 < C < C_2$. Denote

$$p = P\{C_1 < x \le C_2\},$$
$$p_1 = P\{C_1 < x \le C\},$$
$$p_2 = P\{C < x \le C_2\},$$

and

$$h = \frac{\partial p}{\partial \theta}, \quad h_1 = \frac{\partial p_1}{\partial \theta}, \quad h_2 = \frac{\partial p_2}{\partial \theta}.$$

By (14.1), we need only show that

$$\frac{h_1^2}{p_1} + \frac{h_2^2}{p_2} \ge \frac{h^2}{p}. \tag{14.2}$$

Note that $p = p_1 + p_2$ and $h = h_1 + h_2$. Hence, (14.2) can be expressed as

$$\frac{h_1^2}{p_1} + \frac{(h - h_1)^2}{p - p_1} - \frac{h^2}{p} \ge 0.$$

However, elementary derivations show that

$$\frac{h_1^2}{p_1} + \frac{(h - h_1)^2}{p - p_1} - \frac{h^2}{p} = \frac{(h_1 p - p_1 h)^2}{p_1 (p - p_1) p} \ge 0.$$

□

14.2 Binary Sensor Threshold Selection and Input Design: Feasibility Analysis

We now consider the problem of threshold selection and input design. An interval $\mathcal{I}_i = (C_{i-1}, C_i]$ of the output range can provide useful information for system identification only when $\widetilde{h}_i \neq 0$. The contribution of a sensor interval to error reduction depends on the actual parameter θ, the distribution function F, the thresholds, and the input.

Example 14.7. To illustrate, suppose that $F(\cdot)$ is the distribution function of a uniform random variable on $[0, 10]$, namely, $F(x) = x/10$, $0 \leq x \leq 10$. The prior information on θ is that $\theta \in [2, 5]$. If one selects $u_k \equiv u_0 = 1$, and places four sensor thresholds at $C_1 = 1$, $C_2 = 6$, $C_3 = 10$, and $C_4 = 20$, then it can be verified that $F(C_1 - \theta) = 0$, $F(C_2 - \theta) = (6 - \theta)/10$, $F(C_3 - \theta) = (10 - \theta)/10$, and $F(C_4 - \theta) = 1$. These imply

$$\widetilde{p}_1 = 0,\ \widetilde{p}_2 = (6 - \theta)/10,\ \widetilde{p}_3 = 0.4,\ \widetilde{p}_4 = \theta/10.$$

Since $\widetilde{p}_1$ and $\widetilde{p}_3$ do not depend on θ, the intervals $(-\infty, C_1]$ and $(C_2, C_3]$ do not provide information about θ.

Even when the threshold does provide information, the selection of the threshold value will have a significant impact on the convergence speed.

Example 14.8. Suppose that $F(\cdot)$ is the distribution function of a zero-mean Gaussian random variable with variance $\sigma^2 = 625$. The true parameter is $\theta = 100$ and $u_k \equiv 1$. In this case, Corollary 14.1 becomes, adding C in notation,

$$\eta_C(1, 100) = N\sigma^2_{\text{CR}}(1, N, \theta) = \left(\frac{\widetilde{h}_1^2}{\widetilde{p}_1} + \frac{\widetilde{h}_2^2}{\widetilde{p}_2}\right)^{-1} = \frac{F(C-\theta)(1 - F(C-\theta))}{f^2(C-\theta)}.$$

For $C = 20, 50, 80, 100$, we obtain the values $\eta_{20}(1, 100) = 75506$, $\eta_{50}(1, 100) = 4767$, $\eta_{80}(1, 100) = 1244$, and $\eta_{100}(1, 100) = 982$. In fact, it is easy to verify that the optimal threshold choice is $C = \theta$ in this case.

In an identification problem, the parameter θ is unknown. Hence, one must work with the prior uncertainty set on θ. While in most applications, $\theta \in [\theta_{\min}, \theta_{\max}]$ is the typical prior information, we shall use the general prior uncertainty set $\theta \in \Omega$ to include other possibilities. We first concentrate on a binary-valued sensor of threshold C. Similar conclusions will later be derived for general quantized sensors. It is assumed that C and u_0 can be either selected prior to an identification experiment, or tuned during it. Let $\widetilde{p}(\theta) = F(C - \theta u_0)$ and $\widetilde{h}(\theta) = \partial\widetilde{p}(\theta)/\partial\theta = -f(C - \theta u_0)u_0$.

Definition 14.9. An interval $\mathcal{I} = (-\infty, C]$ or a threshold C is said to be (1) *feasible* for θ if the corresponding $\widetilde{h}(\theta) \neq 0$; (2) *robustly feasible* for Ω if the corresponding $\widetilde{h}(\theta) \neq 0$ for all $\theta \in \Omega$.

For a given Ω, the set of all feasible thresholds will be denoted by Γ_Ω. We will derive concretely Γ_Ω for some typical cases.

Theorem 14.10. *Suppose that the prior information on the unknown parameter is $\theta \in \Omega_0 = [\theta_{\min}, \theta_{\max}]$, and the disturbance d_k is zero mean and its density function has support (i.e., strictly positive) in $(-\delta, \delta)$. For $u_0 > 0$, there exists a robustly feasible C if and only if*

$$\delta > \frac{(\theta_{\max} - \theta_{\min})u_0}{2}. \tag{14.3}$$

Under (14.3)*, the set of robustly feasible thresholds is*

$$\Gamma_{\Omega_0} = \{C : \theta_{\max} u_0 - \delta < C < \theta_{\min} u_0 + \delta\}.$$

Proof. Under the hypothesis,

$$p = F(C - \theta u_0) - F(-\infty) = F(C - \theta u_0).$$

Then $h = \partial p / \partial \theta = -f(C - \theta u_0) u_0 \neq 0$ if and only if $-\delta < C - \theta u_0 < \delta$. Hence, for a given θ, C is feasible if and only if $C \in (\theta u_0 - \delta, \theta u_0 + \delta)$. The intersection of all such thresholds

$$\bigcap_{\theta \in \Omega_0} (\theta u_0 - \delta, \theta u_0 + \delta)$$

is nonempty if and only if

$$\theta_{\max} u_0 - \delta < \theta_{\min} u_0 + \delta,$$

or equivalently,

$$\delta > \frac{(\theta_{\max} - \theta_{\min})u_0}{2}.$$

Note that under condition (14.3),

$$(\theta_{\max} u_0 - \delta, \theta_{\min} u_0 + \delta) = \bigcap_{\theta \in [\theta_{\min}, \theta_{\max}]} (\theta u_0 - \delta, \theta u_0 + \delta).$$

Hence, for any $C \in (\theta_{\max} u_0 - \delta, \theta_{\min} u_0 + \delta)$, $C \in (\theta u_0 - \delta, \theta u_0 + \delta)$ for all $\theta \in \Omega_0$. This proves that it is robustly feasible for Ω_0. □

Note that (14.3) can be rewritten as

$$u_0 < \frac{2\delta}{\theta_{\max} - \theta_{\min}}, \tag{14.4}$$

which defines the maximum input value.

14.3 Worst-Case Optimal Threshold Design

For a given prior uncertainty set Ω of the unknown parameter θ, the set Γ_Ω of robustly feasible thresholds can be used to select thresholds to reduce identification errors. Corollary 14.1 provides the main vehicle for this pursuit.

Observe that

$$\widetilde{p}_1 = F(C-\theta u_0), \quad \widetilde{p}_2 = 1 - F(C-\theta u_0),$$
$$\widetilde{h}_1 = -f(C-\theta u_0)u_0, \quad \widetilde{h}_2 = f(C-\theta u_0)u_0.$$

By adding the dependence on C and u_0 in the notation, Corollary 14.1 reduces to

$$\begin{aligned}\eta_{C,u_0}(1,\theta) &= N\sigma^2_{\rm CR}(1,N,\theta) = \left(\frac{\widetilde{h}_1^2}{\widetilde{p}_1} + \frac{\widetilde{h}_2^2}{\widetilde{p}_2}\right)^{-1} \\ &= \frac{F(C-\theta u_0)(1-F(C-\theta u_0))}{f^2(C-\theta u_0)u_0^2}.\end{aligned}$$

If $\theta \in \Omega$ and $C \in \Gamma_\Omega$, then $f(C-\theta u_0) \neq 0$. Therefore, $\eta_{C,u_0}(1,\theta)$ is well defined. The optimal worst-case design is achieved by solving the min-max problem

$$\eta^* = \inf_{C\in\Gamma_\Omega, u_0>0} \sup_{\theta\in\Omega} \eta_{C,u_0}(1,\theta).$$

We shall solve this problem more concretely under some typical situations.

Bounded Disturbances

Suppose that the prior information on the unknown parameter is $\theta \in \Omega_0 = [\theta_{\min}, \theta_{\max}]$, and the disturbance d_k is zero mean and its density function has support in $(-\delta, \delta)$. By Theorem 14.10, the set of robustly feasible thresholds is $\Gamma_{\Omega_0} = (\theta_{\max}u_0 - \delta, \theta_{\min}u_0 + \delta)$. The optimal threshold and input selection is obtained by solving the following min-max optimization problem:

$$\eta^* = \inf_{C\in(\theta_{\max}u_0-\delta,\theta_{\min}u_0+\delta),u_0>0} \max_{\theta\in[\theta_{\min},\theta_{\max}]} \frac{F(C-\theta u_0)(1-F(C-\theta u_0))}{f^2(C-\theta u_0)u_0^2}. \tag{14.5}$$

Theorem 14.11. *Suppose that d_k is uniformly distributed with density function $f(x) = 1/(2\delta)$ for $x \in (-\delta, \delta)$.*

(1) *If*

$$(\theta_{\max} - \theta_{\min})u_0/2 < \delta \leq (\theta_{\max} - \theta_{\min})u_0,$$

then

$$\eta^* = \frac{(\theta_{\max} - \theta_{\min})^2}{4} \tag{14.6}$$

and any $C \in (\theta_{\max}u_0 - \delta, \theta_{\min}u_0 + \delta)$ is optimal in the worst-case sense over $\theta \in [\theta_{\min}, \theta_{\max}]$.

(2) *If* $\delta > (\theta_{\max} - \theta_{\min})u_0$, *then*

$$\eta^* = (\theta_{\max} - \theta_{\min})^2. \tag{14.7}$$

Proof. For any $C \in (\theta_{\max}u_0 - \delta, \theta_{\min}u_0 + \delta)$ and $\theta \in [\theta_{\min}, \theta_{\max}]$, we have

$$C - \theta u_0 < \theta_{\min}u_0 + \delta - \theta_{\min}u_0 = \delta$$

and

$$C - \theta u_0 > \theta_{\max}u_0 - \delta - \theta_{\max}u_0 = -\delta.$$

It follows that in (14.5), $f(C-\theta u_0) = 1/(2\delta)$ and $F(C-\theta u_0) = (C-\theta u_0+\delta)/(2\delta)$. As a result,

$$\begin{aligned}\frac{F(C-\theta u_0)(1-F(C-\theta u_0))}{f^2(C-\theta)u_0^2} &= (2\delta)^2\frac{(C-\theta u_0+\delta)}{2\delta u_0^2}\left(1-\frac{C-\theta u_0+\delta}{2\delta}\right)\\ &= \frac{\delta^2-(C-\theta u_0)^2}{u_0^2}.\end{aligned}$$

(1) If

$$(\theta_{\max} - \theta_{\min})u_0/2 < \delta \le (\theta_{\max} - \theta_{\min})u_0,$$

then

$$\theta_{\min}u_0 \le \theta_{\max}u_0 - \delta, \quad \theta_{\min}u_0 + \delta \le \theta_{\max}u_0.$$

This implies that for any $C \in (\theta_{\max}u_0 - \delta, \theta_{\min}u_0 + \delta)$, there exists $\theta \in [\theta_{\min}, \theta_{\max}]$ such that $\theta u_0 = C$. Consequently, for any given $C \in (\theta_{\max}u_0 - \delta, \theta_{\min}u_0 + \delta)$,

$$\max_{\theta\in[\theta_{\min},\theta_{\max}]}\frac{\delta^2-(C-\theta u_0)^2}{u_0^2} = \frac{\delta^2}{u_0^2}.$$

It follows from this and (14.4) that

$$\begin{aligned}\eta^* &= \inf_{C\in(\theta_{\max}u_0-\delta,\theta_{\min}u_0+\delta),u_0>0}\ \max_{\theta\in[\theta_{\min},\theta_{\max}]}\frac{F(C-\theta u_0)(1-F(C-\theta u_0))}{f^2(C-\theta u_0)u_0^2}\\ &= \inf_{u_0<2\delta/(\theta_{\max}-\theta_{\min})}\frac{\delta^2}{u_0^2}\\ &= \frac{\delta^2}{4\delta^2/(\theta_{\max}-\theta_{\min})^2}\\ &= \frac{(\theta_{\max}-\theta_{\min})^2}{4}.\end{aligned}$$

This proves (14.6).

(2) If $\delta > (\theta_{\max} - \theta_{\min})u_0$, then

$$u_0 < \frac{\delta}{\theta_{\max} - \theta_{\min}} \tag{14.8}$$

and

$$[\theta_{\min}u_0, \theta_{\max}u_0] \subset (\theta_{\max}u_0 - \delta, \theta_{\min}u_0 + \delta).$$

For

$$C \in (\theta_{\max} u_0 - \delta, \theta_{\min} u_0 + \delta) \;\text{ but }\; C \notin [\theta_{\min} u_0, \theta_{\max} u_0],$$

we have that if $C > \theta_{\max} u_0$,

$$\begin{aligned} \max_{\theta \in [\theta_{\min}, \theta_{\max}]} \frac{\delta^2 - (C - \theta u_0)^2}{u_0^2} &= \frac{\delta^2 - \min_{\theta \in [\theta_{\min}, \theta_{\max}]} (C - \theta u_0)^2}{u_0^2} \\ &= \frac{\delta^2 - (C - \theta_{\max} u_0)^2}{u_0^2}; \end{aligned}$$

or if $C < \theta_{\min} u_0$,

$$\max_{\theta \in [\theta_{\min}, \theta_{\max}]} \frac{\delta^2 - (C - \theta u_0)^2}{u_0^2} = \frac{\delta^2 - (\theta_{\min} u_0 - C)^2}{u_0^2}.$$

In the first case, the variance is minimized when C is closest to $\theta_{\min} u_0 + \delta$, and

$$\begin{aligned} \eta^* &= \inf_{u_0 < \delta/(\theta_{\max} - \theta_{\min})} \frac{\delta^2 - (\theta_{\min} u_0 + \delta - \theta_{\max} u_0)^2}{u_0^2} \\ &= \inf_{u_0 < \delta/(\theta_{\max} - \theta_{\min})} \frac{2\delta(\theta_{\max} - \theta_{\min}) u_0 - (\theta_{\max} - \theta_{\min})^2 u_0^2}{u_0^2} \\ &= \inf_{u_0 < \delta/(\theta_{\max} - \theta_{\min})} \frac{2\delta(\theta_{\max} - \theta_{\min})}{u_0} - (\theta_{\max} - \theta_{\min})^2 \\ &= (\theta_{\max} - \theta_{\min})^2. \end{aligned}$$

In the second case, the variance is minimized when C is closest to $\theta_{\max} u_0 - \delta$ and similarly,

$$\eta^* = (\theta_{\max} - \theta_{\min})^2.$$

Thus, (14.7) is proved. □

Unbounded Disturbances

When the disturbance is unbounded such that $f(x) > 0$ for all x, for any given u_0, all $C \in \mathbb{R}$ is robustly feasible. In this case, the optimal threshold selection becomes

$$\eta^* = \inf_{C, u_0 \in \mathbb{R}} \max_{\theta \in \Omega} \frac{F(C - \theta u_0)(1 - F(C - \theta u_0))}{f^2(C - \theta u_0) u_0^2}. \tag{14.9}$$

The solutions to (14.9) can be obtained by first calculating

$$\widetilde{\eta}(C, u_0) = \max_{\theta \in \Omega} \frac{F(C - \theta u_0)(1 - F(C - \theta u_0))}{f^2(C - \theta u_0) u_0^2}, \quad C, u_0 \in \mathbb{R},$$

and then,

$$\eta^* = \inf_{C, u_0 \in \mathbb{R}} \widetilde{\eta}(C, u_0).$$

For example, suppose that the disturbance is Gaussian distributed with zero mean and standard deviation $\sigma = 25$. The left plot of Figure 14.1 is

the distribution function $F(x)$. The middle plot is $F(x)(1-F(x))/f^2(x)$. Now, suppose that $\Omega = [10, 40]$. Then, for given C and u_0, we calculate

$$\widetilde{\eta}(C, u_0) = \max_{\theta \in \Omega} \frac{F(C - \theta u_0)(1 - F(C - \theta u_0))}{f^2(C - \theta u_0) u_0^2}.$$

These functions are plotted in the right plot of Figure 14.1. From the plots, the optimal input and threshold are approximately $u_0 = 1.5$ and $C = 22.5$. The more accurate optimal values can be obtained by numerical search methods and will not be discussed further here.

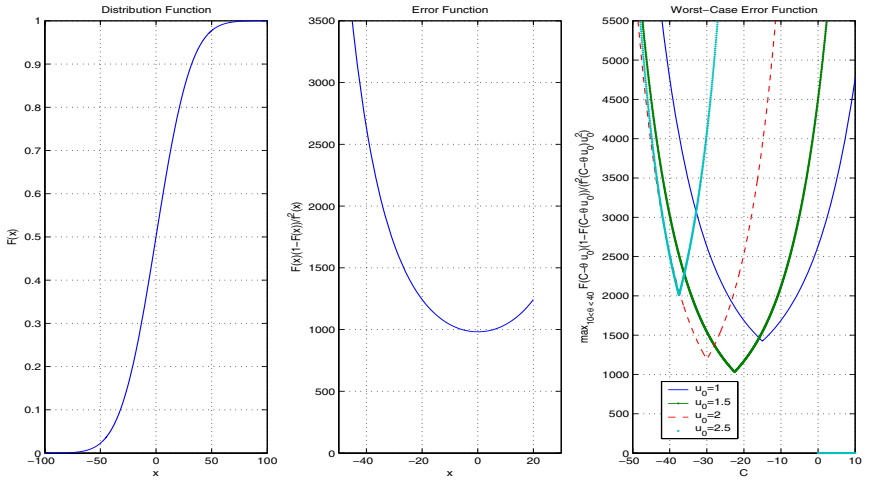


FIGURE 14.1. Optimal worst-case threshold selection and input design for $\theta \in [10, 40]$ with a Gaussian-distributed disturbance of zero mean and standard deviation 25

14.4 Threshold Adaptation

In this subsection, assume $|u_0| \leq u_{\max}$. In the special case of Theorem 14.11, when the range of uncertainty $\theta_{\max} - \theta_{\min}$ is large such that

$$(\theta_{\max} - \theta_{\min}) u_0 > 2\delta,$$

there exists no robustly feasible threshold. In other words, for any threshold $C \in \mathbb{R}$, there exists $\theta \in [\theta_{\min}, \theta_{\max}]$ for which C is not a feasible threshold. In this case, the threshold C must be adaptively selected when more information on θ can be extracted from output observations. More generally, adaptive threshold selection is useful even when a robustly feasible threshold can be found since it can potentially further reduce the errors in Theorem 14.11.

Consider again Example 14.8. Figure 14.2 demonstrates an example of Gaussian-distributed noise and the benefit to use the optimal threshold on enhancing convergence speed. The top plot indicates the estimate trajectory when a non-optimal threshold $C = 50$ is used. The bottom plot shows the estimates when the threshold is optimally selected with $C = 100$. A better convergence speed can be expected when the optimal threshold is used.

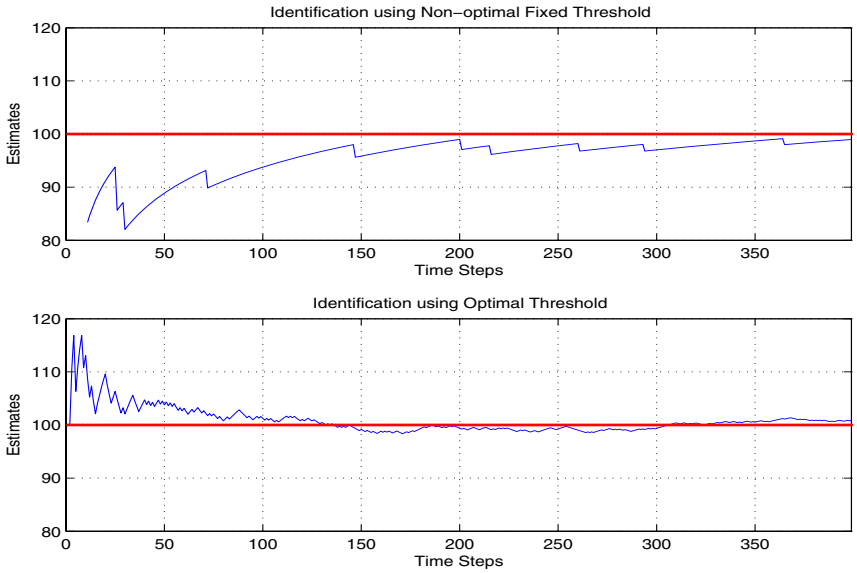


FIGURE 14.2. Comparison of identification accuracy under non-optimal and optimal thresholds: Top plot: estimates when a non-optimal threshold ($C = 50$) is used. Bottom plot: estimates when an optimal threshold ($C = 100$) is used

Conceptually, when C is not feasible, either $s(k) \equiv 0$ with probability 1, which indicates that C is too small, or $s(k) \equiv 1$ with probability 1, which indicates that C is too big. On the other hand, when C is feasible and θ is known, for any fixed u_0, the optimal threshold C^* can be derived from

$$x^* = \arg\min_x \frac{F(x)(1-F(x))}{f^2(x)}, \tag{14.10}$$

and $C^* = x^* + \theta u_0$. In particular, if $f(x)$ is an even function, one can verify that $F(x)(1-F(x))/f^2(x)$ is also an even function. It follows that $x^* = 0$ and $C^* = \theta u_0$.

Consequently, if θ is known, the optimal $u_0 = u_{\max}$ and $C^* = x^* + \theta u_0$,

$$\eta = \frac{F(x^*)(1-F(x^*))}{f^2(x^*)u_{\max}^2}.$$

When θ is unknown, since x^* can be calculated off-line, a potential adaptive threshold selection algorithm is simply

$$C_N = x^* + \theta_N u_{\max}.$$

Without loss of generality, assume $u_{\max} = 1$ in the following algorithm.

When C is adaptively selected, there are two intervening dynamic processes, one for C adaptation and the other for θ estimation. To enhance convergence, we introduce the following two-level adaptation algorithm. Let N_0 be a positive integer, representing the step size for updating the threshold. The time-line is first divided into window blocks of size N_0 indexed by $l = 0, 1, \ldots$ Then within a block, use index τ with $\tau = 0, 1, \ldots, N_0 - 1$. Threshold adaptation is done every N_0 time steps only, while parameter estimation is implemented within each block with index τ. The original time index remains k. Consequently, for a sequence x_k with $k = lN_0 + \tau$ (l and τ are easily obtained by module calculation), we will relabel the empirical measure by $\xi_k = \xi_{lN_0+\tau}$.

We present an algorithm with a two-level structure, which is motivated by an algorithm proposed in [116]. For simplicity, assume $u_0 = u_{\max} = 1$. However, unlike the aforementioned reference, in lieu of restarting the step size, we use two levels for updating parameter estimation and threshold. In view of (14.10), the inner level with index τ estimates θ with its estimate denoted by $\theta_{lN_0+\tau}$, and the outer level with index l adaptively estimates the threshold with the new value denoted by $C_{lN_0+\tau}$. The update algorithm is done inductively.

Suppose that the threshold update C_{l-1} is obtained and the parameter estimate $\theta_{(l-1)N_0+\tau}$ is calculated (the subscript i is omitted since only one threshold is considered). Note that in accordance with our notation, $\theta_{lN_0} = \theta_{(l-1)N_0+N_0}$. Construct

$$\begin{cases} \theta_{(l-1)N_0+\tau} = G(\xi_{lN_0+\tau}), \quad 1 \le \tau \le N_0, \\ C_{(l-1)N_0+\tau} = C_{lN_0}, \quad 0 \le \tau \le N_0 - 1, \\ C_{lN_0} = x^* + \theta_{lN_0}. \end{cases} \tag{14.11}$$

The second equation indicates that the threshold is not updated within a block. The second and third equations may be written together as

$$C_{(l-1)N_0+\tau} = C_{(l-1)N_0} I_{\{\tau \neq N_0\}} + (x^* + \theta_{lN_0})(1 - I_{\{\tau \neq N_0\}}), \quad 0 \le \tau \le N_0.$$

In the asymptotic analysis, we send $N_0 \to \infty$ and $l \to \infty$. As $N_0 \to \infty$, $\theta_{lN_0+\tau} \to \theta$ w.p.1. As a consequence, we also have $C_{lN_0} \to C^* = x^* + \theta$, w.p.1. In actual computations, it suffices to keep N_0 as a fixed and large constant. Then, iteration on C will gradually improve the error variance toward the optimal threshold for the true parameter with a small deviation.

14.5 Quantized Sensors and Optimal Resource Allocation

Bandwidth resources that limit data-flow rates will be denoted by R in bps. R is related to space and time complexities by $R = N \log(m_0 + 1)$. Suppose that the prior uncertainty set on θ is Ω.

Optimal Resource Allocation Problems

To understand the impact of increasing m_0, we revisit the minimal variance in Lemma 6.6:

$$\sigma^2_{\mathrm{CR}}(N, m_0, \theta) = \left(N \sum_{i=1}^{m_0+1} \frac{\widetilde{h}_i^2}{\widetilde{p}_i} \right)^{-1}.$$

The following two optimal resource allocation problems, being natural duals of each other, are introduced, where $\mathbb{Z}_+$ denotes the set of positive integers.

1. **Optimal Uncertainty Reduction:** For a given resource R, it aims at reducing $\sigma^2_{\mathrm{CR}}(N, m_0, \theta)$:

$$\begin{aligned} &\varepsilon(R) = \min_{m \in \mathbb{Z}_+} \max_{\theta \in \Omega} \sigma^2_{\mathrm{CR}}(N, m_0, \theta), \\ &\text{subject to } N \log(m_0 + 1) \le R. \end{aligned} \tag{14.12}$$

2. **Optimal Resource Allocation:** This aims at reducing R for a given error tolerance level ε, i.e., $\sigma^2_{\mathrm{CR}}(N, m_0, \theta) \le \varepsilon$ for a given resource R:

$$\begin{aligned} &R(\varepsilon) = \min_{m_0, N \in \mathbb{Z}_+} N \log(m_0 + 1), \\ &\text{subject to } \max_{\theta \in \Omega} \sigma^2_{\mathrm{CR}}(N, m_0, \theta) \le \varepsilon. \end{aligned} \tag{14.13}$$

We will consider two scenarios to increase space complexity. (1) Structured thresholds: The sets of thresholds are confined to a prespecified class that satisfies the following condition. For $m_1 < m_2$, the corresponding threshold sets $\mathcal{T}_{m_1}$ and $\mathcal{T}_{m_2}$ satisfy the ordered refinement condition: $\mathcal{T}_{m_1} \subset \mathcal{T}_{m_2}$. For instance, in the typical situation of quantization, one may start with a level of quantization. Then space complexity is increased by subdividing each subset $[C_j, C_{j+1})$ by 2 (an increase of space complexity by 1 bit). (2) Unstructured thresholds: For a given m_0, the threshold values in $\mathcal{T}_{m_0} = \{C_1, \ldots, C_{m_0}\}$ can be arbitrarily selected. This is the case, for example, when the selection of the thresholds is considered part of coding for communications. In this case, the thresholds can be designed to minimize communication resource utility.

Resource Allocation with Structured Thresholds

In this scenario, for all m_0 the threshold sets $\mathcal{T}_{m_0}$ are fixed and satisfy the monotone refinement structure $\mathcal{T}_{m_1} \subset \mathcal{T}_{m_2}$ when ever $m_1 < m_2$. We have the following monotonicity in terms of space complexity. Recall from (14.1) that

$$\eta(m_0, \theta) = N\sigma^2_{\mathrm{CR}}(N, m_0, \theta) = \left(\sum_{i=1}^{m_0+1} \frac{\widetilde{h}_i^2}{\widetilde{p}_i} \right)^{-1}.$$

For threshold selection, for a given set of m_0 thresholds $\mathcal{T}_m = \{C_1, \ldots, C_{m_0}\}$, we shall use the notation $\eta(\mathcal{T}_{m_0}, \theta) = \eta(m_0, \theta)$.

Corollary 14.12. *Under the conditions of Theorem* 14.5*,*

$$\eta(\mathcal{T}_{m_2}, \theta) \leq \eta(\mathcal{T}_{m_1}, \theta).$$

For a given resource $R = N \log(m_0+1)$, $N = R/\log(m_0+1)$. As a result, asymptotically

$$\sigma^2_{\mathrm{CR}}(N, m_0, \theta) = \frac{\eta(\mathcal{T}_{m_0}, \theta)}{N} = \frac{\log(m_0+1)\eta(\mathcal{T}_{m_0}, th)}{R}.$$

An optimal resource allocation for the given R is

$$\varepsilon_1(R) = \frac{\min_{1 \leq m \leq 2^R - 1} \log(m_0+1)\eta(\mathcal{T}_{m_0}, \theta)}{R}. \tag{14.14}$$

Example 14.13. Consider the system $y_k = \theta + d_k$. Suppose the disturbance is Gaussian distributed with zero mean and variance 200. Hence, the probability density function is

$$f(x) = \frac{e^{-\frac{x^2}{400}}}{\sqrt{400\pi}}.$$

The actual value of θ is 55. Thresholds are structured as follows. The interval of thresholds is $[-10, 70]$. Initially, one sensor threshold is placed at $C = 30$ (the middle point of the interval) with space complexity $\log(m_0 + 1) = \log 2 = 1$ bit. To increase space complexity, the number m_0 of thresholds is gradually increased by dividing the interval $[-10, 70]$ equally. Figure 14.3 shows $\eta(\mathcal{T}_{m_0}, \theta)$ and $\log(m_0 + 1)\eta(\mathcal{T}_{m_0}, \theta)$ as functions of the space complexity m_0. For this example, the optimal space complexity is $m_0 = 3$ thresholds.

The space complexity depends on the actual values of θ and threshold choices. Its dependence on θ is illustrated in Figure 14.4 in which the space complexities for three different θ values are plotted. Furthermore, the space complexity varies significantly with placement of the thresholds. Figure 14.5 shows the space complexity when the range of thresholds is changed from $[-10, 70]$ to $[-10, 60]$. The optimal number of thresholds becomes $m_0 = 2$.

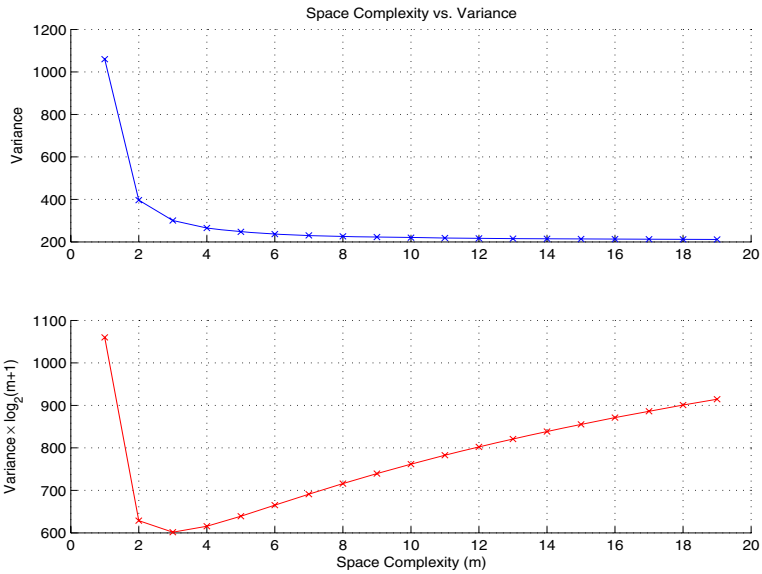


FIGURE 14.3. Space complexity: $\eta(\mathcal{T}_{m_0}, \theta)$ vs. $\log(m_0 + 1)$ (top plot); $\log(m_0 + 1)\eta(\mathcal{T}_{m_0}, \theta)$ vs. $\log(m_0 + 1)$ (bottom plot)

Resource Allocation with Unstructured Thresholds

When the thresholds are design variables, the space complexity is defined as follows. Let

$$\mathcal{Q}_{m_0} = \{\mathcal{T}_{m_0} = \{C_1, \ldots, C_{m_0}\} : y_{\min} < C_1 \leq C_2 \leq \cdots \leq C_{m_0} < y_{\max}\}.$$

This is the set of all possible m_0 thresholds. Note that in this definition, we allow thresholds to be repeated. Consequently, for any $\mathcal{T}_{m_0} \in \mathcal{Q}_{m_0}$, there exists (infinitely many) $\mathcal{T}_{m_0+1} \in \mathcal{Q}_{m_0+1}$ such that $\mathcal{T}_{m_0+1}$ is a refinement of $\mathcal{T}_{m_0}$.

Definition 14.14. $\eta_{m_0}(\theta) = \inf_{\mathcal{T}_{m_0} \in \mathcal{Q}_{m_0}} \eta(\mathcal{T}_m, \theta)$.

By Theorem 14.5, we have the following monotonicity in terms of space complexity.

Corollary 14.15. *If* $m_1 < m_2$, *then*

$$\eta_{m_2}(\theta) \leq \eta_{m_1}(\theta).$$

Proof. For any $\mathcal{T}_{m_1} \in \mathcal{Q}_{m_1}$, there exists a $\mathcal{T}_{m_2} \in \mathcal{Q}_{m_2}$ such that $\mathcal{T}_{m_2}$ is a refinement of $\mathcal{T}_{m_1}$. By Theorem 14.5, we have

$$\eta(\mathcal{T}_{m_2}, \theta) \leq \eta(\mathcal{T}_{m_1}, \theta).$$

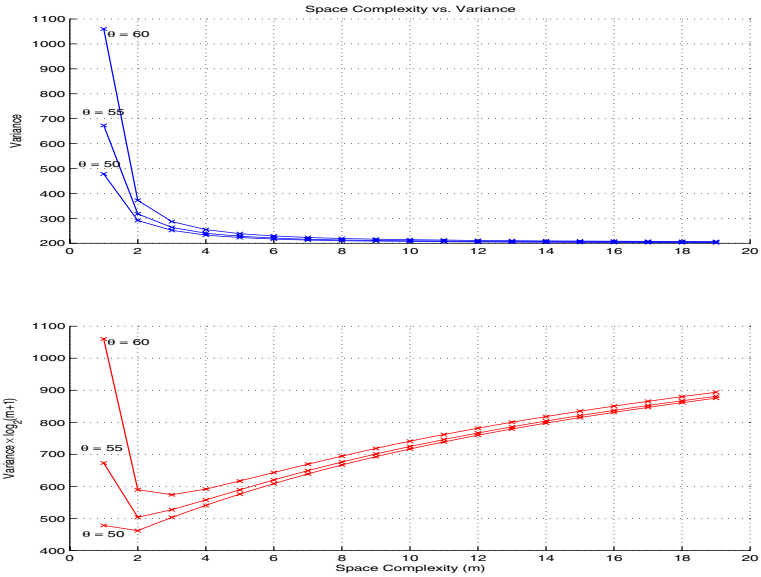


FIGURE 14.4. Different space complexity curves: $\theta = 50, 55, 60$

Consequently,

$$\eta_{m_2}(\theta) = \inf_{\mathcal{T}_{m_2} \in \mathcal{Q}_{m_2}} \eta(\mathcal{T}_{m_2}, \theta) \leq \inf_{\mathcal{T}_{m_1} \in \mathcal{Q}_{m_1}} \eta(\mathcal{T}_{m_1}, \theta) = \eta_{m_1}(\theta).$$

□

For a given resource $R = N \log(m_0 + 1)$, we have $N = R/\log(m_0 + 1)$. As a result, asymptotically

$$\sigma^2_{\mathrm{CR}}(N, m_0, \theta) = \frac{\eta_{m_0}(\theta)}{N} = \frac{\log(m_0 + 1)\eta_{m_0}(\theta)}{R}.$$

An optimal resource allocation for the given R is

$$\varepsilon_2(R) = \frac{\min\limits_{1 \leq m \leq 2^R - 1} \log(m_0 + 1)\eta_{m_0}(\theta)}{R}. \tag{14.15}$$

Example 14.16. Consider the same system setting as in Example 14.13. Suppose the disturbance is Gaussian distributed with zero mean and variance 150. The true value of $\theta = 55$. Two scenarios are compared: (1) Thresholds are structured. The interval of thresholds is $[-20, 60]$. Initially, one sensor threshold is placed at $C = 20$, with space complexity $\log(m_0 + 1) = \log 2 = 1$ bit. To increase space complexity, the number of thresholds is increased, and each time the range is divided equally. (2) Thresholds are optimized for maximum reduction of variances. Figure 14.6 demonstrates the benefit of choosing optimal thresholds for complexity reduction. The

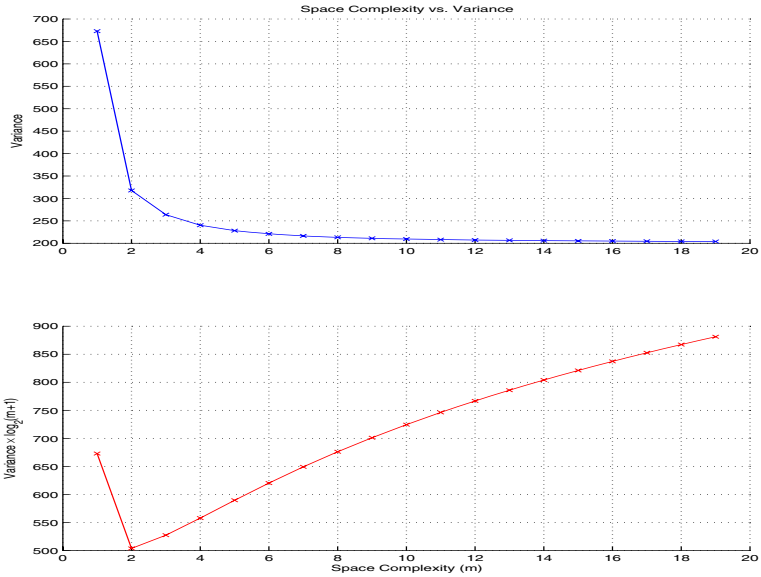


FIGURE 14.5. Space complexity curve varies with thresholds: $\theta = 55$

plots show that (1) optimization of thresholds can greatly reduce identification errors; (2) optimal space complexity can be greatly reduced. For the structured thresholds, the optimal space complexity is $m_0 = 6$ thresholds. For optimized threshold selection, it becomes 1 bit; i.e., a binary sensor is the optimal choice in terms of space complexity. The plots also show the optimized thresholds reduce variance significantly (663 vs. 236 at the respective optimal space complexities).

14.6 Discussions on Space and Time Complexity

The above examples highlight a number of interesting facts about space and time complexities.

1. It is observed that the initial increase in space complexity induces a sharp drop in variance. However, variances soon reach a near-constant level that does not significantly reduce with increased space complexity. The optimal space complexities in these examples are surprisingly low, $m_0 = 3$ in structured thresholds and $m_0 = 1$ in unstructured thresholds. It implies that when observations are corrupted by random noises, many more resources should be devoted to increasing the data size, rather than the data precision.

2. A simple calculation shows that resource allocation is a significant issue in this identification problem. A common quantization scheme

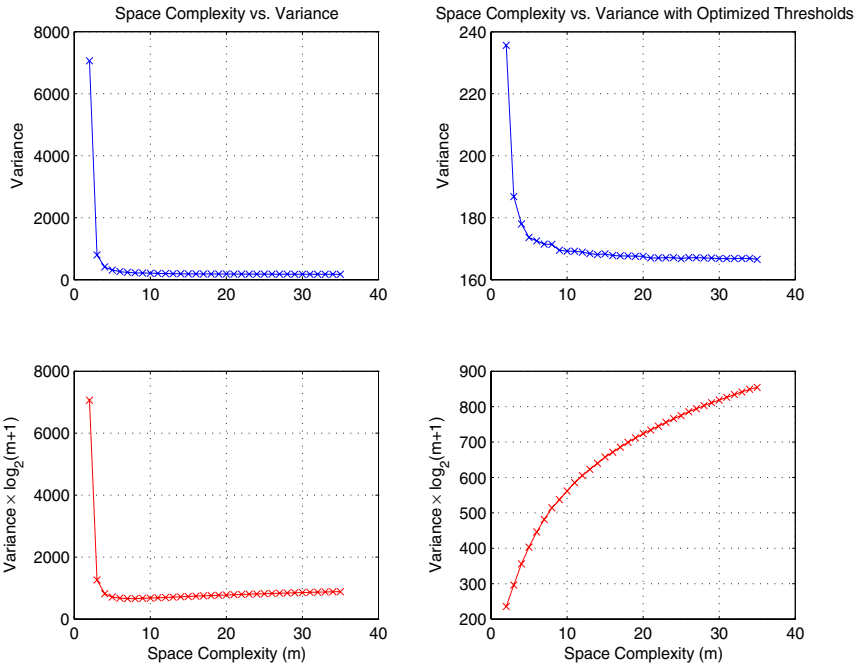


FIGURE 14.6. Comparison of space complexity: (i) Left plots: structured thresholds; (ii) right plots: optimized thresholds

in data processing will carry B bits' precision. Take an example of $B = 10$, namely, $m_0 = 2^{10} - 1 = 1023$ thresholds. From Figure 14.3, $\eta(\mathcal{T}_{1023}, \theta)$ approaches a constant about 200 for large m_0 values. To reduce the variance to, say, 0.1, the observation length N must be larger than $N \geq 200/0.1 = 2000$. Together, this amounts to $R_0 \geq NB = 20K$ bits' resource. For a rational system containing 20 parameters, the total resource will be $R = 20R_0 = 400K$ bits. Optimal resource allocations discussed in this chapter indicate that this resource request can be greatly reduced if one wisely chooses space complexity. For this example, from Figure 14.3 one may choose $m_0 = 3$ as the optimal space complexity. To achieve the same variance of 0.1, we only need $R_0 = 600/0.1 = 6K$ bits' resource, a large reduction from $20K$ bits.

14.7 Notes

The tradeoff between space complexity and time complexity is of fundamental importance in system modeling, identification, and information processing when information processing speed and data-flow rates are limited.

The issue is inherent in all problems involving signal digitization (sampling and quantization), but most relevant in systems involving communications, wireless connections, or computer networks. Following the development of [110], this chapter introduces a basic framework and certain essential tools for analyzing space and time complexities, characterizing the tradeoff in terms of identification accuracy, and optimizing resource utility. Studies on model complexity, time complexity, data compression, and information based complexity can be found in [14, 22, 23, 50, 68, 77, 81, 92, 125, 126].

15

Impact of Communication Channels on System Identification

This chapter deals with the identification of systems whose outputs must be quantized, transmitted through a communication channel, and observed afterwards. Communication errors introduce additional uncertainty that influences identification accuracy. To accomplish an information-oriented and algorithm-independent characterization of communication channels, we compare the Fisher information [or, equivalently, Cramér–Rao (CR) lower bound] of identification errors with and without communication channels. The concept of the Fisher information ratio (FI-R) is introduced. The relationship between the Fisher information ratio and Shannon's mutual information and channel capacity is explained.

The main problem is formulated in Section 15.1. We use the Fisher information to characterize how much information is contained in the observed data about unknown system parameters. Section 15.2 establishes some monotonicity properties of the Fisher information that are related to communication channel uncertainties. Section 15.3 introduces the concept of the Fisher information ratio (FI-R) as a suitable measure for characterization of communication channels in terms of system identification. Section 15.4 extends the previous results to vector parameter cases. Relationships between the FI-R and Shannon's mutual information and channel capacity are explained in Section 15.5. Based on these results, Section 15.6 discusses the tradeoff between space and time complexities in communication channels. Section 15.7 shows a multiplicative property of interconnected communication channels which will be useful for system analysis in metric spaces.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_15,

15.1 Identification with Communication Channels

Consider now the scenario of the system configuration in which sensor outputs are not directly measured, but rather are transmitted through a communication channel. For clarity, we will concentrate on the scalar observations first. That is, the sensor output s_k is scalar and takes values $s_k \in \{1, \ldots, m_0 + 1\}$. When s_k is transmitted through a communication channel, the received sequence $w_k \in \{1, \ldots, m_0 + 1\}$ is subject to channel noise and other uncertainties. When the communication channel is time invariant and memoryless, the relationship between s_k and w_k is characterized by the conditional probabilities

$$\pi^{\{ji\}} = P\{w_k = i | s_k = j\}, \quad i, j = 1, \ldots, m_0 + 1.$$

Denote $p_i^s = P\{s_k = i\}$ and $p_i^w = P\{w_k = i\}$. Then

$$p_i^w \ = P\{w_k = i\} = \sum_{j=1}^{m_0+1} P\{w_k = i | s_k = j\} P\{\widetilde{s}_k = j\} = \sum_{j=1}^{m_0+1} p_j^s \pi^{\{ji\}}.$$

Let

$$p^w = [p_1^w, \ldots, p_{m_0+1}^w]', \ p^s = [p_1^s, \ldots, p_{m_0+1}^s]'.$$

Note that $\mathbb{1}' p^s = 1$ and $\mathbb{1}' p^w = 1$. Then

$$(p^w)' = (p^s)' \Pi, \tag{15.1}$$

where

$$\Pi = \begin{bmatrix} \pi^{\{11\}} & \ldots & \pi^{\{1,m_0+1\}} \\ \vdots & \ddots & \vdots \\ \pi^{\{m_0+1,1\}} & \ldots & \pi^{\{m_0+1,m_0+1\}} \end{bmatrix}. \tag{15.2}$$

(A15.1) (a) Π is invertible. (b) All p_i^s are strictly positive.

Remark 15.1. Under Assumption (A15.1)(a), (15.1) yields $p^s = \Pi^{-T} p^w$, which ensures that the probability information p^w obtained at the receiving side of the communication channel can be used to deduce the probability p^s at the transmission site, which is then used to estimate the system parameters. Since $dp^s/dp^w = \Pi^{-T}$, the variance of the estimation error depends proportionally on the operator norm of Π^{-T}. Furthermore, if $p_i^s = 0$, the corresponding sensor threshold is not used. Such p_i^s can be eliminated from our consideration and the resulting p^s will satisfy Assumption (A15.1)(b).

For notational simplicity, we shall start with the case of scalar parameter estimation. Let $\beta \in \mathbb{R}$ be the parameter and p_i^s be related to β by an invertible mapping

$$p_i^s = K_i^s(\beta), \quad i = 1, \ldots, m_0 + 1,$$

whose inverse is continuously differentiable. Denote

$$h_i^s(\beta) = dp_i^s(\beta)/d\beta \quad \text{and} \quad h^s(\beta) = [h_1^s(\beta), \ldots, h_{m_0+1}^s(\beta)]',$$
$$h_i^w(\beta) = dp_i^w(\beta)/d\beta \quad \text{and} \quad h^w(\beta) = [h_1^w(\beta), \ldots, h_{m_0+1}^w(\beta)]'.$$

Then

$$h^w = \frac{dp^w}{d\beta} = \Pi' \frac{dp^s}{d\beta} = \Pi' h^s.$$

Lemma 15.2 ([104]). *The CR lower bound for estimating β with observations on w_k is*

$$\sigma^2_{\mathrm{CR},w}(N, m_0, \beta) = \left(N \sum_{i=1}^{m_0+1} \frac{(h_i^w)^2}{p_i^w} \right)^{-1}.$$

Similarly, by defining the Fisher information

$$\mathcal{J}_w(N, m_0, \beta) = N \sum_{i=1}^{m_0+1} \frac{(h_i^w)^2}{p_i^w},$$

we have

$$\sigma^2_{\mathrm{CR},w}(N, m_0, \beta) = \frac{1}{\mathcal{J}_w(N, m_0, \beta)}.$$

15.2 Monotonicity of Fisher Information

Define

$$D_s = \mathrm{diag}(p_1^s, \ldots, p_{m_0+1}^s),$$
$$D_w = \mathrm{diag}(p_1^w, \ldots, p_{m_0+1}^w),$$
$$S_s = \sqrt{D_s}, \quad \text{and} \quad S_w = \sqrt{D_w}.$$

Then,

$$\sum_{i=1}^{m_0+1} (h_i^s)^2/p_i^s = (h^s)' D_s^{-1} h^s,$$

and

$$\sum_{i=1}^{m_0+1} (h_i^w)^2/p_i^w = (h^w)' D_w^{-1} h^w = (h^s)' \Pi D_w^{-1} \Pi' h^s.$$

It follows that

$$J_w(N, m_0, \beta) = N (h^s)' \Pi D_w^{-1} \Pi' h^s,$$
$$J_s(N, m_0, \beta) = N (h^s)' D_s^{-1} h^s,$$

and

$$\begin{aligned}\sum_{i=1}^{m_0+1}\frac{(h_i^s)^2}{p_i^s}-\sum_{i=1}^{m_0+1}\frac{(h_i^w)^2}{p_i^w} &= (h^s)'(D_s^{-1}-\Pi D_w^{-1}\Pi')h^s\\ &= (h^s)'S_s^{-1}[I-(S_s\Pi S_w^{-1})(S_w^{-1}\Pi' S_s)]S_s^{-1}h^s\\ &= v'(I-M'M)v,\end{aligned}$$

where

$$v=S_s^{-1}h^s,\ M=S_w^{-1}\Pi' S_s. \tag{15.3}$$

Definition 15.3. $M=S_w^{-1}\Pi' S_s$ is called the *characteristic matrix* of the communication channel.

Lemma 15.4 ([104]). *$\overline{\gamma}(M)=1$, where $\overline{\gamma}(M)$ is the largest singular value of M.*

Theorem 15.5. *Under Assumption* (A15.1), *$\mathcal{J}_w(N,m_0,\beta)\le\mathcal{J}_s(N,m_0,\beta)$, or equivalently, $\sigma^2_{\mathrm{CR},w}(N,m_0,\beta)\ge\sigma^2_{\mathrm{CR},s}(N,m_0,\beta)$.*

Proof. Lemma 15.4 implies that $I-M'M\ge 0$. As a result, $v'(I-M'M)v\ge 0$. Hence,

$$\sum_{i=1}^{m_0+1}\frac{(h_i^s)^2}{p_i^s}\ge\sum_{i=1}^{m_0+1}\frac{(h_i^w)^2}{p_i^w}.$$

Consequently,

$$\begin{aligned}\mathcal{J}_w(N,m_0,\beta) &= N\sum_{i=1}^{m_0+1}\frac{(h_i^w)^2}{p_i^w}\\ &\le N\sum_{i=1}^{m_0+1}\frac{(h_i^s)^2}{p_i^s}\\ &= \mathcal{J}_s(N,m_0,\beta).\end{aligned}$$

□

15.3 Fisher Information Ratio of Communication Channels

From

$$\begin{aligned}\sigma^2_{\mathrm{CR},s} &= 1/(N(h^s)'D_s^{-1}h^s),\\ \sigma^2_{\mathrm{CR},w} &= 1/(N(h^s)'\Pi D_w^{-1}\Pi' h^s),\end{aligned}$$

the error ratio is

$$\chi(p^s,h^s,\Pi)=\frac{\sigma^2_{\mathrm{CR},s}}{\sigma^2_{\mathrm{CR},w}}=\frac{\mathcal{J}_w(N,m_0,\beta)}{\mathcal{J}_s(N,m_0,\beta)}=\frac{(h^s)'\Pi D_w^{-1}\Pi' h^s}{(h^s)'D_s^{-1}h^s}. \tag{15.4}$$

Note that h^s depends on actual function forms of distribution functions which always satisfy $\mathbf{1}'h^s = 0$. Since h^s is not part of the communication channel, we introduce the following concept to characterize the worst-case impact of a communication channel on identification accuracy.

Definition 15.6. The *Fisher information ratio* (FI-R) of a communication channel is defined as

$$\chi(p^s, \Pi) = \min_{h^s \neq 0} \chi(p^s, h^s, \Pi), \quad \text{subject to} \quad \mathbf{1}'h^s = 0.$$

Definition 15.7. The *optimal FI-R* is

$$\chi(\Pi) = \max_{p^s} \chi(p^s, \Pi),$$

where p^s satisfies $p_i^s > 0$ and $\sum_{i=1}^{m_0+1} p_i^s = 1$.

Since the actual code probability into the channel can be modified by source coding, the optimal FI-R indicates the optimal information transfer when the source code is optimally designed.

Definition 15.8. A communication channel is said to be *degenerate* if all singular values of M are equal to 1.

Theorem 15.5 implies that $\chi(p^s, \Pi) \leq 1$. If M is degenerate, then $M'M = I$. As a result,

$$\sum_{i=1}^{m_0+1} \frac{(h_i^s)^2}{p_i^s} - \sum_{i=1}^{m_0+1} \frac{(h_i^w)^2}{p_i^w} = v'(I - M'M)v = 0,$$

and $\mathcal{J}_w(N, m_0, \beta) = \mathcal{J}_s(N, m_0, \beta)$. This is the case when the channel does not introduce uncertainty and $\chi(p^s, \Pi) = 1$.

Theorem 15.9. *Under Assumptions* (A15.1)*, if the channel is not degenerate, then*

$$\chi(p^s, \Pi) = \underline{\gamma}^2(M),$$

where $\underline{\gamma}$ *is the smallest singular value.*

Proof. The following basic relationships will be used in the derivations:

$$p^w = \Pi' p^s; \ \mathbf{1}'\Pi' = \mathbf{1}'; \ \mathbf{1}'\Pi^{-T} = \mathbf{1}'; \ \mathbf{1}'D_s = (p^s)'; \ \mathbf{1}'D_w = (p^w)'.$$

By defining $v = S_s^{-1}h^s$, we have $h^s = S_s v$. Hence, $\mathbf{1}'h^s = \mathbf{1}'S_s v = 0$, and

$$\chi(p^s, h^s, \Pi) = \frac{v'S_s\Pi S_w^{-1}S_w^{-1}\Pi' S_s v}{v'v} = \frac{v'M'Mv}{v'v}.$$

As a result,

$$\chi(p^s, \Pi) = \min_{v'v=1} v'M'Mv, \quad \text{subject to} \quad \mathbf{1}'S_s v = 0.$$

Note that

$$\chi(p^s, \Pi) \geq \underline{\gamma}^2(M).$$

On the other hand, since the communication channel is not degenerate, $\underline{\gamma}^2(M) < 1$. Let $\alpha \in \mathbb{R}$ and $v \in \mathbb{R}^{m_0+1}$ be an eigenvalue/eigenvector pair for

$$M'Mv = \alpha v. \tag{15.5}$$

Note that $\underline{\gamma}^2(M) = \min \alpha$. We will show that if $\alpha < 1$, then v satisfies $\mathbf{1}'S_s v = 0$. This will prove that $\chi(p^s, \Pi) = \underline{\gamma}^2(M)$.

From $M = S_w^{-1}\Pi' S_s$, we have

$$\alpha S_s v = S_s M'Mv = D_s \Pi D_w^{-1} \Pi' S_s v.$$

It follows that

$$\begin{aligned} \alpha \mathbf{1}'S_s v &= \mathbf{1}' D_s \Pi D_w^{-1} \Pi' S_s v \\ &= (p^s)' \Pi D_w^{-1} \Pi' S_s v \\ &= (p^w)' D_w^{-1} \Pi' S_s v \\ &= \mathbf{1}' D_w D_w^{-1} \Pi' S_s v \\ &= \mathbf{1}' \Pi' S_s v \\ &= \mathbf{1}' S_s v. \end{aligned}$$

Since $\alpha < 1$, this implies $\mathbf{1}'S_s v = 0$, as claimed. □

15.4 Vector-Valued Parameters

The previous discussions can be extended to the case of multiple parameters. Let $\beta = [\beta_1, \ldots, \beta_{n_0}]'$ be n unknown parameters. Let s_k be the observation data before communications and w_k be the observation data after communications. Denote the Fisher information matrices for estimating β based on observation data s_k (or w_k) by $\mathcal{J}_s$ (or $\mathcal{J}_w$).

Definition 15.10. The Fisher information ratio for vector-valued parameters is defined by

$$\chi(p^s, h^s, \Pi) = \inf_{x \neq 0} \frac{x' \mathcal{J}_w x}{x' \mathcal{J}_s x}, \tag{15.6}$$

and

$$\chi(p^s, \Pi) = \inf_{h^s} \chi(p^s, h^s, \Pi), \quad \text{subject to} \quad \mathbf{1}'h^s = 0 \tag{15.7}$$

Since both $\mathcal{J}_s$ and $\mathcal{J}_w$ are symmetric and semipositive definite, they permit decomposition $\mathcal{J}_s = M_s' M_s$ and $\mathcal{J}_w = M_w' M_w$.

Theorem 15.11. *If $\mathcal{J}_s$ is positive definite, then*

$$\chi(p^s, h^s, \Pi) = \underline{\gamma}^2(M_w M_s^{-1}),$$

where $\underline{\gamma}$ is the smallest singular value. This implies that

$$\chi(p^s, \Pi) = \inf_{h^s} \underline{\gamma}^2(M_w M_s^{-1}), \quad \textit{subject to} \quad \mathbf{1}' h^s = 0.$$

Proof. By definition,

$$\chi(p^s, h^s, \Pi) = \inf_{x \neq 0} \frac{x' \mathcal{J}_w x}{x' \mathcal{J}_s x}.$$

Since $\mathcal{J}_s$ is nonsingular, M_s is invertible. Define $v = M_s x$. Then

$$x' \mathcal{J}_s x = v' v \quad \text{and} \quad x' \mathcal{J}_w x = (M_w M_s^{-1} v)' M_w M_s^{-1} v.$$

It follows that

$$\chi(p^s, h^s, \Pi) = \inf_{v \neq 0} \frac{v'(M_w M_s^{-1})' M_w M_s^{-1} v}{v' v} = \underline{\gamma}^2(M_w M_s^{-1}).$$

□

Most common system identification problems can be transformed into the following scalar cases:

(A15.2) Suppose that $\{s_k^i\}$ is the sequence of observations for β_i, $i = 1, \ldots, n$; $\{s_k^i\}$ and $\{s_k^j\}$ are independent when $i \neq j$. The communication channel is memoryless.

Let

$$p^s(\beta) = [p^{s^1}(\beta_1), \ldots, p^{s^{n_0}}(\beta_{n_0})]' \quad \text{and}$$
$$h^s(\beta) = [h^{s^1}(\beta_1), \ldots, h^{s^{n_0}}(\beta_{n_0})]'.$$

Then, under Assumption (A15.2), the Fisher information matrix for estimating β based on observation data $\{s_k^i : \ i = 1, \ldots, n; k = 0, \ldots, N-1\}$ is

$$\mathcal{J}_s(\beta) = \operatorname{diag}[J_{s^1}(\beta_1), \ldots, J_{s^{n_0}}(\beta_{n_0})],$$

where $J_{s^i}(\beta_i)$ is the Fisher information for estimating β_i based on observation data $\{s_k^i : \ k = 0, \ldots, N-1\}$. Similarly,

$$\mathcal{J}_w(\beta) = \operatorname{diag}[J_{w^1}(\beta_1), \ldots, J_{w^{n_0}}(\beta_{n_0})].$$

Theorem 15.12. *Under Assumption* (A15.2),

$$\chi(p^s(\beta), h^s(\beta), \Pi) = \min_{i=1,\ldots,n_0} \chi(p^{s^i}(\beta_i), h^{s^i}(\beta_i), \Pi)$$

and

$$\chi(p^s(\beta), \Pi) = \min_{i=1,\ldots,n_0} \chi(p^{s^i}(\beta_i), \Pi).$$

Proof. Under Assumption (A15.2),

$$M_s = \text{diag}[M_{s^1}, \ldots, M_{s^{n_0}}] \text{ and}$$
$$M_w = \text{diag}[M_{w^1}, \ldots, M_{w^{n_0}}].$$

Consequently, by Theorem 15.11,

$$\begin{aligned}\chi(p^s(\beta), h^s(\beta), \Pi) &= \underline{\gamma}^2(M_w M_s^{-1}) \\ &= \min_{i=1,\ldots,n_0} \underline{\gamma}^2(M_{w^i} M_{s^i}^{-1}) \\ &= \min_{i=1,\ldots,n_0} \chi(p^{s^i}(\beta_i), h^{s^i}(\beta_i), \Pi).\end{aligned}$$

Moreover,

$$\begin{aligned}\chi(p^s(\beta), \Pi) &= \inf_{h^{s^1}(\beta_1),\ldots,h^{s^{n_0}}(\beta_{n_0})} \chi(p^s, h^s, \Pi) \\ &= \min_{i=1,\ldots,n_0} \inf_{h^{s^i}(\beta_i)} \chi(p^{s^i}(\beta_i), h^{s^i}(\beta_i), \Pi) \\ &= \min_{i=1,\ldots,n_0} \chi(p^{s^i}(\beta_i), \Pi).\end{aligned}$$

□

Theorem 15.12 shows that under Assumption (A15.2), all discussions and conclusions on the FI-R for the scalar cases are applicable to the vector-valued parameters.

Moreover, it is common that the function forms $p^{s^i}(\cdot)$ are identical, with $p^{s^i}(\cdot) = p(\cdot)$, $i = 1, \ldots, n_0$. The following result shows that this limitation does not change the FI-R. Let

$$p_0^s(\beta) = [p(\beta_1), \ldots, p(\beta_{n_0})]' \text{ and}$$
$$h_0^s(\beta) = [h(\beta_1), \ldots, h(\beta_{n_0})]'.$$

Theorem 15.13. *Under Assumption* (A15.2),

$$\chi(p^s(\beta), \Pi) = \chi(p_0^s(\beta), \Pi).$$

Proof. In Theorem 15.12, let $\chi(p^s(\beta), \Pi)$ be achieved by $\chi(p^{s^l}(\beta_l), h^{s^l}(\beta_l), \Pi)$ for some l and $h^{s^l}(\beta_l)$. By taking $h(\cdot) = h^{s^l}(\cdot)$, we have the conclusion. □

15.5 Relationship to Shannon's Mutual Information

Fisher Information Ratio and Shannon's Channel Capacity

The FI-R characterizes the accuracy of information during communication and is different from Shannon's mutual information and channel capacity

$\mathcal{C}(\Pi)$, which defines a channel's capability in passing a flow of information. For example, when Π is not invertible, the accuracy of information for system identification is lost. But the channel may still allow data to flow through it. In this case, the channel capacity $\mathcal{C}(\Pi) > 0$, but FI-R $\chi(p^s, \Pi) = 0$ for any p. However, the FI-R $\chi(p^s, \Pi)$ and mutual information (channel capacity $\mathcal{C}$ is the maximum mutual information) are closely related.

Theorem 15.14. *If* $\mathcal{C} = 0$*, then* $\chi(\Pi) = 0$*.*

Proof. $\mathcal{C} = 0$ implies that s_k and w_k are independent. Let

$$\Pi = [\Pi_1, \ldots, \Pi_{m_0+1}]',$$

with Π_i' representing the ith row of Π. Then, for $i = 1, \ldots, m_0 + 1$ and $j = 1, \ldots, m_0 + 1$,

$$p_i^w = P\{w_k = i\} = P\{w_k = i | s_k = j\} = \pi_{ji},$$

which implies that

$$(p^s)'\Pi = \Pi_j', \quad j = 1, \ldots, m_0 + 1.$$

In other words, Π has identical rows. Hence, Π is not full rank. Consequently, for any p^s,

$$\chi(p^s, \Pi) = \underline{\gamma}^2(M) = \underline{\gamma}^2(S_w^{-1}\Pi' S_s) = 0.$$

□

15.6 Tradeoff between Time Information and Space Information

The Fisher information $\mathcal{J}_w(N, m_0, \beta)$ is a function of N, representing time complexity, and sensor thresholds representing space information. The total information $\mathcal{J}_w(N, m_0, \beta)$ depends on Π, representing channel uncertainty. In particular, when $\Pi \to I$ (the identity matrix), namely, the channel uncertainty is reduced to zero, the space information becomes

$$\sum_{i=1}^{m_0+1} \frac{(h_i^w)^2}{p_i^w} \to \sum_{i=1}^{m_0+1} \frac{(h_i^s)^2}{p_i^s}.$$

Shannon's celebrated noisy channel theorem provides a vehicle to define a relationship between time complexity and channel uncertainty. Observe that although the channel is subject to channel uncertainty, the amount of uncertainty can be reduced by appropriate channel coding. For example,

instead of sending the binary digit "1" or "0," one may use the longer code "11111" to represent "1" and "00000" to represent "0." It can be shown that if the decoding is to select "1" if the received code contains more 1's than 0's and vice versa, the channel uncertainty can be reduced. This coding scheme, however, will reduce the actual data-flow rate by 5. Shannon's noisy channel theorem claims that using optimal coding schemes, one can reduce channel errors, on average over large data blocks, to nearly zero when the actual data-flow rate is reduced $1/\mathcal{C}$. Consequently, we have the following relationship between the Fisher information and Shannon information.

Theorem 15.15. *If $\chi(\Pi) < \mathcal{C}$, then channel coding can increase the Fisher information ratio by a factor of $\lambda = \mathcal{C}/\chi(\Pi)$.*

Proof. Let L be the average code length. By Shannon's noisy channel theorem, one can find a sequence of coding schemes such that the corresponding channel matrices Π_L have the asymptotic property

$$\Pi_L \to \Pi, \quad \text{as } L \to 1/\mathcal{C}.$$

It follows that the corresponding sequence of the Fisher information ratios satisfies

$$\chi(\Pi_L) = \frac{\mathcal{J}_w(\Pi_L)/L}{\mathcal{J}_s(\Pi)} \to \mathcal{C}, \quad \text{as } L \to 1/\mathcal{C},$$

which implies that

$$\lambda = \lim_{L \to 1/\mathcal{C}} \frac{\chi(\Pi_L)}{\chi(\Pi)} = \frac{\mathcal{C}}{\chi(\Pi)}.$$

□

15.7 Interconnections of Communication Channels

Consider now cascade connections of communication channels. Suppose s_k (with probability vector p) is first communicated through a channel $\mathbf{C}_1$ with probability matrix Π_1 and output w_k (with probability vector p^w). The output w_k is further communicated through another channel $\mathbf{C}_2$ with probability matrix Π_2 and output z_k (with probability vector p^z).

Theorem 15.16. *The following assertions hold.*

(1) *The probability transition matrix of the cascaded system is*

$$\Pi = \Pi_1 \Pi_2.$$

(2) *Let the characteristic matrices for $\mathbf{C}_1$ and $\mathbf{C}_2$ be M_1 and M_2, respectively. Then, the characteristic matrix M of the cascaded system is*

$$M = M_2 M_1.$$

(3) *The Fisher information ratio satisfies*

$$\chi(p^s, \Pi) \geq \chi(p^s, \Pi_1)\chi(p^w, \Pi_2).$$

Proof. The proof is arranged as follows.

(1)

$$(p^z)' = (p^w)'\Pi_2 = (p^s)'\Pi_1\Pi_2,$$

which implies that

$$\Pi = \Pi_1\Pi_2.$$

(2)

$$M = S_z^{-1}\Pi' S_s = S_z^{-1}\Pi_2'\Pi_1' S_s = S_z^{-1}\Pi_2' S_w S_w^{-1}\Pi_1' S_s = M_2 M_1.$$

(3)

$$\begin{aligned}\chi(p^s, \Pi) &= \underline{\gamma}^2(M) = \underline{\gamma}^2(M_2 M_1) \\ &\geq \underline{\gamma}^2(M_2)\underline{\gamma}^2(M_1) \\ &= \chi(p^s, \Pi_1)\chi(p^w, \Pi_2).\end{aligned}$$

□

Remark 15.17. Note Statement (1) above. There is a resemblance of the well-known Chapman–Kolmogorov equation for Markov processes. Statement (2) indicates that such a probabilistic transition property also carries over to the channel property.

15.8 Notes

This chapter concerns impact of communication channels on system identification. While the Shannon channel capacity defines a lower bound on information flow, the Fisher information ratio is a precise characterization of communication channels in terms of their effect on system identification. Also, the Fisher information ratio can be easily computed as the largest singular value of the channel characteristic matrix. The material of this chapter is extracted from [105, 106]. We refer the reader to [22, 50, 81, 92] for further details on the information theory and information based complexity.

Appendix A
Background Materials

This appendix provides certain background materials to facilitate the reading of the book. It presents short reviews on selected topics such as martingales, Markov chains, diffusions, switching diffusions, and weak convergence. Although not all detailed proofs are spelled out, appropriate references are given.

Throughout the book, we work with a probability space $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and $P(\cdot)$ is a probability measure on $\mathcal{F}$. A collection of σ-algebras $\{\mathcal{F}_t\}$, for $t \geq 0$ or $t = 1, 2, \ldots$, or simply $\mathcal{F}_t$, is called a filtration if $\mathcal{F}_s \subset \mathcal{F}_t$ for $s \leq t$. The $\mathcal{F}_t$ is complete in the sense that it contains all null sets. A probability space $(\Omega, \mathcal{F}, P)$ together with a filtration $\{\mathcal{F}_t\}$ is termed a filtered probability space, denoted by $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$.

A.1 Martingales

The origin of martingales can be traced back to a class of betting strategies popular in 18th-century France. In recent years, the idea of martingales has been crucial in modern stochastic analysis and applications. The concept of martingales in probability theory was introduced by Lévy, and much of the original development of the theory was done by Doob.

Let $\{X_n\}$ be a sequence of random variables that can be either real valued or vector valued and that satisfies $E|X_n| < \infty$ for each n. If

$$E[X_{n+1}|X_i, i \leq n] = X_n \text{ w.p.1 for all } n,$$

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2,

then $\{X_n\}$ is said to be a martingale sequence. The difference

$$\delta X_n = X_{n+1} - X_n$$

is called a martingale difference. If, in addition, $E|X_n|^2 < \infty$ for each n, then it is said to be an L_2 martingale. Moreover, the martingale differences are uncorrelated in that for $m \neq n$,

$$E[X_{n+1} - X_n][X_{m+1} - X_m]' = 0.$$

Recall that A' denotes the transpose of A.

We can also define martingales with respect to a sequence of σ-algebras. Let $\{\mathcal{F}_n\}$ be a sequence of sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{F}_n \subset \mathcal{F}_{n+1}$, for all n. Suppose that X_n is measurable with respect to $\mathcal{F}_n$. Denote the expectation conditioned on $\mathcal{F}_n$ as E_n. If

$$E_n X_{n+1} = X_n \text{ w.p.1 for all } n,$$

then we say that either $\{X_n\}$ is an $\mathcal{F}_n$-martingale or $\{X_n, \mathcal{F}_n\}$ is a martingale. If we simply say that $\{X_n\}$ is a martingale without specifying $\mathcal{F}_n$, then we implicitly assume that it is just the σ-algebra generated by $\{X_i, i \leq n\}$. Martingales are one of the important classes of random processes. Note that if an $\mathcal{F}_n$-martingale is vector valued, then each of the real-valued components is also an $\mathcal{F}_n$-martingale. Likewise, a finite collection of real-valued $\mathcal{F}_n$-martingales is a vector-valued $\mathcal{F}_n$-martingale.

Suppose that X_n is real valued. If, in the definition of martingale, we replace the equality by $\leq$, i.e.,

$$E_n X_{n+1} \leq X_n \text{ w.p.1 for all } n,$$

then we say either that $\{X_n, \mathcal{F}_n\}$ is a supermartingale or that $\{X_n\}$ is an $\mathcal{F}_n$-supermartingale. If the $\mathcal{F}_n$ are understood, then we might just say that $\{X_n\}$ is a supermartingale. If

$$E_n X_{n+1} \geq X_n \text{ w.p.1 for all } n,$$

then the process is called a submartingale.

Martingale Inequalities and Convergence Theory. For general discussions on martingales and related issues, we refer the reader to [26] and [41] for further reading. Let $\{X_n, \mathcal{F}_n\}$ be a martingale, which is assumed to be real valued with no loss in generality. Then we have the following inequalities (see [11, Chapter 5], [32, Chapter 1], and [70, Chapter IV.5]). Let $c(\cdot)$ be a nonnegative, nondecreasing convex function. Then for any integers $n < N$ and $\lambda > 0$,

$$P_n\Big\{ \sup_{n \leq m \leq N} |X_m| \geq \lambda \Big\} \leq \frac{E_n c(X_N)}{c(\lambda)}.$$

Commonly used forms of $c(\cdot)$ include $c(x) = |x|$, $c(x) = |x|^2$, and $c(x) = \exp(\alpha x)$ for some positive α. In addition, the following inequality holds:

$$E_n\left[\sup_{n\le m\le N}|X_m|^2\right] \le 4E_n|X_N|^2.$$

If $\{X_n, \mathcal{F}_n\}$ is a nonnegative supermartingale, then for integers $n < N$,

$$P_n\Big\{\sup_{n\le m\le N} X_m \ge \lambda\Big\} \le \frac{X_n}{\lambda}.$$

Use x^- to denote the negative part of the real number x. That is, $x^- = \max\{0,-x\}$. Suppose that $\{X_n, \mathcal{F}_n\}$ is a real-valued submartingale with $\sup_n E|X_n| < \infty$. Then the martingale convergence theorem [11, Theorem 5.14] asserts that $\{X_n\}$ converges with probability one, as $n \to \infty$. A supermartingale $\{X_n\}$ converges with probability one if $\sup_n EX_n^- < \infty$.

A stopping time τ on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ is a nonnegative random variable such that $\{\tau \le t\} \in \mathcal{F}_t$ for all $t \ge 0$. In particular, in discrete time, with $\mathcal{F}_n$ denoting a sequence of nondecreasing σ-algebras, a random variable τ with values in $[0, \infty]$ (the set of extended nonnegative real numbers) is said to be an $\mathcal{F}_n$-stopping time (or a stopping time if the σ-algebras are evident) if $\{\tau \le n\} \in \mathcal{F}_n$ for each n. Let $\mathcal{F}_n$ be the σ-algebra determined by a random sequence $\{\xi_i, i \le n\}$. Then, if τ is an $\mathcal{F}_n$-stopping time, whether or not the event $\{\tau \le n\}$ occurred can be "determined" by looking at ξ_i up to and including time n. If a stopping time is not defined at some ω, we set its value equal to ∞ at that ω. Let $\{X_n, \mathcal{F}_n\}$ be a martingale (resp., a sub- or supermartingale) and let τ be a bounded (uniformly in ω) $\mathcal{F}_n$-stopping time. Define $\tau \wedge n = \min\{\tau, n\}$. Then $\{X_{\tau\wedge n}, \mathcal{F}_n\}$ is a martingale (resp., a sub- or supermartingale).

Burkholder's Inequality and Higher-Moment Conditions for Martingales. An extension of the martingale inequality based on the idea of Burkholder is used often. Define the martingale $X_n = \sum_{i=1}^n \varepsilon_i\xi_i$, where $\varepsilon_n \ge 0$ is $\mathcal{F}_{n-1}$-measurable and $\sum_{n\le N}\varepsilon_n \le 1$. Let $\sup_i E|\xi_i|^p < \infty$ for some even integer $p > 1$. By Burkholder's theorem [88, Theorem 6.3.10], there is a constant β (not depending on p) such that for each N,

$$E\Big|\sup_{n\le N}|X_n|^p\Big]^{1/p} \le \frac{\beta p^{5/2}}{p-1}E\Big[(\sum_{n=1}^N (X_n - X_{n-1})^2)^{p/2}\Big]^{1/p}. \tag{A.1}$$

Define $m = p/2$, and let $s, c_1, \ldots, c_s$ be arbitrary positive integers such that $\sum_{i=1}^s c_i = m$. Then, for some K depending on $\sup_n E|\xi_n|^p$,

$$KE\Big[(\sum_{n=1}^N (X_n - X_{n-1})^2\Big)^{p/2}\Big] \le KE\sum_P \sum_{i\le N}\varepsilon_i^{2c_1}\sum_{i\le N}\varepsilon_i^{2c_2}\cdots\sum_{i\le N}\varepsilon_i^{2c_s}, \tag{A.2}$$

where $\sum_P$ is the sum over all such partitions of $[0, m]$. Consider a typical partition. Rewrite the inner sums in (A.2) as

$$E \sum_{i \le N} \varepsilon_i \varepsilon_i^{2c_1 - 1} \sum_{i \le N} \varepsilon_i \varepsilon_i^{2c_2 - 1} \cdots \sum_{i \le N} \varepsilon_i \varepsilon_i^{2c_s - 1},$$

Now, use Hölder's inequality to get the bound on (A.2):

$$E \Big[\sum_{i \le N} \varepsilon_i \Big]^{l_s} \Big[\sum_{i \le N} \varepsilon_i \varepsilon_i^{(2c_1 - 1)q_1} \Big]^{1/q_1} \cdots \Big[\sum_{i \le N} \varepsilon_i \varepsilon_i^{(2c_s - 1)q_s} \Big]^{1/q_s},$$

where $l_s = \sum_{i=1}^{s} 1/p_i$ and $1/p_i + 1/q_i = 1$. Choose the q_i such that $\sum_{i=1}^{s} 1/q_i = 1$ and the exponents are equal in that, for some $c > 0$, $(2c_i - 1)q_i = c$ for all i. Then $l_s = s - 1$ and

$$\sum_{i=1}^{s} 1/q_i = 1 = \sum_{i=1}^{s} (2c_i - 1)/c = (2m - s)/c.$$

Choose s so that c is the smallest, yielding the bound on (A.1)

$$K_1 E \Big[\sum_{i \le N} \varepsilon_i \Big] \sum_{i \le N} \varepsilon_i \Big[\varepsilon_i \Big]^m,$$

where K_1 depends on K, β, and p; see [5, Proposition 4.2] for a similar calculation, and also [54, Example 6, Section 2.2].

Continuous-Time Martingales. There are definitions of martingale and sub- and supermartingale to continuous-time random processes. Let $X(t)$ be a random process satisfying $E|X(t)| < \infty$ for each $t \ge 0$, and let $\mathcal{F}_t$ be a nondecreasing sequence of σ-algebras such that $X(t)$ is $\mathcal{F}_t$-measurable. If $E_{\mathcal{F}_t} X(t + s) = X(t)$ with probability one for each $t \ge 0$ and $s > 0$, then $\{X(t), \mathcal{F}_t\}$ is termed a martingale. If we only say that $X(\cdot)$ is a martingale (without specifying the filtration $\mathcal{F}_t$), $\mathcal{F}_t$ is taken to be the natural filtration $\sigma\{X(s) : \ s \le t\}$. If there exists a sequence of stopping times $\{\tau_n\}$ such that $0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_n \le \tau_{n+1} \le \cdots$, $\tau_n \to \infty$ w.p.1 as $n \to \infty$, and the process $X^{(n)}(t) := x(t \wedge \tau_n)$ is a martingale, then $X(\cdot)$ is a local martingale.

A.2 Markov Chains

Here, we review some concepts of Markov chains. It is devoted to two subsections. The discussion on discrete-time Markov chains follows that of [122], whereas that of a continuous-time counterpart is from [121]. In applications, one may wish to simulate a Markov chain. A guide on carrying out such simulations may be found in [122, pp. 315–316].

Discrete-Time Markov Chains

Working with discrete time $k \in \{0, 1, \ldots\}$, consider a sequence $\{X_k\}$ of $\mathbb{R}^r$ vectors (or an $\mathbb{R}^r$-valued random variable). If, for each k, X_k is a random vector, we call $\{X_k\}$ a stochastic process and write it as X_k, $k = 0, 1, 2, \ldots$, or simply X_k if there is no confusion. A stochastic process is wide-sense (or covariance) stationary if it has a finite second moment, a constant mean, and a covariance depending only on the time difference. The ergodicity of a stationary sequence $\{X_k\}$ refers to the convergence of the sequence $\sum_{i=1}^{n} X_i/n$ to its expectation in the almost sure or some weak sense; see Karlin and Taylor [47, Theorem 5.6, p. 487] for a strong ergodic theorem of a stationary process. Roughly, this indicates that the ensemble average can be replaced by a time average in the limit. We say that a stochastic process X_k is adapted to a filtration $\{\mathcal{F}_k\}$ if, for each k, X_k is an $\mathcal{F}_k$-measurable random vector.

Suppose that α_k is a stochastic process taking values in $\mathcal{M}$, which is at most countable (i.e., it is either finite $\mathcal{M} = \{1, 2, \ldots, m_0\}$ or countable $\mathcal{M} = \{1, 2, \ldots\}$). We say that α_k is a Markov chain if

$$\begin{aligned} p_{k,k+1}^{\{ij\}} &= P(\alpha_{k+1} = j | \alpha_k = i) \\ &= P(\alpha_{k+1} = j | \alpha_0 = i_0, \ldots, \alpha_{k-1} = i_{k-1}, \alpha_k = i), \end{aligned} \tag{A.3}$$

for any $i_0, \ldots, i_{k-1}, i, j \in \mathcal{M}$. In the above, $p_{k,k+1}^{\{ij\}}$ is the probability of α_{k+1} being in state j given that α_k is in state i and is called a one-step transition probability. The matrix $P_{k,k+1} = (p_{k,k+1}^{\{ij\}})$ is named a transition matrix. The notation indicates that in general the transition probabilities are functions not only of the initial and final states, but also of the time of transition. The defining property (A.3) is known as the Markov property. Thus, α_k is a Markov chain if it has the memoryless property. Roughly, a Markov chain is one that given the values of α_k, α_n for $n > k$ do not depend on the values of α_j for $j < n$.

Given i, j, if $p_{k,k+1}^{\{ij\}}$ is independent of time k, i.e., $p_{k,k+1}^{\{ij\}} = p^{\{ij\}}$, we say that α_k has stationary transition probabilities. The corresponding Markov chains are said to be stationary or time-homogeneous or temporally homogeneous or simply homogeneous. In this case, let $P = (p^{\{ij\}})$ denote the transition matrix. Denote the n-step transition matrix by $P^{(n)} = (p^{(n),\{ij\}})$, with

$$p^{(n),\{ij\}} = P(x_n = j | x_0 = i).$$

Then $P^{(n)} = (P)^n$. That is, the n-step transition matrix is simply the matrix P to the nth power. Note that

(a) $p^{\{ij\}} \geq 0$, $\sum_j p^{\{ij\}} = 1$, and

(b) $(P)^{k_1+k_2} = (P)^{k_1}(P)^{k_2}$, for $k_1, k_2 = 1, 2, \ldots$

The last identity is known as the Chapman–Kolmogorov equation. Working with Markov chains with finite state spaces, certain algebraic properties of Markov chains will be used in the book, some of which are listed next.

Suppose that A is an $r \times r$ square matrix. Denote the collection of eigenvalues of A by Λ. Then the spectral radius of A, denoted by $\rho(A)$, is defined by $\rho(A) = \max_{\lambda \in \Lambda} |\lambda|$. Recall that a matrix with real entries is said to be positive if it has at least one positive entry and no negative entries. If every entry of A is positive, we call the matrix strictly positive. Similarly, for a vector $x = (x^{\{1\}}, \ldots, x^{\{r\}})$, by $x \geq 0$, we mean that $x^{\{i\}} \geq 0$ for $i = 1, \ldots, r$; by $x > 0$, we mean that all entries $x^{\{i\}} > 0$.

Let $P = (p^{\{ij\}}) \in \mathbb{R}^{m_0 \times m_0}$ be a transition matrix. Clearly, it is a positive matrix. Then $\rho(P) = 1$; see Karlin and Taylor [48, p. 3]. This implies that all eigenvalues of P are on or inside the unit circle.

For a Markov chain α_k, state j is said to be accessible from state i if $p^{(k),\{ij\}} = P(\alpha_k = j | \alpha_0 = i) > 0$ for some $k > 0$. Two states i and j, accessible from each other, are said to communicate. A Markov chain is irreducible if all states communicate with each other. For $i \in \mathcal{M}$, let $d(i)$ denote the period of state i, i.e., the greatest common divisor of all $k \geq 1$ such that $P(\alpha_{k+n} = i | \alpha_n = i) > 0$ [define $d(i) = 0$ if $P(\alpha_{k+n} = i | \alpha_n = i) = 0$ for all k]. A Markov chain is called aperiodic if each state has period one. According to Kolmogorov's classification of states, a state i is recurrent if, starting from state i, the probability of returning to state i after some finite time is 1. A state is transient if it is not recurrent. Criteria on recurrence can be found in most standard textbooks on stochastic processes or Markov chains.

Note that (see Karlin and Taylor [48, p. 4]) if P is a transition matrix for a finite-state Markov chain, the multiplicity of the eigenvalue 1 is equal to the number of recurrent classes associated with P. A row vector $\pi = (\pi^{\{1\}}, \ldots, \pi^{\{m_0\}})$ with each $\pi^{\{i\}} \geq 0$ is called a stationary distribution of α_k if it is the unique solution to the system of equations

$$\begin{cases} \pi P = \pi, \\ \sum_{i=1}^{m_0} \pi^{\{i\}} = 1. \end{cases}$$

As demonstrated in [48, p. 85], for i in an aperiodic recurrent class, if $\pi^{\{i\}} > 0$, which is the limit of the probability of starting from state i and then entering state i at the nth transition as $n \to \infty$, then for all j in this class of i, $\pi^{\{j\}} > 0$, and the class is termed positive recurrent or strongly ergodic. The following theorem is concerned with the spectral gaps.

Theorem A.1. *Let $P = (p^{\{ij\}})$ be the transition matrix of an irreducible aperiodic finite-state Markov chain. Then there exist constants $0 < \lambda < 1$ and $c_0 > 0$ such that the kth-step transition probabilities satisfy*

$$|(P)^k - \overline{P}| \leq c_0 \lambda^k \quad \textit{for} \quad k = 1, 2, \ldots,$$

where $\overline{P} = \mathbb{1}_{m_0}\pi$, $\mathbb{1}_{m_0} = (1,\ldots,1)' \in \mathbb{R}^{m_0\times 1}$, *and* $\pi = (\pi^{\{1\}},\ldots,\pi^{\{m_0\}})$ *is the stationary distribution of* α_k. *This implies, in particular,*

$$\lim_{k\to\infty} P^k = \mathbb{1}_{m_0}\pi.$$

Continuous-Time Markov Chains

Working with continuous time, a right-continuous stochastic process is a jump process if it has piecewise-constant sample paths. Suppose that $\alpha(\cdot) = \{\alpha(t) : t \geq 0\}$ is a jump process defined on $(\Omega, \mathcal{F}, P)$ taking values in either $\mathcal{M} = \{1, 2, \ldots, m_0\}$ or $\mathcal{M} = \{1, 2, \ldots\}$. Then $\{\alpha(t) : t \geq 0\}$ is a Markov chain with state space $\mathcal{M}$ if

$$P(\alpha(t) = i|\alpha(r) : r \leq s) = P(\alpha(t) = i|\alpha(s)),$$

for all $0 \leq s \leq t$ and $i \in \mathcal{M}$.

For any $i, j \in \mathcal{M}$ and $t \geq s \geq 0$, let $p^{\{ij\}}(t,s)$ denote the transition probability $P(\alpha(t) = j|\alpha(s) = i)$, and $P(t,s)$ the matrix $(p^{\{ij\}}(t,s))$. We call $P(t,s)$ the transition matrix of the Markov chain $\alpha(\cdot)$, and postulate that

$$\lim_{t\to s^+} p^{\{ij\}}(t,s) = \delta^{\{ij\}},$$

where $\delta^{\{ij\}} = \begin{cases} 1 & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$ It follows that for $0 \leq s \leq u \leq t$,

$$\begin{cases} p^{\{ij\}}(t,s) \geq 0,\ i,j \in \mathcal{M}, \\ \displaystyle\sum_{j\in\mathcal{M}} p^{\{ij\}}(t,s) = 1,\ i \in \mathcal{M}, \\ p^{\{ij\}}(t,s) = \displaystyle\sum_{k\in\mathcal{M}} p^{\{ik\}}(u,s)p^{\{kj\}}(t,u),\ i,j \in \mathcal{M}. \end{cases}$$

The last identity is referred to as the Chapman–Kolmogorov equation.

If the transition probability $P(\alpha(t) = j|\alpha(s) = i)$ depends only on $(t-s)$, then $\alpha(\cdot)$ is stationary. In this case, we define $p^{\{ij\}}(h) := p^{\{ij\}}(s+h,s)$ for any $h \geq 0$. The process is nonstationary otherwise.

Definition A.2 (q-property). Denote $Q(t) = (q^{\{ij\}}(t))$, for $t \geq 0$. It satisfies the q-property if

(a) $q^{\{ij\}}(t)$ is Borel measurable for all $i, j \in \mathcal{M}$ and $t \geq 0$;

(b) $q^{\{ij\}}(t)$ is uniformly bounded; that is, there exists a constant K such that $|q^{\{ij\}}(t)| \leq K$, for all $i, j \in \mathcal{M}$ and $t \geq 0$;

(c) $q^{\{ij\}}(t) \geq 0$ for $j \neq i$ and $q^{\{ii\}}(t) = -\sum_{j\neq i} q^{\{ij\}}(t)$, $t \geq 0$.

For any real-valued function f on $\mathcal{M}$ and $i \in \mathcal{M}$, denote

$$Q(t)f(\cdot)(i) = \sum_{j\in\mathcal{M}} q^{\{ij\}}(t)f(j) = \sum_{j\neq i} q^{\{ij\}}(t)(f(j) - f(i)).$$

We are now ready to define the generator of a Markov chain.

Definition A.3 (Generator). A matrix $Q(t)$, $t \geq 0$, is an infinitesimal generator (or simply a generator) of $\alpha(\cdot)$ if it satisfies the q-property, and for all bounded real-valued functions f defined on $\mathcal{M}$

$$f(\alpha(t)) - \int_0^t Q(u)f(\cdot)(\alpha(u))du \tag{A.4}$$

is a martingale.

Remark A.4. If the Markov chain is time-homogeneous, then the generator $Q(t) = Q$ becomes a constant matrix. The above definition is more general since it allows the Markov chain to be inhomogeneous.

Motivated by the many applications we are interested in, a generator is defined for a matrix satisfying the q-property only in the above definition. In fact, in this book, most of the Markov chains that we consider have finite state spaces. Different definitions, including other classes of matrices, may be found in Chung [20]. To proceed, we note the following equivalence for a finite-state Markov chain generated by $Q(\cdot)$. The proof of this fact can be found in [121, Lemma 2.4]; the details are omitted.

Lemma A.5. *Let $\mathcal{M} = \{1, \ldots, m_0\}$. Then $\alpha(t) \in \mathcal{M}$, $t \geq 0$, is a Markov chain generated by $Q(t)$ if and only if*

$$\left(I_{\{\alpha(t)=1\}}, \ldots, I_{\{\alpha(t)=m_0\}}\right) - \int_0^t \left(I_{\{\alpha(u)=1\}}, \ldots, I_{\{\alpha(u)=m_0\}}\right) Q(u)du \tag{A.5}$$

is a martingale.

It is interesting to note that

$$\begin{aligned} f(\alpha(t)) &= \sum_{i=1}^{m_0} I_{\{\alpha(u)=i\}} f(i) \\ &= \left(I_{\{\alpha(t)=1\}}, \ldots, I_{\{\alpha(t)=m_0\}}\right) \left(f(1), \ldots, f(m_0)\right)' \end{aligned}$$

and

$$\begin{aligned} Q(u)f(\cdot)(\alpha(u)) &= \sum_{i=1}^{m_0} I_{\{\alpha(u)=i\}} [Q(u)f(\cdot)(i)] \\ &= \left(I_{\{\alpha(u)=1\}}, \ldots, I_{\{\alpha(u)=m_0\}}\right) Q(u) \left(f(1), \ldots, f(m_0)\right)'. \end{aligned}$$

It can be shown that for any given $Q(t)$ satisfying the q-property, there exists a Markov chain $\alpha(\cdot)$ generated by $Q(t)$. The construction follows the piecewise-deterministic process approach of Davis [24]. One begins by considering $0 = \tau_0 < \tau_1 < \cdots < \tau_l < \cdots$, a sequence of jump times of $\alpha(\cdot)$ such that the random variables τ_1, $\tau_2 - \tau_1$, ..., $\tau_{k+1} - \tau_k$, ... are independent. Let $\alpha(0) = i \in \mathcal{M}$. Then we can compute the probability distribution $P(\tau_k \in B)$ of the jump time τ_k, where $B \subset [0, \infty)$ is a Borel set, and specify the post jump location of $\alpha(t)$; the details can be found in [121, p. 19]. Moreover, we obtain the following result.

Theorem A.6. *Suppose that the matrix $Q(t)$ satisfies the q-property for $t \geq 0$. Then*

(1) *The process $\alpha(\cdot)$ constructed above is a Markov chain.*

(2) *The process*

$$f(\alpha(t)) - \int_0^t Q(u) f(\cdot)(\alpha(u)) du \tag{A.6}$$

is a martingale for any uniformly bounded function $f(\cdot)$ on $\mathcal{M}$. Thus, $Q(t)$ is indeed the generator of $\alpha(\cdot)$.

(3) *The transition matrix $P(t, s)$ satisfies the forward differential equation*

$$\begin{aligned} &\frac{dP(t,s)}{dt} = P(t,s)Q(t),\ t \geq s, \\ &P(s,s) = I, \end{aligned} \tag{A.7}$$

where I is the identity matrix.

(4) *Assume further that $Q(t)$ is continuous in t. Then $P(t, s)$ also satisfies the backward differential equation*

$$\begin{aligned} &\frac{dP(t,s)}{ds} = Q(s)P(t,s),\ t \geq s, \\ &P(s,s) = I. \end{aligned} \tag{A.8}$$

Diffusion and Switching Diffusion Processes

Diffusion Processes. Diffusion processes are referred to as the solutions of stochastic differential equations (SDEs). The work on SDEs may be traced back to the work of Einstein and Smoluchowski for describing Brownian motions. One of the earliest works related to Brownian motion is in Bachelier's 1900's thesis, "Theory of Speculation." The mathematical foundation of stochastic differential equations was set up by Itô.

Brownian motion, named after the Scottish botanist Robert Brown, was initially used to describe the random movement of particles suspended in a liquid or gas. In the analysis of stochastic systems, a useful way to model the noise is to use a Brownian motion.

Let $W(\cdot)$ be an $\mathbb{R}^d$-valued process with continuous sample paths such that $W(0) = 0$, $EW(t) = 0$; for any set of increasing real numbers $\{t_i\}$ the set $\{W(t_{i+1}) - W(t_i)\}$ is mutually independent; and the distribution of $W(t+s) - W(t)$, $s > 0$, does not depend on t. Then $W(\cdot)$ is called an $\mathbb{R}^d$-valued Wiener process or Brownian motion, and there is a matrix Σ, called the covariance, such that $EW(t)W'(t) = \Sigma t$, and the increments are normally distributed [11]. When $\Sigma = I$, the corresponding Brownian motion is said to be a standard Brownian motion. The next theorem gives a criterion for verifying that a process is a Brownian motion.

Theorem A.7 ([28, Chapter 5, Theorem 2.12])**.** *Let $\{X(t), \mathcal{F}_t\}$ be a vector-valued martingale with continuous sample paths and let there be a matrix Σ such that for each t and $s \geq 0$,*

$$E_{\mathcal{F}_t}\left[X(t+s) - X(t)\right]\left[X(t+s) - X(t)\right]' = \Sigma s \text{ w.p.1.}$$

Then $W(\cdot)$ is a Brownian motion with mean 0 and covariance parameter Σ.

For a stochastic process $X(t)$, and for any $k \in \mathbb{Z}$ and $0 \leq t_1 \leq \cdots \leq t_k$, the distribution of $(X(t_1), \ldots, X(t_k))$ is said to be a finite-dimensional distribution of $X(t)$. Consider a stochastic process $X(t)$, $t \geq 0$. It is a Gaussian process if its finite-dimensional distribution is Gaussian. A random process $X(\cdot)$ is said to have independent increments if, for any $k = 1, 2, \ldots$ and $0 \leq t_1 < t_2 < \cdots < t_k$,

$$(X(t_1) - X(0)),\ (X(t_2) - X(t_1)),\ \ldots,\ (X(t_k) - X(t_{k-1}))$$

are independent. A sufficient condition for a process to be Gaussian is in Skorohod [86, p. 7].

Lemma A.8. *Suppose that the process $X(\cdot)$ has independent increments and continuous sample paths almost surely. Then $X(\cdot)$ is a Gaussian process.*

Suppose that $b(\cdot) : \mathbb{R}^r \mapsto \mathbb{R}^r$ and $\sigma(\cdot) : \mathbb{R}^r \mapsto \mathbb{R}^{r \times d}$ are nonrandom Borel-measurable functions. A stochastic differential equation with drift $b(\cdot)$ and diffusion coefficient $\sigma(\cdot)$ is given by

$$dX(t) = b(X(t))dt + \sigma(X(t))dW(t), \tag{A.9}$$

where $W(\cdot)$ is a standard d-dimensional Brownian motion. For simplicity, we only consider the time-homogeneous case. That is, the $b(\cdot)$ and $\sigma(\cdot)$

do not depend on t explicitly. Since $W(\cdot)$ is known to be continuous everywhere, but nowhere differentiable, the above stochastic differential equation is understood in the integral sense (with the stochastic integrals carefully defined; see [46] for instance).

A process $X(\cdot)$ satisfying

$$X(t) = X(0) + \int_0^t b(X(s))ds + \int_0^t \sigma(X(s))dW(s) \tag{A.10}$$

is called a diffusion. Then $X(\cdot)$ defined in (A.10) is a Markov process in the sense that it verifies the Markov property $P(X(t) \in A|\mathcal{F}_s) = P(X(t) \in A|X(s)), \forall s \in [0,t]$ and for any Borel set A. A more general definition allows $b(\cdot)$ and $\sigma(\cdot)$ to be time-dependent and $\mathcal{F}_t$-measurable processes. Nevertheless, the current definition is sufficient for our purposes.

Associated with the diffusion process, there is an operator $\mathcal{L}$, known as the generator of the diffusion $X(\cdot)$. Denote by C^2 the class of real-valued functions on (a subset of) $\mathbb{R}^r$ whose second-order partial derivatives with respect to x are continuous. Define the operator $\mathcal{L}$ on C^2 by

$$\mathcal{L}f(x) = \sum_{i=1}^{r} b_i(x)\frac{\partial f(x)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{r} a_{ij}(x)\frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \tag{A.11}$$

where $A(x) = (a_{ij}(x)) = \sigma(x)\sigma'(x) \in \mathbb{R}^{r\times r}$ is an $r \times r$ matrix-valued function. The above may also be written as

$$\mathcal{L}f(x) = b'(x)\nabla f(x) + \frac{1}{2}\mathrm{tr}(Hf(x)A(t,x)),$$

where ∇f and Hf denote the gradient and Hessian of f, respectively.

The well-known Itô lemma (see Gihman and Skorohod [35, 36], Ikeda and Watanabe [46], and Liptser and Shiryayev [60]) states that

$$df(X(t)) = \mathcal{L}f(X(t)) + \nabla f'(X(t))\sigma(X(t))dW(t),$$

or, in its integral form,

$$\begin{aligned} &f(X(t)) - f(X(0)) \\ &\quad = \int_0^t \mathcal{L}f(X(s))ds + \int_0^t \nabla f'(X(s))\sigma(X(s))dW(s). \end{aligned}$$

By virtue of Itô's lemma,

$$M_f(t) = f(t, X(t)) - f(0, X(0)) - \int_0^t \mathcal{L}f(s, X(s))ds$$

is a square-integrable $\mathcal{F}_t$-martingale. Conversely, suppose that $X(\cdot)$ is right continuous. Then $X(\cdot)$ is said to be a solution to the martingale problem

with operator $\mathcal{L}$ if $M_f(\cdot)$ is a martingale for each $f(\cdot) \in C_0^2$ (the class of C^2 functions with compact support). For more discussions on multidimensional diffusion processes, we refer the reader to Stroock and Varadhan [89] and references therein.

Switching Diffusion Processes. Recently, switching diffusion processes have gained popularity. In many applications, the usual diffusion processes become inadequate. This is because the systems involve both continuous dynamics and discrete events. The interactions of the continuous and discrete processes provide a more realistic formulation. Nevertheless, the analysis is much more involved.

Suppose that $\alpha(\cdot)$ is a stochastic process with right-continuous sample paths, finite-state space $\mathcal{M} = \{1, \ldots, m\}$, and x-dependent generator $Q(x)$ so that for a suitable function $f(\cdot, \cdot)$,

$$Q(x)f(x,\cdot)(i) = \sum_{j \in \mathcal{M}} q_{ij}(x)(f(x,j) - f(x,i)), \quad \text{for each} \quad i \in \mathcal{M}. \tag{A.12}$$

Let $W(\cdot)$ be an $\mathbb{R}^d$-valued standard Brownian motion defined in the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$. Suppose that $b(\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \mapsto \mathbb{R}^r$ and that $\sigma(\cdot,\cdot) : \mathbb{R}^r \times \mathcal{M} \mapsto \mathbb{R}^d$. The two-component process $(X(\cdot), \alpha(\cdot))$, satisfying

$$\begin{aligned} &dX(t) = b(X(t), \alpha(t))dt + \sigma(X(t), \alpha(t))dw(t), \\ &(X(0), \alpha(0)) = (x, \alpha), \end{aligned} \tag{A.13}$$

and for $i \neq j$,

$$P\{\alpha(t+\Delta) = j | \alpha(t) = i, X(s), \alpha(s), s \leq t\} = q_{ij}(X(t))\Delta + o(\Delta), \tag{A.14}$$

is called a switching diffusion or a regime-switching diffusion. Naturally, for the two-component process $(X(t), \alpha(t))$, we call $X(t)$ the continuous component and $\alpha(t)$ the discrete component, in accordance with their sample path properties. Associated with the process $(X(t), \alpha(t))$, for each $i \in \mathcal{M}$ and each $f(\cdot, i) \in C^2$, we have

$$\begin{aligned} \mathcal{L}f(x,i) &= \nabla f'(x,i)b(x,i) + \text{tr}(Hf(x,i)A(x,i)) + Q(x)f(x,\cdot)(i) \\ &= \sum_{i=1}^{r} b_i(x,i)\frac{\partial f(x,i)}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{r} a_{ij}(x,i)\frac{\partial^2 f(x,i)}{\partial x_i \partial x_j} \\ &\quad + Q(x)f(x,\cdot)(i), \end{aligned} \tag{A.15}$$

where $\nabla f(x,i)$ and $Hf(x,i)$ denote the gradient and Hessian of $f(x,i)$ with

respect to x, respectively,

$$Q(x)f(x,\cdot)(i) = \sum_{j=1}^{m} q_{ij} f(x,j), \quad \text{and}$$
$$A(x,i) = (a_{ij}(x,i)) = \sigma(x,i)\sigma'(x,i) \in \mathbb{R}^{r\times r}.$$

Note that when $Q(x) = Q$ is independent of x, the process becomes the so-called Markovian regime-switching diffusion. That is, $\alpha(\cdot)$ and $W(\cdot)$ are independent. While the Markovian regime-switching diffusions have been examined by many researchers, the switching diffusions with state-dependent switching are much difficult to deal with. For example, if one is interested in the Itô formula, then one can rewrite the pure jump process $\alpha(\cdot)$ as a stochastic integral with respect to a Poisson measure. Then one will get a version of the Itô formula in which it has two martingale terms. One of them, as in the diffusion cases, is a martingale with a driving noise being the Brownian motion. The other is a martingale with respect to the jump measure. Some of the basic properties such as the Feller property, the strong Feller property, and the smooth dependence on the initial data become much more difficult to obtain. We refer the reader to Yin and Zhu [123] for further details.

A.3 Weak Convergence

Convergence in distribution is a basic concept in elementary probability theory. The notion of weak convergence is a generalization of convergence in distribution. In what follows, we present definitions and results, including tightness, tightness criteria, the martingale problem, Skorohod representation, and Prohorov's theorem.

Definition A.9 (Weak convergence). Let P and P_k, $k = 1, 2, \ldots$, be probability measures defined on a metric space $\mathbb{S}$. The sequence $\{P_k\}$ converges weakly to P if

$$\int f dP_k \to \int f dP$$

for every bounded and continuous function $f(\cdot)$ on $\mathbb{S}$. Suppose that $\{X_k\}$ and X are random variables associated with P_k and P, respectively. The sequence X_k converges to X weakly if, for any bounded and continuous function $f(\cdot)$ on $\mathbb{S}$, $Ef(X_k) \to Ef(X)$ as $k \to \infty$.

Let $D([0,\infty);\mathbb{R}^r)$ be the space of $\mathbb{R}^r$-valued functions defined on $[0,\infty)$ that are right continuous and have left-hand limits; let $\mathbb{L}$ be a set of strictly increasing Lipschitz continuous functions $\zeta(\cdot) : [0,\infty) \mapsto [0,\infty)$ such that

the mapping is surjective with $\zeta(0) = 0$, $\lim_{t\to\infty} \zeta(t) = \infty$, and

$$\gamma(\zeta) := \sup_{0\le t<s} \left|\log\left(\frac{\zeta(s)-\zeta(t)}{s-t}\right)\right| < \infty.$$

Similar to $D([0,\infty);\mathbb{R}^r)$, we also use the notation $D([0,T];\mathbb{S})$ to denote the D-space of functions that take values in $\mathbb{S}$.

Definition A.10 (Skorohod topology). For $\xi, \eta \in D([0,\infty);\mathbb{R}^r)$, the Skorohod topology $d(\cdot,\cdot)$ on $D([0,\infty);\mathbb{R}^r)$ is defined as

$$d(\xi,\eta) = \inf_{\zeta\in\mathbb{L}} \left\{\gamma(\zeta) \vee \int_0^\infty e^{-s} \sup_{t\ge 0} \left(1 \wedge |\xi(t\wedge s) - \eta(\zeta(t)\wedge s)|\right) ds\right\}.$$

Analogous definitions and results are available for $D([0,T];\mathbb{S})$; see Ethier and Kurtz [28] and Billingsley [8] for related references. Although we often work with $D([0,T];\mathbb{R}^r)$ in this book, the results are often stated with respect to the space $D([0,\infty);\mathbb{R}^r)$. This enables us to apply them to $t \in [0,T]$ for any $T > 0$.

Definition A.11 (Tightness). A family of probability measures $\mathcal{P}$ defined on a metric space $\mathbb{S}$ is tight if, for each $\delta > 0$, there exists a compact set $K_\delta \subset \mathbb{S}$ such that

$$\inf_{P\in\mathcal{P}} P(K_\delta) \ge 1 - \delta.$$

The notion of tightness is closely related to compactness. The following theorem, known as Prohorov's theorem, gives such an implication. A complete proof can be found in Ethier and Kurtz [28].

Theorem A.12 (Prohorov's theorem). *If $\mathcal{P}$ is tight, then $\mathcal{P}$ is relatively compact. That is, every sequence of elements in $\mathcal{P}$ contains a weakly convergent subsequence. If the underlying metric space is complete and separable, the tightness is equivalent to relative compactness.*

Although weak convergence techniques usually allow one to use weaker conditions and lead to a more general setup, it is often more convenient to work with probability one convergence for purely analytic reasons, however. The Skorohod representation provides us with such opportunities.

Theorem A.13 (The Skorohod representation; Ethier and Kurtz [28]). *Let X_k and X be random elements belonging to $D([0,\infty);\mathbb{R}^r)$ such that X_k converges weakly to X. Then there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{P})$ on which are defined random elements $\widetilde{X}_k$, $k = 1, 2, \ldots$, and $\widetilde{X}$ in $D([0,\infty);\mathbb{R}^r)$ such that for any Borel set B and all $k < \infty$,*

$$\widetilde{P}(\widetilde{X}_k \in B) = P(X_k \in B) \quad \text{and} \quad \widetilde{P}(\widetilde{X} \in B) = P(X \in B),$$

satisfying

$$\lim_{k\to\infty} \widetilde{X}_k = \widetilde{X} \qquad w.p.1.$$

In this book, when we use the Skorohod representation, with a slight abuse of notation, we often omit the tilde notation for convenience and notational simplicity.

Let $C([0,\infty);\mathbb{R}^r)$ be the space of $\mathbb{R}^r$-valued continuous functions that are equipped with the sup-norm topology, and C_0 be the set of real-valued continuous functions on $\mathbb{R}^r$ with compact support. Let C_0^l be the subset of C_0 functions that have continuous partial derivatives up to the order l.

Definition A.14. Let $\mathbb{S}$ be a metric space and A be a linear operator on $B(\mathbb{S})$ (the set of all Borel-measurable functions defined on $\mathbb{S}$). Let $X(\cdot) = \{X(t) : t \geq 0\}$ be a right-continuous process with values in $\mathbb{S}$ such that for each $f(\cdot)$ in the domain of A,

$$f(X(t)) - \int_0^t Af(X(s))ds$$

is a martingale with respect to the filtration $\sigma\{X(s) : s \leq t\}$. Then $X(\cdot)$ is called a *solution of the martingale problem* with operator A.

Theorem A.15 (Ethier and Kurtz [28, p. 174]). *A right-continuous process $X(t)$, $t \geq 0$, is a solution of the martingale problem for the operator A if and only if*

$$E\left(\prod_{j=1}^{i} h_j(X(t_j))\left(f(X(t_{i+1})) - f(X(t_i)) - \int_{t_i}^{t_{i+1}} Af(X(s))ds\right)\right) = 0$$

whenever $0 \leq t_1 < t_2 < \cdots < t_{i+1}$, $f(\cdot)$ in the domain of A, and $h_1, \ldots, h_i \in \mathcal{B}(\mathbb{S})$, the Borel field of $\mathbb{S}$.

Theorem A.16 (Uniqueness of martingale problems; Ethier and Kurtz [28, p. 184]). *Let $X(\cdot)$ and $Y(\cdot)$ be two stochastic processes whose paths are in $D([0,T];\mathbb{R}^r)$. Denote an infinitesimal generator by A. If, for any function $f \in A$ (the domain of A),*

$$f(X(t)) - f(X(0)) - \int_0^t Af(X(s))ds,\ t \geq 0,$$

and

$$f(Y(t)) - f(Y(0)) - \int_0^t Af(Y(s))ds,\ t \geq 0,$$

are martingales, and $X(t)$ and $Y(t)$ have the same distribution for each $t \geq 0$, $X(\cdot)$ and $Y(\cdot)$ have the same distribution on $D([0,\infty);\mathbb{R}^r)$.

A.4 Miscellany

Gronwall-Type Inequalities. Treating dynamic systems, Gronwall's inequality is used most often. The first inequality below is the Gronwall inequality, whereas the second one is the so-called generalized Gronwall inequality. Both of them can be found in Hale [40, p. 36].

Lemma A.17. *If $\gamma \in \mathbb{R}$, $\beta(t) \geq 0$, and $\varphi(t)$ are continuous real-valued functions for $a \leq t \leq b$, which satisfy*

$$\varphi(t) \leq \gamma + \int_a^t \beta(s)\varphi(s)ds, \ t \in [a, b],$$

then

$$\varphi(t) \leq \gamma \exp\left(\int_a^t \beta(s)ds\right), \ t \in [a, b].$$

Lemma A.18. *Suppose that $\varphi(\cdot)$ and $\gamma(\cdot)$ are real-valued continuous functions on $[a, b]$, that $\beta(t) \geq 0$ is integrable on $[a, b]$, and that*

$$\varphi(t) \leq \gamma(t) + \int_a^t \beta(s)\varphi(s)ds, \ t \in [a, b].$$

Then

$$\varphi(t) \leq \gamma(t) + \int_a^t \beta(s)\gamma(s) \exp\left(\int_s^t \beta(u)du\right) ds, \ t \in [a, b].$$

The following lemma is a discrete-time counterpart and is useful for establishing bounds in difference equations. For a proof, see Yin and Zhang [122, pp. 331–332].

Lemma A.19. *Let $\{\phi_k\}$ be a nonnegative sequence satisfying*

$$\phi_{k+1} \leq C_0 + \varepsilon C_1 \sum_{j=0}^{k} \phi_j, \quad k = 0, 1, 2, \ldots, T/\varepsilon, \tag{A.16}$$

for some positive constants C_0 and C_1, and a parameter $\varepsilon > 0$. Then, for $k = 0, 1, 2, \ldots, T/\varepsilon$,

$$\phi_k \leq C_0(1 + \varepsilon C_1)^{T/\varepsilon}.$$

Moreover,

$$\phi_k \leq C_0 \exp(C_1 T). \tag{A.17}$$

Remark A.20. Several points are worth noticing.

1. The inequality (A.16) may be written recursively as

$$\phi_{k+1} \le \phi_k + \varepsilon\phi_k.$$

The parameter $\varepsilon > 0$ above is known as a constant stepsize. Its introduction stems from many scenarios of recursive estimation.

2. There is also a version of the inequality corresponding to the generalized Gronwall's inequality in continuous time (Lemma A.18). We state it as follows. Suppose that $\{\phi_k\}$ is a nonnegative sequence of real numbers satisfying

$$\phi_{k+1} \le \psi_{k+1} + \varepsilon \sum_{j=0}^{k} C_j\phi_j, \quad k = 0, 1, 2, \ldots, T/\varepsilon,$$

for some $\psi_k \ge 0$ and $C_k \ge 0$ and a parameter $\varepsilon > 0$. Then

$$\phi_k \le \psi_k + \varepsilon \sum_{j=0}^{k-1} \prod_{i=j+1}^{k-1} (1 + \varepsilon C_i) C_j\psi_j, \quad k = 0, 1, 2, \ldots, T/\varepsilon.$$

The proof is a modification of Lemma A.19.

3. Often, one is interested in a particular form of the Gronwall's inequality, namely,

$$\phi_{k+1} \le C_0 + C_1 \sum_{j=0}^{k} \phi_j, \quad k = 0, 1, 2, \ldots, N. \tag{A.18}$$

That is, $\varepsilon = 1$ in (A.16). In this case, we obtain

$$\phi_k \le C_0(1 + C_1)^N.$$

Likewise, for the inequality

$$\phi_{k+1} \le \psi_{k+1} + \sum_{j=0}^{k} C_j\phi_j, \quad k = 0, 1, 2, \ldots, N,$$

we have

$$\phi_k \le \psi_k + \sum_{j=0}^{k-1} \prod_{i=j+1}^{k-1} (1 + C_i) C_j\psi_j, \quad k = 0, 1, 2, \ldots, N.$$

Borel–Cantelli Lemma. Let A_n be events (i.e., sets in $\mathcal{F}$) and suppose that

$$\sum_n P\{A_n\} < \infty.$$

Then the *Borel–Cantelli lemma* [11] states that for almost all ω, only finitely many of the events A_n will occur.

Chebyshev's Inequality (see [11]). Let X be a real-valued random variable. Then for any integer m and $\delta > 0$,

$$P\{|X| \geq \delta\} \leq \frac{E|X|^m}{\delta^m}.$$

Hölder's Inequality For an integer k, let X_i, $i \leq k$ be real-valued random variables. for positive p_i, $i \leq k$, satisfying $\sum_i 1/p_i = 1$,

$$E|X_1 \cdots X_k| \leq E^{1/p_1}|X_1|^{p_1} \cdots E^{1/p_k}|X_k|^{p_k}.$$

Cauchy–Schwarz Inequality. This is a special case of the Hölder inequality with $k = 2, p_1 = p_2 = 2$.

Inequality for Sums. There is an analogous inequality for sums. Let $X_{i,n}$ be real-valued random variables and let $a_n \geq 0$ with $\sum_n a_n < \infty$. Then

$$E\left|\sum_n a_n X_n^{\{1\}} \cdots X_n^{\{k\}}\right| \leq \left(E\sum_n a_n \left|X_n^{\{1\}}\right|^{p_1}\right)^{1/p_1} \cdots \left(E\sum_n a_n \left|X_n^{\{k\}}\right|^{p_k}\right)^{1/p_k}.$$

Jensen's Inequality. Let $f(\cdot)$ be a convex function and $\mathcal{G}$ a σ-algebra, and suppose that $E|X| < \infty$ and $E|f(X)| < \infty$. Then

$$Ef(X) \geq f(EX) \quad \text{or with conditioning,}$$

$$E_{\mathcal{G}} f(X) \geq f(E_{\mathcal{G}} X) \text{ w.p.1.}$$

References

[1] H. Abut, Ed., *Vector Quantization*, IEEE Press Selected Reprint Series, IEEE, Piscataway, 1990.

[2] J.C. Aguero, G.C. Goodwin, J.I. Yuz, System identification using quantized data, in *Proc. 46th IEEE Conf. Decision Control*, 4263–4268, 2007.

[3] E.W. Bai, An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear system, *Automatica*, **34** (1998), 333–338.

[4] E.W. Bai, A blind approach to the Hammerstein–Wiener model identification, *Automatica*, **38** (2002), 967–979.

[5] M. Benaïm, Dynamics of stochastic approximation algorithms, in *Lecture Notes in Mathematics: 1769: Séminaire de Probabilities*, 1–69, Springer-Verlag, New York, 1999.

[6] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.

[7] S. Billings, Identification of nonlinear systems–A survey, *Proc. IEE*, Part D, **127** (1980), 272–285.

[8] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[9] T. Bohlin, On the maximum likelihood method of identification, *IBM J. Res. Devel.*, **14** (1970), 41–51.

[10] A.D. Brailsford, M. Yussouff, and E.M. Logothetis, Theory of gas sensors, *Sensors and Actuators B*, **13** (1993), 135–138.

[11] L. Breiman, *Probability Theory*, SIAM, Philadelphia, 1992.

[12] J.-M. Brossier, Egalization Adaptive et Estimation de Phase: Application aux Communications Sous-Marines, Ph.D. thesis, Institut National Polytechnique de Grenoble, France, 1992.

[13] E.R. Caianiello and A. de Luca, Decision equation for binary systems: Application to neural behavior, *Kybernetik*, **3** (1966), 33–40.

[14] M. Casini, A. Garulli, and A. Vicino, Time complexity and input design in worst-case identification using binary sensors, in *Proc. 46th IEEE Conf. Decision Control*, 5528–5533, 2007.

[15] P. Celka, N.J. Bershad, and J.M. Vesin, Stochastic gradient identification of polynomial Wiener systems: Analysis and application, *IEEE Trans. Signal Process.*, **49** (2001), 301–313.

[16] H.F. Chen, Recursive identification for Wiener model with discontinuous piece-wise linear function, *IEEE Trans. Automat. Control*, **51** (2006), 390–400.

[17] H.F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.

[18] H.F. Chen and G. Yin, Asymptotic properties of sign algorithms for adaptive filtering, *IEEE Trans. Automat. Control*, **48** (2003), 1545–1556.

[19] Y.S. Chow and H. Teicher, *Probability Theory*, 3rd ed., Springer-Verlag, New York, 1997.

[20] K.L. Chung, *Markov Chains with Stationary Transition Probabilities*, 2nd ed., Springer-Verlag, New York, 1967.

[21] K.L. Chung, *A Course in Probability Theory*, John Wiley, New York, 1974.

[22] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley, New York, 1991.

[23] M.A. Dahleh, T. Theodosopoulos, and J.N. Tsitsiklis, The sample complexity of worst-case identification of FIR linear systems, *Sys. Control Lett.*, **20** (1993), 157–166.

[24] M.H.A. Davis, *Markov Models and Optimization*, Chapman & Hall, London, 1993.

[25] P.J. Davis, *Circulant Matrices*, 2nd ed., Chelsea, New York, 1994.

[26] J.L. Doob, *Stochastic Processes*, John Wiley, New York, 1990.

[27] C.R. Elvitch, W.A. Sethares, G.J. Rey, and C.R. Johnson Jr., Quiver diagrams and signed adaptive fiters, *IEEE Trans. Acoustics, Speech, Signal Process.*, **30** (1989), 227–236.

[28] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[29] E. Eweda, Convergence analysis of an adaptive filter equipped with the sign-sign algorithm, *IEEE Trans. Automat. Control*, **40** (1995), 1807–1811.

[30] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed., John Wiley, New York, 1968.

[31] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed., John Wiley, New York, 1971.

[32] A. Friedman, *Stochastic Differential Equations and Applications*, Academic Press, New York, 1975.

[33] A. Gersho, Adaptive filtering with binary reinforcement, *IEEE Trans. Information Theory*, **30** (1984), 191–199.

[34] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer, Norwell, 1992.

[35] I.I. Gihman and A.V. Skorohod, *Introduction to the Theory of Random Processes*, W.B. Saunders, Philadelphia, 1969.

[36] I.I. Gihman and A.V. Skorohod, *Stochastic Differential Equations*, Springer-Verlag, Berlin, 1972.

[37] E. Golub and V. Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1993.

[38] K. Gopalsamy and I.K.C. Leung, Convergence under dynamical thresholds with delays, *IEEE Trans. Neural Networks*, **8** (1997), 341–348.

[39] C.S. Güntürk, One-bit sigma-delta quantization with exponential accuracy, *Comm. Pure Appl. Math.*, **56** (2003), 1608–1630.

[40] J.K. Hale, *Ordinary Differential Equations*, 2nd ed., R.E. Krieger, Malabar, 1980.

[41] P. Hall and C.C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press, New York, 1980.

[42] A.L. Hodgkin and A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiology*, **117** (1952), 500–544.

[43] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985.

[44] X.L. Hu and H.-F. Chen, Strong consistence of recursive identification for Wiener systems, *Automatica*, **41** (2005), 1095–1916.

[45] I.W. Hunter and M.J. Korenberg, The identification of nonlinear biological systems: Wiener and Hammerstein cascade models, *Bio. Cybernetics*, **55** (1986), 135–144.

[46] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.

[47] S. Karlin and H.M. Taylor, *A First Course in Stochastic Processes*, 2nd ed., Academic Press, New York, 1975.

[48] S. Karlin and H.M. Taylor, *A Second Course in Stochastic Processes*, Academic Press, New York, 1981.

[49] W. Kim, K. Mechitov, J.Y. Choi, and S.K. Ham, On tracking objects with binary proximity sensors, in *Information Processing in Sensor Networks, Fourth International Symposium*, 301–308, 2005.

[50] A.N. Kolmogorov, On some asymptotic characteristics of completely bounded spaces, *Dokl. Akad. Nauk SSSR*, **108** (1956), 385–389.

[51] M.J. Korenberg and I.W. Hunter, Two methods for identifying Wiener cascades having noninvertible static nonlinearities, *Ann. Biomed. Eng.*, **27** (1998), 793–804.

[52] X. Koutsoukos, Estimation of hybrid systems using discrete sensors, in *Proc. 42nd IEEE Conf. Decision Control*, 155–160, 2003.

[53] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, 1984.

[54] H.J. Kushner and D.S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.

[55] H.J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2003.

[56] S.L. Lacy and D.S. Bernstein, Identification of FIR Wiener systems with unknown, noninvertible polynomial nonlinearities, in *Proc. Amer. Control Conf.*, 893–899, 2002.

[57] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.

[58] Z. Lin and C. Lu, *Limit Theory for Mixing Dependent Radnom Variables*, Kluwer, New York, 1996.

[59] R. Liptster, http://www.eng.tau.ac.il/˜liptser/list.html, *Lectures of Stochastic Processes*.

[60] R.S. Liptser and A.N. Shiryayev, *Statistics of Random Processes I & II*, Springer-Verlag, New York, 2001.

[61] X. Liu and A. Goldsmith, Wireless communication tradeoffs in distributed control, in *Proc. 42nd IEEE Conf. Decision Control*, 688–694, 2003.

[62] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, 1987.

[63] M. Loéve, *Probability Theory*, 4th ed., Springer-Verlag, New York, 1977.

[64] D.G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addision-Wesley, Reading, 1984.

[65] E. Lukacs, *Stochastic Convergence*, Academic Press, New York, 1975.

[66] M. Milanese and G. Belforte, Estimation theory and uncertainty intervals evaluation in the presence of unknown but bounded errors: Linear families of models and estimators, *IEEE Trans. Automat. Control*, **27** (1982), 408–414.

[67] M. Milanese and A. Vicino, Optimal estimation theory for dynamic systems with set membership uncertainty: An overview, *Automatica*, **27** (1991), 997–1009.

[68] M. Milanese and A. Vicino, Information-based complexity and nonparametric worst-case system identification, *J. Complexity*, **9** (1993), 427–446.

[69] K. Narendra and P. Gallman, An iterative method for the identification of nonlinear systems using a Hammerstein model, *IEEE Trans. Automat. Control*, **11** (1966), 546–550.

[70] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, 1965.

[71] B. Ninness and S. Gibson, Quantifying the accuracy of Hammerstein model estimation, *Automatica*, **38** (2002), 2037–2051.

[72] S.J. Norquay, A. Palazoglu, and J.A. Romagnoli, Application of Wiener model predictive control (WMPC) to pH neutralization experiment, *IEEE Trans. Control Sys. Tech.*, **7** (1999), 437–445.

[73] K. Pakdaman and C.P. Malta, A note on convergence under dynamical thresholds with delays, *IEEE Trans. Neural Networks*, **9** (1998), 231–233.

[74] A. Papoulis and S.U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed., McGraw-Hill, Boston, 2002.

[75] E. Pitman, *Some Basic Theory for Statistical Inference*, Chapman & Hall, London, 1979.

[76] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.

[77] K. Poolla and A. Tikku, On the time complexity of worst-case system identification, *IEEE Trans. Automat. Control*, **39** (1994), 944–950.

[78] V.K. Rohatgi, *An Introduction to Probability Theory and Mathematical Statistics*, John Wiley, New York, 1976.

[79] H.L. Royden, *Real Analysis*, 3rd ed., Macmillan, New York, 1988.

[80] K. Sayood, *Introduction to Data Compression*, 2nd ed., Morgan Kaufmann, San Francisco, 2000.

[81] A.M. Sayeed, A signal modeling framework for integrated design of sensor networks, in *IEEE Workshop Statistical Signal Processing*, 7, 2003.

[82] L. Schweibert and L.Y. Wang, Robust control and rate coordination for efficiency and fairness in ABR traffic with explicit rate marking, *Computer Comm.*, **24** (2001), 1329–1340.

[83] R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, John Wiley, New York, 1980.

[84] G.R. Shorack and J.A. Wellner, *Emperical Processes with Applications to Statistics*, John Wiley, New York, 1986.

[85] A.V. Skorohod, *Studies in the Theory of Random Processes*, Dover, New York, 1982.

[86] A.V. Skorohod, *Asymptotic Methods in the Theory of Stochastic Differential Equations*, Amer. Math. Society, Providence, 1989.

[87] W.F. Stout, *Almost Sure Convergence*, Academic Press, New York, 1974.

[88] D.W. Stroock, *Probability Theory, An Analytic View: Revised Edition*, Cambridge University Press, Cambridge, 1994.

[89] D.W. Stroock and S.R.S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, 1979.

[90] J. Sun, Y. Kim, and L.Y. Wang, HEGO signal processing and strategy adaptation for improved performance in lean burn engines with a lean NOx trap, *Int. J. Adaptive Control Signal Process.*, **18** (2004), 145–166.

[91] J. Sur and B.E. Paden, State observer for linear time-invariant systems with quantized output, *ASME J. Dynamic Syst., Measurement, Control*, **120** (1998), 423–426.

[92] J.F. Traub, G.W. Wasilkowski, and H. Wozniakowski, *Information-Based Complexity*, Academic Press, New York, 1988.

[93] D.C.N. Tse, M.A. Dahleh, and J.N. Tsitsiklis, Optimal asymptotic identification under bounded disturbances, *IEEE Trans. Automatic Control*, **38** (1993), 1176–1190.

[94] V.N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, New York, 1998.

[95] V.N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer-Verlag, New York, 2000.

[96] M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks*, 2nd ed., Springer, London, 2003.

[97] L.Y. Wang, Persistent identification of time varying systems, *IEEE Trans. Automatic Control*, **42** (1997), 66–82.

[98] L.Y. Wang, Y. Kim, and J. Sun, Prediction of oxygen storage capacity and stored NOx using HEGO sensor model for improved LNT control strategies, in *2002 ASME International Mechanical Engineering Congress and Exposition*, New Orleans, Nov. 17–22, 2002.

[99] L.Y. Wang, I.V. Kolmanovsky, and J. Sun, On-line identification lean NOx trap in GDI engines, in *Proc. 2000 Amer. Control Conf.*, 1006–1010, 2000.

[100] L.Y. Wang and L. Lin, On metric dimensions of discrete-time systems, *Syst. Control Lett.*, **19** (1992), 287–291.

[101] L.Y. Wang and G. Yin, Towards a harmonic blending of deterministic and stochastic frameworks in information processing, in *Robustness in Identification and Control*, Springer-Verlag Volume LNCS, 102–116, 1999.

[102] L.Y. Wang and G. Yin, Persistent identification of systems with unmodelled dynamics and exogenous disturbances, *IEEE Trans. Automat. Control*, **45** (2000), 1246–1256.

[103] L.Y. Wang and G. Yin, Closed-loop persistent identification of linear systems with unmodeled dynamics and stochastic disturbances, *Automatica*, **38** (2002), 1463–1474.

[104] L.Y. Wang and G. Yin, Asymptotically efficient parameter estimation using quantized output observations, *Automatica*, **43** (2007), 1178–1191.

[105] L.Y. Wang and G. Yin, Information characterization of communication channels for system identification, *J. Sys. Science Complexity*, **20** (2007), 251–261.

[106] L.Y. Wang and G. Yin, Quantized identification under dependent noise and Fisher information ratio for communication channels, *IEEE Transactions on Automatic Control*, **53** (2010), 674–690.

[107] L.Y. Wang, G. Yin, and H. Wang, Identification of Wiener models with anesthesia applications, *Int. J. Pure Appl. Math.*, **1** (2004), 35–61.

[108] L.Y. Wang, G. Yin, and J.F. Zhang, Joint identification of plant rational models and noise distribution functions using binary-valued observations, *Automatica*, **42** (2006), 535–547.

[109] L.Y. Wang, G. Yin, J.F. Zhang, and Y.L. Zhao, Space and time complexities and sensor threshold selection in quantized identification, *Automatica*, **44** (2008), 3014–3024.

[110] L.Y. Wang, G. Yin, Y. Zhao, and J.F. Zhang, Identification input design for consistent parameter estimation of linear systems with binary-valued output observations, *IEEE Trans. Automat. Control*, **53** (2008), 867–880.

[111] L.Y. Wang, J.F. Zhang, and G. Yin, System identification using binary sensors, *IEEE Trans. Automat. Control*, **48** (2003), 1892–1907.

[112] T. Wang, R.E. Soltis, E.M. Logothetis, J.A. Cook, and D.R. Hamburg, Static characteristics of ZrO_2 exhaust gas oxygen sensors, SAE paper 930352, 1993.

[113] T. Wigren, Convergence analysis of recursive identification algorithms based on the nonlinear Wiener model, *IEEE Trans. Automat. Control*, **39** (1994), 2191–2206.

[114] W.M. Wonham, Some applications of stochastic differential equations to optimal nonlinear filtering, *SIAM J. Control*, **2** (1965), 347–369.

[115] L. Xiao, M. Johansson, H. Hindi, S. Boyd, and A. Goldsmith, Joint optimization of communication rates and linear systems, *IEEE Trans. Automat. Control*, **48** (2003), 148–153.

[116] G. Yin, Convergence of a global stochastic optimization algorithm with partial step size restarting, *Adv. Appl. Probab.*, **32** (2000), 480–498.

[117] G. Yin, S. Kan, and L.Y. Wang, Identification error bounds and asymptotic distributions for systems with structural uncertainties, *J. Syst. Sci. Complexity*, **19** (2006), 22–35.

[118] G. Yin, S. Kan, L.Y. Wang, and C.Z. Xu, Identification of systems with regime switching and unmodelled dynamics, *IEEE Trans. Automat. Control*, **54** (2009), 34–47.

[119] G. Yin, V. Krishnamurthy, and C. Ion, Iterate-averaging sign algorithms for adaptive filtering with applications to blind multiuser detection, *IEEE Trans. Information Theory*, **49** (2003), 657–671.

[120] G. Yin, L.Y. Wang, and S. Kan, Tracking and identification of regime-switching systems using binary sensors, *Automatica*, **45** (2009), 944–955.

[121] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbations Approach*, Springer-Verlag, New York, 1998.

[122] G. Yin and Q. Zhang, *Discrete-Time Markov Chains: Two-Time-Scale Methods and Applications*, Springer, New York, 2005.

[123] G. Yin and C. Zhu, *Hybrid Switching Diffusions: Properties and Applications*, Springer, New York, 2010.

[124] S. Yuksel and T. Basar, Optimal signaling policies for decentralized multicontroller stabilizability over communication channels, *IEEE Trans. Automat. Control*, **52** (2007), 1969–1974.

[125] G. Zames, On the metric complexity of causal linear systems: ε-entropy and ε-dimension for continuous time, *IEEE Trans. Automat. Control*, **24** (1979), 222–230.

[126] G. Zames, L. Lin, and L.Y. Wang, Fast identification n-widths and uncertainty principles for LTI and slowly varying systems, *IEEE Trans. Automat. Control*, **39** (1984), 1827–1838.

[127] Y. Zhao, L.Y. Wang, G. Yin, and J.F. Zhang, Identification of Wiener systems with binary-valued output observations, *Automatica*, **43** (2007), 1752-1765.

[128] Y. Zhao, J.F. Zhang, L.Y. Wang, and G. Yin, Identification of Hammerstein systems with quantized observations, to appear in *SIAM J. Control Optim.*

Index